

## Appendix of VL-RewardBench

### A. Additional Experiments

#### A.1. Ablation Studies

**Comparison with Baseline Datasets** To validate our filtering mechanism, we compare VL-RewardBench against two baseline datasets derived from the same source preference datasets (WildVision, VLFeedback, RLHF-V, Povid, and RLAIF): (i) a balanced set (B-Set) with uniform sampling across datasets, and (ii) a distribution-matched set (D-Set) that mirrors VL-RewardBench’s proportion from each source. As shown in Tab. 1, VL-RewardBench demonstrates superior discriminative power between different VL-GenRMs. For instance, while Gemini-series models show minimal performance gaps (0.2% and 0.8%) on B-Set and D-Set with both achieving >95% accuracy, the gap widens significantly to 7.9% on our filtered set. This pattern holds for open-source models as well. Moreover, the similar performance between B-Set and D-Set indicates that the increased challenge stems from our filtering mechanism rather than dataset distribution, validating the effectiveness of our approach.

Models	B-Set	D-Set	VL-RewardBench
Qwen2-VL-7B	68.6	66.5	25.3
LLaVA-OneVision-7B-ov	66.1	66.6	26.2
Molmo-7B	59.6	62.8	31.5
InternVL2-8B	70.2	67.3	38.3
Gemini-1.5-Flash	95.3	96.0	53.7
Gemini-1.5-Pro	95.5	95.2	61.6

Table 1. Model performance comparison on different datasets.

#### A.2. Influence of Decoding Parameters

Temperature / Top-P	0.2	0.4	0.8	1.0
0.2	44.8	46.0	44.1	44.6
0.8	43.5	44.4	47.3	43.2
1.2	47.8	44.8	43.6	44.5

Table 2. Macro average accuracy on our dataset of gpt-4o-mini with different temperature and top-p combinations under  $K = 5$ .

To investigate the influence of decoding parameters, we conduct experiments varying both temperature ( $\tau$ ) and nucleus sampling threshold (top-p,  $p$ ). We explore combinations of  $\tau \in \{0.2, 0.8, 1.2\}$  and  $p \in \{0.2, 0.4, 0.8, 1.0\}$  using GPT-4o-mini with K=5 samples. The results, presented in Table 2, show minimal variation across parameter combinations, with a standard deviation of 1.45 points, suggesting that model performance is robust to these decoding parameter choices.

### B. Model Details

We provide the details of the evaluated models in our main paper in Tab. 3. For open-source models, models are served locally with vLLM [1] using the official checkpoints. Commercial models are queried using the original API interface.

### C. Annotation GUI

The annotation process was facilitated through specialized graphical user interfaces (GUIs). For general and hallucination tasks, we developed an annotation interface shown in Tab. 4, where annotators could categorize errors according to the taxonomy presented in Tab. 5. To handle reasoning tasks specifically, we implemented a dedicated GUI that enables clearer visualization of mathematical rationales. This interface, illustrated in Tab. 6, displays the ground truth alongside the model outputs, allowing annotators to evaluate responses using the guideline detailed in Tab. 7.

### D. Cases of Different Judgment Errors

We show annotated cases corresponding to the following eight error categories: object recognition (Tab. 8), text recognition (Tab. 9), scene recognition (Tab. 10), position recognition (Tab. 11), face recognition (Tab. 12), counting (Tab. 13), attribute (Tab. 14) and existence/hallucination (Tab. 15). All the eight cases are firstly checked as “accurate”, which means “Human Ranking” is validated to be correct and the judge model chooses the incorrect response. In all the cases, “Human Ranking” chooses Response A, so the incorrect response is Response B, which is assigned to the corresponding error category in each case.

VLGen-RMs	LLM (Size)	Vision Encoder	Link
<i>Open-Source Models</i>			
LLaVA-OneVision-7B-ov	Qwen2 (7B)	SigLip-SO400M/14@384px	lmms/llava-onevision-qwen2-7b-ov
InternVL2-8B	Internlm2_5 (7B)	InternViT-300M@448px	OpenGVLab/InternVL2-8B
Phi-3.5-Vision	Phi-3.5-mini (3B)	CLIP-ViT-L/14@336px	microsoft/Phi-3.5-vision-instruct
Qwen2-VL-7B	Qwen2 (7B)	ViT-bigG	Qwen/Qwen2-VL-7B-Instruct
Qwen2-VL-72B	Qwen2 (72B)	ViT-bigG	Qwen/Qwen2-VL-72B-Instruct
Llama-3.2-11B	Llama-3 (8B)	ViT-H/14	meta-llama/Llama-3.2-11B-Vision-Instruct
Llama-3.2-90B	Llama-3 (70B)	ViT-H/14	meta-llama/Llama-3.2-90B-Vision-Instruct
Molmo-7B	Qwen2 (7B)	CLIP-ViT-L/14@336px	allenai/Molmo-7B-D-0924
Molmo-72B	Qwen2 (72B)	CLIP-ViT-L/14@336px	allenai/Molmo-72B-0924
Pixtral-12B	Mistral-NeMo (12B)	Pixtral-ViT	mistralai/Pixtral-12B-2409
NVLM-D-72B	Qwen2 (72B)	InternViT-6B-V1-5@448px	nvidia/NVLM-D-72B
<i>Proprietary Models</i>			
Gemini-1.5-Flash (2024-09-24)	N / A	N / A	Gemini-1.5-Flash
Gemini-1.5-Pro (2024-09-24)	N / A	N / A	Gemini-1.5-Pro
Claude-3.5-Sonnet (2024-06-22)	N / A	N / A	Claude 3.5
GPT-4o-mini (2024-07-18)	N / A	N / A	gpt-4o-mini
GPT-4o (2024-08-06)	N / A	N / A	gpt-4o

Table 3. Specification of the evaluated models in our main paper.

## E. Templates

Tab. 16 provides an evaluation template used for small VLGen-RMs filtering and the evaluation experiments in our main paper. The query used for GPT-4o to judge reasoning sample responses with the ground-truth answer is demonstrated in Tab. 17. In both evaluation templates, we will randomly place candidate answers to avoid positional bias.

## References

- [1] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 1

**Annotating general/hallucination tasks**

**ID: RLAIF-V-59085**

Image Path:images/rlaif-v/RLAIF-V-59085.jpg



**Query:** What are the key features you observe in the image?

**Responses:**

In the image, there is a computer monitor or TV with a speaker nearby on a desk or surface. The monitor or TV is likely being used in conjunction with the speaker for enhanced audio experience. The presence of the speaker and the computer monitor or TV suggest that this could be a setup for multimedia activities, such as watching movies, playing games, or streaming content. The overall scene conveys the idea of a device or system with an integrated audio component, emphasizing the importance of sound in the user experience.

The image features a computer monitor sitting next to a speaker on a desk. The speaker has a close-up view in the image, which highlights its features. The presence of the computer monitor suggests that this setup is likely related to audio production or multimedia. The close-up of the speaker emphasizes the details of the hardware, which can be important in cases where the speaker or audio system plays a critical role in the output sound quality. Overall, the image conveys a message of the importance of audio equipment in computing and multimedia applications.

**Ranking:** 0.1

**Models:** LLaVA-NeXT-34B,LLaVA-v1.5-7B

**Judge:** placeholder\_judge

**Query Source:** COCO

Table 4. Graphical user interface for preference and error annotation of general/hallucination tasks.

#### Guidelines for annotating general/hallucination tasks

You are tasked with evaluating answers to visual general/hallucination questions. Your role is to carefully analyze the provided image, question, and two responses to determine which response is better using the graphical user interface. Notice that “Ranking: 0,1” on the GUI means the original annotated preference is Response 1, while the judge model chooses Response 2 as the better response.

Please evaluate both answers based on the following criteria:

1. Accuracy: Are there any discrepancies or incorrect statements that are not aligned with the image in the response?
2. Completeness: Are there any important details or aspects of the question missing?
3. Clarity: Is the answer easy to read and understand?
4. Relevance: Does the answer directly address the question and the image?

After you evaluate both responses, choose the better one and annotate the error types. If you think both responses are incorrect, please press the “**Inaccurate (Both False)**” button. If you think Response 2 is better than Response 1, please press the “**Inaccurate (Inverted)**” button. If you think Response 2 is worse than Response 1 (which means the judge model chooses the worse response), please follow the categories below to annotate one or more error types for Response 2:

#### Categories of annotated error types

**Counting:** Mistakes in accurately identifying the number of objects or elements in an image.

**Text Recognition:** Difficulty in accurately reading or interpreting text within images.

**Position Recognition:** Misidentifying the relative position or spatial relationships between objects.

**Scene Recognition:** Misinterpreting the overall context or setting of an image.

**Face Recognition:** Inability to accurately recognize or differentiate the features of faces.

**Attribute:** Misidentifying or confusing attributes like color, size, shape, age, posture and state of objects or people.

**Hallucination:** Imagining or asserting the existence of objects, details, or information that are not present in the image.

**Vague (Style):** Cannot judge because the two responses only differ in style of narration (e.g. concise response versus detailed response, or responses of the same meaning using different syntax structures).

**Vague (Knowledge):** Cannot judge because the annotator lacks for the relevant knowledge (e.g. medical images).

**Poor Image Quality:** Cannot judge because the image is hard to distinguish or has resolution issues.

**Others:** The judge mistakes response A as response B ; The image/query has ethical issues so that the responses have no meaningful answer; Other types that can't be distributed to the categories above.

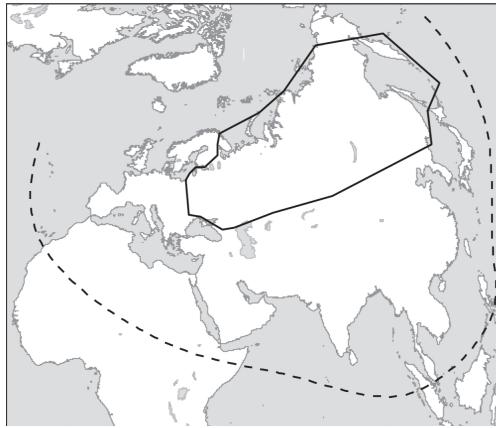
After you annotate the preference and error type, you can press “**Next**” to move on to the next datum, or press “**Previous**” to return to the previous datum. You can enter an “index” and press “**Jump**” button to jump to the (index+1)th datum.

Table 5. Guidelines for annotating general/hallucination tasks.

## Annotating reasoning tasks

ID: mmmu\_pro\_test\_Geography\_42

Image Path: reasoning\_tasks/mmmu\_pro\_test\_Geography\_42.png



**Query:** The model presented in is Options: A. Mackinder's Heartland-Rimland model B. world map as prescribed during the Berlin Conference of 1884 C. the Russian and British Empires in 1904 D. the division between Roman Catholic and Eastern Orthodox religious dominance E. The Cultural Map of 1920 F. The Cold War Division Model G. The European Union Expansion Map H. The World Map from the Treaty of Tordesillas I. The Berlin Wall Separation Map J. Northern and Southern economies separated by the Brandt Line

**Gt:** A

### Responses:

The model presented in the image is: A. Mackinder's Heartland-Rimland model This model is characterized by the delineation of the "Heartland" (the solid line) and the "Rimland" (the dashed line), which are central concepts in Halford Mackinder's geopolitical theory.

The model presented in the image is: A. Mackinder's Heartland-Rimland model This model, proposed by Halford Mackinder, outlines the "Heartland" (central Eurasia) and the surrounding "Rimland" areas, emphasizing the strategic importance of the central region in global geopolitics.

4o:### Evaluation of Answers ##### Criteria Analysis 1. \*\*Accuracy:\*\* - \*\*Answer 1:\*\* Correctly identifies the model as Mackinder's Heartland-Rimland model and describes the delineation of the "Heartland" and "Rimland" using the solid and dashed lines, respectively. This aligns well with the ground truth and the visual information in the image. - \*\*Answer 2:\*\* Also correctly identifies the model as Mackinder's Heartland-Rimland model and provides a description of the "Heartland" (central Eurasia) and the surrounding "Rimland" areas. This is accurate and aligns with the ground truth and the image. 2. \*\*Completeness:\*\* - \*\*Answer 1:\*\* Provides a concise explanation of the model, mentioning the "Heartland" and "Rimland" and their delineation. However, it does not specify the geographical area of the "Heartland." - \*\*Answer 2:\*\* Offers a more detailed explanation by specifying that the "Heartland" is central Eurasia and emphasizes the strategic importance of this region in global geopolitics. This adds depth to the answer and makes it more complete. 3. \*\*Clarity:\*\* - \*\*Answer 1:\*\* The answer is clear and straightforward, making it easy to understand. - \*\*Answer 2:\*\* The answer is also clear and well-articulated, providing additional context that enhances understanding. 4. \*\*Relevance:\*\* - \*\*Answer 1:\*\* Directly relates to the question, the image, and the ground truth by identifying the model and describing its key components. - \*\*Answer 2:\*\* Also directly relates to the question, the image, and the ground truth, but goes a step further by explaining the strategic importance of the "Heartland." #### Reasoning and Judgment - \*\*Answer 1:\*\* correctly identifies the model and provides a clear and accurate description of the "Heartland" and "Rimland." However, it lacks the additional context that could enhance understanding. - \*\*Answer 2:\*\* not only correctly identifies the model but also provides a more detailed explanation, including the geographical area of the "Heartland" and its strategic importance. This makes the answer more complete and informative. #### Overall Judgment \*\*Answer 2\*\* is superior\*\* because it provides a more comprehensive and detailed explanation while maintaining clarity and relevance.

**Models:** qwen-vl-max,gpt-4o-2024-05-13

## Guidelines for annotating reasoning tasks.

You are tasked with evaluating answers to visual reasoning questions. Your role is to carefully analyze the provided image, question, and two responses to determine which response is better using the graphical user interface. Notice that there is no original annotated preference in reasoning tasks.

Please evaluate both answers based on the following criteria:

1. Accuracy/Correctness: Are there any discrepancies or incorrect statements that are not aligned with the image in the response?
2. Clarity: Is the answer easy to read and understand?
3. Relevance: Does the answer directly address the question and the image?

You can refer to the ground truth (**Gt**) to check correctness, and refer to the judge result of GPT-4o to decide which response is better when both are correct.

If you think both responses are incorrect, please press the "**Inaccurate (Both False)**" button. If you think Response 1 is better than Response 2, please press the "**First better**" button. If you think Response 2 is better than Response 1, please press the "**Second better**" button. If you think both are correct and hard to decide which is better, please distribute the datum to following categories:

**Vague (Style):** Cannot judge because the two responses only differ in style of narration (e.g. concise response versus detailed response, or responses of the same meaning using different syntax structures).

**Vague (Knowledge):** Cannot judge because the annotator lacks for the relevant knowledge (e.g. medical images).

**Poor Image Quality:** Cannot judge because the image is hard to distinguish or has resolution issues.

**Others:** Other types that can't be distributed to the categories above.

After you annotate the preference, you can press "**Next**" to move on to the next datum, or press "**Previous**" to return to the previous datum. You can enter an "index" and press "**Jump**" button to jump to the (index+1)th datum.

Previous
Next
Inaccurate(Both False)
First Better
Second Better
Vague(Style)
Vague(Knowledge)
Poor Image Quality
Other
Enter index
Jump to

Table 6. Graphical user interface for preference annotation of reasoning tasks.

Table 7. Guidelines for annotating reasoning tasks.

An example of object recognition error



**Query:** What type of flower is that?

**Response A:**

In the image, there is a flower placed on one of the beds. However, without a closer view or additional details, I cannot definitively identify the type of flower it is. Flowers are often used as decorative elements in interior design to add a touch of natural beauty and freshness to spaces.

**Response B:**

The image features a hotel room with two beds. On one of the beds, there are towels folded into the shape of hearts and swans. These towel art pieces add a touch of elegance and hospitality to the room. As for the type of flower that might be present in the room, I cannot see any flowers in the image. The main focus is on the neatly made beds and the towel art.

**Human Ranking:** Response A is better

**Response Models:** OmniLMM-12B, OmniLMM-12B

**Query Source:** VQAv2

Table 8. An example of object recognition error.

An example of text recognition error



**Query:** What room is represented in this image? What is unusual about it, if anything?

**Response A:**

The image features a large bathroom setting. One unusual aspect of the bathroom is the presence of a sign that reads "I'm on sale." Typically, bathrooms do not display a sign about a sale item or promotion. Moreover, the bathroom appears to be very well-decorated, giving it a luxurious and expensive look.

**Response B:**

The image shows a bathroom. However, it's worth noting that the framed text above the vanity is a humorous touch. The text on the poster says, "I'm on a whiskey diet. I've lost three days already," which is unusual for a bathroom setting as it's a play on words related to dieting and drinking humor. Its presence in a bathroom adds a quirky and unexpected element to the decor.

**Human Ranking:** Response A is better

**Response Models:** llava-v1.5-13b, gpt-4-vision-preview

**Query Source:** wildvision-battle

Table 9. An example of text recognition error.

An example of scene recognition error



**Query:** Is the girl indoors or outdoors with the horse?

**Response A:**

The girl is **indoors** with the horse, as she is standing by indoor fencing and in a barn.

**Response B:**

The girl is clearly **outdoors** with the horse, taking a leisurely stroll near a pond as evidenced by the sandy beach in the background and the bright sunlight casting shadows.

**Human Ranking:** Response A is better

**Response Models:** gpt, corrupted-by-gpt

**Query Source:** POVID

Table 10. An example of scene recognition error.

An example of position recognition error



**Query:** In which direction should I push the red car to bring it closer to the American flag? Up, down, right or left?

**Response A:**

To bring the red car closer to the American flag, you should **push it to the left**.

**Response B:**

To determine the direction in which to push the red car to bring it closer to the American flag, we can analyze the position of the red car relative to the American flag. Observing the flags' arrangement, the American flag is positioned **to the right of the red car**.

**Human Ranking:** Response A is better

**Response Models:** gpt-4-vision-preview, yi-vl-plus

**Query Source:** wildvision-battle

Table 11. An example of position recognition error.

An example of face recognition error



**Query:** What are the key features you observe in the image?

**Response A:**

In the image, there are three young men standing side by side against a white background. The man on the left is wearing a white t-shirt, while the man in the middle is wearing a gray hoodie. The man on the right is dressed in a pink shirt. All three of them have brown hair and are looking directly at the camera. Their expressions are neutral, and they appear to be in good spirits. The overall mood of the image is casual and relaxed.

**Response B:**

The image presents a trio of young men, each exuding their unique style. The man on the left, donning a white t-shirt, stands out with his dark hair and beard. In the middle, another man is seen in a gray hoodie, his light brown hair contrasting with his darker beard. On the right, the third man is dressed in a pink shirt, adding a pop of color to the scene. Their faces are lit up with smiles, reflecting their cheerful dispositions. The background is a stark white, putting the focus entirely on these three individuals. The image captures not just their physical appearances but also their individual styles and personalities.

**Human Ranking:** Response A is better

**Response Models:** LLaVA-NeXT-34B, MiniCPM-Llama3-V-2.5

**Query Source:** LCS-558K

Table 12. An example of face recognition error.

An example of counting error



**Query:** Can you tell me how many cows are eating in this scene?

**Response A:**

In the scene, there are five cows eating grass.

**Response B:**

In the scene, there are six cows eating grass.

**Human Ranking:** Response A is better

**Response Models:** LLaVA-13B, injected-noise

**Query Source:** COCO

Table 13. An example of counting error.

An example of attribute error



**Query:** Are two of the three holding their own hands?

**Response A:**

Yes, two of the three women are holding **their own hands** in the picture.

**Response B:**

Yes, two of the three women are holding **each other's hands**, while the third woman is not holding anyone else.

**Human Ranking:** Response A is better

**Response Models:** LLaVA-NeXT-34B, LLaVA-v1.5-7B

**Query Source:** VQAv2

Table 14. An example of attribute error.

An example of existence error



**Query:** Give a thorough description of the image, focusing on the landmark it shows.

**Response A:**

The image features a scenic view of **a river at night**, with people in boats and floating lanterns. The landmark is likely a popular tourist destination known for its beautiful nighttime views and cultural activities.

**Response B:**

The image depicts a vibrant scene from Hoi An, Vietnam, with **the iconic Japanese Bridge** and surrounding traditional buildings. People are enjoying a boat ride on the river, releasing colorful paper lanterns into the water.

**Human Ranking:** Response A is better

**Response Models:** MiniCPM-Llama3-V-2.5, MiniCPM-Llama3-V-2.5

**Query Source:** sharegpt4v-web-landmark

Table 15. An example of existence error.

**System Prompt**

You are a highly capable multimodal AI assistant tasked with evaluating answers to visual questions.

**Instruction**

Please analyze the following image and question, then determine which of the two provided answers is better.

**Question:** {item['instruction']}

**Answer 1:** {item['answer1']}

**Answer 2:** {item['answer2']}

Please evaluate both answers based on the following criteria:

1. Accuracy: How well does the answer align with the visual information in the image?
2. Completeness: Does the answer fully address all aspects of the question?
3. Clarity: Is the answer easy to understand and well-articulated?
4. Relevance: Does the answer directly relate to the question and the image?

After your evaluation, please:

1. Explain your reasoning for each criterion.
2. Provide an overall judgment on which answer is better (Answer 1 or Answer 2). For example: Overall Judgment: Answer X is better.

Your response should be structured and detailed, demonstrating your understanding of both the visual and textual elements of the task.

Table 16. Evaluation template for pairwise comparison.

**System Prompt**

You are a highly capable multimodal AI assistant tasked with evaluating answers to visual reasoning questions.

**Instruction**

Analyze the following image and question, then determine which of the two provided answers is superior. You will also be given the ground truth answer for reference.

**Question:** {item['question']}

**Ground-truth Answer:** {item['ground\_truth']}

**Answer 1:** {item['answer1']}

**Answer 2:** {item['answer2']}

Please evaluate both answers based on the following criteria:

1. Accuracy: How well does the answer align with the visual information in the image and the ground truth?
2. Completeness: Does the answer fully address all aspects of the question, as compared to the ground truth?
3. Clarity: Is the answer easy to understand and well-articulated?
4. Relevance: Does the answer directly relate to the question, the image, and the ground truth?

After your evaluation, please:

1. Explain your reasoning, citing specific aspects of the image, answers, and ground truth.
2. Provide an overall judgment on which answer is better (Answer 1 or Answer 2).

Your response should be structured and detailed, ending with a conclusive judgment, e.g., Overall Judgment: Answer X is superior.

Table 17. Annotation template for reasoning tasks used by GPT-4o.