



# VL-RewardBench

## A Challenging Benchmark for Vision-Language Generative Reward Models

Lei Li\*, Yuancheng Wei\*, Zhihui Xie\*, Xuqing Yang\*,  
Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin,  
Lingpeng Kong, Qi Liu

*(\* Core Contribution)*

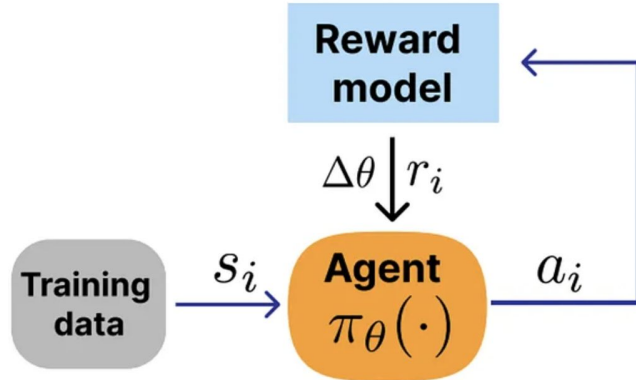


# Outline

- Background & Motivation:
  - Why do we need this benchmark?
- VL-RewardBench:
  - How to build a **high-quality** benchmark **efficiently**?
- Experimental Findings & Analysis
  - Do current MLLMs perform well on VL-RewardBench?
- Future Directions
  - What lessons we have learned?

# Background

*Reward Models (RMs) are key for AI alignment.*



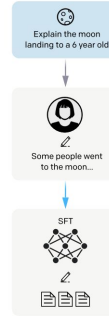
RL (H/AI) Framework

Step 1  
Collect demonstration data,  
and train a supervised policy.

A prompt is  
sampled from our  
prompt dataset.

A labeler  
demonstrates the  
desired output  
behavior.

This data is used  
to fine-tune GPT-3  
with supervised  
learning.

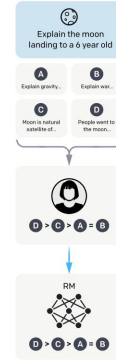


Step 2  
Collect comparison data,  
and train a reward model.

A prompt and  
several model  
outputs are  
sampled.

A labeler ranks  
the outputs from  
best to worst.

This data is used  
to train our  
reward model.



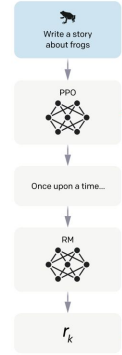
Step 3  
Optimize a policy against  
the reward model using  
reinforcement learning.

A new prompt  
is sampled from  
the dataset.

The policy  
generates an output.

The reward model  
calculates a  
reward for the  
output.

The reward is  
used to update  
the policy  
using PPO.



InstructGPT

## Background

As human feedback is costly and hard to scale,  
Generative Reward Models (LLM-as-a-Judge) are adopted to:  
(i) Model Assessment (ii) Data Curation ...



Application Scenarios of GenRMs [1]

## Motivation



Similarly, Vision-Language Generative Reward Models (VL-GenRMs) are crucial for:

- **The Multimodal Data Flywheel:** Enabling scalable data curation and model evaluation.
- **Advanced Multimodal Reasoning:** Facilitating test-time scaling of VL-GenRMs and enabling Reinforcement Learning (RL) feedback loops.

This critical importance underscores the need for ***a comprehensive benchmark to evaluate and ensure the effectiveness of VL-GenRMs.***

## Motivation

What constitutes an ideal benchmark for evaluating VL-GenRMs?

It should possess the following characteristics:

◆▲ **Domain Diversity:** Cover a wide range of real-world use cases to ensure comprehensive evaluation.



**High Difficulty:** Present a challenge even for state-of-the-art (SOTA) VL models (e.g., Gemini, GPT-4o, Claude 3.5).

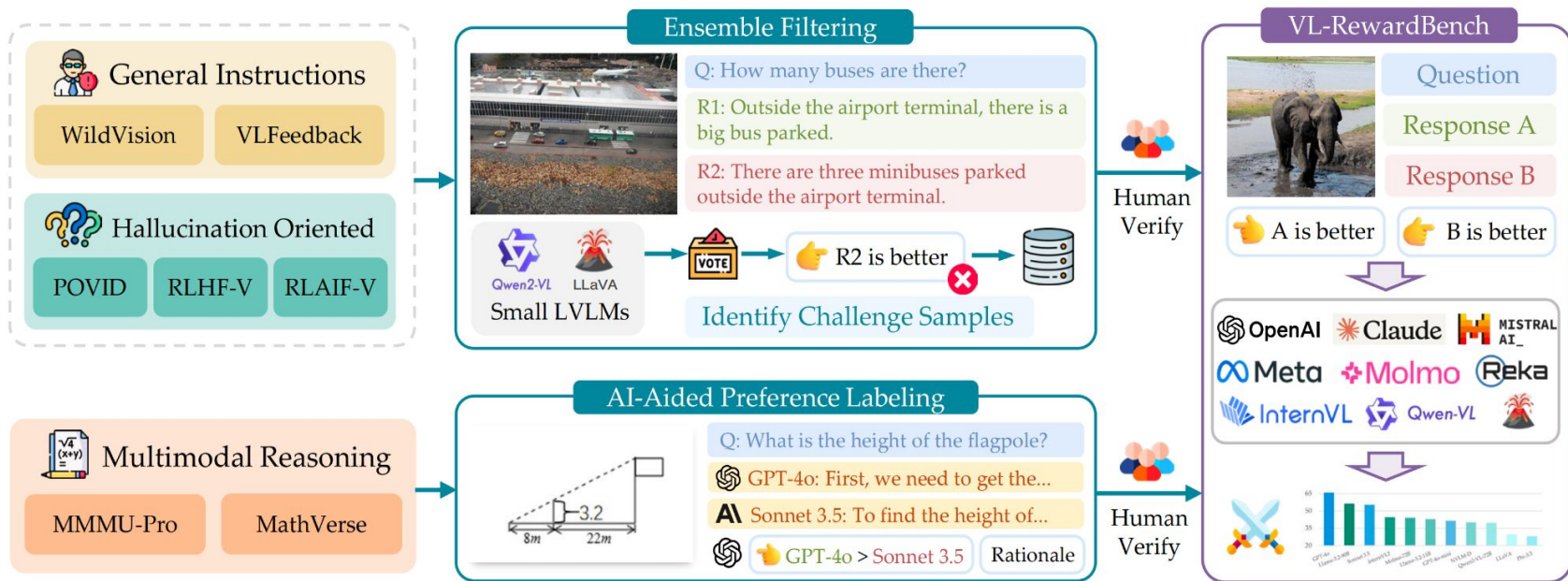


**Objective Labeling:** Employ objective labeling criteria to eliminate common biases, such as verbosity bias.

Current benchmarks **fail to meet these criteria!** 😭

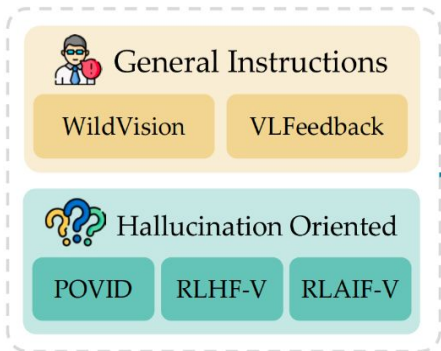
So how can we build such a benchmark **efficiently?** 🤔

# VL-RewardBench: Overall Pipeline



**Curation Pipeline of VL-RewardBench**

## VL-RewardBench: Diverse Sources



Five datasets with annotated preferences:

- *General Queries*: WildVision, VLFeedback
- *Hallucination-oriented Tasks*:  
POVID, RLHF-V, RLAIF-V

Two newly annotated reasoning datasets

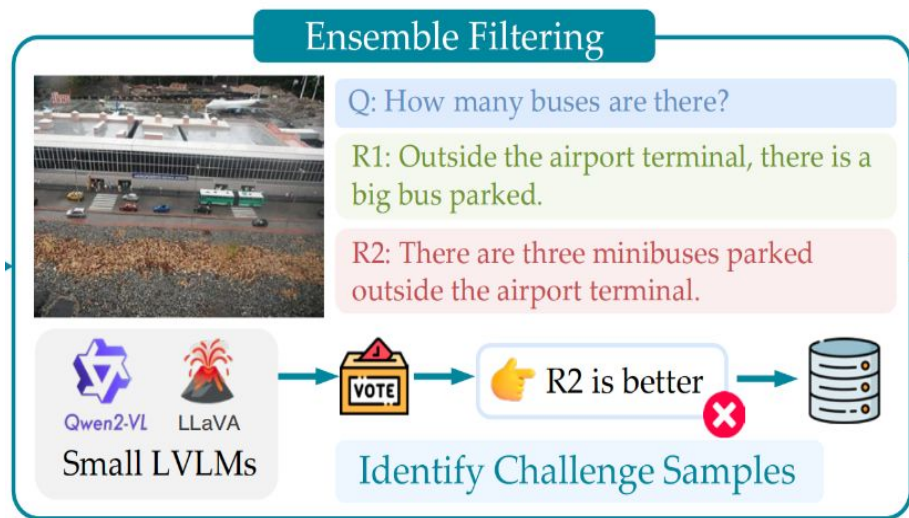
- MMMU-Pro
- MathVerse



# VL-RewardBench: Increasing Difficulty with Ensemble Filtering

## Assumption:

Samples consistently confuse all small models indicate inherent challenges for LVLMs.



## Ensemble Filtering

- Constructing a small models ( ~7B) pools
- Filter out examples all models consistently misjudge

Filter out 3,785 samples from original ~ 100K preferences pairs

## VL-RewardBench: AI-Aided Reasoning Tasks Preference Annotation

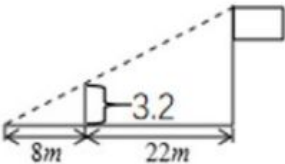
For reasoning samples are already challenging while without preference labels:

**Step 1:** Generate candidate rationales and answers with state-of-the-art LVLMs


**Step 2:** Construct comparison pairs with verbosity difference controlled

**Step 3:** Employ GPT-4o to provide the initial preference labels with rationales



### AI-Aided Preference Labeling



Q: What is the height of the flagpole?

 GPT-4o: First, we need to get the...

**AI** Sonnet 3.5: To find the height of...

  GPT-4o > Sonnet 3.5

Rationale

# VL-RewardBench: Human Verification

All pairs were checked by human annotators.

## *For General & Hallucination Samples:*

- Ambiguous pairs (style difference) were discarded;
- Error tags were annoatated;
- 932 out of 3785 are kept.

## *For Reasoning Samples:*


- Correctness first, then Reasoning Path Quality;
- Responses both containing wrong answers were discarded;
- 318 out of 3114 are left.

## Annotation GUI

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

Previous

Next

Inaccurate(Inverted)

Inaccurate(Both False)

Counting

Recog(Text)

Recog(Position)

Recog(Scene)

Recog(Face)

Recog(Object)

Hallucination

Attribute

Vague(Style)

Vague(Knowledge)

Poor Image Quality

Other

Enter index

Jump to

# VL-RewardBench: Statistics

Table 1: Statistics of VL-RewardBench.

Statistics	Number
Total Preference Pairs	1,250
- General Multimodal Instructions	183
- Hallucination-oriented Queries	749
- Multimodal Reasoning Tasks (newly annotated)	318
Source Datasets	7
- Existing preference datasets	5
- Newly annotated datasets	2
Newly annotated error tags	895
- Existence Error	531
- Recognition Error	184
- Visual Attribute Error	69
- Counting Error	60
- Other Errors	51
Query Word Length Quartile	(6, 9, 31)
Response Word Length Quartile	(48, 99, 136)

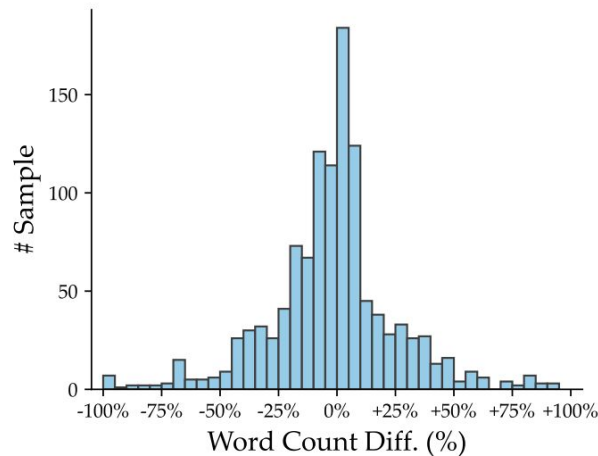


Figure 3: Distribution of the word count difference between the chosen and the rejected response, compared to the word count of the chosen response.

## VL-RewardBench: An Example

Q: How many sinks and mirrors are there in the public restroom?



**Answer 1 (Chosen) ✓:**  
There are **three sinks and three mirrors** in the public restroom.

**Answer 2 (Rejected) ✗:**  
There are **two sinks and four mirrors** equipped with sensor taps alongside a matching set of hand dryers hanging above.

**Qwen2-VL-7B**

**Answer 2 is better**, which accurately states that there are two sinks and four mirrors in the image.

**Llama-3.2-90B**

**Answer 2 is better** because it accurately describes the visual elements in the image, fully addresses all aspects of the question.

**Claude-3.5-Sonnet**

**Answer 2 provides a more accurate, complete, and informative response** to the question while maintaining clarity and relevance. It correctly identifies the number of sinks and mirrors and offers additional details about the restroom that enhance the overall understanding of the space.

**Our pipeline identifies challenging samples for both open-source and commercial models!**

# VL-RewardBench: Evaluation Results

K=5

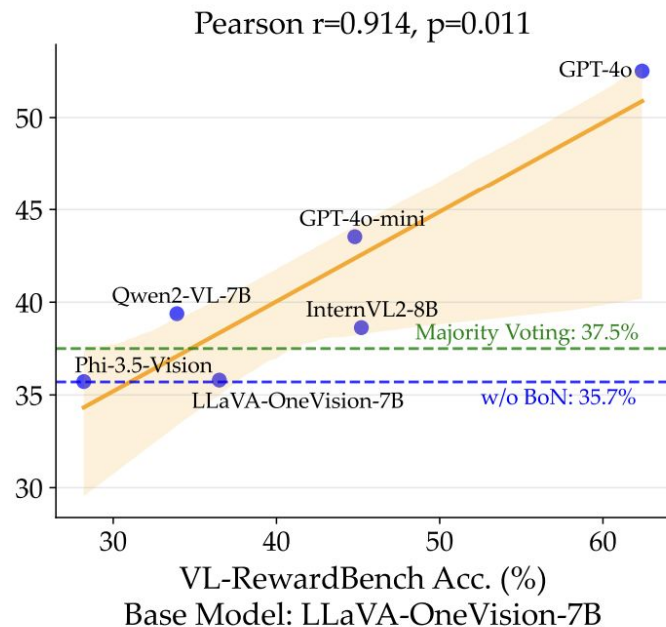
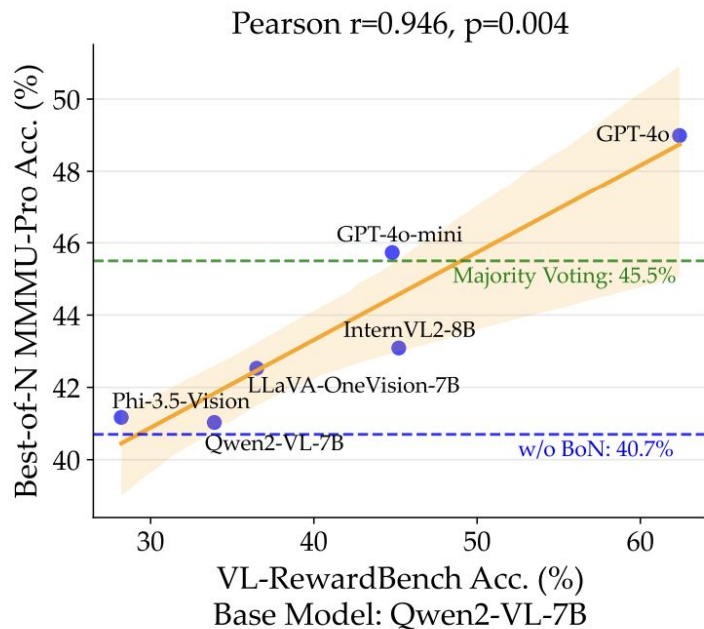
Models	General	Hallucination	Reasoning	Overall Accuracy	Macro Average Accuracy
Open-Source Models					
LLaVA-OneVision-7B-ov	32.2	20.1	57.1	29.6	36.5
InternVL2-8B	35.6	41.1	59.0	44.5	45.2
Phi-3.5-Vision	28.0	22.4	56.6	28.2	35.7
Qwen2-VL-7B	31.6	19.1	51.1	28.3	33.9
Qwen2-VL-72B	38.1	32.8	58.0	39.5	43.0
Llama-3.2-11B	33.3	38.4	56.6	42.9	42.8
Llama-3.2-90B	42.6	57.3	61.7	56.2	53.9
Molmo-7B	31.1	31.8	56.2	37.5	39.7
Molmo-72B	33.9	42.3	54.9	44.1	43.7
Pixtral-12B	35.6	25.9	59.9	35.8	40.4
NVLM-D-72B	38.9	31.6	62.0	40.1	44.1
Proprietary Models					
Gemini-1.5-Flash (2024-09-24)	47.8	59.6	58.4	57.6	55.3
Gemini-1.5-Pro (2024-09-24)	<b>50.8</b>	<b>72.5</b>	64.2	<b>67.2</b>	<b>62.5</b>
Claude-3.5-Sonnet (2024-06-22)	43.4	55.0	62.3	55.3	53.6
GPT-4o-mini (2024-07-18)	41.7	34.5	58.2	41.5	44.8
GPT-4o (2024-08-06)	<u>49.1</u>	<u>67.6</u>	<b>70.5</b>	<u>65.8</u>	<u>62.4</u>

# VL-RewardBench: Evaluation Results

Models	General	Hallucination	Reasoning	Overall Accuracy	Macro Average Accuracy
Open-Source Models					
LLaVA-OneVision-7B-ov	32.2	20.1	57.1	29.6	36.5
InternVL2-8B	35.6	41.1	59.0	44.5	45.2
Phi-3.5-Vision	28.0	22.4	56.6	28.2	35.7
Qwen2-VL-7B	31.6	19.1	51.1	28.3	33.9
Qwen2-VL-72B	38.1	32.8	58.0	39.5	43.0
Llama-3.2-11B	33.3	38.4	56.6	42.9	42.8
Llama-3.2-90B	42.6	57.3	61.7	56.2	53.9
Molmo-7B	31.1	31.8	56.2	37.5	39.7
Molmo-72B	33.9	42.3	54.9	44.1	43.7
Pixtral-12B	35.6	25.9	59.9	35.8	40.4
NVLM-D-72B	38.9	31.6	62.0	40.1	44.1
Proprietary Models					
Gemini-1.5-Flash (2024-09-24)	47.8	59.6	58.4	57.6	55.3
Gemini-1.5-Pro (2024-09-24)	<b>50.8</b>	<b>72.5</b>	64.2	<b>67.2</b>	<b>62.5</b>
Claude-3.5-Sonnet (2024-06-22)	43.4	55.0	62.3	55.3	53.6
GPT-4o-mini (2024-07-18)	41.7	34.5	58.2	41.5	44.8
GPT-4o (2024-08-06)	<u>49.1</u>	<u>67.6</u>	<b>70.5</b>	<u>65.8</u>	<u>62.4</u>

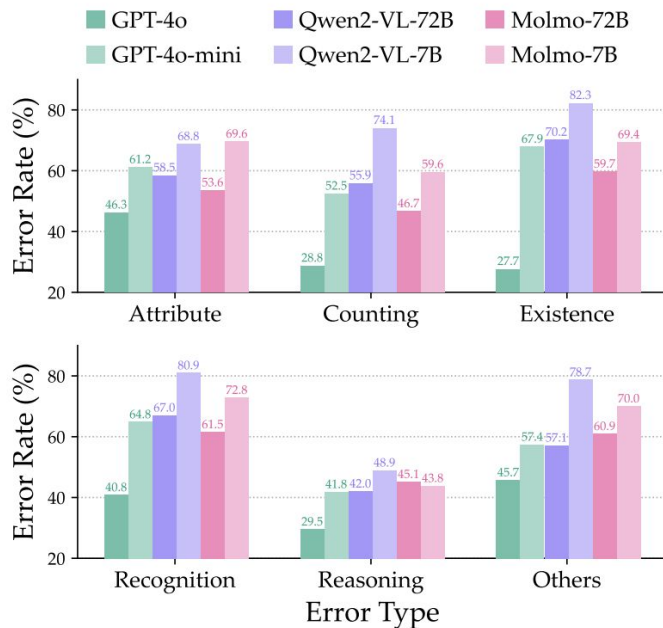


## VL-RewardBench: Correlation with Downstream Tasks



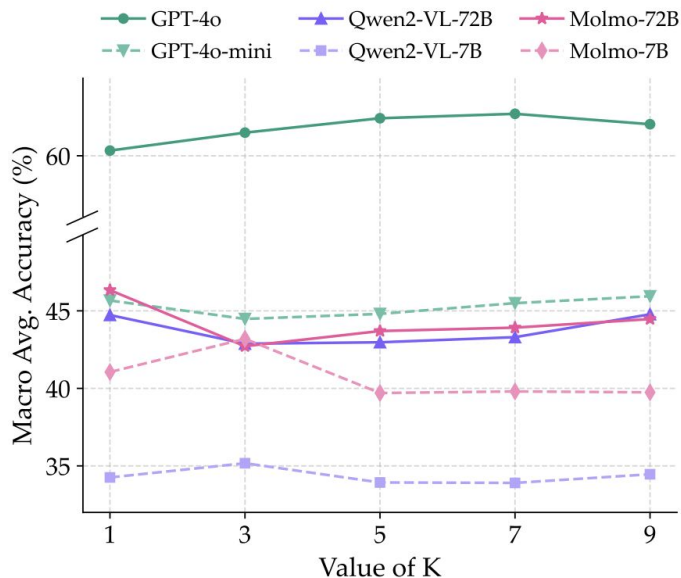


## Findings I: VL-GenRMs make most mistakes on Perception Tasks



- Fundamental perception capabilities emerge as the primary bottleneck
- Reasoning demonstrate relatively better
- Model scaling improvements vary by task type

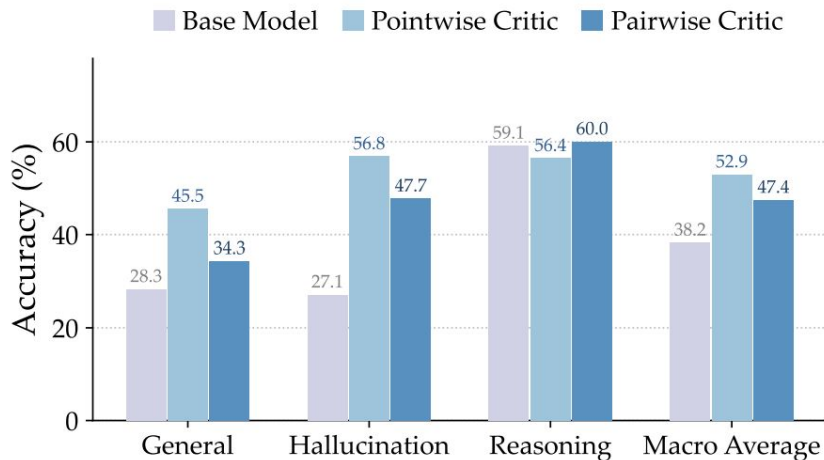
## Findings II: Test-time scaling



(b) Performance changes with varying  $K$ . Increased test-time computation effect varies for different models.

- Text-based scaling law may not directly transfer to visual judgment tasks
- Many open-source LVLMs show performance degradation with increased  $K$

## Findings III: Critic-training yields significant boost



- LLaVA-Critic brings 14.7% ↑
- The pointwise critic achieves better

Figure 6: Evaluation of LLaVA-Critic on VL-RewardBench. Critic training greatly improves judgment accuracy.

## Takeaway

Three key insights for advancing VL-GenRM development:



**Improving Visual Perception:** potentially through visual search mechanisms and vision expert integration.



**Advancing Scaling Strategies:** explore advanced reasoning strategies, incorporating complex planning, process-level supervision or critic training.



**Enabling Co-evolution:** strong LVLMs improve VL-GenRMs, yielding better data and advancing LVLMs.



Thanks for listening.

Q & A

Project Page:

<https://vl-rewardbench.github.io/>



Explore the dataset at:

<https://hf.co/datasets/MMInstruction/VL-RewardBench>

Support [\*\*lmms-eval\*\*](#) & [\*\*VLMEvalKit\*\*](#) for evaluation!



Project



Paper