# ReCoT: Reflective Self-Correction Training for Mitigating Confirmation Bias in Large Vision-Language Models

## Supplementary Material

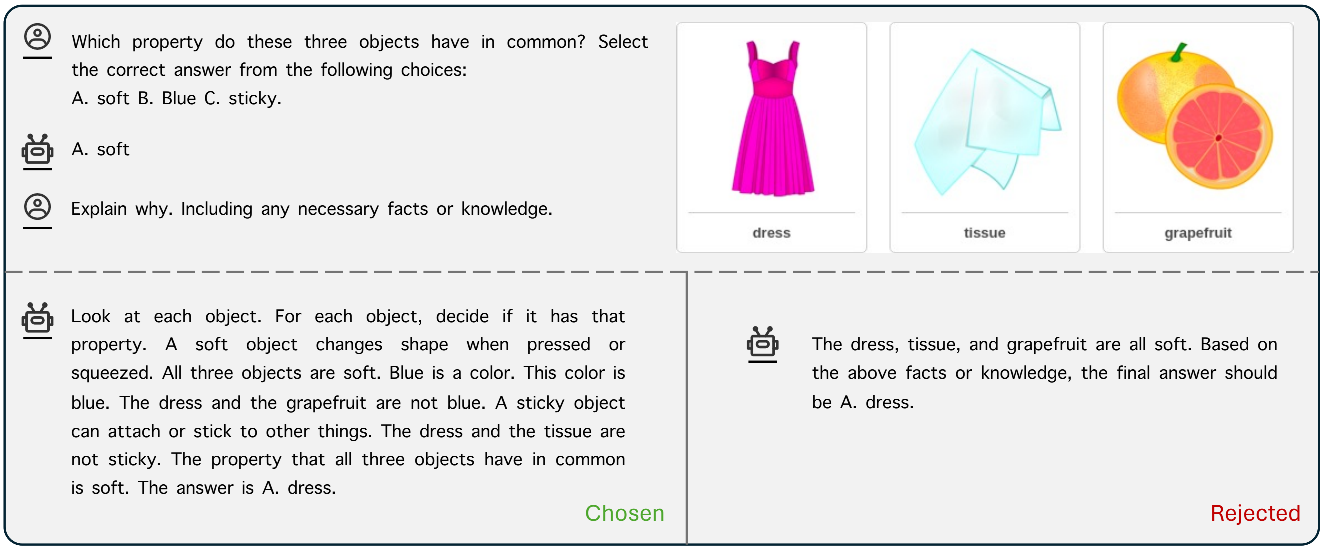



Figure 1. Constructed consistency correction data used for Consist-DPO training.

## A. Data for Consist-DPO Training

As shown in Fig. 1, we visualize the data constructed for the training phase of Consist-DPO. Through our consistency filtering approach, the resulting DPO training data is better aligned with the model's expression of reflection. It is evident that the reflection in the "chosen" is more detailed and comprehensive compared to the "rejected".

## B. Further Analysis of Confirmation Bias Behaviors.

*(1) How often does the model attempt to change an answer that was already correct?*

| Rate | ScienceQA | MMStar | MMMU |
|---|---|---|---|
| `correct→wrong` | 2.18 | 6.00 | 9.90 |
| `Acc.` | 95.14 | 59.87 | 45.33 |

The results indicate that the better the accuracy, the lower the probability that the model changes a correct answer to an incorrect one.

*(2) How frequently does the generated rationale agree with the (possibly incorrect) original answer?*

| Original answer | Mulberry-2B (w_ReCoT) | REVERIE-7B (w_ReCoT) |
|---|---|---|
| incorrect | 1.13 | 3.72 |
| correct | 98.9 | 93.40 |

The frequency of rationale agreement with the original answer is high when the original answer is correct, but drops significantly when the answer is incorrect, suggesting that the model trained with our method exhibits relatively weak confirmation bias in rationale generation.

*Results on MMMU benchmark.* We additionally evaluate ReCoT on the MMMU benchmark, and the results in the table below further demonstrate its effectiveness. We will add the results to the main paper.

| Method | Reflect? | Mulberry-2b [43] | | | Mulberry-7b [43] | | |
|---|---|---|---|---|---|---|---|
| | | `Acc.` | $\mathrm{CSR}_{self}$ | $\mathrm{CSR}_{ctrl}$ | `Acc.` | $\mathrm{CSR}_{self}$ | $\mathrm{CSR}_{ctrl}$ |
| SFT | ✗ | 39.81 | – | – | 46.86 | – | – |
| SFT | ✓ | 38.76 | 2.05 | 6.97 | 46.86 | 2.87 | 14.36 |
| **ReCoT** | ✓ | **40.33** | **18.68** | **39.36** | 45.33 | **18.29** | **45.97** |

## C. Visualization Analysis

We use ReCoT to train the reflection and error-correction capabilities of REVERIE and Mulberry separately, and visualize examples of reasoning tests on the MME benchmark, as shown in Fig 2. The upper one displays the reflective dialogue results of the REVERIE base model, while the lower one presents the reflective dialogue results of Mulberry. From the figure, it can be observed that when the initial answer of the model is tampered with, the model is able to engage in reflective reasoning based on appropriate reflection instructions, correcting the erroneous initial answer to arrive at the correct response.

In the example from the upper one, the user asks whether

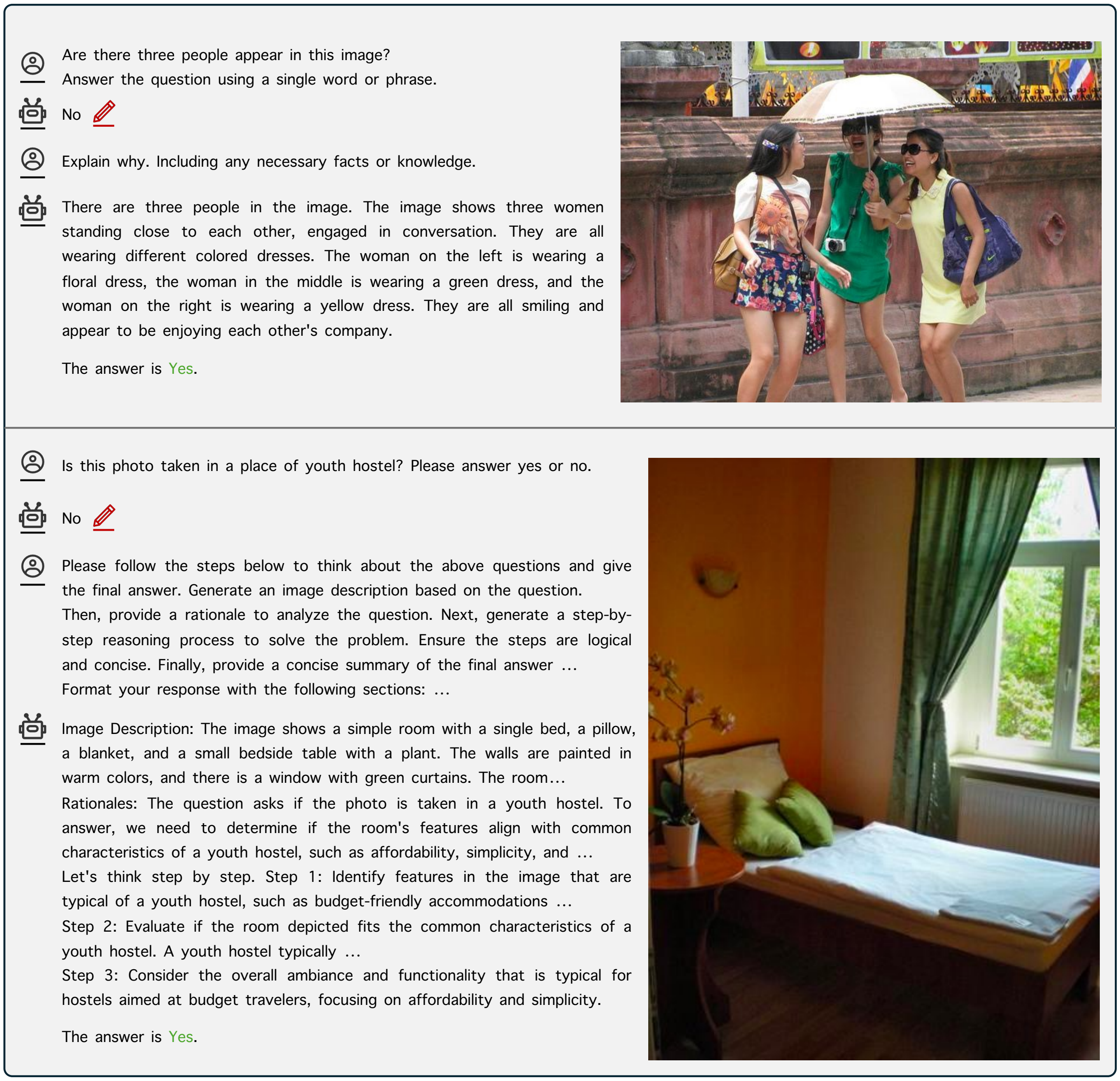



Figure 2. Visualization of ReCoT results from REVERIE (upper) and Mulberry (lower) dataset.

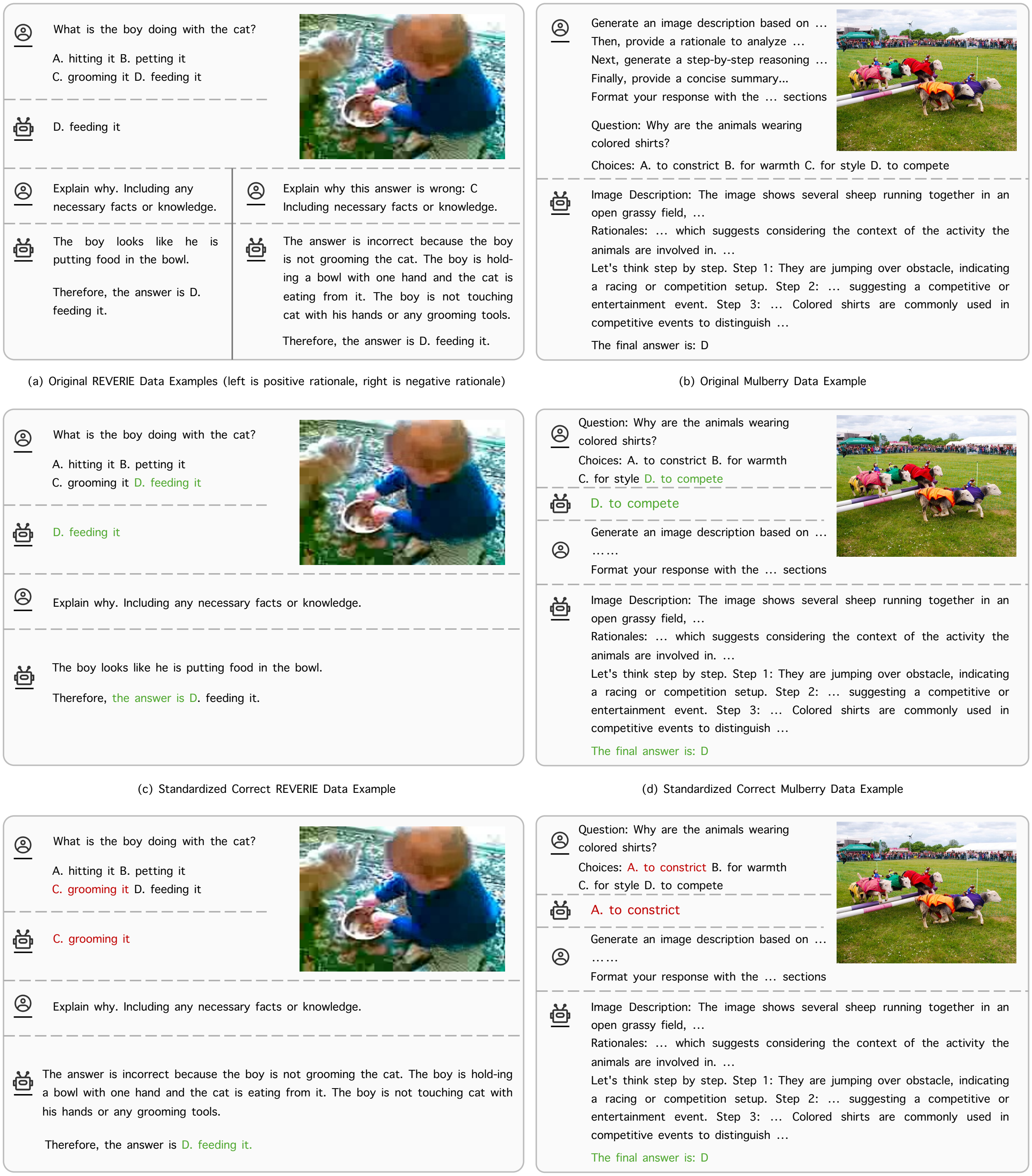
there are three people in the image. Under the tampered incorrect initial answer, the model reflects on the image content, providing a detailed and accurate description of the clothing of each of the three individuals, ultimately yielding the correct answer. In the example from the lower one, the user inquires whether the photo was taken in a youth hostel. Under the tampered incorrect initial answer, the model reflects on the image content and infers the correct answer based on the room's furnishings. These examples demonstrate the effectiveness of the ReCoT framework in enabling models to perform reflective reasoning and correct errors, highlighting its potential for improving the accuracy and robustness of multimodal models in real-world applications.

## D. Formatted Data Example for RFT

We take the actual data from REVERIE and Mulberry as an example to illustrate how data with various structures of rationale can be formatted for ReCoT as in training in Figure 3.

## E. Prompts for RCS

As shown in Fig. 4 and Fig. 5, we present the prompt that is set when gpt-4o is used to evaluate the viewpoint consistency of the reflection and the answer during the reflection-answer consistency evaluation.




(e) Correction-required REVERIE Data Example

(f) Correction-required Mulberry Data Example

Figure 3. Data Examples Visualization. (a) and (b) are examples of actual data from REVERIE and Mulberry dataset. (c) and (d) are examples of standardized correct data format for our ReCoT. (e) and (f) are examples of correction-required data format.

````
Here's a refined prompt tailored to evaluating the accuracy of explanation judgments:

---

# Task
Judge the consistency of the explanation and the initial answer and final answer.
The goal is to assess whether the explanation correctly identifies the correctness of
an initial answer and whether it reaches the correct final answer.

---

# Judgment Step Identification
Each correction explanation consists of an initial answer judgment (assessing the
correctness of the given answer) and a final answer selection (choosing the correct
answer after correction).
If it is not explicitly stated that it is wrong or right, then this explanation is the
one that defaults to the initial answer being correct, and only the correctness of the
final answer needs to be judged.

A proper judgment must:
1. If the final answer and the initial answer do not agree, the explanation should
mention that the initial answer is wrong and state that the final answer is correct.
2. If the final answer and the initial answer agree, there should be no mention in the
explanation that the initial answer is wrong and again confirm that the initial answer
is correct.

---

# Example Description
The initial answer given is A and the final answer is B. Then the explanation should
point out that A is wrong and B is right, then the initial_answer_judgment is 1 and
the final_answer is also 1.

---

# Input Format
```
[Question]
{question}

[Initial Answer]
{initial_answer}

[Explanation]
{explanation}

[Final Answer]
{final_answer}

```

---

# Output Requirements
1. The output must be in the following dictionary format without any additional
content.

### Output Format
```json
[
{{
"initial_answer_judgment": "1 | 0",
"final_answer_judgment": "1 | 0",
}}
]
```

# Rules:
1. Direct JSON output without any other output.
````

Figure 4. Prompt for measuring the reflection-answer consistency of REVERIE data format.

````
Here's a refined prompt tailored to evaluating the accuracy of reasoning judgments:

---

# Task
Judge the consistency of the reasoning and the initial answer and final answer.
The goal is to assess whether the reasoning correctly identifies the correctness of an
initial answer and whether it reaches the correct final answer.

---

# Judgment Step Identification
Each reasoning consists of an image description, rationales and thinking steps. The
thinking steps may contain an evaluation of each option. Assess whether the statements
in the thinking steps match the final answer.

A proper judgment must:
1. If the final answer and the initial answer do not agree, the reasoning should
mention that the initial answer is wrong and state that the final answer is correct.
2. If the final answer and the initial answer agree, there should be no mention in the
reasoning that the initial answer is wrong and again confirm that the initial answer
is correct.

---

# Example Description
The initial answer given is A and the final answer is B. Then the reasoning should
point out that A is wrong and B is right, then the initial_answer_judgment is 1 and
the final_answer is also 1.

---

# Input Format
```
[Question]
{question}

[Initial Answer]
{initial_answer}

[Reasoning]
{reasoning}

[Final Answer]
{final_answer}

```

---

# Output Requirements
1. The output must be in the following dictionary format without any additional
content.

### Output Format
```json
[
{{
"initial_answer_judgment": "1 | 0",
"final_answer_judgment": "1 | 0",
}}
]
```

# Rules:
1. Direct JSON output without any other output.
````

Figure 5. Prompt for measuring the reflection-answer consistency of Mulberry data format.