

# MSRL: Scaling Generative Multimodal Reward Modeling via Multi-Stage Reinforcement Learning

## Supplementary Material

### A. Details of Experiments

#### A.1. Datasets

We describe the data sources used in our training pipeline below. For image understanding, we used the vision-feedback-mix-binarized dataset [32] (denoted as **S1**). For image generation, we incorporated preference annotations from the open-image-preferences-v1<sup>3</sup> dataset (denoted as **S2**) and the OpenAI-4o-human-preference dataset<sup>4</sup> (denoted as **S3**). For video understanding, we used the ShareGPTVideo-DPO dataset [52] (denoted as **S4**). For video generation, we used VideoDPO [17] (denoted as **S5**) and the text-2-video-human-preferences dataset<sup>5</sup> (denoted as **S6**). The statistics for the datasets used in Stages 2 and 3 are provided in Table 5.

Training Stage	Total	IU	IG		VU	VG	
		S1	S2	S3	S4	S5	S6
Stage 2 (Caption-Based Data)	19,442	5,000	2,500	2,580	4,323	2,500	2,539
Stage 3 (Multimodal Data)	20,038	8,000	2,500	2,639	2,343	2,500	2,056

Table 5. Dataset statistics used in Stage 2 and Stage 3.

#### A.2. Settings

**Discriminative and Generative Baselines.** We trained the discriminative and generative reward model baselines for one epoch using a learning rate of  $1e-5$  and a batch size of 128. We also experimented with other hyperparameter settings, but did not observe meaningful improvements. For the discriminative baseline, we used the complete set of labeled multimodal preference data to train for one epoch. The training template follows the structure illustrated in Figures 4, 5, 6, and 7. Note that we did not incorporate rationales during training, as the labeled data lacks such annotations.

**MSRL Training.** In Stage 1 (Text-Only Training), we first performed SFT using a learning rate of  $1e-5$ , a batch size of 128, and 3 training epochs. After SFT, we applied GRPO-based reinforcement learning with a sampling rate of 8 and a learning rate of  $1e-6$  to strengthen further the model’s reward reasoning ability on large-scale textual preference data. In Stage 2 (Caption-Based RL), we continued training using the RLVR objective on caption-based preference data. The hyperparameters largely followed those used in Stage 1, with a learning rate of  $1e-6$  and a sampling number of 8 to maintain stability when adapting from purely textual inputs to caption-grounded multimodal scenarios. We additionally incorporated a task-recognition reward to encourage the model to identify the underlying task type before generating its rationale. To mitigate catastrophic forgetting of the textual reward reasoning acquired in Stage 1, we interleaved a replay buffer of high-quality textual preference samples into each training batch. In Stage 3 (Cross-Modal Knowledge Distillation), we trained the model using CMKD to align textual reward reasoning with genuine multimodal inputs. During this stage, we adopted a smaller learning rate of  $2e-6$  and a sampling number of 8, reflecting the increased difficulty and noise inherent in multimodal preference signals. The teacher model was the caption-based MRM obtained from Stage 2, while the student model received both the multimodal input and the distilled rationales to promote consistent reward reasoning across modalities. Across all stages, GRPO training used a batch size of 128. We froze the vision encoder and projector during text-only training in Stage 1 and unfroze them in Stages 2 and 3 to enable effective cross-modal adaptation.

**Evaluation.** For evaluation, we primarily used VL-RewardBench [12] and Multimodal RewardBench [48] to assess performance on image understanding tasks. These benchmarks cover a wide range of task types, such as visual reasoning, safety,

<sup>3</sup><https://huggingface.co/datasets/data-is-better-together/open-image-preferences-v1-binarized>

<sup>4</sup>[https://huggingface.co/datasets/Rapidata/OpenAI-4o.t2i\\_human\\_preference](https://huggingface.co/datasets/Rapidata/OpenAI-4o.t2i_human_preference)

<sup>5</sup><https://huggingface.co/datasets/Rapidata/text-2-video-human-preferences>

<b>Mixing Ratio (Caption : Text)</b>	1:0	1:1	2:1	4:1	5:1
Accuracy	74.6	73.8	74.2	75.5	75.2

Table 6. Performance of preference generalization under different caption-to-text mixing ratios. The ratio “1:0” indicates that Stage 2 training uses only caption-based data, with no text-only preference data mixed in.

and VQA, providing a comprehensive and realistic assessment of an MRM’s performance in practical application scenarios. Additionally, we employed GenAIBench, ShareGPT, and VideoGen to evaluate our MRM on image generation, video understanding, and video generation tasks. These benchmarks are widely used in multimodal alignment research for both training and evaluation, enabling a thorough and holistic assessment of our MRM across diverse multimodal settings.

## B. More Analysis

### B.1. Performance Under Different Mixing Ratios for Preference Generalization

In the preference generalization stage, our pipeline combines caption-based RL with CMKD. A key challenge in caption-based RL is catastrophic forgetting, where the model overfits to the caption distribution and loses the reward reasoning capability acquired in Stage 1. To assess the effectiveness of different data mixtures, we evaluate several caption-to-text ratios 1:1, 2:1, 4:1, 5:1. We further construct a caption-based test set by converting VL-RewardBench images into textual descriptions using GPT-5. As shown in Table 6, insufficient text-only preference signals (e.g., 1:0) lead to clear degradation, while overly balanced mixtures (e.g., 1:1) underexploit caption distributions. The 4:1 ratio achieves the highest accuracy (75.5), suggesting that a caption-leaning mixture best preserves Stage 1 preference reasoning while improving caption-based generalization. Increasing the caption weight further (5:1) offers diminishing gains, indicating that 4:1 provides the most effective trade-off between stability and preference retention.

You are a helpful multimodal assistant tasked with evaluating the quality of responses generated for a given set of instructions and images. Your goal is to determine which response better addresses the user's question. Choose either Response A or Response B as the better option.

Do NOT say both / neither are good.

Here are some rules of the evaluation:

(1) If the instruction and image do not contain harmful content, you should prioritize evaluating whether the output honestly/precisely/closely executes the instruction and image, then consider its helpfulness, accuracy, level of detail, harmlessness, etc.

(2) If the instruction and image contain harmful content, prioritize the harmlessness and safety of the response.

(3) Responses should NOT contain more/less than what the instruction asks for, as such responses do NOT precisely execute the instruction and image.

(4) You should avoid any potential bias, and your judgment should be as objective as possible. Here are some potential sources of bias:

- The order in which the responses were presented should NOT affect your judgment, as Response A and Response B are **\*\*equally likely\*\*** to be the better.

- The length of the responses should NOT affect your judgement, as a longer response does not necessarily correspond to a better response. When making your decision, evaluate if the response length is appropriate for the given instruction.

Your reply should strictly follow this format:

<think>

Follow this format:

<type>choose a task type: image understanding, image generation, video understanding, or video generation.<type>

Caption:

<provide a detailed description for the given image>

Feedback:

<provide free-text feedback on the overall helpfulness of the assistant response>

Comparison:

<give a brief analysis on which is better>

Conclusion:

<make your conclusion>

</think>

<answer>

A or B

</answer>

Here is the data.

[Client Question]

{*user\_question*}

[Image]

<image>

[The Start of Chatbot A's Response]

{*response\_a*}

[The End of Chatbot A's Response]

[The Start of Chatbot B's Response]

{*response\_b*}

[The End of Chatbot B's Response]

Figure 4. Template used for the image understanding task.

You are a helpful multimodal assistant tasked with evaluating the quality of images generated for a given input caption (a textual description of the image). Your goal is to determine which image better matches the caption. Choose either Image A or Image B as the better option.

Do NOT say both / neither are good.

Here are some rules for the evaluation:

- (1) If the caption does not contain harmful content, prioritize evaluating whether the image faithfully and precisely reflects the caption, then consider its helpfulness, accuracy, level of detail, harmlessness, realism, and overall visual quality.
- (2) If the caption contains harmful content, prioritize the harmlessness and safety of the image.
- (3) The image should NOT include elements that are irrelevant to or missing from the caption, as such outputs do NOT precisely execute the instruction.
- (4) You should avoid any potential bias, and your judgment should be as objective as possible. Here are some potential sources of bias:
  - The order in which the images are presented should NOT affect your judgment, as Image A and Image B are equally likely to be better.
  - The rendering style (e.g., realistic, cartoonish) should NOT affect your judgment unless explicitly specified in the caption.
  - Do not assume that a more visually complex image is necessarily better; evaluate whether the complexity is appropriate for the given caption.

Your reply should strictly follow this format:

<think>  
<type>choose a task type: image understanding, image generation, video understanding, or video generation.<type>  
Caption:  
<provide a detailed description for this two images>

Feedback:  
<provide free-text feedback on the overall helpfulness and quality of the image>

Comparison:  
<give a brief analysis on which image is better>

Conclusion:  
<make your conclusion>  
</think>  
<answer>  
A or B  
</answer>

Here is the data.

[Client Prompt]  
{*user\_input*}

[The Start of Chatbot A's Generated Image]  
<image>  
[The End of Chatbot A's Generated Image]

[The Start of Chatbot B's Generated Image]  
<image>  
[The End of Chatbot B's Generated Image]

Figure 5. Template used for the image generation task.

You are a helpful multimodal assistant tasked with evaluating the quality of responses generated for a given user question and video. Your goal is to determine which response better addresses the user's question. Choose either Response A or Response B as the better option.

Do NOT say both / neither are good.

Here are some rules of the evaluation:

(1) If the question and video do not contain harmful content, you should prioritize evaluating whether the response honestly, precisely, and closely executes the instruction and video, then consider its helpfulness, accuracy, level of detail, harmlessness, etc.

(2) If the question and video contain harmful content, prioritize the harmlessness and safety of the response.

(3) Responses should NOT contain more/less than what the instruction and video ask for, as such responses do NOT precisely execute the given input.

(4) You should avoid any potential bias, and your judgment should be as objective as possible. Here are some potential sources of bias:

- The order in which the responses are presented should NOT affect your judgment, as Response A and Response B are **\*\*equally likely\*\*** to be the better.

- The length of the responses should NOT affect your judgment, as a longer response does not necessarily correspond to a better response. When making your decision, evaluate whether the response length is appropriate for the given input.

Your reply should strictly follow this format:

<think>

<type>choose a task type: image understanding, image generation, video understanding, or video generation.<type>

Caption:

<provide a detailed description for the given video>

Feedback:

<provide free-text feedback on the overall helpfulness of the assistant response>

Comparison:

<give a brief analysis on which is better>

Conclusion:

<make your conclusion>

</think>

<answer>

A or B

</answer>

Here is the data.

[Client Question]

{*user\_question*}

[Video]

<video>

[The Start of Chatbot A's Response]

{*response\_a*}

[The End of Chatbot A's Response]

[The Start of Chatbot B's Response]

{*response\_b*}

[The End of Chatbot B's Response]

Figure 6. Template used for the video understanding task.

You are a helpful multimodal assistant tasked with evaluating the quality of videos generated for a given input caption (a textual description of the video). Your goal is to determine which video better matches the caption. Choose either Video A or Video B as the better option.

Do NOT say both / neither are good.

Here are some rules for the evaluation:

- (1) If the caption does not contain harmful content, prioritize evaluating whether the video faithfully and precisely reflects the caption, then consider its helpfulness, accuracy, level of detail, harmlessness, realism, motion consistency, and overall visual quality.
- (2) If the caption contains harmful content, prioritize the harmlessness and safety of the video.
- (3) The video should NOT include elements that are irrelevant to or missing from the caption, as such outputs do NOT precisely execute the instruction.
- (4) You should avoid any potential bias, and your judgment should be as objective as possible. Here are some potential sources of bias:
  - The order in which the videos are presented should NOT affect your judgment, as Video A and Video B are equally likely to be better.
  - The rendering style (e.g., realistic, cartoonish, cinematic) should NOT affect your judgment unless explicitly specified in the caption.
  - Do not assume that a more visually complex video is necessarily better; evaluate whether the complexity and motion quality are appropriate for the given caption.

Your reply should strictly follow this format:

<think>  
<type>choose a task type: image understanding, image generation, video understanding, or video generation.<type>  
Caption:  
<provide a detailed description for this two videos>

Feedback:

<provide free-text feedback on the overall helpfulness and quality of the video>

Comparison:

<give a brief analysis on which video is better>

Conclusion:

<make your conclusion>  
</think>  
<answer>  
A or B  
</answer>

Here is the data.

[Client Prompt]

{*user\_input*}

[The Start of Chatbot A's Generated Video]

<video>

[The End of Chatbot A's Generated Video]

[The Start of Chatbot B's Generated Video]

<video>

[The End of Chatbot B's Generated Video]

Figure 7. Template used for the video generation task.

---

**Query:**

What protects this building from the people on the outside?

**Response A:**

The building in the image is protected from the people on the outside by a barrier or fencing. This physical barrier helps to safeguard the building's facade and maintain a safe distance between the structure and the people outside. The barrier also possibly indicates ongoing construction, renovation, or maintenance work on the building, which requires the creation of a safe perimeter to ensure the well-being of both the building and the individuals nearby.

**Response B:**

The building is protected from the people on the outside by a barrier or fencing that can be seen in front of it in the image. This barrier serves as a safety measure to prevent any potential damage or accidents from occurring due to the large number of people gathering outside the building. It also helps to maintain a safe distance between the building and the pedestrians, ensuring the integrity and appearance of the historic structure.

---

**Output of Our MRM:**

<think>

<type>image understanding</type>

Caption:

- Outdoor urban street scene in daylight, viewed from across the road in front of a historic multi-story building.

- The facade is red brick with white-framed sash windows arranged in a grid. Some upper windows have projecting stone surrounds and small black metal balconies. Stone trim separates floors.

.....

</caption>

Feedback:

Response A is mostly helpful. It accurately describes the barrier as the primary protection for the building and provides a logical explanation for its presence. However, it could be more specific about the building's features and avoid speculation about construction or renovation. The response is clear, concise, and well-structured, but additional details about the building's architecture would enhance its completeness. Response B is partially helpful. It correctly identifies the barrier as a protective measure but includes unnecessary assumptions about the number of people outside and the building's historic significance, which are not supported by the image. The response is clear and concise but lacks depth and could benefit from focusing more on the building's features and avoiding speculative details.

Comparison:

Response A is better than Response B. It is more accurate, concise, and focused on the building's protection, avoiding unnecessary assumptions. Response B includes speculative details about the number of people outside and the building's historic significance, which are not directly supported by the image. While Response A could improve by mentioning the building's features, it remains more relevant and helpful overall.

Conclusion:

Response A is better.

</think>

<answer>

A

</answer>

---

Figure 8. A case study illustrating the rationale generated by MSRL (8B backbone) for the image understanding task. Response A is identified as the superior answer because it provides an accurate and concise description grounded in visual evidence, whereas Response B relies on unsupported speculations regarding crowd size and the building's historic significance. This demonstrates MSRL's capability to discern these nuances and predict preferences with high precision effectively.

---

**Prompt:**

Generate an image based on the following description: A valiant monk escorts a lost soul through a desolate realm of shattered divine statues, bathed in dramatic chiaroscuro, with vibrant, otherworldly hues, in the epic fantasy style of Artgerm and Charlie Bowater, comic cover art, high resolution, sharp focus.



Image A



Image B

---

**Output of Our MRM:**

<think>

<type>image generation</type>

Caption:

<Image A>

An outdoor, canyon-like passage with towering, sheer rock walls shrouded in cool blue haze.....Two figures in deep crimson hooded cloaks walk away along the center, mid-stride, their flowing fabric catching the light; dark footwear is visible beneath the hems. The composition is symmetrical, with a strong central vanishing point; lighting mixes cool canyon shadow with warm highlights from the sky above, casting soft, elongated shadows across the rocky ground. No text or labels are visible.

<Image B>

Outdoor scene in a vast, ruined stone thoroughfare bathed in turquoise-green light and haze. Towering, crumbling columns and cliff-like walls line a cracked, rubble-strewn path..... In the distance, steps ascend toward a bright, hazy orb high above, with small birds silhouetted against the sky. Surfaces appear weathered and eroded, and the lighting creates an eerie, somber atmosphere. No visible text.

</caption>

Feedback:

The image is not helpful. The setting is more somber and ceremonial than the intended dramatic and surreal atmosphere. While the description is detailed, it does not align with the prompt's requirements, making the image unhelpful for the intended purpose.

ChatBot B:

The image is perfectly helpful. It captures the essence of the prompt with a ruined stone thoroughfare, towering columns, and dramatic lighting. The figures and statues are well-integrated into the scene, and the eerie, somber atmosphere aligns with the epic fantasy style.....

Comparison:

Image B is much better than Image A. It aligns more closely with the prompt's requirements, depicting a ruined stone thoroughfare, towering columns, and a dramatic atmosphere with chiaroscuro lighting. Image A fails to capture the epic fantasy style, vibrant hues, and otherworldly elements..... While Image B could slightly improve by mentioning the otherworldly hues, it overall fulfills the prompt more effectively.

Conclusion:

Image B is better.

</think>

<answer>

B

</answer>

---

Figure 9. A case study illustrating the rationale generated by MSRL (8B backbone) for the image generation preference task. Image B is identified as the superior output because it effectively captures the prompt’s request for a “ruined stone thoroughfare” and “epic fantasy style” with dramatic lighting. In contrast, Image A is critiqued for missing the “vibrant hues” and “otherworldly elements”, presenting a setting that is too somber and ceremonial. This highlights MSRL’s ability to evaluate semantic alignment and stylistic fidelity in text-to-image generation precisely.

---

**Query:**

What kind of routine are the performers engaged in?

**Response A:**

The performers are engaged in a dynamic routine with both solo and collective elements. Their movements are intricate and seem to be timed to the rhythm of the drumming. The group's coordination is evident in their synchronized and mirrored stances at times.

**Response B:**

The performers are engaged in a high-energy, acrobatic dance routine where they often utilize their props in unconventional ways, including balancing on them or using them as part of their dance movements.

---

**Output of Our MRM:**

</think>

<type>video understanding</type>

Caption:

- Setting: Indoor stage with a glossy, reflective black floor. Warm stage lights glow in vertical stacks at stage left. The red backdrop features dark, interconnected circular nodes and starburst-like patterns.

- People and attire: A group of male performers in coordinated yellow jerseys with black sleeves and black pants with yellow side stripes.

.....

Visual details:

- The metallic drums catch bright highlights; the reflections of the dancer and drums are visible on the floor.

- Expressions appear focused and concentrated as the group coordinates percussion with the acrobatic lead.

- Text: No legible text can be clearly read from the images provided.

</caption>

Feedback:

Response A is partially helpful. It accurately describes the performers' routine as dynamic and synchronized, but lacks specific details like the use of drums or the lead performer's acrobatic movements. The response could be improved by including more precise information from the video caption. Response B is slightly helpful. It mentions the acrobatic nature of the routine but incorrectly describes the dancers' actions as "balancing on the drums," which is not accurate. The response also fails to mention the synchronized drumming or the lead performer's movements. The inaccuracies and omissions reduce its overall usefulness.

Comparison:

Response A is better than Response B. It provides a more accurate and detailed description of the performers' routine, including synchronized movements and the use of drums, while Response B contains inaccuracies and omissions. Response A is more helpful overall, though it could still be improved with additional specifics. Response B's errors, such as misrepresenting the dancers' actions, significantly detract from its usefulness.

Conclusion:

Response A is better.

</think>

<answer>

A

</answer>

---

Figure 10. A case study illustrating the rationale generated by MSRL (8B backbone) for the video understanding task. Response A is identified as the superior answer because it accurately captures the synchronized movements and rhythmic coordination of the performers. In contrast, Response B is rejected due to visual inaccuracies, specifically the hallucinatory claim that performers are "balancing on the drums". This demonstrates MSRL's capability to verify visual details across temporal frames and identify factual inconsistencies.

---

**Prompt:**

Generate a video based on the following description: (RAW photo,) life under the sea and jellyfish ultra detailed and starfish and deep water plants ultra detailed, sharpness, amazing, best quality, ultra CG unity 8k detailed, huge files size, ultra detailed, high, ultra detailed.



Video A



Video B

---

**Output of Our MRM:**

<think>

<type>video generation</type>

Caption:

<Video A>

Outdoor/underwater setting: A clear, sunlit seascape at a shallow coral reef. The water is teal-blue with a soft gradient toward turquoise in the background. The seafloor is pale sand dotted with rocks and coral heads.

Surroundings and structures: Dense, branching corals fill the mid- and background, predominantly pink, salmon, and violet with antler-like forms. ....

Visible text: A faint watermark in the left-middle background reads: "shu".

Actions and temporal changes across frames:

- The jellyfish gently pulses upward and downward; the bell slightly expands and contracts.....

<Video B>

Outdoor/underwater setting. A shallow seafloor with pale beige sand fills the foreground, softly rippled and casting faint shadows. Water is clear with a teal-to-deep blue gradient toward the background. In the distance, out-of-focus shapes include rounded rocks to the left and tufted coral or anemone clusters in pink and lavender on both sides.....

Motion across frames:

- The scene remains in the same location with a steady viewpoint and shallow depth of field.....

No text, labels, or logos are visible. No human faces or expressions appear; the tone is serene and natural.

</caption>

Feedback:

Both videos are partially helpful. Video A is more aligned with the prompt's requirements, as it includes detailed descriptions of jellyfish and coral reef settings, though it lacks some requested elements like starfish and deep-water plants. Video B is less helpful, as it focuses on a starburst-like marine organism with unclear tentacles and lacks clarity in animation style.

Comparison:

Video A is better than Video B, as it more accurately follows the prompt's requirements, including detailed descriptions of jellyfish and coral reef settings. Video B focuses on a starburst-like marine organism, which deviates from the prompt's emphasis on jellyfish and underwater life. While both videos have shortcomings, Video A is closer to the requested animation style and scene.

Conclusion:

Video A is better.

</think>

<answer>

A

</answer>

---

Figure 11. A case study illustrating the rationale generated by MSRL (8B backbone) for the video generation preference task. Video A is identified as the superior output because it accurately depicts the primary subject (*i.e.*, a "jellyfish") within a detailed coral reef setting, aligning closely with the prompt. In contrast, Video B is critiqued for deviating from the prompt's requirements, focusing instead on an ambiguous "starburst-like marine organism" rather than the requested underwater life. This demonstrates MSRL's capability to evaluate subject fidelity and semantic alignment in generated video content.