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

Anonymous ACL submission

Abstract

Evaluation is important for multimodal generation tasks. With the rapid progress of MLLMs, there is growing interest in applying MLLMs to build general evaluation systems. However, existing work overlooks two aspects: (1) the development of evaluation capabilities for text-to-image (T2I) generation task, and (2) the incorporation of large-scale human evaluation data. In this paper, we introduce Minos-Corpus, a large-scale multimodal evaluation dataset that combines evaluation data from both human and GPT. The corpus contains evaluation data across both image-to-text(I2T) and T2I generation tasks. Based on this corpus, we propose Data Selection and Balance, Mix-SFT training methods, and apply DPO to develop Minos, a multimodal evaluation model built upon a 7B backbone. Minos achieves state-of-the-art (SoTA) performance among all open-source evaluation models of similar scale on the average of evaluation performance on all tasks, and outperforms all open-source and closed-source models on evaluation of T2I generation task. Extensive experiments demonstrate the importance of leveraging high-quality human evaluation data and jointly training on evaluation data from both I2T and T2I generation tasks.

1 Introduction

Multimodal evaluation(Huang and Zhang, 2024; Zhang et al., 2023; Ge et al., 2023) is crucial for multimodal generation tasks and developing mutlimodal models. A reliable evaluation not only enables more accurate comparison across models, but also reduces the cost of human annotation, facilitates the generation of higher-quality synthetic data, and supports simulation of human feedback during alignment. Although multimodal evaluation is crucial, traditional metrics, though widely used, still face notable limitations(Hessel et al.; 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. However, developing such evaluation system often faces several challenges, such as limited generalizability across diverse multimodal tasks, the scarcity of human-annotated data, and the high cost of data annotation.

With the rapid development of multimodal large language models (MLLMs), recent studies(Chen et al., 2024; Lee et al., 2024; Xiong et al., 2024) begin to explore applying MLLMs as the foundation for building general multimodal evaluation systems. For example, MLLM-as-a-Judge(Chen et al., 2024) follows the LLM-as-a-Judge(Zheng et al., 2023) paradigm to construct a benchmark for multimodal evaluation on many multimodal tasks, testing prompt-based methods across a range of open-source and closed-source models. LLaVA-Critic(Xiong et al., 2024) collects existed multimodal generation data and prompts the GPT to obtain evaluation results among various image-to-text(I2T) generation tasks to train a general evaluation model.

However, the training and testing data used in existing work(Chen et al., 2024; Lee et al., 2024; Xiong et al., 2024) are limited to image-to-text generation tasks such as image captioning, visual question answering, and instruction following. Previous works have not sufficiently considered supporting another important multimodal task which is text-to-image(T2I) generation task. In addition, current multimodal evaluation models typically rely on synthetic labels and have not applied large-scale high-quality human evaluation data. Most training data employs annotations produced by GPT, essentially serving as a distillation of GPT's own judgments. Though LLaVA-Critic applies small amount of pairwise human evaluation data, developing effective evaluation models may require large-scale pointwise evaluation data (Hu et al., 2024), which is overlooked in previous works. Drawing inspiration from textual evaluation models such as Themis(Hu et al., 2024), we argue that the role of large-scale human evaluations in multimodal settings might be important as well.

To alleviate the scarcity of high-quality evaluation data across various multimodal tasks, we introduce Minos-Corpus, a multimodal evaluation corpus covering 16 datasets across 6 common multimodal tasks. Minos-Corpus contains 124k evaluation samples, all annotated by GPT-4o, among which 48k also include human annotations. Each data instance comprises an evaluation input and annotated outputs, which include both analyses and scores. Specifically, for instances without human scores, GPT-4o provides both scores and analyses; for those with human scores, GPT-4o generates only the corresponding analyses. Minos-Corpus incorporates high-quality human evaluation data and expands multimodal evaluation to cover both image-to-text and text-to-image generation tasks, improving generality and data diversity.

Moreover, prior multimodal evaluation models such as LLaVA-Critic typically rely on supervised fine-tuning(SFT) using evaluation data, while overlooking the alignment stage that plays a crucial role in MLLMs development. Alignment techniques such as RLHF(Ouyang et al., 2022) and RLAIF(Lee et al., 2023) have proven effective in aligning models with human intent. In particular, Direct Preference Optimization(DPO)(Rafailov et al., 2023), which leverages human or model-generated preference data, has become a widely adopted method for aligning LLMs. Inspired by this, we further utilize Minos-Corpus to construct preference data from existing human-annotated and GPT-annotated evaluation data, thereby enabling DPO alignment of the evaluation model. Moreover, we observe that preference over evaluation outputs is the pairwise evaluation of the evaluation output, constituting another form of the evaluation task. Therefore, to make more effective usage of preference signals, we construct pairwise comparison data from them and mix the resulting comparison data with the original training data during the SFT stage as Mix-SFT training. We find that this Mix-SFT training strategy further improves the model's evaluation performance on various multimodal tasks. And after Mix-SFT training, DPO can still be applied for alignment training to further improve model performance, providing our final proposed model **Minos**.

Our proposed Minos is an evaluation model on a 7B backbone designed for the evaluation of bidirectional (I2T and T2I) multimodal generation. Our model is capable of evaluating diverse multimodal generation tasks across text-to-image(T2I) and image-to-text(I2T) tasks(**Modality Generalization**) by providing reference-free(**Independence**) scores and generating human-interpretable analysis(**Interpretability**), highlighting its practical value.

Overall, our main contributions are as follows:

- We construct Minos-Corpus, a large-scale multimodal evaluation dataset comprising 124k evaluation samples across 16 datasets and 6 multimodal tasks, including I2T generation task, using human annotations and GPT-4o.
- We propose Minos, a multimodal evaluation model with modality generalization, independence and interpretability, trained using a Mix-SFT training and DPO alignment strategy on Minos-Corpus. Minos outperforms previous state-of-the-art models of the same parameter size, and even surpasses the performance of prior SoTA models GPT-4o on evaluating T2I generation tasks.
- We conduct extensive experiments demonstrating that introducing human annotated data and incorporating multimodal tasks(including T2I and I2T) help enhance evaluation capability, pointing out future research directions.

2 Minos-Corpus

There are a wide range of tasks and associated datasets in multimodal area. We start constructing Minos-Corpus by collecting existing human evaluation data. For tasks lacking human annotations, we leverage GPT-4o to generate corresponding evaluation annotations.

2.1 Data Format

Following previous settings in LLaVA-Critic(Xiong et al., 2024), a multimodal evaluation instance consists of the task input i, the task description d, the model output t, the evaluation criteria c, and an optional reference answer r. The corresponding output of the evaluation instance includes an analysis a and a score s. A single

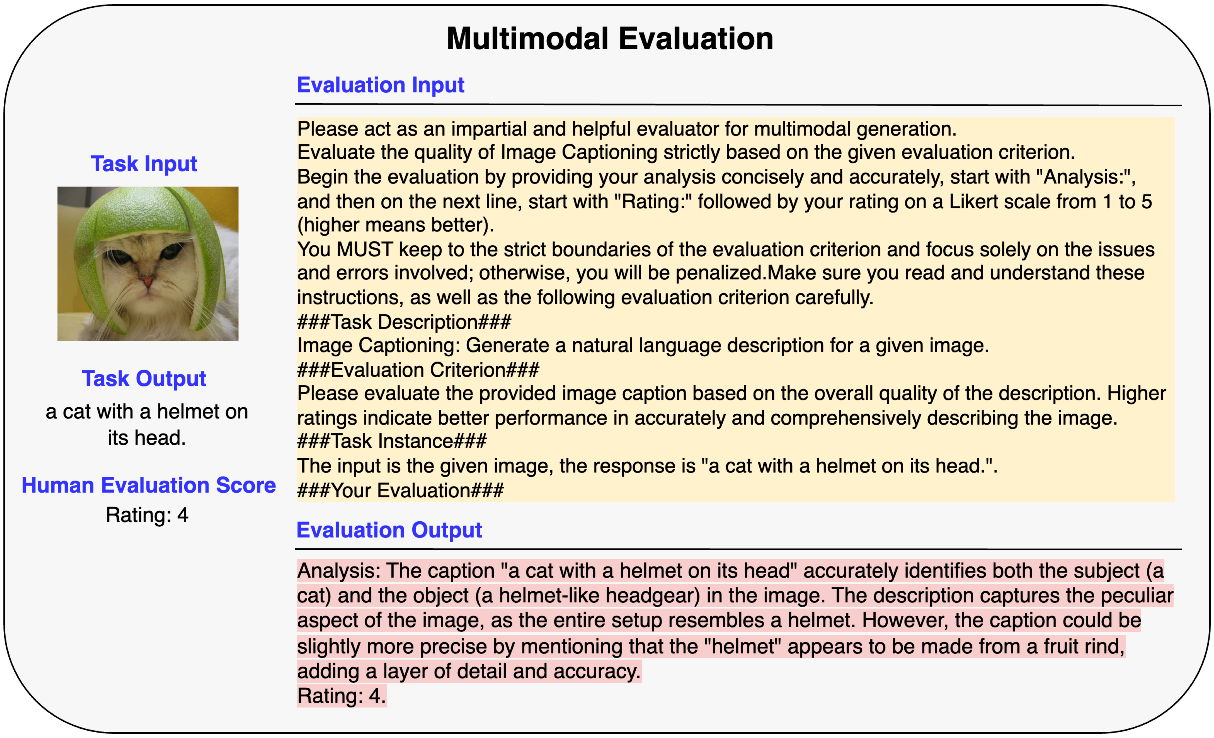

Figure 1: An example of Evaluation instance in Minos-Corpus. We select a sample from the human-annotated image captioning evaluation dataset Polaris to illustrate our data. Note that only instances from human evaluation data contain human evaluation score like presented in the figure. Depending on the 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.

multimodal evalution instance can be represented as: $(i, d, t, c, [r], a, s)$

This standardized format provides a flexible and unified structure for representing evaluation of diverse multimodal generation tasks, enabling consistent model training and assessment across different modalities and objectives. An example of Minos-Corpus evaluation data can be seen in the Figure 1.

2.2 Corpus Construction

We start the data construction process by collecting human annotated evaluation dataset. We collect four large-scale human-annotation evaluation dataset which are 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), RichHF-18K(Liang et al., 2024), which we collect nearly 60k evaluation samples. However, this subset of evaluation data only contains human-provided scores without corresponding analyses. To address this, we follow the prompt design in their original paper and use GPT-4o to generate 10 candidate evaluation outputs which contain both analyses and scores. There are 48k samples left which obtain valid GPT-annotated candidate evaluation outputs.

However, many tasks lack corresponding human-annotated evaluation datasets. Following previous settings(Xiong et al., 2024), to cover more multimodal tasks, we additionally extract relevant samples from several commonly used datasets and obtain model responses from VLFeedback(Li et al., 2024b). We then use GPT-4o to generate 10 evaluation output candidates which contain both analysis and scores for these instances. After applying the selection and balance method, we obtain 76k evaluation samples in the end. The entire Minos-Corpus consists of 124k evaluation samples, each containing the corresponding evaluation input and a set of candidate outputs. Details regarding the source of evaluation, task types, and data source are provided in Table 1.

3 Minos

We designed two main methods to develop our multimodal evaluation model Minos. We first design a Selection and Balancing strategy to filter and refine multimodal evaluation data, aiming to obtain a smaller but higher-quality dataset. Then, we consider the evaluation comparisons as a form of pair-

Evaluation Source	Multimodal Task	Data Source	Data Size
Human	Image Captioning	Polaris	6.7k
	Visual Question Answering	LAVE	13.7k
	Text-to-Image Generation	ImageReward,RichHF-18K	27.4k
GPT-4o	Image Captioning	SViT-detail, LLaVA	10.8k
	Visual Question Answering	LLaVAMed, LLaVA, comvint, SVIT	28.4k
	Text Reading	LLaVAR	13.5k
	Reasoning	LLaVA, SVIT	15.3k
	instruction following	PCAEVAL, M3IT, LRV-Instruction	8.5k

Table 1: The Evaluation Sources, contained Multimodal Tasks, Data Sources,, and corresponding Data Sizes of Minos-Corpus. Minos-Corpus contains a total of 124k evaluation instances constructed from diverse sources.

wise evaluation data and construct corresponding evaluation comparison data. We mix these pairwise samples with the evaluation data for Mix-SFT training. After Mix-SFT training, we apply DPO alignment to obtain final evaluation model Minos.

3.1 Selection and Balance

According to prior research on textual evaluation models such as Themis(Hu et al., 2024), high-quality data Selection and Balance are critical for improving model evaluation capabilities. In the multimodal setting, although LLaVA-Critic(Xiong et al., 2024) explores the usage of evaluation data, it simply samples a subset of evaluiton instances without performing in-depth filtering or balancing method. To enhance the evaluation capability of Minos, we design a Selection and Balance method over the constructed 124k evaluation samples, aiming to achieve better evaluation performance with fewer but higher-quality data.

First, we try to select one evaluation output from the the GPT-generated evaluation output candidates. For human-annotated instances which only contain evaluation scores, we perform quality filtering by comparing the GPT-generated scores with human evaluation score, resulting in a filtered subset. We refer this quality filtering to as Human Selection. Specifically, we sample one result where the GPT-assigned 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 those datasets that lack human-labeled evaluation data, we randomly select a GPT-generated evaluation output whose score matches the mode of the score in GPT-generated candidates.

Score	#1	#2	#3	#4	#5	All
Full	9.5k	15k	16k	18k	65k	124k
	7%	12%	13%	15%	53%	100%
Filtered	9.2k	9.8k	12k	13k	14k	57k
	16%	17%	21%	23%	23%	100%

Table 2: Score distribution of Full and Filtered Corpus. We apply Data Selection and Balance method to filter the Corpus. We calculated the number and corresponding proportion of data samples for each score.

We then analyzed the overall score distribution of the combination of two filtered datasets and observed a significant imbalance of evaluation scores. To address this issue, we manually balanced the score distribution, resulting in the final evaluation data. The score distribution of the Full and Filtered data can be seen in Table 2.

3.2 Mix-SFT Training and DPO Alignment

For MLLM development, it is common to follow the supervised fine-tuning (SFT) stage with an alignment phase, which often leads to further performance improvements. However, preference data in evaluation tasks is more special, as it reflects pairwise judgments over candidate evaluation results. It can also be regarded as a form of evaluation task in itself.

Inspired by this observation, we not only constructed preference data for alignment but also explicitly formulated them as evaluation comparison samples which can be seen in the bottom of the Figure 2.

Specifically, for each instance on the filtered evaluation dataset, we select the previously chosen an-

Model	MLLM-as-a-Judge															RichHF	All
	CO	C.C.	Dif	Graph	Math	Text	WIT	Chart	Vis	CC	M2W	Sci	Aes	MM	Ave.		Ave.
Gemini-Pro-1.5	32.2	38.6	29.4	34.6	45.2	41.9	43.6	43.3	34.7	18.4	5.8	27.2	28.4	32.9	32.6	52.5	33.9
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	43.9	54.5	44.6
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	28.7	48.8	30.0
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	39.3	51.0	40.1
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	15.1	21.4	15.5
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	21.3	28.8	21.8
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	31.4	32.0	31.4
Minos(7B)	24.5	38.4	**32.5**	**34.9**	**43.4**	**47.8**	**37.0**	**46.9**	**42.9**	24.4	23.5	**33.4**	31.7	**36.1**	**35.5**	**57.9**	**37.0**

Table 3: Main Result of Minos and other evaluators on MLLM-as-a-Judge and RichHF-18K. We present the pearson-r between the evaluation scores of models with the evaluation scores of human. We report results across three model categories: closed models, open-source MLLMs built on large-scale backbones, and open-source MLLMs based on 7B backbones. We include results from previous researches(Chen et al., 2024; Xiong et al., 2024), and additionally evaluate all models on the text-to-image evaluation dataset RichHF. For models based on 7B backbones, we highlight in bold the model that achieves the highest consistency with human evaluations.

notation as the good evaluation sample. Among the remaining GPT-annotated 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 bad evaluation sample to form a comparison pair. We discard the instance, where all GPT-annotated scores are identical to the previously selected good evaluation sample. We obtain 38k evaluation comparison samples in the end.

We mix these evaluation comparison data with evaluation data during SFT training of our model, refered it as Mix-SFT. During Alignment stage, we also transform the comparison data into preference data for DPO alignment training by selecting the good evaluation as preferred data and bad evaluation as dispreferred data.

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 datasets. MLLM-as-a-Judge consists of 5k evaluation samples spanning 14 datasets across 9 tasks. Since MLLM-as-a-Judge doesn't contain T2I generation task, We additionally sample 600 examples with balanced score distribution from RichHF-18K as test data for the text-to-image generation task. The T2I generation evaluation data of our training data contain samples from ImageReward only and doesn't contain samples from RichHF-18K. 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.

Baselines We selected Gemini-Pro-1.5(Team et al., 2024) and GPT-4o(Achiam et al., 2023) to evaluate the performance of closed-source models across multiple multimodal tasks. For open-source models, we evaluated general-purpose multimodal models of different scales, including LLaVA-OV (7B) and LLaVA-OV (72B), 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., 2023), while LLaVA-Critic includes both version based on a 7B and a 72B LLaVA-OV(Li et al., 2024a).

Training Details We build our model on top of LLaVA-OneVision (LLaVA-OV)-7B(Li et al., 2024a) as the backbone. During the SFT stage, we train the model for 2 epochs using a batch size of 192 and a learning rate of 1e-5. In the DPO stage, we train for 1 epoch with a learning rate of 5e-7, setting $\beta = 0.1$ and $\gamma = 0$. Throughout both stages, we freeze the vision encoder and update only the multimodal adapter and the language model components of LLaVA-OV. All other training configurations follow the default settings of LLaVA-OV-7B. We train the model with BF16 precision on 4 H100 GPUs. The SFT stage takes approximately 10 hours, while the DPO stage takes around 5 hours.

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 72B backbones, and open-source MLLMs based on 7B backbones. As shown in the results, our model consistently obtains the highest evaluation on the average of all multimodal generation tasks and achieves the highest agreement with human evaluations across 10 diverse multimodal evluation tasks among models with 7B backbones. Notably, on the RichHF-18K dataset for text-to-image generation evaluation, our model Minos largely outperforms all previous closed source and open source models. One of a multimodal tasks in MLLM-as-a-Judge is constructed on DiffusionDB, which is related to text-to-image generation task (though not text-to-image task). Our model outperforms all open-source models and ranks second only to GPT-4o, another evidence supporting our model's capability of evaluating text-to-image generation task.

In summary, Minos sets a new state of the art among 7B-scale evaluation models, outperforming all existing models across 10 datasets spanning 6 multimodal tasks. Compared to prior models of similar scale, Minos achieves an average improvement of 4.1 Pearson-r on image-to-text generation tasks and 5.6 Pearson-r on all multimodal generation tasks including T2I generation task.

Despite based on a 7B backbone, Minos still achieves competitive results when compared with closed-source models and open-source models built on much larger(72B) backbones. Minos outperforms the 72B general-purpose multimodal model LLaVA-OV, demonstrating that developing dedicated evaluation models remains important.

4.3 Analysis of Training Data During SFT

We first investigate the impact of applying different evaluation data as training data on SFT stage for multimodal evaluation models before incorporating Evaluation Comparison data. The Result can be seen in Table 4. Full results can be found in Appendix A.

Analysis of Selection and Balance Firstly, we would like to investigate the influence of our Selection and Balance on the corpus. At a starting point, we train a model using the full set of 124k samples, all of which rely solely on GPT-generated analyses and scores. Specifically, for each evaluation input, we randomly selected one of the GPT-generated outputs whose score matched the mode of the multiple GPT annotations as the evaluation output, with no selection and balance strategy applied. As can be seen from the results, despite the large data volume, this model did not yield the best evaluation performance, neither on image-to-text (I2T) tasks nor on text-to-image (T2I) tasks.

There are two components in our Selection and Balance method. We first apply only the Human Selection strategy on the full set, and we find that the Human Selection strategy only enhances the model's performance on only T2I generation evaluations. We then apply only the Score Balance Strategy and find that applying score balancing alone can slightly improve the model's evaluation capability on image-to-text (I2T) tasks, but it leads to a decline in performance on text-to-image (T2I) tasks. Finally, with the combination of Human Selection and Score Balance, the model achieves the highest evaluation consistency across Bidirectional Generation (T2I and I2T) compared with other settings. Notably, the model's performance under this setting on the T2I evaluation task even surpasses that of GPT-4o, which achieves a Pearson correlation of 54.5.

These results underscore the importance of high-quality, human-annotated multimodal evaluation data. Particularly for text-to-image generation tasks, incorporating human-annotated data leads to better evaluation performance than simply increasing the volume of GPT-generated annotations. Compared to LLaVA-Critic which trained on 113k evalaution samples(73k pointwise and 40k pairwise evaluation data), our model achieves better alignment with human scores using less training data on the same 7B backbone, demonstrating the superior quality of our dataset.

Analysis of relation between T2I and I2T We would like to further investigate the interaction between evaluation tasks of image-to-text (I2T) and text-to-image (T2I) generation. Specifically, we selected all 10k T2I evaluation samples from dataset after applying selection and balance method, and conducted SFT using only this subset. As shown in the results, training solely on 10k T2I evaluation data fails to yield strong performance on image-to-text generation tasks, although it does not lead

Analysis of	Data Size	Contained Tasks I2T	Contained Tasks T2I	Selection with Human Score on Human Evaluation Data	Score Balance	Judge Ave.	RichHF	All Ave.
Selection and Balance	124k	✓	✓	✗	✗	32.3	52.4	33.6
	102k	✓	✓	✓	✗	32.2	55.3	33.7
	57k	✓	✓	✗	✓	33.4	50.2	34.5
	57k	✓	✓	✓	✓	**33.5**	**56.8**	**35.1**
I2T and T2I tasks	10k	✗	✓	✓	✓	19.5	52.7	21.7
	10k	✓	✗	✓	✓	30.8	50.4	32.1
	47k	✓	✗	✓	✓	33.4	52.1	34.6
	57k	✓	✓	✓	✓	**33.5**	**56.8**	**35.1**

Table 4: Result of Minos on MLLM-as-a-Judge and RichHF-18K when training with different multimodal evaluation Data only during SFT stage. We present the pearson-r between the evaluation scores of Minos with the evaluation scores of human.

to bad T2I evaluation performance either. Interestingly, adding I2T evaluation data improves the model's evaluation capability on T2I generation task(from 52.7 to 56.8 in Pearson-r correlation). This suggests that training on I2T evaluation tasks contributes positively to the model's generalization on T2I evaluation tasks.

We also analyzed the impact of training only with I2T evaluation data. To this end, we used rest 47k I2T samples from the 57k total SFT dataset, as well as a sampled subset of 10k I2T samples, for separate SFT experiments. The results show that even training solely on I2T evaluation data enables the model to achieve non-trivial T2I evaluation performance. However, incorporating T2I evaluation data further enhances the model's consistency on text-to-image generation tasks. In summary, we observe that I2T evaluation data contributes positively to the model's performance on the text-to-image generation task, while T2I evaluation data also does not harm the model performance on I2T evaluation tasks. This finding highlights the importance of develop the multimodal evaluation model on image-to-text and text-to-image generation tasks together.

In conclusion, we argue that building a strong multimodal evaluation model requires the use of large-scale, high-quality human-annotated evaluation data. Furthermore, it is crucial to jointly consider both image-to-text and text-to-image evaluation tasks during training. These important aspects have been overlooked in previous research about MLLM evaluators, and our findings effectively point out directions for future research.

Selection & Balance	Mix Train	DPO	Judge Ave.	RichHF	All Ave.
✗	✗	✗	32.3	52.4	33.6
✓	✗	✗	33.5	56.8	35.1
✓	✗	✓	33.6	56.3	35.1
✓	✓	✗	34.7	55.6	36.1
✓	✓	✓	**35.5**	**57.9**	**37.0**

Table 5: Result of Ablation Study on MLLM-as-a-Judge and RichHF-18K. We present the pearson-r between the evaluation scores of models with the evaluation scores of human.

4.4 Ablation Study

We conduct an ablation study on the key methods used in Minos, and the results are summarized in Table 5. Full results can be found in Appendix A.

We adopted selection and balancing strategies on our data to enhance the model's evaluation performance on both image-to-text (I2T) and text-to-image (T2I) tasks, which shows performance improvement over not applying this strategy, as details in the previous part. However, directly applying DPO alignment training on top of this model leads to only marginal improvements, suggesting that preference data pairs alone may be insufficient for effective alignment.

Given that preference pairs in evaluation tasks are inherently a form of pairwise evaluation, they can still be considered valid evaluation data. Therefore, we constructed Evaluation Pair Comparison data and integrated it into the original evaluation data for SFT. We observe that training with this mixed data improve the model's I2T evaluation capabilities. However, it unexpectedly degraded performance on text-to-image generation tasks.

We further applied DPO alignment training on the Mix-Train model. This led to improvements in both I2T and T2I evaluation performance. These findings suggest that when developing multimodal evaluation models, solely using them for DPO alignment might be insufficient; instead, combining the evaluation comparison data during SFT and then applying DPO alignment allows for more effective utilization. Compared to the base model, our final Minos model achieved substantial improvements in consistency with human evaluation scores. Specifically, on the image-to-text tasks in MLLM-as-a-Judge, the average Pearson correlation improves by 3.2; on the RichHF-18k text-to-image dataset, it improves by 5.5; and across all cross-modal tasks, the average gain is 3.4.

5 Related Work

5.1 LMM as a Judge

As large multimodal models (LMMs) 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 LMMs 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 proprietary MLLMs as evaluators. 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 based on this data. However, existing evaluation datasets and MLLM-based evaluators have primarily focused on image-to-text tasks, with little exploration of text-to-image tasks, which are also a core part of multimodal evaluation. Moreover, prior works have largely relied on GPT-generated annotations, without systematically collecting or leveraging high-quality human-labeled evaluation data to develop a general-purpose multimodal evaluation model. This gap is precisely what our work aims to address.

5.2 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 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 image-to-text 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 text-to-image side, datasets such as ImageReward(Xu et al., 2023) and RichHF-18K(Liang et al., 2024) have been developed to support human evaluation. Nonetheless, despite these efforts, high-quality and large-scale human evaluation datasets that can be directly converted into pointwise supervision for training multimodal evaluation models remain limited. Many multimodal tasks still lack sufficient human evaluation data.

6 Conclusion

In this work, we first collect and construct a large-scale, general-purpose multimodal evaluation dataset: Minos-Corpus. Minos-Corpus comprises 124k multimodal evaluation samples spanning six common tasks and 16 datasets, covering both image-to-text and text-to-image settings. Each sample is accompanied by GPT-generated evaluation outputs, and a subset of 48k samples includes high-quality human annotations. Based on Minos-Corpus, we propose a training strategy that incorporates Data Selection and Balance, Evaluation Comparison, and Alignment techniques to develop Minos, a 7B-scale multimodal evaluation model. Averaged across all benchmark tasks, Minos achieves state-of-the-art (SoTA) performance among all open-source evaluation models built on a 7B backbone, and even outperforms several larger-scale open-source and proprietary models. Notably, on tasks such as text-to-image generation, Minos surpasses all models across scales, including GPT-4o.

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.

Moreover, when constructing Minos-Corpus, we utilized a substantial amount of GPT-generated annotations, which may introduce certain biases inherent to GPT models. Nevertheless, given the current scarcity of multimodal evaluation data, we argue that releasing a large-scale, high-quality, and general-purpose multimodal evaluation dataset is of great value to the community. Our experiments demonstrate that training multimodal evaluators on Minos-Corpus leads to stronger performance compared to previous models with the same backbone, thereby establishing Minos-Corpus as a valuable new resource for multimodal evaluation research.

Ethical Considerations

Our evaluation results are annotated by GPT-4o, and may therefore carry certain inherent biases from the model. When applying these annotations in real-world scenarios, it is important to carefully examine whether any such biases are present. Moreover, since our evaluation model is trained on data generated through this annotation process, the model itself may also reflect these biases. Consequently, we recommend conducting manual spot-checks of the model's outputs during deployment to ensure the reliability and fairness of its evaluations. We follow the correct usage of the data and models with the corresponding license.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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*.

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. 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.

Jiaxing Huang and Jingyi Zhang. 2024. A survey on evaluation of multimodal large language models. *Preprint*, arXiv:2408.15769.

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 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

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, and 1 others. 2024. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19401–19411.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

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.

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, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

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 Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, and 1 others. 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.

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.

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.

A Full Evaluation Results

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

We present full evaluation results corresponding to Table 4 and Table 5 in Table 7 and Table 8.

B Example of Evaluation and Evaluation Comparison Data

There is An example of Evaluation instance and its corresponding Evaluation Comparison instance in Minos-Corpus.

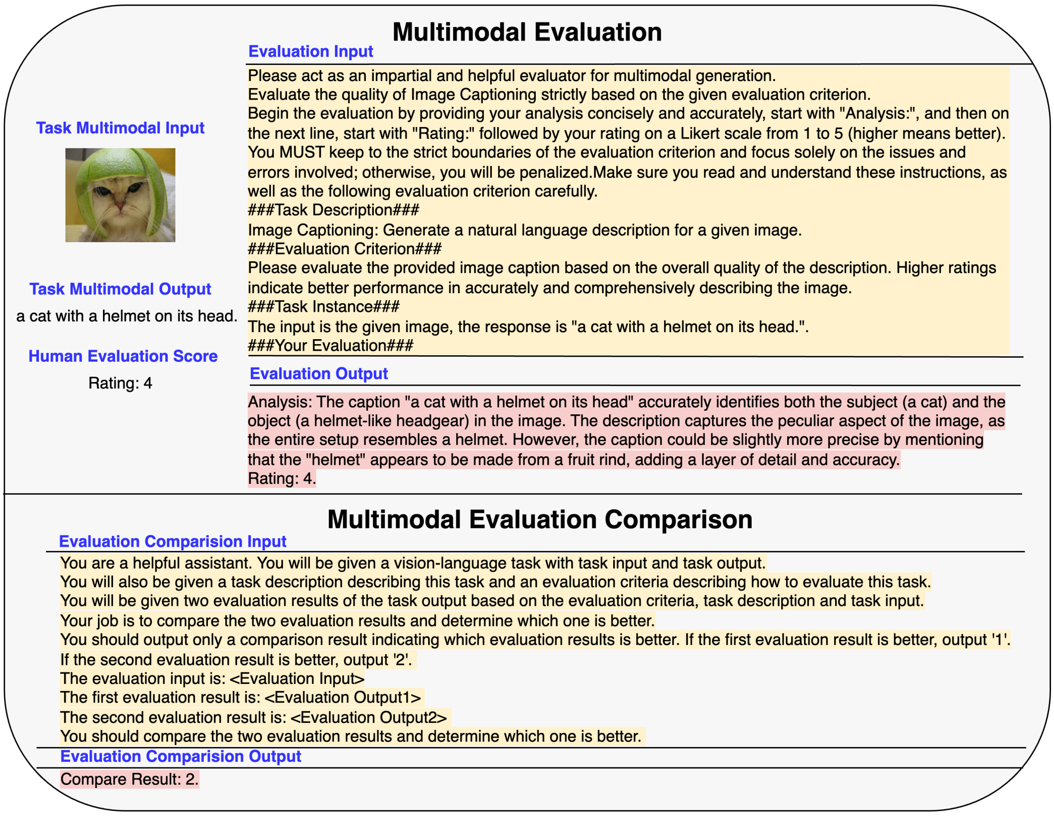

Figure 2: An example of Evaluation instance and its corresponding Evaluation Comparison instance in Minos-Corpus. We select a sample from the human-annotated image captioning dataset Polaris to illustrate our data. In total, 48k samples in Minos-Corpus are accompanied by corresponding human evaluation scores.

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 6: The full name of each task.

Setting	MLLM-as-a-Judge															RichHF	All
	CO	C.C.	Dif	Graph	Math	Text	WIT	Chart	Vis	CC	M2W	Sci	Aes	MM	Ave.		Ave.
S_1	23.2	42.8	16.2	31.7	43.7	41.9	35.4	45.3	37.2	19.9	22.7	24.7	34.6	33.1	32.3	52.4	33.6
S_2	25.6	42.8	16.3	28.7	42.8	42.0	34.1	45.0	41.2	23.0	23.4	22.3	30.2	32.9	32.2	55.3	33.7
S_3	23.4	41.6	17.2	31.2	43.1	44.4	34.8	43.8	42.5	22.2	24.7	31.1	32.9	34.1	33.4	50.2	34.5
S_4	23.2	41.7	16.8	35.1	44.9	46.8	35.9	48.0	37.6	20.7	24.6	24.8	38.0	30.9	33.5	56.8	35.1
S_5	23.8	16.3	8.9	26.1	25.9	35.5	-2.9	26.5	26.5	14.1	14.0	2.99	19.7	35.1	19.5	52.7	21.7
S_6	19.7	38.7	16.0	29.8	41.9	43.5	25.3	42.8	35.8	23.2	29.6	26.4	33.2	24.9	30.8	50.4	32.1
S_7	25.5	42.1	13.9	32.0	42.9	45.2	36.5	42.1	40.9	18.6	26.9	30.0	35.9	35.0	33.4	52.1	34.6
S_8	23.2	41.7	16.8	35.1	44.9	46.8	35.9	48.0	37.6	20.7	24.6	24.8	38.0	30.9	33.5	56.8	35.1

Table 7: 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	All
	CO	C.C.	Dif	Graph	Math	Text	WIT	Chart	Vis	CC	M2W	Sci	Aes	MM	Ave.		Ave.
S_1	23.2	42.8	16.2	31.7	43.7	41.9	35.4	45.3	37.2	19.9	22.7	24.7	34.6	33.1	32.3	52.4	33.6
S_2	23.2	41.7	16.8	35.1	44.9	46.8	35.9	48.0	37.6	20.7	24.6	24.8	38.0	30.9	33.5	56.8	35.1
S_3	22.3	39.5	28.8	34.5	45.2	42.6	32.3	45.6	40.0	21.4	24.9	26.6	33.5	32.6	33.6	56.3	35.1
S_4	25.6	44.0	18.6	36.2	44.7	47.0	34.4	42.4	46.8	24.3	22.9	26.1	38.6	34.3	35.5	55.6	36.1
S_5	24.5	38.4	32.5	34.9	43.4	47.8	37.0	46.9	42.9	24.4	23.5	33.4	31.7	36.1	35.5	57.9	37.0

Table 8: Detailed results of data analysis in Table 5. S_i represents the setting corresponding to the i-th row in Table 5.