InternLM-XComposer2.5-Reward: A Simple Yet Effective Multi-Modal Reward Model

Anonymous ACL submission

Abstract

Despite the promising performance of Large Vision Language Models (LVLMs) in visual understanding, they occasionally generate incorrect outputs. While reward models (RMs) with reinforcement learning or test-time scaling offer the potential for improving generation quality, a critical gap remains: publicly available multi-modal RMs for LVLMs are scarce, and the implementation details of proprietary models are often unclear. We bridge this gap with InternLM-XComposer2.5-Reward (IXC-2.5-Reward), a simple yet effective multi-modal reward model that aligns LVLMs with human preferences. To ensure the robustness and versatility of IXC-2.5-Reward, we set up a high-quality multi-modal preference corpus spanning text, image, and video inputs across diverse domains, such as instruction following, general understanding, text-rich documents, mathematical reasoning, and video understanding. IXC-2.5-Reward achieves excellent results on the latest multi-modal reward model benchmark and shows competitive performance on text-only reward model benchmarks. We further demonstrate three key applications of IXC-2.5-Reward: (1) Providing a supervisory signal for RL training. We integrate IXC-2.5-Reward with Proximal Policy Optimization (PPO) yields IXC-2.5-Chat, which shows consistent improvements in instruction following and multi-modal open-ended dialogue; (2) Selecting the best response from candidate responses for test-time scaling; and (3) Filtering outlier or noisy samples from existing image and video instruction tuning training data.

1 Introduction

"*If you don't know where you are going, you'll end up some place else.*"

— Yogi Berra

Reward Models (RMs) (Cai et al., 2024; Zhu et al., 2023; Liu et al., 2024a; Wang et al., 2024f,b; Yuan et al., 2024a; Lou et al., 2024; Yang et al., 2024b; Yuan et al., 2024b; Shiwen et al., 2024; Wang et al., 2024e) provide the crucial direction guidance about how well an AI model's outputs align with human preference, and benefit Large Language Models (LLMs) in training and inference. During training, RMs are often used with reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022b; Schulman et al., 2017; Rafailov et al., 2024) to penalize undesirable model behaviors and encourage outputs that align with human values. At inference, RMs facilitate test-time scaling strategies (Snell et al., 2024; Gulcehre et al., 2023), such as selecting the best response from candidate outputs or providing step-by-step critiques for complex reasoning tasks (Zelikman et al., 2022; Hosseini et al., 2024).

Despite their crucial role in both training and inference, multi-modal RMs for Large Vision Language Models (LVLMs) remain notably underexplored compared to language-only RMs for LLMs. Because current preference data is predominantly text-based and skewed toward specific domains (e.g., safety), data scarcity poses a significant challenge to training multi-modal RMs for diverse modalities such as images, videos, and text. Consequently, existing multi-modal RMs (Wang et al., 2024a; Xiyao et al., 2024) are largely constrained to narrow domains (e.g., mitigating hallucinations) or rely on prompting LVLMs with evaluation prompts, effectively functioning as generative RMs (Xiong et al., 2024). The limitation of multi-modal RMs subsequently constrains the capabilities of open-source LVLMs such as instruction following and safety-should-refuse, thereby hampering user interaction experience in multi-modal chat scenarios.

The growing community interest in RLHF and test-time scaling highlights the need for multi-modal RMs, which motivates us to present InternLM-XComposer2.5-Reward (IXC-2.5-Reward). Instead of directly transferring uni-

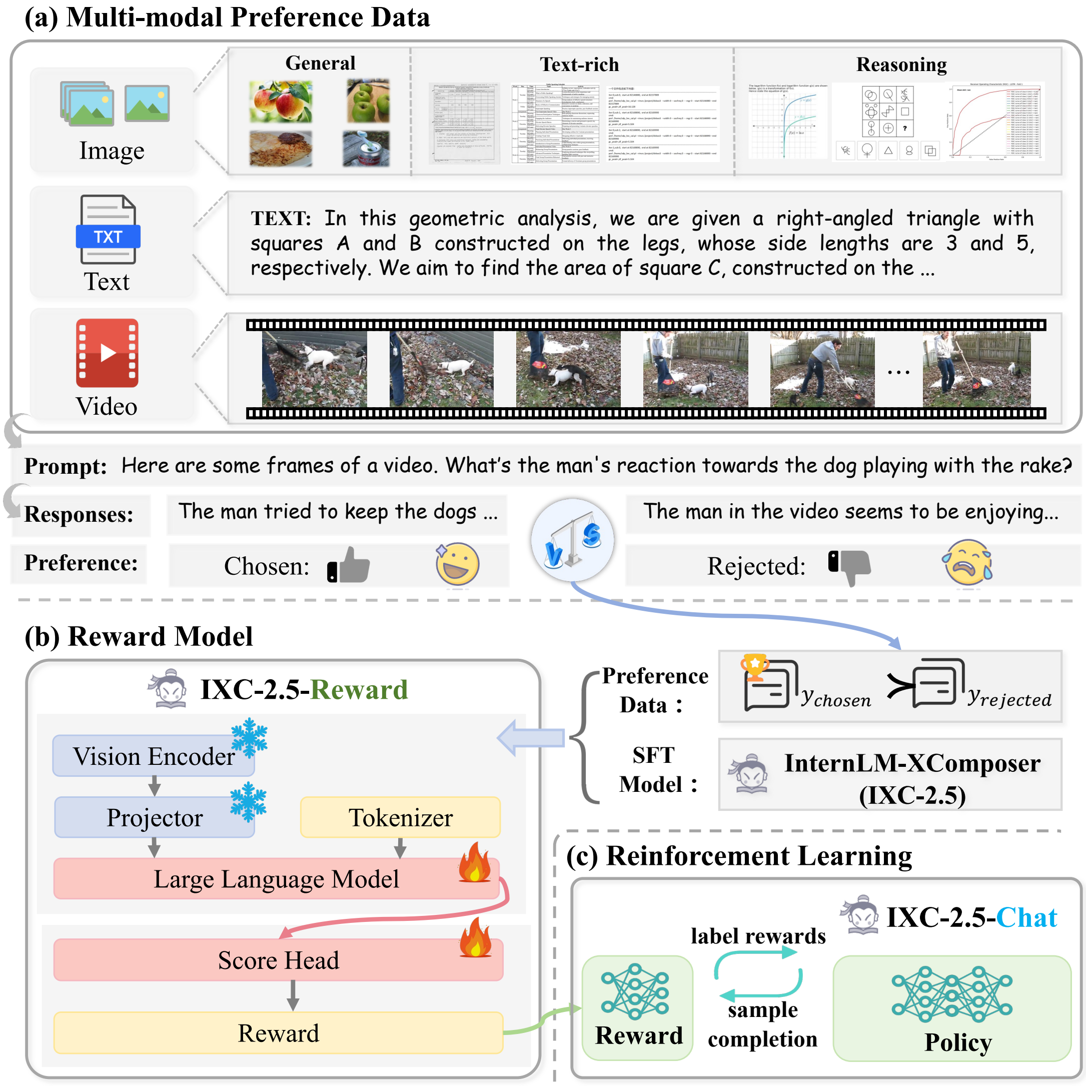

Figure 1: **(a)** To train the IXC-2.5-Reward, we construct a multi-modal preference dataset spanning diverse domains (e.g., natural scenes, text-rich, reasoning) and modalities (image, text, video). **(b)** The framework of IXC-2.5-Reward. **(c)** The IXC-2.5-Reward guides policy training for IXC-2.5-Chat via reinforcement learning.

modal (text) reward models (RMs) to the vision modality, we augment the existing LVLM (InternLM-XComposer2.5) with an additional scoring head to predict reward scores. An effective multi-modal RM should ideally possess two key properties: (1) the ability to predict reward scores for both image, video, and textual inputs and (2) the capacity to generalize across diverse domains, such as instruction following, knowledge, text-rich images (e.g., documents), reasoning tasks, etc. To this end, we develop a pipeline (Fig. 1(a)) to construct multi-modal preference data, and also incorporate existing high-quality datasets. This pipeline selects prompts across diverse domains for text, image, and video inputs, generates corresponding responses, and then uses GPT-4o (Hurst et al., 2024) or verifier (Lambert et al., 2024a) to perform preference judgments. Trained on our preference data, IXC-2.5-Reward effectively evaluates both visual (image and video) and textual inputs (Fig. 1 (b)).

IXC-2.5-Reward achieves best performance on multi-modal VL-RewardBench (Li et al., 2024b) (70.0%) that beat all previous generative RMs including Gemini-1.5-Pro (62.5%) and GPT-4o (62.4%). Even on uni-modal (text) RM benchmarks, IXC-2.5-Reward also demonstrates good results, with an average score of 88.6% on Reward-Bench (Lambert et al., 2024b) and 68.8% on RM-Bench (Liu et al., 2024b).

We further demonstrate the effectiveness of IXC-2.5-Reward in the following three aspects:

(1) IXC-2.5-Reward for RL training. We train a chat model (IXC-2.5-Chat) using the on-policy Proximal Policy Optimization (PPO) algorithm with IXC-2.5-Reward to enhance its ability to follow instructions and provide a better user experience in multi-modal conversations. Our results show clear improvements of IXC-2.5-Chat on multi-modal instruction following and in-the-wild chatting benchmarks, which validate the effectiveness of IXC-2.5-Reward for providing the reward signal during RL training.

(2) IXC-2.5-Reward for Test-Time Scaling. Using best-of-N sampling with IXC-2.5-Reward leads to additional performance gains compared to the RL-trained IXC-2.5-Chat, confirming IXC-2.5-Reward's effectiveness in selecting good responses from candidate responses.

(3) IXC-2.5-Reward for Data Cleaning. We observe a strong correlation between low IXC-2.5-Reward scores and problematic samples, such as those exhibiting hallucinations or mismatched image/video and question/answer content. This suggests that IXC-2.5-Reward can effectively clean LVLM pre-training and post-training data.

2 Related Work

Reward Model in Large Language Models. Reward models (RMs) are crucial for both Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022b) and Test-time Scaling Laws (Snell et al., 2024; Hosseini et al., 2024). RMs have different implementation forms, such as **(1)** discriminative RM (Cai et al., 2024; Zhu et al., 2023; Liu et al., 2024a; Wang et al., 2024f,b; Yuan et al., 2024a; Lou et al., 2024; Yang et al., 2024b), usually a sequence classifier that classifies input sequences into different categories, such as binary classification ("good" or "bad,") or on a more granular scale (Wang et al., 2024f,b). **(2)** generative RM (Kim et al., 2023; Yuan et al., 2024b; Shiwen et al., 2024; Wang et al., 2024e) that are prompted to generate the feedback in the form of text, often a critique or explanation of why a certain output is good or bad. **(3)** implicit RMs (Ivison et al., 2023; Lambert et al., 2024a) that are models optimized using DPO (Rafailov et al., 2024) that the predicted log probabilities are interpreted as implicit reward signal. Besides, RMs can also be divided into Outcome RMs (ORMs) (Cobbe et al., 2021) and Process RMs (PRMs) (Uesato et al., 2022; Lightman et al., 2023; Setlur et al., 2024). Our IXC-2.5-Reward belongs to the discriminative RM and ORM.

Reward Model in Large Vision-Language Models. Previous RMs for LVLMs (Wang et al., 2024a; Xiong et al., 2024; Xiyao et al., 2024) are limited to specific domains (e.g., reducing hallucination) or developed using relatively weak base models, which makes the implemented models significantly inferior to LLM RMs. The lack of effective multi-modal RMs has created a bottleneck in vision RLHF, forcing researchers to merely use the variants of the off-poly DPO algorithm (Rafailov et al., 2024). Previous work using open-source LVLMs as generative RMs (Yu et al., 2024c; Ouali et al., 2025; Xiyao et al., 2024), injection of hallucinations with data augmentation techniques (Deng et al., 2024b; Favero et al., 2024; Zhou et al., 2024b; Zhu et al., 2024; Pi et al., 2025; Jiang et al., 2024; Deng et al., 2024a) and rule-based selection (Cao et al., 2024; Liu et al., 2024f) for DPO data selection, which potentially compromise performance compared to the on-policy RL solutions like PPO (Schulman et al., 2017). Moreover, lacking multi-modal RMs has also led to the reliance on human annotation (Sun et al., 2023; Yu et al., 2024a) or the use of proprietary models (Zhang et al., 2024a; Zhao et al., 2023) like GPT4 as generative RMs for DPO pair selection, which is expensive and unsustainable for large-scale applications. Although open-source RMs for LVLMs have lagged behind their LLM counterparts, the growing community interest highlights the need for multi-modal RMs, which motivates our work. In this work, we demonstrate that IXC-2.5-Reward is capable of combining with the PPO training and for DPO data selection at a low cost.

Reward Model Evaluations. The development of evaluation benchmarks is essential for improving RMs. Several comprehensive benchmarks have been proposed for evaluating RMs of LLMs, such as general abilities (Lambert et al., 2024b; Zhou et al., 2024a; Liu et al., 2024b), multilingual (Son et al., 2024; Gureja et al., 2024), RAG (Jin et al., 2024), and mathematical process reward (Zheng et al., 2024). The limited availability of multi-modal RMs has hampered the development of evaluation benchmarks, with existing benchmark (Li et al., 2024b) focusing solely on generative RMs and lacking the evaluation of process supervision. However, given the critical importance of RMs, we expect significant progress in multi-modal RM benchmarking in the future.

3 IXC2.5-Reward

Data Preparation. Reward models are trained using pairwise preference annotations (e.g., prompts

Table 1: Overview of existing preference datasets used in IXC-2.5-Reward.

Category	Dataset
	Text
IF General	Tulu-3-IF-augmented-on-policy-8b (Lambert et al., 2024a) UltraFeedback (Cui et al., 2024)
Safety	hhh alignment (Askell et al., 2021), PKU-Safe (Dai et al., 2024a) SHP (Ethayarajh et al., 2022), Anthropic-hhrlhf (Bai et al., 2022a)
	Image
Chat	WildVision-Battle (Lu et al., 2024c)
General	LLaVA-Critic (Xiong et al., 2024), VL-Feedback (Li et al., 2024c), RLAIF-V (Yu et al., 2024b) MIA-DPO (Liu et al., 2024e)

Table 2: Overview of the source of newly collected data used in IXC-2.5-Reward.

Category	Dataset
	Image
IF General	in-house (will release) KVQA (Shah et al., 2019), A-OKVQA (Schwenk et al., 2022), PMC-VQA (Zhang et al., 2023)
Text-Rich	AI2D (Kembhavi et al., 2016), IconQA (Lu et al., 2021), TQA (Kembhavi et al., 2017) ChartQA (Masry et al., 2022), DVQA (Kafle et al., 2018), ScienceQA (Lu et al., 2022a)
Reasoning	GeoQA (Chen et al., 2021), CLEVR-Math (Lindström and Abraham, 2022) Super-CLEVR (Li et al., 2023), TabMWP (Lu et al., 2022b)
	Video
General	TrafficQA (Xu et al., 2021), FunQA (Xie et al., 2024), MiraData (Ju et al., 2024)

x with chosen responses y_c and rejected responses y_r) that reflect human preferences. While existing public preference data is primarily textual, with limited image and scarce video examples, we train IXC-2.5-Reward using both open-source data and a newly collected dataset to ensure broader domain coverage.

Tab. 1 lists the open-source pairwise data used in IXC-2.5-Reward, primarily focused on instruction following, safety, and general knowledge. Tab. 2 details the source of our newly collected data, which is initially the supervised fine-tuning (SFT) data consisting of prompts x and corresponding chosen responses y_c across diverse domains: text-rich document understanding, math reasoning, and video understanding. We also collect some in-house data about the instruction following, which will be released in the future. To obtain rejected responses y_r, we prompt the SFT model, InternLM-XComposer-2.5 (IXC-2.5) (Zhang et al., 2024c) to generate multiple outputs for each prompt and then employ distinct selection criteria. For general and text-rich data, we use GPT-4o (Hurst et al., 2024) with pairwise evaluation prompts to determine the rejected response that was evaluated worse than the SFT ground-truth answer. For math reasoning and instruction following data, we build verifier functions (Lambert et al., 2024a) that compare generated responses against ground-truth solutions to label the chosen and rejected data. Our newly collected data complements existing open-source data, creating a comprehensive, high-quality multi-modal preference dataset.

Model Architecture. Our reward model InternLM-XComposer 2.5-Reward (IXC-2.5-Reward) is built upon the SFT model (IXC-2.5) (Zhang et al., 2024d). As shown in Fig. 1 (b), we use the pre-trained weights of IXC-2.5-Chat for most of the parts, such as the visual encoder and the MLP projector, which has aligned the image and video data with text modalities. Thus, the IXC-2.5-Reward is merely required to train preference data to predict the reward score and avoid using other pre-training data for modality alignment.

We replace the final linear layer of IXC-2.5 with a score head f for IXC-2.5-Reward that predicts the reward score. Given the input prompt x and the response y, the score head f transforms the averaged hidden state features of all tokens into a binary scalar $r(x, y)$. This scalar value $r(x, y)$ serves as the predicted reward score for the inputs.

Loss Function. Our reward model is trained via the following loss function:

$$\mathcal{L}_{\text{RM}} = -E(\log(\sigma(r(x, y_w) - r(x, y_l)))), \quad (1)$$

where $r(x, y_w)$ and $r(x, y_l)$ denotes to the reward score assigned to the prompt x with the chosen data y_w and rejected data y_l, respectively.

Training Strategy. As shown in Fig. 1 (b), we froze the model's vision encoder and projector that are initialized from IXC-2.5 (Zhang et al., 2024c), training only the LLM (InternLM (Zhang et al., 2024c)) and the score head. Other components of IXC-2.5, such as the dynamic image partitioning mechanism for high-resolution inputs, remained unchanged.

Length Constraints. We remove data pairs where the length of the chosen response y_w is significantly longer than the length of the rejected response y_l. This helps prevent the reward model from learning to associate length with quality. Notably, we found that the vulnerability of LLM-based evaluation to length bias, a known issue in LLMs (Dubois et al., 2024), has also significant implications for LVLMs. Specifically, open-ended Visual Question Answering (VQA) benchmarks that employ LVLMs (e.g., GPT-4o) as judges are susceptible to inflated scores from overly long responses. Consequently, removing the length constraint on the reward model resulted in improved PPO policy performance. A detailed analysis is provided in Tab. 7.

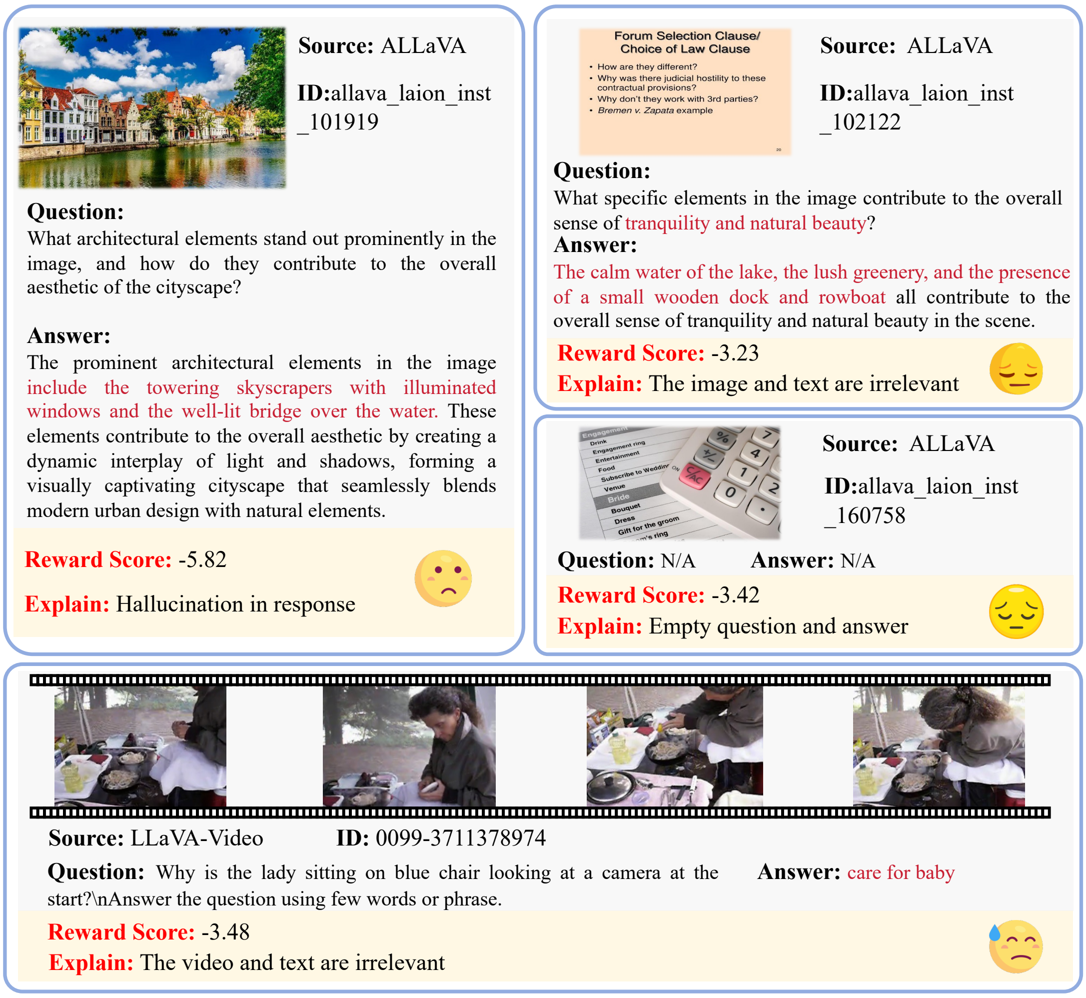

Figure 2: **Using IXC-2.5-Reward for Data Cleaning.** We visualize the outlier and noisy examples detected by IXC-2.5-Reward with low reward scores from existing image and video instruction-tuning datasets, such as ALLaVA (Chen et al., 2024a) and LLaVA-Video-178K (Zhang et al., 2024e). The "Explain" refers to explanations of error causes as identified by human experts, rather than outputs generated by the reward model.

4 The Applications of IXC-2.5-Reward

In this section, we further validate three applications of IXC-2.5-Reward for (1) RL training (Sec. 4.1), (2) test-time scaling (Sec. 4.2), and (3) data cleaning (Sec. 4.3).

4.1 IXC-2.5-Reward for RL training

Having the reward model IXC-2.5-Reward enables the application of on-policy reinforcement learning algorithms (e.g., PPO (Schulman et al., 2017), RLOO (Ahmadian et al., 2024), GRPO (Shao et al., 2024)) to optimize LVLM performance towards desired human preferences directly. Using the PPO (Schulman et al., 2017) algorithm, we subsequently train the policy model (**IXC-2.5-Chat**, π_θ) to maximize expected rewards from our reward model (IXC-2.5-Reward) while staying close to the reference model (IXC-2.5, π_{ref}) for stability. A critic model V, initialized from IXC-2.5-Reward, is trained alongside π_θ to reduce the variance of policy updates.

Data Prepration. Similar to findings in (Hou et al., 2024), we found that average reward scores differ across task domains (e.g., general, text-rich, reasoning). This work focuses on improving the policy model's instruction following and open-ended chat abilities, which are critical for real-world applications such as stream chatting and human-AI interaction (Zhang et al., 2024b). Simultaneously, we ensure that performance in other domains (e.g., text-rich, reasoning) is not degraded relative to the SFT model IXC-2.5. Using our multi-modal preference data (which trains IXC-2.5-Reward), we curate a prompt set that prioritizes general chat and instruction following, while ensuring diversity through the inclusion of text-rich documents, math reasoning, and video understanding.

PPO. The PPO training begins by sampling a prompt from our prompt set. Then, the policy θ_π model generates responses, and the reward model

Table 3: **Evaluation results on VLRewardBench (Li et al., 2024b).** The best and second-best results for proprietary models and open-source models are highlighted in **bold** and underlined, respectively.

Models	#Param	General	Hallucination	Reasoning	Overall Acc	Macro Acc
Proprietary Models						
Gemini-1.5-Flash (2024-09-24) (Team, 2024a)	-	47.8	59.6	58.4	57.6	55.3
Gemini-1.5-Pro (2024-09-24) (Team, 2024a)	-	**50.8**	**72.5**	64.2	**67.2**	**62.5**
Claude-3.5-Sonnet (2024-06-22) (Anthropic, 2024)	-	43.4	55.0	62.3	55.3	53.6
GPT-4o-mini (2024-07-18) (AI, 2024)	-	41.7	34.5	58.2	41.5	44.8
GPT-4o (2024-08-06) (AI, 2024)	-	49.1	67.6	**70.5**	65.8	62.4
Open-Source Models						
LLaVA-OneVision-7B-ov (Li et al., 2024a)	7B	32.2	20.1	57.1	29.6	36.5
Qwen2-VL-7B (Wang et al., 2024d)	7B	31.6	19.1	51.1	28.3	33.9
Molmo-7B (Deitke et al., 2024)	7B	31.1	31.8	56.2	37.5	39.7
InternVL2-8B (Team, 2024c)	8B	35.6	41.1	59.0	44.5	45.2
LLaVA-Critic-8B (Xiong et al., 2024)	8B	54.6	38.3	59.1	41.2	44.0
Llama-3.2-11B (Team, 2024b)	11B	33.3	38.4	56.6	42.9	42.8
Pixtral-12B (Agrawal et al., 2024)	12B	35.6	25.9	59.9	35.8	40.4
Molmo-72B (Deitke et al., 2024)	72B	33.9	42.3	54.9	44.1	43.7
Qwen2-VL-72B (Wang et al., 2024d)	72B	38.1	32.8	58.0	39.5	43.0
NVLM-D-72B (Dai et al., 2024b)	72B	38.9	31.6	62.0	40.1	44.1
Llama-3.2-90B (Team, 2024b)	90B	42.6	57.3	61.7	56.2	53.9
IXC-2.5-Reward (Ours)	7B	**84.7**	**62.5**	**62.9**	**65.8**	**70.0**

computes the reward score r_t at each state s_t at the time-step t. Given the reward score r_t and and the critic model V, we compute the temporal difference error δ_t, the Generalized Advantage Estimation (GAE) (Schulman et al., 2018) A_t, and the Returns R_t as:

$$\begin{aligned} \delta_t &= r_t + \gamma \cdot V(s_{t+1}) - V(s_t), \\ A_t &= \delta_t + \gamma \cdot \beta \cdot A_{t+1}, \\ R_t &= A_t + V(s_t), \end{aligned} \tag{2}$$

where γ is a discount factor that determines how much future rewards are valued compared to immediate rewards, and β is the parameter that controls the trade-off between bias and variance in the advantage estimation. The advantage A refers to how much better the policy model did than expected, and the returns R is the cumulative reward.

Based on the advantage A, we compute the policy gradient loss $\mathcal{L}_{\text{PG}}$ to update the policy model π_θ:

$$\mathcal{L}_{\text{PG}} = \min(\frac{\pi_\theta}{\pi_{\text{ref}}} \cdot A, \text{clip}(\frac{\pi_\theta}{\pi_{\text{ref}}}, 1.0-\epsilon, 1.0+\epsilon) \cdot A), \tag{3}$$

where $\frac{\pi_\theta}{\pi_{\text{ref}}}$ is the log of the probability ratio between the policy model π_θ and the reference model π_{ref}, and ϵ is a hyper-parameter that controls the clipped ratio.

We further update the critic model via the Mean Squared Error (MSE) loss to minimize the difference between the predicted value of a state $V(s_t)$ and the actual return R_t obtained from state t:

$$\mathcal{L}_{\text{critic}} = \sum_t \text{MSE}(V(s_t), R_t). \tag{4}$$

In summary, with the help of IXC-2.5-Reward and PPO, we train the IXC-2.5-Chat to generate responses that improve the quality of multi-modal chat and follow user instructions. The quality of IXC-2.5-Chat also demonstrates the quality of IXC-2.5-Reward that provides the reward scores.

4.2 IXC-2.5-Reward for Test-Time Scaling

We further demonstrate that IXC-2.5-Reward is essential for scaling the inference-time capabilities of LVLMs. We select the Best-of-N (BoN) sampling technique that improves the quality of generated text by using the reward model. Specifically, the IXC-2.5-Chat model generates multiple (N) different text outputs with different random seeds for a given prompt. Subsequently, the reward model IXC-2.5-Reward scores each of these N outputs, and the output with the highest score from the reward model is selected as the final output.

4.3 IXC-2.5-Reward for Data Cleaning

Garbage in, garbage out. Problematic samples in instruction tuning datasets negatively impact LVLM training. While existing methods (Chen et al., 2024c) employ classifiers like CLIP (Radford et al., 2021) for filtering, these approaches have limitations, particularly with long-context inputs (Zhang et al., 2025a), high-resolution images, or videos. As shown in Fig. 2, we observe a strong correlation between low IXC-2.5-Reward scores and problematic samples, including hallucinations, empty answers, and irrelevant image/video-text pairings. Therefore, IXC-2.5-Reward effectively cleans both pre-training and post-training data for LVLMs.

5 Experiments

5.1 Evaluation Results of IXC-2.5-Reward

Benchmarks. To evaluate IXC-2.5-Reward, we use diverse reward model benchmarks: **(1)** VL-RewardBench (Li et al., 2024b), encompassing 1250 multi-modal problems addressing general understanding, hallucination, and reasoning challenges; (2) Reward-Bench (Lambert et al., 2024b), with 2985 language-only problems including chat, chat hard, safety and reasoning; and **(3)** RM-Bench (Liu et al., 2024b), comprising 1237 language-only problems across chat, math, code, and safety. RM-Bench defines three tracks (easy, normal, hard) that evaluate the sensitivity of reward models to subtle content variations and style biases. While Reward-Bench and RM-Bench are designed for reward models of language-only LLMs, we evaluate IXC-2.5-Reward on these benchmarks to demonstrate that our multi-modal reward model maintains strong language capabilities despite also processing image and video inputs.

5.1.1 Results on VL-RewardBench

Main Results. Tab. 3 presents the evaluation results of various multi-modal RMs on the VL-RewardBench (Li et al., 2024b). Unlike previous multi-modal generative reward models, our IXC-2.5-Reward is a discriminative model that predicts a scalar reward. Our proposed IXC-2.5-Reward model, despite being an open-source 7B parameter model, outperforms all other open-source models. Notably, IXC-2.5-Reward achieves the highest overall accuracy (65.8%) among open-source models and the highest Macro Accuracy (70.0%) among all models, indicating its superior performance in handling diverse tasks within the VL-RewardBench.

Strong Performance on General Problems. The results in Table 3 reveal that IXC-2.5-Reward achieves a significantly higher accuracy (84.7%) on general problems compared to other generative RMs. We found the reason is attributed to these problems presenting a considerable challenge, often leading to tied judgments in previous LVLMs, whereas IXC-2.5-Reward demonstrates a greater ability to make correct classifications with different scalar scores.

5.1.2 Results on Reward Bench and RM-Bench

Main Results. We argue that multi-modal RMs should preserve strong language processing abilities despite the incorporation of image and video data during training. Consequently, we evaluate the performance of multi-modal reward models, including IXC-2.5-Reward, on Reward Bench (Tab. 4) and RM-Bench (Tab. 5). The results demonstrate that IXC-2.5-Reward achieves considerable performance and surpasses other multi-modal models on this benchmark.

Sensitivity to Content and Style. Consistent with findings in (Liu et al., 2024b), IXC-2.5-Reward demonstrates sensitivity to subtle content variations and style biases, an issue often overlooked in multi-modal research. We believe further research is needed to enhance the robustness of multi-modal reward models.

5.2 Evaluation Results of IXC-2.5-Chat

Benchmarks. We select four representative benchmarks for evaluating the instruction following and in-the-wild chatting abilities of LVLMs. (1) The WildVision bench (Lu et al., 2024c) uses prompts collected from user submissions, reflecting real-world multimodal interactions. (2) MIA-bench (Qian et al., 2024) that is specially designed to evaluate instruction following. (3) MM-MT (Agrawal et al., 2024) which is an instruction-following benchmark for multi-modal models, exhibits a strong correlation with LMSys-Vision ELO ratings (Chiang et al., 2024). (4) MM-Vet (Yu et al., 2023) that evaluate LVLMs on complex tasks such as language generation. These datasets contain open-ended questions and the referenced answers, and evaluation is performed using an LLM-as-a-Judge (Zheng et al., 2023) approach, which involves using a judge model like GPT-4o (Hurst et al., 2024) to predict scores.

We also report the performance on other categories, such as MMBench (Liu et al., 2025), MMMU (Yue et al., 2024) and MMStar (Chen et al., 2024b) for general knowledge, MathVerse (Zhang et al., 2025b) and MathVision (Wang et al., 2024c) for math reasoning, TextVQA (Singh et al., 2019), ChartQA (Masry et al., 2022) and OCR-bench (Liu et al., 2024d) for text-rich document understanding. These benchmarks utilize multiple-choice questions (MCQ) or visual question answering (VQA), where responses are limited to short

Table 4: **Evaluation results on Reward Bench (Lambert et al., 2024b).** We report the performance of selective representative language-only RMs and previous multi-modal generative RMs.

Model Name	LLM	Chat	Chat Hard	Safety	Reasoning	Avg Score
Language-Only Reward Models						
InternLM2-7B-Reward (Cai et al., 2024)	InternLM2-7B	99.2	69.5	87.2	94.5	87.6
InternLM2-20B-Reward (Cai et al., 2024)	InternLM2-20B	98.9	76.5	89.5	95.8	90.2
Skyword-Reward-Llama3.1-8B (Liu et al., 2024a)	Llama3.1-8B	95.8	87.3	90.8	96.2	<u>92.5</u>
INF-ORM-Llama3.1-70B (Yang et al., 2024a)	Llama3.1-70B	96.6	91.0	93.6	99.1	**95.1**
Multi-Modal Reward Models						
QWen2-VL-7B (Wang et al., 2024d)	QWen2-7B	96.6	57.0	73.9	94.3	83.8
LLaVA-Critic-8B (Xiong et al., 2024)	LLaMA3-7B	96.9	52.8	81.7	83.5	80.0
IXC-2.5-Reward (Ours)	InternL2-7B	90.8	83.8	87.8	90.0	88.6

Table 5: **Evaluation results on RM-Bench (Liu et al., 2024b).** We classify reward models into three types: sequence classifiers (Seq.), generative models, and implicit DPO models. Performance is reported across four domains (Chat, Math, Code, Safety) and three difficulty levels (Easy, Normal, Hard), along with average scores.

Model Name	Type	Chat	Math	Code	Safety	Easy	Normal	Hard	Avg
Language-Only Reward Models									
Tulu-2-dpo-13b (Ivison et al., 2023)	Implicit	66.4	51.4	51.8	85.4	86.9	66.7	37.7	63.8
InternLM2-7B-Reward (Cai et al., 2024)	Seq.	61.7	71.4	49.7	85.5	85.4	70.7	45.1	67.1
InternLM2-20B-Reward (Cai et al., 2024)	Seq.	63.1	66.8	56.7	86.5	82.6	71.6	50.7	68.3
Nemotron-4-340B-Reward (Wang et al., 2024f)	Generative	71.2	59.8	59.4	87.5	81.0	71.4	56.1	69.5
URM-LLaMa-3.1-8B (Lou et al., 2024)	Seq.	71.2	61.8	54.1	93.1	84.0	73.2	53.0	<u>70.0</u>
Skyword-Reward-Llama3.1-8B (Liu et al., 2024a)	Seq.	69.5	60.6	54.5	95.7	89.0	74.7	46.6	**70.1**
Multi-Modal Reward Models									
IXC-2.5-Reward (Ours)	Seq.	65.5	55.9	51.7	93.8	87.5	71.3	47.4	68.8

Table 6: Evaluation results of our IXC-2.5-Chat model against previous SOTA proprietary and open-source models $\leq$10B (results are copied from OpenVLM Leaderboard and Open LMM Reasoning Leaderboard, accessed 01-Jan-2025). **Best** and <u>second best</u> results are highlighted.

Category	Benchmark	Evaluation	Proprietary API Previous-SOTA	Open-Source Model ($\leq$**10B**) Previous-SOTA	IXC-2.5	IXC-2.5-Chat
Instruction Following & Chat	WildVision$_{(0617)}$ (Lu et al., 2024c)	Open	89.2 (Hurst et al., 2024)	67.3 (Xiong et al., 2024)	37.5	74.6
	MIA$_{(val)}$ (Qian et al., 2024)	Open	88.6 (Hurst et al., 2024)	80.7 (Wang et al., 2024d)	80.4	84.0
	MM-MT$_{(val)}$ (Agrawal et al., 2024)	Open	7.72 (Hurst et al., 2024)	5.45 (Wang et al., 2024d)	3.85	5.70
	MM-Vet v2$_{(0613)}$ (Yu et al., 2023)	Open	71.8 (Anthropic, 2024)	58.1 (Chen et al., 2024d)	45.8	54.8
Knowledge	MMBench$_{(v1.1)}$ (Liu et al., 2025)	MCQ	85.7 (SenseTime, 2024)	82.7 (Lu et al., 2024b)	79.4	79.0
	MMMU$_{(val)}$ (Yue et al., 2024)	MCQ	70.7 (Hurst et al., 2024)	56.2 (Chen et al., 2024d)	42.9	44.1
	MMStar (Chen et al., 2024b)	MCQ	72.7 (SenseTime, 2024)	63.2 (Chen et al., 2024d)	59.9	59.6
Reasoning	MathVista$_{(mini)}$ (Lu et al., 2023)	VQA	78.4 (SenseTime, 2024)	66.5 (Lu et al., 2024a)	63.7	63.4
	MathVerse$_{(vision\text{-}only)}$ (Zhang et al., 2025b)	VQA	54.8 (Google, 2024)	26.6 (Liu et al., 2024c)	16.2	19.0
	MathVision$_{(full)}$ (Wang et al., 2024c)	VQA	43.6 (Google, 2024)	22.0 (Liu et al., 2024c)	17.8	18.8
Text-Rich	TextVQA$_{(val)}$ (Singh et al., 2019)	VQA	82.0 (Megvii, 2024)	78.5 (Li et al., 2024a)	78.2	81.3
	ChartQA$_{(test)}$ (Masry et al., 2022)	VQA	81.2 (Megvii, 2024)	82.4 (Yao et al., 2024)	82.2	80.5
	OCRBench (Liu et al., 2024d)	VQA	89.4 (SenseTime, 2024)	82.2 (Chen et al., 2024d)	69.0	70.0

keywords and evaluated based on string matching.

Results on Instruction Following & Chat. Tab. 6 shows that IXC-2.5-Chat outperforms previous SOTA models across multiple benchmarks (WildVision, MIA, and MM-MT), demonstrating significant improvements in multi-modal understanding with instruction following ability and providing more comprehensive information for in-the-wild chat scenarios.

Results on Other Categories. On other categories (Knowledge, Reasoning, and Text-Rich), IXC-2.5-Chat performs comparably to the supervised fine-tuned (SFT) model IXC-2.5, demonstrating that RL training with IXC-2.5-Reward improves instruction following and conversational ability without sacrificing performance in these areas.

6 Conclusion and Future Work

We present IXC-2.5-Reward, a multi-modal reward model that is capable of multi-modal RL training, test-time scaling, and data cleaning. Using IXC-2.5-Reward, we further trained IXC-2.5-Chat via RLHF techniques to optimize the multi-modal user chat experience, focusing on providing detailed explanations and in-depth answers. We believe that exploring multi-modal reward models with on-policy reinforcement learning algorithms holds significant promise for future research, such as exploring reward benchmarks and RL algorithms for video alignment.

7 Limitations

The limitation of our work stems from the composition of our training data, which is primarily sourced from English language corpora. This reliance on English-centric data potentially limits the multilingual capabilities of our reward model. The English language datasets may reflect specific cultural viewpoints and societal biases prevalent in English-speaking communities. Future research should consider the incorporation of multilingual datasets to mitigate these limitations and enhance the generalizability and fairness of the multi-modal reward model.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *Preprint*, arXiv:2402.14740.

Open AI. 2024. Hello gpt-4o.

AI Anthropic. 2024. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 3.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, and 3 others. 2021. A general language assistant as a laboratory for alignment. *Preprint*, arXiv:2112.00861.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and 1 others. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Rui Cao, Yuming Jiang, Michael Schlichtkrull, and Andreas Vlachos. 2024. Decompose and leverage preferences from expert models for improving trustworthiness of mllms. *arXiv preprint arXiv:2411.13697*.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. ALLaVA: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. 2021. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, and 1 others. 2024c. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024d. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. UltraFeedback: Boosting language models with scaled ai feedback. *Preprint*, arXiv:2310.01377.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang.

2024a. Safe RLHF: Safe reinforcement learning from human feedback. In *ICLR*.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024b. Nvlm: Open frontier-class multimodal llms. *Preprint*, arXiv:2409.11402.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *Preprint*, arXiv:2409.17146.

Shijian Deng, Wentian Zhao, Yu-Jhe Li, Kun Wan, Daniel Miranda, Ajinkya Kale, and Yapeng Tian. 2024a. Efficient self-improvement in multimodal large language models: A model-level judge-free approach. *arXiv preprint arXiv:2411.17760*.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang. 2024b. Enhancing large vision language models with self-training on image comprehension. *arXiv preprint arXiv:2405.19716*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. *Preprint*, arXiv:2110.08420.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *CVPR*.

Google. 2024. Gemini-2.0-Flash.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, and 1 others. 2023. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*.

Srishti Gureja, Lester James V Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2024. M-RewardBench: Evaluating reward models in multilingual settings. *arXiv preprint arXiv:2410.15522*.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. V-STaR: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457*.

Zhenyu Hou, Yilin Niu, Zhengxiao Du, Xiaohan Zhang, Xiao Liu, Aohan Zeng, Qinkai Zheng, Minlie Huang, Hongning Wang, Jie Tang, and Yuxiao Dong. 2024. ChatGLM-RLHF: Practices of aligning large language models with human feedback. *Preprint*, arXiv:2404.00934.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *Preprint*, arXiv:2311.10702.

Songtao Jiang, Yan Zhang, Ruizhe Chen, Yeying Jin, and Zuozhu Liu. 2024. Modality-fair preference optimization for trustworthy mllm alignment. *arXiv preprint arXiv:2410.15334*.

Zhuoran Jin, Hongbang Yuan, Tianyi Men, Pengfei Cao, Yubo Chen, Kang Liu, and Jun o Zhao. 2024. Rag-rewardbench: Benchmarking reward models in retrieval augmented generation for preference alignment. *arXiv preprint arXiv:2412.13746*.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. 2024. MiraData: A large-scale video dataset with long durations and structured captions. *Preprint*, arXiv:2407.06358.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: Understanding data visualizations via question answering. In *CVPR*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *ECCV*.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. Solar 10.7b: Scaling large language models with simple yet effective depth up-scaling. *Preprint*, arXiv:2312.15166.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024a. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024b. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. LLaVA-OneVision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, and 1 others. 2024b. VLRewardBench: A challenging benchmark for vision-language generative reward models. *arXiv preprint arXiv:2411.17451*.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. 2024c. VLFeedback: A large-scale ai feedback dataset for large vision-language models alignment. *arXiv preprint arXiv:2410.09421*.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. 2023. Super-CLEVR: A virtual benchmark to diagnose domain robustness in visual reasoning. In *CVPR*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. CLEVR-Math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024a. Skywork-Reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024b. RM-Bench: Benchmarking reward models of language models with subtlety and style. *arXiv preprint arXiv:2410.16184*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2025. MMBench: Is your multi-modal model an all-around player? In *ECCV*.

Yuan Liu, Le Tian, Xiao Zhou, Xinyu Gao, Kavio Yu, Yang Yu, and Jie Zhou. 2024c. POINTS1. 5: Building a vision-language model towards real world applications. *arXiv preprint arXiv:2412.08443*.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024d. OCRBench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12).

Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2024e. Mia-DPO: Multi-image augmented direct preference optimization for large vision-language models. *arXiv preprint arXiv:2410.17637*.

Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2024f. MIA-DPO: Multi-image augmented direct preference optimization for large vision-language models. *arXiv preprint arXiv:2410.17637*.

Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. 2024. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *arXiv preprint arXiv:2410.00847*.

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

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024a. Ovis: Structural embedding alignment for multimodal large language model. `https://huggingface.co/AIDC-AI/Ovis1.6-Gemma2-9B`.

Xudong Lu, Yinghao Chen, Cheng Chen, Hui Tan, Boheng Chen, Yina Xie, Rui Hu, Guanxin Tan, Renshou Wu, Yan Hu, and 1 others. 2024b. BlueLM-V-3B: Algorithm and system co-design for multimodal large language models on mobile devices. *arXiv preprint arXiv:2411.10640*.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024c.

WildVision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Megvii. 2024. Taiyi. https://taiyi.megvii.com/.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. 2025. CLIP-DPO: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *ECCV*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. 2025. Strengthening multimodal large language model with bootstrapped preference optimization. In *ECCV*.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. 2024. Mia-bench: Towards better instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct Preference Optimization: Your language model is secretly a reward model. In *NeurIPS*.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2018. High-dimensional continuous control using generalized advantage estimation. *Preprint*, arXiv:1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *ECCV*.

SenseTime. 2024. SenseNova. https://platform.sensenova.cn/home.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2024. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146*.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: Knowledge-aware visual question answering. In *AAAI*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Tu Shiwen, Zhao Liang, Chris Yuhao Liu, Liang Zeng, and Yang Liu. 2024. Skywork critic model series. https://huggingface.co/Skywork.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *CVPR*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. 2024. MM-Eval: A multilingual meta-evaluation benchmark for llm-as-a-judge and reward models. *arXiv preprint arXiv:2410.17578*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Gemini Team. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Llama Team. 2024b. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

OpenGVLab Team. 2024c. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Chenglong Wang, Yang Gan, Yifu Huo, Yongyu Mu, Murun Yang, Qiaozhi He, Tong Xiao, Chunliang Zhang, Tongran Liu, Quan Du, and 1 others. 2024a. RoVRM: A robust visual reward model optimized via auxiliary textual preference data. *arXiv preprint arXiv:2408.12109*.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024b. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024c. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024d. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024e. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024f. HelpSteer2-Preference: Complementing ratings with preferences. *arXiv preprint arXiv:2410.01257*.

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. 2024. FunQA: Towards surprising video comprehension. *Preprint*, arXiv:2306.14899.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. LLaVA-Critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*.

Wang Xiyao, Yang Zhengyuan, Li Linjie, Lu Hongjin, Xu Yuancheng, Lin Chung-Ching Lin, Lin Kevin, Huang Furong, and Wang Lijuan. 2024. Scaling inference-time search with vision value model for improved visual comprehension. *arXiv preprint arXiv:2412.03704*.

Li Xu, He Huang, and Jun Liu. 2021. SUTD-TrafficQA: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9878–9888.

Minghao Yang, Chao Qu, and Xiaoyu Tan. 2024a. Inf outcome reward model. https://huggingface.co/infly/INF-ORM-Llama3.1-70B.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. 2024b. Regularizing hidden states enables learning generalizable reward model for llms. In *NeurIPS*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. MiniCPM-V: A gpt-4v level mllm on your phone. https://huggingface.co/openbmb/MiniCPM-V-2_6.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024a. RlHF-V: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *CVPR*.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024b. RLAIF-V: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and 1 others. 2024c. RLAIF-V: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. 2024a. Advancing llm reasoning generalists with preference trees. *Preprint*, arXiv:2404.02078.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024b. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STaR: Bootstrapping reasoning with reasoning. In *NeurIPS*.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2025a. Long-CLIP: Unlocking the long-text capability of clip. In *ECCV*.

Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, and 1 others. 2024a. Critic-V: Vlm critics help catch vlm errors in multimodal reasoning. *arXiv preprint arXiv:2411.18203*.

Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding, Qipeng Guo, Haodong Duan, Xin Chen, Han Lv, Zheng Nie, Min Zhang, Bin Wang, Wenwei Zhang, Xinyue Zhang, and 10 others. 2024b. InternLM-XComposer2.5-OmniLive:

A comprehensive multimodal system for long-term streaming video and audio interactions. *Preprint*, arXiv:2412.09596.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, and 8 others. 2024c. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *Preprint*, arXiv:2407.03320.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, and 8 others. 2024d. InternLM-XComposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2025b. Math-Verse: Does your multi-modal llm truly see the diagrams in visual math problems? In *ECCV*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024e. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024. ProcessBench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, and 1 others. 2024a. RMB: Comprehensively benchmarking reward models in llm alignment. *arXiv preprint arXiv:2410.09893*.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024b. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness & harmlessness with rlaif.

Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. 2024. Self-supervised visual preference alignment. In *ACM MM*.

Appendix

A More Experimental Results

Implementation Details For IXC-2.5-Reward, the learning rates were set at 1e-5 with a batch size of 256. As for IXC-2.5-Chat, the learning rates were set at 5e-5 with a batch size of 256. We set the PPO hyper-parameters $\gamma = 0.99$, $\beta = 0.95$, and $\epsilon = 0.2$.

Table 7: **Ablation Studies** of the impact of response length constraints of reward models that guided training IXC-2.5-Chat.

	Avg Tokens	Wild Vision	MIA	MM-MT	MM-Vet v2
w/o Length Constraints	**361**	**76.2**	**87.0**	**5.86**	**56.6**
IXC-2.5-Chat	274	74.6	84.0	5.70	54.8

The Impact of Length Constraints To prevent the chat model from generating overly long responses to artificially inflate rewards, we introduce length constraints on the ratio of chosen to rejected responses during training reward model IXC-2.5-Reward. The ablation study results of length constraints are present in Tab. 7. On the WildVision benchmark, we compute the average token length of the model's responses. We observe a substantial increase in average token length, from 274 to 361, when length constraints were not applied. Surprisingly, removing length constraints yielded substantial improvements in open-ended benchmarks, achieving state-of-the-art results. Such observation is because these benchmarks do not penalize length in their evaluation prompts, judge models (e.g., GPT-4) tend to favor longer responses, even if they contain unnecessary details that detract from the user experience. As our focus is on optimizing user experience, not benchmark scores, we retain the length constraints. Following the precedent set by language-only benchmarks (e.g., (Dubois et al., 2024)), we believe multi-modal Chat Arena and dialogue benchmarks should also address potential length and style biases in their evaluation protocols in future work.

Results on Test-Time Scaling According to Tab. 8, we observe that the Best-of-N sampling further improves the results. The averaged tokens is increased slightly (from 274 to 283), demonstrate that the improvements is bring from the high-quality response, rather than hacking the length bias in Tab. 7.

Table 8: **Results of Best-of-N (BoN) sampling** for test-time scaling with IXC-2.5-Reward.

	N	Avg Tokens	Wild Vision	MIA	MM-MT	MM-Vet v2
IXC-2.5-Chat		274	74.6	84.0	5.70	54.8
IXC-2.5-Chat + BoN	4	283	**77.7**	**87.3**	**6.03**	**56.3**

Visualization Results We present the visualization examples of IXC-2.5-Chat on a series of topics, such as instruction following (Fig. 3) and open-ended questions (Fig. 4). These figures reveal that IXC-2.5-Chat demonstrates several key advantages, including superior organization and presentation, more comprehensive and in-depth answers, and more detailed explanations. These strengths significantly enhance IXC-2.5-Chat's effectiveness in multi-modal chat interactions.

Please output the name, price, P/E ratio of all rising stocks in json format.

Symbol	Name	Price	Change	Change %	Volume	Avg Vol (3M)	Market Cap	P/E Ratio (TTM)
NVDA	NVIDIA Corporation	137.71	+4.14	+3.10%	201.189M	212.155M	3.373T	54.43
INTC	Intel Corporation	21.49	+1.82	+9.25%	166.484M	72.202M	92.686B	-
RGTI	Rigetti Computing, Inc.	9.83	-1.41	-12.54%	155.07M	118.947M	2.753B	-
PLUG	Plug Power Inc.	2.4400	-0.3000	-10.95%	95.902M	65.764M	2.223B	-
TSLA	Tesla, Inc.	426.50	+12.68	+3.06%	94.991M	94.873M	1.369T	116.21
MARA	MARA Holdings, Inc.	19.91	+1.61	+8.80%	88.413M	60.699M	6.757B	23.99
AAPL	Apple Inc.	229.98	+1.72	+0.75%	68.488M	46.061M	3.458T	37.89

IXC-2.5 :

```
[
  "NVIDIA Corporation", 137.71, 54.43],
  [
  "Intel Corporation", 21.49, -9.26],
  [
  "Tesla, Inc.", 426.50, 116.21],
  [
  "MARA Holdings, Inc.", 19.91, 23.99],
  [
  "Apple Inc.", 229.98, 37.89]
]
```

IXC-2.5-Chat:

```
```json
[
  {
   "name": "NVIDIA Corporation",
   "price": 137.71,
   "pe_ratio": 54.43
  },
  {
   "name": "Intel Corporation",
   "price": 21.49,
   "pe_ratio": null
  },
  {
   "name": "Tesla, Inc.",
   "price": 426.50,
   "pe_ratio": 116.21
  },
  {
   "name": "MARA Holdings, Inc.",
   "price": 19.91,
   "pe_ratio": 23.99
  },
  {
   "name": "Apple Inc.",
   "price": 229.98,
   "pe_ratio": 37.89
  }
]
```
```

Note:
- The P/E ratios for Rigetti Computing, Inc., Plug Power Inc., and Intel Corporation are not available.

	IXC-2.5	IXC-2.5-Chat
Name	✅	✅
Price	✅	✅
P/E ratio	❌	✅
All rising stocks	✅	✅
JSON Format	❌	✅

Figure 3: **Visualizations** of multi-modal dialogues generated by IXC-2.5-Chat on **instruction following** abilities.

How can I redesign this to make it look more professional?

IXC-2.5:

To redesign the room to make it look more professional, consider the following steps:

1.Color Scheme: Introduce a cohesive color scheme that complements the books. This could involve painting the walls a neutral color like light gray or beige, which would help the books stand out without overwhelming the space.

2.Book Arrangement: Organize the books by genre, author, or size to create a visually appealing and functional layout. Grouping similar books together can make the room appear more organized and professional.

3.Shelving: Consider upgrading the shelving to a more modern and sturdy design. This could include using metal or glass shelving units that reflect light and give the room a more open and airy feel.

4.Lighting: Improve the lighting by adding adjustable track lighting or recessed lighting. This will not only enhance the visibility of the books but also create a more inviting and professional atmosphere.

5.Furniture: Introduce a desk or table in the center of the room to serve as a focal point. This can be used for reading, working, or displaying decorative items that complement the room's theme.

6.Flooring: Consider replacing the carpet with hardwood or laminate flooring. This will give the room a more polished and professional appearance.

7.Wall Decor: Add artwork or framed prints that are related to literature or the subjects of the books. This will create a thematic connection and add visual interest to the room.

8.Storage Solutions: Incorporate additional storage solutions such as cabinets or drawers to keep the room clutter-free and maintain a professional appearance.

By implementing these changes, the room can be transformed into a more professional and aesthetically pleasing space that reflects the importance of the books and the activities that take place within the room.

IXC-2.5-Chat:

Redesigning a room to look more professional often involves a combination of organization, aesthetic appeal, and functional improvements. Here are some steps you can take to enhance the professional look of this room:

1. **Bookshelf Arrangement:**
- **Consistency:** Ensure that all bookshelves have a uniform look by using the same type of shelves and consistent spacing between books.
- **Labelling:** Label the shelves with categories or genres for easy navigation.
- **Presentation:** Organize books by size and color to create a visually appealing gradient effect.

2. **Storage Solutions:**
- **Bookends:** Use uniform bookends to maintain a neat appearance.
- **Storage Boxes:** For overflow items, consider using uniform storage boxes that blend with the room's color scheme.

3. **Office Equipment:**
- **Printer Placement:** Relocate the printer to a corner or a dedicated desk area to keep it out of the way of the bookshelves.
- **Cables:** Use cable management solutions to keep cables tidy and out of sight.

4. **Lighting:**
- **Task Lighting:** Add a desk lamp or a floor lamp near the printer for better task lighting.
- **Natural Light:** If possible, maximize natural light by ensuring the window is clean and unobstructed.

5. **Color Scheme:**
- **Neutral Colors:** Use a neutral color palette for the walls, shelves, and furniture to create a clean and professional look.
- **Color Accents:** Add a pop of color with a few well-chosen decorative items or artwork.

6. **Wall Decor:**
- **Artwork:** Consider adding framed artwork or prints that are relevant to the room's purpose (e.g., a map for a study room, a motivational quote for a creative space).
- **Bulletin Board:** A bulletin board can be useful for keeping important notes and reminders organized.

7. **Floor Space:**
- **Clear Space:** Ensure there is enough open floor space to make the room feel less cluttered.
- **Area Rugs:** If the floor is bare, consider adding an area rug to define the workspace and add warmth.

8. **Personal Touches:**
- **Plants:** Incorporate a few low-maintenance plants to add life to the room without overwhelming the space.
- **Personal Items:** Keep personal items to a minimum to maintain a professional atmosphere.

9. **Cleanliness:**
- **Dust and Clutter:** Regularly clean the room to ensure it looks well-maintained and professional.
- **Declutter:** Remove any unnecessary items that do not contribute to the room's purpose.

By implementing these changes, you can transform the room into a more organized, functional, and visually appealing space that exudes professionalism.

Figure 4: **Visualizations** of multi-modal dialogues generated by IXC-2.5-Chat on **open-ended** questions.