Appendix

A. Dataset Examples

We present examples of VISCO dataset in Figure 18. We further show the details of each step in the dataset construction process as follows:

Task input collection. We collect images, questions and ground truth answers from existing visual question answering datasets. The distribution of our tasks and datasets are shown in Figure 3 in the main content. Examples from each task are in Figure 19.

Response collection. We use 7 LVLMs to sample model responses, including both the CoT and the final answer. We prompt the LVLMs to generate CoT with less than five sentences, and additionally remove CoT with more than five sentences. The prompt template and an example output are shown in Figure 23.

Response filtering. In the response filtering stage, we filter the responses into three subsets: responses with outcome errors, responses with process errors, and responses with no errors. Examples from each of the three subsets are shown in Figure 20.

Critique collection. We train three human annotators to provide dense and fine-grained critique with a binary label for each step, and a natural language explanation if the step is considered as incorrect. The annotation interface is shown in Figure 24.

B. Metric Design Details

VISCore design. To verify the robustness of VISCore metric design, we conducted human evaluation on 200 data where annotators compare pairs of model generated critique and judge which is better. We find that VISCore correlates the best with human judgment ($\tau = 0.58$), outperforming alternatives like arithmetic mean of F1 ($\tau = 0.42$) and exlanation BLEU ($\tau = 0.41$). This is because VISCore prioritizes explanation correctness, better aligning with human focus, and LLM-based evaluation of explanation is more effective than BLEU.

LLM evaluation. The calculation of explanation-level F1 requires matching model-generated critique explanations against human-annotated explanations using LLMs. We use GPT-4o for this evaluation, and the prompt is shown in Figure 26. To evaluate the reliability of this evaluation, we manually compare 100 model-generated explanations against human explanations, and evaluate the agreement between our manual evaluation and LLM evaluation. We observe a high agreement of 0.80 accuracy and 0.61 Cohen's Kappa, validating the robustness of LLM evaluation.

Metric calculation examples. To better illustrate the metric calculation, we present two examples in Figure 25 and 22, illustrating the calculation of VISCore for critique task and correction gain for correction task respectively.

C. Experimental Details

We evaluate 27 open LVLMs and 3 proprietary LVLMs. For the open LVLMs, we evaluate models from 7 families: DeepSeek-VL [26], LLaVA [23], InternVL [7], Qwen-VL [39], Molmo [9], NVLM [8], Llama-3.2, and MiniCPM [45]. Specifically, we run the inference with fast serving frameworks. Specifically, we evaluate Qwen2-VL, Molmo, Llama-3.2, NVLM and MiniCPM with `vllm`, evaluate LLaVA-OV and LLaVA-Critic with `sglang`, and evaluate InternVL2, DeepSeek-VL, LLaVA-v1.5, LLaVA-v1.6, Qwen-VL and Prometheus-Vision with `lmdeploy`. We set the sampling temperature as 0.7. For LOOKBACK, we use the same set of hyperparameters. The prompt for critique and correction are in Figure 27 and 28.

For the human baseline, we ask one of the trained annotators to establish the human baseline on a randomly selected subset of 265 data points. To reduce annotation costs, we provide the ground truth answers to the annotator. The 100% answer-level F1 shows that the ground truth answers are verified by the annotator to be correct.

D. Additional Results

In this section, we present additional experimental results. Figure 21 is a more complete version of Figure 9 in the main content, showing the correction performance given model-generated or human-generated critiques with different granularity. We further report the detailed critique performance of each model at different granularity and categories in Table 6, and include a few additional models like LLaVA-v1.5.

To measure the importance of CoT to our framework, we perform ablation study by evaluating the critique and correction performance without CoT in the original response. As shown below, LVLMs can effectively perform critique and correction bringing positive correction gains without CoT. However, CoT brings better performance especially for strong models like GPT-4o.

Model	Critique		Correction	
	w. CoT	w/o CoT	w. CoT	w/o CoT
InternVL2-8B	37.1	**47.3**	**5.4**	5.1
Claude-3.5-Sonnet	**61.8**	60.3	**25.6**	**25.6**
GPT-4o	**63.0**	59.0	**28.8**	26.4

Table 5. Additional experiments for ablations of CoT.

E. Details on LookBack

The algorithm for our proposed LookBack method is in Alg. 1. We further show an example of critique generated by LookBack in Figure 29.

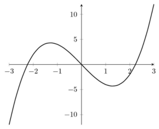
Algorithm 1 LookBack algorithm

```
Require: LVLM, image $I$, question $q$, model response to be critiqued $(\mathbf{T}_{\mathcal{M}}, a_{\mathcal{M}})$
$\mathcal{Q} = \{\}$
for $i = 1$ to $N$ do
    $\mathcal{Q}_i = RaiseQuestion(\text{LVLM}, I, T_i)$
    $\mathcal{Q} = \mathcal{Q} \cup \mathcal{Q}_i$
end for
$\mathcal{Q} = \mathcal{Q} \cup \{q\}$
$\mathcal{A} = \{\}$
for $q'$ in $\mathcal{Q}$ do
    $a' = AnswerQuestion(\text{LVLM}, I, q')$
    $\mathcal{A} = \mathcal{A} \cup \{a'\}$
end for
return $Critique\,(\text{LVLM}, I, q, (\mathbf{T}_{\mathcal{M}}, a_{\mathcal{M}}), \mathcal{Q}, \mathcal{A})$
```

Error cases of LookBack. We further evaluate the error cases of LookBack and identify two typical scenarios: (1) Errors in reasoning critique. LookBack is less effective in critiquing reasoning, as these often go beyond verifying atomic information against the image. This is reflected in the smaller performance boosts in reasoning tasks in Table 4. (2) Referral expression error in atomic information extraction. In an example involving four images, the reasoning discusses a blue car *in the first image*, but the extracted question refers to the car as *the blue car* without specifying the first image, resulting in critique errors.

F. Limitations

The main limitation of this work is its task setting, which constrains the CoT to at most five reasoning steps, potentially affecting its generalization to complex reasoning or multi-turn interactions. While this work provides a valuable initial investigation into LVLM critique and correction, this limitation highlights opportunities for future research.



Question

What is the center of symmetry of this function?

Choices:

(A) (0, 0)

(B) (-1, 0)

(C) (2, 0)

1. A function with a center of symmetry at a point (a, b) implies the graph of the function is symmetric when rotated 180 degrees about this point.

✓ CORRECT

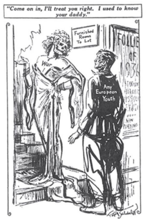
2. Observing the graph, it appears that this symmetry exists around the point (2, 0).

✕ INCORRECT The symmetry exists around (0, 0), not (2, 0).

3. Points on opposite sides of (2, 0) have values equidistant from this center, confirming symmetry around this point.

✕ INCORRECT Points on opposite sides of (2, 0) do not have values equidistant from this center.

Answer: (C) (2, 0) ✕ INCORRECT


“Come on in, I’ll treat you right. I used to know your daddy.”

—C. D. Batchelor, *New York Daily News*, April 25, 1936

Question:

The 1936 cartoon above, from the New York Daily News, is making the point that

A. although European individuals and countries might be seduced into waging another major war, the United States would be wise to avoid participating.

B. the policy of appeasement is a bankrupt policy that can only lead to more death and destruction.

C. munitions manufacturers, the so-called merchants of death, were pushing the world toward war in the name of profits.

D. the weaponry of modern warfare had advanced to such a degree that future military engagements would result in unprecedented carnage.

1. The cartoon is making a statement about the consequences of war and the potential for future military engagements to result in unprecedented carnage.

✓ CORRECT

2. It suggests that European countries might be drawn into another major war, but the United States should be cautious and avoid participating.

✓ CORRECT

3. The cartoon also implies that the advanced weaponry of modern warfare has made such conflicts more destructive.

✕ INCORRECT Such information is not implied by the image.

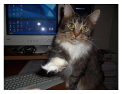
4. This reasoning process leads to answer C, which is not one of the options provided.

✕ INCORRECT Answer C ("munitions manufacturers, the so-called merchants of death, were pushing the world toward war in the name of profits") is provided, but is incorrect. Such information is not implied by the image.

5. Therefore, the correct answer is C, which is not listed as an option.

✕ INCORRECT Answer C ("munitions manufacturers, the so-called merchants of death, were pushing the world toward war in the name of profits") is provided, but is incorrect. Such information is not implied by the image.

Answer: C. munitions manufacturers, the so-called merchants of death, were pushing the world toward war in the name of profits. ✕ INCORRECT



Question: Determine whether the following statement about the image is true or false: The cat is facing the keyboard.

1. The cat is facing the keyboard, with its front paws extended, which indicates that it is interacting with the keyboard.

✕ INCORRECT The cat is not facing the keyboard, but facing the camera.

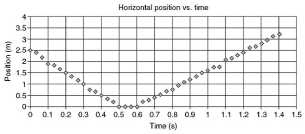
2. Its eyes are also focused on the keyboard area, which suggests that it is either looking at the keyboard or something on it.

✕ INCORRECT The cats' eyes are focused on the camera's direction, not on the keyboard.

3. The position of the cat and its posture imply that it is facing the keyboard.

✕ INCORRECT The cat is not facing the keyboard, but facing the camera.

Answer: True ✕ INCORRECT



Question

A 50.0-g ball moving to the left strikes a wall and bounces back to the right. Slow-motion video analysis produces the following horizontal position versus time graph. Determine the momentum change of the bouncing ball.

A. +0.05 kg × m/s B. +0.45 kg × m/s

C. -0.45 kg × m/s D. -0.05 kg × m/s

1. Before the collision, the ball moves left from about t=0.1s to t=0.5s, showing a change in slope indicating velocity to the left.

✕ INCORRECT There is no change in slope. The ball is moving at the same speed.

2. After the collision around t=0.5s, the ball's position reverses direction, indicating a bounce back or reversal of velocity.

✓ CORRECT

3. The momentum change is calculated as mass times change in velocity, which depends on the slope before and after t=0.5s.

✓ CORRECT

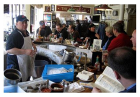
4. The slope magnitude roughly appears similar, suggesting similar speeds, making the approximate velocity change twice the initial speed upon reversal.

✕ INCORRECT The speeds are similar but different. The speed after t = 0.6s can be calculated by the points on the graph.

5. Therefore, the change in momentum is around -0.05 kg*m/s moving left to +0.05 kg*m/s moving right, giving an overall change of +0.1 kg*m/s, aligning with the correct answer of +0.05 kg*m/s given minor calculation or graph estimation issues.

✕ INCORRECT This is incorrect, as it is based on a previous incorrect identification of speed in sentence 4.

Answer: A. +0.05 kg × m/s ✕ INCORRECT



Question: Is there a knife in the image?

1. The image shows a group of people seated around what appears to be an oyster bar.

✓ CORRECT

2. There are some dishes and condiments on the counter.

✓ CORRECT

3. Upon closer inspection, there is a shiny object that resembles a knife near the oysters.

✕ INCORRECT There is no knife near the oysters.

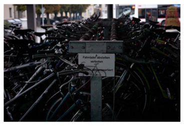
4. The shape and position suggest it is likely used for shucking oysters.

✕ INCORRECT There is no knife near the oysters.

5. Thus, the presence of a knife is confirmed.

✕ INCORRECT There is no knife near the oysters.

Answer: Yes ✕ INCORRECT



Question: what is the first word on the sign?

1. The image shows a sign in German, which reads 'Fahrrad abstellen verboten' in English.

✕ INCORRECT The sign reads "Fahhrader abstellen verboten".

2. This translates to 'Parking bicycles prohibited'.

✓ CORRECT

3. The sign is located in a bike rack area, surrounded by numerous bicycles.

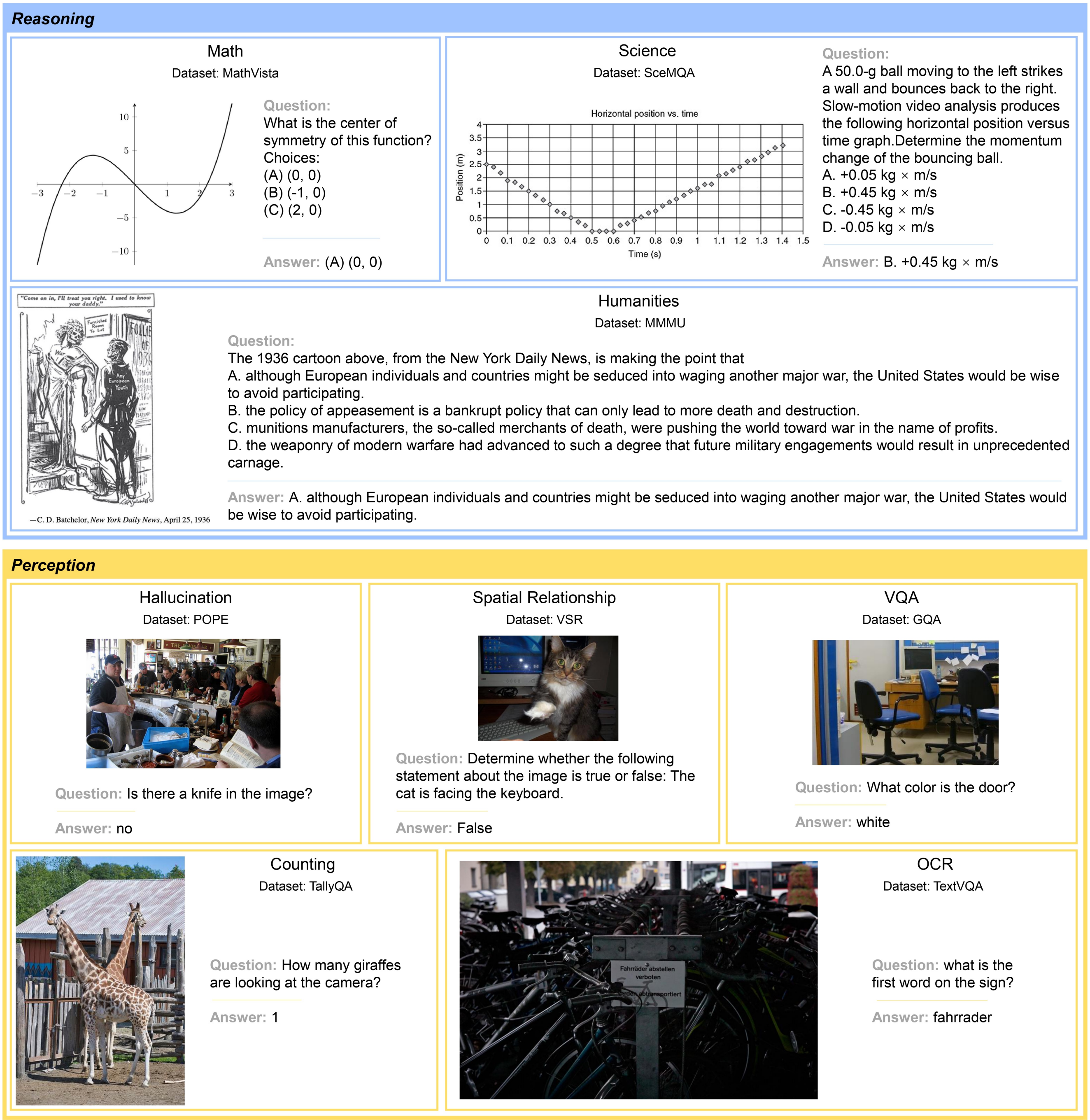
✓ CORRECT

4. The presence of the sign indicates that bicycle parking is not allowed in this particular area.

✓ CORRECT

Answer: Bicycle parking prohibited ✕ INCORRECT

Figure 18. Dataset examples.



Math — Question: What is the center of symmetry of this function? Choices: (A) (0, 0) (B) (-1, 0) (C) (2, 0)

Answer: (A) (0, 0)

Science — Question: A 50.0-g ball moving to the left strikes a wall and bounces back to the right. Slow-motion video analysis produces the following horizontal position versus time graph.Determine the momentum change of the bouncing ball.

A. +0.05 kg × m/s

B. +0.45 kg × m/s

C. -0.45 kg × m/s

D. -0.05 kg × m/s

Answer: B. +0.45 kg × m/s

Humanities — —C. D. Batchelor, *New York Daily News*, April 25, 1936

Question:

The 1936 cartoon above, from the New York Daily News, is making the point that

A. although European individuals and countries might be seduced into waging another major war, the United States would be wise to avoid participating.

B. the policy of appeasement is a bankrupt policy that can only lead to more death and destruction.

C. munitions manufacturers, the so-called merchants of death, were pushing the world toward war in the name of profits.

D. the weaponry of modern warfare had advanced to such a degree that future military engagements would result in unprecedented carnage.

Answer: A. although European individuals and countries might be seduced into waging another major war, the United States would be wise to avoid participating.

Hallucination — Question: Is there a knife in the image?

Answer: no

Spatial Relationship — Question: Determine whether the following statement about the image is true or false: The cat is facing the keyboard.

Answer: False

VQA — Question: What color is the door?

Answer: white

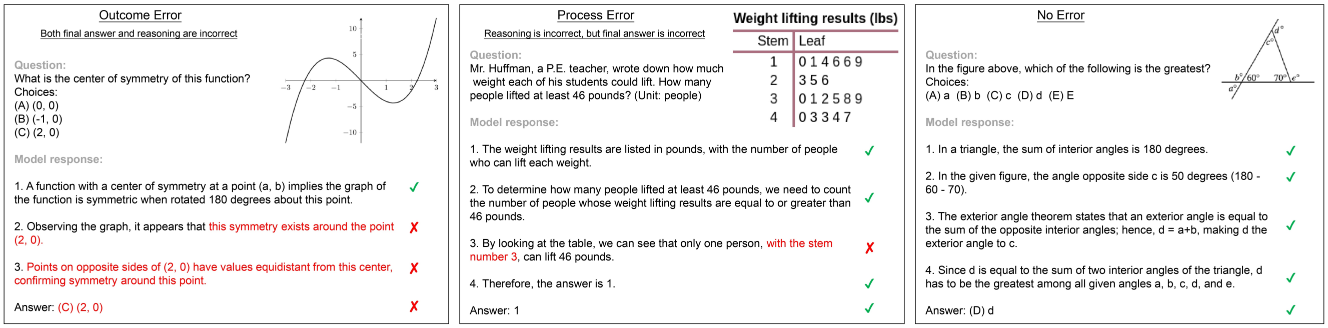
Counting — Question: How many giraffes are looking at the camera?

Answer: 1

OCR — Question: what is the first word on the sign?

Answer: fahrrader

Figure 19. Examples for each task in our dataset.



Outcome Error
Both final answer and reasoning are incorrect
10
5
−5
−10
−3
−2
−1
1
2
3

Question:
What is the center of symmetry of this function?
Choices:
(A) (0, 0)
(B) (-1, 0)
(C) (2, 0)

Model response:

1. A function with a center of symmetry at a point (a, b) implies the graph of the function is symmetric when rotated 180 degrees about this point. ✓

2. Observing the graph, it appears that this symmetry exists around the point (2, 0). ✗

3. Points on opposite sides of (2, 0) have values equidistant from this center, confirming symmetry around this point. ✗

Answer: (C) (2, 0) ✗

Question:
Mr. Huffman, a P.E. teacher, wrote down how much weight each of his students could lift. How many people lifted at least 46 pounds? (Unit: people)

Weight lifting results (lbs)

Stem	Leaf
1	0 1 4 6 6 9
2	3 5 6
3	0 1 2 5 8 9
4	0 3 3 4 7

Model response:

1. The weight lifting results are listed in pounds, with the number of people who can lift each weight. ✓

2. To determine how many people lifted at least 46 pounds, we need to count the number of people whose weight lifting results are equal to or greater than 46 pounds. ✓

3. By looking at the table, we can see that only one person, with the stem number 3, can lift 46 pounds. ✗

4. Therefore, the answer is 1. ✓

Answer: 1 ✓

Question:
In the figure above, which of the following is the greatest?
Choices:
(A) a (B) b (C) c (D) d (E) E

Model response:

1. In a triangle, the sum of interior angles is 180 degrees. ✓

2. In the given figure, the angle opposite side c is 50 degrees (180 - 60 - 70). ✓

3. The exterior angle theorem states that an exterior angle is equal to the sum of the opposite interior angles; hence, d = a+b, making d the exterior angle to c. ✓

4. Since d is equal to the sum of two interior angles of the triangle, d has to be the greatest among all given angles a, b, c, d, and e. ✓

Answer: (D) d ✓

Figure 20. Examples of model responses with outcome errors, process errors, and no errors.

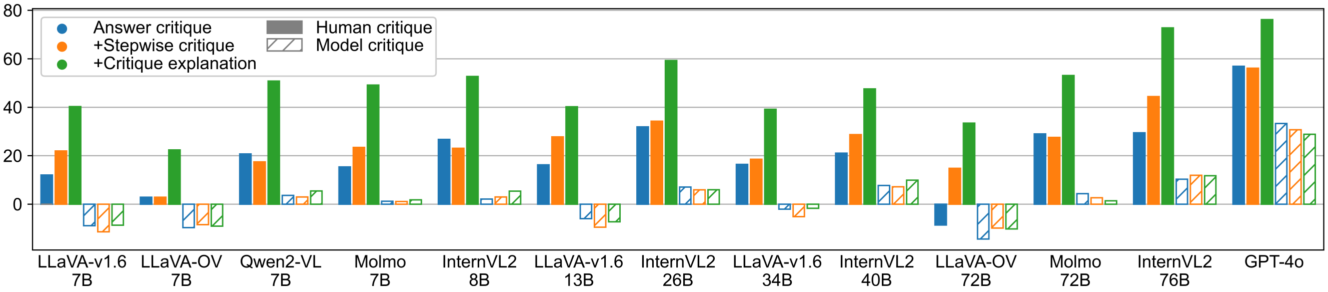
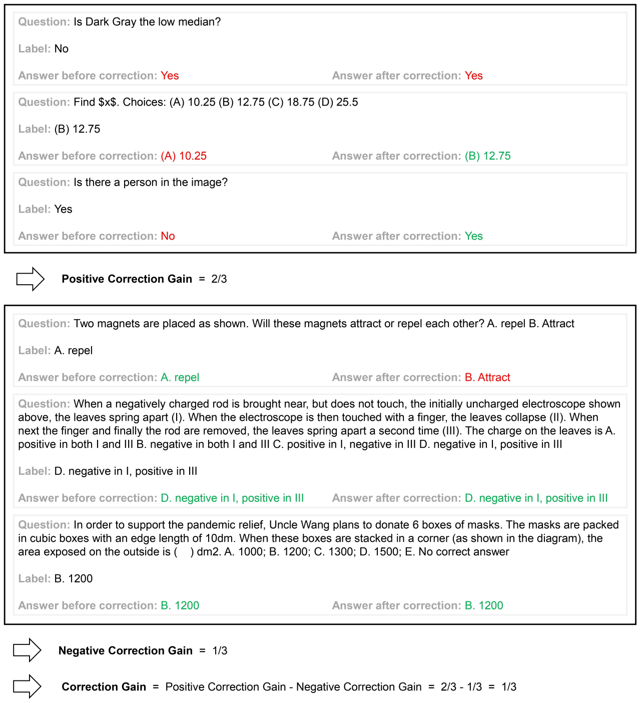


Figure 21. Correction performance given model-generated or human-generated critiques with different granularity.



Question: Is Dark Gray the low median?
Label: No
Answer before correction: Yes — Answer after correction: Yes

Question: Find x. Choices: (A) 10.25 (B) 12.75 (C) 18.75 (D) 25.5
Label: (B) 12.75
Answer before correction: (A) 10.25 — Answer after correction: (B) 12.75

Question: Is there a person in the image?
Label: Yes
Answer before correction: No — Answer after correction: Yes

Positive Correction Gain = 2/3

Question: Two magnets are placed as shown. Will these magnets attract or repel each other? A. repel B. Attract
Label: A. repel
Answer before correction: A. repel — Answer after correction: B. Attract

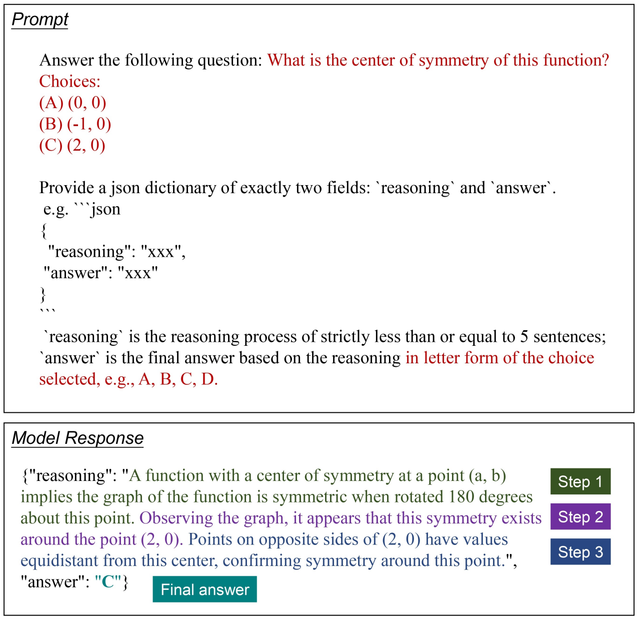
Question: When a negatively charged rod is brought near, but does not touch, the initially uncharged electroscope shown above, the leaves spring apart (I). When the electroscope is then touched with a finger, the leaves collapse (II). When next the finger and finally the rod are removed, the leaves spring apart a second time (III). The charge on the leaves is A. positive in both I and III B. negative in both I and III C. positive in I, negative in III D. negative in I, positive in III
Label: D. negative in I, positive in III
Answer before correction: D. negative in I, positive in III — Answer after correction: D. negative in I, positive in III

Question: In order to support the pandemic relief, Uncle Wang plans to donate 6 boxes of masks. The masks are packed in cubic boxes with an edge length of 10dm. When these boxes are stacked in a corner (as shown in the diagram), the area exposed on the outside is ( ) dm2. A. 1000; B. 1200; C. 1300; D. 1500; E. No correct answer
Label: B. 1200
Answer before correction: B. 1200 — Answer after correction: B. 1200

Negative Correction Gain = 1/3

Correction Gain = Positive Correction Gain - Negative Correction Gain = 2/3 - 1/3 = 1/3

Figure 22. Calculation of correction gain.



Prompt

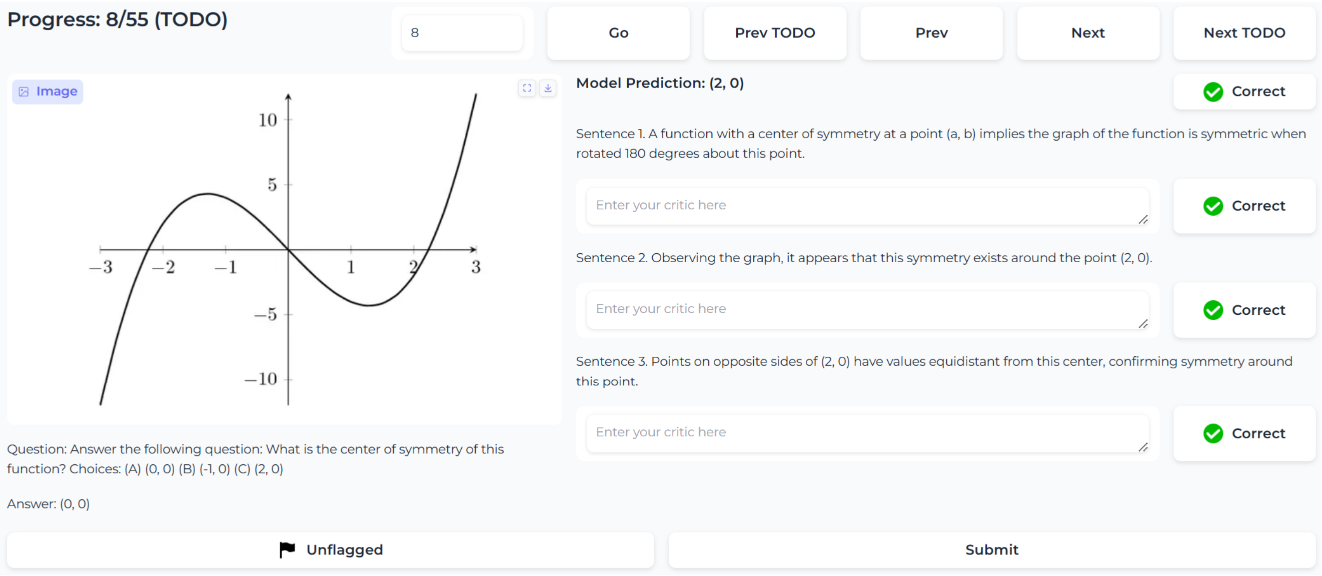
~~~
Answer the following question: What is the center of symmetry of this function?
Choices:
(A) (0, 0)
(B) (-1, 0)
(C) (2, 0)

Provide a json dictionary of exactly two fields: `reasoning` and `answer`.
 e.g. ```json
{
  "reasoning": "xxx",
 "answer": "xxx"
}
```
 `reasoning` is the reasoning process of strictly less than or equal to 5 sentences;
`answer` is the final answer based on the reasoning in letter form of the choice
selected, e.g., A, B, C, D.
~~~

Model Response

{"reasoning": "A function with a center of symmetry at a point (a, b) implies the graph of the function is symmetric when rotated 180 degrees about this point. Observing the graph, it appears that this symmetry exists around the point (2, 0). Points on opposite sides of (2, 0) have values equidistant from this center, confirming symmetry around this point.", "answer": "C"}

Figure 23. Prompt template for collecting model responses and an example response. The highlighted fields will be filled with the information according to each data point. The "reasoning" field in the model response will be split into multiple reasoning steps using `nltk.tokenize.sent_tokenize`.



Progress: 8/55 (TODO)

8

Go | Prev TODO | Prev | Next | Next TODO

Image

Question: Answer the following question: What is the center of symmetry of this function? Choices: (A) (0, 0) (B) (-1, 0) (C) (2, 0)

Answer: (0, 0)

Model Prediction: (2, 0) ✓ Correct

Sentence 1. A function with a center of symmetry at a point (a, b) implies the graph of the function is symmetric when rotated 180 degrees about this point.

Enter your critic here ✓ Correct

Sentence 2. Observing the graph, it appears that this symmetry exists around the point (2, 0).

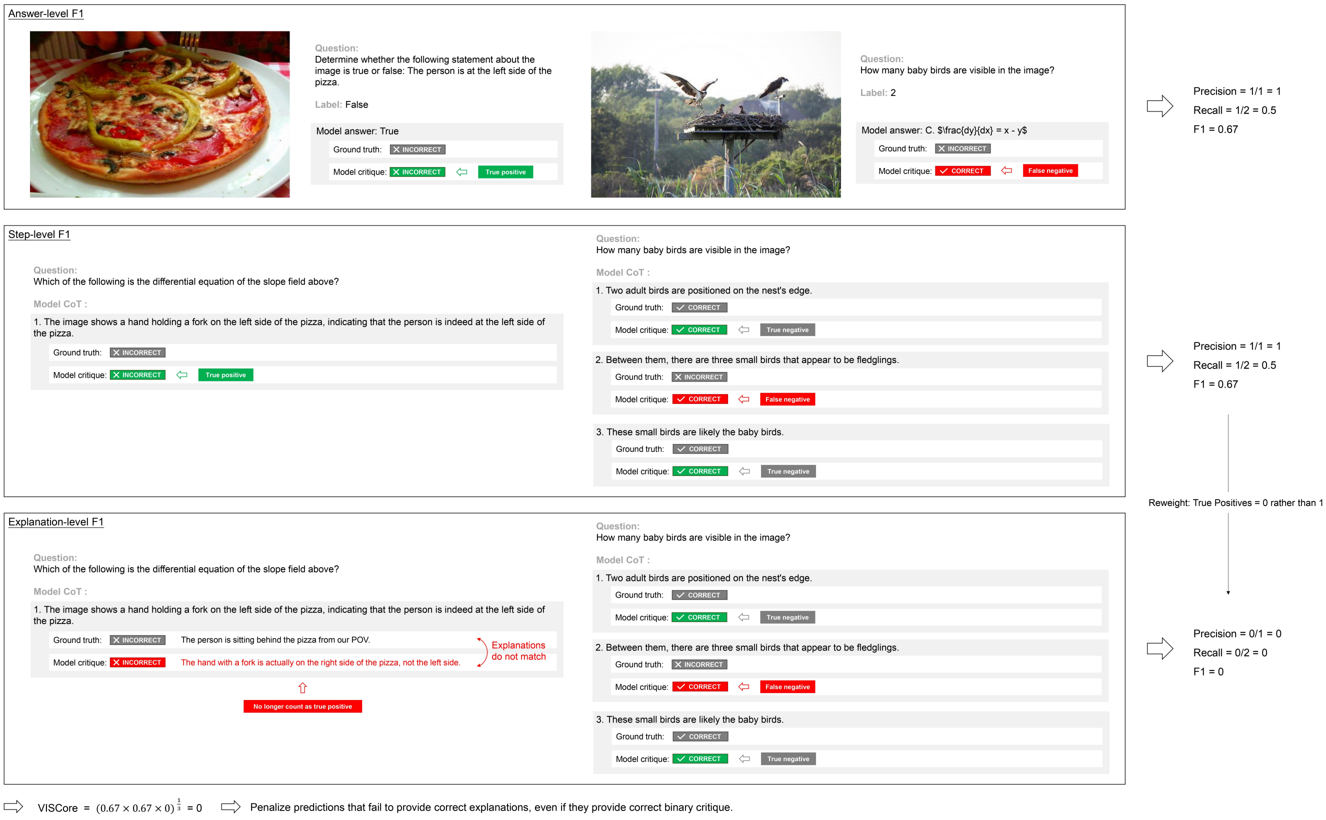
Enter your critic here ✓ Correct

Sentence 3. Points on opposite sides of (2, 0) have values equidistant from this center, confirming symmetry around this point.

Enter your critic here ✓ Correct

Unflagged | Submit

Figure 24. Interface of critique annotation.



Answer-level F1

Question:
Determine whether the following statement about the image is true or false: The person is at the left side of the pizza.

Label: False

Model answer: True
Ground truth: ✗ INCORRECT
Model critique: ✗ INCORRECT ⇦ True positive

Question:
How many baby birds are visible in the image?

Label: 2

Model answer: C. $\frac{dy}{dx} = x - y$
Ground truth: ✗ INCORRECT
Model critique: ✓ CORRECT ⇦ False negative

Precision = 1/1 = 1
Recall = 1/2 = 0.5
F1 = 0.67

Step-level F1

Question:
Which of the following is the differential equation of the slope field above?

Model CoT :

1. The image shows a hand holding a fork on the left side of the pizza, indicating that the person is indeed at the left side of the pizza.
Ground truth: ✗ INCORRECT
Model critique: ✗ INCORRECT ⇦ True positive

Question:
How many baby birds are visible in the image?

Model CoT :

1. Two adult birds are positioned on the nest's edge.
Ground truth: ✓ CORRECT
Model critique: ✓ CORRECT ⇦ True negative

2. Between them, there are three small birds that appear to be fledglings.
Ground truth: ✗ INCORRECT
Model critique: ✓ CORRECT ⇦ False negative

3. These small birds are likely the baby birds.
Ground truth: ✓ CORRECT
Model critique: ✓ CORRECT ⇦ True negative

Precision = 1/1 = 1
Recall = 1/2 = 0.5
F1 = 0.67

Reweight: True Positives = 0 rather than 1

Explanation-level F1

Question:
Which of the following is the differential equation of the slope field above?

Model CoT :

1. The image shows a hand holding a fork on the left side of the pizza, indicating that the person is indeed at the left side of the pizza.
Ground truth: ✗ INCORRECT The person is sitting behind the pizza from our POV.
Model critique: ✗ INCORRECT The hand with a fork is actually on the right side of the pizza, not the left side.

Explanations do not match

No longer count as true positive

Question:
How many baby birds are visible in the image?

Model CoT :

1. Two adult birds are positioned on the nest's edge.
Ground truth: ✓ CORRECT
Model critique: ✓ CORRECT ⇦ True negative

2. Between them, there are three small birds that appear to be fledglings.
Ground truth: ✗ INCORRECT
Model critique: ✓ CORRECT ⇦ False negative

3. These small birds are likely the baby birds.
Ground truth: ✓ CORRECT
Model critique: ✓ CORRECT ⇦ True negative

Precision = 0/1 = 0
Recall = 0/2 = 0
F1 = 0

⇨ VISCore = $(0.67 \times 0.67 \times 0)^{\frac{1}{3}} = 0$ ⇨ Penalize predictions that fail to provide correct explanations, even if they provide correct binary critique.

Figure 25. Calculation of VISCore.

	VISCore	Total			Reasoning			Perception		
		Answer	Thought	Expl.	Answer	Step	Expl.	Answer	Step	Expl.
Random	-	37.91	32.02	-	39.44	33.10	-	36.03	30.54	-
Tiny-size (<3B) Open LVLMs										
Qwen2-VL-2B	9.76	30.7	23.0	1.3	29.5	24.3	1.7	31.9	21.3	0.8
InternVL2-2B	**13.96**	**36.1**	**28.2**	**2.7**	**37.3**	**29.3**	**3.7**	**34.8**	**26.8**	**1.5**
Small-size (∼7B) Open LVLMs										
DeepSeek-VL-7B	7.53	21.8	15.7	1.2	18.7	14.4	1.6	24.9	17.5	0.8
LLaVA-v1.5-7B	6.51	11.9	13.4	1.7	12.7	13.5	2.0	10.9	13.3	1.4
LLaVA-v1.6-7B	21.80	**44.6**	**33.6**	6.9	**46.8**	34.6	7.8	**41.3**	**31.9**	5.6
LLaVA-v1.6-Vicuna-7B	11.45	27.7	24.1	2.3	25.8	24.3	2.2	29.9	23.8	2.4
LLaVA-OV-7B	7.53	14.5	14.9	2.0	16.8	15.7	2.2	11.7	13.8	1.6
Qwen-VL-7B	12.69	33.1	26.4	2.3	31.2	26.7	2.8	35.1	26.0	1.8
Qwen2-VL-7B	21.71	43.0	30.6	7.8	44.5	30.4	7.3	41.1	31.0	**8.4**
Molmo-7B	13.43	35.5	22.0	3.1	35.7	23.6	4.1	35.3	19.8	1.6
InternVL2-8B	**23.33**	37.1	31.1	**11.0**	44.2	**37.4**	**14.1**	26.1	19.7	5.6
MiniCPM-V2.6 (8B)	13.07	27.9	18.2	4.4	32.0	21.7	5.5	22.4	12.4	2.7
Medium-size (10∼70B) Open LVLMs										
Llama-3.2-11B	11.44	29.4	21.1	2.4	31.4	23.0	3.3	26.9	17.6	0.8
LLaVA-v1.6-13B	21.02	40.2	**32.8**	7.1	43.6	36.5	7.6	**34.4**	**25.9**	6.1
InternVL2-26B	25.20	39.4	30.2	13.4	**48.7**	36.8	16.9	24.5	18.5	7.2
LLaVA-v1.6-34B	11.05	23.6	14.3	4.0	29.4	17.4	4.2	15.3	9.5	3.6
InternVL2-40B	**28.48**	**41.6**	31.4	**17.7**	47.8	**37.4**	**20.7**	32.0	20.7	**12.4**
Large-size (>70B) Open LVLMs										
LLaVA-OV-72B	35.27	47.1	42.0	22.2	49.3	44.7	23.2	**43.9**	**37.3**	**20.5**
Qwen2-VL-72B	**37.44**	49.2	41.9	**25.5**	56.3	49.0	**30.8**	38.0	28.6	15.5
NVLM-72B	33.07	44.0	38.6	21.3	52.6	46.1	26.1	30.6	24.7	12.3
Molmo-72B	35.59	**49.4**	39.8	22.9	**57.5**	46.8	26.8	36.5	26.4	15.5
InternVL2-76B	26.38	37.7	28.6	17.0	47.3	35.6	21.2	22.9	16.2	9.6
Llama-3.2-90B	36.40	46.8	**42.5**	24.3	55.7	**49.1**	28.8	32.9	30.5	16.1
Critique LVLMs										
Prometheus-Vision-7B	17.67	37.6	35.8	4.1	37.2	37.6	3.1	38.0	33.4	5.4
LLaVA-Critic-7B	20.02	32.0	28.7	8.8	36.3	30.3	9.5	26.3	26.3	7.7
Prometheus-Vision-13B	19.32	38.0	37.8	5.0	42.0	40.1	4.6	33.1	34.5	5.6
LLaVA-Critic-72B	**42.60**	**53.9**	**50.9**	**28.2**	**56.3**	**54.8**	**28.6**	**50.5**	**44.8**	**27.4**
Proprietary LVLMs										
Gemini-1.5-Pro	45.01	55.6	51.2	32.0	58.8	55.0	35.7	51.2	45.2	26.3
Claude-3.5-Sonnet	51.28	61.8	**58.1**	37.6	**65.4**	**61.6**	43.5	57.4	**53.2**	29.5
GPT-4o	**52.36**	**63.0**	57.2	**39.8**	65.2	61.4	**43.5**	**60.1**	50.3	**34.0**
Human*	86.47	100.0	90.6	71.4	100.0	90.6	66.8	100.0	90.6	76.4

Table 6. **Detailed critique performance VISCO.** In addition to the overall VISCore score, we also report the critique F1 at three granularity: answer-level F1, step-level F1, and explanation-level F1. We also report the scores for the two categories, reasoning and perception respectively. We also include a few additional models compared to Table 3 in the main content.

You task is to evaluate a **critique** of reasoning.

You are given a question about an image, an incorrect chain-of-thought trying to answer the question, and a **critique** that explains why the `STEP_NUM` step of the chain-of-thought is incorrect. You are required to focus on the `STEP_NUM` step, and analyzes whether the critique correctly identifies the source of error.

For reference, you will be provided with the ground truth critique for each individual step. Evaluate the critique by comparing against the ground truth. Focus on whether the critique accurately identifies the **core mistake**. If the critique addresses the core error and does not contain factual or logical error, minor deviations or omissions in reasoning or explanation should be considered as correct.

Think step by step, then provide your judgment. Your response should end with either:

# Judgment: the critique is correct.

or

# Judgment: the critique is incorrect.

`... In-context examples ...`

# Question: `QUESTION`

# Chain-of-thought: `CoT`

Figure 26. **Prompt for LLM-assisted evaluation of critique explanation.** The highlighted fields are to be filled according to each data point. We omit the in-context examples here.

You are given an image, a question about the image, a reasoning process involving multiple steps, and a final answer. Evaluate the accuracy of each reasoning step and the final answer. For each reasoning step, evaluate whether it is correct or incorrect. If it is incorrect, briefly explain why. Then, evaluate the final answer as correct or incorrect.

Your response should be a json in the following format:

```json
{
"step_1": {"correctness": true or false, "explanation": "Write your explanation here"},
...,
"answer_correctness": true or false
}
```

—

Instructions for evaluating reasoning steps:
* For each reasoning step, evaluate whether it is correct or incorrect based on the accuracy of the factual information and logical calculations it contains.
* Evaluate each step in isolation.
* You do not need to evaluate the importance of the step in achieving the correct final answer; focus solely on the correctness within that step itself.

—

# Question: `QUESTION`

# Reasoning: `CoT`

# Answer: `ANSWER`

Figure 27. **Prompt for critique task.** The highlighted fields are to be filled according to each data point.

You are given an image, a question about the image, a multi-step reasoning process leading to an answer, and the critique for each reasoning step. Based on this information, think step by step and provide the correct answer. Your response should end with a json dictionary as follows:

```json
{"answer": ANSWER}
```

ANSWER should be `ANSWER_FORMAT`.

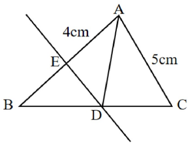
# Question: `QUESTION`

# Reasoning:
1. `REASONING_STEP_1`
  - Critique: `CRITIQUE_FOR_STEP_1`

`...`

The final answer is: `MODEL_RESPONSE_ANSWER`
  - Critique: `CRITIQUE_FOR_ANSWER`

Figure 28. **Prompt for correction task.** The highlighted fields are to be filled according to each data point.



Question:
As shown in the figure, triangle ABC is folded along the straight line DE so that point B coincides with point A. If the perimeter of triangle ADC is 17 cm, what is the perimeter of triangle ABC in cm? A. 17; B. 12; C. 15; D. 25; E. No correct answer

Label: D. 25

Model CoT:

1. Since triangle ABC is folded such that point B coincides with point A, line segment AB is equal to line segment AE.

Ground truth: ✕ INCORRECT AB is not equal to AE, but it is twice as long as AE.

Baseline: ✓ CORRECT

LookBack:

Questions regarding atomic information:		*LVLM answers:*
Is line segment AB equal to line segment AE?	⇨	No, line segment AB is not equal to line segment AE. Line segment AE is given as 4 cm, while no length is provided for line segment AB.

✕ INCORRECT The reasoning incorrectly states that AB equals AE. AE is given as 4 cm, but AB is not specified and appears longer.

2. Given that AD is 5 cm and AC is 4 cm, line segment DC must be equal to DE.

Ground truth: ✕ INCORRECT DC does not equal DE. There is no explicit relationship between DC and DE.

Baseline: ✓ CORRECT

LookBack:

Questions regarding atomic information:		*LVLM answers:*
Is line segment DC equal to line segment DE?	⇨	No, line segment DC is not equal to line segment DE.

✕ INCORRECT The reasoning states DC equals DE, which is incorrect based on the visual information provided.

3. Therefore, ADC's perimeter AD+DC+AC is 17 cm.

Ground truth: ✓ CORRECT

Baseline: ✓ CORRECT

LookBack:

Questions regarding atomic information:		*LVLM answers:*
Is line segment AD labeled as 5 cm?	⇨	Yes, line segment AD is labeled as 5 cm.
Is line segment AC labeled as 4 cm?		No, line segment AC is labeled as 5 cm.

✕ INCORRECT The premise is incorrect since the lengths used (AD, DC, AC) do not add up correctly based on given information.

4. If AE=AB, this implies the same triangle length, making the total perimeter of ABC also 17 cm as position symmetrical.

Ground truth: ✕ INCORRECT AE does not equal AB. The perimeter of ABC is not 17cm.

Baseline: ✓ CORRECT

LookBack:

Questions regarding atomic information:		*LVLM answers:*
Is line segment AE equal to line segment AB?	⇨	No, line segment AE is not equal to line segment AB. AE is 4 cm, while AB is not labeled with a length but appears longer than AE.
Does line segment AD measure 5 cm?		No, line segment AD does not measure 5 cm; it is AC that measures 5 cm.
Does line segment AC measure 4 cm?		No, line segment AC measures 5 cm.

✕ INCORRECT The reasoning states DC equals DE, which is incorrect based on the visual information provided.

5. Thus, perimeter of ABC is the same as ADC.

Ground truth: ✕ INCORRECT AE does not equal AB. The perimeter of ABC is not 17cm.

Baseline: ✓ CORRECT

LookBack:

Questions regarding atomic information:		*LVLM answers:*
Is AE equal to AB in the image?	⇨	No, AE is not equal to AB. AE is 4 cm, while AB is not labeled with a specific measurement.
Is the length of AD labeled as 5 cm in the image?		Yes, the length of AD is labeled as 5 cm in the image.
Is the length of AC labeled as 4 cm in the image?		No, the length of AC is labeled as 5 cm.

✕ INCORRECT The reasoning incorrectly concludes that the perimeters are the same based on incorrect assumptions.

Model answer:

A. 17

Ground truth: ✕ INCORRECT

Baseline: ✓ CORRECT

LookBack: ✕ INCORRECT

Figure 29. Example of LookBack.