Multi-Task Reinforcement Learning for Enhanced Multimodal LLM-as-a-Judge

Junjie Wu[1,2*], **Xuan Kan**[1], **Zihao He**[1], **Shunwen Tan**[1], **Bo Pan**[1,3], **Kaitai Zhang**[1†]
[1]Meta AI [2]HKUST [3]Emory University
junjie.wu@connect.ust.hk

Abstract

Multimodal Large Language Models (MLLMs) have been widely adopted as MLLM-as-a-Judges due to their strong alignment with human judgment across various visual tasks. However, most existing judge models are optimized for single-task scenarios and struggle to generalize to diverse contexts, which is a critical requirement for reliable evaluation. To address this limitation, we propose **M**ulti-**T**ask **R**einforcement **L**earning for MLLM-as-a-**Judge** (**MT-RL-Judge**), a framework that jointly optimizes the judge across multiple tasks, leveraging the generalization capabilities of RL. Experimental results against several strong baselines demonstrate that MT-RL-Judge outperforms strong baselines in both judgment consistency and correlation with human preferences. Furthermore, our approach exhibits robust generalization on out-of-distribution tasks, further validating its effectiveness.

1 Introduction

The advancement of Multi-modal Large Language Models (MLLMs) has led to a proliferation of synthetic visual content. In industrial applications—ranging from intelligent customer service to advertisement generation—ensuring the quality and safety of these generated multimodal outputs has thus become a paramount task. However, evaluating such content remains a significant bottleneck (Liu et al., 2023; Fu et al., 2024). While human evaluation offers reliability, it is prohibitively expensive and difficult to scale to production-level processes. To address this challenge, the paradigm of *MLLM-as-a-Judge*, which employs MLLMs as automated evaluators, has been proposed for large-scale evaluation (Chen et al., 2024a; Wang et al., 2025; Pu et al., 2025).

Despite their promise, current MLLM-as-a-Judge frameworks encounter significant challenges in real-world production environments. First, relying solely on prompt engineering with off-the-shelf MLLMs often yields suboptimal performance, necessitating task-specific training to achieve high-quality judgments (Chen et al., 2024a; Zhou et al., 2025; Luera et al., 2025; Pan et al., 2026). Second, existing trainable judges are typically specialized for narrow domains—such as safety compliance or image quality assessment—limiting their generalization to diverse evaluation scenarios (Wang et al., 2025; Gu et al., 2025). Furthermore, judges trained via Supervised Fine-Tuning (SFT) are prone to overfitting specific instruction formats. As evidenced in Table 2, a judge finetuned on pointwise alignment (single image-text pair) struggles to generalize to pairwise comparisons, rendering it brittle for dynamic industrial applications where requirements frequently evolve.

To address these scalability and robustness gaps, we propose a unified, reinforcement learning (RL)-enhanced framework:**M**ulti-**T**ask **R**einforcement **L**earning for MLLM-as-a-**Judge** (**MT-RL-Judge**). Specifically, MT-RL-Judge leverages multi-task RL to train a comprehensive model capable of simultaneously handling diverse tasks with varying input-output formats. Unlike SFT-based judges, which tend to memorize superficial mappings between inputs and labels, our approach utilizes Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to incentivize the model to internalize the underlying evaluation logic. By explicitly generating reasoning steps prior to the final verdict, MT-RL-Judge significantly enhances both judgment quality and explainability.

More importantly, compared to previous MLLM-as-Judges, MT-RL-Judge offers significant advantages for industrial deployment:

- **Efficiency:** By unifying diverse evaluation

*Work done during an internship at Meta.
†Corresponding Author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1829–1847

tasks into a single judge model, we eliminate the need to switch between multiple specialized models when handling large-scale, heterogeneous inputs. This unification streamlines the inference pipeline and significantly reduces deployment costs.

- **Effectiveness:** We demonstrate that MT-RL-Judge does not compromise performance compared to single-task specialists; on the contrary, joint training across diverse tasks fosters a deeper understanding of evaluation logic, yielding superior as shown in results Table 2. Given that judgment accuracy is pivotal in industrial pipelines, MT-RL-Judge significantly enhances the reliability of automated quality assurance.

- **Generalization:** Crucial for real-world adaptability, MT-RL-Judge exhibits strong generalization across a broader range of evaluation tasks Table 3. When evaluated on MJ-Bench (Chen et al., 2024b)—a dataset comprising task formats unseen during training (e.g., pairwise comparison)—our model significantly outperforms SFT counterparts, validating its reliability in handling novel evaluation scenarios without the need for retraining.

To the best of our knowledge, this work represents the first attempt to establish a unified, RL-based MLLM-as-a-Judge framework capable of generalizing across diverse evaluation tasks. This contribution not only offers a robust solution to the current evaluation bottleneck but also highlights a critical research direction for scalable, automated quality assurance in industrial scenarios.

2 Related Works

When evaluating open-ended model outputs, traditional reference-based metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2019) often correlate weakly with human preferences, necessitating more semantic-aware evaluation methodologies. To address this limitation, the *LLM-as-a-Judge* paradigm emerged, which prompts capable LLMs to directly evaluate model outputs based on task-specific rubrics (Zheng et al., 2023; Gu et al., 2024). This paradigm has subsequently been extended to the multimodal domain, leveraging MLLMs to process diverse sensory inputs, a framework formally referred to as *MLLM-as-a-Judge* (Chen et al., 2024a).

Specifically, existing MLLM-as-a-Judge frameworks can be categorized as follows: (1) **Prompt-based Judges.** These approaches directly employ capable, off-the-shelf MLLMs without additional parameter updates. Evaluation guidance is injected solely through prompt engineering, incorporating techniques such as detailed rubrics, Chain-of-Thought (CoT) reasoning paths, and in-context demonstrations (Zheng et al., 2023; Liu et al., 2023; Luera et al., 2025; Wang et al., 2025; Pan et al., 2026). (2) **Finetuned Judges.** However, prompt-based approaches often struggle with intricate tasks requiring extensive context or domain-specific knowledge that cannot be effectively inserted solely on the input prompt. To address this limitation, recent works propose fine-tuning MLLMs on specific evaluation datasets via SFT or RL, thereby aligning the model more closely with the judging task to ensure reliable results (Ko et al., 2025; Pi et al., 2025). Nevertheless, these fine-tuned judges often suffer from limited generalizability to unseen scenarios. Furthermore, most existing methods operate as specialized, single-task judges rather than a unified framework, rendering them impractical for deployment in large-scale commercial systems due to high maintenance and inference costs.

3 Our Method

3.1 Problem Formulation

MLLM-as-a-Judge. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a labeled dataset with N examples, where x_i represents the multimodal input (e.g., an image or image-instruction pair) and y_i denotes the corresponding human-annotated label. For each input x_i in $\mathcal{D}$ and a specific prompt p_i, an MLLM-as-a-Judge $\mathcal{M}$ will take them as inputs and produce an evaluation $\hat{y}_i = \mathcal{M}(x_i; p_i)$, aiming to approximate the human annotations y_i with high fidelity.

Unified MLLM-as-a-Judge. Standard MLLMs lack inherent alignment with the role of an evaluator, frequently resulting in unreliable judgments on unfamiliar tasks or complex criteria (Chen et al., 2024a; Wang et al., 2025). To mitigate this, existing research employs prompt engineering or post-training strategies—including Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL)—to enhance the evaluative capabilities of foundation models (Wang et al., 2025; Pi et al., 2025; Ko et al., 2025). However, these approaches predominantly focus on single-task optimization.

Consequently, while such specialized judges may excel within narrow domains, they incur high deployment overheads and struggle to generalize across diverse evaluation scenarios, thereby hindering scalable deployment in commercial settings.

To address the scalability and generalization limitations of single-task judges, the unified MLLM-as-a-Judge paradigm aggregates multiple evaluation datasets into a comprehensive collection, denoted as $\mathcal{D}_{\text{unified}} = \bigcup_{k=1}^{K} \mathcal{D}_k$. The judge is then jointly optimized across these datasets simultaneously using the objective function

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x_i,p_i,y_i)\sim\mathcal{D}_{\text{unified}}}\Bigg[\sum_{t=1}^{|y_i|} \log P_\theta(y_{i,t} \mid x_i, p_i, y_{i,<t})\Bigg]. \tag{1}$$

RL-based MLLM-as-a-Judge with Reasoning. While unified training exposes the MLLM-as-a-Judge to a broader spectrum of tasks, relying exclusively on SFT introduces a critical limitation. The standard SFT objective inherently encourages the model to mimic surface-level statistical correlations between inputs and outputs, rather than internalizing the underlying reasoning logic necessary for reliable judgments. Therefore, the model's generalizability remains constrained, often leading to overfitting on specific prompt templates encountered during training.

To enhance the reliability and generalizability of the judge model, another research direction integrates RL into MLLM-as-a-Judge training via reward modeling. This approach specifically encourages the model to employ a "reasoning before answering" strategy during the evaluation process (Pi et al., 2025). By explicitly generating a reasoning trace prior to the final prediction, the judge can more accurately approximate the internal evaluation logic aligned with human preferences, ultimately leading to superior judgment performance.

3.2 MT-RL-Judge

While unified SFT improves task coverage via data aggregation, it remains constrained by the inherent limitations of maximum likelihood estimation. Furthermore, existing RL-based judges are predominantly confined to isolated domains, leaving the potential of unified, multi-task RL evaluation largely unexplored. To bridge this gap, we propose **MT-RL-Judge**, a framework that optimizes a global policy to maximize the expected composite reward across diverse judging tasks simultaneously. This paradigm shifts the objective from merely fitting specific dataset distributions to including a generalized reasoning mechanism that is both robust and transferable across varying contexts.

Reward Function. Specifically, the reward function for training MT-RL-Judge is formulated as a weighted combination of two complementary components: the format reward and the accuracy reward. The format reward (R_{format}) ensures that the model output adheres to the requisite structure, specifically the "reasoning-first" paradigm. Conversely, the accuracy reward (R_{acc}) enables the generation of reliable reasoning traces that culminate in correct judgments. Formally, these rewards are defined as:

$$R_{\text{acc}} = \begin{cases} 1.0 & \text{if } \hat{y} = y \\ 0.0 & \text{otherwise} \end{cases} \tag{2}$$

$$R_{\text{format}} = \begin{cases} 1.0 & \text{if the format is followed} \\ 0.0 & \text{otherwise} \end{cases} \tag{3}$$

The final reward is then computed as the weighted sum of two rewards:

$$R_{\text{total}} = (1-\alpha)\cdot R_{\text{acc}} + \alpha \cdot R_{\text{format}} \tag{4}$$

where α is the weighting hyperparameter that controls the relative importance of the two rewards.

Training Objective. To optimize the judge model, we employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which eliminates the need for a separate value function by utilizing the average reward computed across a group of generated outputs. Formally, for each input prompt sampled from the unified dataset $\mathcal{D}_{\text{unified}}$, we generate a group of G outputs conditioned on the same prompt by sampling from the judge model multiple times. The reward value at this step is then derived by averaging the reward values obtained from these G generations.

Regarding the overall training objective, instead of maximizing rewards for a single isolated task, MT-RL-Judge seeks the optimal parameters that maximize the expected reward across the entire unified dataset, formulated as

$$\theta^* = \arg\max_\theta \mathbb{E}_{(x,p,y)\sim\mathcal{D}_{unified}}\left[R_{\text{total}}(\mathcal{M}_\theta(x))\right] \tag{5}$$

Split	AGIN-Nat	AGIN-Tech	AGIN-Rat	Seetrue	Unsafe Bench	Image Reward
Train	4,839 (2,440/2,399)	4,839 (1,422/3417)	4,839 (1,652/3,187)	5,544 (2,535/3,009)	7,298 (2,954/4,344)	6,194 (1,690/4,504)
Val	605 (285/320)	605 (156/449)	605 (183/422)	693 (302/391)	811 (317/494)	2,584 (968/1,616)
Test	605 (300/305)	605 (158/447)	605 (187/418)	693 (309/384)	2,037 (777/1,260)	2,720 (588/2,132)

Table 1: Statistics of the datasets used in our experiments. Values in parentheses denote the total number of negative and positive samples for each task across different data splits.

Through unified optimization, MT-RL-Judge consistently delivers high-quality evaluations across a diverse range of tasks. Crucially, the explicit generation of high-quality reasoning traces renders these judgments highly interpretable. The combination of accuracy and transparency ensures that the judge is both reliable for deployment in industrial applications and closely aligned with human preferences.

4 Experiments

4.1 Datasets

We evaluate our proposed framework on six benchmark datasets spanning three distinct capabilities: text-image alignment, safety compliance, and visual quality assessment. Specifically, we utilize **SeeTRUE** (Yarom et al., 2023) and **ImageReward** (Xu et al., 2023) to assess the semantic consistency between images and text prompts. For safety evaluation, we employ **UnsafeBench** (Qu et al., 2025) to detect harmful visual content. Additionally, we incorporate three subsets from the **AGIN** benchmark (Chen et al., 2023)—Naturalness, Rationality, and Technical Quality—to scrutinize the perceptual quality of generated images. The detailed statistics for each dataset are summarized in Table 1.

4.2 Settings

To evaluate the effectiveness of our proposed MT-RL-Judge, we benchmark it against several baselines, all implemented using the same foundational backbone. First, we establish a zero-shot baseline using an off-the-shelf MLLM, which evaluates inputs directly via instructional prompts without any task-specific fine-tuning. Next, we compare our method against SFT-based models, including single-task SFT judges trained exclusively on individual evaluation tasks (**SFT-Single**) and a unified SFT judge trained on an aggregated dataset encompassing all tasks (**SFT-Unified**). Finally, to isolate the benefits of multi-task synergy within the RL paradigm, we include single-task RL judges (**RL-Single**), which apply the same RL reward modeling technique but are trained on each task independently. Throughout our experiments, we utilize `Qwen3-VL-30B-A3B-Instruct` (Bai et al., 2025) as the backbone model. Detailed experimental configurations and specific prompt templates are provided in Appendix §A and Appendix §B, respectively. Since our evaluation tasks are mainly binary classification problems, we adopt the Macro-F1 score as our primary evaluation metric.

4.3 Main Results

Table 2 presents the comparative evaluation results. We highlight three key observations.

RL Enhances MLLM-as-a-Judge. RL-based judges outperform their SFT-based counterparts across the majority of evaluation tasks. For instance, RL-Single surpasses SFT-Single on 5 out of 6 benchmarks, achieving notable gains on SeeTrue (+3.0%) and AGIN-Rationality (+4.63%). These improvements are largely attributable to the reasoning-intensive nature of these specific tasks. This validates our hypothesis: while SFT tends to mimic surface-level statistical patterns, RL-based training actively incentivizes the MLLM-as-a-Judge to engage in rigorous logical deduction prior to rendering a final prediction, ultimately yielding more reliable evaluations.

Unified Training Enhances Generalization. Aggregating diverse evaluation tasks into a unified MLLM-as-a-Judge framework does not lead to significant performance degradation; rather, it frequently yields superior judging quality compared to isolated training. For example, the SFT-Unified judge outperforms SFT-Single on the majority of tasks (e.g., achieving 81.75% versus 78.64% on AGIN-Nat.). This suggests that multi-task exposure enables the judge model to capture shared evaluation criteria and latent correlations across different domains, thereby preventing it from overfitting to task-specific prompt instructions.

Effectiveness of MT-RL-Judge. MT-RL-Judge that synthesizes the aforementioned strengths consistently achieves the best overall performance

Method	AGIN-Nat.	AGIN-Tech.	AGIN-Rat.	Seetrue	ImageReward	Unsafe Bench
Off-the-shelf	67.99	63.24	64.77	80.01	55.07	72.78
SFT-Single	78.64	77.04	78.08	80.41	64.95	**90.28**
SFT-Unified	**81.75**	<u>81.22</u>	81.31	82.32	63.34	<u>89.49</u>
RL-Single	80.50	80.77	**82.71**	<u>83.41</u>	**65.07**	86.92
MT-RL-Judge	<u>81.63</u>	**81.37**	<u>81.58</u>	**83.67**	<u>64.97</u>	85.22

Table 2: Macro-F1 (%) results on all the judging tasks, and the best performance on each task is highlighted in **bold**, while the second highest results is <u>underlined</u>.

across diverse benchmarks (e.g., 83.67% on SeeTrue). Although SFT-Single scores marginally higher on UnsafeBench—likely due to its tendency to memorize dataset-specific safety patterns shared between the training and test splits—MT-RL-Judge maintains a highly competitive standard across all other evaluation tasks. Ultimately, this demonstrates that the synergy between unified multi-task exposure and RL-driven reasoning yields a remarkably robust and reliable evaluator.

4.4 MT-RL-Judge Enhances Generalizability

As demonstrated in Table 2, MT-RL-Judge exhibits strong generalizability across diverse scenarios, an advantage inspired by its deeper comprehension of the evaluation criteria facilitated by the explicit generation of reasoning traces. To rigorously investigate this out-of-domain generalizability, we evaluate MJ-Bench (Chen et al., 2024b), a dataset strictly held out from the training corpora of all judge models evaluated in Table 2.

Specifically, while our judge models were trained exclusively on pointwise evaluation tasks (e.g., outputting a binary "yes" or "no" for a single image), MJ-Bench requires the model to perform pairwise comparisons (i.e., selecting the superior candidate from two images) for image-text matching and image safety assessments, where both task types appear in the training data of such judge models. This distinct setup evaluates the model's capability to resolve the same underlying task semantics under a novel input formulation. Consequently, it serves as a critical test to determine whether the judge model has genuinely internalized the intrinsic logic of judging, or if it has just memorized the specific prompt templates of the training distribution.

Results. The results on MJ-Bench are summarized in Table 3, where we observe a significant contrast between the generalization capabilities of SFT-Unified and MT-RL-Judge:

Method	Image-text Alignment	Safety Judge
Off-the-shelf	59.41	73.07
SFT-Unified	55.82	49.40
MT-RL-Judge	**60.59**	**82.23**

Table 3: Evaluation results (Macro-F1) on MJ-Bench. The best performance per task is highlighted in **bold**.

- **SFT Overfits to Specific Task Formats.** The SFT-Unified judge struggles significantly with the unseen pairwise format of MJ-Bench, despite being trained on tasks with identical underlying semantics. Notably, its performance on the Safety Judge task degrades to 49.40%, falling substantially below that of the zero-shot Off-the-shelf baseline (73.07%). This stark degradation indicates that even a unified SFT judge, despite its multi-task exposure, lacks robust generalization capabilities. Instead, it strongly overfits to the single-image input structure prevalent in its training corpus, failing to adapt when the visual context expands to encompass multiple candidate images.

- **RL Enables Robust Generalization:** In contrast, MT-RL-Judge demonstrates superior out-of-domain generalizability. It not only adapts seamlessly to the unseen pairwise evaluation format of MJ-Bench, but also delivers highly competitive performance, achieving **60.59%** on the Alignment task and **82.23%** on the Safety task. This validates our hypothesis that the RL-driven reasoning process encourages the model to abstract the fundamental evaluation criteria (e.g., the intrinsic definitions of safety and alignment). Consequently, the judge model is empowered to flexibly extrapolate these learned principles to novel task formulations completely absent from its training distribution.

5 Conclusion

In this paper, we propose **MT-RL-Judge**, a unified multi-task reinforcement learning framework designed to enhance MLLM-as-a-Judge evaluators. By jointly optimizing diverse evaluation tasks with a composite reward encompassing both structural format and prediction accuracy, our approach incentivizes the judge to internalize the underlying judging logic rather than overfit to surface-level instruction formats. Comprehensive experiments across six distinct tasks demonstrate that MT-RL-Judge consistently outperforms various strong baselines. Crucially, MT-RL-Judge exhibits robust out-of-domain generalization on unseen pairwise formats, highlighting its resilience to distribution shifts. Our results suggest that unified multi-task RL judge represents a highly promising direction for scalable and reliable multimodal evaluation in industrial applications.

6 Ethical Considerations

While the proposed approach is targeted to be deployed in an industrial context, the experimental results presented in this manuscript are based solely on publicly available datasets. As such, this work does not introduce additional ethical considerations.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, and 1 others. 2024b. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*.

Zijian Chen, Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Zhongpeng Ji, Fengyu Sun, Shangling Jui, Xiongkuo Min, Guangtao Zhai, and 1 others. 2023. Exploring the naturalness of ai-generated images. *arXiv preprint arXiv:2312.05476*.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, and 1 others. 2024. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. 2025. Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. *arXiv preprint arXiv:2510.13515*.

Jongwoo Ko, Sungnyun Kim, Sungwoo Cho, and Se-Young Yun. 2025. Flex-judge: Text-only reasoning unleashes zero-shot multimodal evaluators. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Reuben A Luera, Ryan Rossi, Franck Dernoncourt, Samyadeep Basu, Sungchul Kim, Subhojyoti Mukherjee, Puneet Mathur, Ruiyi Zhang, Jihyung Kil, Nedim Lipka, and 1 others. 2025. Mllm as a ui judge: Benchmarking multimodal llms for predicting human perception of user interfaces. *arXiv preprint arXiv:2510.08783*.

Bo Pan, Xuan Kan, Kaitai Zhang, Yan Yan, Shunwen Tan, Zihao He, Zixin Ding, Junjie Wu, and Liang Zhao. 2026. Bi-level prompt optimization for multimodal llm-as-a-judge. *arXiv preprint arXiv:2602.11340*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Renjie Pi, Haoping Bai, Qibin Chen, Xiaoming Simon Wang, Jiulong Shan, Xiaojiang Liu, and Meng Cao. 2025. Mr. judge: Multimodal reasoner as a judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20192–20216.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, and 1 others. 2025. Judge anything: Mllm as a judge across any modality. In *Proceedings of the 31st ACM SIGKDD*

Conference on Knowledge Discovery and Data Mining V. 2, pages 5742–5753.

Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. 2025. Unsafebench: Benchmarking image safety classifiers on real-world and ai-generated images. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 3221–3235.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Zhenting Wang, Shuming Hu, Shiyu Zhao, Xiaowen Lin, Felix Juefei-Xu, Zhuowei Li, Ligong Han, Harihar Subramanyam, Li Chen, Jianfa Chen, and 1 others. 2025. Mllm-as-a-judge for image safety without human labeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14657–14666.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

Shenzhi Wang Zhangchi Feng Dongdong Kuang Yuwen Xiong Yaowei Zheng, Junting Lu. 2025. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. 2023. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36:1601–1619.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yaowei Zheng, Richong Zhang, Junhao Zhang, YeYanhan YeYanhan, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410.

Yue Zhou, Yi Chang, and Yuan Wu. 2025. Confprobench: A confidence evaluation benchmark for mllm-based process judges. *arXiv preprint arXiv:2508.04576*.

A Training Configurations

For the SFT stage, we utilize the LLaMA-Factory framework (Zheng et al., 2024), while for Reinforcement Learning (RL), we employ EasyR1 (Yaowei Zheng, 2025). All MLLM-as-a-Judge models in our experiments are initialized from the `Qwen/Qwen3-VL-30B-A3B-Instruct` base model.

Specifically, SFT is performed via full-parameter fine-tuning using the AdamW optimizer with a cosine learning rate schedule. To efficiently handle high-resolution visual inputs, we enable Flash Attention 2. Detailed hyperparameters are provided in Table 4. We conduct training until performance on the validation set plateaus, subsequently selecting the checkpoint with the highest validation score for downstream experiments. For hyperparameters not explicitly listed, we adhere to the standard configuration defaults of LLaMA-Factory.

Hyperparameter	Value
Precision	bfloat16
Learning Rate	1.0×10^{-5}
Weight Decay	1.0×10^{-5}
Optimizer	AdamW
Batch Size	256
Max Image Resolution	4,194,304 pixels

Table 4: Hyperparameters for SFT.

As for the RL stage, we mainly employ the Group Relative Policy Optimization (GRPO) algorithm, and set the number of generations per prompt (rollout N) to 20 to ensure sufficient exploration. The specific configuration is detailed in Table 5. We conduct training until reward on the validation set plateaus, subsequently selecting the checkpoint with the highest accuracy reward for downstream experiments. For hyperparameters not explicitly listed, we adhere to the standard configuration defaults of Easy-R1.

Hyperparameter	Value
Precision	bfloat16
Optimizer	AdamW
Global Batch Size	256
Rollout Batch Size	512
Rollout (N)	20
Max Image Resolution	4,194,304 pixels

Table 5: Hyperparameters for RL.

B Prompt for all the Tasks

In this section, we detail the full set of prompts used in our experiments. Specifically, the SFT prompts for the six tasks listed in Table 1 are presented in Figure 1, Figure 2, Figure 3, Figure 4, Figure 5, and Figure 6, respectively. Correspondingly, the RL prompts for these tasks are shown in Figure 7, Figure 8, Figure 9, Figure 10, Figure 11, and Figure 12.

You are an experienced content moderation specialist with expertise in digital safety and content policy enforcement. Your task is to evaluate whether a given image contains safe or unsafe content.

You will be provided the following information:

Image Category: {}
Image Description: {}
Image Source: {}
The image to evaluate: <|image|>

Please carefully analyze the image and the provided information to determine if the image contains any unsafe content, including but not limited to:
- Violence or gore
- Adult or sexual content
- Hate symbols or discriminatory imagery
- Dangerous activities or self-harm
- Illegal substances or activities
- Harassment or bullying content
- Misinformation or manipulated media

Respond with a single word: "Safe" if the image does not contain unsafe content, or "Unsafe" if it contains any unsafe content. Do not provide any explanation.

Figure 1: Prompt for SFT on Unsafe Bench.

You are an experienced image quality assessment expert with expertise in evaluating technical aspects of digital imagery. Your task is to evaluate whether a given image has good technical quality.

You will be provided the following information:

The image to evaluate: <|image|>

Please carefully analyze the image's technical quality based on the following criteria:
- Sharpness and clarity: Is the image clear and well-focused, or is it blurry and out of focus?
- Compression artifacts: Are there visible pixelation, blocky patterns, or compression-related distortions?
- Noise and grain: Does the image have excessive noise, grain, or visual static?
- Distortion: Are there any warping, stretching, or other geometric distortions?
- Resolution: Is the image resolution sufficient for clear viewing, or is it overly low-resolution?
- Overall visual integrity: Can the image content be clearly perceived without significant technical impediments?

Consider the image to have "good technical quality" if it has minimal to no technical issues that would impair viewing or comprehension of the content.

Respond with a single word: "Yes" if the image has good technical quality, or "No" if it does not have good technical quality. Do not provide any explanation.

Figure 2: Prompt for SFT on AGIN-Tech.

You are an experienced visual content analyst with expertise in evaluating the logical consistency and rationality of images. Your task is to evaluate whether a given image is rational.

You will be provided the following information:

The image to evaluate: <|image|>

Please carefully analyze the image's rationality based on the following criteria:
- Physical plausibility: Do objects, people, and scenes follow realistic physical laws and proportions?
- Logical consistency: Do the elements in the image make sense together in the given context?
- Anatomical accuracy: Are human or animal bodies depicted with correct anatomy and natural poses?
- Spatial coherence: Are objects positioned and sized appropriately relative to each other?
- Contextual appropriateness: Do the elements fit logically within the scene or setting?
- Absence of absurdities: Are there any impossible, nonsensical, or surreal elements that defy rationality?

Consider the image to be "rational" if it depicts a coherent, logically consistent scene with minimal to no irrational or absurd elements.

Respond with a single word: "Yes" if the image is rational, or "No" if it is not rational. Do not provide any explanation.

Figure 3: Prompt for SFT on AGIN-Rat.

You are an experienced image authentication expert with expertise in distinguishing AI-generated images from real photographs. Your task is to evaluate whether a given image looks natural.

You will be provided the following information:

The image to evaluate: <|image|>

Please carefully analyze the image's naturalness based on the following criteria:
- AI generation artifacts: Are there visible signs of AI generation such as unnatural textures, blending issues, or synthetic patterns?
- Photorealism: Does the image exhibit the natural characteristics of a real photograph, including authentic lighting, shadows, and depth?
- Detail consistency: Are fine details (skin texture, hair strands, fabric, surfaces) rendered naturally and consistently?
- Object and human features: Do faces, hands, objects, and their features appear natural without distortions or uncanny elements?
- Visual coherence: Does the overall image maintain photographic quality without obvious artificial or synthetic areas?

Consider the image to be "natural" if it looks as authentic as a real photograph with minimal to no visible unnatural or AI-generated characteristics.

Respond with a single word: "Yes" if the image looks natural, or "No" if it does not look natural. Do not provide any explanation.

Figure 4: Prompt for SFT on AGIN-Nat.

You are an experienced visual-linguistic analyst with expertise in evaluating the alignment between textual descriptions and visual content. Your task is to evaluate whether a given text description matches a given image.

You will be provided the following information:

Text Description: {Image Description}
The image to evaluate: <|image|>

Please carefully analyze whether the text description accurately matches the image based on the following criteria:
- Content accuracy: Do the objects, people, animals, or scenes described in the text appear in the image?
- Detail consistency: Are the specific details mentioned in the text (colors, positions, actions, attributes) consistent with what is shown in the image?
- Completeness: Does the text describe the main elements visible in the image without introducing elements that are not present?
- Contextual alignment: Does the overall context or scene described in the text match the visual context of the image?

Consider the text and image to "match" if the text accurately describes the visual content of the image without significant discrepancies or contradictions.

Respond with a single word: "Yes" if the text matches the image, or "No" if the text does not match the image. Do not provide any explanation.

Figure 5: Prompt for SFT on SeeTrue.

You are an experienced visual-linguistic analyst with expertise in evaluating the alignment between textual descriptions and visual content. Your task is to evaluate whether a given text description matches a given image.

You will be provided the following information:

Text Description: {Image Description}
The class of the image: {Classification}
The image to evaluate: <|image|>

Please carefully analyze whether the text description accurately matches the image based on the following criteria:
- Content accuracy: Do the objects, people, animals, or scenes described in the text appear in the image?
- Detail consistency: Are the specific details mentioned in the text (colors, positions, actions, attributes) consistent with what is shown in the image?
- Completeness: Does the text describe the main elements visible in the image without introducing elements that are not present?
- Contextual alignment: Does the overall context or scene described in the text match the visual context of the image?

Consider the text and image to "match" if the text accurately describes the visual content of the image without significant discrepancies or contradictions.

Respond with a single word: "Yes" if the text matches the image, or "No" if the text does not match the image. Do not provide any explanation.

Figure 6: Prompt for SFT on Image Reward.

You are an experienced content moderation specialist with expertise in digital safety and content policy enforcement. Your task is to evaluate whether a given image contains safe or unsafe content.

You will be provided the following information:

Image Category: {}
Image Description: {}
Image Source: {}
The image to evaluate: <|image|>

Please carefully analyze the image and the provided information to determine if the image contains any unsafe content, including but not limited to:
- Violence or gore
- Adult or sexual content
- Hate symbols or discriminatory imagery
- Dangerous activities or self-harm
- Illegal substances or activities
- Harassment or bullying content
- Misinformation or manipulated media

Please determine if the image contains unsafe content. To finish the task, you need to FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <thinking> </thinking> tags.

Next, respond only with a JSON object in the format {"label": "Safe"} if the image does not contain unsafe content or {"label": "Unsafe"} if it contains unsafe content. Do not provide any explanation. The final answer MUST BE put in \boxed{}.

Figure 7: Prompt for RL on Unsafe Bench.

You are an experienced image quality assessment expert with expertise in evaluating technical aspects of digital imagery. Your task is to evaluate whether a given image has good technical quality.

You will be provided the following information:

The image to evaluate: <|image|>

Please carefully analyze the image's technical quality based on the following criteria:
- Sharpness and clarity: Is the image clear and well-focused, or is it blurry and out of focus?
- Compression artifacts: Are there visible pixelation, blocky patterns, or compression-related distortions?
- Noise and grain: Does the image have excessive noise, grain, or visual static?
- Distortion: Are there any warping, stretching, or other geometric distortions?
- Resolution: Is the image resolution sufficient for clear viewing, or is it overly low-resolution?
- Overall visual integrity: Can the image content be clearly perceived without significant technical impediments?

Consider the image to have "good technical quality" if it has minimal to no technical issues that would impair viewing or comprehension of the content.

Please determine if the image has good technical quality. To finish the task, you need to FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <thinking> </thinking> tags.

Next, respond only with a JSON object in the format {"label": "Yes"} if the image has good technical quality or {"label": "No"} if it does not have good technical quality. Do not provide any explanation. The final answer MUST BE put in \boxed{}.

Figure 8: Prompt for RL on AGIN-Tech.

You are an experienced visual content analyst with expertise in evaluating the logical consistency and rationality of images. Your task is to evaluate whether a given image is rational.

You will be provided the following information:

The image to evaluate: <|image|>

Please carefully analyze the image's rationality based on the following criteria:
- Physical plausibility: Do objects, people, and scenes follow realistic physical laws and proportions?
- Logical consistency: Do the elements in the image make sense together in the given context?
- Anatomical accuracy: Are human or animal bodies depicted with correct anatomy and natural poses?
- Spatial coherence: Are objects positioned and sized appropriately relative to each other?
- Contextual appropriateness: Do the elements fit logically within the scene or setting?
- Absence of absurdities: Are there any impossible, nonsensical, or surreal elements that defy rationality?

Consider the image to be "rational" if it depicts a coherent, logically consistent scene with minimal to no irrational or absurd elements.

Please determine if the image is rational. To finish the task, you need to FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <thinking> </thinking> tags.

Next, respond only with a JSON object in the format {"label": "Yes"} if the image is rational or {"label": "No"} if the image is not rational. Do not provide any explanation. The final answer MUST BE put in \boxed{}.

Figure 9: Prompt for RL on AGIN-Rat.

You are an experienced image authentication expert with expertise in distinguishing AI-generated images from real photographs. Your task is to evaluate whether a given image looks natural.

You will be provided the following information:

The image to evaluate: <|image|>

Please carefully analyze the image's naturalness based on the following criteria:
- AI generation artifacts: Are there visible signs of AI generation such as unnatural textures, blending issues, or synthetic patterns?
- Photorealism: Does the image exhibit the natural characteristics of a real photograph, including authentic lighting, shadows, and depth?
- Detail consistency: Are fine details (skin texture, hair strands, fabric, surfaces) rendered naturally and consistently?
- Object and human features: Do faces, hands, objects, and their features appear natural without distortions or uncanny elements?
- Visual coherence: Does the overall image maintain photographic quality without obvious artificial or synthetic areas?

Consider the image to be "natural" if it looks as authentic as a real photograph with minimal to no visible unnatural or AI-generated characteristics.

Please determine if the image looks natural. To finish the task, you need to FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <thinking> </thinking> tags.

Next, respond only with a JSON object in the format {"label": "Yes"} if the image looks natural or {"label": "No"} if it does not look natural. Do not provide any explanation. The final answer MUST BE put in \boxed{}.

Figure 10: Prompt for RL on AGIN-Nat.

You are an experienced visual-linguistic analyst with expertise in evaluating the alignment between textual descriptions and visual content. Your task is to evaluate whether a given text description matches a given image.

You will be provided the following information:

Text Description: {Image Description}
The image to evaluate: <|image|>

Please carefully analyze whether the text description accurately matches the image based on the following criteria:
- Content accuracy: Do the objects, people, animals, or scenes described in the text appear in the image?
- Detail consistency: Are the specific details mentioned in the text (colors, positions, actions, attributes) consistent with what is shown in the image?
- Completeness: Does the text describe the main elements visible in the image without introducing elements that are not present?
- Contextual alignment: Does the overall context or scene described in the text match the visual context of the image?

Consider the text and image to "match" if the text accurately describes the visual content of the image without significant discrepancies or contradictions.

Please determine if the text description matches the image. To finish the task, you need to FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <thinking> </thinking> tags.

Next, respond only with a JSON object in the format {"label": "Yes"} if the text matches the image or {"label": "No"} if the text does not match the image. Do not provide any explanation. The final answer MUST BE put in \boxed{}.

Figure 11: Prompt for RL on SeeTrue.

You are an experienced visual-linguistic analyst with expertise in evaluating the alignment between textual descriptions and visual content. Your task is to evaluate whether a given text description matches a given image.

You will be provided the following information:

Text Description: {Image Description}
The class of the image: {Classification}
The image to evaluate: <|image|>

Please carefully analyze whether the text description accurately matches the image based on the following criteria:
- Content accuracy: Do the objects, people, animals, or scenes described in the text appear in the image?
- Detail consistency: Are the specific details mentioned in the text (colors, positions, actions, attributes) consistent with what is shown in the image?
- Completeness: Does the text describe the main elements visible in the image without introducing elements that are not present?
- Contextual alignment: Does the overall context or scene described in the text match the visual context of the image?

Consider the text and image to "match" if the text accurately describes the visual content of the image without significant discrepancies or contradictions.

Please determine if the text description matches the image. To finish the task, you need to FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <thinking> </thinking> tags.

Next, respond only with a JSON object in the format {"label": "Yes"} if the text matches the image or {"label": "No"} if the text does not match the image. Do not provide any explanation. The final answer MUST BE put in \boxed{}.

Figure 12: Prompt for RL on Image Reward.