

MINOS: A Multimodal Evaluation Model for Bidirectional Generation Between Image and Text

Junzhe Zhang¹, Huixuan Zhang¹, Xinyu Hu¹, Li Lin¹
Mingqi Gao¹, Shi Qiu², Xiaojun Wan¹

¹Wangxuan Institute of Computer Technology, Peking University

²School of Physics, Peking University

{junzhezhang, zhanghuixuan, efsotr_l, shiqiu}@stu.pku.edu.cn

{huxinyu, gaomingqi, wanxiaojun}@pku.edu.cn

Abstract

Evaluation is important for multimodal generation tasks, while traditional multimodal evaluation metrics suffer from several limitations. With the rapid progress of MLLMs, there is growing interest in applying MLLMs to build general evaluation systems. However, existing researches often simply collect large-scale evaluation data for training, while overlooking the quality of evaluation data. What's more, current proposed evaluation models often struggle to achieve consistently strong performance across both image-to-text (I2T) and text-to-image (T2I) tasks. In this paper, through rigorous quality control strategies, we construct a comprehensive multimodal evaluation dataset, Minos-57K, with evaluation samples across 15 datasets, for developing the multimodal evaluation model Minos with SFT and preference alignment training strategies. Notably, despite using less than half the scale of the training data of prior work, our model achieves state-of-the-art evaluation performance across 16 out-of-domain datasets covering both I2T and T2I tasks among all open-source multimodal evaluation models and remain competitive with closed-source models. Extensive experiments demonstrate the importance of leveraging quality control process, jointly training on evaluation data from both I2T and T2I generation tasks and further preference alignment.

1 Introduction

Multimodal evaluation (Huang and Zhang, 2024; Zhang et al., 2023; Ge et al., 2023) is crucial for multimodal generation tasks and developing multimodal models. A reliable evaluation not only enables more accurate comparison across models, but also plays a crucial role in the development of multimodal models. Although multimodal evaluation is crucial, traditional metrics still face notable limitations (Hessel et al., 2022; Mañas et al., 2024), such as correlating poorly with

human judgments, requiring reference data, and task-dependent. Therefore, developing a general-purpose multimodal evaluation system is becoming increasingly important.

With the rapid development of multimodal large language models (MLLMs), recent studies (Chen et al., 2024; Lee et al., 2024; Xiong et al., 2024; Wang et al., 2025) begin to explore applying MLLMs for building general multimodal evaluation systems. For example, LLaVA-Critic (Xiong et al., 2024) collects existing multimodal generation data and prompts the GPT-4o (OpenAI et al., 2024) to obtain evaluation results among various image-to-text (I2T) generation tasks to train a general evaluation model. UnifiedReward (Wang et al., 2025) applies data of LLaVA-Critic and other open-source multimodal evaluation data to train a unified multimodal reward model across multiple modalities. However, they simply collected large-scale evaluation data from diverse tasks for training, without subjecting it to rigorous quality control.

We argue that multimodal evaluation models should be developed using high-quality evaluation data curated through a strict and systematic quality control process, which is as important as the task diversity of evaluation data. Following this insight, we first try to collect existing human-annotated evaluation datasets on multi-directional multimodal generation tasks including both I2T and T2I tasks as high-quality evaluation data sources. However, human-annotated evaluation data remain scarce and often provide only evaluation scores without clear analyses, and many multimodal tasks even lack human evaluation data. Therefore, we manually define evaluation guidelines and annotate additional evaluation data for more multimodal tasks. After obtaining the annotated large-scale evaluation data, different from previous researches (Xiong et al., 2024; Wang et al., 2025), instead of directly using them for training, we conduct strict quality control from both the instance-level and dataset-level per-

spectives to construct our evaluation dataset Minos-57K. Minos-57K is a multimodal evaluation dataset constructed from 15 different multimodal datasets including both I2T and T2I.

Moreover, prior trained multimodal evaluation models such as LLaVA-Critic and UnifiedReward typically rely solely on supervised fine-tuning(SFT) using evaluation data, while overlooking the alignment stage(Ouyang et al., 2022; Lee et al., 2023) that plays a crucial role in MLLMs development. Inspired by this, we further construct multimodal preference dataset, thereby enabling Direct Preference Optimization(DPO)(Rafailov et al., 2023) alignment of the multimodal evaluation model. Building upon the SFT evaluation data quality control strategy, we further apply more strict filtering to obtain the high-quality evaluation preference dataset Minos-DPO-5.8K.

After SFT on Minos-57K and DPO alignment on Minos-DPO-5.8K with Qwen3-VL-8B(Bai et al., 2025) backbone, we obtain our multimodal evaluation model Minos. Our proposed Minos is capable of evaluating diverse multimodal generation tasks across T2I and I2T tasks (**Modality Generalization**) by providing reference-free scores (**Independence**) and generating human-interpretable analysis (**Interpretability**), highlighting its practical value. Minos outperforms previous open-source state-of-the-art evaluation models, and is competitive or even surpasses the performance of some closed-source models. Our experiments further demonstrate that, evaluation data quality and task diversity are the key factors influencing multimodal evaluation capability.

Overall, our main contributions are as follows¹:

- We construct **Minos-57K**, a large-scale multimodal evaluation dataset across 15 datasets, including both T2I and I2T generation tasks, with rigorous construction procedure and strict filtering process. We further construct high-quality multimodal evaluation preference dataset **Minos-DPO-5.8K**.
- We propose Minos, a multimodal evaluation model with modality generalization, independence and interpretability, trained using SFT and DPO alignment strategy on Minos-57K and Minos-DPO-5.8K. Minos achieves **state-of-the-art** performance among all open-source evaluation models, and is competitive

or even surpasses the performance of closed-source models.

- Extensive experimental results demonstrate that the quality, task diversity of evaluation data, preference alignment, and evaluation analysis together contribute to improved evaluation capability in evaluation models.

2 Related Work

2.1 MLLM as a Judge

As multimodal large models (MLLMs) are increasingly employed to construct evaluation metrics across various tasks(Huang and Zhang, 2024; Xia et al., 2024), building a unified evaluation model for multiple multimodal tasks based on MLLMs has become a promising direction. The MLLM-as-a-Judge(Chen et al., 2024) benchmark provides human-annotated evaluation data spanning 14 tasks and evaluates the performance of both open-source and closed-source MLLMs as evaluation models. Prometheus-V(Lee et al., 2024) was the first to leverage MLLM to construct a dedicated multimodal evaluation model. LLaVA-Critic(Xiong et al., 2024) further collected a range of pairwise and pointwise evaluation data across multiple tasks annotated by GPT, and trained a larger-scale multimodal evaluation model. Recent studies have proposed several multimodal reward models(Zhang et al., 2025; Wang et al., 2025), which are predominantly designed for pairwise preference evaluation. Among them, UnifiedReward natively supports both pointwise and pairwise evaluation. However, for the development of these pointwise evaluation models, they often simply applying evaluation data for training without strict filtering process. Existing multimodal evaluation models(Lee et al., 2024; Xiong et al., 2024; Wang et al., 2025) lack further DPO alignment after SFT stage, which could further improve the evaluation capacity of model. Main difference between Minos and previous evaluation models are listed in table 1.

2.2 Multimodal Human Evaluation

With the rapid advancement of multimodal research, increasing attention has been paid to the evaluation of multimodal tasks. Early works, such as CLIP-Score(Hessel et al.), introduced human evaluation datasets for image captioning. However, these early datasets often exhibited diverse and inconsistent formats, making them difficult to consolidate into a unified training resource for evaluation

¹Our code is available at <https://github.com/reroze/MINOS>

Method	Task Diversity	Task-Specific Guideline	Instance-level Filtering	Dataset-level Balancing	Preference Alignment
Prometheus-Vision	I2T	Generated	✗	✓	✗
LLaVA-Critic	I2T	✗	✗	✗	✗
UnifiedReward	Multi-Directional	No I2T	✗	✗	✗
Minos	Multi-Directional	✓	✓	✓	✓

Table 1: Comparison of Minos and previous researches of mainstream multimodal evaluation models.

models. More recently, several studies(Wada et al., 2024; Mañas et al., 2024; Xu et al., 2023; Liang et al., 2024) have collected human-annotated multimodal evaluation data across a variety of image-to-text and text-to-image tasks. For I2T evaluation, Polaris(Wada et al., 2024) introduced the image captioning dataset, comprising 131k human ratings annotated by 550 unique annotators. Similarly, LAVE(Mañas et al., 2024) proposed a human evaluation dataset for visual question answering, which includes 29k human-labeled instances. On the T2I side, datasets such as ImageReward(Xu et al., 2023), RichHF-18K(Liang et al., 2024) and GenAI-Bench(Jiang et al., 2024) have been developed to support human evaluation. Nonetheless, despite these efforts, high-quality and large-scale human evaluation datasets remain limited, many multimodal tasks still lack sufficient human evaluation data.

3 Minos

To develop multimodal evaluation model Minos, we first build structured evaluation inputs from raw multimodal task inputs and model responses with pre-defined evaluation guidelines, which provide task description and evaluation criterion, following our general multimodal evaluation data definition. We apply GPT-4o to generate evaluation output candidates and apply Consistency Filtering and Score Balancing to improve the quality of evaluation data, yielding the high-quality evaluation dataset Minos-57K. We then derive a preference alignment dataset by constructing evaluation preference pairs from GPT-annotated candidates, and perform Delta Score Filtering to produce preference dataset Minos-DPO-5.8K. Minos is developed in two stages: SFT(supervised finetuning) on Minos-57K, and DPO alignment training on Minos-DPO-5.8K. Construction workflow is illustrated in the figure 1.

3.1 Data Definition

We first introduce a more general definition of multimodal evaluation instance based on previous I2T evaluation data definition in LLaVA-Critic(Xiong et al., 2024). Specifically, LLaVA-Critic define a evaluation instance as (Image, Question, Response, Reference, Evaluation Criterion, Score, Reason). We generalize the image and question fields into a unified formulation consisting of a task description d , a task input query q , and a model output generation g , which allows for a consistent representation of evaluation data across both image-to-text (I2T) and text-to-image (T2I) tasks. Overall, a multimodal evaluation instance consists of a task description d , a task input query q , a model output generation g , an evaluation criterion k , and an optional reference answer r . The corresponding output of the evaluation instance includes an evaluation analysis a and a 1–5 Likert scale pointwise evaluation score s . A single multimodal evaluation instance can be represented as: $(q, d, g, k, [r], a, s)$. This standardized format provides a flexible and unified structure for representing evaluation of diverse multimodal generation tasks. An specific example can be seen in the figure 2 in appendix C.

3.2 Data Collection

We start the data construction process by collecting raw multimodal task input query q and its corresponding model output generation g . We first extract $\langle q, g \rangle$ from existing high-quality human annotated evaluation dataset on both I2T and T2I tasks. We investigate three large-scale human-annotated evaluation datasets, including Image Captioning dataset Polaris(Wada et al., 2024), Visual Question Answering dataset LAVE(Mañas et al., 2024) and Text-to-Image Generation dataset ImageReward(Xu et al., 2023). However, these evaluation data only contains scores without corresponding evaluation analyses which are an important component of the evaluation data, and

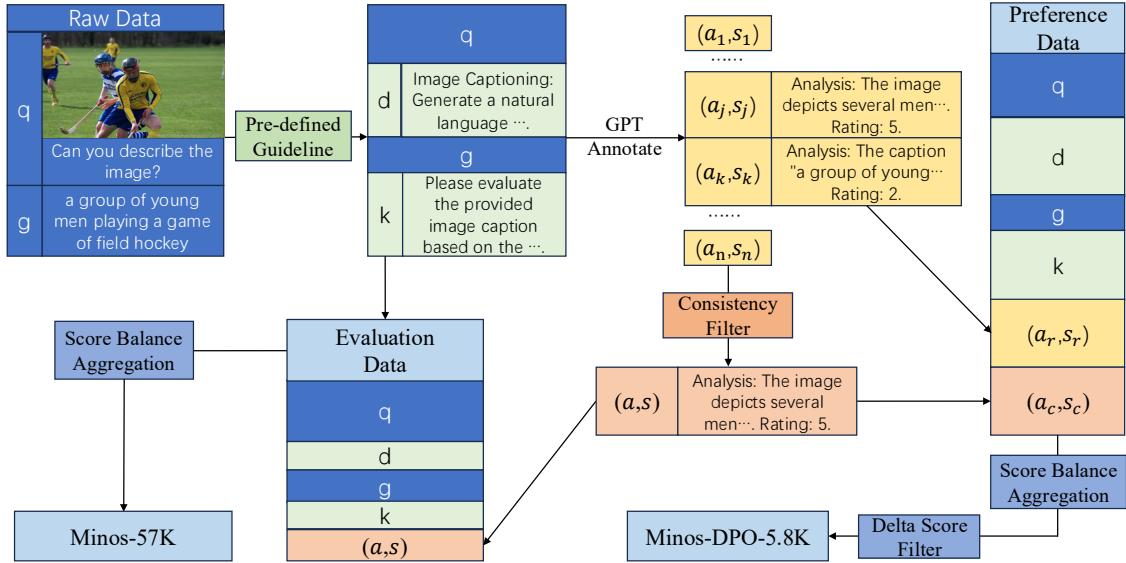


Figure 1: An overview of our dataset construction pipeline for Minos-57K and Minos-DPO-5.8K.

many other tasks even lack corresponding human-annotated evaluation datasets. Therefore, we first extract some of the model responses from VLFeedback(Li et al., 2024b) which contains responses across multiple multimodal tasks from different MLLMs, following previous settings(Xiong et al., 2024). We design the evaluation guidelines for different multimodal tasks to form evaluation inputs. We then use GPT-4o to generate 10 evaluation output candidates which contain both analyses and scores for these evaluation inputs. We obtain 48k raw multi-response annotations for human-annotated evaluation datasets and 76k raw multi-response annotations on other tasks. In the end, we obtain Minos-Raw which contains 124k raw multi-response annotations, each annotation containing the corresponding evaluation input and a set of candidate evaluation outputs. Details about the raw dataset are provided in table 9 in appendix B.

3.3 Quality Control

Several existing multimodal evaluation models(Xiong et al., 2024; Wang et al., 2025) focuses on increasing the variety and scale of evaluation data, while paying limited attention to rigorous evaluation data quality control. Based on prior research(Hu et al., 2024), we argue that evaluation performance depends more on data quality than on data scale, high-quality data Selection and Balance are critical for improving model evaluation capabilities.

Therefore, we further impose both instance-level

and dataset-level quality control strategies. We try to obtain high quality evaluation data through Pre-defined Guideline, Consistency Filter and Score Balance, separately, as shown in figure 1.

Pre-defined Guideline Multimodal evaluation models are typically expected to assess a wide range of multimodal tasks, which may differ substantially from one another. Even for human evaluators, assessing different multimodal tasks typically requires specialized training to understand both the target task and the corresponding evaluation criterion which explains how to evaluate the model response. We therefore believe that, when evaluating a specific multimodal task instance, a multimodal evaluation model must understand both the target task and its associated evaluation criterion. Accordingly, we design evaluation guidelines which consists task descriptions and evaluation criterion for each of the six multimodal task categories in our dataset. Detailed guidelines are provided in the appendix H.

Consistency Filter To enhance the evaluation capability of Minos, we first design Consistency Filter method to obtain high-quality evaluation data over Minos-Raw. Specifically, for those raw evaluation multi-response annotations with human evaluation scores, we sample one evaluation result where the GPT-generated score matches the human score from 10 generated evaluation results. If none of the ten GPT scores align with the human judgment, the instance is discarded. For rest raw evaluation multi-response annotations that lack human-labeled evaluation scores, we first find the mode of the score

Evaluation Source	Multimodal Task	Dataset	Data Size
Human	Image Captioning	Polaris	3.2K
	Visual Question Answering	LAVE	7.1K
	Text-to-Image Generation	ImageReward	10.8K
GPT-4o	Image Captioning	SViT-D, LLaVA-D	7.2K
	Visual Question Answering	LLaVAMed, LLaVA-C, comvint, SVIT-C	11.4k
	Text Reading	LLaVAR	4.8k
	Reasoning	LLaVA-R, SVIT-CR	7.8k
	Instruction Following	PCAEVAL, M3IT, LRV-Instruction	5.2k

Table 2: The Evaluation Source, Multimodal Task, Dataset, and corresponding Data Size of Minos-57K. More details about the Dataset can be seen in Table 9 in appendix B.

in GPT-generated output candidates as the labeled evaluation score, and randomly select one evaluation analysis whose corresponding score matches the mode. We end up obtain 102k evaluation samples after Consistency Filter stage.

Score Balance After applying the two preceding quality control procedures, we are able to ensure the quality of individual evaluation instances. However, when considered as a whole evaluation dataset, the dataset may still exhibit distributional imbalance. We first analyzed the overall score distribution of the combination of the consistency filtered dataset and observed a significant imbalance of evaluation scores as shown in Table 10. To address this issue, we manually balanced the score distribution by random removing some evaluation samples, resulting in the final evaluation dataset Minos-57K. The score distribution of the evaluation data can be seen in Table 10. More analysis about data can be seen in appendix B and I.

3.4 Minos Training

Firstly, we perform supervised-finetuning (SFT) using Minos-57K to construct an MLLM with basic evaluation capabilities. Further, for MLLM development, it is common to follow the SFT stage with an alignment phase, which often leads to further performance improvements. We believe that this observation also holds for training multimodal evaluation models. Therefore, we introduce DPO Alignment stage after SFT stage to further improve the evaluation ability of Minos.

To perform DPO alignment, we first obtain preference pairs from Minos-Raw. After applying the same quality control strategies, for each instance in the evaluation dataset, we apply the previously

selected annotation through Consistency Filter as the chosen data. Among the remaining evaluation output candidates for the same sample, we identify the one with the largest score discrepancy from the selected sample and treat it as the rejected evaluation data to form a comparison pair. We discard the instance, where all candidate evaluation scores are identical to the previously selected good evaluation data. We obtain 38k evaluation preference pairs in the end.

Unlike other tasks, the evaluation data contains evaluation scores. This enables us to quantify the strength of preference by comparing the score difference between the chosen and rejected instances within each preference pair. We compute the score gap between the chosen and rejected instances, and retain only pairs with a score difference no less than 2. The resulting 5.8K high-confidence preference pairs are used as the final dataset Minos-DPO-5.8K for DPO Alignment. Through further DPO alignment on Minos-DPO-5.8K after SFT on Minos-57K, we obtain our multimodal evaluation model Minos. More training details can be seen in appendix D.

4 Experiments

4.1 Experimental Setup

Benchmark Following the evaluation protocol established in LLaVA-Critic(Xiong et al., 2024), we adopt MLLM-as-a-Judge(Chen et al., 2024) to assess the generalization performance of our evaluation model on various out-of-domain test datasets. MLLM-as-a-Judge consists of 5k evaluation samples spanning 14 datasets. More experimental details can be seen in appendix A. Since MLLM-as-a-Judge doesn’t contain T2I generation task, we

Model	Scale	MLLM-as-a-Judge														RichHF	GenAI	All Ave.
		CO	C.C.	Dif	Graph	Math	Text	WIT	Chart	Vis	CC	M2W	Sci	Aes	MM			
Gemini-2.5-Pro	/	38.9	40.7	43.4	56.8	49.5	59.6	34.0	60.7	50.1	22.7	-0.5	40.0	20.0	37.3	39.7	70.3	41.5
GPT-4o*	/	39.6	45.2	34.1	46.4	46.0	56.4	40.8	57.3	58.9	30.5	26.2	56.9	42.1	34.2	31.1	60.9	44.2
LLaVA-OV*	7B	22.4	2.4	6.30	18.9	9.70	26.5	-13.5	27.4	22.7	8.10	3.0	26.1	24.9	26.2	5.85	16.4	14.6
LLaVA-OV*	72B	26.4	39.0	4.6	26.2	35.8	32.7	19.5	29.0	41.5	14.4	35.9	26.7	44.4	25.3	27.2	51.6	30.0
Qwen3-VL	8B	26.4	37.3	28.5	57.0	49.0	51.2	43.2	56.4	43.3	27.2	-0.1	39.6	24.8	29.3	38.9	61.6	38.4
Prometheus-V*	7B	28.9	34.2	10.6	17.2	18.2	21.4	20.9	22.4	22.6	22.8	8.90	17.4	36.8	15.7	8.19	18.6	20.3
LLaVA-Critic*	7B	38.2	45.0	10.3	31.6	35.6	37.8	17.9	42.1	32.2	24.6	30.1	26.9	39.5	27.2	18.4	33.0	30.7
LLaVA-Critic*	72B	33.3	46.3	14.6	45.2	47.4	55.9	39.6	54.5	48.8	27.3	25.9	33.4	40.3	37.4	33.0	53.2	39.8
UnifiedReward_L	7B	25.0	37.8	19.6	34.4	40.5	44.0	22.6	39.5	35.9	22.9	22.0	36.9	32.4	21.8	39.9	62.6	33.6
UnifiedReward_Q	8B	29.3	35.1	22.3	44.0	46.6	43.4	29.0	55.3	37.9	25.0	17.2	43.1	30.6	33.6	40.4	62.7	37.2
Minos	8B	32.8	41.1	25.1	54.1	51.3	58.3	41.4	62.1	47.1	24.3	23.6	50.7	28.7	39.8	36.0	60.2	42.3

Table 3: Main Result of Minos and other evaluation models on MLLM-as-a-Judge, RichHF-18K and GenAI-Bench. We present the **pearson-r** between the evaluation scores of models and the evaluation scores of human. We report results across two model categories: closed-source models, open-source MLLMs. We include results (which are referred as *) from previous researches(Chen et al., 2024; Xiong et al., 2024), and additionally evaluate all models on the text-to-image evaluation dataset RichHF-18K and GenAI-Bench. In each categories, we highlight in bold the model that achieves the highest consistency with human evaluations.

additionally select two extra text-to-image generation evaluation dataset RichHF-18K(Liang et al., 2024) and GenAI-Bench(Jiang et al., 2024) as test data for text-to-image generation task. More details about data integrity can be seen in appendix K. Following previous setting(Xiong et al., 2024), we apply Pearson-r to measure the consistency between the model’s evaluation scores and human evaluation scores. We also show the results with correlation coefficient Kendall’s Tau, details can be seen in appendix F.

Baselines We selected Gemini-2.5-Pro(Comanici et al., 2025) and GPT-4o(OpenAI et al., 2024) to evaluate the performance of closed-source models. For open-source models, we evaluated base MLLMs of different scales, including LLaVA-OV-7B, LLaVA-OV-72B(Li et al., 2024a) and Qwen3-VL-8B(Bai et al., 2025), as well as prior multimodal evaluation models such as Prometheus-V(Lee et al., 2024) and LLaVA-Critic(Xiong et al., 2024). Specifically, Prometheus-V is built upon LLaVA-v1.5-7B(Liu et al., 2023a), while LLaVA-Critic includes both versions based on a 7B and a 72B LLaVA-OV. Recent work has introduced several multimodal reward models such as UnifiedReward(Wang et al., 2025), but most of them only support pairwise evaluation. We select UnifiedReward(Wang et al., 2025), which natively supports pointwise scor-

ing, as a representative reward model for testing. We evaluate both the earlier version of the UnifiedReward presented in its paper, which is built on LLaVA-OV-7B(UnifiedReward_L), as well as the recently released version based on Qwen3-VL-8B(UnifiedReward_Q).

4.2 Main Results

The main results of Minos with other models on different multimodal tasks can be seen in Table 3. We evaluate our model against closed-source models, open-source MLLMs built on different backbones. As shown in the results, our model obtains the highest performance on the average of all multimodal generation tasks and achieves the highest agreement with human evaluations across 16 diverse multimodal evaluation tasks among open-source models. Compared with closed-source models, Minos achieves higher average evaluation consistency than Gemini-2.5-Pro, only performing slightly below GPT-4o with an average gap of 1.9 pearson-r on all multimodal generation tasks. In summary, Minos achieve state of the art on average evaluation consistency among open-source evaluation models, outperforming all existing open-source models. Compared to prior SoTA of open-source evaluation model LLaVA-Critic(72B), Minos achieves an average improvement of 2.5 Pearson-r on all multimodal generation tasks with smaller back-

bone. Compared to prior SoTA of open-source evaluation model UnifiedReward_Q with the same scale, Minos achieves an average improvement of 5.1 Pearson-r on all multimodal generation tasks. More results of T2I specialized evaluation models on T2I evaluation dataset can be seen in table 11 in appendix G.

4.3 Ablation Study

Ablation Study of Quality Control We conduct an ablation study to analyze the impact of quality control on the training data, and the results are presented in the table 4. Starting from the Minos-Raw dataset, we first randomly sample one output from 10 evaluation output candidates to construct each evaluation sample, without applying any explicit evaluation criterion and task description when building the evaluation sample. We use such samples to simulate and evaluate a low-control setting, where the evaluation data lacks predefined evaluation guidelines, Consistency Filtering, and Score Balancing constraints. This corresponds to the configuration reported in the first row of the table 4.

We sequentially obtain the Minos-57K by progressively enabling the three modules of our quality control, Pre-defined Guideline, Consistency Filter, and Score Balancing. After full quality control is applied, the resulting dataset contains less than half of Minos-RAW and the data used in prior work, which is 113K in LLaVA-Critic(Xiong et al., 2024) and 236K in UnifiedReward(Wang et al., 2025). However, the experimental findings indicate that the quality of evaluation data is more critical than its scale. Minos trained on the 57k quality-controlled evaluation samples exhibits stronger evaluation capability than Minos trained on the larger, unfiltered raw dataset. As more quality control methods are applied, the dataset scale generally decreases, yet the model’s average evaluation capability consistently improves across multiple multimodal tasks. More importantly, our experiments show that training an evaluation model on large-scale but low-quality evaluation data may harm its original evaluation capability, potentially making it underperform the base model prior to evaluation model training. This finding further highlights the critical importance of evaluation data quality.

Ablation Study of DPO Alignment When developing large language models, a common practice is to perform preference alignment training after SFT (Supervised Fine-Tuning), which typically leads to further improvements in model capability. How-

Score Balance	Consistency Filter	Evaluation Guideline	SFT Data Size	All Ave.
✗	✗	✗	124k	36.3
✗	✗	✓	124k	37.1
✗	✓	✓	102k	39.0
✓	✓	✓	57k	40.9

Table 4: Result of Minos with pearson-r on MLLM-as-a-Judge, RichHF-18K and GenAI-Bench when training with different multimodal evaluation Data during SFT stage.

DPO	Delta Score Filter	DPO Data Size	All Ave.
✗	✗	0k	40.9
✓	✗	35k	40.1
✓	✓	5.8k	42.3

Table 5: Result of Minos with pearson-r on MLLM-as-a-Judge, RichHF-18K and GenAI-Bench when training with different dpo settings. 0k means original SFT model.

ever, prior multimodal evaluation models, such as LLaVA-Critic and UnifiedReward, are trained solely via supervised fine-tuning on multimodal evaluation data, without applying subsequent alignment training such as DPO (Direct Preference Optimization) on evaluation model itself.

Based on result of our ablation study, we identify a potential explanation: naively constructing preference pairs and simply applying DPO alignment does not always yield clear gains in evaluation capability. In some cases, applying DPO alignment may even degrade the evaluation model’s evaluation performance relative to the model before alignment training, such as 40.9 to 40.1 in table 5. However, this does not imply that preference alignment is a wrong approach for training multimodal evaluation models. Our experimental results demonstrate that when we leverage the evaluation score of evaluation data to further filter DPO preference pairs by the score gap between chosen and rejected outputs, we can obtain a smaller set of high-quality preference pairs that provide a stronger alignment signal. Training Minos using these fewer but higher-quality DPO pairs leads to additional gains in evaluation capability, improving the average evaluation performance from 40.9 to 42.3 in table 5. The ablation results on DPO alignment further demonstrate

that evaluation performance is more sensitive to data quality than data scale, which holds for both the evaluation data used in SFT and the preference pairs used in DPO alignment. Full experimental results can be seen in appendix E.

5 Analysis

5.1 Evaluation between I2T and T2I

To analyze the impact of unifying I2T (Image-to-Text) and T2I (Text-to-Image) tasks, we conduct extra experiments on the T2I and I2T training data during the SFT stage, whose results are shown in table 6. We conduct extra experiments with three configurations: training on T2I-only evaluation data, training on I2T-only evaluation data with the same scale (10K samples), and training on full I2T-only evaluation data. 0k means original base model Qwen3-v1-8B. All these subsets are sampled from Minos-57K. In addition to reporting the overall average evaluation performance (All Ave.), we further compute the specific average performance on I2T and T2I tasks separately, namely I2T Ave. and T2I Ave.

The experimental results reveal that training on T2I evaluation data alone may impair the model’s evaluation capability on I2T tasks, with average performance on I2T tasks dropping from 36.7 to 25.4. A similar negative transfer is observed in the opposite direction, where training on I2T data only reduces T2I evaluation performance from 50.3 to 46.1. In contrast, when we perform joint training using both T2I and I2T evaluation data, the evaluation model exhibits mutual performance enhancement between the two task categories, suggesting that the training from T2I and I2T evaluation samples can be complementary rather than conflicting. Compared to the T2I-only training setting, adding 47K I2T evaluation samples yields a 1.8 improvement in average T2I evaluation capability. Likewise, compared to the 47k I2T-only training baseline, incorporating T2I evaluation data leads to a 4.2 improvement in average I2T evaluation performance. These findings indicate that building evaluation MLLMs capable of unified assessment across both T2I and I2T tasks provides further benefits than developing category-specific evaluation models.

5.2 Impact of Evaluation Analysis

Many previous researches (Xiong et al., 2024; Wang et al., 2025) treat rationale generation dur-

I2T	T2I	SFT Data Size	I2T Ave.	T2I Ave.	All Ave.
✗	✗	0k	36.7	50.3	38.4
✗	✓	10k	25.4	49.1	28.3
✓	✗	10k	34.1	46.2	35.6
✓	✗	47k	35.3	46.1	36.6
✓	✓	57k	39.5	50.9	40.9

Table 6: Results of Minos with pearson-r on MLLM-as-a-Judge, RichHF-18K and GenAI-Bench when training with different multimodal evaluation task only during SFT stage after applying quality control strategies.

ing evaluation as an auxiliary capability to justify the reasonableness of model evaluations. We conduct an additional experiment to analyze whether including evaluation analysis supervision in evaluation data affects the performance of the evaluation model’s final scoring behavior. As shown in the table 7, evaluation analysis generation not only improves the interpretability of model evaluations, but also enhances the agreement between the model’s final scores and human judgments.

More analysis can be seen in appendix J and L.

SFT Data	All Ave.
w/o evaluation analysis	38.8
w evaluation analysis	40.9

Table 7: Result of Minos with pearson-r on MLLM-as-a-Judge, RichHF-18K and GenAI-Bench when supervised finetuning with/without evaluation analysis on Minos-57K.

6 Conclusion

In this work, we first collect and construct a high-quality, general multimodal evaluation dataset: Minos-57K. Minos-57K comprises multimodal evaluation samples spanning 6 common tasks and 15 datasets, covering both image-to-text and text-to-image settings. Each sample is accompanied by evaluation analysis and score after strict quality control process. We further apply more strict filtering to obtain the high-quality evaluation preference dataset Minos-DPO-5.8K. We build our multimodal evaluation model Minos with supervised finetuning on Minos-57K and dpo alignment on Minos-DPO-5.8K. Averaged across all benchmark

tasks, Minos achieves state-of-the-art (SoTA) performance among all open-source evaluation models, and even outperforms several closed-source models. Extensive experimental results demonstrate that the quality, task diversity of evaluation data, preference alignment, and evaluation analysis supervision together contribute to improved evaluation capability in evaluation MLLMs.

Limitations

Some early multimodal human evaluation datasets are no longer accessible due to broken links, and more human-annotated datasets are continuously being proposed. Our work represents a snapshot collection of the currently available multimodal human evaluation datasets. As the field progresses, we anticipate the emergence of larger and higher-quality human-labeled datasets, which can support more reliable evaluation results and enable more comprehensive experimental analysis. Considering the limitations of computational resources, we did not build our method upon extremely large foundation models such as Qwen3-VL-70B. Although larger backbone models may yield stronger performance, they typically demand substantially higher inference compute and longer evaluation latency, which can diminish their practical utility in real-world deployment scenarios.

Ethical Considerations

We follow the correct usage of the data and models with the corresponding license. Since real-world tasks are much different from well-defined evaluation instance, we recommend conducting manual spot-checks of the model’s outputs during deployment on real-world tasks to ensure the reliability of its evaluations. Additionally, Minos is primarily developed as an evaluation model. If it is instead used as a reward model to optimize other models, it is important to consider whether this may lead to issues of over-optimization.

Acknowledgements

This work was supported by Beijing Natural Science Foundation (L253001), Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology) and National Engineering Research Center of New Electronic Publishing Technologies. We appreciate the anonymous reviewers for their

helpful comments. Xiaojun Wan is the contact author.

References

- Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. [no-caps: novel object captioning at scale](#). In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 8947–8956. IEEE.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond](#). *Preprint*, arXiv:2308.12966.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, and 1 others. 2025. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. 2024. [Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark](#). *arXiv preprint arXiv:2402.04788*.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. [Meshed-memory transformer for image captioning](#). In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587.
- Wentao Ge, Shunian Chen, Guiming Chen, Junying Chen, Zhihong Chen, Shuo Yan, Chenghao Zhu, Ziyue Lin, Wenya Xie, Xidong Wang, and 1 others. 2023. [Mllm-bench, evaluating multi-modal llms using gpt-4v](#). *arXiv preprint arXiv:2311.13951*.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. [Clipscore: A reference-free evaluation metric for image captioning](#). *Preprint*, arXiv:2104.08718.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. [Clipscore](#).

- Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024. [Themis: A reference-free NLG evaluation language model with flexibility and interpretability](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15924–15951, Miami, Florida, USA. Association for Computational Linguistics.
- Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.
- Jiaxing Huang and Jingyi Zhang. 2024. [A survey on evaluation of multimodal large language models](#). *Preprint*, arXiv:2408.15769.
- Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhua Chen. 2024. Genai arena: An open evaluation platform for generative models. *arXiv preprint arXiv:2406.04485*.
- Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023a. [Pick-a-pic: An open dataset of user preferences for text-to-image generation](#). *Preprint*, arXiv:2305.01569.
- Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023b. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.
- Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11286–11315.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. [Llava-onevision: Easy visual task transfer](#). *Preprint*, arXiv:2408.03326.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. 2024b. [Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment](#). *arXiv preprint arXiv:2410.09421*.
- Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. 2024. [Rich human feedback for text-to-image generation](#). *Preprint*, arXiv:2312.10240.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179.
- Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. 2022. Grit: Faster and better image captioning transformer using dual visual features. In *Computer Vision – ECCV 2022*, pages 167–184, Cham. Springer Nature Switzerland.
- OpenAI. 2023. [GPT-4V\(ision\) System Card](#).
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoní Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13559–13568.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, and 1 others. 2024. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. 2025. Unified reward model for multimodal understanding and generation. *Preprint*, arXiv:2503.05236.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2023. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 893–911.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, Lijuan Wang, and Huaxiu Yao. 2024. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. *arXiv preprint arXiv:2410.10139*.

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.

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.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023. Gpt-4v(ision) as a generalist evaluator for vision-language tasks. *Preprint*, arXiv:2311.01361.

Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, Haojie Ding, Jiankang Chen, Fan Yang, Zhang Zhang, Tingting Gao, and Liang Wang. 2025. R1-reward: Training multimodal reward model through stable reinforcement learning. *Preprint*, arXiv:2505.02835.

A Experimental Details

We present the full name of each task mentioned in Table 3. The full name remains the same as (Chen et al., 2024).

Task Name (Short)	Task Name (Full)
CO	MS COCO
C.C.	Conceptual Captions
Dif	DiffusionDB
Graph	InfographicVOA
Math	MathVista
Text	TextVOA
WIT	WIT
Chart	ChartOA
Vis	VisIT-Bench
CC	CC-3M Concept-balanced
M2W	Mind2Web
Sci	ScienceOA
Aes	AesBench
MM	MMvet

Table 8: The full name of each task.

B Dataset Details

Here, we present the details of the dataset distribution before and after filtering in table 9. The score distribution is shown in table 10.

Data Source	Task Type	Minos-RAW	Minos-57K
Polaris	Image Captioning	6.7k(5.4%)	3.2k(5.6%)
LAVE	Visual Question Answering	13.7k(11.0%)	7.1k(12.3%)
ImageReward	Text-to-Image Generation	27.4k(22.1%)	10.8k(18.8%)
Human-Eval	Mixed	47.8k(38.5%)	21.1k(36.6%)
SViT-detail	Image Captioning	4.9k(3.9%)	4.0k(6.9%)
LLaVA-detail	Image Captioning	5.8k(4.7%)	3.2k(5.6%)
LLaVAMed	Visual Question Answering	5.7k(4.6%)	2.7k(4.7%)
LLaVA-conversation	Visual Question Answering	7.4k(6.0%)	2.2k(3.8%)
comvint	Visual Question Answering	7.7k(6.2%)	3.3k(5.7%)
SVIT-conversation	Visual Question Answering	7.6k(6.1%)	3.2k(5.6%)
LLaVAR	Text Reading	13.5k(10.9%)	4.8k(8.3%)
LLaVA-reasoning	Reasoning	7.7k(6.2%)	4.3k(7.5%)
SVIT-complex_reasoning	Reasoning	7.6k(6.1%)	3.5k(6.1%)
PCAEVAL	instruction following	0.8k(0.6%)	0.4k(0.7%)
M3IT	instruction following	1.5k(1.2%)	0.8k(1.4%)
LRV-Instruction	instruction following	6.3k(5.1%)	4.0k(6.9%)
GPT-Eval	Mixed	76.5k(61.6%)	36.5k(63.4%)
All	Mixed	124.2k	57.6k

Table 9: The Data Source, Task Type and corresponding Data Size of Minos-57K and Minos-RAW are shown in table. Minos-RAW contains a total of 124k evaluation instances constructed from diverse sources. In Table 2, SViT-D refers to SViT-detail, SVIT-C refers to SViT-conversation, SVIT-CR refers to SVIT-complex_reasoning. LLaVA-D refers to LLaVA-detail, LLaVA-C refers to LLaVA-conversation, LLaVA-R refers to LLaVA-reasoning.

Score	#1	#2	#3	#4	#5	All
RAW*	9.7k	15k	16k	19k	64k	124k
	8%	12%	13%	16%	51%	100%
Consistency Filtered	9.2k	9.8k	12k	13k	57k	102k
	9%	10%	12%	13%	56%	100%
Final	9.2k	9.8k	12k	13k	14k	57k
	16%	17%	21%	23%	23%	100%

Table 10: Score distribution of RAW* and Final dataset. We apply Data Selection and Balance method to filter the Corpus. RAW* means we random selection one of ten GPT output candidates in Minos-RAW. We calculated the number and corresponding proportion of data samples for each score.

Metrics	Model	RichHF	GenAI
pearson-r	HPS v2	19.5	31.0
	PickScore	22.4	34.7
kendall’s tau	HPS v2	13.1	23.1
	PickScore	18.3	26.0

Table 11: Result of specialized T2I evaluation models on T2I evaluation dataset RichHF-18K and GenAI-Bench. We present the pearson-r and kendall’s tau between the evaluation scores of Minos and the evaluation scores of human.

Gold Score	#1	#2	#3	#4	#5
accuracy	17.1%	27.7%	57.9%	50.5%	4.4%

Table 12: The scoring accuracy of Minos conditioned on different gold score levels.

C Example of Evaluation Data

Figure 2 shows an example of Evaluation instance in Minos-57K.

D Training Details

We build our model on top of Qwen3-VL-8B. During the SFT stage, we train the model for 2 epochs using a batch size of 64 and a learning rate of $1e-5$. In the DPO stage, we train for 1 epoch with a learning rate of $2e-6$, setting $\beta = 0.03$ and $\gamma = 0$. All other training configurations follow the default settings of Qwen3-vl-8B. We train the model with BF16 precision on 4 H100 GPUs. The SFT stage takes approximately 10 hours, while the DPO stage takes around 2 hours.

E Full Evaluation Results

We present full evaluation results in table 13 (corresponding to table 4), table 14 (corresponding to table 5), table 15 (corresponding to table 6) and table 16 (corresponding to table 7).

F Results with kendall’s tau

We show the result of models in table 3 with correlation coefficient kendall’s tau. The results are shown in table 17. Since previous researches (Chen et al., 2024; Xiong et al., 2024) don’t include result with kendall’s tau, we reproduce all models to obtain evaluation results with kendall’s tau. We also present the pearson-r of our re-implemented models in table 18. For rigor, the results compared in the main text are primarily cited from previously published papers whenever possible.

G More Results with T2I evaluators

To enable a more comprehensive comparison among multimodal evaluation models, we additionally benchmarked specialized text-to-image (T2I) evaluation models, such as HPS v2 (Wu et al., 2023) and PickScore (Kirstain et al., 2023a), on out-of-domain T2I evaluation datasets RichHF-18K and GenAI-Bench. As shown in table 11, even a specialized T2I evaluation model does not surpass Minos in evaluation capability on T2I tasks, which further demonstrates the advantage of jointly training text-to-image and image-to-text evaluation abilities.

Setting	MLLM-as-a-Judge														RichHF	GenAI	All Ave.
	CO	C.C.	Dif	Graph	Math	Text	WIT	Chart	Vis	CC	M2W	Sci	Aes	MM			
$S_1(RAW)$	25.6	33.7	2.7	53.0	48.8	52.7	37.0	59.1	46.1	20.5	8.8	36.9	27.8	32.3	37.4	58.0	36.3
S_2	25.5	40.1	17.1	53.4	47.7	50.0	43.3	58.7	44.6	17.4	11.5	27.6	23.4	33.2	38.8	60.9	37.1
S_3	30.7	41.9	22.6	54.4	48.3	51.4	47.6	58.4	46.8	20.5	11.3	26.8	23.2	39.0	38.3	62.4	39.0
$S_4(Final)$	28.8	45.1	25.3	52.1	52.9	58.9	45.9	60.3	48.0	23.8	16.3	31.1	30.2	34.5	39.0	62.8	40.9

Table 13: Detailed results of data analysis in table 4. S_i represents the setting corresponding to the i -th row in table 4.

Setting	MLLM-as-a-Judge														RichHF	GenAI	All Ave.
	CO	C.C.	Dif	Graph	Math	Text	WIT	Chart	Vis	CC	M2W	Sci	Aes	MM			
S_1	28.8	45.1	25.3	52.1	52.9	58.9	45.9	60.3	48.0	23.8	16.3	31.1	30.2	34.5	39.0	62.8	40.9
S_2	24.8	39.9	24.7	52.0	52.0	54.1	35.9	61.0	40.0	28.1	18.3	46.1	33.0	38.5	36.8	57.0	40.1
S_3	32.8	41.1	25.1	54.1	51.3	58.3	41.4	62.1	47.1	24.3	23.6	50.7	28.7	39.8	36.0	60.2	42.3

Table 14: Detailed results of data analysis in table 5. S_i represents the setting corresponding to the i -th row in table 5.

Setting	MLLM-as-a-Judge														RichHF	GenAI	All Ave.
	CO	C.C.	Dif	Graph	Math	Text	WIT	Chart	Vis	CC	M2W	Sci	Aes	MM			
S_1	26.4	37.3	28.5	57.0	49.0	51.2	43.2	56.4	43.3	27.2	-0.1	39.6	24.8	29.3	38.9	61.6	38.4
S_2	22.7	25.8	12.6	24.3	41.5	36.6	15.2	30.1	34.8	19.6	20.9	14.6	16.8	39.4	40.3	57.9	28.3
S_3	27.3	37.9	22.2	34.8	47.7	50.2	42.6	47.8	41.8	20.2	17.5	21.2	27.3	38.5	37.9	54.5	35.6
S_4	23.7	45.8	27.8	31.2	48.0	51.6	43.0	44.3	40.1	25.5	20.7	20.2	27.6	44.0	36.5	55.6	36.6
S_5	28.8	45.1	25.3	52.1	52.9	58.9	45.9	60.3	48.0	23.8	16.3	31.1	30.2	34.5	39.0	62.8	40.9

Table 15: Detailed results of data analysis in table 6. S_i represents the setting corresponding to the i -th row in table 6.

Setting	MLLM-as-a-Judge														RichHF	GenAI	All Ave.
	CO	C.C.	Dif	Graph	Math	Text	WIT	Chart	Vis	CC	M2W	Sci	Aes	MM			
S_1	20.8	43.6	30.4	42.6	51.0	52.5	37.3	43.9	38.9	28.9	14.7	39.5	30.5	37.9	42.6	65.1	38.8
S_2	28.8	45.1	25.3	52.1	52.9	58.9	45.9	60.3	48.0	23.8	16.3	31.1	30.2	34.5	39.0	62.8	40.9

Table 16: Detailed results of data analysis in table 7. S_i represents the setting corresponding to the i -th row in table 7.

Model	MLLM-as-a-Judge														RichHF	GenAI	All Ave.
	CO	C.C.	Dif	Graph	Math	Text	WIT	Chart	Vis	CC	M2W	Sci	Aes	MM			
Gemini-2.5-Pro	29.9	34.1	36.0	48.1	42.1	47.4	28.6	51.2	40.5	18.9	-0.2	34.5	18.8	23.2	33.9	61.6	34.3
GPT-4o	23.1	28.8	16.1	36.8	35.9	39.6	30.3	47.2	31.9	18.3	1.7	28.8	20.4	29.0	27.1	53.6	29.3
LLaVA-OV(7B)	15.3	16.9	10.4	12.9	11.5	19.8	3.3	28.2	19.6	15.9	17.5	6.6	26.9	26.0	3.15	13.6	15.5
LLaVA-OV(72B)	20.2	13.3	7.5	14.2	24.1	22.2	4.6	19.7	27.2	21.9	19.0	13.8	19.6	26.9	24.3	46.3	20.3
Qwen3-VL(8B)	17.3	31.2	23.6	49.1	40.9	41.0	37.9	49.3	36.4	22.0	-0.7	31.7	22.9	22.6	33.7	53.6	32.0
Prometheus-V(7B)	11.2	9.10	9.40	12.7	11.1	17.8	18.0	8.40	11.5	15.5	11.4	8.20	24.1	11.0	6.55	16.0	12.6
LLaVA-Critic(7B)	19.2	31.6	10.3	30.1	23.3	29.1	8.3	25.0	25.4	15.6	14.8	7.7	28.6	25.3	16.8	29.8	21.3
LLaVA-Critic(72B)	30.3	33.8	2.9	34.8	38.1	37.4	13.5	43.1	34.8	20.4	7.6	25.6	17.0	29.3	28.9	46.8	27.8
UnifiedReward_L(7B)	23.7	31.9	13.8	29.5	34.2	35.0	20.2	35.7	30.0	18.3	12.5	31.4	27.0	14.9	33.4	54.4	27.9
UnifiedReward_Q(8B)	26.9	30.3	17.6	37.7	37.8	35.4	29.5	48.7	29.2	21.4	7.9	34.8	20.9	28.3	33.6	54.7	30.9
Minos(8B)	24.9	35.7	23.1	48.5	41.4	44.5	36.6	50.9	35.0	25.2	12.0	38.7	23.8	25.8	31.5	55.9	34.6

Table 17: Main Result of Minos and other evaluation models on MLLM-as-a-Judge, RichHF-18K and GenAI-Bench. We present the **kendall’s tau** between the evaluation scores of models and the evaluation scores of human. We report results across two model categories: closed-source models, open-source MLLMs. Since previous researches(Chen et al., 2024; Xiong et al., 2024) don’t include result with kendall’s tau, we reproduce all models to obtain evaluation results with kendall’s tau. For models in each categories, we highlight in bold the model that achieves the highest consistency with human evaluations.

Model	MLLM-as-a-Judge														RichHF	GenAI	All Ave.
	CO	C.C.	Dif	Graph	Math	Text	WIT	Chart	Vis	CC	M2W	Sci	Aes	MM			
Gemini-2.5-Pro	38.9	40.7	43.4	56.8	49.5	59.6	34.0	60.7	50.1	22.7	-0.5	40.0	20.0	37.3	39.7	70.3	41.5
GPT-4o	30.1	33.0	22.2	44.3	44.3	52.8	32.8	57.4	40.1	22.4	6.5	36.2	29.3	31.2	31.1	60.9	35.9
LLaVA-OV(7B)	20.6	19.4	13.6	21.1	14.8	28.2	4.9	30.3	29.6	16.9	34.2	13.5	42.3	39.1	5.85	16.4	21.9
LLaVA-OV(72B)	27.5	12.4	9.7	20.1	31.6	31.3	8.9	26.0	35.1	25.6	31.5	27.2	26.3	37.2	27.2	51.6	26.8
Qwen3-VL(8B)	26.4	37.3	28.5	57.0	49.0	51.2	43.2	56.4	43.3	27.2	-0.1	39.6	24.8	29.3	38.9	61.6	38.4
Prometheus-V(7B)	13.4	10.3	12.2	15.4	13.7	21.3	19.7	13.5	14.2	16.1	18.1	11.9	36.3	15.8	8.19	18.6	16.2
LLaVA-Critic(7B)	25.9	34.3	15.1	37.9	28.5	38.4	17.6	32.0	34.1	20.1	23.1	15.6	39.2	32.3	18.4	33.0	27.8
LLaVA-Critic(72B)	37.0	36.0	8.6	43.1	45.7	46.8	20.7	53.0	44.0	26.5	14.7	35.2	27.1	42.2	33.0	53.2	35.4
UnifiedReward_L(7B)	25.0	37.8	19.6	34.4	40.5	44.0	22.6	39.5	35.9	22.9	22.0	36.9	32.4	21.8	39.9	62.6	33.6
UnifiedReward_Q(8B)	29.3	35.1	22.3	44.0	46.6	43.4	29.0	55.3	37.9	25.0	17.2	43.1	30.6	33.6	40.4	62.7	37.2
Minos(8B)	32.8	41.1	25.1	54.1	51.3	58.3	41.4	62.1	47.1	24.3	23.6	50.7	28.7	39.8	36.0	60.2	42.3

Table 18: Main Result of Minos and other evaluation models on MLLM-as-a-Judge, RichHF-18K and GenAI-Bench. We present the **pearson-r** between the evaluation scores of models and the evaluation scores of human. We report results across two model categories: closed-source models, open-source MLLMs. We reproduce all models to obtain evaluation results with pearson-r. For models in each categories, we highlight in bold the model that achieves the highest consistency with human evaluations.

H Detailed Evaluation Guidelines

We present the evaluation guidelines for six distinct tasks, as shown in the figure 3, 4, 5, 6, 7 and 8.

I Human Study

To validate the effectiveness of our filtering pipeline, we randomly sampled 50 evaluation instances from the filtered GPT-generated data and conducted human assessment. First, we measured the agreement between GPT scores and human scores on this subset, obtaining a consistency of 63.9, indicating that our filtering strategy successfully retains samples with relatively reliable judgments.

We further evaluated the quality of the generated rationales along three dimensions: Factual Accuracy, Relevance to the Evaluation Criterion, and Coherence. Under a 1–5 scoring scale, the average scores are 4.62, 4.64, and 4.56, respectively. These results demonstrate that the filtered synthetic data maintains high-quality rationales in addition to reasonably aligned scoring. The evaluation guidelines can be seen in figure 9. Annotators were informed about the purpose of the task and consented to perform the annotation. The task did not involve sensitive or personal data.

J Analysis of Failure Modes

We further analyze Minos’s scoring accuracy conditioned on different gold score levels. The detailed results are shown in the table 12. It can be observed that Minos achieves relatively high accuracy on samples with gold scores of 3 and 4, while its accuracy is lower for scores of 1 and 5, particularly for score 5. Through case studies, we find that Minos tends to be conservative when assigning a model response corresponding to perfect score 5, especially for complex tasks. It is often reluctant to give full marks even when the output is of high quality, which may explain a substantial portion of the observed errors and point out promising direction for developing better multimodal evaluation models. We show an example of this failure from MLLM-as-a-Judge in figure 10.

Meanwhile, we observe that Minos exhibits relatively low evaluation consistency on certain tasks, such as AesBench. We select a failure case for analysis, as illustrated in the figure 11. In the example, Minos correctly identifies some inaccuracies in the model’s response; however, it fails to point out other aspects that cannot be reliably inferred, such

as the description about the “man,” which is not clearly supported by the image. Moreover, even when the model’s response contains partially incorrect analysis, Minos still assigns a score of 4. This suggests that, in certain instances, the identification of errors does not substantially affect its final scoring. Such behavior is inconsistent with the evaluation criterion of “how well the analysis captures key elements.” A possible explanation is that the task itself involves evaluating the model’s evaluation of image aesthetic quality, which may introduce additional complexity and challenge compared to other non-evaluation tasks.

K Data Integrity

The definition of an evaluation sample follows Section 3.1: (q, d, g, k, [r], a, s). Our training data consists of human-annotated evaluation datasets and GPT-4o-annotated data. Since all references in the test sets are human scores, we only need to verify whether the human-annotated evaluation data in training overlaps with the human-annotated data in testing. The human scores (s) in both training and test sets are independently annotated in their respective original papers.

We further detail the data sources. MLLM_Judge(Chen et al., 2024): MLLM_Judge generates model responses mostly using LLaVA-1.5-13b, LLaVA-1.6-34b(Liu et al., 2023b), CogVLM(Wang et al., 2024), GPT-4V(OpenAI, 2023), Qwen-VL-Max(Bai et al., 2023) and Gemini(Team et al., 2023). Therefore, many evaluation instance in our human-annotated evaluation data of training dataset differs from this test data in terms of model response (g) which comes from different applied models.

Polaris(Wada et al., 2024): Polaris uses MSCOCO(Lin et al., 2014) and NoCaps(Agrawal et al., 2019) and generates model responses with SAT(Xu et al., 2015), M^2 -Transformer(Cornia et al., 2020), VinVL(Zhang et al., 2021), GRIT(Nguyen et al., 2022), BLIP(Li et al., 2022), and OFA(Wang et al., 2022) for captioning tasks. While MLLM_Judge also contains data derived from MSCOCO, those correspond to VQA tasks rather than Polaris’s captioning setting. Moreover, the models used to generate responses are different, resulting in different (g), (a), and (s).

LAVE(Mañías et al., 2024): LAVE collects human judgments for answers generated by BLIP-2(Li et al., 2023), PromptCap(Hu et al., 2022) and

BLIP(Li et al., 2022). Since most of the underlying generation models differ, the corresponding responses (g), annotations (a), and scores (s) are also different.

ImageReward(Xu et al., 2023): ImageReward uses real user prompts from DiffusionDB(Wang et al., 2023). In contrast, our test set RichHF(Liang et al., 2024) is derived from the Pick-a-Pic(Kirstain et al., 2023b) dataset, annotated via the Pick-a-Pic web application. Additionally, GenAI-Bench(Jiang et al., 2024) consists of 1,600 challenging real-world prompts sourced from professional designers. These data sources are distinct from ImageReward, and thus do not overlap. Although Dif dataset in MLLM_Judge apply data from DiffusionDB, it is used for I2T task which is different from T2I task in ImageReward.

L Perturbation Experiments

We additionally conduct a common multimodal adversarial experiment. Specifically, we randomly replace the input images to construct corrupted samples and evaluate Minos under this setting. After random image replacement, Minos’s overall scoring pearson-r consistency drops sharply to 8.5. This substantial degradation indicates that Minos is highly sensitive to visual information and does not rely solely on textual cues. Instead, it performs evaluation by jointly considering both the input image and the text, demonstrating genuine multimodal evaluation behavior.

We further conduct a perturbation experiment on the MLLM_Judge test set. Specifically, we randomly replace 25% of the spans in the evaluated model responses to construct corrupted versions. We then compare the scores assigned by Minos before and after perturbation. The results show that 71.27% of the perturbed responses receive lower scores by Minos than their original counterparts. This indicates that Minos is able to distinguish between higher and lower quality responses and is sensitive to content degradation, rather than assigning scores arbitrarily.

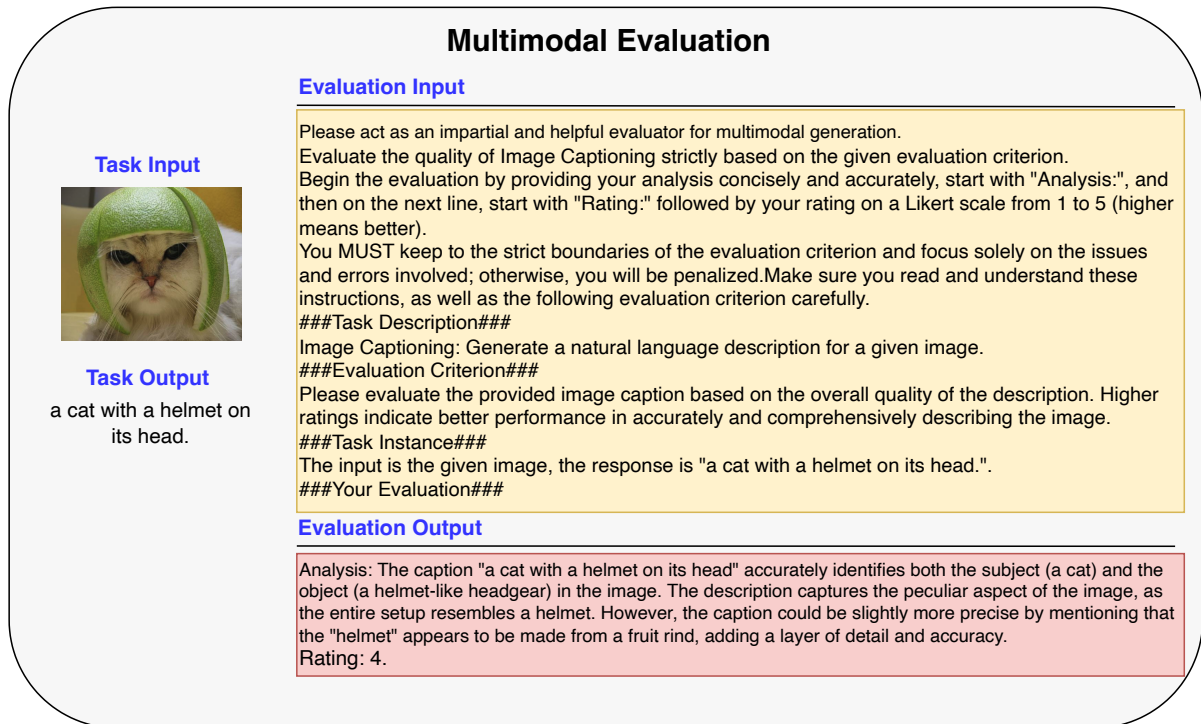


Figure 2: An example of Evaluation instance in Minos-57K. We select a sample constructed from a image captioning evaluation dataset Polaris to illustrate our data. Depending on the specific multimodal task, the task input can be an image, text, or a combination of both, while the task output can be either text or an image.

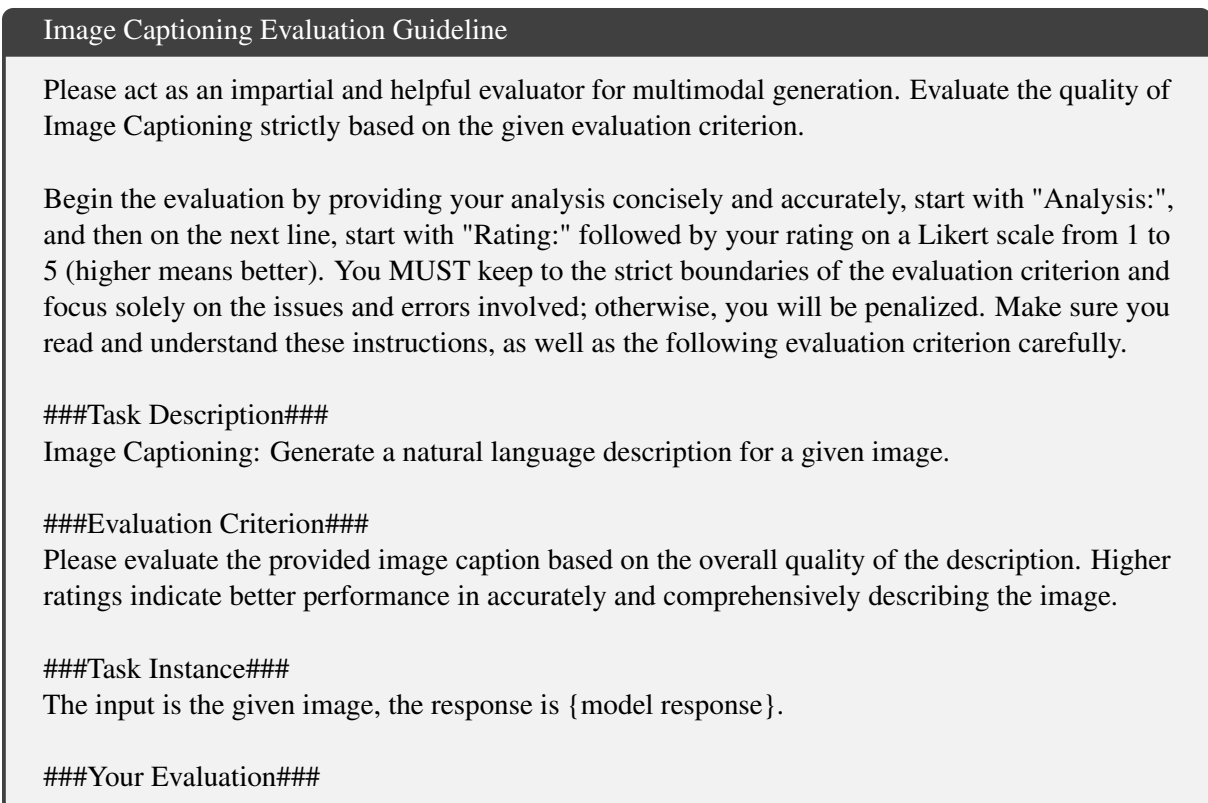


Figure 3: Image Captioning Evaluation Guideline.

Visual Question Answering Evaluation Guideline

Please act as an impartial and helpful evaluator for multimodal generation. Evaluate the quality of Visual Question Answering strictly based on the given evaluation criterion.

Begin the evaluation by providing your analysis concisely and accurately, start with "Analysis:", and then on the next line, start with "Rating:" followed by your rating on a Likert scale from 1 to 5 (higher means better). You MUST keep to the strict boundaries of the evaluation criterion and focus solely on the issues and errors involved; otherwise, you will be penalized. Make sure you read and understand these instructions, as well as the following evaluation criterion carefully.

###Task Description###

Visual Question Answering: Answer the natural language question based on the content of a given image.

###Evaluation Criterion###

Please evaluate the provided answer to the visual question based on the overall quality of the response, considering its helpfulness, relevance, and accuracy. Higher ratings indicate better performance in addressing the question effectively and correctly.

###Task Instance###

The input image is the given image, the question is {VQA question}, the response is {model response}.

###Your Evaluation###

Figure 4: Visual Question Answering Evaluation Guideline.

Image Generation Evaluation Guideline

Please act as an impartial and helpful evaluator for multimodal generation. Evaluate the quality of Image Generation strictly based on the given evaluation criterion.

Begin the evaluation by providing your analysis concisely and accurately, start with "Analysis:", and then on the next line, start with "Rating:" followed by your rating on a Likert scale from 1 to 5 (higher means better). You MUST keep to the strict boundaries of the evaluation criterion and focus solely on the issues and errors involved; otherwise, you will be penalized. Make sure you read and understand these instructions, as well as the following evaluation criterion carefully.

###Task Description###

Image Generation: Generate the image following the input text.

###Evaluation Criterion###

Please evaluate the generated image based on how well the generated image matches the given text description and the overall quality of the image.

###Task Instance###

The input text is {input text}, the generated response image is the given image.

###Your Evaluation###

Figure 5: Image Generation Evaluation Guideline.

Text Reading Evaluation Guideline

Please act as an impartial and helpful evaluator for multimodal generation. Evaluate the quality of Text Reading strictly based on the given evaluation criterion.

Begin the evaluation by providing your analysis concisely and accurately, start with "Analysis:", and then on the next line, start with "Rating:" followed by your rating on a Likert scale from 1 to 5 (higher means better). You MUST keep to the strict boundaries of the evaluation criterion and focus solely on the issues and errors involved; otherwise, you will be penalized. Make sure you read and understand these instructions, as well as the following evaluation criterion carefully.

###Task Description###

Text Reading: Given an image containing text information and a question related to text information in the image, analyze this figure in detail and answer the question.

###Evaluation Criterion###

Please evaluate the provided answer to the visual question based on the overall quality of the response, considering its helpfulness, relevance, and accuracy. Higher ratings indicate better performance in addressing the question effectively and correctly.

###Task Instance###

The input image is the given image, the question is {input question}, the response is {model response}.

###Your Evaluation###

Figure 6: Text Reading Evaluation Guideline.

Reasoning Evaluation Guideline

Please act as an impartial and helpful evaluator for multimodal generation. Evaluate the quality of Reasoning strictly based on the given evaluation criterion.

Begin the evaluation by providing your analysis concisely and accurately, start with "Analysis:", and then on the next line, start with "Rating:" followed by your rating on a Likert scale from 1 to 5 (higher means better). You MUST keep to the strict boundaries of the evaluation criterion and focus solely on the issues and errors involved; otherwise, you will be penalized. Make sure you read and understand these instructions, as well as the following evaluation criterion carefully.

###Task Description###

Reasoning: Given an image with a question, generate a detailed analysis with the answer of the question.

###Evaluation Criterion###

Please evaluate the provided answer focusing on the alignment and coherence of the model's reasoning concerning the image. The evaluation measures the extent to which the explanation responds effectively to the question, accurately represents the image's content, and provides a useful answer.

###Task Instance###

The input image is the given image, the question is {input question}, the response is {model response}.

###Your Evaluation###

Figure 7: Reasoning Evaluation Guideline.

Instruction Following Evaluation Guideline

Please act as an impartial and helpful evaluator for multimodal generation. Evaluate the quality of Instruction Following strictly based on the given evaluation criterion.

Begin the evaluation by providing your analysis concisely and accurately, start with "Analysis:", and then on the next line, start with "Rating:" followed by your rating on a Likert scale from 1 to 5 (higher means better). You MUST keep to the strict boundaries of the evaluation criterion and focus solely on the issues and errors involved; otherwise, you will be penalized. Make sure you read and understand these instructions, as well as the following evaluation criterion carefully.

###Task Description###

Instruction Following: Given an image and text instructions, generate responses following the instructions.

###Evaluation Criterion###

Please evaluate mainly based on whether the response is informative, and whether the response contains any hallucination. Hallucination, in this context, refers to a situation where the MLLM generates a response that includes information not present or implied in the image or previous conversation. A hallucination could be a false claim about an object, action, emotion, or any other detail that is not grounded in the image.

###Task Instance###

The input image is the given image, the input instruction is {input instruction}, the response is {model response}.

###Your Evaluation###

Figure 8: Instruction Following Evaluation Guideline.

Human Evaluation Guideline of Rationale

Please evaluate the evaluation rationales from the following three perspectives.

(1) Factual Accuracy

Definition: Measures whether the rationale is factually correct with respect to the input (image and/or text) and free from hallucinations.

Annotation Instructions: Annotators verify that all claims in the rationale are supported by the input and do not contradict observable content.

Scoring:

- 5 – Fully accurate; no factual errors.
- 4 – Minor inaccuracies but overall correct.
- 3 – Some noticeable unsupported claims.
- 2 – Multiple factual errors.
- 1 – Largely incorrect or hallucinated.

(2) Relevance to Evaluation Criteria

Definition: Measures whether the rationale properly addresses the specified evaluation criteria and justifies the assigned score.

Annotation Instructions: Annotators assess whether the explanation focuses on the required evaluation dimensions and avoids irrelevant commentary.

Scoring:

- 5 – Directly and clearly aligned with the evaluation criteria.
- 4 – Mostly aligned with minor irrelevance.
- 3 – Partially aligned but incomplete or unfocused.
- 2 – Weak connection to the evaluation criteria.
- 1 – Not aligned with the evaluation criteria.

(3) Coherence

Definition: Measures the clarity, logical flow, and internal consistency of the rationale.

Annotation Instructions: Annotators evaluate whether the explanation is well-structured, logically connected, and easy to follow.


Scoring:

- 5 – Clear, well-organized, and logically consistent.
- 4 – Generally clear with minor structural issues.
- 3 – Understandable but somewhat disorganized.
- 2 – Difficult to follow.
- 1 – Incoherent or contradictory.

Figure 9: Human Evaluation Guideline of Rationale.

Please act as an impartial and helpful evaluator for multimodal generation. Your task is to evaluate the quality of Visual Question Answering strictly based on the given evaluation criterion. Begin the evaluation by providing your analysis concisely and accurately, start with "Analysis:", and then on the next line, start with "Rating:" followed by your rating on a Likert scale from 1 to 5 (higher means better). You MUST keep to the strict boundaries of the evaluation criterion and focus solely on the issues and errors involved; otherwise, you will be penalized. Make sure you read and understand these instructions, as well as the following evaluation criterion carefully.

###Task Description###
 Visual Question Answering: Answer the natural language question based on the content of a given image.
 ###Evaluation Criterion###
 Please evaluate the provided answer to the visual question based on the overall quality of the response, considering its helpfulness, relevance, and accuracy. Higher ratings indicate better performance in addressing the question effectively and correctly.
 ###Task Instance###
 The input image is the given image, the question is "What time of day might this be?", the response is "The image depicts a beach scene during sunset. This conclusion is based on several visual cues. Firstly, the sun is near the horizon, which typically indicates either sunrise or sunset. However, the warm hues and the position of the shadows suggest that this is likely sunset rather than sunrise. The sky is filled with orange and yellow colors which are more commonly associated with the evening. Additionally, the people on the beach are casting long shadows towards the camera, indicating that the light source (the sun) is behind them, which is consistent with the sun setting over the ocean. The calmness of the scene with a person walking a dog and another person sitting on the shore gives a leisurely impression that is often associated with the evening hours. Finally, the presence of the bird flying in the sky adds to the serene end-of-day atmosphere."
 ###Your Evaluation###




Minos:
 Analysis: The response provides a detailed and thoughtful explanation for identifying the time of day in the image. It correctly identifies the scene as a beach during sunset, supported by visual cues such as the sun's position near the horizon, the warm hues in the sky, and the long shadows cast by the figures. These observations are accurate and relevant to the question. However, the response overcomplicates the explanation by including speculative reasoning (e.g., associating the calmness of the scene with the evening) and unnecessary details (e.g., the bird flying in the sky), which slightly detract from the focus on the primary task. Despite this, the response remains highly helpful and accurate overall.
 Rating: 4.

Figure 10: Score discrepancy in an example from MLLM-as-a-Judge: Human Evaluation (5) compared with Minos Evaluation (4).

Please act as an impartial and helpful evaluator for multimodal generation. Your task is to evaluate the quality of Image Aesthetics Evaluation strictly based on the given evaluation criterion. Begin the evaluation by providing your analysis concisely and accurately, start with "Analysis:", and then on the next line, start with "Rating:" followed by your rating on a Likert scale from 1 to 5 (higher means better). You MUST keep to the strict boundaries of the evaluation criterion and focus solely on the issues and errors involved; otherwise, you will be penalized. Make sure you read and understand these instructions, as well as the following evaluation criterion carefully.

###Task Description###
 Image Aesthetics Evaluation: Given an image and its aesthetic quality evaluation, generate reasons of this evaluation.
 ###Evaluation Criterion###
 Please evaluate the provided answer focusing on the model's ability to accurately assess the aesthetic quality of an image. The evaluation measures how well the analysis captures key elements, the depth of explanation, and the overall alignment with the aesthetic quality of an image.
 ###Task Instance###
 The input image is the given image, the aesthetic quality of the image is "high", the response is "The image is a painting of a coastal scene. The sky is a light blue and the ocean is a deep blue. The waves are crashing against the rocks on the shore. There are two people walking on the beach. The woman is wearing a red shirt and the man is wearing a blue shirt. The dog is running ahead of them. The painting is done in a realistic style and the colors are vibrant. The artist has used a variety of brushstrokes to create texture and depth in the painting. The overall effect of the painting is one of peace and tranquility."
 ###Your Evaluation###



Minos:
 Analysis: The response provides a detailed description of the image, capturing several elements such as the coastal scene, the colors of the sky and ocean, the presence of people and a dog, and the realistic style of the painting. However, while the description is thorough, it does not explicitly address the aesthetic quality of the image as "high." The analysis mentions the vibrant colors and texture created by brushstrokes, which align with aesthetic qualities, but it does not sufficiently tie these observations to an overarching assessment of high aesthetic quality. Additionally, some details in the response, such as the claim that the dog is "running ahead of them," are inaccurate based on the image. This detracts from the overall alignment and accuracy of the evaluation.
 Rating: 4.

Figure 11: Score discrepancy in an example from AesBench: Human Evaluation (2) compared with Minos Evaluation (4).