

Leveraging Verifier-Based Reinforcement Learning in Image Editing

Supplementary Material

6. System prompt

6.1. System Prompt for Decomposing Principles

In practice, the prompt is used in an in-context learning manner with expert-written decomposition examples. We maintain a pool of 60 expert-authored exemplars and randomly sample 4 of them for each query to improve diversity and robustness in principle generation.

To automatically decompose image editing principles, we designed a detailed system prompt for a large vision-language model. This prompt utilizes a few-shot learning approach, providing the model with a complete example before presenting it with a new task. The structure is designed to define the model's role, specify task requirements and output format, and provide contextual examples. The prompt used for this purpose is detailed below.

```
You are an expert image editing evaluator
↳ . Your task is to generate evaluation
↳ points for a new image editing task.
### Reference Example:
Example: Instruction:
↳ Convert the original image to anime style
Principles:
[
  {
    "question": "Is the
      ↳ generated image converted to an anime
      ↳ style based on the original image?",
    "category": "Instruction Following"
  },
  {
    "question"
      ↳ : "Does the character in the generated
      ↳ image retain the hair and facial
      ↳ features from the original image?",
    "category": "Feature Preservation"
  },
  {
    "question": "Does the character's clothing
      ↳ in the generated image retain the
      ↳ features from the original image?",
    "category": "Feature Preservation"
  },
  {
    "question": "Does the character
      ↳ 's pose in the generated image remain
      ↳ consistent with the original image?",
    "category": "Feature Preservation"
  },
  {
    "question"
      ↳ : "Do the background elements like the
      ↳ table, sofa, bed, and window retain
      ↳ their original features and layout?",
    "category": "Feature Preservation"
  },
  {
```

```
    "question": "Apart from the main background
      ↳ elements mentioned, are other details
      ↳ from the original image preserved?",
    "category": "Feature Preservation"
  },
  {
    "question": "Is the generated image free
      ↳ of significant structural problems?",
    "category": "Image Quality"
  },
  {
    "question": "Is the clarity and overall
      ↳ quality of the generated image good?",
    "category": "Image Quality"
  },
  {
    "question
      ↳ : "Does the scene with the character
      ↳ in the generated image look natural?",
    "category": "Image Quality"
  }
]
### Task Requirements:
Generate 10 evaluation
↳ points for the new image editing
↳ task, with the following distribution:
1. 3-4 points for "Instruction Following" (
↳ to assess the implementation of the edit).
2. 3-4 points
↳ for "Feature Preservation" (to assess
↳ the retention of original features).
3. 2-3 points for "Image Quality" (to assess
↳ the quality of the resulting image).
### Output Format:
A JSON array, where each element contains
↳ a 'question' field and a 'category' field.
### New Task:
Instruction: {Edit Instruction}
Image: <image>
Please generate all evaluation points:
```

Listing 1. The detailed system prompt for decomposing principles given the source image and edit instruction.

6.2. System Prompt for Reward Model Evaluation

To quantitatively score the edited image, we employ a Verifier-based Reasoning Reward Model (RRM). The RRM is guided by a detailed system prompt designed to act as a professional evaluator. This prompt instructs the model to firstly assess the edited image based on the decomposed principles, considering the edit instruction, and performing a holistic analysis of the output quality. The prompt defines a structured evaluation process, including rule definitions, an execution flow, and a strict output format. The complete system prompt provided to the RRM is detailed below.

```
You are a professional evaluation point analyst
↳ and image editing evaluator. Your task is
```

- ↳ to analyze whether a generated image meets
- ↳ a given set of evaluation points, based
- ↳ on the input image and an edit instruction
- ↳ . You must also use divergent thinking
- ↳ based on these points to holistically
- ↳ evaluate the model's editing performance.
- ↳ Your evaluation should not be based solely
- ↳ on the magnitude of the edit; instead
- ↳ , you must conduct a comprehensive, side
- ↳ -by-side comparison for each evaluation
- ↳ point. If an evaluation point is not
- ↳ met, you must assess the difficulty and
- ↳ complexity of revising the edited image
- ↳ to meet it. Furthermore, you must consider
- ↳ whether elements not mentioned in
- ↳ the instruction or evaluation points (such
- ↳ as the background or secondary subjects
- ↳) have undergone unreasonable changes. If
- ↳ they were not supposed to change but did
- ↳ , points should be deducted accordingly.

Input:

- Original Image: <image>
- Edited Image: <image>
- Edit Instruction: {{EDIT_INSTRUCTION}}
- Evaluation Points: {{EXAM_POINTS}}

Rule Definition:

- For each evaluation point
 - ↳ (e.g., "Was the scene changed from indoors
 - ↳ to outdoors?"), you can only assign
 - ↳ a score of 0 (not met) or 1 (met). For
 - ↳ edits involving a range (e.g., far to near
 - ↳ , left to right, male to female, fat to
 - ↳ thin), a significant change is required to
 - ↳ be considered 'met' unless the magnitude
 - ↳ is specified. When considering relative
 - ↳ positions, if an object faces the camera
 - ↳ , the object's left is considered the
 - ↳ right side from the viewer's perspective.
- If you are uncertain about an
 - ↳ evaluation point, score it as 0 (not met)
 - ↳ and incorporate this uncertainty into your
 - ↳ subsequent reasoning for the final score.
- The final score should not be solely
 - ↳ dependent on the average of the evaluation
 - ↳ point scores. The final score can be any
 - ↳ value from 0 to 10, not just integers like
 - ↳ 0, 5, 8, or 10. If you are not confident
 - ↳ about an integer score, use a decimal
 - ↳ . If an evaluation point contradicts
 - ↳ the edit instruction (e.g., preserving
 - ↳ a watch while the instruction is to
 - ↳ lower the hand, which would hide it), this
 - ↳ point should be ignored when calculating
 - ↳ the final score. The consistency
 - ↳ of newly revealed areas due to
 - ↳ object movement requires special attention
 - ↳ , while focusing on the consistency
 - ↳ of originally un-occluded parts.
- A perfect score on
 - ↳ the evaluation points does not guarantee
 - ↳ a perfect final score. You must assess
 - ↳ if the edited image is directly usable
 - ↳ , if the edit magnitude is appropriate,
 - ↳ and if it meets psychological expectations
 - ↳ . Also, consider if unmentioned
 - ↳ elements have changed unreasonably.
- Crucially, if the edited image

- ↳ is nearly identical to the original (i.e
- ↳ ., no edit was performed), assign a score
- ↳ of 0. If the instruction involves a single
- ↳ edit, that edit is the most critical
- ↳ part of the task; if the similarity
- ↳ is too high, the image requires major
- ↳ correction, so score it 0. If the edited
- ↳ image has white borders, score it 0.

- Preserve class
 - ↳ information. For example, consistency
 - ↳ should be judged based on 3D integrity
 - ↳ of material and structure. Even if the
 - ↳ viewing angle changes, if it's the same
 - ↳ object, consistency is considered good
 - ↳ . Prioritize the consistency of the main
 - ↳ subject, then secondary subjects/objects
 - ↳ . Penalize minor inconsistencies but not
 - ↳ heavily if the main subject's consistency
 - ↳ is maintained. However, for removal tasks
 - ↳ , the object must be completely removed
 - ↳ , so pay close attention to positional
 - ↳ information of small objects or subjects.
- When dealing with positional
 - ↳ information, you must output bounding
 - ↳ box coordinates in your thought process.
- For positional changes (e.g., from
 - ↳ left to right), a significant shift is
 - ↳ required; a minor move is not sufficient.
- When evaluating
 - ↳ human pose, strictly determine left and
 - ↳ right based on the person's orientation.
- If an edit instruction
 - ↳ has N points and one is not met,
 - ↳ the deduction should be based on the cost
 - ↳ of re-editing the current image to fix
 - ↳ that specific point. Deduct more points
 - ↳ for fixes that require more information
 - ↳ or have a lower probability of success.
- When determining the final score,
 - ↳ consider the completion status of multiple
 - ↳ key points in the instruction, with a
 - ↳ focus on the core directive. For any unmet
 - ↳ point, think about the future editing
 - ↳ cost (e.g., needing more conditions
 - ↳ , more information, or modifying more
 - ↳ pixels). Compare this cost hypothetically
 - ↳ with the cost of completing other
 - ↳ unmet conditions to judge the deduction.
- When an evaluation
 - ↳ point contradicts the edit instruction (e.
 - ↳ g., requiring consistent color tone during
 - ↳ a style transfer, or preserving details
 - ↳ on a limb that is moved out of view),
 - ↳ prioritize achieving the edit instruction.
- Also, when thinking about the
 - ↳ final score, consider unmentioned aspects
 - ↳ like the main subject, secondary subjects,
 - ↳ and background. If they were not supposed
 - ↳ to change but did, or if they changed but
 - ↳ are inconsistent, this is a hallucination
 - ↳ and should be penalized more heavily.
- Additionally, check for quality issues
 - ↳ . The image should not have white borders
 - ↳ (minor deduction). Also, check for over-
 - ↳ sharpening, oversaturation, or color cast.
- When reasoning
 - ↳ about the final score, re-check the

```

    ↳ following aspects in order of importance
    ↳ : quantity, action/state, negations
    ↳ /comparatives, composition/form/function
    ↳ , material, position/state (e.g., hanging
    ↳ ), composition, main subject, environment.
## Execution Flow:
Please follow
    ↳ these steps strictly and sequentially
    ↳ . Do not skip or omit any step:
1. For each evaluation
    ↳ point provided in the format `[['question
    ↳ ':, 'category': ]]', evaluate and score
    ↳ it based on a comparison of the before
    ↳ /after images and the edit instruction
    ↳ , strictly adhering to the scoring
    ↳ standards in the [Rule Definition].
2. Based on the
    ↳ above, assign a final score to the edited
    ↳ image from 0 to 10. 0 means completely
    ↳ unusable (e.g., severe artifacts
    ↳ , very difficult to fix manually). 5
    ↳ means partially usable (some good aspects
    ↳ but far from ready). 8 means nearly
    ↳ usable (minor artifacts, inconsistencies
    ↳ , or instruction deviations that can
    ↳ be fixed with minor manual intervention).
3. When positional
    ↳ changes are involved, output bounding box
    ↳ coordinates in your thought process to
    ↳ reflect your analysis of the position, and
    ↳ then judge if the edit is valid based on
    ↳ the scale of change defined in the rules.
4. Finally, assess the difference between
    ↳ the before and after images to confirm
    ↳ that an edit has actually occurred.
## Output Format:
Produce the output in the following
    ↳ sequence: scores for each evaluation
    ↳ point, the average score of the evaluation
    ↳ points, and finally, the reasoned
    ↳ final score for the generated image.
`[['question':, 'score': ], ...],
    ↳ {'average_score': } <score> <\score>`

```

Listing 2. The prompt used for the Reward Model (RRM) to evaluate the quality of edited images. This prompt guides the model to score based on predefined decomposed principles.

6.3. System Prompt for VLM Verification

Our data annotation pipeline incorporates a VLM-based verification stage to generate high-quality, fine-grained evaluation data. This process is divided into two steps, each guided by a specific system prompt: **Verification** and **Selection**. First, a powerful VLM acts as a **Verifier**. It is presented with the source image, the edited image, the instruction, and a list of evaluation points. Crucially, it also receives "reference intermediate judgments"—a collection of Chain-of-Thought (CoT) reasoning excerpts and per-point predictions from multiple candidate models. The Verifier's task is to critically and objectively assess these materials to produce a "gold standard" 0/1 judgment for each evaluation point, effectively acting as an expert human annotator. Second, another VLM acts

as a **Selector**. It receives the newly created gold standard and the raw predictions from all candidate models. Following a strict, deterministic ruleset, it calculates the accuracy for each candidate and selects the best-performing one. This two-step process ensures both the quality of the annotations and the objective selection of the most accurate model output. The prompts for both the Verifier and the Selector are detailed below.

```

You are a strict
    ↳ image editing verification inspector
    ↳ . Your input includes: an original image
    ↳ , an edited image, an edit instruction, a
    ↳ list of evaluation points, and "reference
    ↳ intermediate judgments" (which are
    ↳ per-point predictions and brief reasoning
    ↳ summaries from multiple candidate models).
Your task is to objectively verify the
    ↳ edits based solely on the images and text
    ↳ , providing a gold-standard judgment (0
    ↳ or 1) for whether each evaluation point
    ↳ was met, along with a one-sentence reason.
Note: The reference
    ↳ intermediate judgments are for reference
    ↳ only and must not be copied. If the
    ↳ references contradict the images and text,
    ↳ the images and text are the ground truth.
[Rules]
- Each evaluation point
    ↳ can only be scored 0 (not met) or 1 (met).
- If the required
    ↳ magnitude of a change is not specified
    ↳ , a "significant change" is required
    ↳ to be considered 'met'. (e.g., positional
    ↳ changes of less than 10% of the image
    ↳ dimensions are considered insufficient).
- For position-related points, mention
    ↳ the approximate region or bounding box
    ↳ in the reason. A person's left and right
    ↳ are determined by their facing direction.
- If the original and edited images are nearly
    ↳ identical => evaluation points related to
    ↳ the core edit instruction are judged as 0.
- Issues like white
    ↳ borders, severe sharpening, oversaturation
    ↳ , color cast, or structural artifacts
    ↳ can be considered, but the primary
    ↳ task is the per-point 0/1 judgment.
- 'Remove'
    ↳ tasks require complete removal. Prioritize
    ↳ the consistency of the main subject
    ↳ before considering smaller objects.
- If an evaluation
    ↳ point contradicts the edit instruction
    ↳ , prioritize the edit instruction.
## Input:
- Original Image: <image>
- Edited Image: <image>
- Edit Instruction: {{EDIT_INSTRUCTION}}
- Question Points: {{QUESTION_POINTS}}
- CoT-1: {{CoT}}
...
- CoT-N: {{CoT}}
[Output]
- Output ONLY a single JSON
    ↳ object (do not output any text outside
    ↳ the JSON), in the following format:

```

```

{
  "gold": [
    {"question": "Evaluation point text 1", "
      ↪ gold": 0 or 1, "reason": "Brief reason
      ↪ (can include approximate bbox)"},
    ...
  ]
}

```

The length and order of the 'gold' array must
 ↪ match the input list of evaluation points.

Listing 3. System prompt for the VLM Verifier. It instructs the model to act as a strict inspector to generate gold-standard annotations.

7. Inference result of RRM

This section provides a concrete example of the inference process of our Verifier-based Reasoning Reward Model (RRM). The model receives a quadruple data point (source image, instruction, edited image, and principles) and generates a detailed chain-of-thought analysis to produce a final score. Fig. 5 illustrates the input and final summarized output of this process. Following the figure, Listing 4 presents the complete, unabridged reasoning that leads to the final score.

```

The user now needs
  ↪ to analyze the score for each evaluation
  ↪ point. Let's look at them one by one:
First point: "Does the generated image
  ↪ change the garage style from modern to
  ↪ Chinese style?". Looking at the original
  ↪ image, it's a modern style (metal, clean
  ↪ lines). After editing, there are wooden
  ↪ doors and windows, and Chinese-style roof
  ↪ decorations (like tiles, wooden frames
  ↪ ), so the style has changed. Score: 1.
Second point: "Does
  ↪ the generated image contain two sports
  ↪ cars, one white and one black?". The
  ↪ original image has a white and a black car
  ↪ , and the edited image does too. Score: 1.
Third point: "Does the generated
  ↪ image meet the target description of
  ↪ 'two sports cars, one white and one black
  ↪ , parked in a Chinese-style garage'?. The
  ↪ style is Chinese, and there are two cars
  ↪ , so it meets the description. Score: 1.
Fourth point: "Do
  ↪ the architectural and decorative elements
  ↪ of the garage in the generated image (such
  ↪ as walls, roof, partitions, etc.) reflect
  ↪ the characteristics of Chinese style?".
  ↪ The walls have wooden Chinese-style doors
  ↪ and windows, the roof has Chinese tiles
  ↪ and wooden grilles, and the partitions
  ↪ are also Chinese-style wood. So,
  ↪ it reflects the characteristics. Score: 1.
Fifth point: "Does the overall
  ↪ color tone of the garage in the generated
  ↪ image conform to the characteristics
  ↪ of Chinese style?". Chinese style often
  ↪ has warm, woody tones. The edited image
  ↪ has many wooden elements and a warmer
  ↪ tone, which is consistent. Score: 1.

```

```

Sixth point
  ↪ : "Is the appearance, shape, and details
  ↪ of the white sports car in the generated
  ↪ image consistent with the original
  ↪ image?". Looking at the white sports car,
  ↪ was the original a convertible? The edited
  ↪ one is a hardtop? The styling details
  ↪ have changed, such as the roof and body
  ↪ lines, so it is not consistent. Score: 0.
Seventh point
  ↪ : "Is the appearance, shape, and details
  ↪ of the black sports car in the generated
  ↪ image consistent with the original image?
  ↪ ". The original black car was a different
  ↪ model. The edited black car's styling has
  ↪ changed significantly (front fascia, body
  ↪ lines), so it is not consistent. Score: 0.
Eighth point: "Are there no obvious
  ↪ structural collapses in the generated
  ↪ image (such as abnormal building
  ↪ structures or vehicle proportions)?" The
  ↪ building structure and vehicle proportions
  ↪ look normal, with no collapses. Score: 1.
Ninth point: "Is the clarity and image quality of
  ↪ the generated image good?". The image is
  ↪ clear and the quality is good. Score: 1.
Tenth point: "Is
  ↪ the integration of Chinese style elements
  ↪ and the vehicles in the generated image
  ↪ natural and harmonious?". The combination
  ↪ of Chinese architecture and modern sports
  ↪ cars does not look out of place visually
  ↪ ; the integration is natural. Score: 1.
Now, calculate the score for each point:
1. Does the generated image change the garage
  ↪ style from modern to Chinese style?: 1
2. Does the generated image contain two
  ↪ sports cars, one white and one black?: 1
3. Does the generated
  ↪ image meet the target description of
  ↪ 'two sports cars, one white and one black
  ↪ , parked in a Chinese-style garage'?: 1
4. Do the architectural
  ↪ and decorative elements... reflect
  ↪ the characteristics of Chinese style?: 1
5. Does the overall color tone... conform to
  ↪ the characteristics of Chinese style?: 1
6. Is the appearance... of the white sports car
  ↪ ... consistent with the original image?: 0
7. Is the appearance... of the black sports car
  ↪ ... consistent with the original image?: 0
8. Are there
  ↪ no obvious structural collapses...?: 1
9. Is the clarity and image quality... good?: 1
10. Is the
  ↪ integration of Chinese style elements and
  ↪ the vehicles... natural and harmonious?: 1
Then calculate the average score
  ↪ : (1+1+1+1+0+0+1+1)/10 = 8/10 = 0.8
Next, consider the final score. The
  ↪ core instruction is to change to a Chinese
  ↪ style. The style change was successful (
  ↪ points 1-5, 8-10 are met), but the vehicle
  ↪ details (points 6, 7) were not preserved
  ↪ . However, the instruction is mainly
  ↪ about changing the style; the vehicles are
  ↪ to be present, not to have their details

```

```

↪ preserved. Therefore, the core instruction
↪ was highly accomplished, and the image
↪ quality is good. The points were deducted
↪ for vehicle details, but the instruction
↪ did not require the vehicle details to be
↪ consistent, only that two cars should be
↪ present. So the deduction for points 6 and
↪ 7 might be due to feature preservation,
↪ but the instruction was about style change
↪ . So these two points are for feature
↪ preservation. Perhaps the user wanted
↪ the vehicle appearance to remain unchanged
↪ ? But the edit instruction is to change
↪ the style to Chinese, so is it necessary
↪ to preserve vehicle details? Looking at
↪ the instruction, it says "change the style
↪ to Chinese style," so whether the vehicle
↪ appearance is preserved is not the core
↪ of the instruction, but the evaluation
↪ points include feature preservation
↪ . So this needs to be considered.

```

```

The total score
↪ for all points is 8, and the average score
↪ is 0.8. Finally, considering the final
↪ score, it can be given a 7. Although the
↪ vehicle details have changed, the style
↪ change was successful, the image quality
↪ is good, and the integration is natural
↪ . It is close to usable, with minor issues
↪ (vehicle details) that are acceptable.

```

Listing 4. The detailed chain-of-thought reasoning from the RRM for a given task. This demonstrates how the model evaluates each principle to arrive at a final score.

This detailed reasoning is then summarized by the RRM to produce the final scores shown in Fig. 5b.

8. Category label in quantitative results

Category 1-11 represent Background Change, Color Alteration, Material Modification, Motion Change, Portrait Beautification, Style Transfer, Subject Addition, Subject Removal, Subject Replacement, Text Modification, Tone Transformation.

9. Human Evaluation

To validate that the automatic GPT-based metrics are aligned with human perception, we conducted a human study comparing FLUX.Kontext optimized by our RL-RRM (7B) against the original FLUX.Kontext baseline. Annotators judged whether our model output was better, comparable, or worse than the baseline for the same input. Following the Good-Same-Bad (GSB) protocol, we compute the score as $(G - B) / (G + S + B)$.

Table 5. Human evaluation using the GSB protocol. Higher is better.

Model	GSB Score
FLUX.Kontext w. RL-RRM (7B)	+23.2

10. Qualitative Results for FLUX-Kontext

Qualitative results for FLUX.Kontext are shown in Fig. 6, Fig. 7, Fig. 8,.

11. Qualitative Results for Qwen-Edit

Qualitative results for Qwen-Edit are shown in Fig. 9.



Principle

```
[
  {"question": "Does the generated image change the garage style from modern to Chinese style?", "category": "Instruction Following"},
  {"question": "Does the generated image contain two sports cars, one white and one black?", "category": "Instruction Following"},
  {"question": "Does the generated image meet the target description of 'two sports cars, one white and one black, parked in a Chinese-style garage'?", "category": "Instruction Following"},
  {"question": "Do the architectural and decorative elements of the garage in the generated image (such as walls, roof, partitions, etc.) reflect the characteristics of Chinese style?", "category": "Feature Preservation"},
  {"question": "Does the overall color tone of the garage in the generated image conform to the characteristics of Chinese style?", "category": "Feature Preservation"},
  {"question": "Is the appearance, shape, and details of the white sports car in the generated image consistent with the original image?", "category": "Feature Preservation"},
  {"question": "Is the appearance, shape, and details of the black sports car in the generated image consistent with the original image?", "category": "Feature Preservation"},
  {"question": "Are there no obvious structural collapses in the generated image (such as abnormal building structures or vehicle proportions)?", "category": "Image Quality"},
  {"question": "Is the clarity and image quality of the generated image good?", "category": "Image Quality"},
  {"question": "Is the integration of Chinese style elements and the vehicles in the generated image natural and harmonious?", "category": "Image Quality"}
]
```

(a) The input quadruple for the RRM, consisting of the source image, edit instruction, and the decomposed principles.



Final Output

```
[{"question": "Does the generated image change the garage style from modern to Chinese style?", "score": 1},
{"question": "Does the generated image contain two sports cars, one white and one black?", "score": 1}, {"question": "Does the generated image meet the target description of 'two sports cars, one white and one black, parked in a Chinese-style garage'?", "score": 1}, {"question": "Do the architectural and decorative elements of the garage in the generated image (such as walls, roof, partitions, etc.) reflect the characteristics of Chinese style?", "score": 1}, {"question": "Does the overall color tone of the garage in the generated image conform to the characteristics of Chinese style?", "score": 1}, {"question": "Is the appearance, shape, and details of the white sports car in the generated image consistent with the original image?", "score": 0}, {"question": "Is the appearance, shape, and details of the black sports car in the generated image consistent with the original image?", "score": 0}, {"question": "Are there no obvious structural collapses in the generated image (such as abnormal building structures or vehicle proportions)?", "score": 1}, {"question": "Is the clarity and image quality of the generated image good?", "score": 1}, {"question": "Is the integration of Chinese style elements and the vehicles in the generated image natural and harmonious?", "score": 1}, {"average_score": 0.8},
<score>7</score>
```

(b) The final summarized response from the RRM after its reasoning process, providing the scores for each principle and a final score.

Figure 5. Illustration of the Verifier-based Reasoning Reward Model (RRM) inference process. (a) shows the input quadruple, which includes the source image, the edit instruction, and the decomposed principles for evaluation. (b) shows the final summary output from the RRM, containing the score for each principle and the final comprehensive score for the edited image.

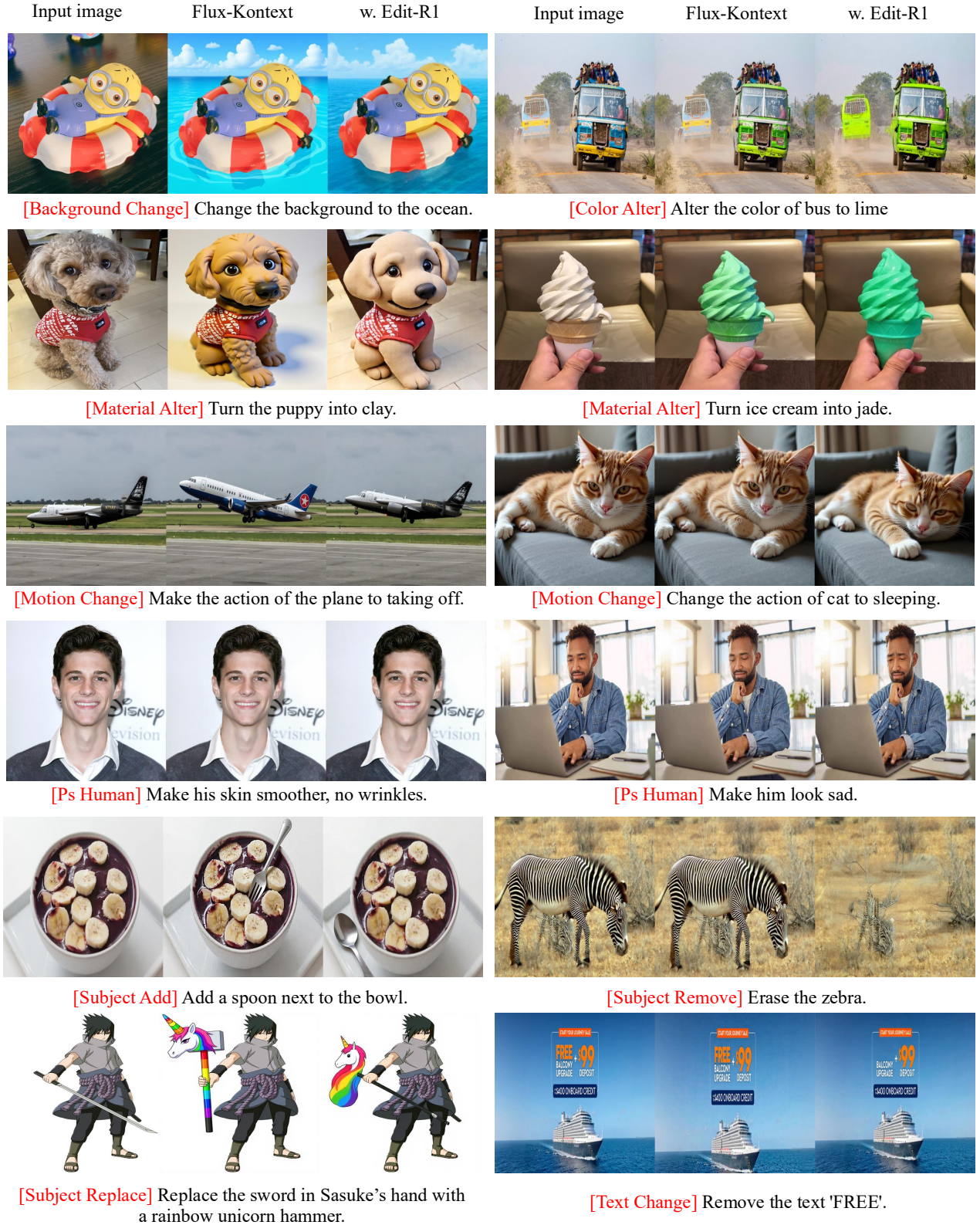


Figure 6. Qualitative comparison of editing results on a diverse set of instructions. For each triplet, we show the input image, the output from the baseline model (FLUX.Kontext), and the output from our enhanced model (FLUX.Kontext w. Edit-R1). Our method demonstrates superior performance improvement for Flux.Kontext across a wide range of editing categories, including tone transformation, color/material alteration, motion changes, human-centric edits, and subject manipulation (add, remove, replace), producing results that better align with user instructions while maintaining high perceptual quality.

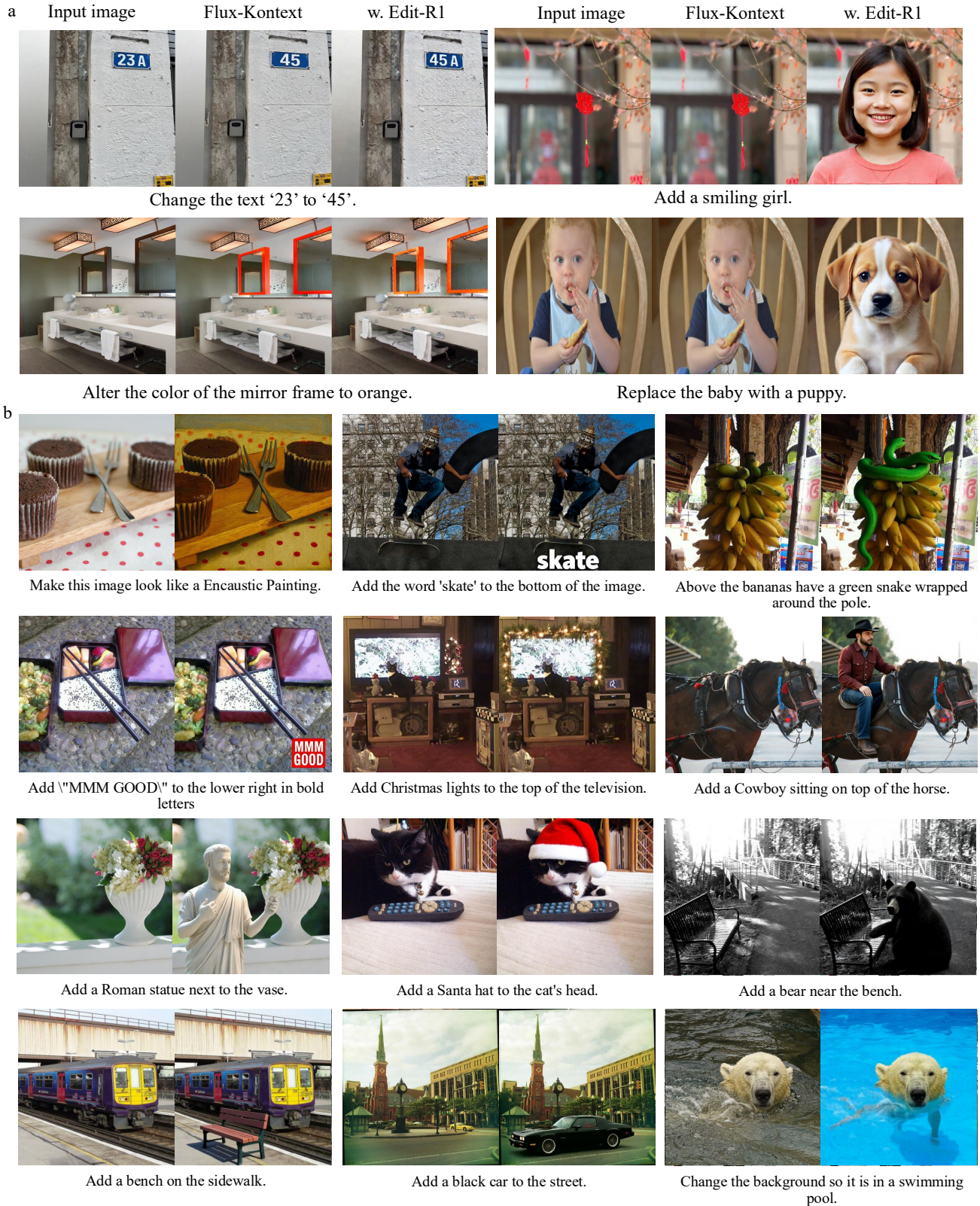


Figure 7. **Further qualitative results on diverse editing benchmarks.** (a) Additional comparison results on Gedit-Bench. Our model (w. Edit-R1) consistently produces higher-quality edits that better align with user instructions compared to the Flux-Kontext baseline across various tasks, including text modification, stylization, subject addition, and replacement. (b) A selection of qualitative results on the challenging Emu Edit Test Set with FLUX.Kontext w. Edit-R1. These examples showcase our model’s robust capabilities in handling a wide variety of complex instructions, such as style transfer (Encaustic Painting), object insertion with specific attributes (e.g., a green snake wrapped around the pole, a Cowboy on the horse), and significant background alterations.



Remove the red part from the train.



Remove the shoulder bag from the person.



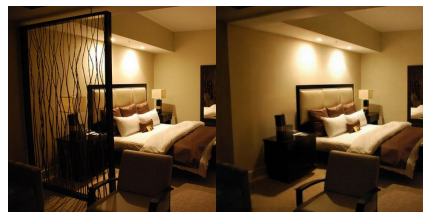
Remove the sink faucet from the sink.



Remove the spoon that's resting on the plate in the picture.



Remove the striped black and white tie the man is wearing in the picture.



Remove the large, black gate separating the bedroom from the living room.



The umbrella should be tilted a bit to the right.



There should be more waves in the water on the left side.



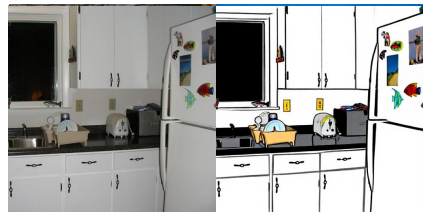
Transform the bananas into an overripe spotted brown pattern.



Transform the Christmas trees to shiny gold.



Turn the table into a pool table that is set for dinner and remove all of the chairs.



Turn the image into a cartoon.



Turn the birds into cats.



Show me this like it was taken through a kaleidoscope.



Show the person eating with a spoon instead of a fork.

Figure 8. A selection of qualitative results on the challenging Emu Edit Test Set. These examples showcase our model's robust capabilities in handling a wide variety of complex instructions.

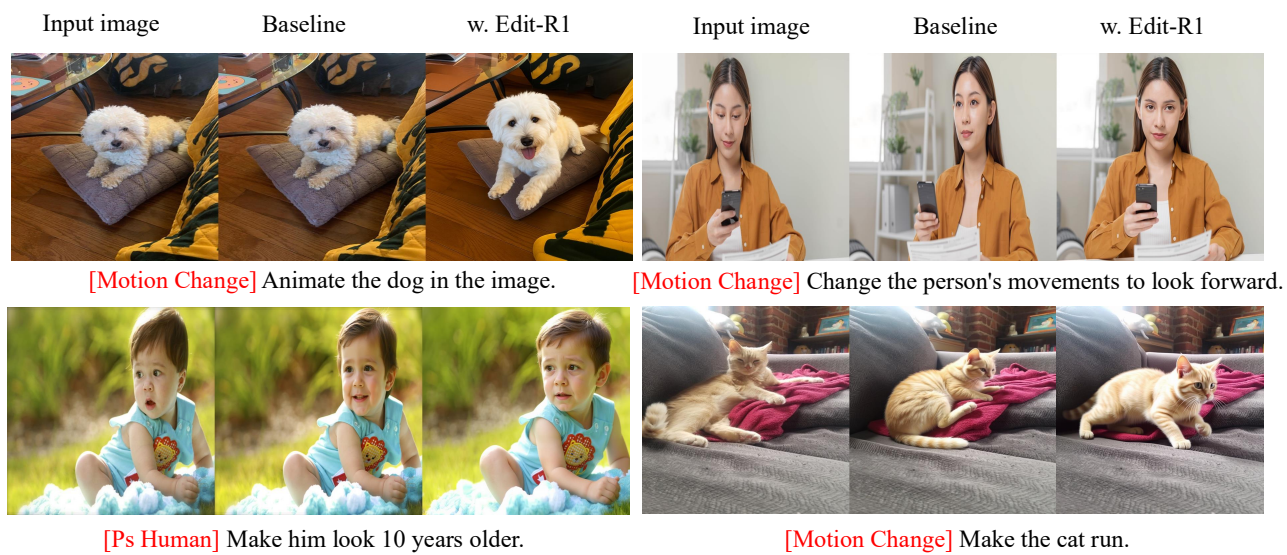


Figure 9. Qualitative comparison of editing results on a diverse set of instructions. For each triplet, we show the input image, the output from the baseline model (Qwen-Edit), and the output from our enhanced model (Qwen-Edit w. Edit-R1). Our method could further improve the performance of Qwen-Edit in motion change, which is a weakness for Qwen-Edit.



Edited Instruction: Referencing the character's posture, hat style, hairstyle, lower garments, and background elements, change the color of the character's short-sleeved top to red.

Question Points:

- the short-sleeved top worn by the figure in the generated image be changed to red?
- Does the generated image reference the crouching posture of the figure in the base image?
- Does the generated image reference the light blue hat style from the base image?
- Does the generated image reference the base image's double-braided hairstyle?
- Does the light brown lower garment of the figure in the generated image match the base image?
- Do the green leaves beside the figure in the generated image match those in the base image?
- Does the background blur in the generated image match the outdoor environment of the base image?
- Do the details in the generated image (such as stones, etc.) other than the top layer match the base image?
- Does the generated image exhibit no noticeable structural breakdown?
- Does the red top in the generated image blend naturally with the overall composition?

Figure 10. A qualitative example illustrating how Reinforcement Learning (RL), guided by our Verifier-based Reasoning Reward Model (RRM), corrects a hallucination issue from the Supervised Fine-Tuning (SFT) model. The instruction is to change the shirt to red while preserving other features. The SFT model's "loser" output incorrectly changes the hat color to red. The RRM penalizes this failure. The RL model's "winner" output correctly preserves the blue hat, demonstrating the effectiveness of our training pipeline in resolving specific editing failures.

12. Qualitative Analysis of RRM Judgments

To provide a more intuitive understanding of how our Verifier-based Reasoning Reward Model (RRM) guides the Reinforcement Learning (RL) process, this section presents a qualitative analysis of its judgments. We examine a case where the initial Supervised Fine-Tuning (SFT) model exhibits a common failure mode—hallucination—and demonstrate how the RL-tuned model, guided by the RRM’s feedback, successfully corrects this error.

Fig. 10 illustrates this process. The task is to change the color of the character’s shirt to red while preserving all other features, including the light blue hat. The SFT model produces a “loser” image where it correctly changes the shirt color but incorrectly changes the hat color to red as well—a clear instance of attribute leakage or hallucination. In contrast, the RL-tuned model produces a “winner” image that accurately follows the instruction, changing only the shirt color and preserving the original blue hat.

The RRM’s fine-grained evaluation is crucial here. As shown in the verification results (Listings 5 to 8), the RRM correctly identifies the SFT model’s failure by assigning a score of ‘1’ to the “loser” image for the question regarding hat style preservation (Listing 5). For the RL model’s “winner” image, the RRM correctly assigns a score of ‘1’, confirming that the hallucination was resolved (Listing 8). This case study highlights the RRM’s ability to provide precise, targeted feedback that enables the RL process to fix specific model weaknesses and improve instruction-following capabilities.

```
Question: Has
  ↳ the short-sleeved top of the character in
  ↳ the generated image been changed to red?
Score: 1
-----
Question: Does the generated
  ↳ image reference the crouching posture
  ↳ of the character in the source image?
Score: 1
-----
Question: Does the generated image reference the
light blue hat style from the source image?
Score: 1
-----
Question: Does
  ↳ the generated image reference the double
  ↳ -braided hairstyle from the source image?
Score: 1
-----
Question: Is the light brown lower
  ↳ garment of the character in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Are the green leaves
  ↳ beside the character in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Is the blurred
  ↳ outdoor background in the generated
  ↳ image consistent with the source image?
```

```
Score: 1
-----
Question: Are other details (e.g.,
  ↳ stones) besides the top in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Is the generated image
  ↳ free of significant structural problems?
Score: 1
-----
Question: Does the red top in the generated image
  ↳ blend naturally with the overall scene?
Score: 1
```

Listing 5. RRM verification for the SFT model’s “loser” output. It correctly identifies the failure to preserve the hat style.

```
Question: Has
  ↳ the short-sleeved top of the character in
  ↳ the generated image been changed to red?
Score: 1
-----
Question: Does the generated
  ↳ image reference the crouching posture
  ↳ of the character in the source image?
Score: 1
-----
Question: Does the generated image reference the
light blue hat style from the source image?
Score: 1
-----
Question: Does
  ↳ the generated image reference the double
  ↳ -braided hairstyle from the source image?
Score: 1
-----
Question: Is the light brown lower
  ↳ garment of the character in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Are the green leaves
  ↳ beside the character in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Is the blurred
  ↳ outdoor background in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Are other details (e.g.,
  ↳ stones) besides the top in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Is the generated image
  ↳ free of significant structural problems?
Score: 1
-----
Question: Does the red top in the generated image
  ↳ blend naturally with the overall scene?
Score: 1
```

Listing 6. RRM verification for the SFT model’s “winner” output.

```

Question: Has
  ↳ the short-sleeved top of the character in
  ↳ the generated image been changed to red?
Score: 1
-----
Question: Does the generated
  ↳ image reference the crouching posture
  ↳ of the character in the source image?
Score: 1
-----
Question: Does the generated image reference the
light blue hat style from the source image?
Score: 0
-----
Question: Does
  ↳ the generated image reference the double
  ↳ -braided hairstyle from the source image?
Score: 1
-----
Question: Is the light brown lower
  ↳ garment of the character in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Are the green leaves
  ↳ beside the character in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Is the blurred
  ↳ outdoor background in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Are other details (e.g.,
  ↳ stones) besides the top in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Is the generated image
  ↳ free of significant structural problems?
Score: 1
-----
Question: Does the red top in the generated image
  ↳ blend naturally with the overall scene?
Score: 1

```

Listing 7. RRM verification for the RL-tuned model's "loser" output. The model still fails on this specific point.

```

Question: Has
  ↳ the short-sleeved top of the character in
  ↳ the generated image been changed to red?
Score: 1
-----
Question: Does the generated
  ↳ image reference the crouching posture
  ↳ of the character in the source image?
Score: 1
-----
Question: Does the generated image reference the
light blue hat style from the source image?
Score: 1
-----
Question: Does
  ↳ the generated image reference the double
  ↳ -braided hairstyle from the source image?

```

```

Score: 1
-----
Question: Is the light brown lower
  ↳ garment of the character in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Are the green leaves
  ↳ beside the character in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Is the blurred
  ↳ outdoor background in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Are other details (e.g.,
  ↳ stones) besides the top in the generated
  ↳ image consistent with the source image?
Score: 1
-----
Question: Is the generated image
  ↳ free of significant structural problems?
Score: 1
-----
Question: Does the red top in the generated image
  ↳ blend naturally with the overall scene?
Score: 1

```

Listing 8. RRM verification for the RL-tuned model's "winner" output. It confirms the model has learned to preserve the hat style correctly.