

758 A Training Hyperparameters

759 In this section, Tables 4, 5, 6, and 7 show the hyperparameters when training models with different RL
760 algorithms (GRPO [33], REINFORCE++-baseline [14], REINFORCE++ [14], and RLOO [19]). For
761 all algorithms, we maintain identical hyperparameter configurations across experimental conditions,
762 differing only in the inclusion/exclusion of our SCS method. For each experiments, we save a
763 checkpoint every 10 steps and select the one with the highest average score.

Table 4: Hyperparameter settings for RLOO experiments.

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

Table 5: Hyperparameter settings for GRPO experiments.

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

Table 6: Hyperparameter settings for REFERENCE++-baseline experiments.

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

Table 7: Hyperparameter settings for REFERENCE++ experiments.

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

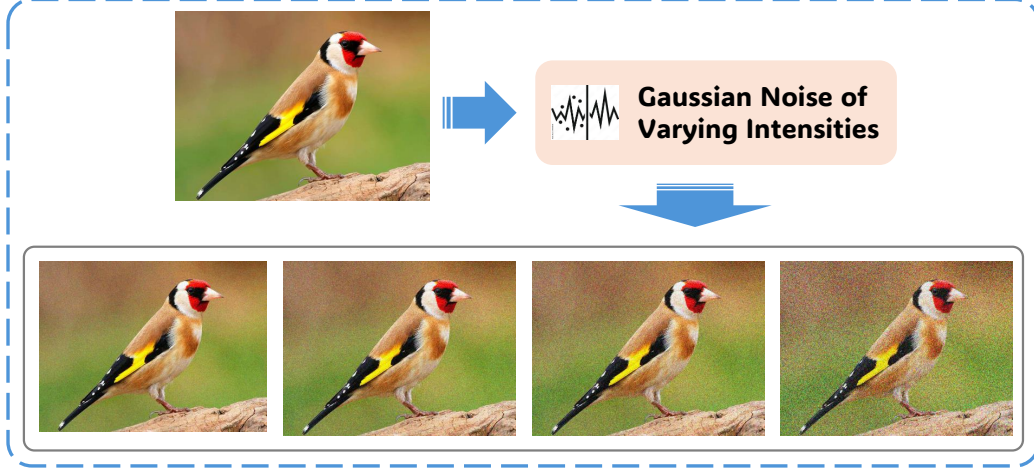


Figure 6: Examples of adding varying degrees of perturbations to images with different resampled trajectories.

B Method Details

B.1 Theoretical Details

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

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

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

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

$$S = \sum_{i=1}^N X_i. \quad (8)$$

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

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

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

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

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

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

Therefore:

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

779 Summing over all such i , we obtain:

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

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

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

781 B.2 Core Codes for SCS

782 In this section, we show the core codes of two components of SCS: Truncation-Resampling and
 783 Visual-Perturbation. Listing 1 illustrates the process of Truncation-Resampling utilizing vLLM, and
 784 Listing 2 showcases how to add varying degrees of Gaussian noise to each resampled trajectories.
 785 Figure 6 is an example of Visual-Perturbation process.

Listing 1: Core codes for Truncation-Resampling.

```

786 # 1. Expand prompts according to number of resampled trajectories
787 all_outputs_to_trunc_expand = sum([[prompt] * args.n_truncs_per_prompt for
788     prompt in all_outputs_to_trunc], [])
789 trunc_batch_size = (len(all_outputs_to_trunc_expand) + len(llms) - 1) //
790     len(llms)
791
792 # 2. Truncation-Resampling.
793 messages = all_outputs_to_trunc_expand[i * trunc_batch_size : (i + 1) *
794     trunc_batch_size]
795 if messages:
796     vllm_inputs = []
797     for output in messages:
798         # Response truncation for each init response
799         prompt_ids_list = list(output["response"].prompt_token_ids)
800         response_ids_list = list(output["response"].outputs[0].token_ids)
801         prompts_text =
802             self.data_processor.tokenizer.batch_decode(prompt_ids_list,
803                 skip_special_tokens=False)
804         trunc_idx = int(len(response_ids_list) * args.trunc_keep_rate)
805         trunc_response_ids = response_ids_list[:trunc_idx]
806         trunc_response =
807             self.data_processor.tokenizer.batch_decode(trunc_response_ids,
808                 skip_special_tokens=False)
809         if args.trunc_keep_rate > 0:
810             p = ''.join(prompts_text) + ''.join(trunc_response)
811         else:
812             p = ''.join(prompts_text)
813
814         if self.data_processor.image_aug:
815             imgs =
816                 self.image_augment_from_PIL(output['image_input']['image'])
817             vllm_inputs.append({
818                 "prompt": p,
819                 "multi_modal_data":{"image": imgs} if imgs else None,
820                 "prompt_id":output['prompt_id']
821             })
822

```

Listing 2: Core codes for Visual-Perturbation.

```

824 def add_image_noise(image, noise_level=10):
825     """
826     Add noise to an image
827
828     Parameters:
829     - image: str or PIL.Image.Image, input image path or PIL Image object
830     - noise_level: int, noise intensity (default 10, range 0-100)
831
832     Returns:
833     - PIL.Image.Image object with added noise
834     """
835
836     # 1. Read image and convert to RGB mode
837     if not isinstance(image, Image.Image):
838         img = Image.open(image).convert('RGB')
839     else:
840         img = image
841     img_array = np.array(img).astype(np.float32)
842
843     # 2. Generate noise
844     noise = None
845     max_pixel = 255.0
846     scaled_noise_level = noise_level * max_pixel / 100
847
848     # Gaussian noise (mean=0, standard deviation=noise level)
849     noise = np.random.normal(
850         loc=0,
851         scale=scaled_noise_level,
852         size=img_array.shape)
853
854     # 3. Add noise and clip pixel values
855     noisy_array = img_array + noise
856     noisy_array = np.clip(noisy_array, 0, max_pixel).astype(np.uint8)
857
858     # 4. Convert back to PIL Image
859     noisy_img = Image.fromarray(noisy_array)
860
861     return noisy_img
862
863 # Add varying degrees of Gaussian noise to each resampled trajectories
864 for img in image:
865     noise_level = random.randint(1,40)
866     img = add_image_noise(img,noise_level=noise_level)
867

```

871 B.3 Illustration of SCS Pipeline

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

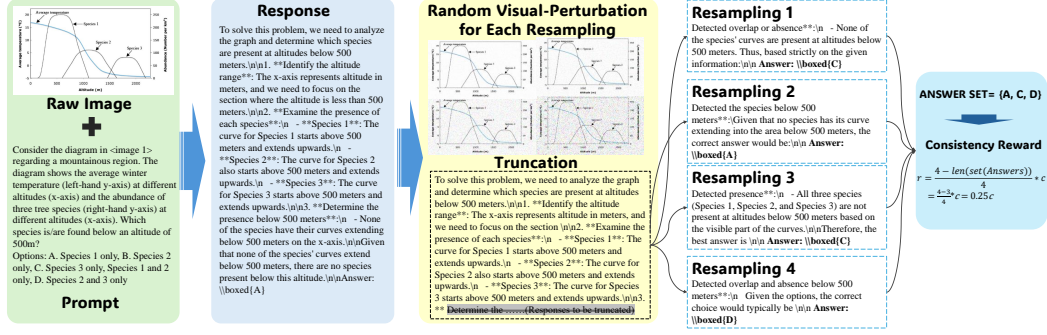


Figure 7: Pipeline of our SCS method.

C Evaluation Details

C.1 Prompt Template

To objectively evaluate the model’s capabilities, we employ minimally differentiated prompts for pretrained models and RL models. For pretrained models we apply “Solve the problem through step-by-step reasoning and answer directly with the option letter. Think about the reasoning process first and answer the question following this format: Answer: \boxed{\$LETTER}.” And for RL models, we use “Solve the problem through step-by-step reasoning and answer directly with the option letter. Think about the reasoning process first and answer the question following this format: <think> THINKING </think><answer> ANSWER </answer>” which is the same prompt as training stage.

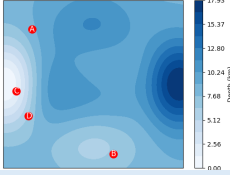
C.2 Supplementary for Evaluation in Figure 2(c)

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

D Additional Results

D.1 More Examples of Unfaithful Reasoning

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



Question:
Could you pinpoint the area in the picture that exhibits the greatest depth?

A. Point C
B. Point B
C. Point A
D. Point D.

QwenVL2.5-7B-Instruct

To determine the area with the greatest depth, we need to look at the color gradient and the contour lines on the map. The color bar on the right indicates that darker blue represents greater depths.

Point A: Located in a medium blue region.

Point B: Located in a very light blue region, indicating shallower depth.

Point C: Located in a medium blue region similar to Point A.

Point D: Located in a medium blue region similar to Points A and C.

From the color gradient, the darkest blue is at the bottom right of the map, which corresponds to the greatest depth according to the color bar.

Point A is in a medium blue region.

Point B is in a very light blue region.

Point C is in a medium blue region.

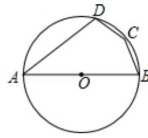
Point D is in a medium blue region.

The darkest blue region is not marked by any point but is located towards the bottom right of the map. However, since the points provided do not include this area, we should consider the points given.

Answer: C ✓

Figure 8: Phenomenon of unfaithful reasoning with correct option.

As shown in the figure, the quadrilateral ABCD is inscribed in $\odot O$, AB is the diameter of $\odot O$, and point C is the midpoint of \widehat{BD} . If $\angle A = 40^\circ$, then the degree of $\angle B$ is ()



Question:

(A) 40°
(B) 50°
(C) 60°
(D) 70°

QwenVL2.5-7B-Instruct

To solve this problem, we need to use properties of cyclic quadrilaterals and the fact that AB is the diameter of the circle.

Identify the given information:

- Quadrilateral ABCD is inscribed in circle O.
- AB is the diameter of circle O.
- Point C is the midpoint of arc BD.
- $\angle A = 40^\circ$.

Use the property of the diameter: Since AB is the diameter, $\angle ADB = 90^\circ$ (angle subtended by a diameter in a semicircle).

Use the property of the midpoint of an arc: Since C is the midpoint of arc BD, $\angle BCD = \angle BAC$ (angles subtended by the same arc are equal).

Calculate $\angle BCD$: $\angle BCD = \angle BAC = 40^\circ$.

Use the fact that the sum of opposite angles in a cyclic quadrilateral is 180° :

$\angle A + \angle C = 180^\circ$

$40^\circ + \angle C = 180^\circ$

$\angle C = 140^\circ$

Calculate $\angle B$:

$\angle B + \angle C = 180^\circ$

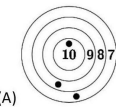
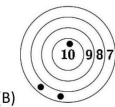
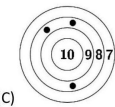
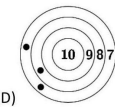
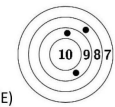
$\angle B + 140^\circ = 180^\circ$

$\angle B = 40^\circ$

Therefore, the degree of $\angle B$ is 40° .

The correct answer is: A ✓

Figure 9: Phenomenon of unfaithful reasoning with correct option.

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

QwenVL2.5-7B-Instruct

To determine which target Ricky scored the most points on, we need to analyze the number of points each target has and identify the one with the highest score.

Target A: The target has 3 points.

Target B: The target has 4 points.

Target C: The target has 5 points.

Target D: The target has 6 points.

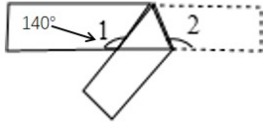
Target E: The target has 7 points.

Since Ricky scored the most points, we need to find the target with the highest number of points.

From the analysis above, Target E has the highest number of points (7).

Therefore, the target that Ricky scored the most points on is E ✓

Figure 10: Phenomenon of unfaithful reasoning with correct option.



Question:
degree of angle 2 is()


A:100°
B:110°
C:120°
D:140°

QwenVL2.5-7B-Instruct

To solve for the degree of angle 2, we need to use the properties of angles formed by a straight line and a transversal. 1. Identify the given information: Angle 1 is 140 degrees. 2. The two angles (angle 1 and angle 2) form a straight line with the dashed line. 3. Recall that the sum of angles on a straight line is 180 degrees. 4. Set up the equation: $\text{Angle 1} + \text{Angle 2} = 180^\circ$. 5. Substitute the value of angle 1 into the equation: $140^\circ + \text{Angle 2} = 180^\circ$. 6. Solve for angle 2: $\text{Angle 2} = 180^\circ - 140^\circ = 40^\circ$. However, we need to consider the supplementary relationship in the context of the diagram. Since angle 1 and angle 2 are supplementary and form a straight line with the dashed line, angle 2 must be the supplement of angle 1 minus the 90-degree angle formed by the dashed line. Therefore, angle 2 is: $\text{Angle 2} = 180^\circ - 140^\circ = 40^\circ$. But since the options provided do not match our calculated value, we need to re-evaluate the problem considering the supplementary relationship correctly. The correct approach is to recognize that angle 2 is the supplementary angle to angle 1 minus the 90-degree angle formed by the dashed line. Thus, the correct answer is: B

Figure 11: Phenomenon of unfaithful reasoning with correct option.


<image1>




Question:
Which piece is missing?

<image2>


(A)




(B)




(C)



(D)



(E)

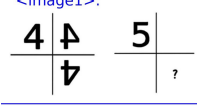


QwenVL2.5-7B-Instruct

To determine which piece is missing from the larger triangle in <image1>, we need to analyze the pattern of the smaller triangles within it. 1. The larger triangle is divided into 9 smaller triangles. 2. Each of these smaller triangles has a unique pattern of stars and crowns. 3. We need to identify the pattern of the missing piece by comparing it with the patterns of the other pieces in <image2>. Let's examine the pattern of the missing piece: The missing piece has a star in the top left corner and a crown in the bottom right corner. This pattern matches the pattern of the piece labeled (A) in <image2>. Therefore, the missing piece is: A

Figure 12: Phenomenon of unfaithful reasoning with correct option.

<image1>



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

<image2>

A) 2 B) 2 C) 5 D) 5 E) 5

QwenVL2.5-7B-Instruct

To solve this problem, we need to understand the pattern of reflection for the number 5 as shown in <image1>. 1. The number 4 is reflected horizontally and vertically. 2. The result of reflecting 4 horizontally is 4. 3. The result of reflecting 4 vertically is also 4. Now, let's apply the same reflections to the number 5. 1. Reflecting 5 horizontally: The top part (5) remains the same. The bottom part (5) remains the same. So, the result is 5. 2. Reflecting 5 vertically: The left part (5) remains the same. The right part (5) remains the same. So, the result is 5. Since both horizontal and vertical reflections of 5 yield the same result, the correct answer is: C

Figure 13: Phenomenon of unfaithful reasoning with correct option.

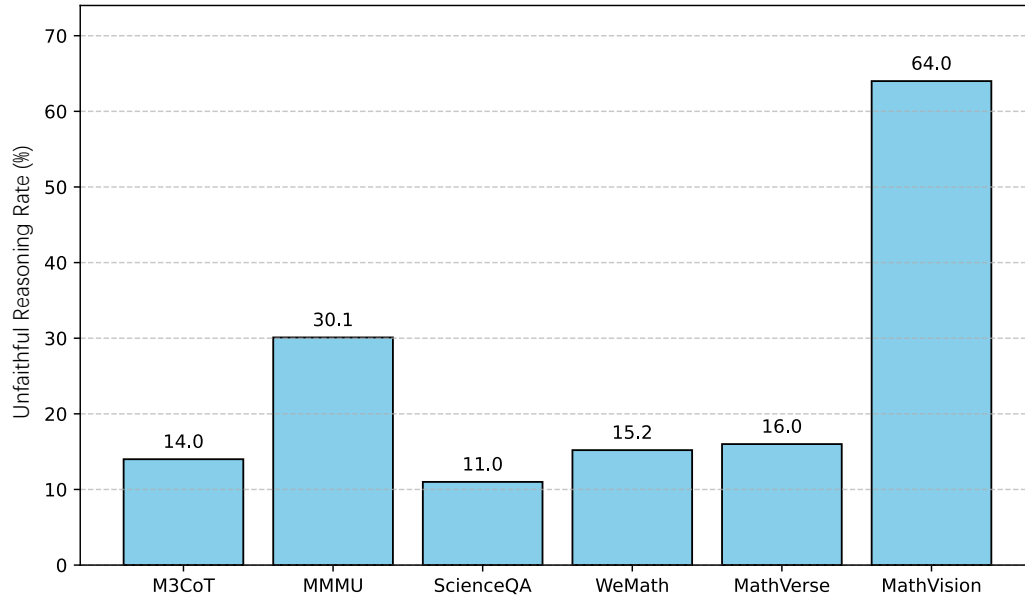


Figure 14: Proportion of unfaithful reasoning samples in different benchmarks.

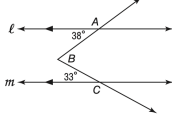
897 D.2 Human Evaluation of Unfaithful Reasoning

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

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

905 D.3 Examples of Models' Outputs

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



Question:
 In the figure below, what is the measure of $\angle ABC$? Express your answer in degrees.
 Choices: A: 33 B: 38 C: 61 D: 71


QwenVL2.5-7B-Instruct ❌

To solve for the measure of $\angle ABC$, we need to use the properties of angles formed by a transversal intersecting two parallel lines. Identify the given angles: $\angle BAC = 38^\circ$ and $\angle BCA = 33^\circ$. Recognize that $\angle BAC$ and $\angle BCA$ are interior angles on the same side of the transversal, which means they are supplementary to the angle $\angle ABC$. The sum of the angles in a triangle is always 180° . Therefore, we can write the equation for the triangle $\angle BAC + \angle ABC + \angle BCA = 180^\circ$. Substitute the known values into the equation: $38^\circ + \angle ABC + 33^\circ = 180^\circ$. Simplify the equation: $71^\circ + \angle ABC = 180^\circ$. Solve for $\angle ABC$: $\angle ABC = 180^\circ - 71^\circ = 109^\circ$. However, there seems to be a misunderstanding in the problem statement or the reference solution. Based on the standard geometric principles, the correct measure of $\angle ABC$ should be 109° , but since this is not one of the options, we should re-evaluate the problem constraints or the reference solution. Given the options provided, the closest reasonable answer based on the standard geometric principles would be: Answer: D

QwenVL2.5-7B-Instruct-SCS ✅

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

Figure 15: Phenomenon of unfaithful reasoning with correct option.



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

QwenVL2.5-7B-Instruct ❌

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

QwenVL2.5-7B-Instruct-SCS ✅

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

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

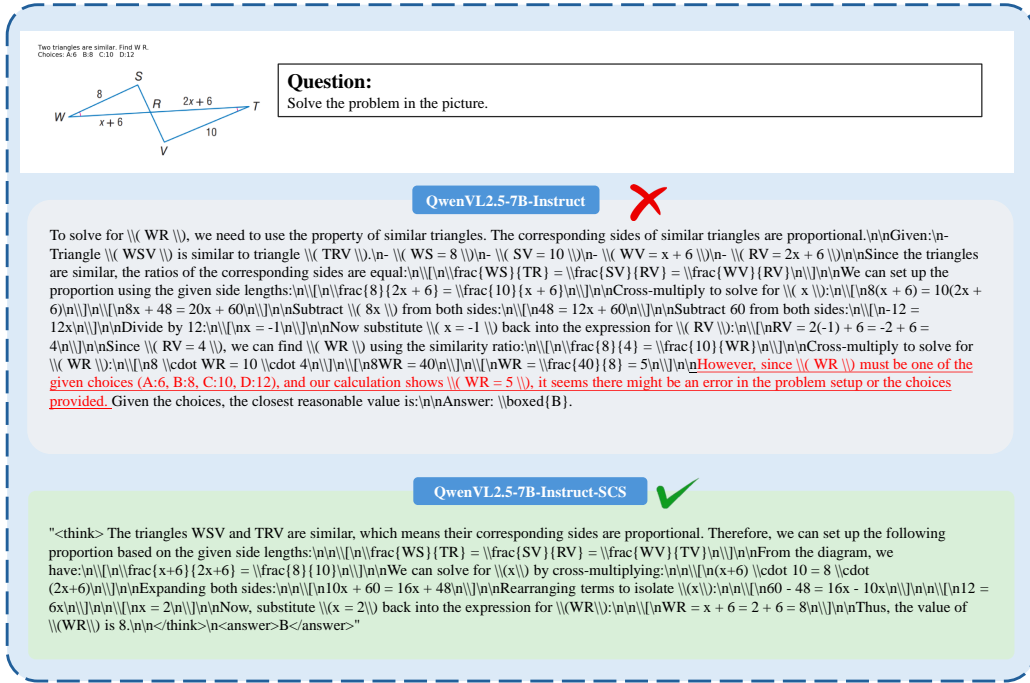

QwenVL2.5-7B-Instruct-SCS
✓

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

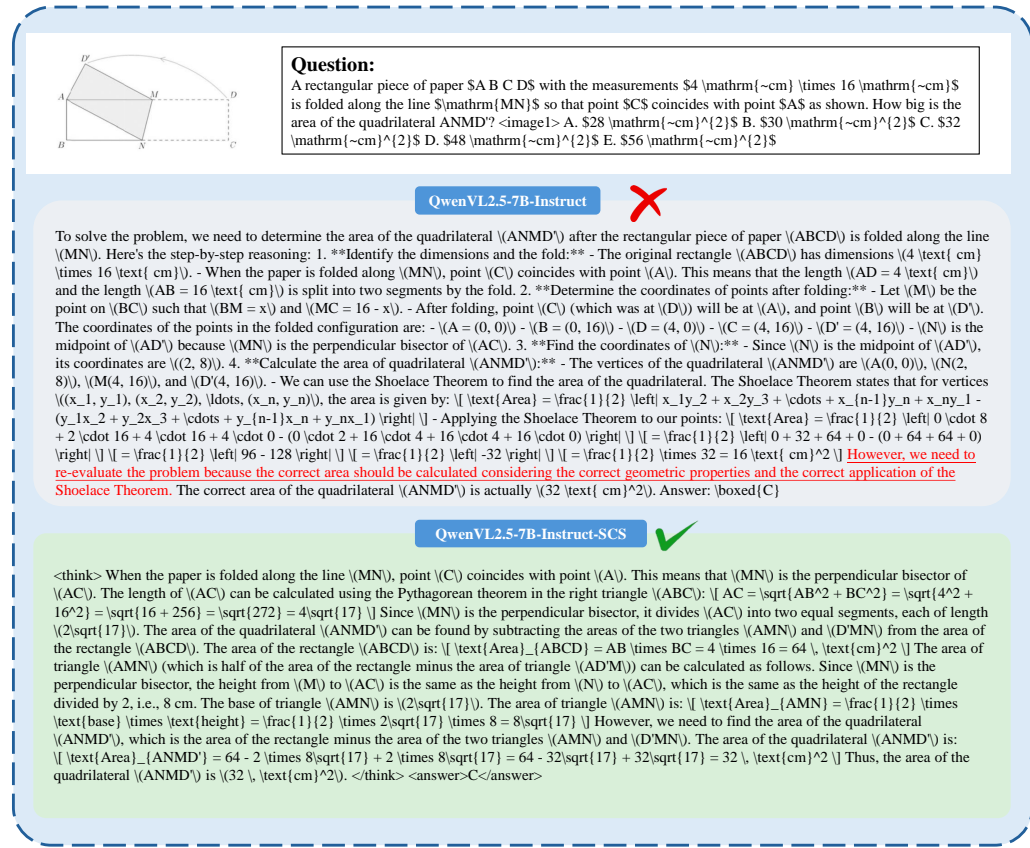

QwenVL2.5-7B-Instruct-SCS
✓

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

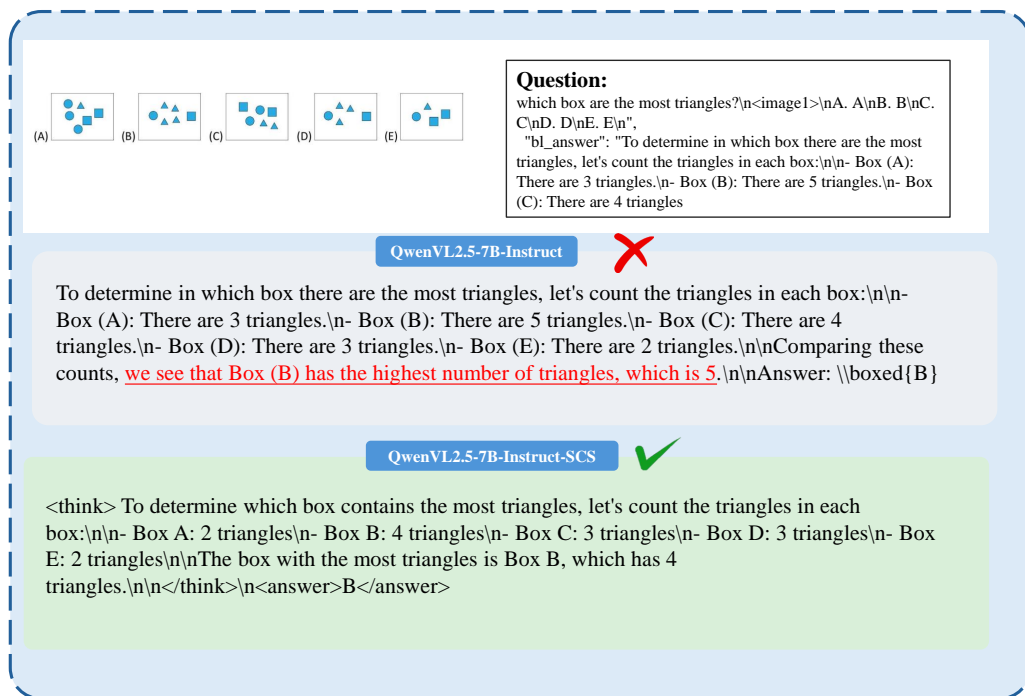


Figure 19: Comparison of model outputs before and after RL training through SCS.

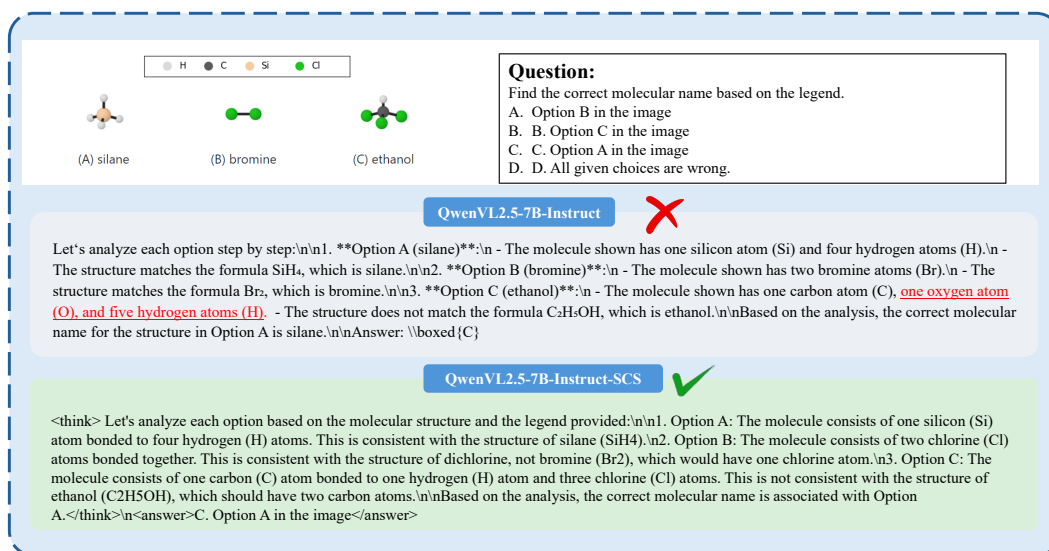


Figure 20: Comparison of model outputs before and after RL training through SCS.

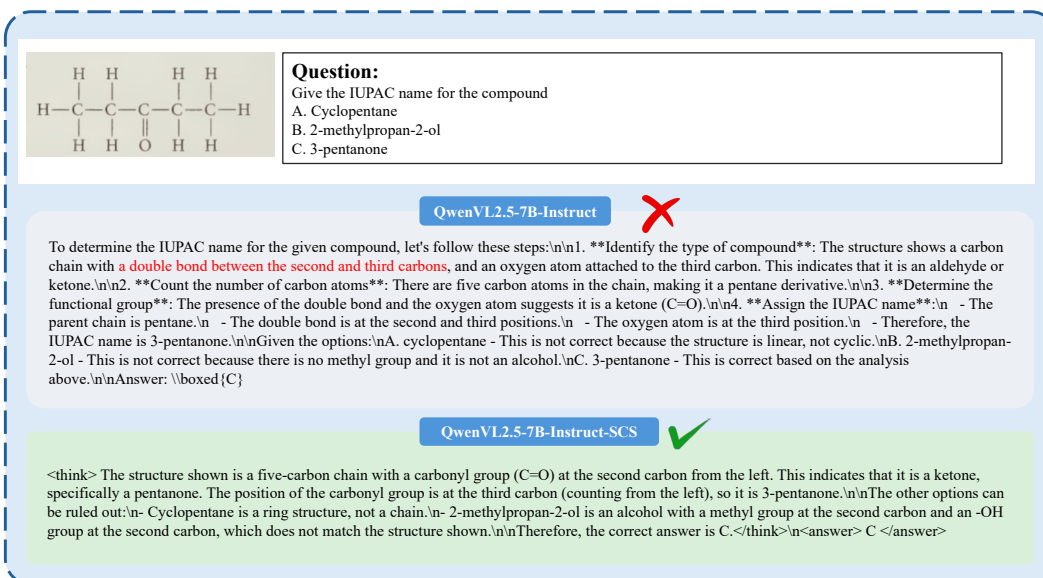


Figure 21: Comparison of model outputs before and after RL training through SCS.

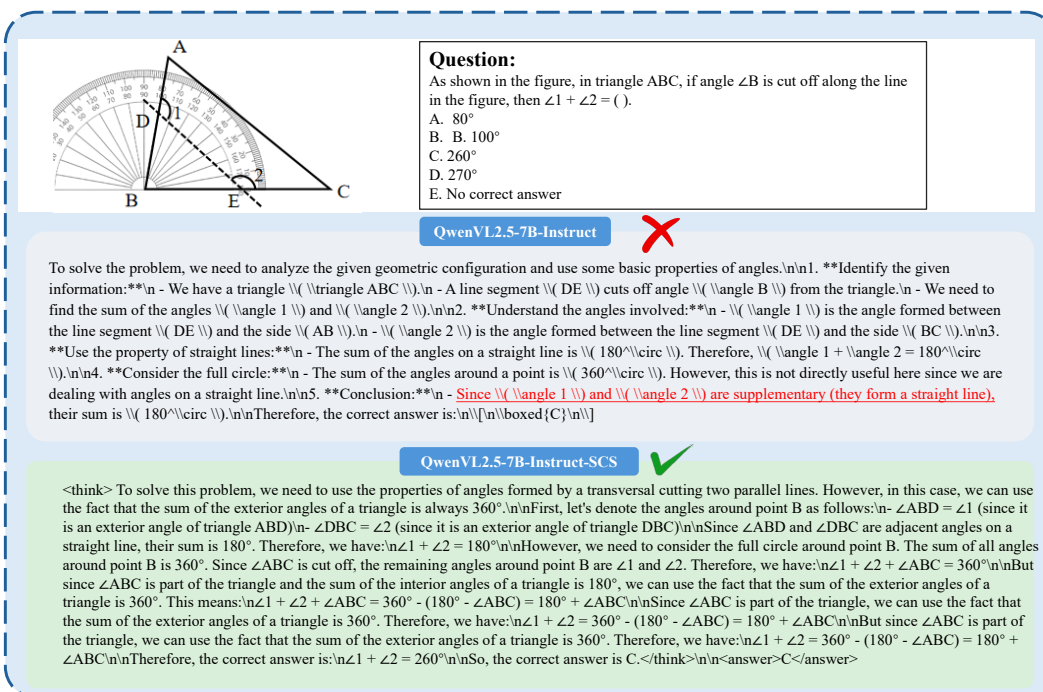


Figure 22: Comparison of model outputs before and after RL training through SCS.

910 **E Other**

911 All benchmark datasets used for evaluation are properly cited within the manuscript. For all evaluated
912 models, we strictly comply with their respective licenses: open-source models are employed in
913 accordance with their designated usage terms. The training pipeline is implemented based on the
914 open-source framework OpenRLHF¹, while the evaluation is conducted using established open-
915 source libraries, including Transformers².

¹<https://github.com/OpenRLHF/OpenRLHF>

²<https://github.com/huggingface/transformers>