Step-1: Initial Caption Generation

Input

Model: LLaVA-1.5-13B
System Prompt: You are an expert ... answer the question based on the given image. For the question, generate several candidate answers with different temperature.
Given the question: [Can you provide a comprehensive caption for this image?]

Output

Candidate 1: This image shows a twin bed with two side-by-side white headboards ...
Candidate 2: I can observe a twin bed with ...

Step-2: VLM-Based Reward Scoring

Input

Model: GPT4-o / Gemini
Rating Criteria: 1. **Authenticity** (4 points): The answer should ...
2. **Correctness** (2 points): ...
Candidate Answers: 1. In the image, we can see ...
2. I can observe a twin ...

Output

Scores: [{Authenticity: 3, Correctness: 2, ..., Final score: 7}, {Authenticity: 4, ..., Final score: 8}, ...]
Reason: [The generated answer incorrectly describes the headboards as white..., The reason is that..., ...]

Step-3: Rule-Based Reward Scoring

Input

Model: Factual parser
Two Reference Captions (from GPT-4o and Gemini-Pro)
Candidate Answers: 1. In the image, we can see ...
2. I can observe a twin ...

Matching (WordNet, BERT)

Output

Visual Elements (object, Attribute, Relations):
For each reference: {obj: [bed, ...], attr: [white,...], rel: [in the center, ...]}
For each candidates: {obj: [pillow, ...], attr: [red,...], rel: [on the bed, ...]}
Scores for Candidates: [0.55, 0.63, ...]

Step-4: Multi-Turn Reflective Dialogue

Input

Candidate Answers: 1. In the image, we can see ...
2. I can observe a twin ...
VLM-based Reward and Reasons: [(7, The generated answer incorrectly...), (8, The reason is that...), …]
Rule-based Reward: [0.55, 0.63, ...]

Filter

Filter the Data: Score Gap > Threshold
Candidate Answers Rank: Turn1 Answer -> Turn2 Answer -> Turn3 Answer

Pre-defined Template

Output

Multi-turn Reflective Dialogue:
Question: Given an image, can you provide a comprehensive caption for this image?
Turn1 Answer: This image shows a twin bed with two side-by-side white ...
Turn1 Feedback: A score of 7 is given to this caption. The description inaccurately mentions different pillowcases on each side; both visible pillows have red pillowcases ...
...