LLaVA-Critic: Learning to Evaluate Multimodal Models

Tianyi Xiong[♮,♡], Xiyao Wang[♮,♡], Dong Guo[†], Qinghao Ye[†], Haoqi Fan[†], Quanquan Gu[†],
Heng Huang[♮], Chunyuan Li[†]
[†]ByteDance, [♮]University of Maryland, College Park
https://llava-vl.github.io/blog/2024-10-03-llava-critic

Abstract

We introduce LLaVA-Critic, the first open-source large multimodal model (LMM) designed as a generalist evaluator to assess performance across a wide range of multimodal tasks. LLaVA-Critic is trained using a high-quality critic instruction-following dataset that incorporates diverse evaluation criteria and scenarios. Our experiments demonstrate the model's effectiveness in two key areas: (i) LMM-as-a-Judge, where LLaVA-Critic provides reliable evaluation scores, performing on par with or surpassing GPT models on multiple evaluation benchmarks; and (ii) Preference Learning, where it generates reward signals for preference learning, enhancing model alignment capabilities. This work underscores the potential of open-source LMMs in self-critique and evaluation, setting the stage for future research into scalable, superhuman alignment feedback mechanisms for LMMs.

1. Introduction

The ability of learning to evaluate is increasingly taking on a pivotal role in the development of modern large multimodal models (LMMs), as pre-training on existing web data reaches maturity and the focus is shifting towards post-training with AI-enhanced synthetic data, which shows growing potential. Reliable AI evaluation is essential, not only for offering scalable solutions to reduce human labor in complex task assessments, but also for generating effective reward signals in reinforcement learning and guiding inference-time search [31, 33, 36].

Existing LMMs have made tremendous progress in recent research community [17], primarily on improving the performance of various real-world vision tasks in single-image [3, 8, 24], multi-image [13, 19] and video scenarios [16, 22, 41]. It remains unexplored to develop open LMMs to play the role of a judge and evaluate the performance of multimodal models. For instance, a model can follow a well-designed, itemized evaluation criterion to provide a score between 1 and 10 for rating different model responses in a visual chat task [24]. Along with the score, it would also offer the associated reasoning behind the evaluation, ensuring transparency and consistency in assessing model performance. In this paper, we present the first attempt to curate the instruction-following data particularly for evaluation, based on which we develop a LMM, LLaVA-Critic. Two primary scenarios/goals of building LLaVA-Critic are highlighted:

- *Scenario 1: LMM-as-a-Judge*. Open-source LMMs that can deliver reliable evaluation scores, comparable to or surpassing proprietary models like GPT-4V [30]/-4o [32]. These models offer a free alternative to replace commercial GPT models in various evaluation benchmarks.
- *Scenario 2: Preference Learning*. A scalable solution for generating effective reward signals, reducing the need for costly human feedback collection. This approach enhances preference alignment with AI-generated feedback.

Our experimental results demonstrate that: (i) As a judge model, the evaluation scores and rankings provided by LLaVA-Critic show a high correlation with commercial GPT models, making it a cost-effective alternative for model developers in resource-constrained settings; (ii) In preference learning, LLaVA-Critic offers AI-generated feedback in iterative Direct Preference Optimization (DPO) [35], outperforming the preference signals provided by the reward model in LLaVA-RLHF [38], which relies on human feedback for training the reward model.

In summary, our contributions are as follows:

- *Critic Instruction-Following Data*: We present a high-quality dataset tailored to follow instructions in complex evaluation setting to provide quantitative judgment and the corresponding reasoning process. It consists of 46k images with 113k evaluation instruction samples, including both pointwise and pairwise evaluation settings.
- *Large Multimodal Models*: We develop LLaVA-Critic, LMMs that expand the capabilities of open models to act as critic, providing effective evaluation and feedback.

♡ Work collaborated with ByteDance

- *Open-Source*: In an effort to support the development of general-purpose visual assistants, we release our critic instruction data, codebase, model checkpoints, and the trained visual chat demo to the public.

2. Related Work

LMM-as-a-judge. Strong proprietary LMMs such as GPT-4V/-4o have been demonstrated to serve as generalist evaluators for vision-language tasks [11, 50]. Specifically, for complex scenarios related to human judgment, such as visual chat and detailed captioning, LMMs have been utilized in evaluation benchmarks to judge the model responses, in both pointwise [15, 24, 38, 46, 48, 49] and pairwise settings [28, 45]. Our LLaVA-Critic are employed in the these evaluation scenarios as open-source alternative, with advantages in cheap and customized evaluation. For open-source models, Prometheus-Vision [14] is the first VLM trained as an evaluator for specific user-designed scoring criteria. While sharing the same open-source spirit, LLaVA-Critic stands out as the first open generalist evaluator. Note that GPT is also utilized to extract answers from LMM responses for subsequent evaluation in some benchmarks [12, 27, 43]. This extractive functionality for evaluation is out of the scope of this paper.

Preference learning for LMMs. Reinforcement learning from human feedback (RLHF) is a proven method to align large language models (LLMs) with human intentions. DPO [35] introduces a new parameterization of the reward model in RLHF, enabling direct optimization using pairwise preference datasets. CriticGPT [29] trains "critic" models to evaluate model-generated code, providing feedback signals to enhance code LLMs. The concept of preference learning has recently expanded from language models to the multimodal space. LLaVA-RLHF [38], the first open-source study in this area, improves visual chat abilities of LMMs using human-scored rankings. BPO [34] constructs preference data by introducing negative responses generated by the model itself, using distorted images or text-based LLMs to inject errors. Wang et al. [40] proposes mDPO, which uses conditional preference optimization to emphasize image information. Other works apply preference alignment to reduce hallucinations and enhance the overall capabilities of vision-language models (VLMs), either through human feedback (e.g., RLHF-V [44]) or AI feedback (e.g., Silkie: VLFeedback [20]). Several approaches use self-rewarding mechanisms to minimize dependence on external preference pairs, such as divide-and-conquer strategies [45] (RLAIF-V), sentence-level beam search [56], deliberate hallucination injection [55], or metric-based self-critic prompts [42]. As a generalist evaluator, LLaVA-Critic can provide valuable feedback for LMM alignment, paving the way for self-improving AI models.

3. Data Collection

We now introduce the data collection process for the LLaVA-Critic training dataset. The use of GPT-4/4V as a generalist evaluator for LMMs can be broadly categorized into two settings: (i) **Pointwise scoring**: GPT assigns a score to an individual candidate response, either by directly evaluating it based on predefined criteria or by scoring it relative to a fixed reference answer. This setting can be regarded as a combination of the *single-answer grading* and *reference grading* methods discussed in Zheng et al. [54]. (ii) **Pairwise ranking**: GPT compares two candidate responses to determine their relative quality (or declares a tie). To equip LLaVA-Critic with a generalized evaluation capacity as with GPT-4V, we design a GPT-assisted pipeline to curate our training dataset for both settings. An example of LLaVA-Critic training data is shown in Table 1.

3.1. Pointwise Data

To train a generalist critic model for the evaluation of individual LMM responses, each pointwise training data instance consists of the following components: a multimodal instruction (i.e., a question-image pair), a model response, an optional reference answer (depending on the evaluation prompt), an evaluation prompt, a judgment score, and the corresponding justification for the score. By organizing them in a sequence, the training sample is:

`(Image, Question, Response, Reference, Evaluation Criteria, Score, Reason)`

where green parts are treated as model output to compute the auto-regressive loss, the order of `Score` and `Reason` is specified by the evaluation prompt. We select multimodal instructions from 8 multimodal instruction tuning datasets, spanning across a wide range of tasks including: (1) general visual conversation, detailed captioning and reasoning (LLaVA-Instruction-150k [24], SVIT [53]); (2) more challenging tasks such as complex reasoning (ComVint [10]), text-rich understanding (LLaVAR [51]) and robustness-oriented instructions (LRV-Instruction [23]); and (3) various specific domains such as academic question answering (M3IT [21]), medical image understanding (LLaVA-Med [18]) and embodied decision-making (PCA-EVAL [7]). For each multimodal instruction, we select one or more model responses from VLFeedback [20], which collects multiple responses from 12 off-the-shelf LMMs. Additionally, we generate responses using GPT-4o, a leading commercial LMM, to serve as high-quality reference answers.

To equip LLaVA-Critic with general evaluation capacities across various tasks, we construct an evaluation prompt pool from 7 widely used multimodal benchmarks that utilize GPT-as-a-judge, including LLaVA-in-the-Wild [24], LLaVA-Wilder [15], Image Detailed Caption-

Multimodal Input for LMM:
Question: What are the specifics visible in the image?

Setting 1: Pointwise Scoring
Response: The image shows a small train with four red cars, traveling on a track. The train is located in a park setting, and there are potted plants nearby.
Evaluation Prompt: From 0 to 100, how much do you rate for this Text Caption in terms of the correct and comprehensive description of the image? Do not dominate the rating by a single attribute such as recognition correctness, but a overall rating on the object/scene appearance, position, pose, action, shape, etc., and contents in the background. Do not consider the appropriateness or sensitive descriptors, such as "middle-aged western man", judge based on if it has correct specifications of the object and scenes in image. Provide a few lines for explanation and the rate number at last after "Final Score:".
LLaVA-Critic: The text caption provides a generally accurate description of the image. It correctly identifies the small train with four red cars and mentions the park setting and the presence of potted plants. However, there are a few details that could be improved for a more comprehensive description. For instance, the caption could mention that the train is stationary rather than traveling, and it could describe the surrounding structures, such as the building and the wooden fence, to give a fuller picture of the scene. Final Score: 85.

Setting 2: Pairwise Ranking
Response A: The image shows a small train on a track, with a total of six cars attached to it. The train is traveling on a small track, and there are potted plants nearby. The scene appears to be set in a park or a similar outdoor area.
Response B: The image shows a small train with four red cars, traveling on a track. The train is located in a park setting, and there are potted plants nearby.
Evaluation Prompt: "Given an image and a corresponding question, please serve as an unbiased and fair judge to evaluate the quality of the answers provided by a Large Multimodal Model (LMM). Determine which answer is better and explain your reasoning with specific details. Your task is provided as follows:
Question: []\n Response A: []\n Response B: []\n ASSISTANT:\n"
LLaVA-Critic: Response B is better than response A because it provides a more accurate description of the objects and their attributes in the image. Response B correctly identifies the train as having four red cars, whereas Response A inaccurately states that there are six cars. Both responses mention the potted plants and the park setting, but Response B's accuracy in the number of train cars makes it more precise. Additionally, Response B avoids unnecessary repetition and provides a clearer depiction of the scene, making it more helpful and accurate overall.

Table 1. An example of LLaVA-Critic training data. The top block shows pointwise scoring, where LLaVA-Critic predicts a score to evaluate a single response's quality; the bottom block illustrates pairwise ranking, where it rank response pairs. In both settings, LLaVA-Critic learns to provide reasons for its judgments.

ing [15], MMHal-Bench [38], MMVet [47], WildVision-Bench [28] and RefoMB [45]. [1] Prompts that require additional textual context—since they use text-only GPT-4 as the evaluator—are adjusted to focus on the input image, better aligning with the LMM evaluator setting. To construct training data based on each evaluation prompt, we select multimodal instructions and model responses according to the specified evaluation scenario, and include reference answers from GPT-4o when necessary. These components are then assembled into the evaluation prompt and used as input for GPT-4o (as-a-judge) to provide high-quality judgment scores and detailed justifications for model responses. Finally, our pointwise training dataset comprises a total of 18,915 question-image pairs and 72,782 critic data samples.

3.2. Pairwise Data

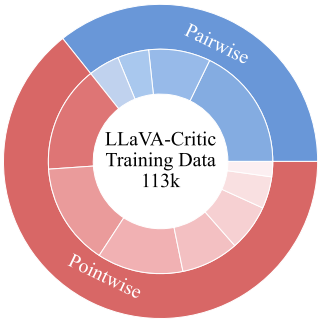
The pairwise data consists of responses with known preference relationships. We collect pairwise data from three datasets: VLFeedback [20], RLHF [38], and RLHF-V [44]. In the VLFeedback dataset, each (question, response) pair is rated across three different dimensions by GPT-4V. For the same question, responses generated by different LMMs can form multiple response pairs for that question. We randomly select 20k pairs where the average score gap between responses is greater than 0.6. Besides, to ensure diversity in the preferences, we randomly sample 5k pairs where the two responses had identical scores across all three dimensions to serve as "Tie" training data. In the RLHF dataset, each question is annotated with preference relationships between different responses by human evaluators. In contrast, the RLHF-V dataset consists of responses generated by LMM, which have been manually refined to produce improved responses. From these two datasets, we collect 9.4k (RLHF) and 5.7k (RLHF-V) response pairs, each annotated with human preferences. This results in a total of 40.1k pairwise data samples.

To enable LLaVA-Critic to provide useful detailed feedback in addition to the preference relation, we utilize GPT-4o to generate reasons behind the given preference judgment. The training sample for pairwise data is structured in the following sequence:

```
(Image, Question, Response 1&2, Evaluation)
(Criteria, Preference, Reason)
```

where the evaluation criteria is from carefully designed prompt templates. To allow LLaVA-Critic to handle diverse

[1] Although RefoMB and WildVision-Bench use pairwise evaluation prompts, only one response is evaluated, with the other from a fixed reference model (GPT-4V and Claude-3-Sonnet, respectively), making them pointwise evaluations.



Setting	Prompt source	Data source	Data size
Pointwise	LLaVA-in-the-Wild	LLaVA, SVIT, LLaVAR, LLaVAMed, ComVint	17.5k
	LLaVA-Wilder	SVIT, LLaVAR, LLaVAMed, ComVint, M3IT, PCAEval	16.6k
	WildVision-Bench	VLFeedback	14.0k
	MMVet	LLaVAR, LLaVAMed, M3IT, PCAEval	9.3k
	MMHal-Bench	LRV-Instruction	7.6k
	ImageDC	SVIT-detail	5.3k
	RefoMB	VLFeedback	2.5k
Pairwise	30 manually crafted prompt templates	VLFeedback	20.0k
		LLaVA-RLHF	9.4k
		VLFeedback (Tie)	5.0k
		RLHF-V	5.7k

Figure 1. Data statistic of LLaVA-Critic-113k training dataset. In the pointwise setting, we categorize datasets by instruction sources and select data based on the task type corresponding to each evaluation prompt. Note that all our training data is sourced from public instruction-following training sets and does not overlap with with any evaluation benchmarks.

pairwise ranking, we develop a set of 30 evaluation prompt templates (see Appendix B.1). Each preference pair is randomly assigned a template from this set to form the final training data.

Data statistics. Our training dataset comprises a total of 46k images and 113k data samples. As shown in Figure 1, we curate our training set with diverse instruction-response pairs, spanning multiple evaluation tasks and domains.

4. LLaVA-Critic

4.1. Model

To train the LLaVA-Critic model, we fine-tune a pre-trained LMM that already possesses strong capabilities in following diverse instructions. This is crucial, as it ensures that the model has already been equipped to handle a wide range of vision tasks in the wild with high quality. The evaluation ability is treated as an additional discriminative ability closely tied to these scenarios. During training, LLaVA-Critic takes an evaluation prompt—assembling the multimodal instruction input, model response(s), and an optional reference response—as input. It is trained to predict quantitative pointwise scores or pairwise rankings based on the criteria in the evaluation prompt, and provide detailed justifications for the assigned judgments. Standard cross-entropy loss is applied to both judgments and justifications.

In experiment, we start with the LLaVA-OneVision(OV) 7B/72B pretrained checkpoint and fine-tune it on the proposed LLaVA-Critic-113k dataset for 1 epoch to develop LLaVA-Critic. We apply a learning rate of 2e-6 and a batch size of 32 for training, with other hyperparameters set to the defaults from Li et al. [16]. We also curate a subset with 53k samples (42k pointwise, 11k pairwise) that cover fewer instruction sources and domains. The model trained on this reduced subset is referred to as LLaVA-Critic (v0.5).

4.2. Scenario 1: LMM-as-a-Judge

Evaluating complex tasks often requires human judges to provide feedback, which can be labor-intensive. LLaVA-Critic can serve as a general evaluator for LMM responses, reducing labor costs by automating the evaluation process. LLaVA-Critic consistently provides reliable judgments and justifications aligned with GPT-4o or human evaluations across a range of widely used multimodal benchmarks. This consistency holds true for both instance-level scoring and model-level ranking, as demonstrated in Sec. 5.1.

Specifically, we consider the following evaluation scenarios: (i) *Visual Chat*. This task involves handling daily-life visual tasks through multimodal dialogue, requiring evaluation of task completion quality in a conversation setting. Examples include LLaVA-Bench [24] and LLaVA-in-the-Wild [24], which focus on simpler scenarios, while LLaVA-Wilder [15] addresses more challenging cases. (ii) *Integrated capabilities*. Real-world tasks require integration of multiple basic abilities of LMMs. MM-Vet [47] offers a comprehensive benchmark, evaluating core vision-language capabilities including recognition, OCR, knowledge integration, language generation, spatial awareness, and math. The Multimodal Live-Bench tests the model's ability to generalize to new, unobserved knowledge by leveraging continuously updated news and online forums. (iii) *Preferences*. This task simulates real-world multimodal interactions where models are expected to align their behavior with human preferences. The WildVision-Bench [28] is a prime example, replicating scenarios from the online platform WildVision-Arena (WV-Arena) to evaluate preference-based interactions. (iv) *Detailed Description*. This task assesses models on their ability to provide comprehensive and detailed descriptions of images and videos. Image Detailed Captioning [15] evaluates detailed descriptions in images, while Video Detailed Captioning [52] extends these capabilities to video contexts. (v) *Hallucination*: This task focuses on the model's ability to provide grounded responses based on the given context, ensuring that it avoids generating inaccurate or fabricated information, exemplified by MMHal-Bench [38].

4.3. Scenario 2: Preference Learning

Leveraging a generalist evaluator as a critic to generate reward signals for reinforcement learning is a promising re-

search direction. In this work, we employ LLaVA-Critic to produce AI-generated feedback datasets for diverse tasks, thereby improving the performance of supervised fine-tuned LMMs through preference alignment. Notably, the reward signals generated by our critic can be utilized in any preference learning algorithms, including RLHF and DPO. To quickly assess the effectiveness of the reward data, we focus on how LLaVA-Critic is incorporated into the iterative DPO training process.

- ***Step 1: Response generation***. The iterative DPO process begins with a pretrained LMM π_0 as the initial checkpoint and a set of multimodal instructions $\{(x_k, v_k)\}_{k=1}^{N}$, where each x_k is a question and v_k is the corresponding image. For each question-image pair (x, v), the pretrained LMM π_0 randomly generates K responses $\{y_1, y_2, \ldots, y_K\}$, sampled independently from its distribution.
- ***Step 2: Scoring***. To mitigate order-related variance in LLaVA-Critic's preferences, we form all possible ordered pairs from these responses, resulting in $K \times (K-1)$ pairs. For each response pair (y_i, y_j), we apply LLaVA-Critic with an evaluation prompt to generate a relative score a_{ij}, which normalizes the score of y_j based on y_i.
- ***Step 3: Reward Preference***. The overall reward score r_i for each response y_i is calculated by aggregating these preference scores: $r_i = \sum_{k \neq i} a_{ki} - \sum_{l \neq i} a_{il}$ This calculation effectively measures how much better or worse y_i is compared to all other responses. We then select the responses with the highest and lowest reward scores as the best and worst responses, denoted as y^+ and y^-, respectively. These form the pairwise feedback data (y^+, y^-), which is used for DPO training to enhance the LMM's alignment with LLaVA-Critic's preferences.

Iterative Improvement. After each round of DPO training, the updated LMM becomes the new starting checkpoint. The process is then iterated for another $M-1$ rounds, using LLaVA-Critic to progressively improve the model's performance based on its self-generated responses.

5. Experimental Results

5.1. LMM-as-a-Judge

To comprehensively assess LLaVA-Critic's capacity in evaluating LMM responses across different scenarios, we consider two primary experimental settings: (1) *In-domain Judgments*: where we measure LLaVA-Critic's consistency with GPT-4o or human evaluators on evaluation tasks/prompts included in the LLaVA-Critic-113k dataset; and (2) *Out-of-domain Judgments*: where we apply LLaVA-Critic on tasks and prompts that are unseen during training. For the second setting, we use the MLLM-as-a-Judge [6] benchmark to assess the alignment between LLaVA-Critic and human evaluators in generalized scenarios.

In-domain Pointwise Scoring To evaluate the consistency between LLaVA-Critic and GPT-4o [32] in pointwise scoring across different evaluation scenarios, as described in Sec. 4.2, we select 7 popular multimodal benchmarks and collect candidate responses from 13 commonly used LMMs alongside their GPT-4o evaluations, resulting in a total of 14174 examples (see details in Appendix B.2). LLaVA-Critic is then tasked with providing judgments on theses samples. We report Pearson correlation to measure the degree of alignment with GPT-4o in instance-level scoring.

We conduct experiments based on three different baseline models: LLaVA-NeXT (LLaMA-8B) [15, 26], LLaVA-OneVision-7B, and LLaVA-OneVision-72B. As shown in Table 2, LLaVA-Critic variants significantly outperform their corresponding baseline models across all models and benchmarks. (i) *Data scaling*. By comparing the performance between v0.5 and full data trained LLaVA-Critic-7B, it concludes the necessity of larger size and diversity of instruction in training data. (ii) *Model scaling*. The best performance in terms of Pearson-r is achieved by LLaVA-Critic-72B with an average score of 0.754, which significantly outperforms the LLaVA-OV-72B baseline (0.634). This indicates that LLaVA-Critic-72B already possesses pointwise scoring capabilities that are quite aligned with GPT-4o. Despite a considerable reduction in model size, LLaVA-Critic-7B retains very strong point-wise scoring capabilities. With a score of 0.732, its shows minimal performance decline compared to LLaVA-Critic-72B, and significantly outperforms other advanced open-source LMMs of similar size, such as Qwen2-VL (0.352) and LLaMA3.2-Vision (0.359). This presents an advantage for deploying and utilizing LLaVA-Critic in resource-constrained environments. We also provide Kendall's Tau results in Appendix C.2 to assess model-level ranking, which reveal similar patterns and conclusions.

Figure 2 presents a qualitative comparison between LLaVA-Critic and other LMM evaluators. While LLaVA-OneVision often assigns fixed scores (e.g., "Tie" on WildVision-Bench or "6" on MMHal-Bench), LLaVA-Critic produces more diverse and balanced scores that closely align with GPT-4o, leading to consistent rankings of response models. Notably, even without training on critic data, LLaVA-OneVision-72B demonstrates model-level rankings that partially align with GPT-4o across four multimodal benchmarks.

In-domain Pairwise Ranking To assess the consistency between LLaVA-Critic and human evaluators in pairwise ranking, we use the battle data from WildVision Arena [28], which comprises 11k human-annotated preference relations among LMM response pairs. Each relation includes a question-image pair and two responses generated by different models, accompanied by a human-annotated preference (including ties). From this dataset, we randomly sample

LMM Evaluator	Pearson-r (↑)							
	ImageDC	MMVet	WildVision	LLaVA-B	LLaVA-W	L-Wilder	MMHal	Avg.
LLaVA-NeXT (LLaMA-8B)	0.262	0.317	0.147	0.211	0.345	0.156	0.472	0.273
LLaVA-Critic (LLaVA-NeXT)	0.673	0.706	0.580	0.529	0.820	0.936	0.748	0.713
Qwen2-VL-7B-Instruct	0.199	0.463	0.096	0.208	0.476	0.694	0.329	0.352
LLaMA3.2-11B-Vision-Instruct	0.069	0.450	0.224	0.356	0.499	0.531	0.387	0.359
LLaVA-OV-7B	0.056	0.349	0.251	0.335	0.533	0.592	0.433	0.364
LLaVA-Critic-7B (v0.5)	0.737	0.718	0.571	0.494	0.789	0.932	0.746	0.712
LLaVA-Critic-7B	0.735	0.733	0.616	0.510	0.843	0.940	0.748	0.732
LLaVA-OV-72B	0.718	0.680	0.446	0.436	0.716	0.824	0.620	0.634
LLaVA-Critic-72B	0.802	0.723	0.705	0.524	0.782	0.951	0.790	0.754

Table 2. Results on in-domain pointwise scoring. LLaVA-Critic consistently outperforms baselines across 7 multimodal benchmarks.

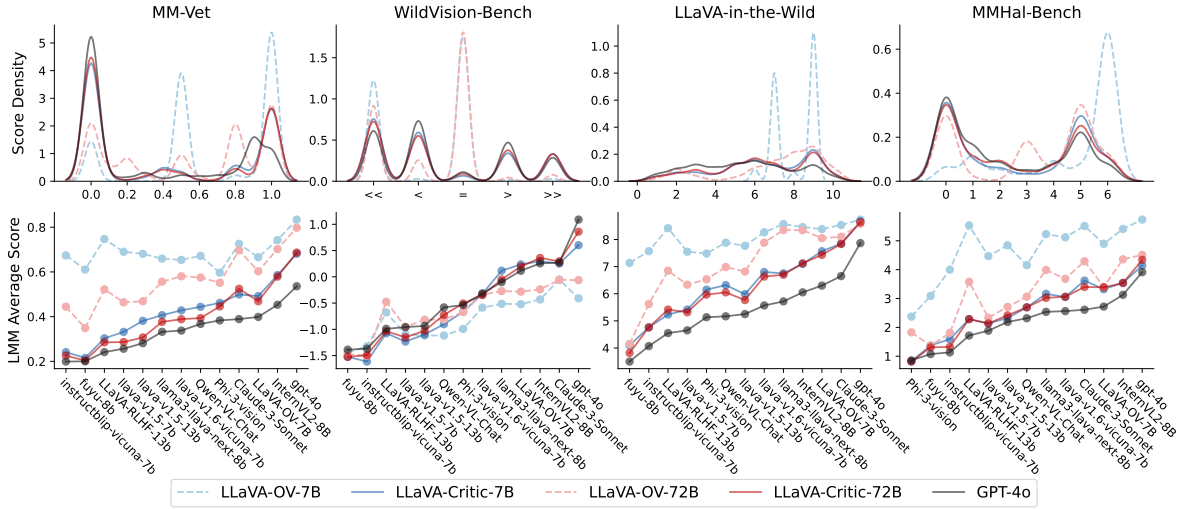


Figure 2. *(Top):* Overall distribution of evaluation scores across 4 benchmarks. *(Bottom):* Calculated average evaluation score for each response model on each benchmark. Each color represents a different LMM evaluator. Leveraging high-quality critic training data, LLaVA-Critic closely aligns with GPT-4o in delivering balanced evaluation scores and accurately ranking response LMMs.

2k response pairs and assign them to evaluation prompts from the pairwise ranking prompt template set mentioned in Section 3.2, creating the in-domain evaluation dataset. We report average accuracy, with and without ties, to assess alignment with human evaluators at the instance level. For model-level consistency, we calculate the Elo rating for each response LMM and report Kendall's Tau to measure the overall ranking correlation with human preferences.

Experimental results are reported in Table 3. While existing open-source LMMs exhibit initial pairwise ranking ability, there is a notable performance gap compared to GPT-4V/4o. After training with critic data, LLaVA-Critic achieves significant improvements. Specifically, LLaVA-Critic-72B achieves an average accuracy of 73.6% in pairwise comparisons without tie, outperforming both GPT-4o and GPT-4V. For pairwise accuracy with tie and model-level ranking (Kendall's Tau), LLaVA-Critic-72B shows only a marginal gap compared to GPT-4V/4o, with an accuracy of 60.5% and a score of 0.779, respectively. Notably, despite a substantial reduction in the number of parameters, LLaVA-Critic-7B still achieves an average accuracy of 59.6% in pairwise ranking with ties and 72.2% without ties, alongside a Kendall's tau of 0.763. These results underscore the strong alignment between LLaVA-Critic and human evaluators in pairwise ranking LMM responses.

MLLM-as-a-Judge MLLM-as-a-Judge [6] is a comprehensive benchmark to evaluate the degree of alignment between model-based evaluation and human evaluation. It collects approximately 17k image-instruction-response triplets across 14 multimodal benchmarks and 6 LMM response models. Human annotators are then employed to assess model responses under scoring, pairwise comparison and batch ranking settings, resulting in 7756, 5719, 1469 examples respectively. In our experiments, we evaluate LLaVA-Critic in both (pointwise) scoring and pair comparison settings to assess its general alignment with human evaluators. We report the average Pearson correlation for scoring and average accuracy for pairwise comparison, following the metrics used in the original benchmark.

We compare LLaVA-Critic with commercial models (GPT-4V/4o, Gemini-Pro [39]), open-source LMMs, as well as Prometheus-Vision [14], which trains a LLaVA

Model	Acc(w. Tie)↑	Acc(w.o. Tie)↑	Kendall's τ ↑
GPT-4o	0.617	0.734	0.819
GPT-4V	0.620	0.733	0.787
LLaVA-NeXT (LLaMA-8B)	0.473	0.569	0.605
LLaVA-OV-7B	0.531	0.640	0.715
Qwen2-VL-7B-Instruct	0.550	0.678	0.699
LLaMA3.2-V (11B-Instruct)	0.513	0.673	0.737
LLaVA-OV-72B	0.594	0.708	0.763
LLaVA-Critic-7B (v0.5)	0.580	0.692	0.755
LLaVA-Critic (LLaVA-NeXT)	0.582	0.686	0.755
LLaVA-Critic-7B	0.596	0.722	0.763
LLaVA-Critic-72B	0.605	0.736	0.779

Table 3. Results on in-domain pairwise ranking. LLaVA-Critic is comparable with GPT-4V in alignment with human evaluators.

model on a curated LMM-as-a-judge dataset comprising 15k GPT-generated rubrics and 150k GPT-4V feedback data. As demonstrated in Table 4, LLaVA-Critic-7B surpasses all baselines of comparable model size by a substantial margin. Built on a stronger base model, LLaVA-Critic-72B further achieves the Pearson similarity with human annotators from 0.314 to 0.393 in pointwise scoring. For pairwise comparisons, it achieves accuracy rates of 57.8% and 71.5% with and without ties, respectively, reaching a level of alignment with human evaluators comparable to GPT-4V/4o. We also compare different variants of LLaVA-Critic and observe performance gains with both stronger base models and larger training data, consistent with previous findings. This again highlights the critical role of model and data scaling in building an effective and generalist open-source LMM evaluator. More comprehensive results are provided in Appendix C.3.

Qualitative Comparison We present example comparisons of the evaluation scores and reasons generated by LLaVA-Critic and other LMMs, with detailed examples provided in Appendix D. The key findings are as follows: Compared to LLaVA-OneVision, LLaVA-Critic delivers more accurate judgments (Table 14), and provides more concrete, image-grounded justifications (Table 15). The latter is crucial for reliable AI [4], as offering well-supported reasons for evaluations establishes LLaVA-Critic as a transparent evaluator of LMM responses.

Critic training preserves original visual capacities. As shown in Appendix C.4, LLaVA-Critic largely preserves LLaVA-OV's original ability in handling diverse visual tasks and achieves modest gains in visual chat performance.

5.2. Preference Learning

We further evaluate LLaVA-Critic's performance in providing reward signals for iterative DPO. LLaVA-OneVision's supervised fine-tuned checkpoint is used as the base policy model, and question-image pairs from LLaVA-RLHF [38] serve as the multimodal instructions. For each pair, $K = 5$ candidate responses are generated through random decoding (with a temperature of 0.7 and top-p of 0.9) to ensure response diversity. LLaVA-Critic is employed as described in Sec. 4.3 to construct the pairwise feedback data, which is then used for one epoch of DPO training. We perform iterative DPO for $M = 3$ rounds in total.

Model	Score↑	Pair(w. Tie)↑	Pair (w.o. Tie)↑
GPT-4V*	0.490	0.636	0.773
GPT-4o†	0.439	0.577	0.736
GPT-4V†	0.424	0.538	0.717
Gemini-pro*	0.304	0.509	0.615
LLaVA-v1.5-7B	0.158	0.439	0.576
Prometheus-V (LLaVA-v1.5-7B)	0.213	–	–
LLaVA-NeXT (LLaMA-8B)	0.198	0.461	0.586
LlaVA-OV-7B	0.151	0.426	0.550
Qwen2-VL-7B-Instruct	0.253	0.348	0.645
LLaMA3.2-V (11B-Instruct)	0.237	0.529	0.658
LlaVA-OV-72B	0.287	0.513	0.701
LLaVA-Critic (LLaVA-v1.5-7B)	0.228	0.528	0.656
LLaVA-Critic (LLaVA-NeXT)	0.272	0.547	0.677
LLaVA-Critic-7B (v0.5)	0.312	0.546	0.675
LLaVA-Critic-7B	0.314	0.556	0.689
LLaVA-Critic-72B	0.393	0.578	0.715

Table 4. Results on MLLM-as-a-Judge [6]. *: the results as reported in the original paper [6]; †: results from our evaluation of GPT-4V/4o based on their codebase. Note that Prometheus-Vision cannot follow the pairwise evaluation prompt. LLaVA-Critic significantly narrows the gap between open-source LMMs and GPT-4V/4o in their ability to evaluate LMM responses across a wide range of evaluation scenarios.

To assess the effectiveness of LLaVA-Critic's reward signals, we evaluate the final LMM checkpoint on 6 open-ended multimodal benchmarks: four image-based tasks (LLaVA-in-the-Wild [24], LLaVA-Wilder [15], WildVision-Bench [28] and LiveBench [48]), one video-based task (Video Detailed Captioning [15]), and one hallucination benchmark (MMHal-Bench [38]). We compare LLaVA-Critic with two baselines: (1) reward model from LLaVA-RLHF [38], which is trained on human preferences, and (2) a naive baseline that replaces LLaVA-Critic with LLaVA-OneVision as a zero-shot reward model.

As shown in Table 6, preferences provided by LLaVA-Critic significantly improve LLaVA-OneVision's visual chat capacities and reduce hallucination across challenging tasks. LLaVA-Critic consistently surpasses other baseline reward models on 5 out of 6 benchmarks for the 7B base model and all 6 benchmarks for the 72B base model. Despite the preference alignment conducted solely with images, LLaVA-Critic also enhances LLaVA-OneVision's performance in Video Detailed Captioning (+0.12 on OV-7B and +0.26 on OV-7B), demonstrating its ability to generalize to both image and video contexts. Additionally, we observe that Critic-7B outperforms Critic-7B-v0.5 on 5 out of 6 benchmarks, highlighting the importance of stronger reward models—trained on more diverse critic instructions—

Method	#Prompts	LLaVA-W	L-Wilder	WildVision	LiveBench	MMHal*	MME^P	MME^C	MMB-en	MM-Vet	MMStar
LLaVA-v1.5-7B	–	63.4	54.2	20.4	45.6	1.94	1510.7	348.2	64.3	31.1	33.3
+ RLHF	9.4k	63.7	54.5	19.8	46.2	1.90	1508.2	360.2	60.4	31.1	33.0
+ SIMA	17k	66.1	52.3	17.6	47.9	1.81	1507.7	**379.3**	64.9	31.6	34.7
+ CSR	15k	71.1	55.9	20.0	45.0	1.96	**1524.2**	367.9	**65.4**	**33.9**	33.6
+ RLAIF-V	33.8k	72.7	56.4	19.2	**50.4**	**3.04**	1362.7	302.9	62.6	26.7	**35.4**
+ LLaVA-Critic	9.4k	**73.5**	**57.2**	**29.2**	50.0	2.07	1500.4	350.7	64.1	32.2	34.2

Table 5. Comparison with other preference learning algorithms on LLaVA-v1.5-7B. Apart from benchmarks in Table 6, we also report the results on 4 comprehensive multimodal benchmarks for reference. The best and second best results are shown in **bold** and underlined respectively. *OpenAI's *gpt-4-0613* is used for the MMhal-Bench evaluation due to the deprecation of the original API.

Base	Reward	LLaVA-W	L-Wilder	WV-B	Live-B	V-DC	MMHal
GPT-4V	–	98.0	81.0	79.8	73.7	4.00	3.83
OV-7B	–	90.7	67.8	54.0	77.1	3.75	3.19
	OV-7B	98.6	70.9	66.6	84.0	3.77	3.79
	LLaVA-RLHF	97.5	70.3	64.1	83.1	3.84	**4.01**
	Critic-7B (v0.5)	98.1	70.5	67.2	**85.1**	3.83	3.85
	Critic-7B	**100.3**	**71.6**	**67.3**	84.5	**3.87**	3.91
OV-72B	–	93.5	72.0	51.7	81.5	3.60	3.61
	LLaVA-RLHF	103.2	75.2	65.2	86.2	3.85	3.67
	Critic-72B	**104.4**	**75.9**	**70.0**	**88.5**	**3.86**	**3.77**

Table 6. Comparison between LLaVA-Critic and baselines in preference alignment. "Base": the initial LMM checkpoint for DPO.

to deliver more accurate reward signals and further enhance preference learning. (See Appendix C.5 for additional results and Table 16 for a visual-chat example.) Notably, while using OpenAI's GPT-4o as a reward model for 3 rounds of iterative DPO would cost approximately $690, LLaVA-Critic provides a reliable, cost-free alternative for providing reward signals.

Comparison We take LLaVA-v.1.5-7B as the base policy model, and compare LLaVA-Critic with 4 previous methods that apply preference optimization with self-generated candidate responses. These methods primarily vary in the source of reward signals: LLaVA-RLHF [38] leverages a pretrained reward model based on human feedback; SIMA [42] develops an in-context self-critic prompt for providing pairwise judgments; CSR [56] incorporates sentence-level beam search with CLIP-score calibration; and RLAIF-V [45] adopts a divide-and-conquer strategy to calculate the overall reward score by combining sentence-level judgments. For our method, we utilize the prompts (question-image pairs) from the LLaVA-RLHF dataset and perform DPO training for 3 epoches.

As illustrated in Table 5, with only 9.4k input prompts, the reward signal provided by LLaVA-Critic substantially improve the base model's performance across various open-ended visual chat benchmarks. It achieves the best improvements of +10.1 on LLaVA-W, +3.0 on LLaVA-Wilder, +8.8 on WildVision-Bench, along with the second-highest gains of + 4.4 on LiveBench and +0.13 on MMHal-Bench, respectively. At the same time, the overall capacities of LLaVA-v1.5-7B are largely preserved, as demonstrated on other comprehensive benchmarks. This is superior to other competing methods, which either result in smaller performance gains or achieve improvements by compromising the overall capabilities on other benchmarks.

Sampling	LLaVA-W	L-Wilder
Random	100.3	71.6
Best-of-5 (w. Critic-7B)	102.0	74.8

Table 7. Results of BoN sampling. Responses are generated by the OV-7B checkpoint after 3-round iterative DPO training, with LLaVA-Critic providing the reward scores.

Inference Time Search Applying LLaVA-Critic for best-of-n sampling [37] further enhances LMM performance during inference. For the LLaVA-OV-7B checkpoint after iterative DPO training, we generate $n = 5$ responses for each question with a temperature of 0.7 and top-p of 0.9, then use LLaVA-Critic-7B to select the best responses. As shown in Table 7, this results in additional gains of +1.7 on LLaVA-W and +3.2 on LLaVA-Wilder.

6. Conclusions

We have presented LLaVA-Critic, an open-source LMM that is trained to evaluate model performance in a wide range of multimodal scenarios. To achieve this, we curated a high-quality critic instruction-following dataset with diverse evaluation criteria. We demonstrated the effectiveness of LLaVA-Critic in two key areas: (1) as a generalized evaluator, LLaVA-Critic provides pointwise scores and pairwise rankings that closely align with human and GPT-4o preferences across multiple evaluation tasks, presenting a viable open-source alternative to commercial GPT models for autonomous assessment of open-ended LMM responses; (2) in preference learning, LLaVA-Critic functions as a reliable reward model, supplying preference signals that enhance the visual chat capabilities of LMMs, surpassing the LLaVA-RLHF reward model built with human feedback. This work represents an important step toward harnessing the self-critique capabilities of open-source LMMs, and we hope it will inspire further research into developing strong LMMs with scalable and superhuman alignment feedback.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 12

[2] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1, 2024. 12

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Technical Report*, 2023. 1, 12

[4] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. 7

[5] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 12

[6] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024. 5, 6, 7, 14, 15

[7] Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond. *arXiv preprint arXiv:2310.02071*, 2023. 2

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 1, 12

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2024. 12

[10] Yifan Du, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, Mingchen Cai, Ruihua Song, and Ji-Rong Wen. What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning. *arXiv preprint arXiv:2311.01487*, 2023. 2

[11] Wentao Ge, Shunian Chen, Guiming Chen, Junying Chen, Zhihong Chen, Shuo Yan, Chenghao Zhu, Ziyue Lin, Wenya Xie, Xidong Wang, et al. Mllm-bench, evaluating multi-modal llms using gpt-4v. *arXiv preprint arXiv:2311.13951*, 2023. 2

[12] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 2

[13] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 1

[14] Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. Prometheusvision: Vision-language model as a judge for fine-grained evaluation. *arXiv preprint arXiv:2401.06591*, 2024. 2, 6

[15] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, 2024. 2, 3, 4, 5, 7, 12

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 4, 12

[17] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 2023. 1

[18] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 2

[19] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next: Tackling multi-image, video, and 3d in large multimodal models, 2024. 1

[20] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023. 2, 3

[21] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023. 2

[22] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 1

[23] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 2

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 2, 4, 7, 12

[25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 12

[26] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 5, 12

[27] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *arXiv preprint arXiv:2310.02255*, 2023. 2

[28] Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*, 2024. 2, 3, 4, 5, 7, 12

[29] Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*, 2024. 2

[30] OpenAI. Gpt-4v. https://openai.com/index/gpt-4v-system-card/, 2023. 1

[31] OpenAI. Openai o1. https://openai.com/o1/, 2024. 1

[32] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 1, 5, 12

[33] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 1

[34] Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization. *arXiv preprint arXiv:2403.08730*, 2024. 2

[35] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2024. 1, 2

[36] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. 1

[37] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020. 8

[38] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 1, 2, 3, 4, 7, 8, 12

[39] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6

[40] Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models, 2024. 2

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

[42] Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, et al. Enhancing visual-language modality alignment in large vision language models via self-improvement. *arXiv preprint arXiv:2405.15973*, 2024. 2, 8

[43] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 416–442. Association for Computational Linguistics, 2024. 2

[44] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024. 2, 3

[45] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024. 2, 3, 8, 12

[46] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2

[47] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2023. 3, 4, 12

[48] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024. 2, 7

[49] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024. 2

[50] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023. 2

[51] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 2

[52] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 4

[53] Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 2

[54] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, and Eric Xing. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2024. 2

[55] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024. 2

[56] Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*, 2024. 2, 8