

---

# VL-RewardBench: A Challenging Benchmark for Vision-Language Generative Reward Models

---

Lei Li<sup>♡\*</sup>, Yuancheng Wei<sup>◇\*</sup>, Zhihui Xie<sup>♡\*</sup>, Xuqing Yang<sup>♣\*</sup>, Yifan Song<sup>†</sup>, Peiyi Wang<sup>†</sup>, Chenxin An<sup>♡</sup>, Tianyu Liu<sup>‡</sup>, Sujian Li<sup>‡</sup>, Bill Yuchen Lin<sup>♣†</sup>, Lingpeng Kong<sup>♡</sup>, Qi Liu<sup>♡</sup>  
♡HKU    ◇SCUT    ♣SJTU    ‡PKU    ♠UW    †AI2  
{nlp.lilei, wyc528813339, zhxieml, catalpabungeiyang}@gmail.com

## Abstract

Vision-language generative reward models (VL-GenRMs) play a crucial role in aligning and evaluating multimodal AI systems, yet their own evaluation remains under-explored. Current assessment methods primarily rely on AI-annotated preference labels from traditional VL tasks, which can introduce biases and often fail to effectively challenge state-of-the-art models. To address these limitations, we introduce VL-RewardBench, a comprehensive benchmark spanning general multimodal queries, visual hallucination detection, and complex reasoning tasks. Through our AI-assisted annotation pipeline combining sample selection with human verification, we curate 1,250 high-quality examples specifically designed to probe model limitations. Comprehensive evaluation across 16 leading large vision-language models, demonstrates VL-RewardBench’s effectiveness as a challenging testbed, where even GPT-4o achieves only 65.4% accuracy, and state-of-the-art open-source models such as Qwen2-VL-72B, struggle to surpass random-guessing. Importantly, performance on VL-RewardBench strongly correlates (Pearson’s  $r > 0.9$ ) with MMU-Pro accuracy using Best-of-N sampling with VL-GenRMs. Analysis experiments uncover three critical insights for improving VL-GenRMs: (i) models predominantly fail at basic visual perception tasks rather than reasoning tasks; (ii) inference-time scaling benefits vary dramatically by model capacity; and (iii) training VL-GenRMs to learn to judge substantially boosts judgment capability (+14.7% accuracy for a 7B VL-GenRM). We believe VL-RewardBench along with the experimental insights will become a valuable resource for advancing VL-GenRMs. Project page: <https://vl-rewardbench.github.io>.

## 1 Introduction

Large vision-language models (LVLMs) such as GPT-4o [34] and Gemini-1.5 [16] have demonstrated remarkable capabilities across diverse multimodal perception and cognition tasks [14, 61]. Building on these capabilities, LVLMs are increasingly being deployed as vision-language generative reward models (VL-GenRMs) to automatically assess model responses [65, 25, 55]. This LVLM-as-a-Judge paradigm has emerged as a scalable solution for model alignment, enabling efficient model ranking and high-quality sample selection [19, 60, 71].

The development of reliable VL-GenRMs is fundamental to three key aspects of LVLM advancement. First, robust evaluation metrics enable systematic and scalable assessment of model performance, crucial for tracking progress and identifying areas for improvement [15, 17]. Second, high-quality automated judges facilitate synthetic training data generation by identifying the most instructive examples, accelerating the data flywheel for better alignment [12, 51]. Third, as the field progresses toward reinforcement learning from human feedback (RLHF) for LVLMs [43, 25, 22], objective

---

\*Core contributors.

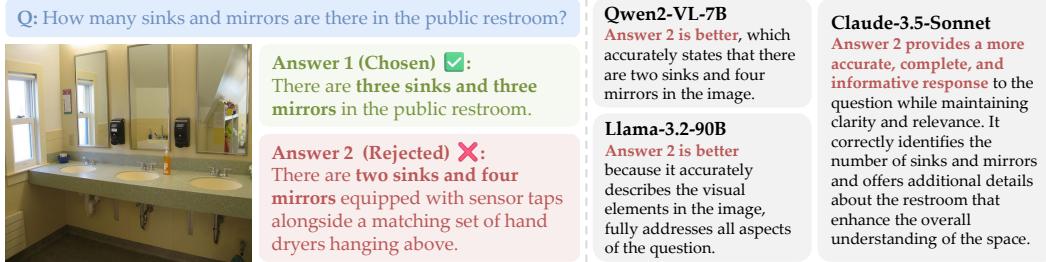


Figure 1: An example from our VL-RewardBench asking the visual details in a restroom. Open-source VL-GenRMs (Qwen2-VL-7B [48] and Llama-3.2-90B [12]) and the commercial Claude-3.5-Sonnet [3] all fail to provide accurate judgments.

and automated evaluation becomes essential for reward modeling. Despite these critical needs and growing applications, the community lacks a comprehensive benchmark for assessing VL-GenRMs' reliability and effectiveness.

Prior work in evaluating VL-GenRMs has taken two main approaches, each with significant limitations. The first approach relies on AI-generated preferences, such as using GPT-4V annotations for assessment [55, 7]. The second approach adapts traditional academic benchmarks with pre-defined labels [6], focusing on preference alignment in traditional vision-language tasks such as image captioning [29]. However, these methods face critical challenges. AI-annotated preferences introduce systematic biases, either favoring model-generated responses [37, 56] or exhibiting stylistic preferences [52, 13]. Meanwhile, conventional task-based evaluations often use simplistic queries that fail to capture the nuanced requirements of real-world applications [67] and therefore lack the complexity needed to differentiate between rapidly evolving LVLMs. Ideally, an effective benchmark for VL-GenRMs should satisfy three key requirements: (a) diverse coverage of real-world applications [69], (b) sufficient difficulty to expose current models' limitations [44], and (c) objective ground truth labels [22]. However, existing studies fall short of these criteria. In this paper, we present VL-RewardBench, designed to meet all three requirements through careful dataset curation and validation. To satisfy criterion (a), our benchmark evaluates VL-GenRMs across three key application domains: (1) general multimodal queries from real users [33, 25], (2) visual hallucination detection tasks [70, 60, 59], and (3) multimodal knowledge and mathematical reasoning [62, 64]. To ensure criterion (b), we employ targeted curation strategies. For source datasets with preference pairs, we employ small LVLMs collaboratively to filter out challenging samples, which our evaluation shows remain difficult even for much larger models. For reasoning tasks without annotated labels, we leverage strong commercial models to generate responses with explicit reasoning paths, followed by GPT-4o's quality assessment. To fulfill (c), all preference labels undergo human verification to eliminate ambiguous or incorrect pairs. The resulting VL-RewardBench comprises 1,250 high-quality samples from 7 diverse datasets, offering a rigorous and comprehensive testbed for advancing VL-GenRM development.

We conduct a comprehensive evaluation of 16 state-of-the-art VL-GenRMs on VL-RewardBench, ranging from open-source models (7B to 90B parameters) to commercial systems including Gemini-1.5-Pro, Claude-3.5-Sonnet, and GPT-4o. Our benchmark reveals significant challenges for current VL-GenRMs: even leading commercial models achieve only moderate performance (GPT-4o: 62.4%, Gemini-1.5-Pro: 62.5%), while state-of-the-art open-source models like Qwen2-VL-72B and LLaMA-3.2-90B struggle to surpass chance level (43.0% and 53.9%, respectively). Notably, performance on VL-RewardBench strongly correlates (Pearson's  $r > 0.9$ ) with downstream MMMU-Pro results when using VL-GenRMs for Best-of-N sampling guidance [42]. Our analysis uncovers three critical insights for advancing VL-GenRMs: (1) The primary performance bottleneck lies in visual perception rather than reasoning - models show significantly higher error rates on existence/recognition tasks ( $> 67\%$ ) compared to reasoning tasks (41.8%); (2) The effectiveness of test-time scaling varies with model capacity, providing benefits to larger models while potentially degrading smaller models' performance; (3) Training VL-GenRMs to learn to judge [55] substantially improves judgment capabilities, demonstrated by a 14.7% accuracy gain for LLaVA-OneVision-7B-ov [24], with pointwise evaluation slightly outperforming pairwise scoring on average. These findings establish VL-

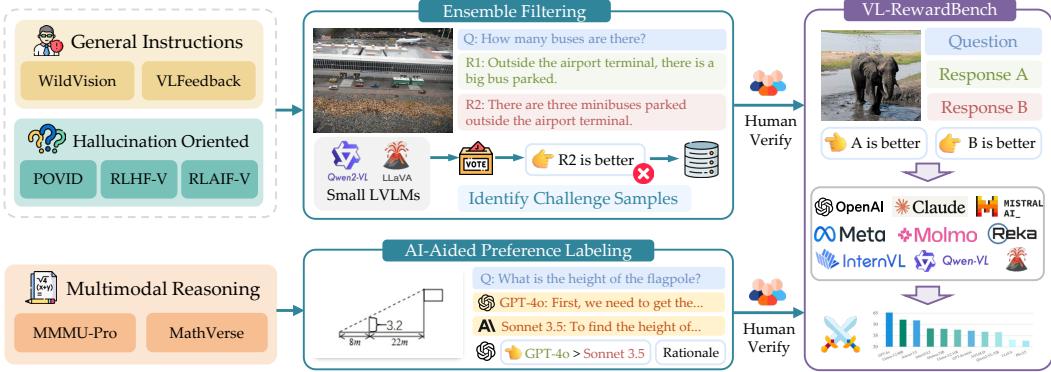


Figure 2: Construction process overview of VL-RewardBench. Two strategies for different datasets: (1) Ensemble filtering process using small LLMs to identify challenging samples from general and hallucination queries (Section 2.2.1); (2) AI-aided preference labeling for multimodal reasoning tasks, where commercial models generate candidate responses and preference labels (Section 2.2.2). All labels are verified by human annotators to ensure correctness.

RewardBench as a valuable benchmark for advancing VL-GenRMs while providing clear directions for future improvements.

## 2 VL-RewardBench

In this section, we introduce the construction process of VL-RewardBench as illustrated in Fig. 2. Following previous benchmarks [25, 22], VL-RewardBench consists of preference pairs  $(x, y_w, y_l)$ , where  $x$  represents a multimodal query containing an image and a user prompt, and  $(y_w, y_l)$  denotes the preferred (chosen) and rejected responses, respectively. For simplicity, we focus on single-image, single-turn interactions, though our framework can be extended to multi-turn dialogues and multiple images. In later subsections, we first describe our diverse data sources (Section 2.1), followed by specific AI-aided pipelines for datasets with originally annotated preference labels and reasoning tasks without labels (Section 2.2). Finally, we present comprehensive dataset statistics and analysis (Section 2.3).

### 2.1 Dataset Source

To ensure our benchmarks cover real-world scenarios, we choose datasets from three domains: general multimodal instructions, hallucination-oriented tasks, and multimodal reasoning tasks.

**General Multimodal Instructions (General)**, which encompass diverse multimodal instructions from different domains to ensure comprehensive coverage of general queries. VLFeedback [25] and WildVision [33] are two general multimodal preference datasets with annotated AI/human feedback. VLFeedback employs GPT-4V as the preference annotator with predefined templates and WildVision collects human preferences by hosting online demos for real-world users. We use the human-verified subset of VLFeedback with 681 samples, and select 6,484 samples in English from the WildVision dataset, for later processing.

**Hallucination-oriented Queries (Hallucination)**, which focus on the hallucination issues of LLMs, involving questions regarding the visual content in the image. We select public available Povid [70], RLAIF-V [60], and RLHF-V [59] with preference labels to reduce the annotation cost. For the preference annotation, Povid injects noise into oracle image descriptions, RLAIF-V innovatively develops a divide-and-conquer framework for annotating the faithfulness of responses, and RLHF-V relies on humans to annotate the preference. All samples in these datasets are included to serve as initial candidates.

**Multimodal Reasoning Tasks (Reasoning)**, aims to evaluate the LLMs with challenging multimodal reasoning tasks. We select the recent MMMU-Pro [62] and MathVerse [64] to reduce the dataset contamination risk. MMMU-Pro is a robust version of the massive multi-discipline multi-

modal understanding and reasoning (MMMU) benchmark [61], where 1,568 single-image samples are adopted. MathVerse serves as a comprehensive visual mathematical reasoning benchmark for LVLMs. We select subsets of Vision Dominant and Vision Intensive to ensure high reliance on the visual inputs, resulting in 1,546 samples.

## 2.2 Preference Annotation

To ensure our VL-RewardBench presents meaningful challenges and separability for current state-of-the-art models given the rapidly evolving landscape, we design an AI-assisted framework with two strategies to obtain challenging samples and annotate preferences: (i) For general multimodal instructions and hallucination-oriented queries that become easier for advanced LVLMs, we design a collaborative filtering strategy to filter out challenging samples with small models. (ii) For challenging multimodal reasoning tasks without preference labels, we design an AI-assisted preference labeling framework to curate high-quality preference pairs for these samples. We elaborate on the detailed process below.

### 2.2.1 Ensemble Filtering with Small Models

Creating effective benchmarks requires careful calibration - tasks that are too easy or impossibly difficult fail to differentiate model capabilities [36]. However, simply selecting tasks that current frontier models cannot solve poses two key challenges: (i) these tasks might become obsolete as models rapidly evolve, and (ii) model-specific failures do not reflect fundamental visual reasoning challenges. Instead, we propose using an ensemble of small models to identify examples that are universally challenging to LVLMs. Our hypothesis is that when multiple small models fail in certain cases, these failures likely stem from fundamental limitations rather than model-specific model weaknesses - a claim validated by our experiments showing these cases challenge even state-of-the-art models. The strategy consists of three steps, *ensemble construction*, *difficulty assessment*, and *human verification*.

**Ensemble Construction** We begin by assembling a diverse ensemble of small vision-language models as weak judges, including LLaVA-1.5-7B [30], LLaVA-1.6-7B [31], LLaVA-OneVision-7B-si [24], and Qwen2-VL-7B [48]. These models evaluate preference pairs from a large initial pool comprising multiple datasets of general multimodal instructions and hallucination-oriented queries. Each model in our ensemble independently evaluates these pairs with a template (see Appendix), yielding preference predictions as the foundation for our subsequent filtering steps.

**Difficulty Assessment** To identify genuinely challenging cases, each preference pair is judged by the small model three times with randomized response positions to mitigate position bias [47]. We construct a “common set” by identifying pairs that all models consistently misjudge (based on majority voting of the three evaluations). This approach ensures that our selected cases represent fundamental challenges rather than model-specific limitations, with 3,785 challenging pairs left.

**Human Verification** To ensure our benchmark represents meaningful challenges rather than annotation artifacts or ambiguous cases, we conduct a rigorous three-stage human verification process. Three authors (graduate students in CS/AI) familiar with the source datasets and problem setup perform the initial annotations, with two additional authors conducting a final verification. Our verification process consists of: (i) **Label Accuracy Check:** We first examine the preference labels, discarding pairs where either the preferred response is worse than the rejected one, or both responses are incorrect; (ii) **Quality and Ambiguity Filtering:** We remove pairs that could lead to ambiguous evaluations, including (a) Responses that are both correct but differ only in style (e.g., verbosity) (b) Images with poor quality or resolution issues (c) Questions requiring domain expertise beyond graduate-level knowledge. (iii) **Error Type Classification:** For the remaining challenging pairs with clear preference labels, we categorize the errors into (1) Recognition errors (text, position, scene, face identification), (2) Counting errors, (3) Visual attribute identification errors, (4) Object existence errors, and (5) Other uncategorized errors. Typical samples for each category could be found in Appendix D. Each sample takes approximately 65 seconds to annotate following our detailed guidelines (Appendix C). The two additional authors review all remaining samples to ensure consistency, with disagreements resolved through discussion. This multi-stage process reduced our initial 3,785 pairs to a final set of 932 high-quality pairs with clear, unambiguous preference labels.

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)

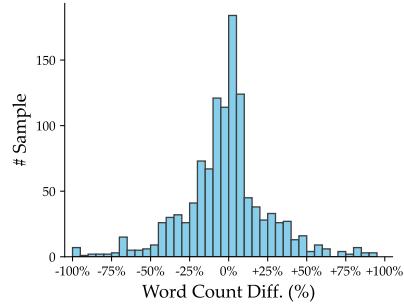


Figure 3: Distribution of the word count difference between the chosen and the rejected response, compared to the word count of the chosen response.

### 2.2.2 AI-aided Preference Labeling for Reasoning Tasks

The original samples from MMMU-Pro and MathVerse provide only oracle answers without preference pairs needed for VL-GenRMs evaluation. To efficiently create such preference pairs for these knowledge-intensive reasoning tasks, we develop a two-stage AI-aided labeling process. First, we generate candidate responses using commercial models (GPT-4o, GPT-4o-mini, and Claude 3.5 Sonnet) given their superior performance compared to open-source LVLMs [62]. To ensure fair comparison, we control for response verbosity by only pairing responses where  $\frac{|l_1 - l_2|}{\min(l_1, l_2)} < \tau$ , where  $l_1$  and  $l_2$  are word lengths. Setting  $\tau = 0.1$  balances controllability and pair diversity, yielding 1,204 candidate pairs. The preference labeling process consists of two steps: (i) **Initial Label Generation:** GPT-4o analyzes each response pair using a structured template (detailed in Appendix) to generate draft preference labels with supporting rationales. (ii) **Human Verification:** Three authors review these AI-generated labels following these criteria: (a) Discard pairs where both responses are incorrect according to the oracle answer; (b) Retain pairs with at least one (partially) correct answer; (c) For pairs with one correct and one incorrect answer, prefer the correct response; and (d) For pairs with two correct answers, prefer responses with more complete and logical reasoning steps. Two additional authors perform a final verification round, with conflicting cases resolved through discussions. This process yields 318 high-quality challenging samples with validated preference labels for multimodal reasoning tasks.

### 2.3 Dataset Statistics

We analyze the composition and characteristics of our VL-RewardBench with detailed statistics in Table 1. **Task Distribution:** The benchmark comprises three main categories: hallucination-related queries (749 pairs, 59.9%), multimodal reasoning prompts (318 pairs, 25.4%), and general instructions (183 pairs, 14.7%). Given this inherent task imbalance, we recommend using macro-average metrics for a more comprehensive evaluation of model performance across different task types. **Error Type Distribution:** Among the 895 pairs annotated with error tags, we find existence errors dominate the distribution at 59.3% (531/895), indicating significant challenges in correctly identifying the presence or absence of objects in images. Recognition errors account for 20.6% (184/895) of cases, while attribute identification and counting errors comprise 7.7% (69/895) and 6.7% (60/895) respectively. This diverse error distribution demonstrates VL-RewardBench’s coverage of various failure modes.

**Length Difference Analysis:** To examine potential length-based preference biases [13], we analyze word count differences between preferred and rejected responses (Fig. 3). The resulting zero-centered bell-shaped distribution confirms that preference labels are not biased by response length, enabling evaluation based on response quality rather than verbosity.

Table 2: Evaluation results on VL-RewardBench. The challenging cases filtered out by small models pose consistent challenges even for different and larger VL-GenRMs. The best results are shown in **bold** and the second best is with underline.

Models	General	Hallucination	Reasoning	Overall Accuracy	Macro Average Accuracy
<i>Open-Source Models</i>					
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
<i>Proprietary Models</i>					
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	<u>62.3</u>	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>

### 3 Experiments

We conduct extensive experiments to evaluate state-of-the-art LVLMs on VL-RewardBench using a systematic evaluation framework. This section presents our experimental setup and findings, organized as follows: evaluated models (Section 3.1), evaluation methodology (Section 3.2), and results analysis (Section 3.3).

#### 3.1 Evaluated Models

We evaluate 16 state-of-the-art LVLMs, encompassing both open-source and commercial models. Open-source models with parameters ranging from 4B to 90B are selected, including LLaVA-OneVision-7B-ov [24], InternVL2-8B [8], Phi-3.5-Vision (4.2B) [1], Qwen2-VL (7B/72B) [48], Llama-3.2 (11B/90B) [12], Molmo- (7B/72B) [11], Pixtral-12B [2], and NVLM-D-72B [10]. For the commercial models, we include prominent options such as GPT-4o/4o-mini [34], Gemini-1.5-Flash/Pro [16], and Claude-3.5-Sonnet [3]. Additionally, we incorporate LLaVA-Critic models [55] on examining the impact of learning to judge for VL-GenRMs (Section 4.3). Model details are provided in the Appendix B.

#### 3.2 Evaluation Settings

We adopt a rigorous evaluation protocol following the LLM-as-a-Judge paradigm [47, 68]. For each test sample, we provide the model with a multimodal input query and two candidate responses (preferred and rejected) through a standardized evaluation template (detailed in the Appendix). To mitigate positional bias [47], where models favor responses based on their presentation order, we conduct  $K$  independent evaluations for each preference pair with randomized response ordering. The final preference is determined through majority voting across these  $K$  runs. We calculate two primary metrics: *Overall Accuracy*: percentage of model decisions matching human preferences, and *Macro Average Accuracy*: mean accuracy across different task categories, addressing task distribution imbalance. Our main results use  $K = 5$ , with detailed analysis of different  $K$  values presented in Section 4.2. All experiments use fixed decoding parameters (temperature=0.2, top-p=0.2), as our validation shows minimal impact from these settings.

#### 3.3 Evaluation Results

**Main Results.** Table 2 presents comprehensive evaluation results across various VL-GenRMs on VL-RewardBench. We have the following observations: (i) The benchmark reveals a clear performance gap among current models, with Gemini-1.5-Pro and GPT-4o leading at 62.5% and 62.4% macro

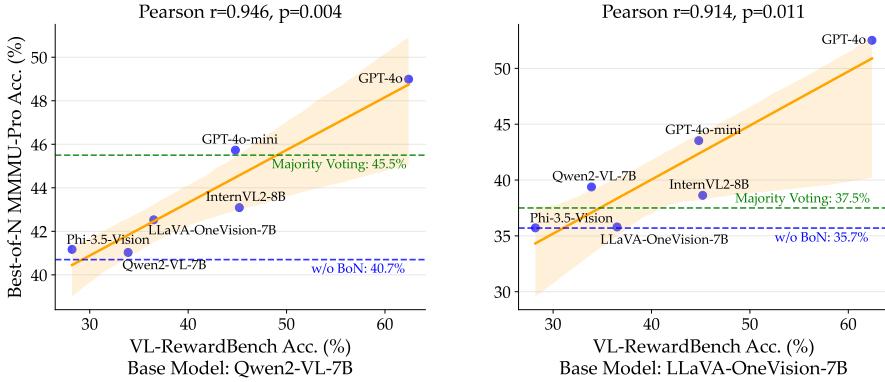


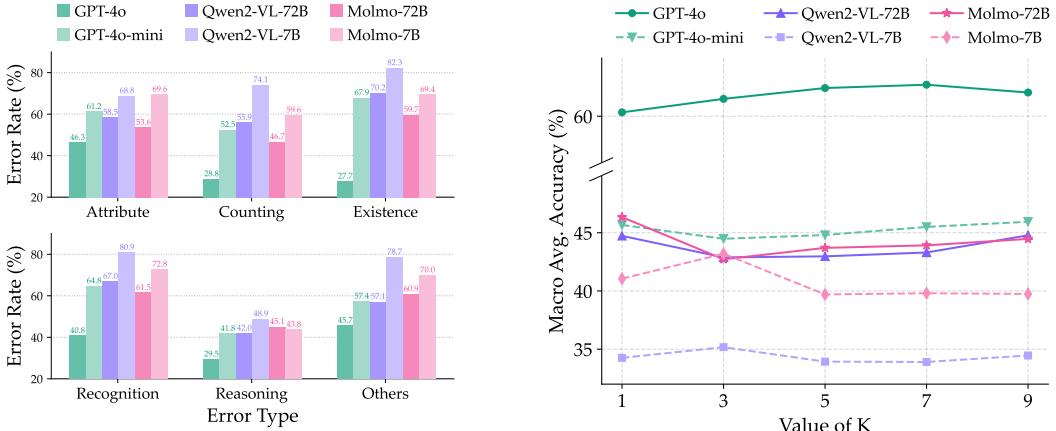
Figure 4: VL-GenRMs accuracy on VL-RewardBench correlates positively with the improvements to serve as Best-of-N selector.

average accuracy, followed by open-source models like Llama-3.2-90B at 53.9%, while most 7B-scale models barely exceed random chance. This gap validates the effectiveness of our ensemble filtering process to find universally challenging samples for VL-GenRMs. (ii) Performance stratification is consistently observed across task categories with varying difficulty levels. Multimodal reasoning tasks see the highest accuracies (51.1% to 70.5%), suggesting models have developed certain capabilities in judging responses involving reasoning paths. In contrast, general instructions prove most challenging (28.0% to 50.8%), indicating significant room for improvement in open-ended queries. Particularly notable is the hallucination task performance, where even top models struggle, validating our dataset’s effectiveness in capturing inherent biases that persist across model scales. (iii) Model scale emerges as a key performance driver, evidenced by consistent improvements across model families: Llama-3.2 (11B to 90B: 42.8% to 53.9%), Qwen2-VL (7B to 72B: 33.9% to 43.0%), and Molmo series (7B to 72B: 39.7% to 43.7%). This scaling effect extends to commercial models, with GPT-4o (62.4%) substantially outperforming GPT-4o-mini (44.8%). In summary, our evaluation demonstrates that VL-RewardBench presents unique challenges beyond conventional datasets, with even state-of-the-art models achieving modest performance. To validate that these challenges stem from our targeted example selection rather than task distribution, we conduct ablation studies with randomly sampled examples following the same task distribution (detailed in our Appendix). VL-GenRMs consistently achieve better results on random samples, e.g., Gemini models achieve accuracy scores higher than 95%, confirming the effectiveness of our data curation strategy in identifying challenging cases.

**Downstream Task Correlation.** A key question is whether VL-RewardBench’s evaluation of VL-GenRMs predicts their real-world utility [69]. To investigate this, we focus on Best-of-N (BoN) sampling [41], a crucial technique for improving model outputs through preference-based selection. We conduct experiments with two base models (Qwen2-VL-7B and LLaVA-OneVision-7B-ov) on the MMMU-Pro benchmark. For each query, we generate  $N = 8$  candidate responses and use six different VL-GenRMs for pairwise scoring and selection. This setup allows us to measure how each VL-GenRM’s preference judgment capability translates to practical performance gains. The results reveal a clear relationship between VL-RewardBench performance and downstream effectiveness. The strongest model, GPT-4o, improves LLaVA-OneVision-7B-ov’s accuracy substantially (35.7% to 52.5%). This pattern generalizes across all VL-GenRMs, showing strong correlations between VL-RewardBench accuracy and BoN performance gains (Pearson  $r = 0.946$  for Qwen2-VL-7B,  $r = 0.914$  for LLaVA-OneVision-7B-ov). These findings validate VL-RewardBench as a reliable predictor of VL-GenRM effectiveness in practical alignment tasks, offering concrete guidance for model selection and development.

## 4 Analysis

To better understand VL-GenRM capabilities and limitations, we conduct three investigations: (1) an error pattern analysis across model scales (Section 4.1), (2) a study of inference-time scaling effects on performance (Section 4.2), and (3) an exploration of potential improvements through critic training and scoring methods (Section 4.3).



(a) Error rate analysis across different types. VL-GenRMs suffer more from perception-related errors than reasoning tasks, and the model scale consistently reduces errors.

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

Figure 5: (a) Analysis of error rates on different tags, and (b) performance with varying  $K$ .

#### 4.1 Error Analysis

To understand systematic patterns in model failures, we analyze judgment errors using manually annotated error types (Section 2.2.1) and treat all samples from the reasoning subset as a Reasoning type. For each error type  $t$ , we calculate the error rate as  $\frac{E_{\text{wrong}}}{E_t}$ , where  $E_t$  represents the total samples of type  $t$  and  $E_{\text{wrong}}$  denotes incorrect predictions within that type. Figure 5a presents these error distributions across GPT-4o, Qwen2-VL, and Molmo model series. Our analysis reveals two key patterns in current VL-GenRM limitations and their relationship to model scale. First, fundamental perception capabilities emerge as the primary bottleneck. Tasks involving Existence (object presence detection) and Recognition (attribute discernment) consistently show the highest error rates across all models. For instance, even advanced models struggle with basic perception: GPT-4o-mini shows a 67.9% error rate on Existence tasks, while Qwen2-VL-7B fails on 80.9% of Recognition tasks. In contrast, Reasoning tasks demonstrate relatively better performance with an average error rate of 41.8%, suggesting that higher-level reasoning capabilities are more robust than basic perception in current models. Second, while model scaling brings consistent improvements, these gains vary significantly across task types. The most substantial improvements occur in basic perception tasks, e.g., scaling Qwen2-VL from 7B to 72B reduces Counting errors by 18.2 percentage points (74.1% to 55.9%). However, Reasoning tasks show more modest gains, with only a 6.0% average error reduction through scaling. This pattern suggests that while scaling effectively addresses some perception limitations, more fundamental architectural innovations may be needed to advance complex reasoning capabilities.

#### 4.2 Does Inference-time Scaling Help?

Given the success of inference-time scaling in improving large language models performance [5, 40, 54], we investigate whether similar benefits extend to VL-GenRMs. Our analysis focuses on the impact of multiple independent judgments per query, implementing a majority voting strategy that has proven effective in text-only scenarios [50]. For each evaluation, we collect  $K$  independent judgments ( $K$  ranging from 1 to 9), carefully randomizing response ordering to minimize positional bias [47]. The results, presented in Fig. 5b, reveal three distinct scaling patterns across different models: (i) GPT-4o demonstrates traditional scaling advantages, with macro accuracy improving from 60.3% to 62.7% as  $K$  increases from 1 to 7, suggesting robust judgment capabilities that benefit from additional computation; (ii) GPT-4o-mini maintains relatively constant performance across  $K$  values, indicating that additional judgments neither help nor harm its decision-making process; (iii) Surprisingly, many open-source LVLMs show performance degradation with increased  $K$ . Notable examples include Qwen2-VL-72B and Molmo-72B, which experience accuracy drops of 1.7 and 2.6

percentage points respectively when scaling from  $K = 1$  to  $K = 5$ . These divergent patterns suggest that successful inference-time scaling strategies from text-only domains may not directly transfer to visual judgment tasks, highlighting the need for specialized scaling approaches for VL-GenRMs.

### 4.3 Critic Training Improves VL-GenRMs

Given the limited benefits of inference-time scaling for most models, we explore an alternative enhancement strategy: critic training [55]. This approach involves specifically training LVLMs to judge response quality through carefully curated instruction tuning samples. We evaluate two variants of LLaVA-OneVision critics [55], each employing distinct judgment strategies: (i) a pointwise critic that independently scores individual answers, and (ii) a pairwise critic that directly compares two candidate answers. Using official model weights and scoring templates, we evaluate both critics on VL-RewardBench, deriving preferences from extracted scores for the pointwise critic. Our analysis<sup>1</sup> reveals two significant findings (Fig. 6): First, critic training substantially enhances judgment capabilities, with both approaches showing remarkable improvements over the base model, i.e., +14.7% and +9.2% for pointwise and pairwise critics, respectively. Second, the pointwise critic achieves better overall performance (52.9% vs 47.4%), while each approach shows distinct advantages: the pointwise critic excels at the hallucination subset (+9.1% over pairwise), and the pairwise critic demonstrates a superior 60.0% accuracy on reasoning tasks. These results suggest that critic training offers a reliable path to improving VL-GenRMs, with different scoring paradigms providing complementary benefits across evaluation scenarios.

### 4.4 Takeaways

Our analysis reveals three key insights for advancing VL-GenRM development: (i) **Improving Visual Perception:** While VL-GenRMs show promise in reasoning, they struggle with basic perception and recognition. Enhancing visual perception capabilities should be prioritized [45], potentially through visual search mechanisms [53] and vision expert integration [38, 18] [66]. (ii) **Advancing Scaling Strategies:** Given that traditional test-time scaling benefits only the largest models (e.g., GPT-4o), future work should explore advanced reasoning strategies incorporating complex planning [58], process-level supervision [49], and specialized approaches like critic training [55]. (iii) **Enabling Co-evolution:** The strong correlation between MMMU-Pro and VL-RewardBench performance suggests a promising improvement cycle: strong LVLMs enable better VL-GenRMs, which curate higher-quality training data, leading to further LVLM improvements. This cycle offers a systematic framework for addressing both perception and scaling challenges through iterative enhancement.

## 5 Related Work

**Large Vision-Language Models** LVLMs have rapidly evolved by combining LLMs [57, 46] with vision encoders [39, 63], showing impressive capabilities across diverse tasks [14, 61, 32, 27]. Key advances include architectural innovations [9, 48], high-quality dataset curation [23, 4, 26], alignment through feedback [43, 59, 60], and systematic design space exploration [20, 45]. Our study evaluates these state-of-the-art LVLMs and demonstrates their limited capability to serve as VL-GenRMs.

**Vision-Language Generative Reward Models** LVLMs have emerged as VL-GenRMs for preference alignment [35, 43] and data curation [28]. Recent studies have investigated the reliability of this approach. Chen et al. [6] developed a benchmark of academic tasks showing that VL-GenRMs achieve substantial agreement with human annotators in comparative assessments, corroborating the

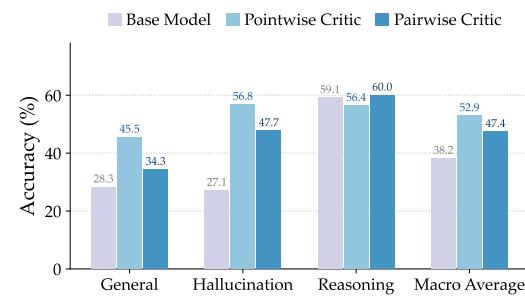


Figure 6: Evaluation of LLaVA-Critic on VL-RewardBench. Critic training greatly improves judgment accuracy.

<sup>1</sup>Results are reported on a subset where pointwise scores were successfully extracted to ensure fair comparison.

findings of VLFeedback [25]. Additionally, LLaVA-Critic [55] established a framework for training generalist VL-GenRMs through a curated critic instruction-following dataset. Our VL-RewardBench advances this research by providing broader coverage of real-world queries and reasoning tasks, with a novel difficulty elevation pipeline targeting challenging cases where even state-of-the-art models struggle. Furthermore, our detailed analysis provides insights for future VL-GenRMs improvements.

## 6 Conclusions

In this paper, we present VL-RewardBench, a benchmark that raises the bar for evaluating VL-GenRMs through systematically curated challenging cases and complex multimodal reasoning tasks. Our comprehensive analysis of 16 state-of-the-art LVLMs identifies crucial limitations in current approaches while highlighting promising directions: the benefits of increased model scale, the variable effectiveness of test-time scaling techniques, and the potential of specialized critic training. These insights, combined with our benchmark, provide concrete pathways for developing more capable and reliable VL-GenRMs.

## Limitations

(i) Our preference pairs, derived from existing public benchmarks, may face potential data contamination, potentially affecting the benchmark’s reliability for evaluating newer models. (ii) While our human verification process ensures quality, the benchmark’s scope is limited by the inherent biases in the source datasets and the specific capabilities of the proprietary model used in our AI-aided annotation process. Despite these limitations, we believe VL-RewardBench takes a critical step for VL-GenRMs evaluation. We are committed to evolving VL-RewardBench alongside LVLM developments.

## References

- [1] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv preprint*, abs/2404.14219, 2024. [6](#)
- [2] P. Agrawal, S. Antoniak, E. B. Hanna, D. Chaplot, J. Chudnovsky, S. Garg, T. Gervet, S. Ghosh, A. Héliou, P. Jacob, et al. Pixtral 12b. *ArXiv preprint*, abs/2410.07073, 2024. [6](#)
- [3] Anthrophic. Claude 3.5 sonnet, 2023. [2](#), [6](#)
- [4] A. Awadalla, L. Xue, O. Lo, M. Shu, H. Lee, E. K. Guha, M. Jordan, S. Shen, M. Awadalla, S. Savarese, C. Xiong, R. Xu, Y. Choi, and L. Schmidt. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *ArXiv preprint*, abs/2406.11271, 2024. [9](#)
- [5] B. Brown, J. Juravsky, R. Ehrlich, R. Clark, Q. V. Le, C. R’e, and A. Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *ArXiv preprint*, abs/2407.21787, 2024. [8](#)
- [6] D. Chen, R. Chen, S. Zhang, Y. Liu, Y. Wang, H. Zhou, Q. Zhang, P. Zhou, Y. Wan, and L. Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *International Conference on Machine Learning*, 2024. [2](#), [9](#)
- [7] Z. Chen, Y. Du, Z. Wen, Y. Zhou, C. Cui, Z. Weng, H. Tu, C. Wang, Z. Tong, Q. Huang, C. Chen, Q. Ye, Z. Zhu, Y. Zhang, J. Zhou, Z. Zhao, R. Rafailov, C. Finn, and H. Yao. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation?, 2024. [2](#)
- [8] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *ArXiv preprint*, abs/2404.16821, 2024. [6](#)

- [9] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. C. H. Hoi. Instructclip: Towards general-purpose vision-language models with instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 9
- [10] W. Dai, N. Lee, B. Wang, Z. Yang, Z. Liu, J. Barker, T. Rintamaki, M. Shoeybi, B. Catanzaro, and W. Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint*, 2024. 6
- [11] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *ArXiv preprint*, abs/2409.17146, 2024. 6
- [12] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Srivankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Rozière, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. M. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, and et al. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783, 2024. 1, 2, 6
- [13] Y. Dubois, B. Galambosi, P. Liang, and T. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *ArXiv preprint*, abs/2404.04475, 2024. 2, 5
- [14] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, K. Li, X. Sun, and R. Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv preprint*, abs/2306.13394, 2023. 1, 9
- [15] S. Gehrmann, T. Adewumi, K. Aggarwal, P. S. Ammanamanchi, A. Aremu, A. Bosselut, K. R. Chandu, M.-A. Clinciu, D. Das, K. Dhole, W. Du, E. Durmus, O. Dušek, C. C. Emezue, V. Gangal, C. Garbacea, T. Hashimoto, Y. Hou, Y. Jernite, H. Jhamtani, Y. Ji, S. Jolly, M. Kale, D. Kumar, F. Ladzhak, A. Madaan, M. Maddela, K. Mahajan, S. Mahamood, B. P. Majumder, P. H. Martins, A. McMillan-Major, S. Mille, E. van Miltenburg, M. Nadeem, S. Narayan, V. Nikolaev, A. Niyongabo Rubungo, S. Osei, A. Parikh, L. Perez-Beltrachini, N. R. Rao, V. Raunak, J. D. Rodriguez, S. Santhanam, J. Sedoc, T. Sellam, S. Shaikh, A. Shimorina, M. A. Sobrevilla Cabezudo, H. Strobel, N. Subramani, W. Xu, D. Yang, A. Yerukola, and J. Zhou. The GEM benchmark: Natural language generation, its evaluation and metrics. In A. Bosselut, E. Durmus, V. P. Gangal, S. Gehrmann, Y. Jernite, L. Perez-Beltrachini, S. Shaikh, and W. Xu, editors, *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, 2021. doi: 10.18653/v1/2021.gem-1.10. 1
- [16] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint*, abs/2403.05530, 2024. 1, 6
- [17] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi. CLIPScore: A reference-free evaluation metric for image captioning. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021. doi: 10.18653/v1/2021.emnlp-main.595. 1
- [18] Y. Hu, O. Stretcu, C.-T. Lu, K. Viswanathan, K. Hata, E. Luo, R. Krishna, and A. Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models, 2023. 9
- [19] L. Jing and X. Du. Fgaif: Aligning large vision-language models with fine-grained ai feedback, 2024. 1

- [20] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *International Conference on Machine Learning (ICML)*, 2024. 9
- [21] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 16
- [22] N. Lambert, V. Pyatkin, J. D. Morrison, L. J. V. Miranda, B. Y. Lin, K. R. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, N. A. Smith, and H. Hajishirzi. Rewardbench: Evaluating reward models for language modeling. *ArXiv preprint*, abs/2403.13787, 2024. 1, 2, 3
- [23] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, M. Cord, and V. Sanh. OBELICS: an open web-scale filtered dataset of interleaved image-text documents. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 9
- [24] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, and C. Li. Llava-onevision: Easy visual task transfer. *ArXiv preprint*, abs/2408.03326, 2024. 2, 4, 6
- [25] L. Li, Z. Xie, M. Li, S. Chen, P. Wang, L. Chen, Y. Yang, B. Wang, and L. Kong. Silkie: Preference distillation for large visual language models. *ArXiv preprint*, abs/2312.10665, 2023. 1, 2, 3, 10
- [26] L. Li, Y. Yin, S. Li, L. Chen, P. Wang, S. Ren, M. Li, Y. Yang, J. Xu, X. Sun, L. Kong, and Q. Liu. M<sup>3</sup>IT: A large-scale dataset towards multi-modal multilingual instruction tuning. *ArXiv preprint*, abs/2306.04387, 2023. 9
- [27] L. Li, Y. Liu, L. Yao, P. Zhang, C. An, L. Wang, X. Sun, L. Kong, and Q. Liu. Temporal reasoning transfer from text to video. *ArXiv preprint*, abs/2410.06166, 2024. 9
- [28] L. Li, Z. Xie, M. Li, S. Chen, P. Wang, L. Chen, Y. Yang, B. Wang, L. Kong, and Q. Liu. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. *ArXiv preprint*, abs/2410.09421, 2024. 9
- [29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2
- [30] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning, 2023. 4
- [31] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 4
- [32] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv preprint*, abs/2310.02255, 2023. 9
- [33] Y. Lu, D. Jiang, W. Chen, W. Y. Wang, Y. Choi, and B. Y. Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *ArXiv preprint*, abs/2406.11069, 2024. 2, 3
- [34] OpenAI. Gpt-4o system card, 2024. 1, 6
- [35] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 9

- [36] P. Padlewski, M. Bain, M. Henderson, Z. Zhu, N. Relan, H. Pham, D. Ong, K. Aleksiev, A. Ormazabal, S. Phua, et al. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models. *ArXiv preprint*, abs/2405.02287, 2024. 4
- [37] A. Panickssery, S. R. Bowman, and S. Feng. Llm evaluators recognize and favor their own generations. *ArXiv preprint*, abs/2404.13076, 2024. 2
- [38] R. Pi, J. Zhang, J. Zhang, R. Pan, Z. Chen, and T. Zhang. Image textualization: An automatic framework for creating accurate and detailed image descriptions, 2024. 9
- [39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021. 9
- [40] C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *ArXiv preprint*, abs/2408.03314, 2024. 8
- [41] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 7
- [42] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. J. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano. Learning to summarize from human feedback. *ArXiv preprint*, abs/2009.01325, 2020. 2
- [43] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, K. Keutzer, and T. Darrell. Aligning large multimodal models with factually augmented rlhf. *ArXiv preprint*, abs/2309.14525, 2023. 1, 9
- [44] S. Tan, S. Zhuang, K. Montgomery, W. Y. Tang, A. Cuadron, C. Wang, R. A. Popa, and I. Stoica. Judgebench: A benchmark for evaluating llm-based judges. *ArXiv preprint*, abs/2410.12784, 2024. 2
- [45] S. Tong, E. Brown, P. Wu, S. Woo, M. Middepogu, S. C. Akula, J. Yang, S. Yang, A. Iyer, X. Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *ArXiv preprint*, abs/2406.16860, 2024. 9
- [46] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971, 2023. 9
- [47] P. Wang, L. Li, L. Chen, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui. Large language models are not fair evaluators. *ArXiv preprint*, abs/2305.17926, 2023. 4, 6, 8
- [48] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *ArXiv preprint*, abs/2409.12191, 2024. 2, 4, 6, 9
- [49] P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, 2024. doi: 10.18653/v1/2024.acl-long.510. 9
- [50] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023. 8

- [51] Z. Wang, A. Bukharin, O. Delalleau, D. Egert, G. Shen, J. Zeng, O. Kuchaiev, and Y. Dong. Helpsteer2-preference: Complementing ratings with preferences, 2024. 1
- [52] M. Wu and A. F. Aji. Style over substance: Evaluation biases for large language models. *ArXiv preprint*, abs/2307.03025, 2023. 2
- [53] P. Wu and S. Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 9
- [54] Y. Wu, Z. Sun, S. Li, S. Welleck, and Y. Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *ArXiv preprint*, abs/2408.00724, 2024. 8
- [55] T. Xiong, X. Wang, D. Guo, Q. Ye, H. Fan, Q. Gu, H. Huang, and C. Li. Llava-critic: Learning to evaluate multimodal models. *ArXiv preprint*, abs/2410.02712, 2024. 1, 2, 6, 9, 10
- [56] W. Xu, G. Zhu, X. Zhao, L. Pan, L. Li, and W. Wang. Pride and prejudice: LLM amplifies self-bias in self-refinement. In L.-W. Ku, A. Martins, and V. Srikanth, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, 2024. doi: 10.18653/v1/2024.acl-long.826. 2
- [57] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, and Z. Fan. Qwen2 technical report. *ArXiv preprint*, abs/2407.10671, 2024. 9
- [58] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 9
- [59] T. Yu, Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, J. Hu, Z. Liu, H.-T. Zheng, M. Sun, and T.-S. Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arxiv*, 2023. 2, 3, 9
- [60] T. Yu, H. Zhang, Y. Yao, Y. Dang, D. Chen, X. Lu, G. Cui, T. He, Z. Liu, T.-S. Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *ArXiv preprint*, abs/2405.17220, 2024. 1, 2, 3, 9
- [61] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024. 1, 4, 9
- [62] X. Yue, T. Zheng, Y. Ni, Y. Wang, K. Zhang, S. Tong, Y. Sun, B. Yu, G. Zhang, H. Sun, Y. Su, W. Chen, and G. Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *ArXiv preprint*, abs/2409.02813, 2024. 2, 3, 5
- [63] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952, 2023. doi: 10.1109/ICCV51070.2023.01100. 9
- [64] R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, P. Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *ArXiv preprint*, abs/2403.14624, 2024. 2, 3
- [65] X. Zhang, Y. Lu, W. Wang, A. Yan, J. Yan, L. Qin, H. Wang, X. Yan, W. Y. Wang, and L. R. Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks. *ArXiv preprint*, abs/2311.01361, 2023. 1

- [66] Z. Zhang, R. A. Rossi, T. Yu, F. Dernoncourt, R. Zhang, J. Gu, S. Kim, X. Chen, Z. Wang, and N. Lipka. Vipact: Visual-perception enhancement via specialized vlm agent collaboration and tool-use. *ArXiv preprint*, abs/2410.16400, 2024. 9
- [67] W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. 2
- [68] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 6
- [69] E. Zhou, G. Zheng, B. Wang, Z. Xi, S. Dou, R. Bao, W. Shen, L. Xiong, J. Fan, Y. Mou, et al. Rmb: Comprehensively benchmarking reward models in llm alignment. *ArXiv preprint*, abs/2410.09893, 2024. 2, 7
- [70] Y. Zhou, C. Cui, R. Rafailov, C. Finn, and H. Yao. Aligning modalities in vision large language models via preference fine-tuning, 2024. 2, 3
- [71] Y. Zhou, Z. Fan, D. Cheng, S. Yang, Z. Chen, C. Cui, X. Wang, Y. Li, L. Zhang, and H. Yao. Calibrated self-rewarding vision language models. *ArXiv preprint*, abs/2405.14622, 2024. 1

## 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 Table 3, 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.

### A.2 Influence of Decoding Parameters

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 4, 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.

Table 3: Performance comparison on different datasets. VL-RewardBench provides a much better separability for VL-GenRMs.

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

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

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 5: Specification of the evaluated models in our main paper.

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

## B Model Details

We provide the details of the evaluated models in our main paper in Table 5. For open-source models, models are served locally with vLLM [21] 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 Table 6, where annotators could categorize errors according to the taxonomy presented in Table 7. To handle reasoning tasks specifically, we implemented a dedicated GUI that enables clearer visualization of mathematical rationales. This interface, illustrated in Table 8, displays the ground truth alongside the model outputs, allowing annotators to evaluate responses using the guideline detailed in Table 9.

## D Cases of Different Judgment Errors

We show annotated cases corresponding to the following eight error categories: object recognition (Table 10), text recognition (Table 11), scene recognition (Table 12), position recognition (Table 13), face recognition (Table 14), counting (Table 15), attribute (Table 16) and existence/hallucination (Table 17). 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.

## E Templates

Table 18 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

Table 6: Graphical user interface for preference and error annotation of general/hallucination tasks.

### Annotating general/hallucination tasks

**ID: RLAIF-V-59085**

Image Path:images/rif-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

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

the ground-truth answer is demonstrated in Table 19. In both evaluation templates, we will randomly place candidate answers to avoid positional bias.

Table 7: Guidelines for annotating general/hallucination tasks.

<b>Guidelines for annotating general/hallucination tasks</b>
<p>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.</p> <p>Please evaluate both answers based on the following criteria:</p> <ol style="list-style-type: none"> <li>1. Accuracy: Are there any discrepancies or incorrect statements that are not aligned with the image in the response?</li> <li>2. Completeness: Are there any important details or aspects of the question missing?</li> <li>3. Clarity: Is the answer easy to read and understand?</li> <li>4. Relevance: Does the answer directly address the question and the image?</li> </ol> <p>After you evaluate both responses, choose the better one and annotate the error types. If you think both responses are incorrect, please press the "<b>Inaccurate (Both False)</b>" button. If you think Response 2 is better than Response 1, please press the "<b>Inaccurate (Inverted)</b>" 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:</p> <p><b>Categories of annotated error types</b></p> <p><b>Counting:</b> Mistakes in accurately identifying the number of objects or elements in an image.</p> <p><b>Text Recognition:</b> Difficulty in accurately reading or interpreting text within images.</p> <p><b>Position Recognition:</b> Misidentifying the relative position or spatial relationships between objects.</p> <p><b>Scene Recognition:</b> Misinterpreting the overall context or setting of an image.</p> <p><b>Face Recognition:</b> Inability to accurately recognize or differentiate the features of faces.</p> <p><b>Attribute:</b> Misidentifying or confusing attributes like color, size, shape, age, posture and state of objects or people.</p> <p><b>Hallucination:</b> Imagining or asserting the existence of objects, details, or information that are not present in the image.</p> <p><b>Vague (Style):</b> 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).</p> <p><b>Vague (Knowledge):</b> Cannot judge because the annotator lacks the relevant knowledge (e.g. medical images).</p> <p><b>Poor Image Quality:</b> Cannot judge because the image is hard to distinguish or has resolution issues.</p> <p><b>Others:</b> 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.</p> <p>After you annotate the preference and error type, you can press "<b>Next</b>" to move on to the next datum, or press "<b>Previous</b>" to return to the previous datum. You can enter an "index" and press "<b>Jump</b>" button to jump to the (index+1)th datum.</p>

Table 8: Graphical user interface for preference annotation of reasoning 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 the image 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 related 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

Previous

Next

Inaccurate(Both False)

First Better

Second Better

Vague(Style)

Vague(Knowledge)

Poor Image Quality

Other

Enter index

Jump to

Table 9: Guidelines for annotating reasoning tasks.

**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.

Table 10: An example of object recognition error.

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 11: An example of text 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 12: An example of scene 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 13: An example of position 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 14: An example of face 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 15: An example of counting 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 16: An example of attribute 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 17: An example of existence 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 18: Evaluation template for pairwise comparison.

**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 19: Annotation template for reasoning tasks used by GPT-4o.

**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.