

A. Author Contributions

- Tianyi Xiong primarily executes the LLaVA-Critic project. He collects pointwise samples for the LLaVA-Critic-113k training dataset, trains the models, implements and experiments with iterative DPO, and conducts evaluation of LLaVA-Critic across various scenarios.
- Xiyao Wang contributes to developing pairwise ranking capacity, including organizing response pairs and preferences, designing pairwise prompt templates, and building in-domain pairwise evaluation benchmarks.
- Dong Guo consistently provides technical support throughout the project and contributes to launching ablation study experiments.
- Qinghao Ye and Haoqi Fan offer helpful suggestions and efforts for constructing the critic training dataset.
- Quanquan Gu gives expert guidance on preference learning and helps refine the related sections of the paper.
- Heng Huang provides support for the project and helps with revising the paper.
- Chunyuan Li initiates the idea and scope, directs the project, outlines the roadmap and milestones, supervises the experimental execution, and leads the paper writing.

We would like to extend our gratitude to Renrui Zhang for his insightful discussions and constructive feedback on the evaluation scope of LLaVA-Critic. Special thanks also go to Bo Li for his significant support with LLaVA-OneVision training, SGLang inference, and contributions to the early version of the DPO codebase.

B. Implementation Details

B.1. Evaluation Prompts For Critic Training

Pointwise prompts To construct pointwise critic training data, we adopt the existing evaluation prompts in 7 widely used multimodal evaluation benchmarks that employ GPT-as-a-judge. For further details, please refer to their papers or codebases as listed below:

- *LLaVA-in-the-Wild* [24]: arxiv.org/abs/2304.08485
- *LLaVA-Wilder* [15]: lmms.eval/tasks/llava_wilder
- *ImageDC* [15]: lmms.eval/tasks/internal_eval/dc100_en
- *MMHal-Bench* [38]: arxiv.org/abs/2309.14525
- *MM-Vet* [47]: arxiv.org/abs/2308.02490
- *WildVision-Bench* [28]: arxiv.org/abs/2406.11069
- *RefoMB* [45]: arxiv.org/abs/2405.17220

LLaVA-in-the-Wild, MM-Vet, MMHal-Bench and RefoMB use text-only GPT models for evaluation. We slightly adjust their evaluation prompts to guide the evaluator to fo-

cus on visual information from the input image.

Pairwise Prompt Pool To fully develop LLaVA-Critic’s capacity of ranking LMM responses pairs across diverse scenarios, with varying format requirements and evaluation criteria, we design a set of 30 pairwise evaluation prompt templates for constructing our pairwise critic training data. Two representative prompts are provided in Table 9.

B.2. In-domain Benchmarks for LMM-as-a-Judge

In-domain Pointwise Scoring To evaluate the ability of LLaVA-Critic in judging LMM-generated responses across varying performance levels, we select 13 off-the-shelf LMMs spanning across a wide range of visual chat capabilities, then collect their responses on 7 multimodal benchmarks. The selected response models are listed below:

GPT-4o [32], Claude3-Sonnet [2], LLaVA-NeXT (LLaMA-8B) [26], LLaVA-NeXT (Vicuna-7B) [26], LLaVA-OneVision-7B [16], LLaVA-RLHF-13B [38], LLaVA-v1.5-7B [25], LLaVA-v1.5-13B [25], InstructBLIP-Vicuna-7B [9], InternVL2-8B [8], Phi-3-Vision-128k-Instruct [1], fuyu-8B [5] and Qwen-VL-Chat [3]

C. Additional Experimental Results

C.1. Ablation on Data Curation

Training Data	Score \uparrow	Pair (w. Tie) \uparrow	Pair (w.o. Tie) \uparrow
–	0.151	0.426	0.550
pointwise only	0.311	0.543	0.664
pairwise only	0.249	0.541	0.659
pointwise + pairwise (w.o. Tie)	0.312	0.552	0.683
all	0.314	0.556	0.689

Table 8. Ablation on data curation. We start from LLaVA-OV-7B to train critic models and evaluate them on MLLM-as-a-Judge.

We analyze the impact of different data components in the LLaVA-Critic-113k dataset for critic training. Results are reported in Table 8, presenting the following insights: 1) Pointwise and pairwise critic data are mutually beneficial. 2) Pointwise only data improves pairwise ranking more than the reverse, as relative ranking lacks absolute judgments. 3) Tie pairs further enhance performance by avoiding fixation on minor quality differences.

C.2. Results on In-domain Pointwise Scoring

In Table 10, we report the Kendall’s Tau results to assess the alignment between LLaVA-Critic and GPT-4o in ranking response LMMs. LLaVA-Critic-72B achieves the highest average Kendall’s Tau of 0.933, outperforming the LLaVA-OV-72B baseline (0.802). LLaVA-Critic-7B scores 0.911, significantly surpassing existing advanced LMMs of similar scale. As with the Pearson-r results in Table 2, LLaVA-Critic benefits from both model and data scaling.

Prompt Template 1:

You will be given an image and a corresponding question. Your task is to evaluate the responses provided, select the better response, and give reasons for your choice. Consider the following factors when making your decision.

1. *Accuracy in Object Description*: Evaluate the accuracy of the descriptions concerning the objects mentioned in the ground truth answer. Responses should minimize the mention of objects not present in the ground truth answer, and inaccuracies in the description of existing objects.
2. *Accuracy in Depicting Relationships*: Consider how accurately the relationships between objects are described compared to the ground truth answer. Rank higher the responses that least misrepresent these relationships.
3. *Accuracy in Describing Attributes*: Assess the accuracy in the depiction of objects' attributes compared to the ground truth answer. Responses should avoid inaccuracies in describing the characteristics of the objects present.
4. *Helpfulness*: Consider whether the generated text provides valuable insights, additional context, or relevant information that contributes positively to the user's comprehension of the image. Assess whether the language model accurately follows any specific instructions or guidelines provided in the prompt. Evaluate the overall contribution of the response to the user experience.
5. *Ethical Considerations*: Identify if the model gives appropriate warnings or avoids providing advice on sensitive topics, such as medical images. Ensure the model refrains from stating identification information in the image that could compromise personal privacy. Evaluate the language model's responses for fairness in treating individuals and communities, avoiding biases. Assess for harmfulness, ensuring the avoidance of content that may potentially incite violence, be classified as NSFW (Not Safe For Work), or involve other unmentioned ethical considerations. Consider any content that could be deemed offensive, inappropriate, or ethically problematic beyond the explicitly listed criteria.

The question and responses are given as follows:

Question: [] \n Response A: [] \n Response B: [] \n ASSISTANT:

Prompt Template 2:

As an expert, you are asked to evaluate two responses to the given image-based question. Provide a professional assessment of responses and decide which one is better. Support your decision with detailed reasons. Here are the question and responses:

Question: [] \n The first response: [] \n The second response: [] \n ASSISTANT:

Table 9. Example prompt templates for constructing pairwise training data. The first template covers scenarios with specific *user-defined criteria*, while the second is for general evaluations without additional instructions.

LMM Evaluator	Kendall's τ (\uparrow)							
	ImageDC	MMVet	WildVision	LLaVA-B	LLaVA-W	L-Wilder	MMHal	Avg.
LLaVA-NeXT (LLaMA-8B)	0.452	0.436	0.615	0.487	0.503	0.231	0.590	0.473
LLaVA-Critic (LLaVA-NEXT)	0.787	0.974	0.846	0.839	0.923	0.974	0.923	0.895
Qwen2-VL-7B-Instruct	0.667	0.462	0.128	0.513	0.667	0.974	0.487	0.557
LLaMA3.2-11B-Vision-Instruct	0.590	0.718	0.872	0.667	0.615	0.615	0.452	0.647
LLaVA-OV-7B	0.539	0.154	0.795	0.667	0.641	0.839	0.590	0.603
LLaVA-Critic-7B (v0.5)	0.813	0.897	0.872	0.846	0.949	0.974	0.923	0.896
LLaVA-Critic-7B	0.897	0.949	0.897	0.839	0.923	0.974	0.897	0.911
LLaVA-OV-72B	0.872	0.795	0.821	0.667	0.769	0.949	0.744	0.802
LLaVA-Critic-72B	0.949	0.949	0.949	0.821	0.923	0.994	0.949	0.933

Table 10. Kendall's Tau results on in-domain pointwise scoring. LLaVA-Critic consistently provides model-level rankings more aligned with GPT-4o, surpassing baseline methods across all evaluation tasks.

Settings	MLLM	COCO	C.C.	Diff.	Graphics	Math	Text	WIT	Chart	VisIT	CC-3M	M2W	SciQA	Aes	MM-Vet	Ave.
Score (↑)	GPT-4V*	0.454	0.507	0.458	0.645	0.606	0.624	0.579	0.645	0.620	0.431	0.185	0.383	0.401	0.326	0.490
	Gemini-Pro*	0.262	0.408	-	0.400	0.228	0.222	0.418	0.343	0.336	0.374	0.324	0.073	0.360	0.207	0.304
	LLaVA-1.5-13b*	0.247	0.227	0.060	0.242	0.093	0.245	0.109	0.237	0.177	0.071	0.424	0.279	0.414	0.322	0.225
	GPT-4o [‡]	0.396	0.452	0.341	0.464	0.460	0.564	0.408	0.573	0.589	0.305	0.262	0.569	0.421	0.342	0.439
	GPT-4V [†]	0.410	0.444	0.361	0.449	0.486	0.506	0.457	0.585	0.554	0.266	0.267	0.315	0.472	0.367	0.424
	LLaVA-v1.5-7B	0.205	0.084	0.094	-0.028	0.099	0.235	-0.076	0.212	0.300	0.167	0.392	0.033	0.354	0.147	0.158
	Prometheus-V (LLaVA-v1.5-7B)	0.289	0.342	0.106	0.172	0.182	0.214	0.209	0.224	0.226	0.228	0.089	0.174	0.368	0.157	0.213
	LLaVA-NeXT (LLaMA-8B)	0.288	0.279	0.066	0.249	0.077	0.145	-0.051	0.197	0.304	0.198	0.364	0.128	0.362	0.165	0.198
	LLaVA-OV-7B	0.224	0.024	0.063	0.189	0.097	0.265	-0.135	0.274	0.227	0.081	0.030	0.261	0.249	0.262	0.151
	Qwen2-VL-7B-Instruct	0.295	0.294	0.114	0.196	0.199	0.323	0.075	0.366	0.259	0.143	0.313	0.201	0.434	0.331	0.253
	LLaMA3.2-11B-Vision-Instruct	0.269	0.244	0.107	0.365	0.254	0.305	0.024	0.333	0.264	0.067	0.197	0.370	0.275	0.245	0.237
	LLaVA-OV-72B	0.264	0.390	0.046	0.262	0.358	0.327	0.195	0.290	0.415	0.144	0.359	0.267	0.444	0.253	0.287
	LLaVA-Critic (LLaVA-v1.5-7B)	0.283	0.295	0.095	0.225	0.246	0.257	0.191	0.194	0.253	0.241	0.198	0.188	0.327	0.198	0.228
	LLaVA-Critic (LLaVA-NeXT)	0.272	0.406	0.118	0.257	0.309	0.275	0.292	0.354	0.374	0.225	0.224	0.091	0.432	0.176	0.272
	LLaVA-Critic-7B (v0.5)	0.369	0.456	0.108	0.291	0.325	0.340	0.228	0.388	0.303	0.205	0.232	0.270	0.511	0.338	0.312
	LLaVA-Critic-7B	0.382	0.450	0.103	0.316	0.356	0.378	0.179	0.421	0.322	0.246	0.301	0.269	0.395	0.272	0.314
	LLaVA-Critic-72B	0.333	0.463	0.146	0.452	0.474	0.559	0.396	0.545	0.488	0.273	0.259	0.334	0.403	0.374	0.393
Pair w. Tie (↑)	GPT-4V*	0.696	0.824	0.847	0.639	0.564	0.673	0.679	0.657	0.640	0.612	0.521	0.415	0.606	0.529	0.636
	Gemini-Pro*	0.616	0.787	-	0.650	0.436	0.664	0.605	0.500	0.660	0.560	0.370	0.262	0.190	0.312	0.509
	Qwen-vl-plus*	0.479	0.507	0.650	0.450	0.328	0.522	0.500	0.380	0.453	0.383	0.577	0.321	0.601	0.457	0.472
	GPT-4o [‡]	0.582	0.665	0.829	0.625	0.433	0.477	0.565	0.355	0.577	0.586	0.581	0.427	0.873	0.505	0.577
	GPT-4V [†]	0.539	0.634	0.668	0.632	0.459	0.495	0.536	0.369	0.591	0.544	0.544	0.389	0.620	0.517	0.538
	LLaVA-v1.5-7B	0.460	0.506	0.577	0.452	0.302	0.454	0.375	0.383	0.518	0.478	0.495	0.281	0.466	0.402	0.439
	LLaVA-NeXT (LLaMA-8B)	0.422	0.595	0.435	0.404	0.343	0.431	0.428	0.384	0.508	0.496	0.571	0.336	0.588	0.512	0.461
	LLaVA-OV-7B	0.334	0.471	0.539	0.397	0.318	0.398	0.324	0.374	0.444	0.438	0.556	0.334	0.577	0.456	0.426
	Qwen2-VL-7B-Instruct	0.391	0.358	0.213	0.413	0.368	0.383	0.231	0.308	0.391	0.406	0.324	0.438	0.281	0.367	0.348
	LLaMA3.2-11B-Vision-Instruct	0.493	0.643	0.638	0.601	0.422	0.512	0.485	0.384	0.544	0.591	0.579	0.380	0.662	0.466	0.529
	LLaVA-OV-72B	0.464	0.593	0.667	0.531	0.434	0.485	0.447	0.394	0.549	0.497	0.557	0.428	0.596	0.541	0.513
	LLaVA-Critic (LLaVA-v1.5-7B)	0.564	0.674	0.633	0.505	0.422	0.528	0.538	0.386	0.583	0.608	0.577	0.294	0.681	0.404	0.528
	LLaVA-Critic (LLaVA-NeXT)	0.583	0.684	0.704	0.562	0.438	0.504	0.579	0.339	0.635	0.599	0.581	0.315	0.693	0.441	0.547
	LLaVA-Critic-7B (v0.5)	0.575	0.677	0.730	0.556	0.427	0.521	0.537	0.366	0.568	0.620	0.571	0.353	0.703	0.435	0.546
	LLaVA-Critic-7B	0.593	0.687	0.707	0.587	0.432	0.544	0.564	0.338	0.596	0.628	0.591	0.370	0.686	0.464	0.556
	LLaVA-Critic-72B	0.587	0.672	0.860	0.588	0.475	0.536	0.618	0.366	0.628	0.608	0.568	0.390	0.721	0.473	0.578
Pair w.o. Tie (↑)	GPT-4V*	0.804	0.870	0.922	0.807	0.801	0.805	0.734	0.849	0.761	0.703	0.699	0.647	0.755	0.659	0.773
	Gemini-Pro*	0.717	0.840	-	0.770	0.678	0.793	0.688	0.658	0.711	0.652	0.471	0.358	0.265	0.400	0.615
	LLaVA-1.6-34b*	0.607	0.824	0.855	0.402	0.587	0.750	0.758	0.381	0.503	0.564	0.712	0.679	0.694	0.762	0.648
	GPT-4o [‡]	0.774	0.776	0.934	0.835	0.628	0.618	0.737	0.513	0.741	0.770	0.706	0.722	0.887	0.660	0.736
	GPT-4V [†]	0.729	0.772	0.884	0.853	0.665	0.661	0.760	0.495	0.785	0.707	0.697	0.639	0.741	0.654	0.717
	LLaVA-v1.5-7B	0.617	0.571	0.637	0.598	0.411	0.544	0.452	0.554	0.653	0.562	0.672	0.600	0.558	0.631	0.576
	LLaVA-NeXT (LLaMA-8B)	0.565	0.684	0.473	0.526	0.460	0.526	0.516	0.549	0.634	0.592	0.641	0.648	0.673	0.716	0.586
	LLaVA-OV-7B	0.462	0.562	0.588	0.530	0.434	0.473	0.400	0.543	0.563	0.527	0.639	0.670	0.633	0.679	0.550
	Qwen2-VL-7B-Instruct	0.662	0.797	0.579	0.753	0.576	0.600	0.418	0.444	0.735	0.712	0.712	0.642	0.848	0.550	0.645
	LLaMA3.2-11B-Vision-Instruct	0.647	0.726	0.688	0.774	0.570	0.599	0.577	0.567	0.667	0.692	0.654	0.708	0.684	0.662	0.658
	LLaVA-OV-72B	0.691	0.780	0.811	0.714	0.623	0.634	0.625	0.570	0.737	0.685	0.755	0.702	0.746	0.736	0.701
	LLaVA-Critic (LLaVA-v1.5-7B)	0.732	0.757	0.665	0.659	0.574	0.623	0.650	0.544	0.710	0.719	0.649	0.589	0.708	0.602	0.656
	LLaVA-Critic (LLaVA-NeXT)	0.763	0.775	0.762	0.720	0.599	0.606	0.705	0.491	0.756	0.716	0.682	0.598	0.725	0.577	0.677
	LLaVA-Critic-7B (v0.5)	0.747	0.758	0.771	0.716	0.580	0.625	0.661	0.525	0.692	0.729	0.697	0.632	0.728	0.585	0.675
	LLaVA-Critic-7B	0.771	0.774	0.755	0.758	0.596	0.658	0.680	0.488	0.727	0.742	0.692	0.658	0.715	0.635	0.689
	LLaVA-Critic-72B	0.762	0.762	0.904	0.755	0.637	0.648	0.763	0.528	0.769	0.718	0.693	0.708	0.742	0.624	0.715

Table 11. Comprehensive results on MLLM-as-a-Judge. *: the results of GPT-4V, Gemini-pro, and the best open-source LMM as reported in the original paper [6]; [‡]: results from our evaluation of GPT-4V/4o using its original codebase.

Method	LLaVA-W	L-Wilder	WildVision	LiveBench	MMHal	MME ^P	MME ^C	MMB-en	MM-Vet	MMStar
LLaVA-OV-7B	90.7	67.8	54.0	77.1	3.19	1580.4	418.2	80.8	57.5	61.7
LLaVA-Critic-7B	90.9	69.4	54.5	76.2	3.45	1599.0	455.4	79.3	58.4	58.9
LLaVA-OV-72B	93.5	72.0	51.7	81.5	3.61	1683.2	578.9	85.9	63.7	66.1
LLaVA-Critic-72B	93.8	73.4	54.8	81.9	3.74	1726.7	597.1	84.8	62.3	64.2

Table 12. Comparison between LLaVA-Critic and LLaVA-OneVision on multimodal evaluation benchmarks. Results are reported for 5 open-ended visual chat benchmarks (left) and 4 comprehensive benchmarks (right).

Method	LLaVA-W	L-Wilder	WildVision	LiveBench	VideoDC	MMHal	MME ^P	MME ^C	MMB-en	MM-Vet	MMStar
GPT-4V	98.0	81.0	79.8	73.7	4.00	3.83	1409.4	517.1	75.0	49.9	57.1
LLaVA-OV-7B	90.7	67.8	54.0	77.1	3.75	3.19	1580.4	418.2	80.8	57.5	61.7
+ LLaVA-Critic-7B iter-1	96.7	70.6	60.5	81.2	3.77	3.62	1561.8	420.7	80.8	54.5	62.1
+ LLaVA-Critic-7B iter-2	97.0	72.2	65.2	83.9	3.82	3.67	1565.5	415.4	80.7	54.6	62.1
+ LLaVA-Critic-7B iter-3	100.3	71.6	67.3	84.5	3.87	3.91	1555.3	414.6	80.3	54.4	62.3
+ GPT-4o iter-3	100.6	71.3	67.1	86.4	—	3.79	1565.3	405.0	80.3	54.8	59.0
LLaVA-OV-72B	93.5	72.0	51.7	81.5	3.60	3.61	1683.2	578.9	85.9	63.7	66.1
+ LLaVA-Critic-72B iter-1	99.3	75.3	65.7	86.4	3.83	3.75	1683.2	584.3	85.6	67.0	66.4
+ LLaVA-Critic-72B iter-2	104.1	75.6	68.4	86.6	3.86	3.75	1681.1	586.4	85.7	66.7	66.5
+ LLaVA-Critic-72B iter-3	104.4	75.9	70.0	88.5	3.86	3.77	1686.1	586.4	85.4	67.1	66.4

Table 13. Performance of difference rounds of iterative DPO on LLaVA-OneVision. With the high-quality feedback from LLaVA-Critic, both LLaVA-OneVision 7B and 72B learn to refine its self-generated responses in a progressive manner, leading to overall better performance across various open-ended multimodal benchmarks.

C.3. Results on MLLM-as-a-Judge

We include the comprehensive results on MLLM-as-a-Judge [6] in Table 11. Since the complete evaluation pipeline has not been released, we re-implement the evaluation code based on their official codebase.

C.4. Results on General Visual Tasks

We evaluate LLaVA-Critic on comprehensive multimodal benchmarks, including both short-answer and open-ended visual tasks. As shown in Table 12, training on critic instruction-following data effectively preserves the original capabilities for handling a wide range of visual tasks, with slight improvements in visual chat performance. The LLaVA-Critic-7B model outperforms LLaVA-OV-7B on 4 of 5 open-ended visual chat benchmarks, while the Critic-72B model outperforms OV-72B on all 5.

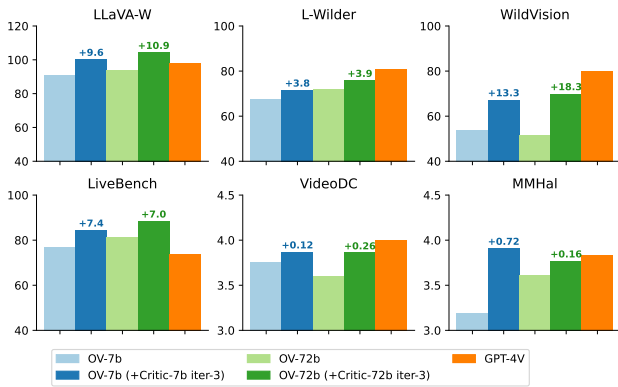


Figure 3. Visualization on the performance gain of LLaVA-OneVision obtained through LLaVA-Critic across 6 open-ended multimodal benchmarks.

C.5. Results on Preference Learning

We present the comprehensive results of applying multi-round iterative DPO on the base model, using LLaVA-Critic for reward signals. As shown in Table 13, this ap-

proach enhances LLaVA-OneVision’s performance across six open-ended visual chat benchmarks, with consistent gains per round. Notably, on the LLaVA-OV-7B base model, LLaVA-Critic boosts improvement comparable to GPT-4o while saving approximately \$690 in 3 iterative DPO rounds. For other comprehensive benchmarks, the iterations offer minimal benefit, while LLaVA-OneVision’s overall capabilities are largely preserved. Figure 3 visualizes the performance gains on visual chat benchmarks.

D. More Qualitative Comparisons

We provide detailed examples of LLaVA-Critic in serving as a generalist evaluator for LMM responses, and improving visual chat through preference alignment.

- Table 14 shows an example of LLaVA-Critic ranking a pair of LMM responses. By accurately recognizing the visual content of the input image and grounding the differences between the responses, LLaVA-Critic offers judgments consistent with human evaluators, along with clear justifications.
- Table 15 presents an example of LLaVA-Critic scoring an LMM response using a reference answer from GPT-4. LLaVA-Critic closely follows the evaluation prompt and, by referring to the image content, accurately identifies the strengths and weaknesses of the response at both overall and fine-grained levels. Consequently, it delivers a fair evaluation score and detailed justification aligned with GPT-4o.
- Table 16 illustrates LLaVA-Critic’s effectiveness in enhancing LLaVA-OneVision’s visual chat performance through iterative DPO training. By progressively aligning with LLaVA-Critic’s preferences, LLaVA-OneVision learns to generate responses that are more structured, detailed and better tailored to user needs.

Example 1: LLaVA-Critic in Judging LMM Responses (Pairwise)

Critic Input	<div data-bbox="435 562 654 779" data-label="Image"> </div> <p>Question: What this image presents?</p> <p>Response 1 (model_a): The image is a black and white sketch of a line that appears to be in the shape of a cross. The line is a simple and straightforward representation of the cross shape, with two straight lines intersecting at a point.</p> <p>Response 2 (model_b): This is a handwritten number seven.</p> <p>System: 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:</p> <p>Question: []\n Response 1: []\n Response 2: []\n ASSISTANT:</p>
Human	The winner is Response 2.
LLaVA-Critic-7B	<p>Response 2 is better than Response 1 because it accurately identifies the image as a handwritten number seven. Response 1 incorrectly describes the image as a black and white sketch of a line in the shape of a cross, which does not match the visual content of the image. Response 2 correctly interprets the image, providing a precise and accurate description that aligns with the actual content of the image. This makes Response 2 more helpful and accurate for understanding what the image presents.</p>
LLaVA-OV-7B	<p>Response 1 is better because it accurately describes the image as a simple and straightforward representation of a cross shape, which is what the image actually depicts. Response 2, while also describing the image, does not accurately convey the intended meaning of the image, as it suggests that the image is a handwritten number seven, which is not the case.</p>

Table 14. Example of LLaVA-Critic in ranking LMM response pairs. While LLaVA-OV misinterprets the handwritten number and gives inaccurate judgment, LLaVA-Critic aligns with human assessments and offers clear, visually grounded justifications.

Example 2: LLaVA-Critic in Judging LMM Responses (Pointwise)



Critic Input

Question: Can you describe the weather conditions and overall atmosphere of the scene depicted in the image?

Assistant 1 (GPT-4): The image depicts a calm and serene beach scene. The sky appears to be clear with no signs of clouds, indicating good weather conditions. The calmness of the ocean, suggested by the gentle waves crashing on the shore, further enhances the tranquility. The presence of a colorful beach umbrella provides a vibrant contrast to the more subdued hues of the sea and sand, implying a relaxed and leisurely atmosphere commonly associated with a sunny day at the beach.

Assistant 2 (Evaluated LMM response): The image depicts a sunny and pleasant day at the beach, with a colorful umbrella providing shade on the sandy shore. The presence of the umbrella suggests that it is likely a hot day, and people might be seeking relief from the sun by sitting under the umbrella or going for a swim in the ocean. The atmosphere is likely relaxed and enjoyable, with people spending time at the beach to unwind and take in the beautiful surroundings. The sunny weather and the calming sound of the waves further contribute to the overall pleasant and leisurely vibe of the scene.

System: We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above. The user asks the question on observing an image shown to you. Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance. Assume assistant 1 always receive a score of 10 and is the correct answer. Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

GPT-4o

10 9

Assistant 1 provided a detailed and accurate description of the weather conditions and overall atmosphere of the scene depicted in the image. The response **highlighted the clear sky, calm ocean, and the presence of a colorful beach umbrella, which all contribute to a serene and tranquil atmosphere.** The description was comprehensive and captured the essence of the scene effectively. Assistant 2 also provided a good description, **noting the sunny and pleasant weather, the colorful umbrella, and the relaxed atmosphere.** However, the response **included some assumptions, such as people seeking relief from the sun and going for a swim, which were not directly observable in the image.** This slightly reduced the accuracy and relevance of the response.

LLaVA-Critic-7B

10 9

Assistant 1 provided a detailed and accurate description of the weather conditions and overall atmosphere depicted in the image. The response **highlighted the clear sky, calm ocean, and the presence of a colorful beach umbrella, which all contribute to a serene and tranquil beach scene.** The description was relevant, accurate, and included a good level of detail. Assistant 2 also gave a good description, **mentioning the sunny day, the colorful umbrella, and the relaxed atmosphere.** However, **it made an assumption about people seeking relief from the sun and going for a swim, which is not directly supported by the image. This slightly detracts from the accuracy and relevance of the response.** Therefore, Assistant 2 receives a slightly lower score.

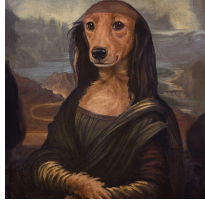
LLaVA-OV-7B

10 9

Assistant 1 provided a clear and detailed description of the image, including the weather conditions and overall atmosphere. **The response was relevant to the question asked and accurately described the scene depicted in the image. The level of detail was appropriate for the task, providing enough information for the user to understand the context of the image.** Assistant 2 also provided a clear and detailed description of the image, but it was slightly less concise than Assistant 1's response. **While the response was still relevant to the question asked and accurately described the scene depicted in the image, it did not provide as much detail as Assistant 1's response. However, the level of detail provided by Assistant 2 was still sufficient for the task, and the response was still helpful and accurate. Overall, both assistants performed well in their responses, but Assistant 1 provided a slightly more detailed and comprehensive response, which is why it receives a score of 10, while Assistant 2 receives a score of 9.**

Table 15. Example of LLaVA-Critic conducting pointwise scoring on an LMM response for LLaVA-Wilder. Unlike LLaVA-OV, which offers **vague and uninformative explanations**, LLaVA-Critic delivers a thorough and fair assessment with clear justifications on both the **strengths** and **weaknesses** of the evaluated response, closely resembling the depth and fairness of GPT-4o's assessments.

Example 3: LLaVA-Critic Enhances LLaVA-OneVision in Visual Chat



Question	Discuss how this creative twist on a classic work of art might be interpreted differently by various audiences.
LLaVA-OV-72B	Different audiences might interpret this creative twist on the Mona Lisa in various ways. Art enthusiasts might appreciate the skillful execution and the playful reinterpretation of a classic, seeing it as a commentary on the universality of beauty and expression across species. Dog lovers could find joy in seeing their favorite breed portrayed with such dignity and elegance. Some might view it as a humorous take on high art, making it more accessible and relatable. Others might see it as a statement on the anthropomorphism of pets, reflecting how we often attribute human-like qualities to animals. Overall, the varied interpretations highlight the versatility of art in engaging diverse perspectives and emotions.
LLaVA-OV-72B + Critic-72b iter-3	<p>This creative twist on the Mona Lisa, featuring a dog's face, can be interpreted in various ways by different audiences.</p> <p>For art enthusiasts, it might be seen as a playful commentary on the nature of portraiture and the human tendency to anthropomorphize animals. It challenges traditional notions of beauty and subjectivity in art, inviting viewers to reconsider what makes a portrait compelling.</p> <p>For dog lovers and pet owners, the image could evoke a sense of joy and familiarity, as they see their beloved pets reflected in one of the most iconic artworks in history. It might also highlight the deep bond between humans and their pets, suggesting that animals share in our cultural heritage.</p> <p>In a broader social context, this image could be interpreted as a reflection of contemporary society's obsession with internet culture and memes. The substitution of the Mona Lisa's face with a dog's could be seen as a satirical take on how viral content often overshadows classical art in modern media.</p> <p>Academically, this piece could spark discussions about copyright, artistic originality, and the public domain. It raises questions about the boundaries between homage and appropriation in art, encouraging critical thinking about intellectual property rights and creative freedom.</p> <p>Overall, the diverse interpretations of this image underscore the power of art to engage, provoke, and connect people across different backgrounds and perspectives.</p>

Table 16. Example of LLaVA-Critic in preference learning. Leveraging reward signals from LLaVA-Critic, LLaVA-OneVision improves its visual chat ability through iterative DPO, delivering more detailed, valuable, and structured point-by-point responses.