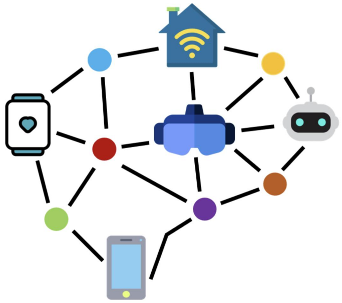
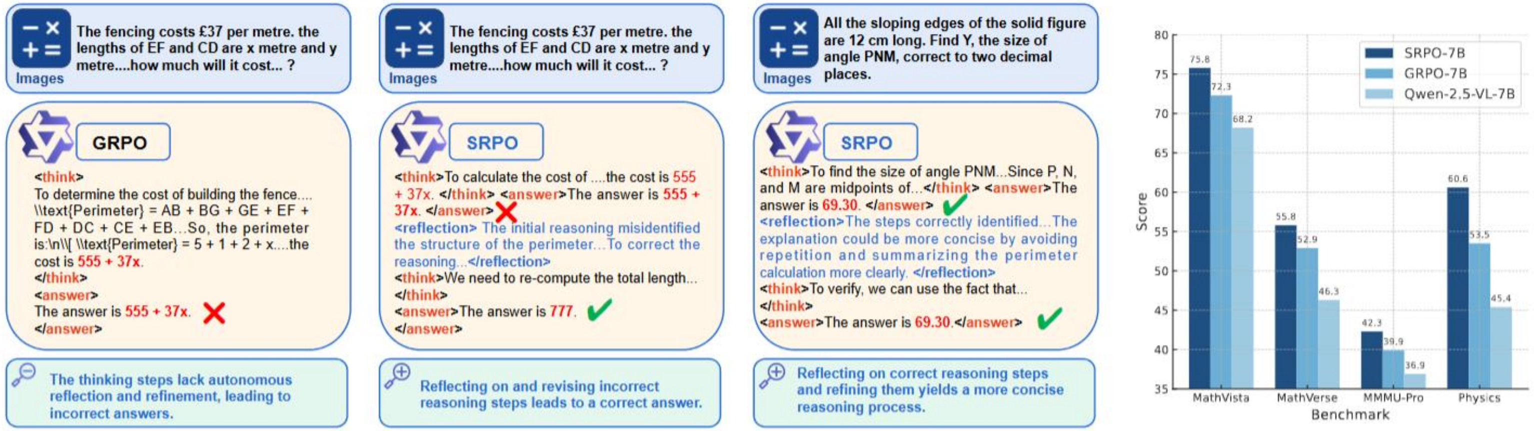


SRPO: Enhancing Multimodal LLM Reasoning via Reflection-Aware Reinforcement Learning

Zhongwei Wan,
The Ohio State University,

https://arxiv.org/abs/2506.01713
https://srpo.pages.dev/

Background and Motivation



The fencing costs £37 per metre. the lengths of EF and CD are x metre and y metre....how much will it cost... ?

<think>
To determine the cost of building the fence....
\\text{Perimeter} = AB + BG + GE + EF + FD + DC + CE + EB...So, the perimeter is:\n\\[\\text{Perimeter} = 5 + 1 + 2 + x....the cost is 555 + 37x.
</think>
<answer>
The answer is 555 + 37x. ✗
</answer>

The thinking steps lack autonomous reflection and refinement, leading to incorrect answers.

The fencing costs £37 per metre. the lengths of EF and CD are x metre and y metre....how much will it cost... ?

<think>To calculate the cost of ....the cost is 555 + 37x. </think> <answer>The answer is 555 + 37x. </answer> ✗
<reflection> The initial reasoning misidentified the structure of the perimeter...To correct the reasoning...</reflection>
<think>We need to re-compute the total length...
</think>
<answer>The answer is 777. ✓
</answer>

Reflecting on and revising incorrect reasoning steps leads to a correct answer.

All the sloping edges of the solid figure are 12 cm long. Find Y, the size of angle PNM, correct to two decimal places.

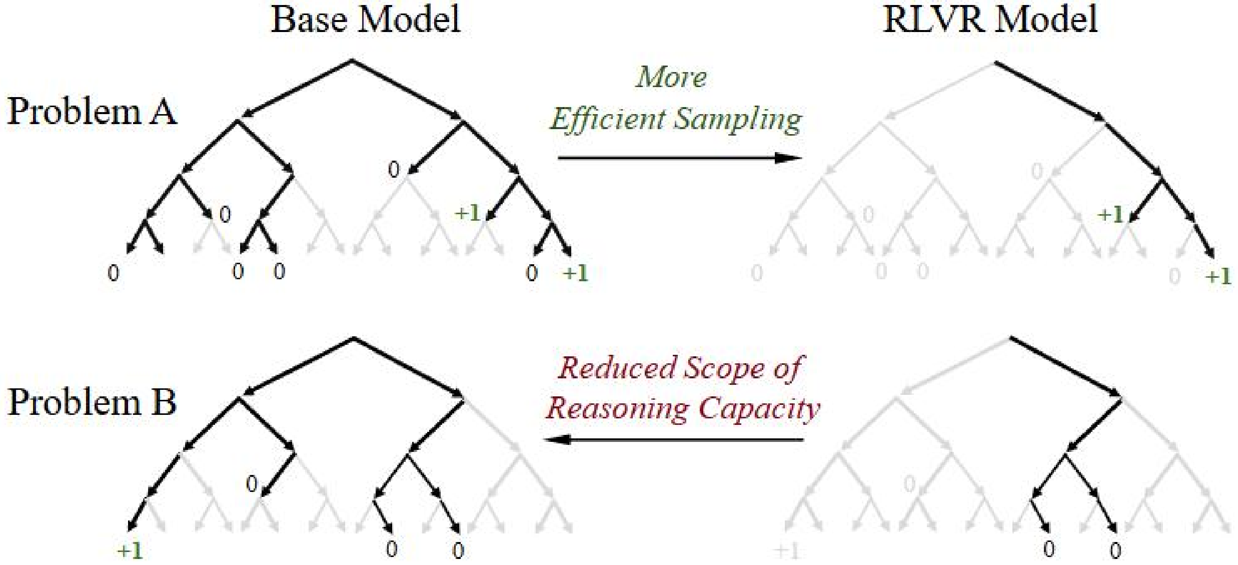
<think>To find the size of angle PNM...Since P, N, and M are midpoints of...</think> <answer>The answer is 69.30. </answer> ✓
<reflection>The steps correctly identified...The explanation could be more concise by avoiding repetition and summarizing the perimeter calculation more clearly. </reflection>
<think>To verify, we can use the fact that...
</think>
<answer>The answer is 69.30.</answer> ✓

Reflecting on correct reasoning steps and refining them yields a more concise reasoning process.

Figure 1: Left: Illustrative examples of reflection improving reasoning. Right: Quantitative comparison on benchmark datasets.

- MLLMs struggle on complex visual-text reasoning.
- Outputs often verbose or incorrect.
- Missing piece: explicit self-reflection.
- We propose SRPO to inject reflection in SFT + RL.

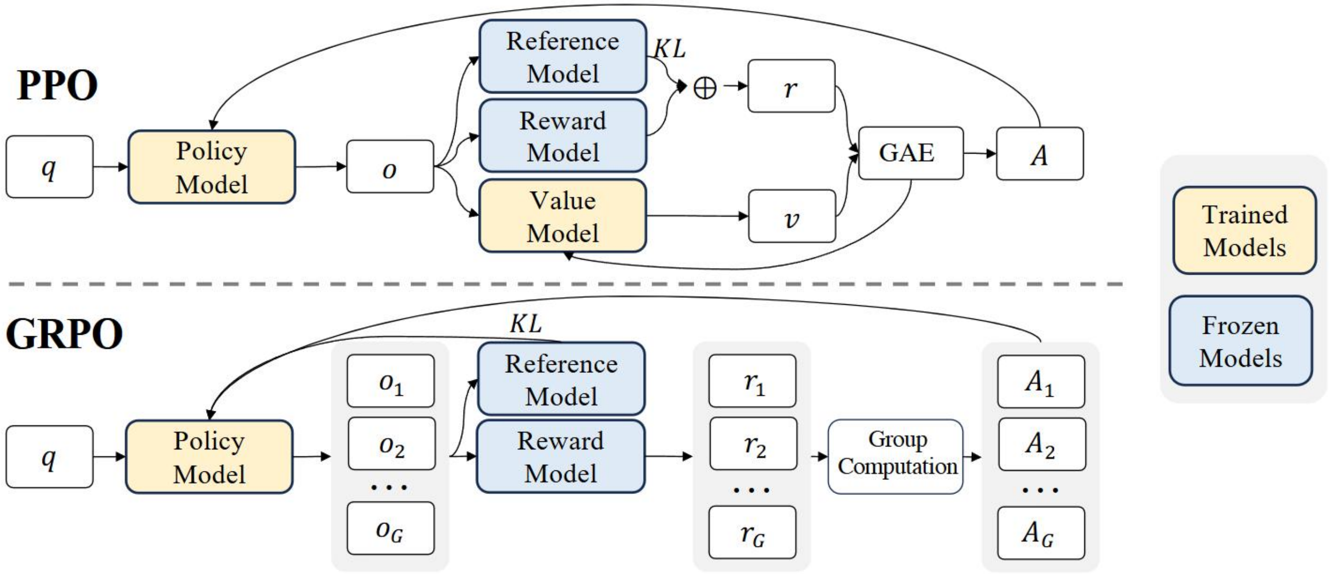
Problem



- Local token dependency → drift, verbosity.
- Pretraining constrains behaviors; vanilla RL can’t teach reflection well.
- Need a unified way to learn reflect–revise.

[1] Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?

State of the Arts & Limitations



- Large-scale RL (R1, GRPO variants) boosts CoT but seldom rewards reflection.
- Multimodal RL (Vision-R1, MM-Eureka, VL-Rethinker) lacks reflection utility checks.
- Vulnerable to length redundancy and reward gaming.

[2] DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

Overview of SRPO

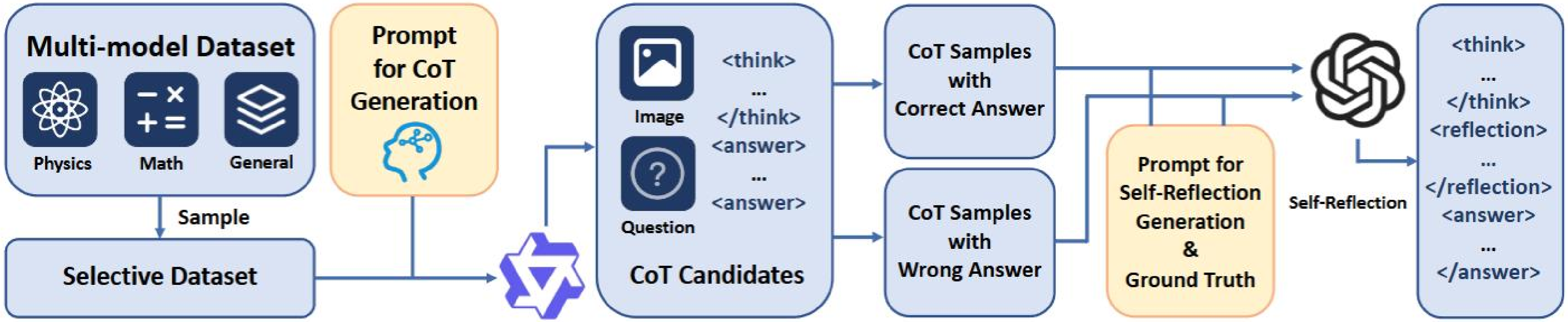
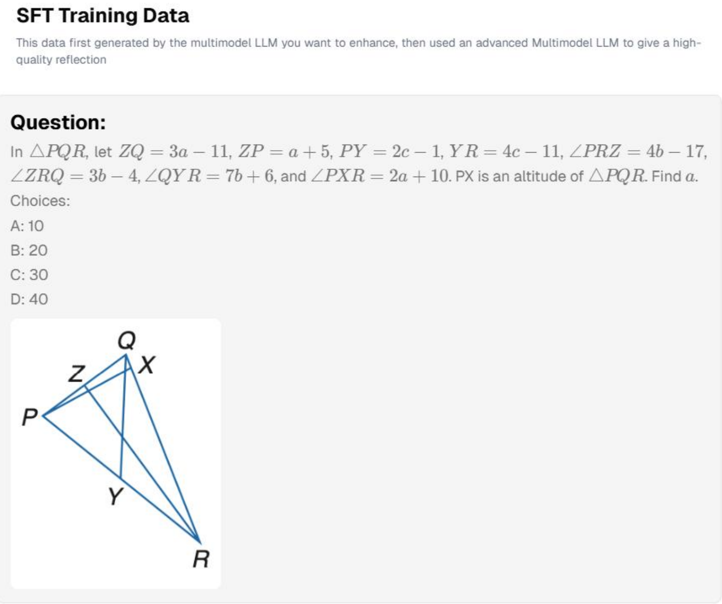


Figure 2: Pipeline of Self-Reflection SFT data construction, including CoT and self-reflection generation.

- Two stages: reflection-SFT → reflection-aware RL.
- Pattern: first solution → reflection → refined solution.
- Reward = task + reflection.

Key Technique (1/3): Reflection-Oriented SFT



SFT Training Data

This data first generated by the multimodel LLM you want to enhance, then used an advanced Multimodel LLM to give a high-quality reflection

Question:

In $\triangle PQR$, let $ZQ = 3a - 11$, $ZP = a + 5$, $PY = 2c - 1$, $YR = 4c - 11$, $\angle PRZ = 4b - 17$, $\angle ZRQ = 3b - 4$, $\angle QYR = 7b + 6$, and $\angle PXR = 2a + 10$. PX is an altitude of $\triangle PQR$. Find a.

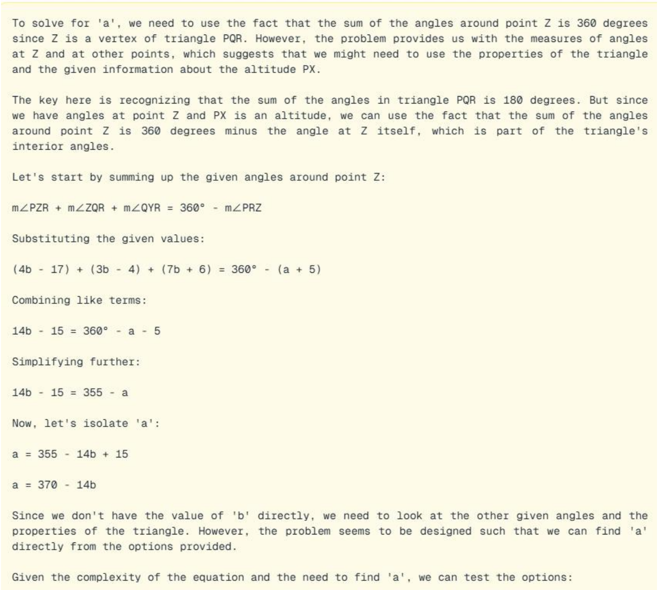
Choices:

A: 10

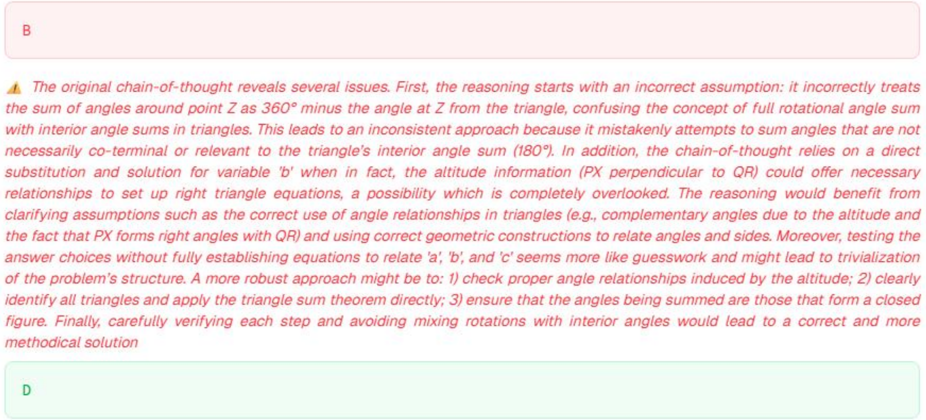
B: 20

C: 30

D: 40



```
To solve for 'a', we need to use the fact that the sum of the angles around point Z is 360 degrees
since Z is a vertex of triangle PQR. However, the problem provides us with the measures of angles
at Z and at other points, which suggests that we might need to use the properties of the triangle
and the given information about the altitude PX.

The key here is recognizing that the sum of the angles in triangle PQR is 180 degrees. But since
we have angles at point Z and PX is an altitude, we can use the fact that the sum of the angles
around point Z is 360 degrees minus the angle at Z itself, which is part of the triangle's
interior angles.

Let's start by summing up the given angles around point Z:

m∠PZR + m∠ZQR + m∠QYR = 360° - m∠PRZ

Substituting the given values:

(4b - 17) + (3b - 4) + (7b + 6) = 360° - (a + 5)

Combining like terms:

14b - 15 = 360° - a - 5

Simplifying further:

14b - 15 = 355 - a

Now, let's isolate 'a':

a = 355 - 14b + 15

a = 370 - 14b

Since we don't have the value of 'b' directly, we need to look at the other given angles and the
properties of the triangle. However, the problem seems to be designed such that we can find 'a'
directly from the options provided.

Given the complexity of the equation and the need to find 'a', we can test the options:
```



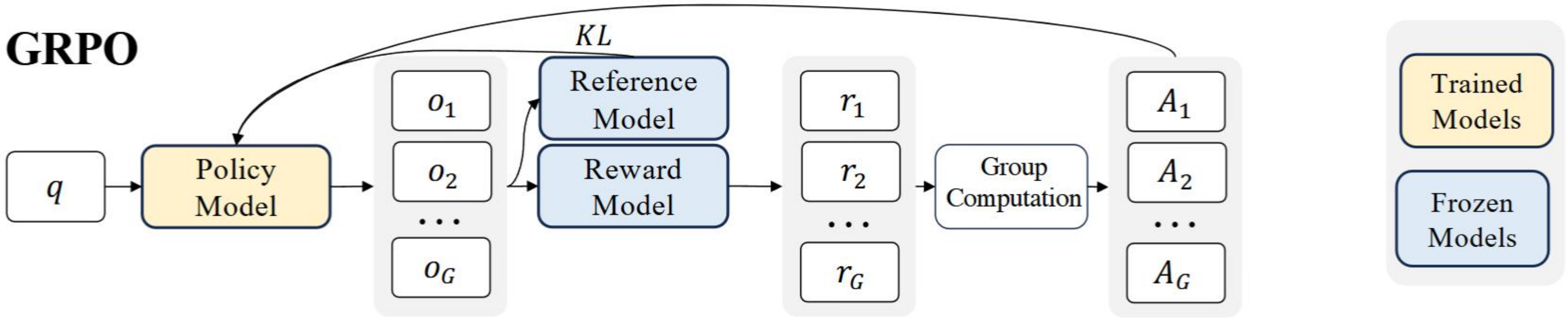
B

⚠ *The original chain-of-thought reveals several issues. First, the reasoning starts with an incorrect assumption: it incorrectly treats the sum of angles around point Z as 360° minus the angle at Z from the triangle, confusing the concept of full rotational angle sum with interior angle sums in triangles. This leads to an inconsistent approach because it mistakenly attempts to sum angles that are not necessarily co-terminal or relevant to the triangle's interior angle sum (180°). In addition, the chain-of-thought relies on a direct substitution and solution for variable 'b' when in fact, the altitude information (PX perpendicular to QR) could offer necessary relationships to set up right triangle equations, a possibility which is completely overlooked. The reasoning would benefit from clarifying assumptions such as the correct use of angle relationships in triangles (e.g., complementary angles due to the altitude and the fact that PX forms right angles with QR) and using correct geometric constructions to relate angles and sides. Moreover, testing the answer choices without fully establishing equations to relate 'a', 'b', and 'c' seems more like guesswork and might lead to trivialization of the problem's structure. A more robust approach might be to: 1) check proper angle relationships induced by the altitude; 2) clearly identify all triangles and apply the triangle sum theorem directly; 3) ensure that the angles being summed are those that form a closed figure. Finally, carefully verifying each step and avoiding mixing rotations with interior angles would lead to a correct and more methodical solution*

D

- Build ~10k high-quality reflective samples.
- Two types: refine correct CoT; revise incorrect CoT.
- Distill reflective skills from stronger MLLMs.

Key Technique (2/3): GRPO Recap & Group Advantage



$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim Q,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}}$$

$$\left[\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(\frac{\pi_\theta(o_{i,t}|q)}{\pi_{\theta_{\text{old}}}(o_{i,t}|q)}A_i,\ \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q)}{\pi_{\theta_{\text{old}}}(o_{i,t}|q)},1-\epsilon,1+\epsilon\right)A_i\right)-\beta D_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right] \quad (2)$$

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}, \quad \text{where} \quad \{r_i\}_{i=1}^{G} \quad \text{are rewards from the group.}$$

- Grouped sampling; intra-group advantage.
- Clipped ratio + KL to reference.
- No critic network needed.

Key Technique (3/3): Reflection-Aware Rewards

- Structure: think → reflect → rethink.

$$R_{\text{total}} = R_{\text{task}} + R_{\text{reflection}}.$$

$$R_{\text{task}} = R_{\text{format}}(0/0.5) + R_{\text{accuracy}}(0/0.5).$$

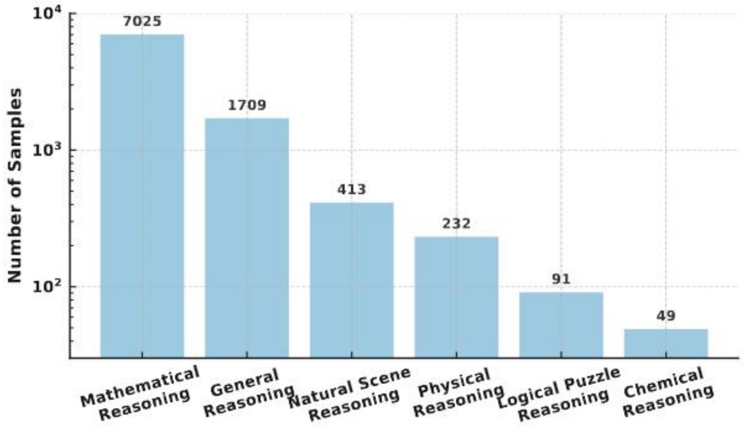
$$R_{\text{reflection}} = I_{\text{eff}} + I_{\text{ref}} + \alpha f_{\text{len}}(L_{\text{response}}).$$

$$f_{\text{len}}(L) = \left(\exp(-|L - T_{\text{target}}|/(T_{\max} - T_{\text{target}}))\right)^2.$$

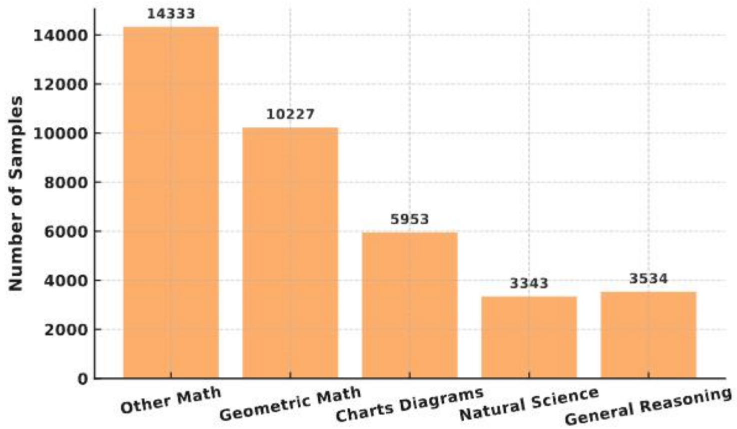
$$I_{\text{eff}} \in \{-0.25, 0, 0.25, 0.5\} \text{ (penalize harmful reflection).}$$

$$I_{\text{eff}} = \begin{cases} 0.25, & \text{if reflection keeps a corrected answer,} \\ 0.5, & \text{if reflection corrects the wrong answer,} \\ 0, & \text{if reflection fails to correct the wrong answer,} \\ -0.25 & \text{if reflection misconducts the right into wrong answer.} \end{cases}$$

Experimental Settings



(a) Self-reflection SFT data statistic

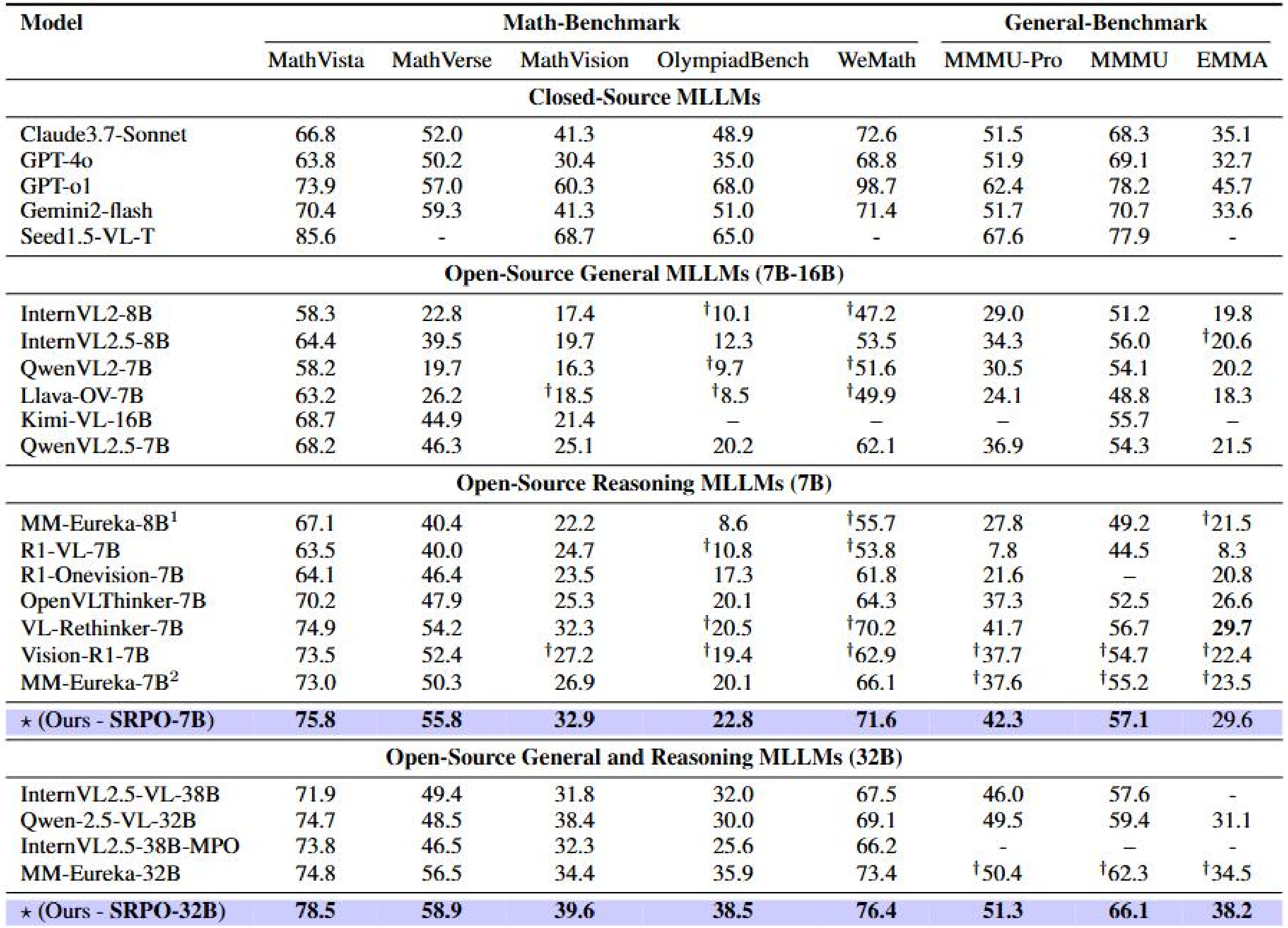


(b) RL training data statistic

- SFT: ~10k reflective samples from LLaVA-CoT, Mulberry, MathV360K.
- RL: ~30k diverse multimodal reasoning tasks.
- Models: Qwen-2.5-VL-7B/32B; OpenRLHF; 3 epochs; α=0.1; T=1.0; Adam 1e-6.

Results (1/3): Main Benchmarks

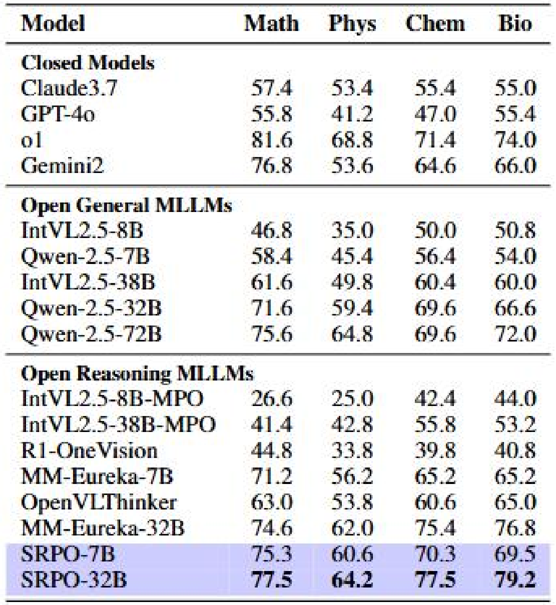
- SRPO-7B tops open-source peers across many tasks.
- SRPO-32B competitive with strong closed models.
- Strong on MathVista/MathVerse/MMMU-Pro/EMMA.



Model	Math-Benchmark					General-Benchmark		
	MathVista	MathVerse	MathVision	OlympiadBench	WeMath	MMMU-Pro	MMMU	EMMA
Closed-Source MLLMs								
Claude3.7-Sonnet	66.8	52.0	41.3	48.9	72.6	51.5	68.3	35.1
GPT-4o	63.8	50.2	30.4	35.0	68.8	51.9	69.1	32.7
GPT-o1	73.9	57.0	60.3	68.0	98.7	62.4	78.2	45.7
Gemini2-flash	70.4	59.3	41.3	51.0	71.4	51.7	70.7	33.6
Seed1.5-VL-T	85.6	-	68.7	65.0	-	67.6	77.9	-
Open-Source General MLLMs (7B-16B)								
InternVL2-8B	58.3	22.8	17.4	†10.1	†47.2	29.0	51.2	19.8
InternVL2.5-8B	64.4	39.5	19.7	12.3	53.5	34.3	56.0	†20.6
QwenVL2-7B	58.2	19.7	16.3	†9.7	†51.6	30.5	54.1	20.2
Llava-OV-7B	63.2	26.2	†18.5	†8.5	†49.9	24.1	48.8	18.3
Kimi-VL-16B	68.7	44.9	21.4	–	–	–	55.7	–
QwenVL2.5-7B	68.2	46.3	25.1	20.2	62.1	36.9	54.3	21.5
Open-Source Reasoning MLLMs (7B)								
MM-Eureka-8B[1]	67.1	40.4	22.2	8.6	†55.7	27.8	49.2	†21.5
R1-VL-7B	63.5	40.0	24.7	†10.8	†53.8	7.8	44.5	8.3
R1-Onevision-7B	64.1	46.4	23.5	17.3	61.8	21.6	–	20.8
OpenVLThinker-7B	70.2	47.9	25.3	20.1	64.3	37.3	52.5	26.6
VL-Rethinker-7B	74.9	54.2	32.3	†20.5	†70.2	41.7	56.7	**29.7**
Vision-R1-7B	73.5	52.4	†27.2	†19.4	†62.9	†37.7	†54.7	†22.4
MM-Eureka-7B[2]	73.0	50.3	26.9	20.1	66.1	†37.6	†55.2	†23.5
⋆ (Ours - **SRPO-7B**)	**75.8**	**55.8**	**32.9**	**22.8**	**71.6**	**42.3**	**57.1**	29.6
Open-Source General and Reasoning MLLMs (32B)								
InternVL2.5-VL-38B	71.9	49.4	31.8	32.0	67.5	46.0	57.6	-
Qwen-2.5-VL-32B	74.7	48.5	38.4	30.0	69.1	49.5	59.4	31.1
InternVL2.5-38B-MPO	73.8	46.5	32.3	25.6	66.2	-	–	-
MM-Eureka-32B	74.8	56.5	34.4	35.9	73.4	†50.4	†62.3	†34.5
⋆ (Ours - **SRPO-32B**)	**78.5**	**58.9**	**39.6**	**38.5**	**76.4**	**51.3**	**66.1**	**38.2**

Results (2/3): Cross-Disciplinary & RL Variants

Table 2: Performance comparison across different disciplines in MMK12.



Model	Math	Phys	Chem	Bio
Closed Models				
Claude3.7	57.4	53.4	55.4	55.0
GPT-4o	55.8	41.2	47.0	55.4
o1	81.6	68.8	71.4	74.0
Gemini2	76.8	53.6	64.6	66.0
Open General MLLMs				
IntVL2.5-8B	46.8	35.0	50.0	50.8
Qwen-2.5-7B	58.4	45.4	56.4	54.0
IntVL2.5-38B	61.6	49.8	60.4	60.0
Qwen-2.5-32B	71.6	59.4	69.6	66.6
Qwen-2.5-72B	75.6	64.8	69.6	72.0
Open Reasoning MLLMs				
IntVL2.5-8B-MPO	26.6	25.0	42.4	44.0
IntVL2.5-38B-MPO	41.4	42.8	55.8	53.2
R1-OneVision	44.8	33.8	39.8	40.8
MM-Eureka-7B	71.2	56.2	65.2	65.2
OpenVLThinker	63.0	53.8	60.6	65.0
MM-Eureka-32B	74.6	62.0	75.4	76.8
SRPO-7B	75.3	60.6	70.3	69.5
SRPO-32B	**77.5**	**64.2**	**77.5**	**79.2**

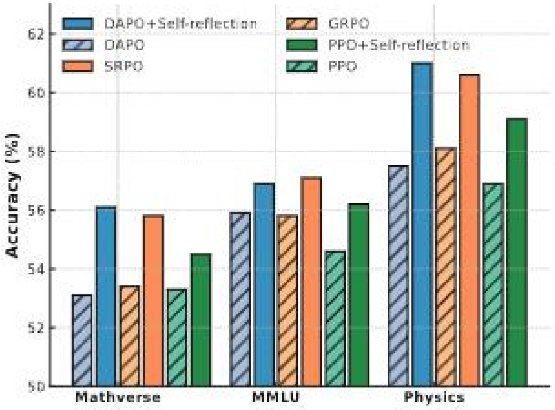
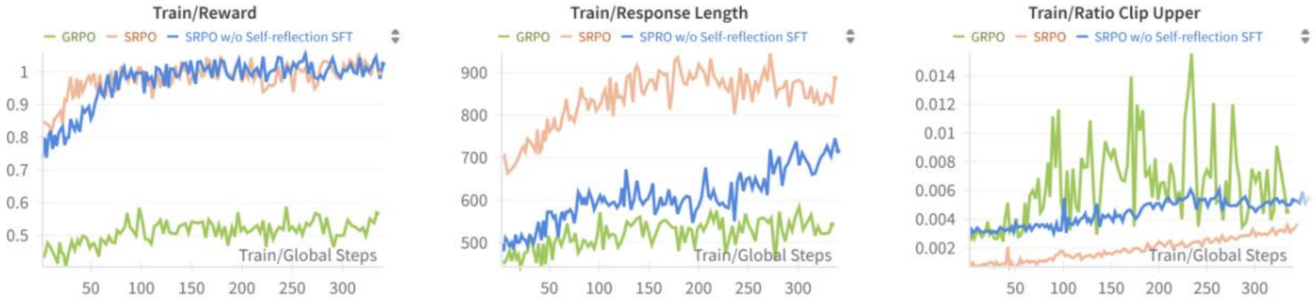


Figure 5: Performance of various RL methods with and without self-reflection.

Results (3/3): Ablations & Dynamics

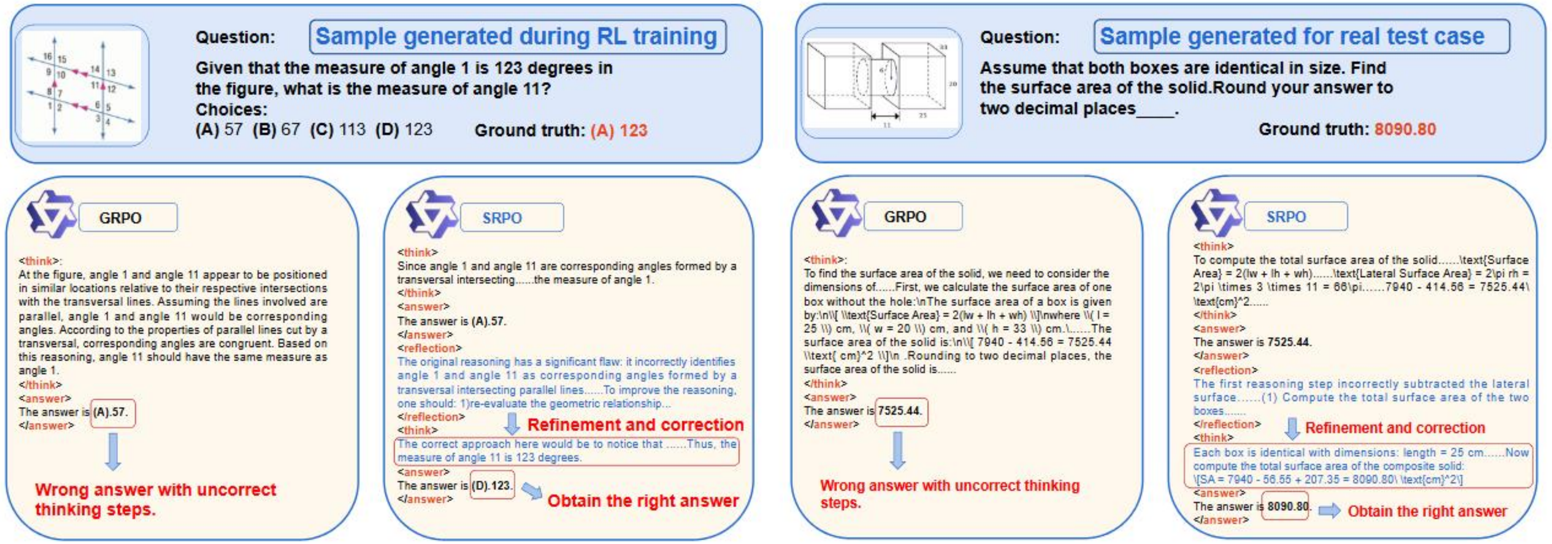
Table 3: Ablation study of SRPO-7B on RL training data size and self-reflection components.

Model Components	RL Data Size	MathVista	MathVerse	MathVision	MMMU-Pro	Physics	Avg.
Qwen-2.5-VL-7B	-	68.2	46.3	25.1	36.9	45.4	44.4
+ GRPO	37K	72.3	52.9	30.3	39.9	53.5	49.8
⋆ (Ours - **SRPO-7B**)	37K	**75.8**	**55.8**	**32.9**	**42.3**	**60.6**	**53.5**
SRPO-7B	15K	74.5	54.9	32.2	41.4	60.1	52.6
SRPO-7B	5K	73.7	53.6	31.2	40.3	57.7	51.3
w/o *Self-Reflection SFT*	37K	74.2	53.3	30.3	39.7	58.6	51.2
w/o *Self-Reflection RL*	37K	70.3	48.2	27.2	38.7	48.5	46.6
- no Length Reward ($f_{\text{len}}(\cdot)$)	37K	75.3	56.2	32.4	41.7	60.1	53.1
- no Effectiveness Reward (I_{eff})	37K	73.9	54.7	31.6	40.9	58.8	52.0



- More RL data → steady gains; even 5k > GRPO.
- Removing SFT or reflection-RL hurts; I_eff is crucial.
- SRPO converges faster; length well-controlled.

Samples Analysis



Question: **Sample generated during RL training**

Given that the measure of angle 1 is 123 degrees in the figure, what is the measure of angle 11?
Choices:
(A) 57 **(B)** 67 **(C)** 113 **(D)** 123 **Ground truth:** **(A) 123**

GRPO

<think>:
At the figure, angle 1 and angle 11 appear to be positioned in similar locations relative to their respective intersections with the transversal lines. Assuming the lines involved are parallel, angle 1 and angle 11 would be corresponding angles. According to the properties of parallel lines cut by a transversal, corresponding angles are congruent. Based on this reasoning, angle 11 should have the same measure as angle 1.
</think>
<answer>
The answer is (A).57.
</answer>

Wrong answer with uncorrect thinking steps.

SRPO

<think>
Since angle 1 and angle 11 are corresponding angles formed by a transversal intersecting......the measure of angle 1.
</think>
<answer>
The answer is (A).57.
</answer>
<reflection>
The original reasoning has a significant flaw: it incorrectly identifies angle 1 and angle 11 as corresponding angles formed by a transversal intersecting parallel lines......To improve the reasoning, one should: 1)re-evaluate the geometric relationship...
</reflection>
<think>

Refinement and correction

The correct approach here would be to notice that ......Thus, the measure of angle 11 is 123 degrees.
<answer>
The answer is (D).123.
</answer>

Obtain the right answer

Question: **Sample generated for real test case**

Assume that both boxes are identical in size. Find the surface area of the solid.Round your answer to two decimal places____.

Ground truth: **8090.80**

GRPO

<think>:
To find the surface area of the solid, we need to consider the dimensions of......First, we calculate the surface area of one box without the hole:\nThe surface area of a box is given by:\n\\[\\text{Surface Area} = 2(lw + lh + wh) \\]\nwhere \\(l = 25 \\) cm, \\(w = 20 \\) cm, and \\(h = 33 \\) cm.\......The surface area of the solid is:\n\\[7940 - 414.56 = 7525.44 \\text{ cm}^2 \\]\n .Rounding to two decimal places, the surface area of the solid is......
</think>
<answer>
The answer is 7525.44.
</answer>

Wrong answer with uncorrect thinking steps.

SRPO

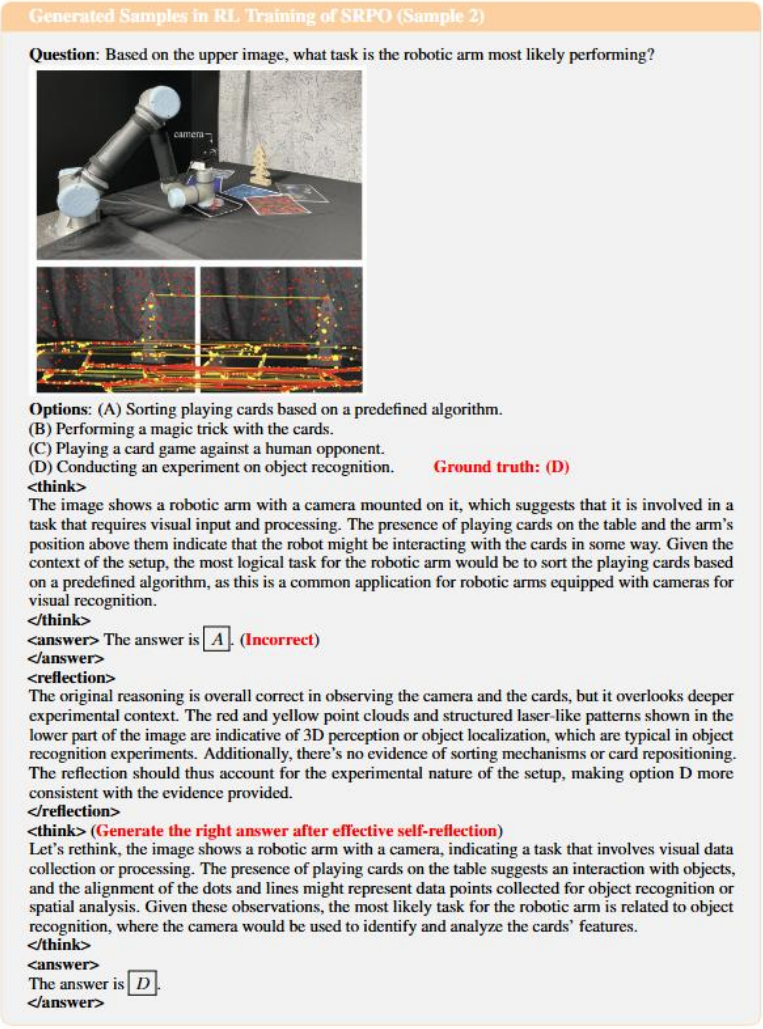
<think>
To compute the total surface area of the solid......\text{Surface Area} = 2(lw + lh + wh)......\text{Lateral Surface Area} = 2\pi rh = 2\pi \times 3 \times 11 = 66\pi......7940 - 414.56 = 7525.44\ \text{cm}^2......
</think>
<answer>
The answer is **7525.44**.
</answer>
<reflection>
The first reasoning step incorrectly subtracted the lateral surface......(1) Compute the total surface area of the two boxes.......
</reflection>
<think>

Refinement and correction

Each box is identical with dimensions: length = 25 cm......Now compute the total surface area of the composite solid:
\[SA = 7940 - 56.55 + 207.35 = 8090.80\ \text{cm}^2\]
<answer>
The answer is **8090.80**.
</answer>

Obtain the right answer

Figure 3: Generated samples in RL training (**left**) and generated samples in real test case (**right**).



Generated Samples in RL Training of SRPO (Sample 2)

Question: Based on the upper image, what task is the robotic arm most likely performing?

Options: (A) Sorting playing cards based on a predefined algorithm.
(B) Performing a magic trick with the cards.
(C) Playing a card game against a human opponent.
(D) Conducting an experiment on object recognition. **Ground truth: (D)**

<think>
The image shows a robotic arm with a camera mounted on it, which suggests that it is involved in a task that requires visual input and processing. The presence of playing cards on the table and the arm's position above them indicate that the robot might be interacting with the cards in some way. Given the context of the setup, the most logical task for the robotic arm would be to sort the playing cards based on a predefined algorithm, as this is a common application for robotic arms equipped with cameras for visual recognition.
</think>
<answer> The answer is $\boxed{A}$. (**Incorrect**)
</answer>
<reflection>
The original reasoning is overall correct in observing the camera and the cards, but it overlooks deeper experimental context. The red and yellow point clouds and structured laser-like patterns shown in the lower part of the image are indicative of 3D perception or object localization, which are typical in object recognition experiments. Additionally, there's no evidence of sorting mechanisms or card repositioning. The reflection should thus account for the experimental nature of the setup, making option D more consistent with the evidence provided.
</reflection>
<think> (**Generate the right answer after effective self-reflection**)
Let's rethink, the image shows a robotic arm with a camera, indicating a task that involves visual data collection or processing. The presence of playing cards on the table suggests an interaction with objects, and the alignment of the dots and lines might represent data points collected for object recognition or spatial analysis. Given these observations, the most likely task for the robotic arm is related to object recognition, where the camera would be used to identify and analyze the cards' features.
</think>
<answer>
The answer is $\boxed{D}$.
</answer>

Conclusion & Future Work

- SRPO unifies reflection in SFT + RL.
- Rewards: structure, brevity, effectiveness.
- Next: scale (MoE/larger), harder multimodal tasks, better reflection data generation, dynamic self-reflection.