Skywork-VL Reward: An Effective Reward Model for Multimodal Understanding and Reasoning

Xiaokun Wang[*] Chris[*] Jiangbo Pei[*] Yunzhuo Hao, Weijie Qiu,
Ai Jian, Tianyidan Xie, Xuchen Song[†] Yang Liu[†] Yahui Zhou

Skywork AI, Kunlun Inc.
xuchen.song@kunlun-inc.com

Abstract

We propose Skywork-VL Reward, a multimodal reward model that provides reward signals for both multimodal understanding and reasoning tasks. Our technical approach comprises two key components: First, we construct a large-scale multimodal preference dataset that covers a wide range of tasks and scenarios, with responses collected from both standard vision-language models (VLMs) and advanced VLM reasoners. Second, we design a reward model architecture based on Qwen2.5-VL-7B-Instruct, integrating a reward head and applying multi-stage fine-tuning using pairwise ranking loss on preference-labeled response pairs. Experimental evaluations show that Skywork-VL Reward achieves state-of-the-art results on VL-RewardBench (vision-language) and exhibits competitive performance on the text-only RewardBench benchmark. Furthermore, preference data constructed based on our Skywork-VL Reward proves highly effective for training Mixed Preference Optimization (MPO), leading to significant improvements in multimodal reasoning capabilities. Our results underscore Skywork-VL Reward as a significant advancement toward general-purpose, reliable reward models for multimodal alignment. Our model has been publicly released to promote transparency and reproducibility[‡].

1 Introduction

Recent advancements in large language models (LLMs) and vision-language models (VLMs) have shown significant capabilities, yet aligning these models effectively with human preferences remains challenging [1–3]. Reward models (RMs) have emerged as critical tools to address this alignment issue, playing essential roles during both training and inference phases of LLMs and VLMs [4–6]. For instance, RMs guide model training through reinforcement learning from human feedback (RLHF), providing signals that steer models toward desirable behaviors [7]. At inference time, they function as evaluators or judges of model outputs, supporting methods like best-of-N sampling and reranking to enhance output quality [8].

Although RMs for purely text-based LLMs are relatively well-explored, multimodal RMs face significant challenges, primarily due to the scarcity of high-quality human-annotated preference data across both image and text modalities. Human annotation of multimodal preferences, such as ranking responses involving images and texts, is expensive and time-consuming, with few public datasets available. Consequently, existing multimodal RMs remain underdeveloped and typically constrained

[*] Equal contribution
[†] Corresponding author
[‡] https://huggingface.co/Skywork/Skywork-VL-Reward-7B

Tech report. Preprint.

to narrow domains [6]. There is, therefore, a pressing need for robust multimodal RMs capable of broadly addressing diverse user queries and visual inputs encountered by general-purpose VLMs.

In this paper, we introduce Skywork-VL Reward, a multimodal RM designed to serve as a comprehensive and robust evaluator for VLM outputs. Our approach addresses previous limitations in domain coverage and reasoning capacity by incorporating two critical improvements: (i) creating a carefully curated multimodal preference dataset derived from various sources and tasks, and (ii) developing a strong base model and training paradigm to enable effective vision-language understanding and reasoning. Specifically, we compile high-quality preference pairs from both publicly available datasets and internal annotations, covering a spectrum of tasks ranging from everyday image descriptions to complex reasoning scenarios. The collected preference pairs includes the image (when applicable), textual prompt, and candidate responses sourced from both standard VLMs [9, 10] and advanced VLM reasoners [3]. Building on this dataset, we construct Skywork-VL Reward based on Qwen2.5-VL-7B-Instruct, with an integrated reward head designed to output scalar scores aligned with human preferences. The model is trained using a two-stage training paradigm that combines both pure-text and multimodal data, which enhances its generalization and performance across a wide range of multimodal scenarios. Experimental evaluations confirm that Skywork-VL Reward achieves state-of-the-art results on emerging multimodal RM benchmarks while maintaining competitive performance in text-only scenarios. Furthermore, preference data constructed based on our Skywork-VL Reward proves highly effective for training Mixed Preference Optimization (MPO) [11], leading to significant improvements in multimodal reasoning capabilities. We have made our model publicly available to encourage further research and application.

2 Related Work

Reward Models for Large Language Models. Reward modeling has become a cornerstone of aligning LLM behavior with human preferences [12–14]. Generally, RMs are trained on comparisons of outputs (chosen vs. rejected responses) to predict which output is better, often using data collected from human raters or AI assistants. Existing RMs can be categorized along two axes: (1) the model form [7] (discriminative RMs vs. generative RMs vs. implicit RMs) and (2) the feedback target (outcome-based vs. process-based supervision). Discriminative RMs [15, 2, 16] treat preference prediction as a binary (or scalar) regression problem: given an input and a candidate response, the model directly outputs a score (or probability of being the preferred response). By contrast, generative RMs use a language-model head to generate an evaluation or verdict based on a specific prompt [17, 7, 18], rather than directly outputting a numeric score. A third category, implicit RMs [19], effectively reparameterize preference learning within the model itself via Direct Preference Optimization (DPO) [20], enables construct preference pairs without requiring an explicit RM.

Orthogonal to the above, RM can be divided into outcome-level RMs (ORMs) and process-level RMs (PRMs). ORMs [21] output a single score at the end of response, while PRMs [22] output a trajectory of scores or a cumulative reward. Our Skywork-VL Reward belongs to the discriminative RM and ORM.

Reward Models for Vision-Language Models. Extending reward modeling to multimodal models is an active area of research, motivated by the observation that VLMs can similarly benefit from preference-based alignment [23]. Early efforts in this space have mirrored techniques from the text domain, albeit on a smaller scale. LLaVA-RLHF [24] was one of the first open attempts to apply RLHF to a vision-language model, using human rankings of model responses to improve an image-dialogue agent. Other works followed by introducing data augmentation to generate "bad" responses [25, 26]. These studies showed that preference tuning can reduce visual hallucinations and produce more helpful visual chatbots.

Recently, the community has started addressing the data scarcity issue for multimodal reward modeling through a variety of strategies. One approach is to synthesize large multi-domain preference datasets using powerful models as proxies for human annotators [27]. This dataset covers pointwise scoring and pairwise ranking on a range of tasks (from captioning and visual QA to complex reasoning), and was used to train LLaVA-Critic as a generalist open-source evaluator. The resulting model can assign quality scores to image responses and even produce textual justifications, approaching the evaluation reliability of GPT-4o [28] itself. In recent work, IXC-2.5-Reward [6] utilizes GPT-4o in

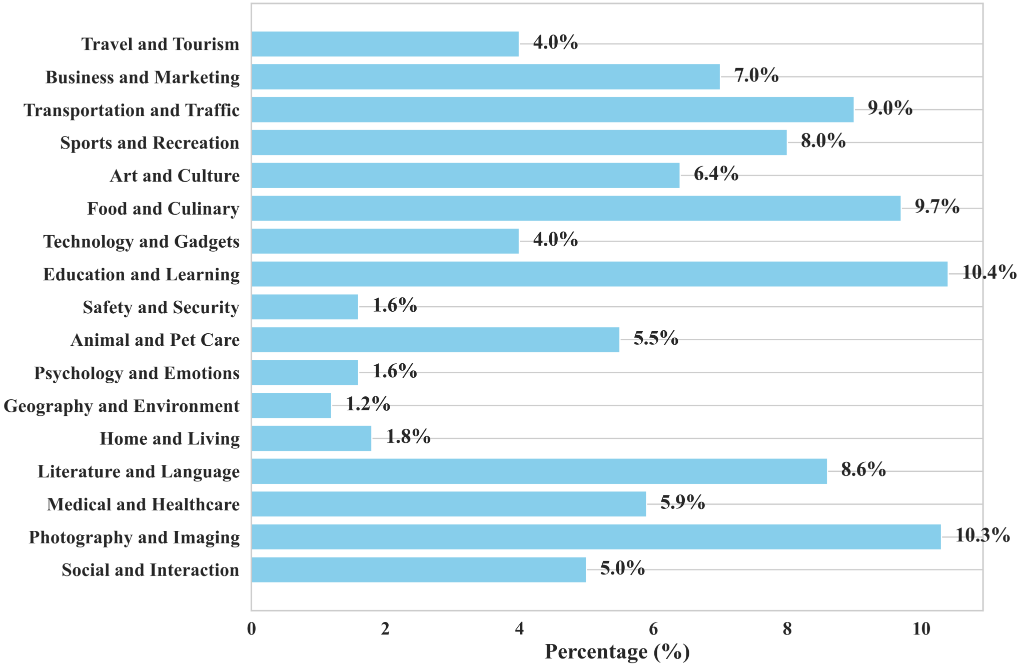


Figure 1: Distribution of Training Data from Open-Source Sources.

conjunction with verifier functions [7] to label preference data, resulting in impressive performance gains.

Another tactic is data curation: rather than sheer quantity, focus on high-quality, diverse preferences. The Skywork-Reward [4] exemplifies this by assembling a carefully filtered set of 80k preference pairs from various open sources. Even though 80k is smaller than some earlier corpora, the strict quality filtering (removing inconsistent or low-informative samples) led to RMs that achieved state-of-the-art on text-only benchmarks. Inspired by this, some multimodal RM efforts incorporate filtered text-only data to bolster learning [6]. By training on text-only comparisons and visual comparisons, they mitigate the lack of visual preference examples. Our approach in Skywork-VL Reward combines several of these insights: we aggregate multiple open datasets and augment them with our own human-annotated cases in hard domains like math reasoning. We also perform extensive cleaning to ensure the final training data is balanced and high-quality.

3 Method

Our goal is to train a multimodal RM, Skywork-VL Reward, that can take an image (optional), a textual prompt, and a candidate response (provided by either a multimodal understanding model or reasoning model), and output a scalar reward score indicating the quality or preference-worthiness of the response. We achieve this by fine-tuning a pretrained vision-language model on a curated set of preference comparison data. In this section, we describe our dataset construction pipeline, the model architecture and modifications for reward modeling, the loss function used for pairwise preference training, and the overall training strategy.

3.1 Dataset Construction

Open-Source Data. We construct a comprehensive training dataset for Skywork-VL Reward by integrating multiple open-source preference datasets and additional in-house annotations. The dataset primarily includes three sources: (1) LLaVA-Critic-113k [27], (2) Skywork-Reward-Preference-80K-v0.2 [4], and (3) RLAIF-V-Dataset [29].

Table 1: Distribution of In-house Training Data.

Field	Mathematics	Physics	Biology	Chemistry	Others
Percentage (%)	35.4	24.6	14.7	20.2	5.1

Table 2: Percentage of Generation Approaches of In-house Training Data.

Approach	Direct generation	Two-stage generation
Percentage (%)	47.4	52.6

LLaVA-Critic-113k is an open dataset consisting of 113k multimodal instruction-response examples. Each example contains an image, a user query, and one or more model responses with associated quality judgments. This dataset uniquely provides both pointwise scores and pairwise rankings generated by GPT-4o, covering tasks from straightforward image descriptions to complex reasoning challenges. Each pair is often accompanied by explanatory annotations, enriching our understanding of judgment criteria.

Skywork-Reward-Preference-80K-v0.2 is a high-quality dataset comprising 80k pairs of human-preferred textual responses, covering diverse domains such as general QA and creative writing. Carefully filtered to eliminate noisy or inconsistent judgments, this dataset significantly boosts the text comprehension and alignment capabilities of Skywork-VL Reward, enabling it to effectively handle purely textual inputs.

RLAIF-V-Dataset is a large-scale multimodal feedback dataset containing 83,132 preference pairs. Instructions in this dataset are sourced from diverse datasets. Incorporating this dataset greatly enhances the general multimodal understanding abilities of Skywork-VL Reward, enabling robust performance across varied tasks and contexts.

Finally, we implemented a two-stage data cleaning and refinement procedure. In the first stage, we deduplicated and filtered the aggregated dataset. Specifically, we identified and removed duplicate pairs that appeared across multiple sources, as well as filtered out highly similar samples based on semantic similarity. Additionally, we discarded pairs exhibiting ambiguous or low-confidence preference judgments with the judgments provided by GPT-4o. Any pair that is identified as "equal quality" was treated as ambiguous. Such ambiguous pairs were excluded from the training data to ensure clarity in training. After completing the first stage, we obtained approximately 200,000 distinct, high-confidence preference pairs. This refined dataset was then employed to train a surrogate RM, which subsequently scored the entire dataset.

In the second stage, we further refined the collected dataset based on scores predicted by the surrogate RM, with the following principles:

- For preference pairs where the chosen response received a low score, we regenerated the response using GPT-4o, replacing the original version.
- For pairs where the score difference between the chosen and rejected responses was minimal, we also regenerated the chosen response to improve clarity.

After this refinement process, a total of 150,000 data samples were retained. Figure 1 presents the distribution of the resulting training data.

In-house Reasoning Data. We augment existing datasets with a proprietary in-house dataset consisting of approximately 50,000 preference comparisons focused on complex reasoning tasks. The tasks primarily involved carefully curated multimodal problems spanning mathematics, physics, biology, and chemistry (Table 1). These comparisons were collected through human annotation, where annotators assessed the correctness and reasoning quality of various VLM-generated reasoning-style responses.

The reasoning-style responses were generated using two distinct methods (Table 2):

- Direct generation: Using our vision-language reasoning model Skywork R1V to directly produce reasoning-style responses conditioned on the original multimodal input.

- Two-stage generation: First employing a standard VLM model (we use the InternVL series [23]) to generate a textual description of the image, which serves as a substitute for the original visual input. Then, this textual description is fed into llm-reasoner (we use the Deepseek R1 series [30]) to generate the final reasoning-style response.

After the above procedure, our final training dataset consists of approximately $\mathbf{190}$k comparison samples (each a prompt with two responses and a preference), of which about 70% include images. We believe this dataset is one of the largest and most diverse multimodal preference collections to date, and it is a key enabler of Skywork-VL Reward's strong generalization.

3.2 Model Architecture

Our RM is built upon the Qwen2.5-VL-7B-Instruct [9]. Qwen2.5-VL-7B-Instruct is a 7-billion-parameter vision-language model recently released by Alibaba Cloud, which extends the Qwen-7B LLM with visual understanding capabilities. The architecture consists of three main components: a Vision Transformer (ViT) as the visual encoder, a vision-language adapter/projector, and the Qwen language model decoder. The visual encoder processes an input image into a sequence of patch features. These raw image features are then passed through a small adapter module that projects them into a compact set of visual tokens that the language model can attend to. The language model then takes these visual tokens, along with text tokens from the prompt, and generates a response.

For Skywork-VL Reward, we modify this architecture to output a scalar reward instead of a text sequence. We remove Qwen's causal LM head used for token prediction and replace it with a reward head that produces a single score. Concretely, we attach a small fully-connected layer on top of the final hidden state to predict the reward. This reward head reads the final hidden state (after the answer's last token) and outputs a raw score r_θ. During training, r_θ for different answers will be used to compute a preference loss (described next). At inference, Skywork-VL Reward can take a prompt and a response and output a score to judge the quality of answer.

3.3 Loss Function

We train Skywork-VL Reward using a standard pairwise preference loss commonly used in RLHF reward modeling. For a given comparison example, we have a prompt (with optional image) x, a preferred response y^+ (the one chosen by the annotator or judge), and a dispreferred response y^- (the one that was rejected). The model produces a scalar score for each: $s^+ = r_\theta(x, y^+)$ and $s^- = r_\theta(x, y^-)$, where r_θ denotes the RM's forward pass. We then define the loss to push s^+ higher than s^-. Specifically, we use the logistic sigmoid loss:

$$\mathcal{L}_{\text{RM}}(\theta) \;=\; -\log \sigma\Big(r_\theta(x, y^+) \;-\; r_\theta(x, y^-)\Big)\,, \tag{1}$$

where $\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function. Intuitively, if the model assigns a higher score to the chosen response than the rejected response, the term inside $\log \sigma(\cdot)$ is positive, leading to a low loss; if the model mistakenly scores the bad response higher, the loss increases. Optimizing this encourages $r_\theta(x, y^+) - r_\theta(x, y^-)$ to be positive, i.e. $r_\theta(x, y^+) > r_\theta(x, y^-)$. This loss is equivalent to maximizing the probability $\sigma(s^+ - s^-)$ that the model ranks the pair correctly.

We found it important to handle cases of equal preferences: if an example was labeled "tie" or had scores indicating near-equality, we either remove it (as mentioned in data cleaning) or treat it as $s^+ \approx s^-$ so that it contributes little to the gradient. We do not perform any explicit calibration of the absolute scores during training – the model is only trained to get the relative ordering correct. This pairwise loss has been shown to be effective and is used by many recent RMs, including those on complex multi-domain data.

3.4 Training Details

Parameter Freezing. To efficiently fine-tune the model as a reward scorer, we adopt a partial parameter freezing strategy. In particular, we freeze the entire visual encoder of Qwen2.5-VL-7B-Instruct during training. This component was pretrained on massive image-text data and already provides strong visual feature extraction. Freezing it reduces GPU memory usage and speeds up training, preserves the pretrained visual capabilities and alleviates the risk of forgetting. Preliminary experiments showed no loss in RM accuracy from freezing the vision module, confirming that the language model can learn to make use of the frozen visual features effectively. Thus, the trainable weights in Skywork-VL Reward are the projector, Qwen language model parameters and the small reward head.

Two-Stage Fine-Tuning Procedure. We formulate preference learning as a supervised learning task over the constructed comparison dataset. The fine-tuning follows a two-stage training strategy. In the first stage, the model is trained exclusively on multimodal preference data, allowing it to develop strong vision-language alignment capabilities. In the second stage, we additionally incorporate pure-text preference data to further enhance the model's generalization and reasoning performance in language-only settings.

We use AdamW with a moderate learning rate for optimization in the first stage (10^{-5}) and a lower learning rate in the second stage (10^{-6}). A batch size of 1 is used throughout, and the model is fine-tuned for two epochs per stage, which we find sufficient for convergence.

4 Experiment

4.1 Evaluation Benchmarks

We evaluate Skywork-VL Reward on two complementary benchmarks: VL-RewardBench [31] and RewardBench [7]. VL-RewardBench is designed to assess vision-language reward modeling. It contains 1,250 carefully curated examples spanning general multimodal queries, visual hallucination detection, and complex reasoning tasks involving images. For each example, a vision-language model provides two candidate responses (with one labeled as preferred), and the RM must determine which response is better.

RewardBench is a recent benchmark targeting reward functions for language models. It comprises prompt–response comparison triples across diverse domains, including conversational dialogue, reasoning tasks (such as math and coding), and safety-critical scenarios. The benchmark evaluates whether a RM can reliably prefer the human-preferred response in challenging cases featuring subtle but verifiable distinctions—such as factual correctness or relevance. We report results on both benchmarks across their key evaluation dimensions and the overall accuracy.

4.2 Comparison Methods

For VL-RewardBench, we compare Skywork-VL Reward against a broad range of RMs, including both cutting-edge proprietary models and leading open-source alternatives.

The proprietary multimodal RMs (closed-source) in our evaluation include GPT-4o [28], Claude 3.5 [32] with Vision, and Google Gemini 1.5 [33]. These models represent top-performing industrial models and serve as upper-bound references for RM performance.

In addition, we evaluate several prominent open-source models, including Qwen2-VL-7B-Instruct [34], MAmmoTH-VL-8B [35], Qwen2.5-VL-7B-Instruct [9], InternVL3-8B [36], Qwen2-VL-72B-Instruct [34], IXC-2.5-Reward-7B [6], Molmo-72B [37], QVQ-72B-Preview [38], Qwen2.5-VL-72B-Instruct [9], and InternVL3-78B [36].

For RewardBench, we evaluate several advanced language-only RMs, including InternLM2-7B-Reward [39], Skywork-Reward-Llama3.1-8B [4], Skywork-Reward-Llama3.1-8B-v0.2 [4], and QRM-Llama3.1-8B-v2 [40]. We also consider VLM RMs that are comparable in size to our own models including Qwen2-VL-7B-Instruct, InternVL3-8B, IXC-2.5-Reward-7B, and Qwen2.5-VL-7B-Instruct.

In all experiments, Qwen2.5-VL-7B-Instruct, InternVL3-8B, Qwen2.5-VL-72B-Instruct, IXC-2.5-Reward-7B, and InternVL3-78B are reproduced by ourselves, while the results of the remaining models are taken from official reports.

Table 3: Evaluation Results on VL-RewardBench.

Models	Model Size	General	Hallucination	Reasoning	Overall Accuracy	Macro Average
Proprietary Models						
Claude-3.5-Sonnet(2024-06-22)	-	43.4	55.0	62.3	55.3	53.6
Gemini-1.5-Flash (2024-09-24)	-	47.8	59.6	58.4	57.6	55.3
GPT-4o(2024-08-06)	-	49.1	67.6	70.5	65.8	62.4
Gemini-1.5-Pro(2024-09-24)	-	50.8	72.5	64.2	67.2	62.5
Gemini-2.0-flash-exp(2024-12)	-	50.8	72.6	70.1	**68.8**	**64.5**
Open-Source Models						
Qwen2-VL-7B-Instruct	7B	31.6	19.1	51.1	28.3	33.9
MAmmoTH-VL-8B	8B	36.0	40.0	52.0	42.2	42.7
Qwen2.5-VL-7B-Instruct	7B	43.4	42.0	63.0	48.0	49.5
InternVL3-8B	8B	60.6	44.0	62.3	57.0	55.6
IXC-2.5-Reward-7B	7B	80.3	65.3	60.4	66.3	68.6
Qwen2-VL-72B-Instruct	72B	38.1	32.8	58.0	39.5	43.0
Molmo-72B-0924	72B	33.9	42.3	54.9	44.1	43.7
QVQ-72B-Preview	72B	41.8	46.2	51.2	46.4	46.4
Qwen2.5-VL-72B-Instruct	72B	47.8	46.8	63.5	51.6	52.7
InternVL3-78B	78B	67.8	52.5	64.5	63.3	61.6
Skywork-VL Reward (Ours)	7B	66.0	80.0	61.0	**73.1**	**69.0**

Table 4: Evaluation Results on RewardBench.

Models	Chat	Chat Hard	Safety	Reasoning	Avg Score
Language-Only Reward Models					
InternLM2-7B-Reward	99.2	69.5	87.2	94.5	87.6
Skywork-Reward-Llama3.1-8B	95.8	87.3	90.8	96.2	92.5
Skywork-Reward-Llama-3.1-8B-v0.2	94.7	88.4	92.7	96.7	93.1
QRM-Llama3.1-8B-v2	96.4	86.8	92.6	96.8	**93.1**
Multi-Modal Reward Models					
Qwen2-VL-7B-Instruct	65.1	50.9	55.8	68.3	60.0
InternVL3-8B	97.2	50.4	83.6	83.9	78.8
Qwen2.5-VL-7B-Instruct	94.3	63.8	84.1	86.2	82.1
IXC-2.5-Reward-7B	90.8	83.8	87.8	90.0	88.1
Skywork-VL Reward (Ours)	90.0	87.5	91.1	91.8	**90.1**

4.3 VL-RewardBench Evaluation

We begin by evaluating our model on VL-RewardBench, a comprehensive benchmark for multimodal reward modeling. Table 3 presents comparisons with both proprietary and open-source models across four key evaluation dimensions.

In the general category, skywork-VL Reward achieves a score of 66.0%, significantly outperforming even the strongest proprietary model, Gemini-2.0-flash-exp (50.8%). However, there is still a gap compared to IXC-2.5-Reward-7B (80.3%). In the hallucination category, our model achieves the best score (80.0%), which surpasses proprietary models (e.g., Gemini-2.0-flash-exp at 72.6%), and open-source models (e.g., IXC-2.5-Reward-7B 65.3%). This demonstrates our model's strong capability in mitigating factual inconsistencies. Our model also shows robust performance in the reasoning category. Despite having 10× fewer parameters, our model achieves a reasoning score of 61.0%, which is comparable to that of the much larger InternVL3-78B (64.5%), underscoring the efficiency and strong reasoning capabilities of our method.

Our model achieves an overall accuracy of 73.1% and a Macro Average of 69.0%, demonstrating superior performance across diverse task types and surpassing the best proprietary model, Gemini-2.0-flash-exp. The performance improvement is particularly notable given the compact size of our model (7B parameters), especially when compared to much larger models such as InternVL3-78B (78B parameters).

4.4 RewardBench Evaluation

Table 4 reports the results on RewardBench, a language-focused benchmark, comparing both language-only and multimodal RMs.

Table 5: Performance Evaluation on the MathVista Benchmark Using MPO with Different Reward Models.

Model	Base (Skywork R1V)	Qwen2.5-VL-7B-Instruct	InternVL3-8B	Ours
Performance (%)	69.2	71.2	71.8	73.5

Our model achieves an average score of 90.1 on RewardBench, the highest among all open-source multimodal RMs of comparable scale—2.0% higher than the second-best, IXC-2.5-Reward-7B. It is also competitive with top-performing language-specific RMs such as QRM-Llama3.1-8B-v2 (93.1%).

In the Chat Hard category, our model scores 87.5%, and it leads all comparable multimodal RMs in safety (91.1%) and reasoning (91.8%), outperforming the next best by 3.7%, 3.3%, and 1.8%, respectively. These results highlight the model's strong robustness and well-rounded performance across challenging and safety-critical tasks. The results validate that our model is not only effective in handling multimodal data, but also demonstrates strong capabilities on pure-text inputs.

5 Analysis

We present two illustrative examples that highlight the efficacy of our Skywork-VL Reward across distinct reasoning scenarios. For each example we supply a multimodal prompt together with a *good* and a *bad* answer. Skywork-VL Reward is queried once per answer and returns a scalar reward; higher values indicate stronger alignment with human preferences.

The first example (Figure 2) is a geometry problem that asks for the area of a circular sector. Although both candidate answers reach the same numeric conclusion, the good answer delivers a accurate derivation, whereas the bad answer meanders through repeated, self-correcting remarks. Skywork-VL Reward sharply favors the concise solution, confirming its sensitivity to reasoning quality rather than mere correctness.

The second example (Figure 3) involves reading a bar chart of extreme-poverty rates and identifying the country with the longest bar. Here the good answer states the correct label once and briefly cites the relevant percentages, while the bad answer redundantly lists the same numbers several times. Again, Skywork-VL Reward assigns a markedly higher reward to the compact explanation, demonstrating robustness to a very different visual domain.

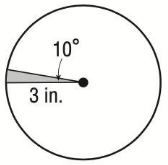
These two cases from two different domains show that Skywork-VL Reward consistently distinguishes well-structured reasoning from verbose or confused discourse. The clear reward gap between good and bad responses in both settings suggests that the model captures an alignment signal valuable for downstream reinforcement learning.

6 Skywork-VL Reward for Mixed Preference Optimization

Finally, we examine Skywork-VL Reward's usefulness as a reward signal for MPO [11], a recent strategy to further improve model alignment. MPO refers to optimizing a model's behavior using a mixture of preference signals rather than a single RM. This approach was proposed to stabilize and enhance training, especially for complex reasoning tasks that benefit from diverse feedback.

Furthermore, preference data constructed based on our Skywork-VL Reward proves highly effective for training Mixed Preference Optimization (MPO) [11], leading to significant improvements in multimodal reasoning capabilities. We utilize our Skywork-VL Reward to construct preference data, using it to fine-tune a strong VLM reasoner (Skywork R1V) via MPO and evaluate the impact on MathVista [41], a challenging benchmark for mathematical reasoning in visual contents. The results (Table 5) show that using Skywork-VL Reward as an additional reward yields a notable improvement: the model's MathVista score increases from 69.2% to 73.5%. This underscores Skywork-VL Reward's potential as a critical component in training the next generation of VLMs capable of long-sequence reasoning.

Source: OneVision



Hint:

Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.

Question:

What is the area of the sector given the following choices:

(A)$\pi/6$ (B) $\pi/4$ (C) $3\pi/5$ (D) $9\pi/10$

Good Answer :

Okay, let's see. The problem is asking for the area of a sector with a central angle of 10 degrees and a radius of 3 inches. Hmm, I remember that the formula for the area of a sector is something like $(\theta/360)$multiplied by π r squared. Let me make sure I recall that correctly.

First, the central angle θ here is 10 degrees. The radius r is given as 3 inches. So the area should be (10/360) times π times (3)^2. Let me write that out step by step to avoid mistakes. Calculating the fraction of the circle first: 10 divided by 360. That simplifies to 1/36, right? Because 10 divided by 10 is 1, and 360 divided by 10 is 36. So 10/360 = 1/36. Wait, is that correct? Let me check again. 10 divided by 360: yes, dividing numerator and denominator by 10 gives 1/36. Okay, that part seems right. Now, the area of the entire circle would be π r², which is $\pi * 3^2$. 3 squared is 9, so that's 9π. So multiplying the fraction by the area of the whole circle: (1/36) * 9 π. Let me compute that. 9 divided by 36 is 0.25, so that's 0.25 π? Wait, 9 divided by 36 is actually 1/4. Because 9 divided by 9 is 1, and 36 divided by 9 is 4. So yeah, 9/36 reduces to 1/4. Therefore, the area would be (1/4) π. Hmm, but looking at the answer choices, option B is π/4 .That seems to match. But let me double-check everything. Maybe I made a mistake in the formula. The formula for the area of a sector is indeed $\left(\frac{\theta}{360}\right) * \pi r^2$. So θ is 10 degrees, radius 3. So substituting in, it's$\left(\frac{\theta}{360}\right) * \pi * 3^2$.

Let's compute this step by step again. First, 10 divided by 360 is 1/36. Then 3 squared is 9. Multiply those together: (1/36)*9 = 9/36 = 1/4. So yes, 1/4 times π is π/4 . So that's answer choice B. Wait a second, but let me make sure that the angle is indeed in degrees and that I didn‘t mix up radians. The problem states the angle is 10 degrees, so that’s correct. If it were in radians, the formula would be different, but since it‘s specified in degrees, the formula I used is appropriate.

Alternatively, maybe I should think in terms of proportions. A full circle is 360 degrees, so a 10-degree sector is 10/360 of the circle. The area of the circle is $\pi r^2 = \pi * 9,$ so multiplying gives the sector area. Yep, that still gives π/4 .

Looking at the answer choices again: (A)π/6 (B) π/4 (C) 3π/5 (D) 9π/10 . So π/4 is option B. That seems right. Wait, but hold on, let me check if I messed up the multiplication somewhere. Let me recompute: 10 divided by 360 is 1/36. Then 1/36 multiplied by 9 π :9fivided by 36 is indeed 0.25, which is 1/4. So 1/4 π. So yes, that's correct. Alternatively, maybe the question is a trick question? Let me check the figure again. The problem mentions a figure with a central angle of 10 degrees and radius 3 inches. The sector is the shaded part, which is 10 degrees. So there's no trick here. Hmm, I think that's solid. So the answer should be B, π/4 . </think>

The area of the sector is calculated using the formula: \[\text{Area} = \frac{\theta}{360^\circ} \times \pi r^2 \] Substituting the given values (\(\theta = 10^\circ\), \(r = 3\) in): \[\text{Area} = \frac{10}{360} \times \pi \times 3^2 = \frac{1}{36} \times 9\pi = \frac{\pi}{4} \] Thus, the correct answer is **B**.

Answer: B --- **Final Answer** \boxed{B}

Reward Score: 5.86 🤩

Bad Answer:

Okay, let‘s see. Okay, let’s tackle this problem step by step. The question is asking for the area of a sector in a circle, and the options are given. First, I need to recall the formula for the area of a sector. I remember that the area of a sector is a fraction of the area of the entire circle. The formula is something like (θ /360)multiplied by πr², where θ is the central angle in degrees and r is the radius of the circle. Let me confirm that.

So, the problem mentions a sector with a central angle of 10 degrees and a radius of 3 inches. **Wait,** the image shows a central angle of 10 degrees and the radius is labeled as 3 inches. That's the key information here. The fraction of the circle that the sector represents is θ divided by 360 degrees. So that would be 10/360. Simplifying that fraction: both numerator and denominator are divisible by 10, so that becomes 1/36. **Wait,** 10 divided by 360 is indeed 1/36. Hmm, but let me check the arithmetic again. 10 divided by 360 is the same as 1/36. Yes, that's correct.

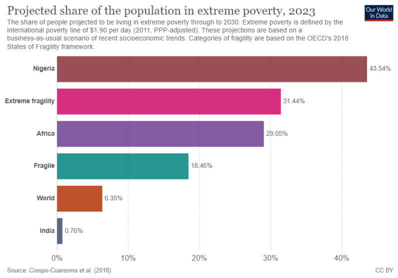
Therefore, the area of the sector should be (1/36) multiplied by the area of the whole circle. The area of the whole circle is 9 $\pi,$ so multiplying those together: (1/36)*9 $\pi.$ Let me compute that. 9 divided by 36 is 0.25, so 0.25 π? **Wait,** but 9 divided by 36 is actually 1/4. **Wait,** hold on. Let me recalculate: **Wait,** 9 divided by 36 is indeed 1/4? **Wait,** no. 9 divided by 36 is 0.25, which is 1/4. **Wait,** but 10 degrees over 360 is 1/36, so 1/36 times 9 $\pi.$ Let me do the math again step by step. First, the area of the sector formula is $\left(\frac{\theta}{360}\right) * \pi r^2$. Plugging in the numbers: θ = 10, so 10/360 = 1/36,r = 3 inches, so r² = 9. Therefore, the area is $\frac{1}{36} * \pi * 9 = \frac{1}{4} \pi.$ **Wait,** so that would be $\pi/4$? But hold on, π/4 is one of the options, option B. However, I need to check if I made a mistake here because sometimes there might be a different approach. **Wait,** 10 degrees is the central angle. The formula is$\left(\frac{\theta}{360}\right) * \pi r^2$. So 10 divided by 360 is indeed 1/36. Multiply that by $\pi * 3^2$which is $9\pi.$ So$\frac{1}{36} * \pi * 9 = \frac{1}{4} \pi.$ So $\pi/4.$ That's correct. But **wait,** let me check if the radius is indeed 3 inches. The problem states "3 in." which is the radius? The image might have the radius labeled as 3 inches. So yes, radius is 3. Therefore, the calculation seems right. Therefore, the area is $\pi/4,$ which is option B. **But wait**, let me check the options again to confirm. The options are (A)π/6 (B) π/4 (C) 3π/5 (D) 9π/10 . Yes, B is π/4. So that's the answer. **Wait a second**, but let me think again. Let me recast the problem in another way. Suppose the sector area formula is (1/2)*r²* $\theta,$ but in radians. **Wait,** that's another formula. So maybe if the angle was in radians, but here it's given in degrees, so that formula would require converting θ to radians first. **Wait,** let me clarify the formulas. The area of a sector can be calculated in two ways depending on whether the central angle is in degrees or radians. Therefore, the answer should be π/4, which is option B. Hmm, but let me check once more. **Wait,** let me compute the numerical value to see if it matches any of the options. Let's see: Option A: $\pi/6 \approx 0.5236$ Option B: $\pi/4 \approx 0.7854$ Option C:$3\pi/5 \approx 1.885$ Option D:$9\pi/10 \approx 2.827$. The area of the sector we calculated is $\pi/4 \approx 0.7854,$ which matches option B. So that's correct.

Thus, the area of the sector is \(\boxed{B}\).

Reward Score: -15.5 😢 **Explanation: The word ‘wait’ appears multiple times and is redundant.**

Figure 2: Evaluating Skywork R1V on Mathematical Problems.

Source: ChartQA



Hint:
Please answer the question and provide the final answer at the end.

Question:

Which country has longest bar?

Good Answer :

Okay, let me look at the data provided. The question is asking which country has the longest bar in the bar chart. The bars represent the projected share of the population in extreme poverty for 2023.

First, I need to check the values for each country. The bars are color-coded, so each country's value is along a horizontal bar. The values are as follows:

- Nigeria: 43.54% - Extreme fragility: 31.44% - Africa: 29.05% - Fragile: 18.46% - World: 6.35% - India: 0.76%

The length of each bar corresponds to these percentages. The longer the bar, the higher the percentage. Comparing all the values, Nigeria has the highest percentage at 43.54%, which would make its bar the longest. The next is Extreme fragility at 31.44%, then Africa at 29.05%, and so on.

So, the country with the longest bar is Nigeria because it has the highest value. The answer should be straightforward based on the given data.

The country with the longest bar is Nigeria, as it has the highest projected share of the population in extreme poverty at 43.54% in 2023.

Answer: Nigeria.

Reward Score: 7.53 🤩

Bad Answer:

Okay, let's see. The user is asking which country has the longest bar in the given bar chart. The chart shows the projected share of the population in extreme poverty for 2023 across different countries and regions. The data is presented as percentages for each category.

First, I need to look at the values for each country. The bars are horizontal, so the length of each bar corresponds to the percentage value. The countries listed are Nigeria, Extreme Fragility, Africa, Fragile, World, and India.

The values are as follows: - Nigeria: 43.54% - Extreme Fragility: 31.44% - Africa: 29.05% - Fragile: 18.46% - World: 6.35% - India: 0.76%

The longest bar would be the one with the highest percentage value. Comparing all the values, Nigeria has the highest at 43.54%, so that's the longest bar. The next one is Extreme Fragility at 31.44%, then Africa at 29.05%, and so on. The user might be confused if they think "longest" refers to physical length, but in a bar chart, the length is proportional to the value. So Nigeria's bar is the longest here.

The country with the longest bar is Nigeria, with a projected share of 43.54% of its population in extreme poverty in 2023. This is the highest value among all the listed categories, making its bar the longest in the chart.

Key Values for Clarity: - Nigeria: 43.54% - Extreme Fragility: 31.44% - Africa: 29.05% - Fragile: 18.46% - World: 6.35% - India: 0.76%

The length of the bars directly corresponds to these percentage values, so Nigeria's bar is the longest.

Reward Score: -10.36 😢 **Explanation: Repeating the answer multiple times makes it verbose.**

Figure 3: Evaluating Skywork R1V on Chart Problems.

7 Conclusion

In this work, we introduced Skywork-VL Reward, a multimodal reward model tailored for vision-language models (VLMs), addressing the critical need for reliable and general-purpose evaluators aligned with human judgment in multimodal understanding and reasoning tasks. Through the construction of a large-scale, meticulously curated preference dataset encompassing diverse tasks and scenarios, coupled with a two-stage training paradigm, our model establishes a comprehensive framework for evaluating responses generated by both standard VLMs and VLM reasoners. Empirical results demonstrate Skywork-VL Reward's state-of-the-art performance on the VL-RewardBench benchmark and its competitive capabilities on text-only RewardBench, validating its versatility as a reasoning-aware evaluator. Furthermore, integrating Skywork-VL Reward into MPO significantly enhances multimodal reasoning ability of VLMs, highlighting its practical value.

References

[1] Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*, 2024. 1

[2] Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. Regularizing hidden states enables learning generalizable reward model for llms. *arXiv preprint arXiv:2406.10216*, 2024. 2

[3] Yi Peng, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, et al. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*, 2025. 1, 2

[4] Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms, 2024. 1, 3, 6

[5] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark, 2025.

[6] Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025. 1, 2, 3, 6

[7] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024. 1, 2, 3, 6

[8] Yuu Jinnai, Tetsuro Morimura, Kaito Ariu, and Kenshi Abe. Regularized best-of-n sampling with minimum bayes risk objective for language model alignment, 2025. 1

[9] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 5, 6

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 2

[11] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization, 2025. 2, 8

[12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[13] Gemma Team. Gemma. 2024.

[14] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023. 2

[15] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 2

[16] Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. In *ACL*, 2024. 2

[17] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 2

[18] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[19] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. 2

[20] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[21] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. 2

[22] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. 2

[23] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 2, 5

[24] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023. 2

[25] Shijian Deng, Wentian Zhao, Yu-Jhe Li, Kun Wan, Daniel Miranda, Ajinkya Kale, and Yapeng Tian. Efficient self-improvement in multimodal large language models: A model-level judge-free approach, 2024. 2

[26] Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension, 2024. 2

[27] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models, 2025. 2, 3

[28] OpenAI. Gpt-4 technical report, 2023. 2, 6

[29] Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness, 2024. 3

[30] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 5

[31] Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. Vlrewardbench: A challenging benchmark for vision-language generative reward models, 2024. 6

[32] Anthropic. Claude-3.5, 2024. 6

[33] Gemini Team. Gemini: A family of highly capable multimodal models, 2024. 6

[34] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6

[35] Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale, 2024. 6

[36] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 6

[37] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. 6

[38] Qwen Team. Qvq: To see the world with wisdom, December 2024. 6

[39] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, et al. Internlm2 technical report, 2024. 6

[40] Nicolai Dorka. Quantile regression for distributional reward models in rlhf. *arXiv preprint arXiv:2409.10164*, 2024. 6

[41] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 8