

# 000 SPARK: SYNERGISTIC POLICY AND REWARD CO- 001 002 EVOLVING FRAMEWORK 003 004

005 **Anonymous authors**

006 Paper under double-blind review

## 007 008 ABSTRACT 009

010 Recent Large Language Models (LLMs) and Large Vision-Language Models  
011 (LVLMs) increasingly use Reinforcement Learning (RL) for post-pretraining,  
012 such as RL with Verifiable Rewards (RLVR) for objective tasks and RL from  
013 Human Feedback (RLHF) for subjective tasks. However, RLHF incurs high costs and  
014 potential reward–policy mismatch due to reliance on human preferences, while  
015 RLVR still wastes supervision by discarding rollouts and correctness signals af-  
016 ter each update. To address these challenges, we introduce the **Synergistic Policy**  
017 **And Reward Co-Evolving Framework** (SPARK), an efficient, on-policy, and sta-  
018 ble method that builds on RLVR. Instead of discarding rollouts and correctness  
019 data, SPARK recycles this valuable information to simultaneously train the model  
020 itself as a generative reward model. This auxiliary training uses a mix of objec-  
021 tives, such as pointwise reward score, pairwise comparison, and evaluation condi-  
022 tioned on further-reflection responses, to teach the model to evaluate and improve  
023 its own responses. Our process eliminates the need for a separate reward model  
024 and costly human preference data. SPARK creates a positive co-evolving feed-  
025 back loop: improved reward accuracy yields better policy gradients, which in turn  
026 produce higher-quality rollouts that further refine the reward model. Our unified  
027 framework supports test-time scaling via self-reflection without external reward  
028 models and their associated costs. We show that SPARK achieves significant per-  
029 formance gains on multiple LLM and LVLM models and multiple reasoning, re-  
030 ward models, and general benchmarks. For example, SPARK-VL-7B achieves an  
031 average 9.7% gain on 7 reasoning benchmarks, 12.1% on 2 reward benchmarks,  
032 and 1.5% on 8 general benchmarks over the baselines, demonstrating robustness  
033 and broad generalization.

## 034 1 INTRODUCTION

035 Reinforcement learning (RL) is a standard step of post-pretraining improvement and alignment for  
036 Large Language Models (LLMs) and Large Vision-Language Models (LVLMs). In practice, current  
037 RL systems rely on two complementary routes: **(1)** RL with verifiable rewards (**RLVR**) (Lambert  
038 et al., 2024a; Guo et al., 2025; Team et al., 2025), which uses a verifier to address objective and ver-  
039 ifiable problems like math and code. **(2)** Reward-model-based pipelines such as RL from Human  
040 Feedback (**RLHF**) (Ouyang et al., 2022; Bai et al., 2022), which distill human or synthetic prefer-  
041 ences into a learned reward model to guide policy optimization on subjective tasks. These two RL  
042 stages have yielded significant gains in reasoning quality, safety, and truthfulness, and have become  
043 a cornerstone in modern LLM/LVLM training.

044 Despite impressive progress, current RL pipelines for LLMs/LVLMs still exhibit several limi-  
045 tations. Approaches based on verifiable rewards (RLVR) are effective only for tasks with ex-  
046 plicit verifiers, leaving open-ended objectives like helpfulness and safety unaddressed. Conversely,  
047 reward-model-based pipelines (RLHF) can handle subjective tasks with reward models (Su et al.,  
048 2025a) or LLM-as-a-judge (Zheng et al., 2023; Gunjal et al., 2025) but demand substantial and  
049 costly curated human preference data. Furthermore, training the reward model as a separate com-  
050 ponent causes it to lag the evolving policy, inducing reward–policy mismatch, reward hacking, and  
051 brittle generalization under out-of-distribution queries (Skalse et al., 2022; Gao et al., 2023). Finally,  
052 dependence on external reward models or judge models introduces significant latency and serving  
053 costs during both training and test-time scaling (Zhao et al., 2025).

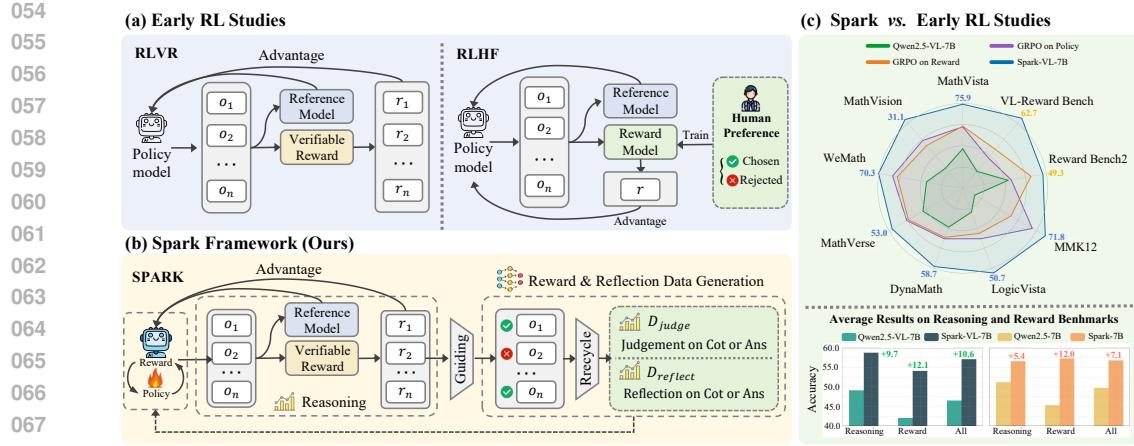


Figure 1: (a) Early studies of RL with Verifiable Rewards (RLVR) and RL from Human Feedback (RLHF) that rely on reward models. (b) We propose SPARK that recycles the rollouts from the RLVR, and further trains the model itself as a generative reward model. (c) SPARK consistently outperforms early RL approaches in both reasoning and reward model benchmarks.

To mitigate the limitations of early RL studies, such as costs of human preference labeling and deployment, we turn to an *internalized* source of supervision. Our method builds on RL with Verifiable Rewards (RLVR), where  $n$  candidate responses or rollouts  $\{o_1, o_2, \dots, o_n\}$  are generated, and score them against a ground-truth label to update the policy model (see Fig. 1 (a)). However, these valuable rollouts are typically **discarded** after this single use. Our key insight is to **recycle** the rollouts and correctness data to further train the model itself as a generative reward model simultaneously. We use the RLVR-derived correctness scores to train the model on a mix of objectives: a pointwise objective to determine if a response is correct, a pairwise objective to identify which response is better, and a reflection objective to learn how to fix an incorrect response to get the correct one.

The proposed *auxiliary* training paradigm for RLVR, the **Synergistic Policy And Reward Co-Evolving Framework (SPARK)**, enhances reward accuracy, yielding stronger policy gradients and improving the model’s reasoning abilities (see Fig. 1 (b)). We further use this internal judge for self-reflection at test time, extending alignment to tasks beyond strictly verifiable domains while retaining the robustness of verifiable feedback. SPARK has four advantages: (1) **Data- and compute-efficient**: no extra human preference data annotation or separate reward model training loop is required, as the signals come “for free” from RLVR training rollouts. (2) **On-policy and stable**: reward data are continually sampled from and calibrated to the model’s current behavior, reducing reward–policy mismatch. (3) **Co-evolving**: improved reward accuracy yields better gradients for the policy, which produces high-quality rollouts, further refining the reward. (4) **Unified development**: our framework enables RL training and test-time scaling, removing the dependency on an external reward model, and thereby saving GPU memory and reducing the communication overhead.

Our SPARK is applicable to both LLMs (e.g., Qwen2.5 (Yang et al., 2025a)) and LVLMs (e.g., Qwen2.5-VL (Bai et al., 2025)). As shown on Fig. 1 (c), SPARK achieves clear improvements on various mathematical reasoning and reward model benchmarks. For LVLMs, SPARK-VL-7B improves by **9.7%** on 7 reasoning and **12.1%** on 2 reward benchmarks, in addition to an average **1.5%** gain on 8 general benchmarks. These improvements are observed also with larger LVLM models (SPARK-VL-32B) and pure LLMs (SPARK-7B), demonstrating the robustness across different model scales and architectures.

Our key contributions are: (1) We introduce an efficient, on-policy, and stable framework SPARK that builds on RLVR but recycles the valuable rollouts that are typically discarded after policy updates. We use the RLVR-derived correctness scores to train the model itself to become a generative reward model, which eliminates the need for human preference data to train a separate, external reward model with additional development costs. (2) Our SPARK is designed as a **co-evolving** mechanism. Improved reward accuracy yields better gradients for the policy, which in turn produces higher-quality rollouts. These high-quality rollouts further refine the reward model, creating a positive feedback loop that leads to stronger overall performance and stability. (3) Extensive experiments show that SPARK achieves substantial improvements on multiple LVLM and LLM models. SPARK-

108 VL-7B achieves average **9.7%** gains on 7 reasoning benchmarks, **12.1%** on 2 reward benchmarks,  
 109 and **1.5%** on 8 general benchmarks, demonstrating the strong generalization ability.  
 110

## 111 2 RELATED WORKS

113 **Reinforcement Learning with Verifiable Reward.** Following the success of DeepSeek-R1 (Guo  
 114 et al., 2025), the GRPO (Shao et al., 2024) algorithm—driven by verifiable rewards—has demon-  
 115 strated strong potential across a variety of reasoning-intensive tasks, particularly in mathematics and  
 116 programming. Moreover, this RLVR paradigm has been successfully extended to a wide range of  
 117 domains, including perception (Zheng et al., 2025; Su et al., 2025b; Liu et al., 2025a; Peng et al.,  
 118 2025), agent (Jin et al., 2025; Liu et al., 2025b) and so on. In this work, we adopt a GRPO-based al-  
 119 gorithm to build our synergistic policy and reward co-evolving framework. Through reinforcement  
 120 learning, our framework jointly enhances the policy’s reasoning and the reward’s judging abilities  
 121 in a unified model, breaking the isolation between policy and reward models in prior approaches.

122 **Reinforcement Learning from Human Feedback.** Reinforcement Learning from Human Feed-  
 123 back (RLHF) optimizes policy models using human preference data. These data are either directly  
 124 collected from human annotations or generated by teacher models, and are typically used to first train  
 125 an independent reward model. The reward model then provides feedback signals that guide policy  
 126 optimization (Cai et al., 2024; Zhu et al., 2023; Zang et al., 2025; Kim et al., 2023; Yuan et al.,  
 127 2024; Lambert et al., 2024b; Ivison et al., 2023). However, a key limitation of existing paradigms  
 128 is that policy and reward models are usually developed in isolation, which restricts their interac-  
 129 tion and reduces the potential for mutual improvement. In this work, we instead treat policy and  
 130 reward as complementary capabilities, and introduce SPARK, a unified framework where the two  
 131 evolve jointly, reinforcing each other and ultimately achieving stronger overall performance. We  
 132 also discuss related work on self-reward and self-reflection; please refer to Appendix. A.3.

## 133 3 METHODS

135 In this section, we provide a detailed introduction to the SPARK approach. Specifically, Sec. 3.1  
 136 presents the SPARK training framework with verifiable reward, Sec. 3.2 outlines the on-policy re-  
 137 ward&reflection data generation of SPARK, and finally, Sec. 3.3 details the test-time scaling eval-  
 138 uation strategy used in SPARK.  
 139

### 140 3.1 SPARK TRAINING WITH VERIFIABLE REWARD

142 Fig. 2 (a) illustrates our Synergistic Policy and Reward Co-Evolving Framework. In contrast to prior  
 143 approaches, our method integrates the training of policy and reward into a unified framework, where  
 144 both components are optimized within a single model under the guidance of verifiable rewards. In  
 145 this section, we detail how SPARK employs verifiable rewards to guide optimization during training,  
 146 enabling the model to co-evolve its policy and reward capabilities. Through this process, the model  
 147 develops not only into a strong reasoning system but also into an effective reward model.

148 **Step 1: Sampling an answer group.** As shown in Fig. 2(a), given a Visual Question Answering  
 149 (VQA) sample  $d = (q, a, I)$ , or  $d = (q, a)$  in the case of a language-only LLM without visual input,  
 150 the model generates an answer group of size  $n$ , denoted as

$$152 G = \{(o_{\text{cot}}^i, o_{\text{ans}}^i)\}_{i=1}^n, \quad (1)$$

153 where  $q$  denotes the input prompt,  $a$  the ground-truth answer,  $I$  the input image,  $o_{\text{cot}}^i$  the  $i$ -th reason-  
 154 ing trace, and  $o_{\text{ans}}^i$  the corresponding final answer. To facilitate a clear separation between reasoning  
 155 and the final output, we design a prompt that requires the answer to be enclosed in  $\backslash \text{box}\{\}$ , as  
 156 illustrated in Appendix A.2.

157 **Step 2: Verifiable reward.** Each final answer is evaluated by a rule-based, verifiable reward:

$$159 \mathbb{R}(q, o) = \begin{cases} 1, & \text{if } o = a, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

161 For the  $i$ -th sample in the answer group  $G$ , we denote its reward as  $r^i = \mathbb{R}(q, o_{\text{ans}}^i)$ .

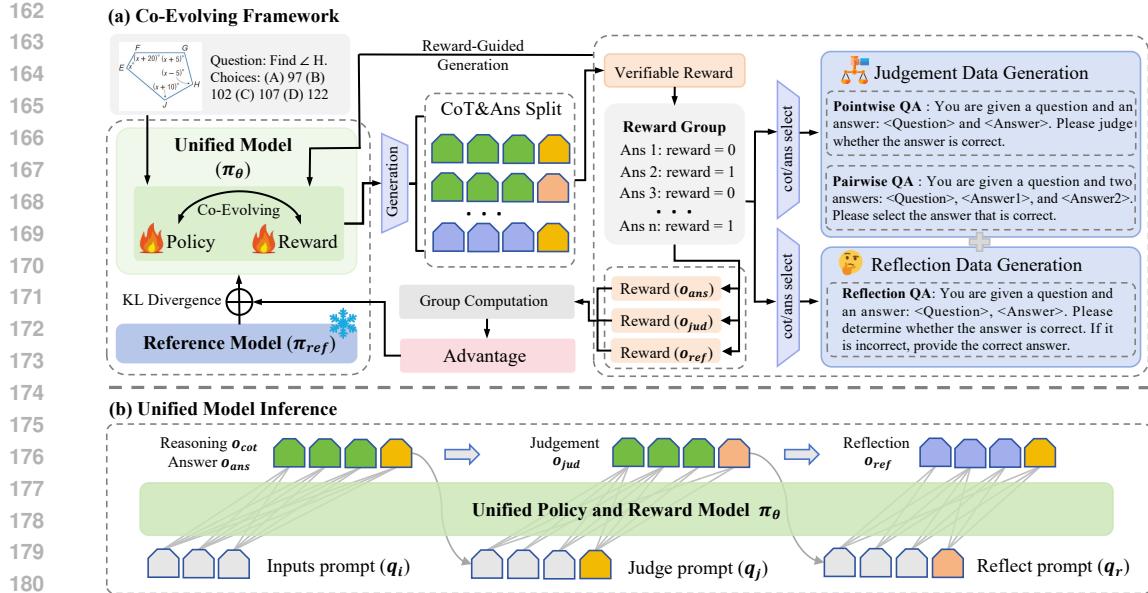


Figure 2: **SPARK Framework.** (a) **Training:** Our method recycles the valuable rollouts from verifiable reward-guided generation to simultaneously train a unified policy model  $\pi_\theta$ , also as a generative reward model. (b) **Inference:** at test time, the single unified model can handle reasoning, judgment, and reflection tasks for test-time scaling, eliminating the need for external reward or judge models.

**Step 3: Advantage computation.** Following GRPO-style normalization, we compute a standardized advantage for each candidate:

$$\bar{r} = \frac{1}{n} \sum_{j=1}^n r^j, \quad s = \sqrt{\frac{1}{n} \sum_{j=1}^n (r^j - \bar{r})^2 + \epsilon}, \quad A^i = \frac{r^i - \bar{r}}{s}, \quad (3)$$

where  $\bar{r}$  is the mean reward across the  $n$  candidates,  $s$  is the standard deviation with a small constant  $\epsilon > 0$  for numerical stability, and  $A^i$  is the normalized advantage used for policy gradient updates.

**Step 4: Overall objective.** The training objective maximizes the expected verifiable reward while regularizing the learned policy  $\pi_\theta$  towards a reference policy  $\pi_{\text{ref}}$ :

$$\mathbb{E}_{o \sim \pi_\theta(\cdot | q)} [\mathbb{R}(q, o)] - \lambda \text{KL}(\pi_\theta(\cdot | q) \| \pi_{\text{ref}}(\cdot | q)), \quad (4)$$

where  $\lambda$  is a hyperparameter controlling the KL-divergence. This formulation ensures that the optimization signal comes directly from task-defined correctness, enabling efficient and stable training whenever outputs are objectively verifiable.

### 3.2 SPARK ON-POLICY DATA GENERATION

Unlike early RL methods that optimize only the policy model, the advantage of SPARK lies in the joint co-evolution of policy and reward within a single model. To achieve this, SPARK generates reward and reflection data on-policy during the reasoning process, which requires neither additional preference annotations nor teacher models, making it highly efficient.

Beyond computing the advantage for gradient optimization in Eq. 3, the reward values  $r^i$  also guide the on-policy generation of reward and reflection data, as illustrated in Fig. 2. Specifically, we categorize the generated data into three forms: pointwise, pairwise, and reflection, each contributing to different aspects of judgment and self-reflection.

**Pointwise.** We reorganize  $(q, G)$  into binary judgment samples of the form

$$\mathbb{D}_{\text{pointwise}} = \{(q, o_{\text{ans}}^i, \mathbb{R}(q, o_{\text{ans}}^i))\}, \quad (5)$$

where the model is asked to determine whether a single candidate answer  $o_{\text{ans}}^i$  is correct. This formulation directly trains the model’s ability to judge the validity of individual answers.

216 **Pairwise.** We also construct comparison-style samples:  
 217

$$\mathbb{D}_{\text{pairwise}} = \left\{ (q, \mathbf{o}_{\text{ans}}^i, \mathbf{o}_{\text{ans}}^j, \mathbb{R}(q, \mathbf{o}_{\text{ans}}^i), \mathbb{R}(q, \mathbf{o}_{\text{ans}}^j)) \right\}, \quad (6)$$

220 where two candidate answers  $\mathbf{o}_{\text{ans}}^i$  and  $\mathbf{o}_{\text{ans}}^j$  are drawn from  $G$ , and the model must select the better  
 221 one. This encourages preference-style judgment, allowing the model to distinguish between rela-  
 222 tively stronger and weaker outputs. Notably, in both pointwise and pairwise settings,  $\mathbf{o}_{\text{ans}}$  can be  
 223 replaced with reasoning traces  $\mathbf{o}_{\text{cot}}$  to shift supervision toward intermediate steps.

224 **Reflection.** Finally, we construct reflection-style samples:  
 225

$$\mathbb{D}_{\text{reflect}} = \left\{ (q, \mathbf{o}_{\text{ans}}^i, \mathbb{R}(q, \mathbf{o}_{\text{ans}}^i)) \right\}, \quad (7)$$

226 where the model first verifies correctness and, if  $\mathbb{R}(q, \mathbf{o}_{\text{ans}}^i) = 0$ , the incorrect answer is then fed  
 227 back to the model for reflection and refinement. This process explicitly stimulates the model’s self-  
 228 reflection capability.  
 229

230 The combined dataset is then given by  
 231

$$\mathbb{D}_{\text{on-policy}} = \mathbb{D}_{\text{pointwise}} \cup \mathbb{D}_{\text{pairwise}} \cup \mathbb{D}_{\text{reflect}}, \quad (8)$$

232 which is used to further optimize the unified policy–reward model, strengthening both its judgment  
 233 and self-reflection abilities. Representative prompt templates used for data generation are provided  
 234 in Appendix A.2.  
 235

### 237 3.3 TEST TIME SCALING WITH SELF-REFLECTION

238 Benefiting from the co-evolution of policy and reward capabilities, SPARK functions not only as  
 239 a strong policy model but also as a strong reward model. The synergy between these two abilities  
 240 further enhances the model’s capacity for self-reflection, which proves especially valuable in the  
 241 context of test-time scaling (TTS).  
 242

243 As illustrated in Fig. 2(b), we adopt a TTS procedure to evaluate the model’s reasoning, judgment,  
 244 and reflection abilities. Formally, given a question  $q$  and image  $I$ , the model generates a candidate  
 245 answer at step  $t$  as

$$\mathbf{o}_t = \pi_{\theta}(q, I), \quad (9)$$

246 where  $\pi_{\theta}$  denotes the model, and  $\mathbf{o}_t = (c_t, a_t)$  consists of a reasoning chain  $c_t$  and a final prediction  
 247  $a_t$ . The model then assesses its own output through a judgment prompt:

$$\mathbf{r}_t = \pi_{\theta}(q, I, \text{judge}(c_t, a_t)), \quad \mathbf{r}_t \in \{0, 1\}, \quad (10)$$

248 where  $\text{judge}(c_t, a_t)$  instructs the model to verify whether  $(c_t, a_t)$  is correct.  
 249

250 Based on this evaluation, the model either accepts the result or performs iterative refinement:  
 251

$$\mathbf{o}_{t+1} = \begin{cases} \mathbf{o}_t, & \text{if } \mathbf{r}_t = 1, \\ \pi_{\theta}(q, I, \text{reflect}(c_t, a_t)), & \text{if } \mathbf{r}_t = 0, \end{cases} \quad (11)$$

252 where  $\text{reflect}(c_t, a_t)$  prompts the model to critique its prior reasoning and generate a revised solu-  
 253 tion. The process terminates once the model produces an answer it judges correct, and accuracy is  
 254 computed by comparing this final prediction with the ground truth.  
 255

## 259 4 EXPERIMENTS

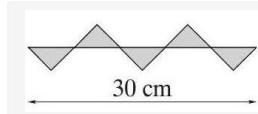
### 260 4.1 EXPERIMENTAL SETUP

261 **Benchmarks** To comprehensively evaluate the effectiveness of SPARK, we conduct experiments  
 262 on three categories of benchmarks: mathematical, reward-related, and general. For mathematical  
 263 benchmarks, we assess both multimodal and language-only reasoning using representative datasets  
 264 such as MathVista (Lu et al., 2023) and GSM8k (Cobbe et al., 2021). For reward-related evalua-  
 265 tion, we employ RewardBench2 (Malik et al., 2025) and VL-RewardBench (Li et al., 2025), including  
 266 their mathematical subsets for fine-grained analysis. For general capabilities, we test on widely used  
 267 multimodal benchmarks such as MMBench (Liu et al., 2023) and MMStar (Chen et al., 2024). For  
 268 the complete list of benchmarks used in this work, please refer to Sec. A.1.2.  
 269

270

271 **Table 1: Evaluation Results on SPARK-VL-7B.** We evaluate SPARK on multiple mathematical and reward-  
272 related benchmarks. Here, RB2 denotes RewardBench2, and VL-RB denotes VL-RewardBench.

Model	VLM Math Benchmark							Avg-M	Reward Benchmark			Avg-R	Avg-All		
	MathVista	MathVision	WeMath	MathVerse	DynaMath	LogicVista	MMK12		RB2	RB2-Math	VL-RB				
<i>Baseline</i>															
Qwen2.5-VL-7B	68.2	25.1	62.1	49.2	53.3	40.4	45.1	49.1	45.8	38.8	47.7	35.5	42.0	46.5	
OpenVLThinker-7B	70.2	25.3	64.3	47.9	-	-	60.6	-	48.5	37.3	33.2	33.1	38.0	-	
Vision-R1-7B	73.5	27.4	<b>75.0</b>	52.4	54.9	37.1	36.7	51.0	32.6	28.1	-	-	-	-	
R1-OneVision-7B	64.1	<u>29.9</u>	61.8	47.1	-	39.1	39.8	-	35.7	33.1	37.6	37.4	36.0	-	
VL-ReThinker-7B	73.7	28.4	67.9	<b>54.0</b>	<u>57.3</u>	42.7	64.9	55.6	42.3	35.0	47.1	32.5	39.2	49.6	
MM-Eureka-7B	73.0	26.9	66.1	50.3	56.9	<b>48.9</b>	64.5	55.2	44.9	<u>41.0</u>	48.8	36.8	42.9	50.7	
<i>Owen2.5-VL-7B + GRPO</i>														-	
+ Policy-Only	72.0	28.5	67.9	51.2	54.9	44.9	66.9	55.2	46.1	40.4	51.5	62.1	50.0	53.3	
+ Reward-Only	72.1	27.9	67.1	51.0	54.7	44.0	58.8	53.7	48.1	39.3	53.9	62.7	51.0	52.7	
+ Policy&Reward	74.2	28.9	<u>70.9</u>	51.3	56.3	46.2	<u>67.9</u>	<u>56.5</u>	48.9	<b>43.7</b>	<u>54.4</u>	<u>63.9</u>	<u>52.7</u>	<u>55.1</u>	-
<i>Ours</i>														-	
SPARK-VL-7B	<b>75.9</b>	<b>31.1</b>	70.3	<u>53.0</u>	<b>58.7</b>	<b>50.7</b>	<b>71.8</b>	<b>58.8</b>	<b>49.3</b>	39.2	<b>62.7</b>	<b>65.1</b>	<b>54.1</b>	<b>57.1</b>	
$\Delta$	<b>+7.7</b>	<b>+6.0</b>	<b>+8.2</b>	<b>+3.8</b>	<b>+5.4</b>	<b>+10.3</b>	<b>+26.7</b>	<b>+9.7</b>	<b>+3.5</b>	<b>+0.4</b>	<b>+15.0</b>	<b>+29.6</b>	<b>+12.1</b>	<b>+10.6</b>	



The figure in the drawing consists of five isosceles right triangles of the same size. Find the area (in  $\text{cm}^2$ ) of the shaded figure. Please first conduct reasoning, and then answer the question. Repeat the final answer using a `\boxed{}`.

#### Reasoning Trace:

To determine the area of the shaded figure, we start by analyzing the given information and the geometric configuration.....Since the triangles are arranged in a way that their hypotenuses form the base of the figure, the total length of the base is  $\sqrt{5x^2}$ . Thus, we have:  $5x = 30$ . Solving for  $x$ , we get:  $x = 6$ . The area of one isosceles right triangle is given by:  $\text{Area} = \frac{1}{2}x^2$ . Since there are five such triangles, the total area of the figure is:  $5 \times 36 = 90$ . However, the shaded figure consists of three of these triangles. Therefore, the area of the shaded figure is:  $3 \times 36 = 54$ . The final answer is:  $\boxed{54}$ .

#### Judgement:

The given answer is  $54 \text{ cm}^2$ , which is incorrect. Thus, the judgment is `\boxed{No}`.

#### Self-Reflection:

1. Since the figure consists of five isosceles right triangles of the same size, let's denote the length of the legs of each triangle as  $x$ . The hypotenuse of each triangle will be  $\sqrt{2}x$ . Since there are five such triangles, the total area of the figure is:  $5 \times \frac{1}{2}x^2 = 90$ . However, the figure is divided into two equal parts by the central line, and the shaded part is half of the total area of the figure. Therefore, the area of the shaded figure is:  $\frac{90}{2} = 45$ . The correct final answer is:  $\boxed{45}$ .

302 **Figure 3: Math Reasoning Case.** We illustrate the reasoning process of SPARK on a mathematical task,  
303 covering reasoning, judgment, and reflection. For brevity, parts of the content are omitted.  
304

**Baseline Methods** Our experiments are built upon the Qwen2.5-VL (Bai et al., 2025) and Qwen2.5 (Yang et al., 2025a) model series. For comparison, we include representative RL-based baselines such as VL-Rethinker (Wang et al., 2025), MM-Eureka (Meng et al., 2025), Vision-R1 (Huang et al., 2025), as well as the standard GRPO baselines, where *Policy-Only* and *Reward-Only* denote models trained to improve reasoning or judgment in isolation. Full details of the baseline models are provided in Sec. A.1.1 of the supplementary material.

312 

## 4.2 RESULTS ON MATHEMATICAL AND REWARD BENCHMARKS

**Results on SPARK-VL-7B.** Tab. 1 reports the results of SPARK-VL-7B on both mathematical and reward-related benchmarks. Compared with the Qwen2.5-VL-7B baseline, our model achieves consistent and substantial gains. Specifically, SPARK delivers an average improvement of **9.7%** on mathematical benchmarks and **12.1%** on reward benchmarks, resulting in an overall gain of **10.6%**.

The *Qwen2.5-VL-7B+GRPO* ablations further provide insight into the effect of different training signals. Training with *Policy-Only* data slightly favors mathematical reasoning, while *Reward-Only* training better enhances judgment ability. When both sources are combined (+*Policy&Reward*), the model surpasses either single-source variant, indicating the complementarity of policy and reward supervision. Building on this, SPARK-VL-7B advances through the co-evolution of policy and reward capabilities, and further incorporates reflection-driven data generation, which strengthens the integration between reasoning and judgment and brings additional performance improvements.

324

325 **Table 2: Evaluation Results on SPARK-7B.** We evaluate SPARK on multiple mathematical and reward-  
326 related benchmarks. Here, RB2 denotes RewardBench2.

Model	LLM Math Benchmark						Avg-M		Reward Benchmark		Avg-R		Avg-All	
	AIME24	AIME25	AMC23	GSM8k	Math-500	MMLU-STEM	RB2	RB2-Math						
<i>Baseline</i>														
Qwen2.5-7B	6.7	6.7	50.0	91.9	76.2	75.8	51.2	49.5	41.0	45.3	49.7			
Simple-RL-Zero	16.7	6.7	<b>62.5</b>	92.0	78.0	64.5	53.4	31.0	38.7	34.9	48.8			
Eurus-2-7B-PRIME	<b>26.7</b>	-	<u>57.8</u>	-	79.2	71.0	-	-	-	-	-			
Open-Reasoner-Zero-7B	13.3	-	47.0	-	79.2	70.3	-	31.4	37.3	34.4	-			
Qwen2.5-Math	13.3	-	50.6	-	<b>79.8</b>	60.2	-	-	-	-	-			
<i>Qwen2.5-7B + GRPO</i>														
+ Policy-Only	6.7	<b>6.7</b>	52.5	92.4	76.0	76.2	51.8	48.4	40.4	44.4	49.9			
+ Reward-Only	13.3	3.3	50.0	90.2	73.6	79.3	51.6	43.6	40.4	42.0	49.2			
+ Policy&Reward	16.7	3.3	50.0	<u>92.7</u>	76.6	<u>80.1</u>	53.2	48.3	<u>43.7</u>	46.0	<u>51.4</u>			
<i>Ours</i>														
SPARK-7B	<u>16.7</u>	<b>6.7</b>	<b>62.5</b>	<b>93.2</b>	<u>79.4</u>	<b>81.1</b>	<b>56.6</b>	<b>58.8</b>	<b>55.7</b>	<b>57.3</b>	<b>56.8</b>			
$\Delta$	<b>+10.0</b>	<b>+0.0</b>	<b>+12.5</b>	<b>+1.3</b>	<b>+3.2</b>	<b>+5.3</b>	<b>+5.4</b>	<b>+9.3</b>	<b>+14.7</b>	<b>+12.0</b>	<b>+7.1</b>			

339

340 **Table 3: Evaluation Results on SPARK-VL-32B.** We evaluate SPARK on multiple mathematical and reward-  
341 related benchmarks. Here, RB2 denotes RewardBench2, and VL-RB denotes VL-RewardBench.

Model	VLM Math Benchmark						Avg-M		Reward Benchmark		Avg-R		Avg-All	
	MathVista	MathVision	WeMath	MathVerse	DynaMath	LogicVista	MMK12		RB2	RB2-Math	VL-RB	VL-RB-Math		
<i>Baseline</i>														
Qwen2.5-VL-32B	74.7	38.4	69.1	48.5	61.3	55.4	52.9	57.2	<b>57.0</b>	59.6	<b>59.2</b>	56.0	<b>58.0</b>	57.5
VL-ReThinker-32B	<b>78.8</b>	<b>40.5</b>	76.7	56.9	<b>62.9</b>	51.8	<u>72.9</u>	62.9	53.9	51.8	49.9	23.5	44.8	56.3
Vision-R1-32B	73.2	35.7	<b>78.9</b>	53.8	<b>62.9</b>	54.2	55.2	59.1	53.4	<b>68.9</b>	-	-	-	-
MM-Eureka-32B	74.8	34.4	73.4	56.5	<u>62.1</u>	53.4	72.2	61.0	56.4	58.5	58.3	<u>56.6</u>	57.5	59.7
<i>Qwen2.5-VL-32B + GRPO</i>														
+ Policy&Reward	78.2	<b>40.2</b>	<u>77.1</u>	<u>57.7</u>	60.9	<u>57.4</u>	72.3	<u>63.4</u>	56.3	56.4	57.1	53.6	55.9	<b>60.7</b>
<i>Ours</i>														
SPARK-VL-32B	<b>79.1</b>	<b>40.2</b>	76.7	<b>59.2</b>	<b>62.9</b>	<b>59.4</b>	<b>77.4</b>	<b>65.0</b>	<b>60.3</b>	<b>62.7</b>	<b>61.4</b>	<b>59.6</b>	<b>61.0</b>	<b>63.5</b>
$\Delta$	<b>+4.4</b>	<b>+1.8</b>	<b>+7.6</b>	<b>+10.7</b>	<b>+1.6</b>	<b>+4.0</b>	<b>+24.5</b>	<b>+7.8</b>	<b>+3.3</b>	<b>+3.1</b>	<b>+2.2</b>	<b>+3.6</b>	<b>+3.0</b>	<b>+6.0</b>

355

356 Notably, from Tab. 1 we also observe that SPARK-VL-7B achieves significant improvements on both  
357 reward-related benchmarks, including RewardBench2 (**+3.5%**) and VL-RewardBench (**+15.0%**).  
358 Although all reward-related data generated during training are mathematics-specific, these two re-  
359 ward benchmarks span diverse domains. This indicates that SPARK is able to generalize its judg-  
360 ment ability beyond mathematics and perform strongly on broader tasks.

361

362 Overall, these results demonstrate that policy and reward are not in competition but instead mutually  
363 reinforcing. Joint optimization, when augmented with reflection, produces a synergistic effect that  
364 drives simultaneous and significant improvements in both reasoning and judgment. Representative  
365 reasoning cases on mathematical are illustrated in Fig. 3. More cases can be found in Appendix A.4.  
366

367

**Results on SPARK-7B** To further assess the generalizability of our approach, we conduct ad-  
368 dditional experiments on the LLM Qwen2.5-7B. As shown in Tab. 2, SPARK-7B achieves average  
369 improvements of **5.4%** on mathematical benchmarks, **12.0%** on reward benchmarks, and **7.1%**  
370 overall, demonstrating consistent gains across diverse evaluation settings.

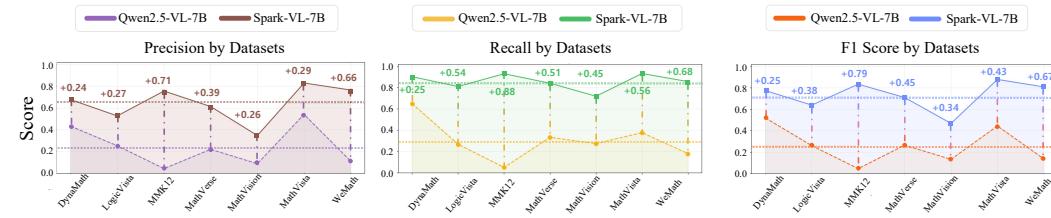
371

372 Models such as Simple-RL-Zero (Zeng et al., 2025) and the *Policy-Only* variant of Qwen2.5-7B  
373 show declines on reward benchmarks, which reflects the trade-off between reasoning and judgment  
374 ability. Moreover, we observe that *Reward-Only* GRPO training performs worse than *Policy-Only*  
375 training on both reasoning and reward tasks. This suggests that the model tends to overfit to re-  
376 ward signals, which in turn weakens its reasoning ability and prevents it from excelling in either  
377 skill. The *Policy&reward* GRPO variant partially alleviates this issue by training both capabilities  
378 simultaneously. Building on this, SPARK-7B further leverages on-policy co-evolution and reflection  
379 mechanisms to more effectively fuse and reinforce reasoning and judgment, ultimately achieving a  
380 substantial performance leap.

378

379  
380  
Table 4: **Evaluation Results on General Multimodal Benchmarks.** We select multiple general multimodal  
benchmarks to assess the generalization and robustness of our method.

Models	MMBench	MMStar	MMMU	MMVet	ScienceQA	POPE	SeedBench	RealWorldQA	Average
Qwen2.5-VL-7B	82.2	64.1	58.0	<u>69.7</u>	89.0	85.9	<u>77.0</u>	<u>68.4</u>	74.3
VL-Rethinker-7B	82.3	<u>65.4</u>	<b>59.0</b>	69.3	87.8	<u>86.1</u>	76.3	69.3	<u>74.4</u>
MM-Eureka-7B	<u>84.2</u>	65.3	57.8	68.9	88.8	85.8	76.8	65.1	74.1
<b>SPARK-VL-7B</b>	<b>84.4</b>	<b>67.3</b>	<b>58.7</b>	<b>71.5</b>	<b>90.8</b>	<b>88.2</b>	<b>77.2</b>	<b>68.5</b>	<b>75.8</b>
$\Delta$	<b>+2.2</b>	<b>+3.2</b>	<b>+0.7</b>	<b>+1.8</b>	<b>+1.8</b>	<b>+2.3</b>	<b>+0.2</b>	<b>+0.1</b>	<b>+1.5</b>

355  
356  
Figure 4: **Study on Model's Reward Accuracy.** We evaluate the model's judgment ability by measuring its  
357 accuracy in determining whether its own answers are correct.395  
396  
397  
**Results on Qwen2.5-VL-32B** To further examine the scalability of our approach, we conduct ex-  
398  
periments with Qwen2.5-VL-32B as the backbone. As shown in Tab. 3, SPARK-VL-32B achieves  
399  
improvements of **+7.8%** on mathematical benchmarks and **+3.0%** on reward benchmarks. These  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1299  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
138

**Table 5: Ablation on Test-Time Scaling.** We conduct ablation studies on the TTS. Specifically, we apply judge-reflection-based TTS to both the Qwen2.5-VL-7B model and the GRPO-trained model to evaluate its effectiveness.

Model	MathVista	MathVision	WeMath	MathVerse	DynaMath	LogicVista	MMK12	Average
<i>Baseline</i>								
Qwen2.5-VL-7B	68.2	25.1	62.1	49.2	53.3	40.4	45.1	49.1
Qwen2.5-VL-7B+TTS	68.5	18.4	24.1	29.9	52.7	43.1	42.8	39.9
<i>Qwen2.5-VL-7B + GRPO</i>								
+ Policy	72.0	28.5	67.9	51.2	54.9	44.9	66.9	55.2
+ Policy & TTS	73.1	28.7	67.5	51.9	57.7	48.9	68.9	56.6
SPARK-VL-7B	75.9	31.1	70.3	53.0	58.7	50.7	71.8	58.8

**Table 6: Ablation Study on Answer- vs. CoT-based Data Generation.** We generate data on-policy using either final answers, chains of thought (CoT), or a combination of both, and evaluate the impact of these strategies on performance.

Model	VLM Math Benchmark						Avg-M		Reward Benchmark				Avg-R Avg-All	
	MathVista	MathVision	WeMath	MathVerse	DynaMath	LogicVista	MMK12	RB2	RB2-Math	VL-RB	VL-RB-Math			
SPARK + Ans	73.8	29.1	69.0	51.9	58.1	46.2	69.9	56.9	47.2	41.8	57.9	63.9	52.7	55.3
SPARK + CoT	73.6	29.7	67.4	50.4	57.3	48.4	71.3	56.9	52.5	44.0	62.3	60.8	54.9	56.2
SPARK + Ans&CoT	75.9	31.1	70.3	53.0	58.7	50.7	71.8	58.8	49.3	39.2	62.7	65.1	54.1	57.1

stantially larger gains, benefiting from its inherently enhanced reasoning, judgment, and reflection capabilities.

**Ablation Study on Answer- vs. CoT-based Data Generation** In SPARK training, on-policy reward data can be generated either from final answers or from chains of thought (CoT). As shown in Tab. 6, we compare three settings: using only answer-based data (55.3), using only CoT-based data (56.2), and combining both (57.1). The results indicate that integrating both sources leads to the best performance, suggesting that the complementary nature of answer- and CoT-based data provides richer training signals and ultimately enhances the model’s learning effectiveness.

**Cost Analysis** Unlike traditional RM-based RL methods, SPARK removes the need for an additional reward model and extra preference data. As shown in Tab. 7, RM-based RL requires a separate RM training stage with large-scale human or teacher annotations, and during RL optimization, it repeatedly calls the RM for reward inference, which doubles GPU usage and slows training. In contrast, SPARK directly employs lightweight rule-based verifiable rewards to generate feedback on-policy, all and reward. This design not only reduces data and a more scalable training pipeline.

Table 7: Comparison between RM-based RL and SPARK.

	RM-based RL	SPARK (Ours)
Extra Data (Preference)	✓	✗
Extra RM Training	✓	✗
GPU Cost	~2×	1×
Reward Signal	RM inference	Rule-based signal
Efficiency	Slower	Faster

## 5 CONCLUSION

We presented the **Synergistic Policy And Reward Co-Evolving FrameworkK (SPARK)**, an efficient, on-policy, and stable paradigm that unifies policy optimization and reward modeling within a single model. Unlike prior RL pipelines that treat policy and reward in isolation or rely on costly external reward models, SPARK recycles RLVR rollouts into judgment and reflection objectives, enabling the model itself to function as both a strong policy and a generative reward model. This co-evolving mechanism establishes a positive feedback loop: improved reward accuracy enhances reasoning ability, while stronger reasoning in turn refines reward judgment, fostering self-reflection and stability. Demonstrating substantial improvements on mathematical, reward, and general benchmarks, SPARK offers a scalable and generalizable solution for RL, advancing a new paradigm where reasoning, judgment, and reflection evolve synergistically.

486 ETHICS STATEMENT  
487488 This work adheres to the ICLR Code of Ethics. Our study focuses on improving reinforcement  
489 learning methods for large language and vision–language models without involving human subjects  
490 or sensitive personal data. All datasets used in our experiments are publicly available and widely  
491 adopted in the community, and we carefully follow their respective licenses and intended usage. The  
492 proposed framework is designed to enhance reasoning, judgment, and reflection capabilities in AI  
493 models, and does not introduce new risks of bias, privacy violations, or harmful applications. We  
494 are committed to responsible research practices and transparent reporting.  
495496 REPRODUCIBILITY STATEMENT  
497498 We have made significant efforts to ensure the reproducibility of our work. The paper provides de-  
499 tailed descriptions of our framework, along with the formal definitions and equations. Experimental  
500 settings, datasets, and models are specified in the main text and supplementary materials. To fur-  
501 ther support reproducibility, we will release our training and evaluation code, along with processed  
502 datasets and prompts, upon acceptance. These resources allow researchers to replicate all reported  
503 results and extend our framework to new tasks.  
504505 REFERENCES  
506507 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang,  
508 Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*,  
509 2025.510 Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones,  
511 Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI:  
512 Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.513 Zheng Cai, Maosong Cao, Haojong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui  
514 Chen, Zhi Chen, Pei Chu, Xiao wen Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao,  
515 Jiaye Ge, Chenna Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu,  
516 Ting Huang, Tao Jiang, Penglong Jiao, Zhen Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li,  
517 Shuaibin Li, Wei Li, Yining Li, Hong wei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kui-  
518 Jie Liu, Xiaoran Liu, Chen Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang  
519 Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song,  
520 Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang,  
521 Rui Wang, Yudong Wang, Ziyi Wang, Xing Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao  
522 Xu, Rui Ze Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing  
523 Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang,  
524 Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao,  
525 Qian Zhao, Xiaomeng Zhao, Fen-Fang Zhou, Zaida Zhou, Jingming Zhuo, Yi-Ling Zou, Xipeng  
526 Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report. *ArXiv*, abs/2403.17297, 2024. URL  
527 <https://api.semanticscholar.org/CorpusID:268691939>.528 Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi  
529 Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language  
530 models? *arXiv preprint arXiv:2403.20330*, 2024.531 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,  
532 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to  
533 solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.534 Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu  
535 Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint*  
536 *arXiv:2502.01456*, 2025.538 Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker:  
539 An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv*  
540 *preprint arXiv:2503.17352*, 2025.

540 Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In  
 541 *ICML*, 2023.

542

543 Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Bing Liu, and Sean Hendryx. Rubrics as  
 544 rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*,  
 545 2025.

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

549

550 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and  
 551 Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint*  
 552 *arXiv:2009.03300*, 2020.

553 Haitao Hong, Yuchen Yan, Xingyu Wu, Guiyang Hou, Wenqi Zhang, Weiming Lu, Yongliang  
 554 Shen, and Jun Xiao. Cooper: Co-optimizing policy and reward models in reinforcement  
 555 learning for large language models. *ArXiv*, abs/2508.05613, 2025. URL <https://api.semanticscholar.org/CorpusID:280546264>.

556

557 Jingcheng Hu, Yinmin Zhang, Qi Han, Dixin Jiang, Xiangyu Zhang, and Heung-Yeung Shum.  
 558 Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base  
 559 model. *arXiv preprint arXiv:2503.24290*, 2025.

560

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

564

565 Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew E. Peters, Pradeep  
 566 Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hanna Hajishirzi. Camels in  
 567 a changing climate: Enhancing lm adaptation with tulu 2. *ArXiv*, abs/2311.10702, 2023. URL  
 568 <https://api.semanticscholar.org/CorpusID:265281298>.

569

570 Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and  
 571 Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement  
 572 learning. *arXiv preprint arXiv:2503.09516*, 2025.

573

574 Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeyonwoo  
 575 Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee,  
 576 Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. Solar 10.7b:  
 577 Scaling large language models with simple yet effective depth up-scaling. *ArXiv*, abs/2312.15166,  
 2023. URL <https://api.semanticscholar.org/CorpusID:266550918>.

578

579 Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman,  
 580 Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. T\ ulu 3: Pushing  
 581 frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024a.

582

583 Nathan Lambert, Jacob Daniel Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze  
 584 Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya  
 585 Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher  
 586 Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hanna Hajishirzi. T\ ulu 3:  
 587 Pushing frontiers in open language model post-training. *ArXiv*, abs/2411.15124,  
 2024b. URL <https://api.semanticscholar.org/CorpusID:274192505>.

588

589 Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu  
 590 Liu, Sujian Li, Bill Yuchen Lin, et al. Vl-rewardbench: A challenging benchmark for vision-  
 591 language generative reward models. In *Proceedings of the Computer Vision and Pattern Recognition  
 592 Conference*, pp. 24657–24668, 2025.

593

594 Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object  
 595 hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in  
 596 Natural Language Processing*, 2023.

594 Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan  
 595 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint*  
 596 *arXiv:2305.20050*, 2023.

597

598 Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan,  
 599 Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around  
 600 player? *arXiv preprint arXiv:2307.06281*, 2023.

601 Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi  
 602 Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025a.

603

604 Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong  
 605 Duan, Dahua Lin, and Jiaqi Wang. Visual agentic reinforcement fine-tuning. *arXiv preprint*  
 606 *arXiv:2505.14246*, 2025b.

607

608 Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord,  
 609 Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for  
 610 science question answering. *Advances in Neural Information Processing Systems*, 2022.

611

612 Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-  
 613 Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of  
 614 foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

615

616 Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou,  
 617 Nan Du, and Jia Li. S2r: Teaching llms to self-verify and self-correct via reinforcement  
 618 learning. *ArXiv*, abs/2502.12853, 2025. URL <https://api.semanticscholar.org/CorpusID:276421568>.

619

620 Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A Smith, Hannaneh Ha-  
 621 jishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. *arXiv*  
 622 *preprint arXiv:2506.01937*, 2025.

623

624 Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng  
 625 Han, Botian Shi, Wenhui Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multi-  
 626 modal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

627

628 Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong  
 629 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kel-  
 630 ton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan  
 631 Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback.  
 632 In *NeurIPS*, 2022.

633

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

637

638 Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma  
 639 GongQue, Shanglin Lei, Zhe Wei, MiaoXuan Zhang, et al. We-math: Does your large multi-  
 640 modal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*,  
 641 2024.

642

643 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,  
 644 Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical  
 645 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

646

647 Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and character-  
 648 izing reward gaming. In *NeurIPS*, 2022.

649

650 Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu.  
 651 Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *arXiv*  
 652 *preprint arXiv:2503.23829*, 2025a.

648 Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie  
 649 Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via  
 650 visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025b.

651 Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun  
 652 Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with  
 653 llms. *arXiv preprint arXiv:2501.12599*, 2025.

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

657 Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hong-  
 658 sheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in  
 659 Neural Information Processing Systems*, 37:95095–95169, 2024.

660 Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning  
 661 benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

662 An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jian-  
 663 hong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical  
 664 expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

665 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu,  
 666 Chang Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint  
 667 arXiv:2505.09388*, 2025a.

668 Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng  
 669 Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal rea-  
 670 soning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025b.

671 Jie Ying, Zihong Chen, Zhefan Wang, Wanli Jiang, Chenyang Wang, Zhonghang Yuan, Haoyang Su,  
 672 Huanjun Kong, Fan Yang, and Nanqing Dong. Seedbench: A multi-task benchmark for evaluating  
 673 large language models in seed science. *arXiv preprint arXiv:2505.13220*, 2025.

674 Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang,  
 675 and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv  
 676 preprint arXiv:2308.02490*, 2023.

677 Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason E.  
 678 Weston. Self-rewarding language models. *ArXiv*, abs/2401.10020, 2024. URL <https://api.semanticscholar.org/CorpusID:267035293>.

679 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens,  
 680 Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multi-  
 681 modal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF  
 682 Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

683 Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo  
 684 Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective  
 685 multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025.

686 E. Zelikman, Yuhuai Wu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning.  
 687 2022. URL <https://api.semanticscholar.org/CorpusID:247762790>.

688 Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-  
 689 zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv  
 690 preprint arXiv:2503.18892*, 2025.

691 Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou,  
 692 Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the  
 693 diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186.  
 694 Springer, 2024.

702 Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian,  
703 Biqing Qi, Xiu Li, et al. GenPRM: Scaling test-time compute of process reward models via  
704 generative reasoning. *arXiv preprint arXiv:2504.00891*, 2025.

705

706 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,  
707 Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and  
708 chatbot arena. In *NeurIPS*, 2023.

709

710 Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and  
711 Xing Yu. Deepeyes: Incentivizing” thinking with images” via reinforcement learning. *arXiv*  
712 preprint *arXiv:2505.14362*, 2025.

713 Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. Starling-7b: Improving llm  
714 helpfulness & harmlessness with rlaif, November 2023.

715 Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A  
716 dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language  
717 models. *arXiv preprint arXiv:2411.00836*, 2024.

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756 **A APPENDIX**  
757758 **USAGE OF LARGE LANGUAGE MODELS**  
759760 In this work, we made limited use of a large language model (LLM) as a writing assistant. Specifi-  
761 cally, the LLM was employed solely for language polishing and minor grammatical improvements  
762 of the manuscript text. It was not used for research ideation, experimental design, data analysis, or  
763 substantive content generation. All technical contributions, results, and conclusions were developed  
764 entirely by the authors.  
765766 **OUTLINE**  
767768 In the appendix, we provide additional supporting materials to facilitate a deeper understanding of  
769 our work. First, in Sec. A.1, we summarize all the models, datasets, and benchmarks used in the  
770 experiments of SPARK. Second, in Sec. A.2, we present the prompt templates employed in our  
771 study, including those used during evaluation as well as the on-policy prompt designs adopted by  
772 SPARK for reward and reflection data generation. Third, Sec. A.3 discusses related works on self-  
773 reward and self-reflection. Finally, in Sec. A.4, we provide several illustrative reasoning cases from  
774 SPARK on mathematical and reward benchmarks.  
775776 **A.1 MODEL, DATASET AND BENCHMARK STATISTIC**  
777778 **A.1.1 MODELS**779 In our study, we adopt the Qwen family of models as the backbone, including Qwen2.5-VL-7B (Bai  
780 et al., 2025), Qwen2.5-VL-32B (Bai et al., 2025), and Qwen2.5-7B (Yang et al., 2025a). Based  
781 on these backbones, we train three corresponding variants of our proposed framework: SPARK-  
782 VL-7B, SPARK-VL-32B, and SPARK-7B. These variants allow us to comprehensively evaluate the  
783 effectiveness of SPARK across both multimodal and text-only settings, as well as across different  
784 model scales.785 For comparison, we benchmark against a wide range of existing RL-based approaches. In the  
786 multimodal domain, we include VL-Rethinker-7B (Wang et al., 2025), MM-Eureka-7B (Meng  
787 et al., 2025), OpenVLThinker-7B (Deng et al., 2025), Vision-R1-7B (Huang et al., 2025), and  
788 R1-OneVision-7B (Yang et al., 2025b), all of which represent recent efforts to strengthen reason-  
789 ing capacity in vision-language models through reinforcement learning. In the language domain,  
790 we compare with Qwen2.5-Math-7B-Instruct (Yang et al., 2024), Simple-RL-Zero-7B (Zeng et al.,  
791 2025), Eurus-2-7B-PRIME (Cui et al., 2025), and Open-Reasoner-Zero-7B (Hu et al., 2025), which  
792 focus primarily on mathematical or general reasoning tasks within the NLP setting.793 To further examine the generalization ability of SPARK across different scales, we additionally  
794 evaluate against larger multimodal baselines, including VL-Rethinker-32B (Wang et al., 2025), MM-  
795 Eureka-32B (Meng et al., 2025), and Vision-R1-32B (Huang et al., 2025). These larger-scale models  
796 provide an important reference point to test whether the improvements introduced by SPARK are  
797 preserved when scaling up.798 A complete summary of all models used in our experiments, along with their categories (multimodal  
799 vs. language-only, 7B vs. 32B scale), is provided in Tab. 8 of the supplementary material.  
800801 **A.1.2 BENCHMARKS**  
802803 We evaluate SPARK across three major categories of benchmarks: mathematical reasoning, reward-  
804 related evaluation, and general multimodal understanding. This comprehensive setup ensures that  
805 our analysis covers not only specialized domains but also broader tasks.806 Mathematical Benchmarks. For multimodal mathematical reasoning, we adopt MathVista (Lu et al.,  
807 2023), MathVision (Wang et al., 2024), WeMath (Qiao et al., 2024), MathVerse (Zhang et al., 2024),  
808 DynaMath (Zou et al., 2024), LogicVista (Xiao et al., 2024), and MMK12 (Meng et al., 2025).  
809 For text-only reasoning, we include AIME24, AIME25, AMC23, GSM8k (Cobbe et al., 2021),  
Math500 (Lightman et al., 2023), and the MMLU\\_STEM (Hendrycks et al., 2020). These datasets

810  
 811 **Table 8: Model Sources.** We have compiled a list of all the models involved in the experiments along with  
 812 their parameter scale.

	Models	Parameter
<i>Baseline</i>		
	Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	7B
	Qwen2.5-VL-32B-Instruct (Bai et al., 2025)	32B
	Qwen2.5-7B-Instruct (Yang et al., 2025a)	7B
<i>Multimodal</i>		
	VL-Rethinker-7B (Wang et al., 2025)	7B
	VL-Rethinker-32B (Wang et al., 2025)	32B
	MM-Eureka-7B (Meng et al., 2025)	7B
	MM-Eureka-32B (Meng et al., 2025)	32B
	OpenVLThinker-7B (Deng et al., 2025)	7B
	Vision-R1-7B (Huang et al., 2025)	7B
	Vision-R1-32B (Huang et al., 2025)	32B
	R1-OneVision-7B (Yang et al., 2025b)	7B
<i>Language-Only</i>		
	Qwen2.5-Math-7B-Instruct (Yang et al., 2024)	7B
	Simple-RL-Zero-7B (Zeng et al., 2025)	7B
	Eurus-2-7B-PRIME (Cui et al., 2025)	7B
	Open-Reasoner-Zero-7B (Hu et al., 2025)	7B

832 **Table 9: Benchmark Sources.** We have included information for all the benchmarks tested in the paper in  
 833 the table.

	Setting	Models
834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863	836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863	MathVista (Lu et al., 2023) MathVision (Wang et al., 2024) WeMath (Qiao et al., 2024) MathVerse (Zhang et al., 2024) DynaMath (Zou et al., 2024) LogicVista (Xiao et al., 2024) MMK12 (Meng et al., 2025)
		RewardBench2 (Malik et al., 2025) VL-RewardBench (Li et al., 2025)
		844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863
836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863	836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863	844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863
		844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863

collectively test numerical reasoning, symbolic manipulation, logical deduction, and competition-style problem solving under both textual and multimodal settings.

Reward-related Benchmarks. We adopt RewardBench2 (RB2) (Malik et al., 2025) and VL-RewardBench (VL-RB) (Li et al., 2025) as representative reward evaluation benchmarks. Both focus on assessing models’ ability to judge correctness and quality of generated outputs. In addition, we separately report results on the mathematical subsets of these benchmarks, in order to analyze the interplay between reward judgment and mathematical reasoning.

General-purpose Multimodal Benchmarks. To assess generalization and robustness beyond math and reward domains, we evaluate on a diverse set of multimodal benchmarks, including MMMU (Yue et al., 2024), MMVet (Yu et al., 2023), MMBench (Liu et al., 2023), MMStar (Chen et al., 2024), POPE (Li et al., 2023), ScienceQA (Lu et al., 2022), SeedBench (Ying et al., 2025),

864  
865  
866  
867  
868  
869  
870

## Prompt for Mathematical Evaluation

"Hint: Please answer the question requiring an integer answer and provide the final value, e.g., 1, 2, 3, at the end.\nQuestion:  
How many objects are preferred by more than 90 percent of people in at least one category?  
Please first conduct reasoning, and then answer the question. Repeat the final answer using a '\\boxed{}'."

871

Figure 5: Mathematical Prompt. Prompt suffix used for mathematical benchmark evaluation.

872

873

## Prompt for VL-RewardBench Evaluation

You are a highly capable multimodal AI assistant tasked with evaluating answers to visual questions. Please analyze the following image and question, then determine which of the two provided answers is better.

874

Question: {query}

875

Answer 1: {answers[0]}

876

Answer 2: {answers[1]}

877

Please evaluate both answers based on the following criteria:

1. Accuracy: How well does the answer align with the visual information in the image?
2. Completeness: Does the answer fully address all aspects of the question?
3. Clarity: Is the answer easy to understand and well-articulated?
4. Relevance: Does the answer directly relate to the question and the image?

878

After your evaluation, please:

879

1. Explain your reasoning for each criterion.
2. Provide an overall judgment on which answer is better (Answer 1 or Answer 2). For example: Overall Judgment: Answer X is better.

880

Your response should be structured and detailed, demonstrating your understanding of both the visual and textual elements of the task.

881

882

883

Figure 6: Prompt for VL-RewardBench. In the figure, the placeholders *query* and *answer* should be replaced with the specific content of each task.

884

885

and RealWorldQA. These benchmarks cover a wide spectrum of multimodal reasoning and understanding tasks, ranging from knowledge-intensive QA to real-world perception challenges.

886

A complete statistical summary of all benchmarks is provided in Tab. 9.

887

888

## A.1.3 TRAINING DATA PREPARATION

889

Previous approaches that aimed to train a model's judgment or reflection capabilities typically relied on either reward data generated by a teacher model or manually annotated reflection traces. In contrast, SPARK is able to generate the required reward and reflection training data on-policy during policy's RFT, guided by verifiable reward signals. This approach not only greatly reduces the cost of data collection, but also ensures that the generated data evolves alongside model optimization and remains consistently aligned with the model's current policy distribution.

890

891

892

Leveraging SPARK's ability to generate reward and reflection data on-policy, we only need to collect VQA triples consisting of images, questions, and answers. In our experiments, 19k randomly sampled instances from ViRL-39k (Wang et al., 2025) are used to train SPARK-VL-7B, while 24k difficulty-filtered instances from the same dataset are used for SPARK-VL-32B. For the language-only variant SPARK-7B, we employ the Simple-RL-Zero-25k dataset (Zeng et al., 2025). All data are represented in the form of  $(q, a, I)$ , where  $q$  denotes the question,  $a$  the ground-truth answer, and  $I$  the corresponding image. Notably, this setup requires no manually annotated reward data, reflection traces, or judgment-oriented CoT trajectories.

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

### Prompt for On-Policy Data Generation

You're given a question and two possible answers. (A) and (B) may differ in correctness. Please analyze both answers carefully and decide which one is correct. After reasoning, restate your final choice using  $\boxed{\text{}}.$  \n\n<Questions>: \n{question}\n\n(A) {answer\_a}\n(B) {answer\_b}\n<Judgement>:

You are a specialist in answer evaluation. For the given question and answer, perform logical reasoning to assess the answer's correctness. Then state your judgment as 'Yes' or 'No'. Finally, repeat your judgment inside  $\boxed{\text{}}.$  \n\n<Questions>: \n{question}\n\n<Answers>: \n{answer}\n<Judgements>:

You are given a question and a reasoning process (CoT). Your task is to judge whether both the reasoning and the final answer are correct. If correct, summarize the logic and repeat the answer. If not, provide a correct reasoning process and answer. \n\nQuestion: \n{question}\nProvided CoT: \n{cot}\nPlease first conduct reasoning, and then answer the question. Repeat the final answer using a ' $\boxed{\text{}}$ '.

You are given a question and a proposed answer. Judge whether the answer is correct. If it is, briefly confirm it. If it is incorrect, provide correct reasoning and the correct answer. \n\nQuestion: \n{question}\nAnswer: \n{answer}\nPlease first conduct reasoning, and then answer the question. Repeat the final answer using a ' $\boxed{\text{}}$ '.

Figure 7: **On-Policy Data Generation Prompts.** The figure illustrates four different prompt templates used for generating reward and reflection data.

## A.2 PROMPTS

**Evaluation Prompts** During dataset evaluation, we appended an additional prompt at the end of each mathematical problem to facilitate the separation of reasoning steps from the final answer. To avoid over-constraining the model with rigid output formats (e.g., <think><answer>), we instead instructed the model to enclose the final answer within  $\boxed{\text{}}$  after completing its reasoning process. A concrete example is provided in Fig. 5.

For reward-related benchmarks, we followed the official evaluation prompts provided by each benchmark. For instance, in VL-RewardBench (Li et al., 2025), we adopted the original prompt format as illustrated in Fig. 6.

**On-Policy Data Generation Prompts** A key step in the policy–reward co-evolving training of SPARK is the on-policy generation of reward and reflection data. During GRPO optimization, the reward signals serve two purposes: on the one hand, they are used to compute the advantage for updating model parameters; on the other hand, they guide the construction of reward and reflection data. Specifically, chain of thought (CoT) or final answers filtered by reward values are wrapped with carefully designed prompts and reorganized into new training samples. These samples further enhance the model's judgment and reflection abilities. Examples of the prompts used for reward and reflection data generation are shown in Fig. 7.

## A.3 RELATED WORKS

**Self-Reward and Self-Reflection.** Early studies have explored incorporating *self-reward* and *self-reflection* into supervised fine-tuning (SFT) pipelines by generating reward or reflection data to enhance reasoning abilities. For example, STaR (Zelikman et al., 2022) iteratively generates chain-of-thought traces to improve its own reasoning capability. S2R (Ma et al., 2025) employs pre-annotated self-verification and self-correction data for both SFT and RL training. COOPER (Hong et al., 2025) leverages an external assistant to generate preference data, which are then used to train a reward model. While these approaches demonstrate the potential of self-reward and self-reflection for improving reasoning, they still rely on either external annotation data or independently trained reward models. In contrast, our method is the first to unify policy and reward capabilities within a single model by optimizing the GRPO framework. This co-evolving design breaks the conventional paradigm of separately trained reward models, enabling policy and reward to mutually reinforce each other, and integrates reasoning, judgment, and self-reflection into a unified process—without the need for preference annotation or external reward modeling.

972

973

974

975

976

977

978



You are a highly capable multimodal AI assistant tasked with evaluating answers to visual questions. Please analyze the following image and question, then determine which of the two provided answers is better.....

**Choice:**

[“A kitchen counter with a sink, blue plate, and knife block, along with cabinets below and above it; a dishwasher at the base of the cabinetry; a window with curtains on the left side; a towel rack on the wall to the right of the sink; a bottle on the counter; and a silver exhaust hood above the stove.”, “A clean and well-organized kitchen showcases a stainless steel gas cooktop, positioned centrally on the counter. Nearby, a stack of plates rests on a shelf above the dishwasher. The wall-mounted air vent is to the left of the sink and gas stove, while the paper towel holder containing white paper towels sits on the right, closer to the dishwasher. A chopping board with a knife lies on the counter, next to a green coffee mug and a bottle, both placed near the edge of the counter. The window’s view reveals a glimpse of trees in the distance.”]

**Judgement:**

To evaluate the answers, let's break down each criterion:  
 1. **Accuracy**: This answer accurately describes the kitchen counter with a sink, blue plate, and knife block. It mentions cabinets, a dishwasher, a window, a towel rack, a bottle on the counter, and a silver exhaust hood. However, it does not mention the presence of a stove or a gas cooktop, which is a significant omission.  
 2. **Relevance**: This answer includes a stainless steel gas cooktop, a stack of plates, a wall-mounted air vent, a paper towel holder, a chopping board, a green coffee mug, and a bottle. It also correctly identifies the window and its view. However, it does not mention the sink, the blue plate, or the knife block, which are present in the image.

Both answers are relevant to the question and the image, but Answer 1 is more accurate and complete.

**Overall Judgment**: **Answer 1 is better.**

The reasoning is that Answer 1 is more accurate and complete, providing a clearer and more detailed description of the kitchen's contents and layout without including any inaccuracies. Answer 2, while attempting to be more descriptive, introduces errors and omissions that do not align with the visual information in the image.

979

980

981

982

983

984

985

986

Figure 8: **Reward Reasoning Case.** The example is taken from VL-RewardBench (Li et al., 2025).

#### A.4 CASE STUDY

In Fig. 9, we present several reasoning cases on mathematical problems, which provide an intuitive demonstration of SPARK’s integrated capabilities in reasoning, judgment, and reflection. The selected examples highlight scenarios where the model engages in both reasoning and self-judgment.

Furthermore, in Fig. 8, Fig. 10 and Fig. 11, we showcase cases from VL-RewardBench. These examples demonstrate that the judgment ability acquired by our method in mathematical tasks can directly transfer and generalize to broader visual domains.

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

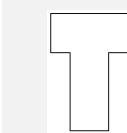
1075

1076

1077

1078

1079

**Prompt:**

The letter T is formed by placing two  $2 \times 4$  inch rectangles next to each other, as shown. What is the perimeter of the T, in inches?

Please first conduct reasoning, and then answer the question. Repeat the final answer using a '`\boxed{}`'.

**<Reasoning>**

To determine the perimeter of the letter T formed by placing two  $2 \times 4$  inch rectangles next to each other, we need to visualize the shape and calculate the total perimeter step-by-step.

\*\*Visualize the T Shape:\*\* - When two  $2 \times 4$  inch rectangles are placed next to each other, they form a T shape. - The T shape has a vertical rectangle on the right and a horizontal rectangle on top. - The vertical rectangle is 4 inches wide and 2 inches tall. - The horizontal rectangle is 2 inches wide and 4 inches tall.

\*\*Calculate the Perimeter:\*\* - The perimeter of a shape is the sum of the lengths of all its sides. - For the T shape, the perimeter is the sum of the lengths of the outer edges. - The T shape has two vertical sides of 4 inches each, two horizontal sides of 2 inches each, and one additional side that is the overlap of the two rectangles, which is 2 inches. - Therefore, the perimeter is calculated as follows:

- Two vertical sides:  $2 \times 4 = 8$  inches

- Two horizontal sides:  $2 \times 2 = 4$  inches

- One additional side (overlap): 2 inches

- Adding these together gives:  $8 + 4 + 2 = 14$  inches

**<Judgement>**

The given answer is 16 inches. The answer is incorrect.

**<Reflection>**

To determine the perimeter of the letter T formed by placing two  $2 \times 4$  inch rectangles next to each other, we need to calculate the total perimeter of the resulting shape.

\*\*Identify the dimensions of the T-shaped figure:\*\* - The T-shaped figure consists of two  $2 \times 4$  inch rectangles placed one above the other. - The top rectangle has a length of 4 inches and a width of 2 inches. - The bottom rectangle also has a length of 4 inches and a width of 2 inches.

\*\*Calculate the perimeter:\*\* - When the two rectangles are placed next to each other, the top rectangle's right side and the bottom rectangle's left side will be aligned, forming a straight line. - The perimeter of the T-shaped figure will be the sum of the outer edges of the two rectangles, minus the overlapping edges. - The total perimeter can be calculated as follows:

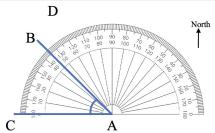
- The top rectangle contributes 2 lengths (4 inches each) and 2 widths (2 inches each) to the perimeter.

- The bottom rectangle contributes 2 lengths (4 inches each) and 2 widths (2 inches each) to the perimeter.

- However, the overlapping edges (the 4-inch sides of the rectangles) are not part of the perimeter of the T-shaped figure.

- Therefore, the perimeter is  $2 \times 4 + 2 \times 2 = 12$  inches.

The correct final answer is `\boxed{12}`.

**Prompt:**

As shown in the diagram, there are several points around the protractor. Which point is located to the northwest of point A?

Please first conduct reasoning, and then answer the question. Repeat the final answer using a '`\boxed{}`'.

**<Reasoning>**

To determine which point is located to the northwest of point A, we need to understand the direction indicated by the protractor. The protractor is oriented with the north direction at the top, which means that angles measured clockwise from the north direction will give us the directions.

1. Identify the direction of point B relative to point A: - Point B is located at an angle of 30 degrees clockwise from the north direction.

- Therefore, point B is located 30 degrees east of north.

2. Identify the direction of point C relative to point A: - Point C is located at an angle of 90 degrees clockwise from the north direction.

- Therefore, point C is directly north of point A.

3. Identify the direction of point D relative to point A: - Point D is located at an angle of 45 degrees clockwise from the north direction.

- Therefore, point D is located 45 degrees north of east.

Since the northwest direction is 45 degrees north of west, and point D is 45 degrees north of east, point D is not in the northwest direction. Point B is 30 degrees east of north, which is not northwest. Point C is directly north, which is not northwest. Therefore, none of the points are located to the northwest of point A.

**<Judgement>**

The given answer D is incorrect.

**<Reflection>**

To determine which point is located to the northwest of point A, we need to understand the direction indicated by the protractor and the compass rose.

1. The protractor is oriented with the top of the protractor aligned with the north direction. This means that the angles measured from the north direction will give us the directions of the points relative to north.

2. Point A is the origin of the protractor, and we need to find a point that is northwest of A. Northwest is a direction that is 45 degrees west of north.

3. Looking at the protractor, we see that point B is located at an angle of 45 degrees from the north direction. This means that point B is indeed northwest of point A.

4. Point C is located at an angle of 90 degrees from the north direction.

5. Point D is located at an angle of 30 degrees from the north direction.

Since point B is the only point that is northwest of point A, the correct answer is `\boxed{B}`.

Figure 9: Math Inference Case.

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

**Prompt:**

You are a highly capable multimodal AI assistant tasked with evaluating answers to visual questions. Please analyze the following image and question, then determine which of the two provided answers is better.

**Question:**

How do the elements in the image relate to each other in terms of positioning or composition?

**Answer 1:**

"In the image, a child is reaching out and attempting to catch a frisbee. The child's positioning suggests that they are standing on a sidewalk, while their arm extends out in an attempt to grasp the flying disc....In addition to the child and frisbee, there is a person near the end of the sidewalk, possibly observing or involved in the play. There is also a car parked further away from the main action, providing context of the urban environment surrounding the scene."

**Answer 2:**

"In the image, a young boy wearing blue jeans and a black jacket is actively engaged in catching a red frisbee. He is positioned on a sidewalk that stretches across the scene, with another person walking further down the pathway. .... A fire hydrant can also be seen near one of the parked cars, indicating the presence of urban infrastructure. Additionally, there's a trash bag placed behind the first car on the right side of the image. Overall, these elements come together to create an active and lively street scene centered around the boy playing frisbee."

Please evaluate both answers based on the following criteria:

1. Accuracy: How well does the answer align with the visual information in the image?
2. Completeness: Does the answer fully address all aspects of the question?
3. Clarity: Is the answer easy to understand and well-articulated?
4. Relevance: Does the answer directly relate to the question and the image?

After your evaluation, please:

1. Explain your reasoning for each criterion.
2. Provide an overall judgment on which answer is better (Answer 1 or Answer 2). For example: Overall Judgment: Answer X is better.

Your response should be structured and detailed, demonstrating your understanding of both the visual and textual elements of the task.

**Spark Judgement:**

### Evaluation Criteria\n\n

#### \*\*1. Accuracy\*\*\n- \*\*Answer 1\*\*: \n - The description of the child attempting to catch a frisbee is accurate.\n - The mention of a person near the end of the sidewalk is somewhat speculative, as the image does not clearly show another person walking.\n - The car parked further away is accurately described, providing context for the urban environment.\n- \*\*Answer 2\*\*: \n - The description of the child (young boy in blue jeans and a black jacket) is accurate.\n - The frisbee is correctly identified as red and positioned above the child's head.\n - The sidewalk, another person walking further down the pathway, and the parked cars are accurately described.\n - The mention of a fire hydrant and a trash bag introduces elements not clearly visible in the image, which reduces accuracy.\n\n\*\*Reasoning\*\*: Answer 1 is slightly more accurate because it avoids introducing elements (like the fire hydrant and trash bag) that are not clearly visible in the image. However, both answers are generally accurate in describing the main elements.\n\n

#### \*\*2. Completeness\*\*\n- \*\*Answer 1\*\*: \n - Focuses on the child, the frisbee, and the urban environment (sidewalk, car).\n - Mentions a person near the end of the sidewalk, which adds context but is not explicitly shown in the image.\n - Does not mention details like the child's clothing or the color of the frisbee, which are visible in the image.\n- \*\*Answer 2\*\*: \n - Provides detailed descriptions of the child's clothing, the frisbee's color, and the surrounding environment (sidewalk, cars, fire hydrant, trash bag).\n - Includes more specific details about the scene, such as the parallel street and the positioning of objects.\n - While some elements (fire hydrant, trash bag) are not clearly visible, the answer is more comprehensive in its description.\n\n\*\*Reasoning\*\*: Answer 2 is more complete as it provides detailed descriptions of the child, the frisbee, and the surrounding environment. However, the inclusion of elements not clearly visible in the image slightly detracts from its completeness.\n\n

#### \*\*3. Clarity\*\*\n- \*\*Answer 1\*\*: \n - The language is clear and easy to understand.\n - The description is well-structured, focusing on the child's action and the urban setting.\n - The flow of the answer is smooth, making it easy to follow.\n- \*\*Answer 2\*\*: \n - The language is also clear and easy to understand.\n - The description is detailed and provides a vivid picture of the scene.\n - The answer is well-structured, but the inclusion of elements not clearly visible in the image (fire hydrant, trash bag) slightly complicates the clarity.\n\n\*\*Reasoning\*\*: Both answers are clear and well-articulated. However, Answer 1 is slightly clearer because it avoids introducing elements that are not clearly visible in the image.\n\n

#### \*\*4. Relevance\*\*\n- \*\*Answer 1\*\*: \n - Directly addresses the question by focusing on the relationship between the child, the frisbee, and the urban environment.\n - The description of the child's action and the surrounding elements is relevant to the question.\n - The mention of a person near the end of the sidewalk is somewhat tangential but still related to the scene.\n- \*\*Answer 2\*\*: \n - Addresses the question by describing the child's action, the frisbee, and the surrounding environment.\n - Introduces elements (fire hydrant, trash bag) that are not clearly visible in the image, which reduces relevance.\n - The detailed description of the scene is relevant but includes unnecessary elements.\n\n\*\*Reasoning\*\*: Answer 1 is more relevant because it focuses on the elements that are clearly visible in the image and directly relates to the question. Answer 2 introduces elements that are not clearly visible, which reduces its relevance.\n\n

#### \*\*Overall Judgment\*\*\n- \*\*Answer 1\*\*: is better because it is more accurate, relevant, and clear. It avoids introducing elements that are not clearly visible in the image and focuses on the main elements of the scene (child, frisbee, urban environment) in a straightforward manner.\n- \*\*Answer 2\*\*: while more complete in its description, introduces elements that are not clearly visible in the image, which reduces its accuracy and relevance.\n\n\*\*Final Answer: Answer 1 is better.\*\*

Figure 10: VL-RewardBench Inference Case 1.

1134

1135

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

**Prompt:**

You are a highly capable multimodal AI assistant tasked with evaluating answers to visual questions. Please analyze the following image and question, then determine which of the two provided answers is better.

**Question:**

Create a detailed description describing the elements of the given worldwide landmark image. Identify the name of the landmark and describe the objects, colors, perspectives, etc.

**Answer 1:**

"The image showcases a breathtaking view of a mountainous region with a deep, green lake nestled amidst the rocky terrain. The perspective is from a high vantage point, providing a panoramic view of the landscape. The landmark, identified as \"sa\_1563\", is a testament to the natural beauty and diversity of our world. The combination of the rugged mountains, the clear blue sky, and the vibrant green lake creates a serene and awe-inspiring scene."

**Answer 2:**

"The image showcases a serene mountain lake, nestled amidst rocky terrains. The deep green color of the water contrasts with the surrounding landscape, creating a tranquil scene. This landmark is a natural beauty, located in a hilly or mountainous region. The perspective of the image gives a sense of being high above the lake, looking down upon it. However, without specific identifiable features, it's challenging to determine the exact landmark from the description."

Please evaluate both answers based on the following criteria:

1. Accuracy: How well does the answer align with the visual information in the image?
2. Completeness: Does the answer fully address all aspects of the question?
3. Clarity: Is the answer easy to understand and well-articulated?
4. Relevance: Does the answer directly relate to the question and the image?

After your evaluation, please:

1. Explain your reasoning for each criterion.
2. Provide an overall judgment on which answer is better (Answer 1 or Answer 2). For example: Overall Judgment: Answer X is better.

Your response should be structured and detailed, demonstrating your understanding of both the visual and textual elements of the task.

**Spark Judgement:**

#### Evaluation Criteria

####1. Accuracy\*\*\n- \*\*Answer 1\*\*: The description mentions a "deep, green lake" and "rugged mountains," which aligns with the image. However, the mention of a "clear blue sky" is not evident in the image, as the sky is not visible. The name "sa\_1563" is provided but is not a recognized landmark name, which introduces an element of inaccuracy.\n- \*\*Answer 2\*\*: This answer accurately describes the deep green color of the water and the rocky terrains surrounding the lake. It also correctly notes the perspective, which appears to be from a higher vantage point. Unlike Answer 1, it avoids mentioning elements (like the sky) that are not present in the image.\n\*\*Reasoning\*\*: Answer 2 is more accurate because it sticks to the visual elements present in the image without introducing unnecessary or incorrect details.\n\n

####2. Completeness\*\*\n- \*\*Answer 1\*\*: While it provides a detailed description of the landscape, including the lake, mountains, and perspective, it includes the name "sa\_1563," which is not a recognized landmark. This adds incomplete or irrelevant information. Additionally, it does not explicitly address the question of identifying the landmark, which is a key part of the prompt.\n- \*\*Answer 2\*\*: This answer focuses on the elements visible in the image, such as the lake, rocks, and perspective. It acknowledges the challenge of identifying the exact landmark due to the lack of specific features, which is a honest and complete response given the information available.\n\*\*Reasoning\*\*: Answer 2 is more complete because it addresses the limitations of the image (e.g., the inability to identify the landmark) and focuses on the visible elements without adding unnecessary details.\n\n

####3. Clarity\*\*\n- \*\*Answer 1\*\*: The answer is generally clear and well-articulated, with a smooth flow of description. However, the inclusion of the name "sa\_1563" without explanation can be confusing and detracts from the clarity.\n- \*\*Answer 2\*\*: This answer is also clear and well-structured. It uses simple language to describe the scene and acknowledges the ambiguity regarding the landmark's identity, which adds to its transparency and clarity.\n\*\*Reasoning\*\*: Both answers are clear, but Answer 2 is slightly better because it avoids confusion by not introducing unclear terms like "sa\_1563".\n\n

####4. Relevance\*\*\n- \*\*Answer 1\*\*: The answer is somewhat relevant, as it describes the image's elements. However, the inclusion of the name "sa\_1563" is irrelevant and distracts from the actual content of the image. The description of the "clear blue sky" is also irrelevant since the sky is not visible.\n- \*\*Answer 2\*\*: This answer is highly relevant. It directly addresses the visual elements in the image, such as the lake, rocks, and perspective, and honestly discusses the difficulty of identifying the landmark. It stays focused on the image and the question.\n\*\*Reasoning\*\*: Answer 2 is more relevant because it stays focused on the image and the question, avoiding irrelevant details.\n\n

## Overall Judgment\n\*\*Answer 2 is better.\*\*\n\*\*Reasoning\*\*: Answer 2 is more accurate as it avoids introducing elements not present in the image (e.g., the sky, the name "sa\_1563").\n\*\*Completeness\*\*: Answer 2 is more complete because it honestly addresses the limitations of identifying the landmark and focuses on the visible elements.\n\*\*Clarity\*\*: Both answers are clear, but Answer 2 is slightly better for its transparency about the unknowns.\n\*\*Relevance\*\*: Answer 2 is more relevant as it stays focused on the image and the question without introducing irrelevant details.\nOverall, Answer 2 provides a more precise, honest, and focused description of the image and its elements.

Figure 11: VL-RewardBench Inference Case 2.