PROMETHEUS-VISION: Vision-Language Model as a Judge for Fine-Grained Evaluation

Seongyun Lee[1*] **Seungone Kim**[1,2,3*] **Sue Hyun Park**[1] **Geewook Kim**[1,4] **Minjoon Seo**[1]

KAIST AI[1] Carnegie Mellon University[2] NAVER AI Lab[3] NAVER Cloud AI[4]
{seongyun, suehyunpark, geewook, minjoon}@kaist.ac.kr seungone@cmu.edu

Abstract

Assessing long-form responses generated by Vision-Language Models (VLMs) is challenging. It not only requires checking whether the VLM follows the given instruction but also verifying whether the text output is properly grounded on the given image. Inspired by the recent approach of evaluating LMs with LMs, in this work, we propose to evaluate VLMs with VLMs. For this purpose, we present a new feedback dataset called the PERCEPTION COLLECTION, encompassing 15K customized score rubrics that users might care about during assessment. Using the PERCEPTION COLLECTION, we train PROMETHEUS-VISION, the first open-source VLM evaluator model that can understand the user-defined score criteria during evaluation. PROMETHEUS-VISION shows the highest Pearson correlation with human evaluators and GPT-4V among open-source models, showing its effectiveness for transparent and accessible evaluation of VLMs. We open-source our code, dataset, and model at `https://anonymous.4open.science/r/prometheus-vision-9D37`.

1 Introduction

While recently developed Vision-Language Models (VLMs) are capable of generating long-form text from a combination of an image and instruction, assessing the quality of the output remains a significant challenge (Liu et al., 2023a; Dai et al., 2023; Gao et al., 2023; Ye et al., 2023a; Zhu et al., 2023a; OpenAI, 2023). Traditional metrics, which rely on text-based exact matches or edit distances, fall short in adhering to the granular evaluation criterion of interest and capturing the rich context within the outputs (Agrawal et al., 2023; Mañas et al., 2023; Bai et al., 2023). For instance, as shown in Figure 1, conventional metrics fail to explain what is missing within the response compared to the answer.

Consequently, the role of high-quality human evaluations remains pivotal for a comprehensive assessment. However, human evaluators are prone to biases and scaling up is expensive in terms of time and cost (Ye et al., 2023c; Kim et al., 2023b).

To address the need for flexible and automatic text evaluation, the 'LM-as-a-Judge' paradigm proposes using language models (LMs) as evaluators, where initial findings suggest its potential to emulate human judgement (Liu et al., 2023a; Zheng et al., 2023; Li et al., 2023; Ye et al., 2023c; Kim et al., 2023d; Zhu et al., 2023b; Bai et al., 2023). However, LMs cannot perceive visual contexts, which necessitates an additional model that could convert the image to text. As a result, such a multistage pipeline could potentially suffer from error propagation and also require multiple inference calls. This situation calls for the direct utilization of VLMs, referred to as **VLM-as-a-Judge**.

However, despite GPT-4V's (OpenAI, 2023) potential as an evaluator, its closed-source nature limits transparent evaluation (Kim et al., 2023d). On the other hand, our initial tests indicate that open-source VLMs are not appropriate to be used as evaluators to assess based on granular aspects, demonstrating a low score correlation with both human evaluators and GPT-4V. To address these challenges, we propose PROMETHEUS-VISION, a 7B & 13B VLM evaluator that could assess based on fine-grained criteria. As shown in Figure 1, PROMETHEUS-VISION could evaluate based on the given criteria, pinpointing the differences between the parody artwork and the original masterpiece.

To develop PROMETHEUS-VISION, we construct the PERCEPTION COLLECTION, the first multimodal feedback dataset that includes 15K fine-grained score rubrics, thus going beyond traditional coarse-grained criteria such as helpfulness, relevance, accuracy, and comprehensiveness. Using

[*] denotes equal contribution. Work was done when Seongyun was an intern at KAIST AI and Seungone was an intern at NAVER AI Lab.

Findings of the Association for Computational Linguistics: ACL 2024, pages 11286–11315
August 11-16, 2024

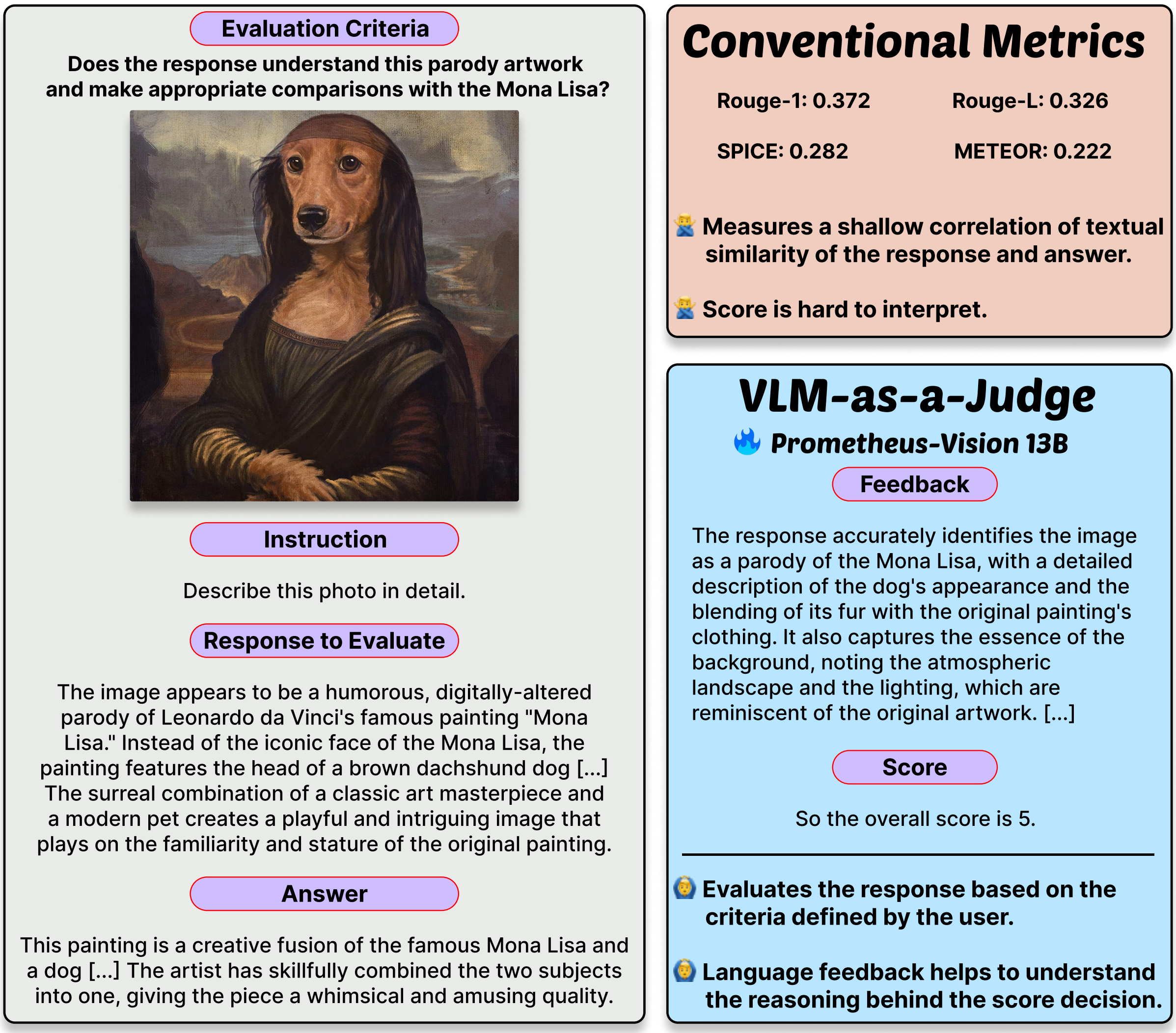


Figure 1: Conventional metrics measure the similarity between the response and ground-truth answer, which is not expressive enough. Moreover, it could not pinpoint what is missing within the response with respect to the evaluation criteria. In contrast, our **VLM-as-a-Judge** pipeline provides not only the flexibility to adhere to arbitrary evaluation criteria but also provides detailed language feedback that specifically pinpoints the deficiencies.

PERCEPTION COLLECTION, we fine-tune LLaVA-1.5 to obtain PROMETHEUS-VISION.

Surprisingly, PROMETHEUS-VISION shows the highest correlation with both human evaluators and GPT-4V on 3 visual instruction tuning benchmarks, 3 visual question answering datasets, and 2 captioning benchmarks among the open-source models. This highlights PROMETHEUS-VISION's potential to serve as an inexpensive yet effective open-source alternative to GPT-4V evaluation. For instance, PROMETHEUS-VISION-13B exhibits a Pearson correlation of 0.639 with human evaluators and 0.786 with GPT-4V on the LLaVA-Bench, one of the most widely used visual instruction tuning benchmark. Moreover, PROMETHEUS-VISION also exhibits high accuracy on human preference datasets, indicating that PROMETHEUS-VISION could potentially function as a reward model in the Reinforcement Learning from Human Feedback (RLHF) pipeline.

Our contributions are summarized as follows:

- We introduce PERCEPTION COLLECTION, the first multimodal feedback dataset that could be used to train an evaluator VLM. In contrast to existing multimodal feedback, critique, and preference datasets that use coarse-grained criteria, PERCEPTION COLLECTION includes 15K fine-grained criteria that determine the crucial aspect for each instance.

- We introduce PROMETHEUS-VISION, the first open-source VLM specialized for evaluation purposes. PROMETHEUS-VISION shows a high correlation with both GPT-4V and human evaluators, indicating its potential to be used as a cheap alternative for GPT-4V evaluation.

2 Related works

2.1 Evaluating Vision Language Models

In prior works, Vision-Language Models (VLMs) are typically evaluated using specific metrics tailored to each task. For image captioning, performance is measured with metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015), focusing on how well the generated text aligns with reference captions. Similarly, Visual Question Answering (VQA) is evaluated using accuracy metrics based on the exact match between the model's answers and human-annotated answers (Agrawal et al., 2023; Mañas et al., 2023).

However, traditional metrics often fall short of capturing the nuanced details of the response generated by VLMs in complex or subjective situations. A more comprehensive approach has been human evaluation, accounting for contextual and creative aspects not captured by automated metrics. Nonetheless, cost and consistency constraints associated with human evaluations render it a less feasible method for scaling to a lot of instances.

2.2 Language Model as a Judge for Fine-grained Evaluation

The difficulty in evaluating long-form responses often arises from the ambiguity in defining what constitutes a good output. For instance, discerning whether a given response is helpful or harmless is often subjective. Recent works have proposed the concept of 'Fine-grained Evaluation', utilizing LM-as-a-judge for assessing granular aspects. Ye et al. (2023c) defines 12 core skill sets that are crucial for evaluating LMs. Kim et al. (2023d) further extends this concept and employs thousands of fine-grained criteria to assess LMs on user-defined criteria. Wu et al. (2023) and Jang et al. (2023) utilize fine-grained criteria to align LMs. Lastly, Kim et al. (2023e) proposes an interactive framework in which users could test LMs on fine-grained criteria.

To the best of our knowledge, we are first to expand the notion of 'Fine-grained Evaluation' for assessing VLMs. Specifically, recent work has proposed to evaluate VLMs using LMs or VLMs (Bai et al., 2023; Ge et al., 2023), yet are still confined to high-level coarse-grained criteria such as helpfulness, relevance, accuracy, and comprehensiveness. We construct the PERCEPTION COLLECTION which encompasses 15K of fine-grained criteria and use it to train PROMETHEUS-VISION.

3 The PERCEPTION COLLECTION

In contrast to the language domain, to the best of our knowledge, there do not exist any available feedback, critique, or preference datasets applicable to train an evaluator VLM that could assess in a *fine-grained* manner. For this purpose, we first construct a comprehensive multimodal feedback dataset called the PERCEPTION COLLECTION.

As shown in Figure 2, each instance in the PERCEPTION COLLECTION consists of five input components (image, instruction, response to evaluate, customized score rubric, reference answer) and two output components (language feedback and score decision). The number of each component in the PERCEPTION COLLECTION is shown in Table 1.

Specifically, the five input components are:

- **Image**: A real-world image that the user would provide to the VLM.
- **Instruction**: A text instruction that the user would prompt the VLM. It is also related to the provided image.
- **Response to Evaluate**: A text response that the VLM would generate based on the image and instruction. The evaluator VLM has to assess this response.
- **Customized Score Rubric**: A detailed scoring criteria that the VLM should refer to for assessment. We use fine-grained criteria in contrast to coarse-grained ones such as helpfulness, relevance, accuracy, and comprehensiveness. The rubric consists of (1) a description of the criteria and (2) a description of each scoring decision on a scale of 1 to 5.
- **Reference Answer**: A reference answer that would achieve a score of 5. While this component could be hand-crafted by human annotators, in our experiments, we utilize GPT-4V.

Moreover, the two output components are:

- **Feedback**: A rationale pinpointing what is good and bad about the response under assessment. Instead of directly providing a scoring decision, this component makes the judgement process more interpretable.

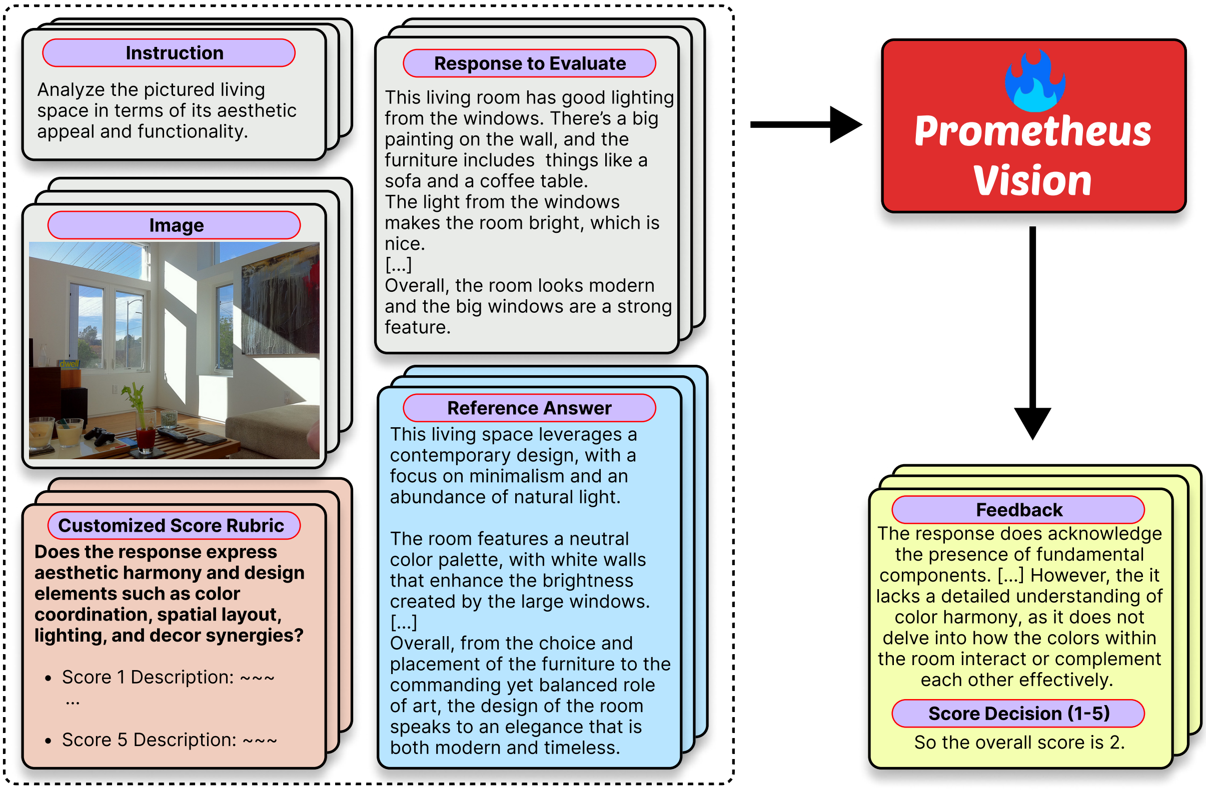


Figure 2: Previous automatic metrics could not capture whether a VLM's response is aware of *aesthetic harmony*. With PROMETHEUS-VISION, users could define customized score rubrics that they care about instead of assessing based on coarse-grained criteria such as helpfulness, relevance, accuracy, and comprehensiveness. Each component within the PERCEPTION COLLECTION consists of 5 input components: an instruction, a real-world image, a response to evaluate, a customized score rubric, and a reference answer. Based on this, PROMETHEUS-VISION is trained to generate a language feedback and a score decision.

Components	# Components	# Components per Image
Images	5,000	1
Score Rubrics	15,000	3
Instructions	30,000	6
Reference Answers	30,000	6
Responses	150,000	30
Feedback & Score	150,000	30

Table 1: The number of each component included in the PERCEPTION COLLECTION. Note that the feedback and score are evenly distributed, leading to 30K instances per score between 1 and 5.

- **Score**: An integer value on a scale of 1 to 5 that represents the quality of the response given the criteria mentioned in the score rubric.

3.1 PERCEPTION COLLECTION Construction

We construct a multimodal feedback dataset called the PERCEPTION COLLECTION. We mainly follow the construction process of Kim et al. (2023d). While creating the PERCEPTION COLLECTION, we utilize 5K real-world images sampled from MS COCO 2017 Challenge (Lin et al., 2014) and the MMMU benchmark (Yue et al., 2023).

Concretely, the augmentation process consists of 4 stages: (1) hand-crafting 50 seed score rubrics, (2) brainstorming 15K fine-grained score rubrics, (3) augmenting 30K instructions and reference answers closely tied with the rubric, and (4) augmenting 150K responses and language feedback. We include a detailed analysis of the PERCEPTION-COLLECTION in terms of diversity and quality in Appendix E and all the prompts used for augmentation in Appendix J.

Step 1: Hand-Crafting Score Rubrics We first start by manually writing 50 examples of fine-grained score rubrics that are more detailed than coarse-grained ones (*e.g.,* helpfulness, relevance, accuracy, and comprehensiveness). For 50 images, the authors write a score rubric that includes a criteria and an description for each score that pinpoints the crucial aspects to consider during the assessment. This serves as our seed data for expanding it in the next step.

Step 2: Brainstorming Score Rubrics By prompting GPT-4V, we expand the number of our score rubrics from 50 to 15K. Using an arbitrary image among the 5K pool and the 50 examples as demonstrations, GPT-4V generates 3 score rubrics for each image. To ensure quality, we go through an additional stage of prompting GPT-4V to inspect whether the generated score rubric *aligns* with the image. If it does not, we iteratively prompt it again until we acquire 3 candidates per image.

Step 3: Augmenting Instructions and Reference Answers related to the Score Rubric Afterwards, we use the 15K score rubrics and prompt GPT-4V to generate 2 novel instructions for each score rubric, leading to a total number of 30K. This process ensures that the instruction is closely tied to the score rubric since the instruction was conditioned on the score rubric during augmentation. During this process, we also prompt GPT-4V to generate a reference answer (*i.e.*, a response that would get a score of 5 according to the score rubric) along with the instruction.

Step 4: Augmenting Training Instances Lastly, we augment the remaining components which are the response to evaluate and language feedback. We use the score rubric, instruction, and reference answer generated from the previous stages and prompt GPT-4V to write a response that would get a score of i ($1 \leq i \leq 5$). This leads to a total number of 150K responses and 150K feedback where each score within a range of 1 to 5 has an even number of 30K instances.

Moreover, we include our analysis of the PERCEPTION COLLECTION in terms of its quality, diversity, and whether there is a length bias among score decisions in Appendix E.

3.2 Fine-tuning a VLM as an Evaluator

Using the PERCEPTION COLLECTION, we use LLaVA-1.5 (7B & 13B) (Liu et al., 2023a) as our backbone model and train PROMETHEUS-VISION (7B & 13B). Training on the PROMETHEUS COLLECTION is analogous to Chain-of-Thought fine-tuning which requires generating a rationale (which is the feedback in our case) and then the score in a sequential manner (Ho et al., 2022; Kim et al., 2023c). We include a fixed phrase '`So the overall score is`' in between the feedback and the score, which we found to prevent degeneration during inference. The detailed hyper-parameters used during training are included in Appendix H.1.

4 Experimental Settings

4.1 Protocol for Evaluating Evaluator VLMs

In this section, we explain our experimental setting used to assess the fine-grained judgement capabilities of evaluator VLMs. As it is a non-trivial problem to directly measure '*How well a VLM is evaluating*', we indirectly compare with two different standards: (1) how closely PROMETHEUS-VISION could simulate human evaluators (Section 5.1) and (2) how closely PROMETHEUS-VISION could simulate the best VLM, which is GPT-4V, for nuanced assessment purposes (Section 5.2).

4.2 Evaluator VLM & LM Baselines

We employ 9 VLMs as our evaluator VLM baselines, namely LLaVA-1.5 (7B & 13B) (Liu et al., 2023a); LLaVA-RLHF (7B & 13B) (Sun et al., 2023); ShareGPT4V (7B) (Chen et al., 2023); Fuyu (8B) (Bavishi et al., 2023); GPT-4V (OpenAI, 2023); and PROMETHEUS-VISION (7B & 13B).

In addition, we also compare with using LMs as a judge for evaluating responses from VLMs as in previous work (Bai et al., 2023). We add 4 LMs as our evaluator LM baselines, namely Prometheus (7B & 13B) (Kim et al., 2023d); GPT-3.5-Turbo (OpenAI, 2022); and GPT-4 (OpenAI, 2023). Since LMs could not receive images as input, we prompt LLaVA-1.5 to generate a caption for the given image and provide the caption as additional input for LM evaluators. In contrast, for VLM evaluator baselines, we directly provide the image as input. The hyper-parameters used to inference evaluator LMs and evaluator VLMs are included in Appendix H.1.

4.3 Response VLMs

We utilize 3 different VLMs to sample the outputs that our VLM evaluators would assess. We denote these 3 VLMs as 'Response VLMs'. We utilize Fuyu (8B), LLaVA-1.5 (13B), and GPT-4V as our response VLM. The hyper-parameters used to inference response VLMs are included in Appendix H.1.

4.4 Benchmarks

Our evaluation benchmarks are mainly divided into 3 categories:

- **Visual Instruction Following Benchmarks**: Tasks that require to write a long-form text output given an image and a text instruction. We

Benchmarks	# Instances	# Score Rubrics
LLaVA-Bench	15	15 (Hand-crafted)
VisIT-Bench	15	15 (Hand-crafted)
PERCEPTION-BENCH	15	15 (Hand-crafted)
Total	45	45

Table 2: The number of the instances and score rubrics included in our evaluation setting in Section 5.1. We randomly sample 15 instances from each benchmark and hand-craft a instance-wise fine-grained score rubric. Each instance originally has an image and an instruction.

use LLaVA-Bench (Liu et al., 2023a), VisIT-Bench (Bitton et al., 2023), and a held-out test set of the PERCEPTION COLLECTION called the PERCEPTION BENCH.

- **Visual Question Answering Benchmarks**: Tasks that require to write a text output given an image and a text question. Compared to instruction following benchmarks, one notable difference is that we use the short-form answers originated from each dataset as reference answers in the input. We use the test set of the OKVQA dataset (Marino et al., 2019), VQAv2 dataset (Goyal et al., 2017), and TextVQA dataset (Singh et al., 2019).
- **Captioning Benchmarks**: Tasks that require to write a text caption of the given image. Similar to the visual question answering benchmarks, the ground truth answers tend to be short compared to the reference responses in the instruction following benchmarks. We use the test set of the COCO-Captions dataset (Chen et al., 2015) and No-Caps dataset (Agrawal et al., 2019).

The number of instances and score rubrics for each benchmark is shown in Table 2 and Table 3. Note that while the datasets in the VQA and captioning benchmarks originally have ground-truth answers, the instruction following benchmarks inherently does not have a reference answer. Using the same augmentation process mentioned in Section 3.1, we augment a reference answer and a fine-grained score rubric for each instance within the LLaVA-Bench, VisIT-Bench, and PERCEPTION-BENCH. For the PERCEPTION-BENCH, which is our held-out test set, we also generate new instructions. For the VQA and captioning benchmarks, we generate 5 score rubrics with the original ground-truth answer in consideration. The authors manually checked the quality of the added components.

Benchmarks	# Instances	# Score Rubrics
LLaVA-Bench	60	60 (Machine-generated)
VisIT-Bench	500	500 (Machine-generated)
PERCEPTION-BENCH	500	500 (Machine-generated)
OKVQA	500	5 (Machine-generated)
VQAv2	500	5 (Machine-generated)
TextVQA	500	5 (Machine-generated)
COCO-Captions	500	5 (Machine-generated)
No-Caps	500	5 (Machine-generated)
TOTAL	3560	1085

Table 3: The number of the instances and score rubrics included in our evaluation setting in Section 5.2. Except for LLaVA-Bench, we randomly sample 500 instances from each benchmark. Each instance originally has an image and an instruction. We additionally add a fine-grained score rubric and reference answer by prompting GPT-4V as explained in Section 3.1.

4.5 Setups & Metrics

Our evaluation setup is divided into 2 parts.

Setup #1 (Table 2) In Section 5.1, we utilize 45 instances with instance-wise hand-crafted score rubrics (15 instances each for LLaVA-Bench, VisIT-Bench, and PERCEPTION-BENCH). We ask 9 human annotators proficient in English to provide a scoring decision as PROMETHEUS-VISION. Then, we measure the correlation of the scoring decision by employing **Pearson**, **Kendall-Tau**, and **Spearman** as our metrics. Next, we ask human annotators to compare 2 language feedbacks that are sampled from either GPT-4, GPT-4V, or PROMETHEUS-VISION (13B) and choose which one is better. Then, we measure the **Pairwise Preference Win-rate** between the 3 candidates. Details of the annotation setting are explained in Appendix H.2.

Setup #2 (Table 3) In Section 5.2, we expand the number of instances and utilize 1,085 fine-grained score rubrics tied across 3,560 instances in total. In this setting, we prompt GPT-4V three times and compare the correlation of the scoring decision by also prompting evaluator VLMs and evaluator LMs three times. As Setup #1, we use **Pearson**, **Kendall-Tau**, and **Spearman** as our metrics.

5 Experimental Results

5.1 Can PROMETHEUS-VISION Closely Simulate Human Evaluators?

To verify whether PROMETHEUS-VISION can emulate human evaluators, we measure the correlation between scores annotated by humans and those predicted by evaluator VLMs.

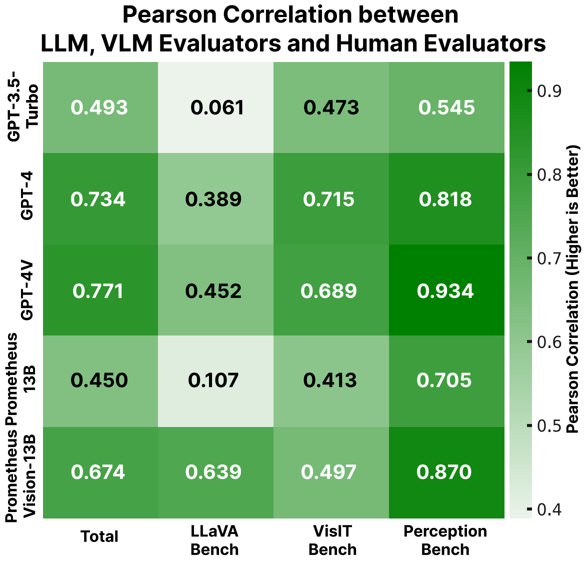


Figure 3: Pearson Correlation between score decisions from human evaluators and score decisions from either GPT-4V, GPT-4, GPT-3.5-Turbo, PROMETHEUS-13B and PROMETHEUS-VISION-13B on 45 customized score rubrics from LLaVA-Bench, VisIT-Bench, and PERCEPTION-BENCH. PROMETHEUS-VISION shows a high correlation with human evaluators on instances with real-world images.

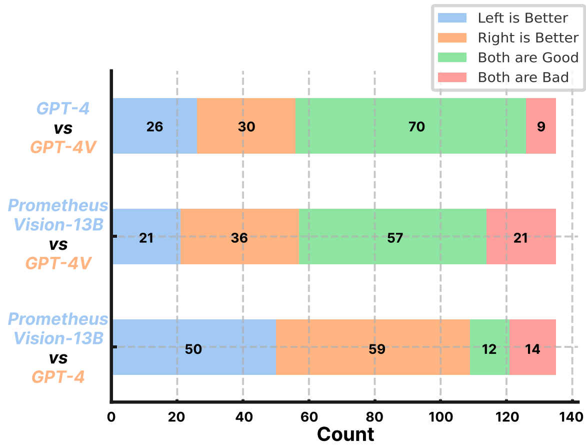


Figure 4: Pairwise comparison of the quality of the language feedback generated by GPT-4V, GPT-4, and PROMETHEUS-VISION-13B. Results show that PROMETHEUS-VISION's feedback is as good as or better than GPT-4V's feedback 57.78% of the time.

5.1.1 Correlation with Human Evaluators

The overall results are shown in Figure 3. PROMETHEUS-VISION 13B exhibits a high correlation with human evaluators on the LLaVA-Bench and the PERCEPTION-BENCH, achieving correlations of 0.639 and 0.870, respectively. These statistics are even comparable to GPT-4 and GPT-4V. However, on the VisIT-Bench, while outperforming GPT-3.5-Turbo and Prometheus 13B, PROMETHEUS-VISION shows a lower correlation when compared to GPT-4 and GPT-4V.

We posit that this disparity primarily originates from the differing characteristics of VisIT-Bench and other benchmarks. The former contains a higher proportion of **text-rich images**, such as graphs and charts, compared to the latter two datasets. Even though our training data, the PERCEPTION COLLECTION, also includes some text-rich images, the amount is relatively limited.

To explore whether this issue originates from the limitations of VLMs to read text-rich images, we conduct additional experiments of including OCR results during both training and inference in Appendix D.1. Yet, we found that it is not effective to mitigate this problem.

Nevertheless, recent works on vision-language models (Zhang et al., 2023; Ye et al., 2023b; Kim et al., 2022, 2023a) show promising capabilities for handling text-rich images, which could potentially serve as a backbone model for future iterations of PROMETHEUS-VISION. In consideration of these findings, the use of text-rich datasets, along with the integration of new methods drawn from recent architectural advancements, could alleviate these limitations. We leave this as future work.

Also, it is worthwhile to compare where GPT-4 and GPT-4V excel at each benchmark. Similar to PROMETHEUS-VISION, on VisIT-Bench, GPT-4 shows a slightly higher correlation with human evaluators compared to GPT-4V. This could mainly be because processing text is as important when assessing responses from text-rich images such as diagrams, charts, and graphs. On the other hand, GPT-4V shows a higher correlation with human evaluators on the LLaVA-Bench and PERCEPTION-BENCH which includes diverse real-world images.

5.1.2 Comparison of the Quality of the Feedback

Next, we compare the quality of the language feedback generated by GPT-4, GPT-4V, and PROMETHEUS-VISION 13B across 135 instances by hiring 9 human annotators. The experimental setting is detailed in Appendix I and the results are shown in Figure 4.

Surprisingly, PROMETHEUS-VISION 13B model is capable of generating feedback of a quality comparable to GPT-4. Among the 135 instances, human annotators determine that 57.78 % of the time, PROMETHEUS-VISION's feedback

Evaluator LM	LLaVA-Bench			VisIT-Bench			PERCEPTION-BENCH		
	Pearson	Kendall-Tau	Spearman	Pearson	Kendall-Tau	Spearman	Pearson	Kendall-Tau	Spearman
LLaVA-RLHF 7B	0.328	0.379	0.412	0.317	0.193	0.215	0.415	0.337	0.374
LLaVA-RLHF 13B	0.296	0.238	0.246	0.384	0.166	0.185	0.335	0.162	0.174
LLaVA-1.5 7B	0.278	0.226	0.254	0.408	0.188	0.214	0.602	0.383	0.419
LLaVA-1.5 13B	-0.005	0.097	0.105	0.597	0.347	0.376	0.505	0.254	0.270
ShareGPT4V 7B	0.366	0.222	0.247	0.360	0.222	0.256	0.474	0.338	0.378
Fuyu 8B	-0.023	0.049	0.052	0.059	0.079	0.087	0.011	-2.15E-04	4.29E-06
GPT-3.5-Turbo-0613	0.107	0.221	0.243	**0.685**	**0.539**	**0.592**	0.563	0.379	0.417
Prometheus 7B	0.233	0.192	0.210	0.482	0.363	0.419	0.723	0.491	0.534
Prometheus 13B	0.376	0.327	0.365	0.514	0.352	0.406	0.705	0.468	0.513
GPT-4-0613	0.712	0.500	0.530	0.494	0.352	0.394	0.808	0.626	0.661
PROMETHEUS-VISION 7B	0.411	0.214	0.233	0.662	0.424	0.478	0.700	0.471	0.502
PROMETHEUS-VISION 13B	**0.786**	**0.630**	**0.660**	0.574	0.378	0.425	**0.832**	**0.655**	**0.690**
GPT-4V-Preview	0.769	0.636	0.669	0.824	0.718	0.761	0.870	0.699	0.727

Table 4: Pearson, Kendall-Tau, Spearman correlation with scores sampled from GPT-4V across 3 inferences on visual instruction following benchmarks. Note that GPT-4V was sampled 6 times in total to measure self-consistency. The best comparable statistics are in **bold** and second best are underlined among baselines. We include GPT-4V as reference to show its self-consistency when inferenced multiple times.

is better or as good as GPT-4V's feedback. Also, human annotators determine that 45.93 % of the time, PROMETHEUS-VISION's feedback is better or as good as GPT-4's feedback. These results indicate that PROMETHEUS-VISION could also be utilized as an critique model for assisting assessment by humans (Saunders et al., 2022).

5.2 Can PROMETHEUS-VISION Closely Simulate GPT-4V as a Judge?

To check whether PROMETHEUS-VISION could be used as a reliable evaluator on various multi-modal tasks (visual instruction following, visual question answering, captioning), we compare the score correlation between GPT-4V and other VLM evaluator baselines, including PROMETHEUS-VISION.

The results in Table 4 show that PROMETHEUS-VISION demonstrates a higher correlation with GPT-4V compared to that of its backbone model, LLaVA-v1.5, on 3 visual instruction following benchmarks. This indicates that training with PERCEPTION COLLECTION enhances the VLM's evaluation capabilities. Furthermore, on the LLaVA-Bench and PERCEPTION-BENCH, PROMETHEUS-VISION 13B exhibits a higher correlation than the LM evaluators GPT-3.5-Turbo and GPT-4.

Due to space limitations, we include the results for visual question answering and captioning tasks, which show similar trends, in Appendix A.

6 Discussions and Analysis

Due to space limitations, we include our discussion and analysis section in the Appendix. This includes analysis of potential biases from VLM evaluators (Appendix B), the potential of PROMETHEUS-VISION as a reward model for visual instruction tuning (Appendix C), and ablation experiments (Appendix D). We find that using (1) an absolute grading scheme, (2) fine-grained score rubrics, and (3) language feedback distillation enables PROMETHEUS-VISION mitigate length biases and also operate in a relative grading settings.

7 Conclusion

In this paper, we expand the 'LM-as-a-Judge' paradigm to the multimodal space and introduce 'VLM-as-a-Judge'. We first propose a multimodal feedback dataset called the PERCEPTION COLLECTION, which contains fine-grained score criteria for each instance, unlike existing multimodal datasets that do not sufficiently consider critical factors in each instance. Using the PERCEPTION COLLECTION, we train PROMETHEUS-VISION, an open-source model specialized for evaluation purposes. The uniqueness of PROMETHEUS-VISION is that it can adhere to user-defined criteria during evaluation. Through experiments, we show that PROMETHEUS-VISION is an effective open-source alternative for GPT-4V evaluation. We hope our work advances the development of accessible and transparent evaluators for diverse modalities.

Limitations

One limitation of PROMETHEUS-VISION is that it does not show optimal performance when evaluating instances that include text-rich images including diagrams, charts, and graphs. This is heavily reliant on the performance of the visual encoder used during visual instruction tuning of the backbone model, LLaVA-v1.5 (Liu et al., 2023b,a). In the future, better VLM backbones could possibly resolve this issue. Moreover, another reason might come from the fact that the PERCEPTION COLLECTION is heavily skewed towards real-world images, not text-rich images. Adding more feedback data that includes text-rich images could be an interesting line of future work.

Also, our work does not consider cases when images generated by image generation models are given as input. Future work could consider exploring whether VLM evaluators could assess text outputs conditioned on AI-generated images.

Next, since our PERCEPTION COLLECTION is generated with OpenAI API, PROMETHEUS-VISION is only for non-commercial use and is subject to OpenAI's Terms of Use for the generated data. Also, since our base model, LLAVA-V1.5 uses LLAMA as their base model, PROMETHEUS-VISION is also subject to the LLaMA license. Future work could explore replicating our data generation pipeline and evaluator fine-tuning pipeline with fully open-source models.

Moreover, although our PERCEPTION COLLECTION includes score rubrics related to harmlessness, it is potentially risky to only rely on evaluator VLMs to assess based on such values. Our intention for assessing based on such values is that PROMETHEUS-VISION should be used as a tool for helping human evaluators instead of fully operating in an automatic fashion. Although we manually inspected and checked through 200 samples within our data to check whether there is any sign of harmful phrases within the generated data, it is difficult to look at all the data manually which is a limitation of using machine-generated data.

Lastly, as mentioned in our motivation for creating the PERCEPTION COLLECTION, currently there are not a lot of multimodal feedback datasets available for public use, compared to the text-only domain. Investigation of different forms of feedback, preference, and critique datasets would be an interesting line of future work.

Acknowledgements

This work was partly supported by KAIST-NAVER Hypercreative AI Center (60%) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00113, Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework, 20%; No.2021-0-02068, Artificial Intelligence Innovation Hub, 20%).

References

Aishwarya Agrawal, Ivana Kajic, Emanuele Bugliarello, Elnaz Davoodi, Anita Gergely, Phil Blunsom, and Aida Nematzadeh. 2023. Reassessing evaluation practices in visual question answering: A case study on out-of-distribution generalization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1171–1196.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and

C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*.

Geewook Kim, Hodong Lee, Daehee Kim, Haeji Jung, Sanghee Park, Yoonsik Kim, Sangdoo Yun, Taeho Kil, Bado Lee, and Seunghyun Park. 2023a. Visually-situated natural language understanding with contrastive reading model and frozen large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Seungone Kim, Se June Joo, Yul Jang, Hyungjoo Chae, and Jinyoung Yeo. 2023b. Cotever: Chain of thought prompting annotation toolkit for explanation verification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 195–208.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023c. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023d. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2023e. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. *arXiv preprint arXiv:2309.13633*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Lik Xun Yuan. 2023. distilbert-base-multilingual-cased-sentiments-student (revision 2e33845).

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2023. Improving automatic vqa evaluation using large language models. *arXiv preprint arXiv:2310.02567*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report.

OpenAI. 2023. GPT-4V(ision) system card. https://openai.com/research/gpt-4v-system-card.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

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

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023a. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023c. Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2023. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023b. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

Evaluator LM	OKVQA			VQA v2			TextVQA		
	Pearson	Kendall-Tau	Spearman	Pearson	Kendall-Tau	Spearman	Pearson	Kendall-Tau	Spearman
LLaVA-RLHF 7B	0.562	0.330	0.368	0.111	0.061	0.074	0.208	0.163	0.187
LLaVA-RLHF 13B	0.615	0.377	0.411	0.072	0.066	0.079	0.362	0.291	0.320
LLaVA-1.5 7B	0.605	0.405	0.464	0.200	0.134	0.152	0.290	0.201	0.247
LLaVA-1.5 13B	0.548	0.373	0.404	0.346	0.286	0.309	0.409	0.352	0.408
ShareGPT4V 7B	0.528	0.385	0.445	0.281	0.258	0.293	0.300	0.233	0.271
Fuyu 8B	0.143	0.147	0.162	0.193	0.163	0.179	0.176	0.174	0.193
GPT-3.5-Turbo-0613	0.371	0.307	0.374	0.370	0.345	0.391	0.436	0.350	0.424
Prometheus 7B	0.422	0.206	0.240	0.253	0.260	0.296	0.501	0.412	0.483
Prometheus 13B	0.482	0.284	0.325	0.178	0.122	0.145	0.417	0.343	0.400
GPT-4-0613	0.594	**0.509**	**0.584**	**0.605**	**0.527**	**0.606**	**0.723**	**0.642**	**0.718**
PROMETHEUS-VISION 7B	<u>0.608</u>	0.261	0.290	<u>0.455</u>	<u>0.395</u>	0.298	0.487	0.413	0.485
PROMETHEUS-VISION 13B	**0.653**	<u>0.401</u>	<u>0.441</u>	0.393	0.389	<u>0.428</u>	<u>0.512</u>	<u>0.445</u>	<u>0.523</u>
GPT-4V-Preview	0.795	0.735	0.810	0.681	0.610	0.684	0.791	0.705	0.796

Table 5: Pearson, Kendall-Tau, Spearman correlation with scores sampled from GPT-4V across 3 inferences on visual question answering benchmarks. Note that GPT-4V was sampled 6 times in total to measure self-consistency. We include GPT-4V as reference to show its self-consistency when inferenced multiple times. For all questions, we provided the Evaluator VLM with a fine-grained rubrics.

Evaluator LM	COCO-Captions	NoCaps
	Pearson	Pearson
LLaVA-RLHF 7B	0.148	0.210
LLaVA-RLHF 13B	0.198	0.171
LLaVA-1.5 7B	0.248	0.155
LLaVA-1.5 13B	0.157	0.111
ShareGPT4V 7B	0.184	0.185
Fuyu 8B	0.191	0.064
GPT-3.5-Turbo-0613	0.233	0.242
Prometheus 7B	0.335	0.165
Prometheus 13B	0.215	0.279
GPT-4-0613	<u>0.470</u>	**0.427**
PROMETHEUS-VISION 7B	0.434	0.327
PROMETHEUS-VISION 13B	**0.508**	<u>0.417</u>
GPT-4V-Preview	0.579	0.638

Table 6: Pearson, Kendall-Tau, Spearman correlation with scores sampled from GPT-4V across 3 inferences on captioning benchmarks. Note that GPT-4V was sampled 6 times in total to measure self-consistency. We include GPT-4V as reference to show its self-consistency when inferenced multiple times. For all questions, we provide the evaluator VLM with a fine-grained rubrics.

A Extension of Experimental Results: Correlation with GPT-4V

A.1 Visual Question Answering Benchmarks

Table 5 presents correlation results in visual question answering (VQA) benchmarks. In this benchmark, PROMETHEUS-VISION significantly outperforms other open-source models, including its backbone model, LLaVA-v1.5. Also, we observe that PROMETHEUS-VISION's correlation is generally lower in VQA benchmarks compared to visual instruction following benchmarks. We attribute this to our training data, PERCEPTION COLLECTION, which generally involves longer responses, while the answers in the VQA benchmark are mostly short. Future works could consider adding more diversity to the training data to obtain a stronger VLM evaluator.

A.2 Captioning Benchmarks

Unlike visual instruction following or VQA benchmarks, captioning benchmarks do not have a direct question but rather require writing a description of a given image in a short sentence. Therefore, we created prompts such as '`Generate a coco-style caption.`' and fed them to our evaluator VLM baselines during experiments. The results are shown in Table 6. While most evaluators, including proprietary LMs, show low correlation, PROMETHEUS-VISION 13B surprisingly stands out by showing a correlation above 0.5 in the COCO-Captions, indicating it could generalize to evaluate other visual-language tasks beyond its training data.

B Analysis of Potential Biases from VLM Evaluators

Previous works have highlighted a phenomenon known as *length bias* in models, which refers to a tendency of evaluator models to prefer longer responses (Li et al., 2023; Dubois et al., 2023; Zheng et al., 2023). This is a critical factor to consider during evaluation, as evaluators with length bias could give higher scores simply based on the length of the response, regardless of its actual content. To verify if this is the case, we plot and analyze the lengths of responses using our results from Section 5.1.

B.1 Is there a Length Bias?

The box plot in Figure 5 showcases GPT-4V and PROMETHEUS-VISION do not indiscriminately favor longer answers, indicating an absence of length bias. This is likely because our experimental setting is in an absolute grading setting where the evaluator VLM assesses the given responses with an absolute score rather than comparing two responses. This also aligns with the previous finding from Zheng et al. (2023) and Kim et al. (2023d). We provide more details of our analysis in Appendix E.3 and Appendix I.

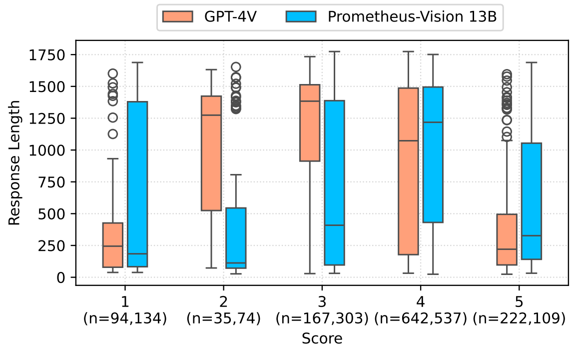
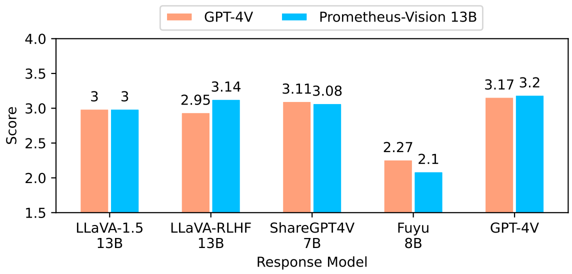


Figure 5: Distribution of length of responses by GPT-4V across different scores, as evaluated by GPT-4V and PROMETHEUS-VISION 13B, in all test sets. Each score category on the x-axis is annotated with the quantity of responses that received that particular score from each evaluator VLM. Individual test set results are in Figure 13.

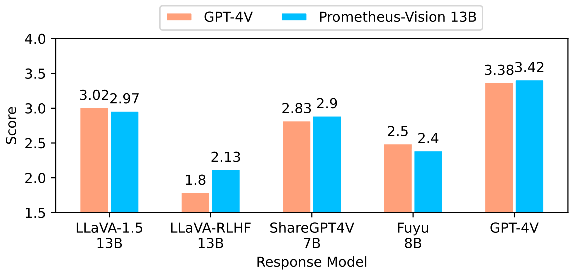
B.2 Is there a Self-Enhancement Bias?

Self-enhancement bias is another type of well-known bias where evaluators tend to prefer its own response (Zheng et al., 2023). Since PROMETHEUS-VISION is a model specialized for evaluation purposes only, it does not directly suffer from this bias. However, since we train PROMETHEUS-VISION with data augmented from GPT-4V and use LLaVA-v1.5 as our base model, this could indirectly influence the direction of evaluation by PROMETHEUS-VISION. To investigate whether there is a self-enhancement bias, we analyze the trends of which score was given to different response VLMs on the LLaVA-Bench and PERCEPTION BENCH.

Figure 6 illustrates the results. Overall, the results show that PROMETHEUS-VISION and GPT-4V exhibit similar evaluation patterns across the two benchmarks, reinforcing the findings from previous correlation studies with GPT-4V. Notably, PROMETHEUS-VISION gives a higher score to other models compared to its backbone model (LLaVA-v1.5) on the LLaVA-Bench, indicating that evaluator VLMs might not always prefer the



(a) LLaVA-Bench



(b) PERCEPTION-BENCH

Figure 6: Evaluation of 5 VLMs on (a) LLaVA-Bench and (b) PERCEPTION-BENCH using either PROMETHEUS-VISION or GPT-4V as an evaluator VLM. Trends show that PROMETHEUS-VISION could closely simulate GPT-4V evaluation. In addition, the open-source nature of PROMETHEUS-VISION provides accessible and transparent evaluation for those developing state-of-the-art VLMs.

responses from its backbone model.

While PROMETHEUS-VISION does give the highest score to GPT-4V, it is hard to determine if this is because PROMETHEUS-VISION was trained on data augmented from GPT-4V, or GPT-4V is distinctively better than the open-source VLMs. We leave analysis of this to future research.

C Can Prometheus-Vision Function as a Reward Modeling for Visual Instruction Tuning?

Ideally, an evaluator VLM that is trained to assess responses could also function as a reward model in the RLHF pipeline (Kim et al., 2023d) or as a data filter model in the DPO pipeline (Rafailov et al., 2023). In this section, we explore the former aspect by conducting experiments on 2 visual human preference datasets: LLaVA-Human-Preference (Sun et al., 2023) and RLHF-V (Yu et al., 2023). Note that both datasets use a ranking grading theme where two responses are presented and an evaluator VLM should distinguish which response would be preferred by the majority of people in terms of its helpfulness. During experiments, we inference

Evaluator LM	LLaVA-Human-Pref	RLHF-V
	Accuracy	Accuracy
Random	0.5	0.5
LLaVA-1.5 7B	0.48	0.45
LLaVA-1.5 13B	0.52	0.50
ShareGPT4V 7B	0.55	0.51
GPT-4-0613	<u>0.64</u>	0.6
PROMETHEUS-VISION 7B	0.58	0.55
PROMETHEUS-VISION 13B	<u>0.64</u>	<u>0.61</u>
GPT-4V-Preview	**0.73**	**0.70**

Table 7: Accuracy on 2 visual human preference datasets. The best comparable performance are **bolded** and second best <u>underlined</u>.

each response to acquire a score (inferencing 2 times) and then measure the accuracy by checking if the evaluator VLM assigned a higher score to the chosen response compared to the rejected response.

In Table 7, results show that PROMETHEUS-VISION, GPT-4, and GPT-4V are the only baselines that show an accuracy above random performance. Moreover, PROMETHEUS-VISION-13B shows +0.12% and +11% improvement over its base model (LLAVA-1.5 (13B)) on each dataset, respectively. This is unexpected since PROMETHEUS-VISION was never trained to assess the helpfulness between two responses since it is trained in an absolute grading scheme. As in the findings from Kim et al. (2023d), these results show the possiblity of using a generative VLM (PROMETHEUS-VISION) directly as a reward model for RLHF and also as a data filter model for DPO. We leave the exploration of this research to future work.

D Ablation Experiments

In this section, we conduct a series of ablation experiments to identify factors influencing the evaluation quality of PROMETHEUS-VISION. The experiments show the correlation with the evaluation results of GPT-4V, and are carried out on visual instruction following benchmarks. Table 8 shows the results of ablation experiments.

D.1 OCR ablation

Recent VLMs still face challenges in processing text-rich images. Since the training data for PROMETHEUS-VISION did not include a significant number of text-rich images and related questions, its ability to handle such images is comparatively lower. We investigate the impact of augmenting PROMETHEUS-VISION with text from external OCR tools, to see how it affects the handling of

Evaluator LM	LLaVA-Bench			VisIT-Bench			PERCEPTION-BENCH		
	Pearson	Kendall-Tau	Spearman	Pearson	Kendall-Tau	Spearman	Pearson	Kendall-Tau	Spearman
PROMETHEUS-VISION 7B	**0.411**	**0.214**	**0.233**	0.662	**0.424**	0.478	**0.699**	**0.471**	**0.502**
OCR Ablation									
w/ OCR (inference)	0.399	0.202	0.214	**0.686**	0.419	**0.488**	**0.699**	0.461	0.481
w/ OCR (train+inference)	0.379	0.196	0.206	0.653	0.401	0.455	0.678	0.455	0.478
Training Ablation									
w/o score rubric	0.265	0.066	0.202	0.530	0.389	0.347	0.617	0.403	0.277
w/o feedback	0.123	0.035	0.198	0.462	0.196	0.332	0.543	0.224	0.171
w/o reference answer	0.085	0.001	0.173	0.433	0.029	0.279	0.245	0.058	0.041
Model Ablation									
ShareGPT4V 7B	0.233	0.197	0.126	0.564	0.230	0.294	0.244	0.116	0.369
LLaVA-RLHF 7B	0.263	0.193	0.225	0.584	0.384	0.329	0.267	0.104	0.425
Fuyu 8B	0.143	0.067	0.181	0.211	0.133	0.106	0.073	0.104	0.219

Table 8: Ablation experiments. Pearson, Kendall-Tau, Spearman correlation with scores sampled from GPT-4V across 3 inferences on visual instruction following benchmarks. Note that GPT-4V was sampled 6 times in total to measure self-consistency. The best comparable statistics are **bolded** and second best underlined among baselines.

images containing text. We observe that incorporating OCR results only during the inference process does not significantly increase the correlation with GPT-4V compared to when OCR is not used. Additionally, introducing OCR results as input during the training phase actually leads to a further decrease in correlation.

D.2 Training ablation

In this section, we examine the impact of each component of the dataset used during training on the evaluation performance. Initially, the removal of the score rubric, which serves as the basis for evaluation, resulted in an average decrease of -0.12 in the Pearson correlation score across all benchmarks. Furthermore, eliminating the feedback, which explains the rationale behind the model's evaluation, led to a decrease of -0.21. Lastly, when the reference answer was removed, there was a decrease of -0.33. These results demonstrate that the absence of a properly established score rubric and reference answer during training can diminish the quality of evaluation. Moreover, training the model to provide feedback explaining its evaluations, rather than merely generating scores, significantly influences the quality of the evaluation.

D.3 Model ablation

Next, we conduct experiments to investigate how the evaluation performance varies depending on the backbone VLM of PROMETHEUS-VISION. In addition to the default model, LLaVA-v1.5, three additional VLMs are trained in the PERCEPTION-

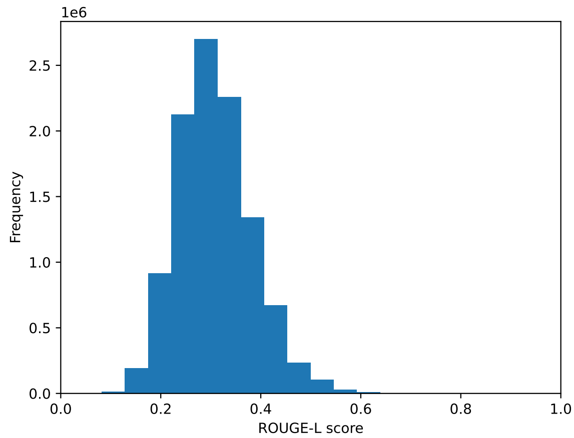


Figure 7: Distribution of ROUGE-L similarities in pairs of score rubric descriptions within PERCEPTION-COLLECTION.

COLLECTION: ShareGPT4V-7B, LLaVA-RLHF 7B, and Fuyu 8B. The results show that all three models exhibit lower evaluation correlation with GPT-4V compared to the default model, LLaVA-v1.5.

E Analysis of PERCEPTION COLLECTION

E.1 Diversity of Score Rubrics

When hand-crafting seed rubrics and generating new fine-grained score rubrics through brainstorming, for each rubric, we tag keywords that best describe the criteria. Figure 11 and Figure 12 show word clouds of keywords in general-purpose rubrics and domain-specific rubrics in-

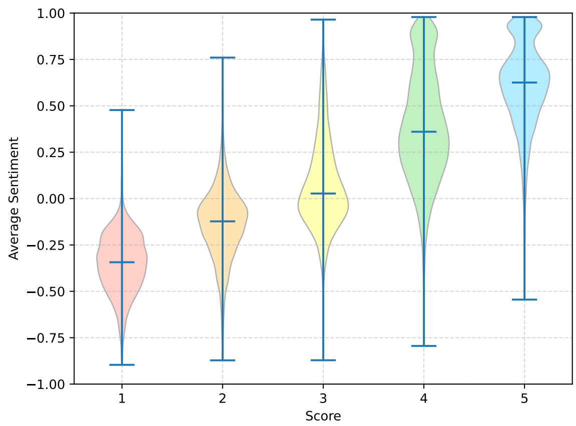


Figure 8: Average sentiment of descriptions for each score in PERCEPTION-COLLECTION Rubrics. Sentiment of +1 signifies positivity, 0 neutrality, and -1 negativity.

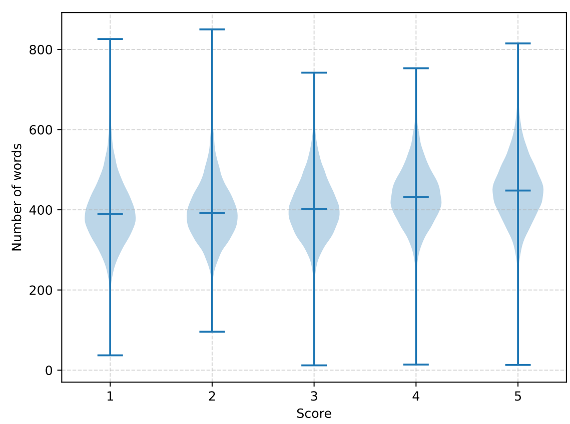


Figure 9: Distribution of length of responses scoring from 1 to 5 provided for training.

cluded in PERCEPTION-COLLECTION, respectively. General-purpose rubrics encourage a broader, more holistic perspective into the image as noted by the prominence of the words 'environmental', 'scene', 'social', *etc.*. Domain-specific rubrics bring more attention to the visual aspects of the image and data, specifying long-tail subfields of various subjects which are shown by the words 'scientific', 'artistic','anatomical', *etc.*.

Following previous works on machine-generated instructions (Wang et al., 2023; Honovich et al., 2023; Kim et al., 2023d), we quantify the overlap of the generated score rubrics in our training data. Specifically, we compute ROUGE-L similarities between score rubric descriptions for every possible pair within PERCEPTION-COLLECTION. The ROUGE-L distribution is plotted in Figure 7, with the average ROUGE-L score being 0.31 and the distribution being left-skewed. This low similarity score underscores the unique and varied nature of the PERCEPTION COLLECTION.

E.2 Decisiveness of Score Descriptions

We examine whether each level of the scoring system in the rubric is clear and distinct. Following Kim et al. (2023d), we compute the average sentiment in the description of each score in rubrics within PERCEPTION-COLLECTION. We use a publicly available DeBERTa-distilled DistilBERT for sentiment analysis tasks (Lik Xun Yuan, 2023). The results can be found in Figure 8, where descriptions corresponding to a score of 1 are generally more negative, while those with a score of 5 are more positive. This suggests that the training data is appropriately interpolated according to scores and PROMETHEUS-VISION trained on this dataset can conduct absolute scoring clearly and effectively.

E.3 Length Bias of Responses per Score Provided for Training

As explained in Section 3.1, given an instruction, rubric, and reference answer, a response corresponding to score i is generated for all $1 \leq i \leq 5$ to provide an evaluator model under training responses to practice assessment on. To nullify the tendency of recent LMs to give higher scores to longer responses (Li et al., 2023; Dubois et al., 2023; Zheng et al., 2023), during PERCEPTION-COLLECTION construction, we aim to maintain similar length of responses across the score range (See Appendix J.1 for the exact prompt). The distribution of length of responses by score is plotted in Figure 9. Response lengths are distributed evenly across the score range, with an 417 words in average.

F Analysis of PERCEPTION-BENCH

F.1 Validity of Unseen Score Rubrics

To ensure that PERCEPTION-BENCH contains rubrics *unseen* in PERCEPTION-COLLECTION, we plot the ROUGE-L distribution between score rubric descriptions in PERCEPTION-BENCH and PERCEPTION-COLLECTION in Figure 10. The average ROUGE-L similarity between descriptions in our test set and train set is 0.29 and the distribution is left-skewed. We claim that the train-test overlap in our proposed dataset is low and that PERCEPTION-BENCH contains many novel score

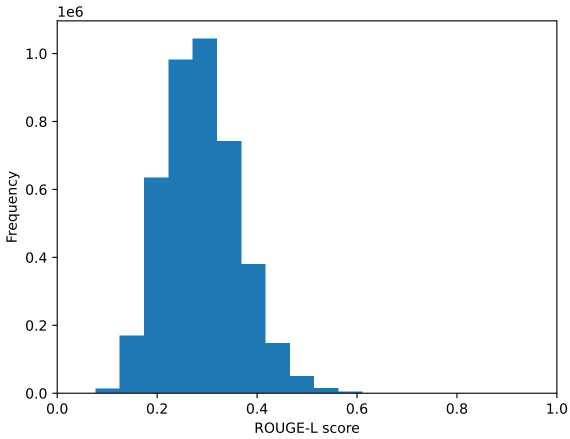


Figure 10: Distribution of ROUGE-L scores between score rubric descriptions in PERCEPTION-BENCH and PERCEPTION-COLLECTION.

Evaluator LM	LLaVA-VisIT-PERCEPTION
	Pearson
Rouge-1	0.314
Rouge-L	0.308
SPICE	0.340
METEOR	0.489
GPT-3.5-Turbo	0.493
Prometheus 13B	0.450
GPT-4	**0.734**
PROMETHEUS-VISION 13B	0.674
GPT-4V-Preview	0.771

Table 9: Pearson correlation with scores from human on 45 samples from 3 visual instruction following benchmarks. The best comparable statistics are **bolded** and second best underlined among baselines.

rubrics.

G Comparison with conventional metrics

Traditional VLM response evaluation metrics, which measure similarity solely between the reference answer and the response without considering the image, struggle to account for the varied information in images. Consequently, these conventional metrics can diverge significantly from human evaluations. As shown in Table 9, there is a low Pearson correlation between human-predicted scores and conventional metrics. Notably, even METEOR, the conventional metric with the highest correlation, only achieves around 0.489, whereas PROMETHEUS-VISION 13B demonstrates a higher correlation of 0.674. Moreover, conventional metrics often lack explainability. As Figure 1 indicates, they typically represent response quality with a simple value between 0 and 1. Model response, although it adequately depicts the image without employing expressions used in the reference answer, still receives a low score from conventional metrics due to their inability to perceive the image. In contrast, PROMETHEUS-VISION not only provides a proper numeric score but also generates feedback that elucidates the reasons behind the score. This dual output can be instrumental in identifying ways to improve the model.

H Experimental Details

H.1 Implementation Details and Computation

Training We employ LLaVA-1.5 7B / 13B as the backbone VLM for PROMETHEUS-VISION. For the language model component, we utilize vicuna-13b-v1.5, and for the vision encoder, we use clip-vit-large-patch-14-336px. We freeze both the language model and the vision encoder, focusing our training solely on an MLP based alignment network. The training is conducted for one epoch, with a batch size per device set at 32. We set the learning rate at 1e-3, with no weight decay and a warmup ratio of 0.03. A cosine scheduler is utilized as the learning rate scheduler. To enhance training efficiency, we incorporate gradient checkpointing and deepspeed zero 2 in our training process.

Inference We use three response VLMs to generate responses to given images and questions in each dataset. Then, an evaluator VLM generates feedback and scores indicating how the response might improve given these responses, along with the image, question, reference answer, and a guiding rubric. This approach allows us to measure the correlation between scores from GPT-4V and those from other models. In the process of generating feedback, the model employs sampling with a temperature set to 1.0 and top-p set at 0.9, while the maximum number of tokens is configured to 2048. Regarding the resources utilized for training and inference, the GPU setup includes 8 NVIDIA A100 80GB. For the CPU, an AMD EPYC 7543 32-Core Processor is used.

H.2 Details in Human Evaluation

We recruit 9 undergraduate students proficient in English to conduct a human evaluation. The dataset used for the human evaluation is exclusively drawn from the visual instruction tuning benchmarks. Additionally, we randomly sample 15 items each from LLaVA-Bench, VisIT-Bench, and Perception-Bench, creating a total of 45 problems. For the pair-

Figure 11: Word cloud of keywords in general-purpose score rubrics within PERCEPTION-COLLECTION

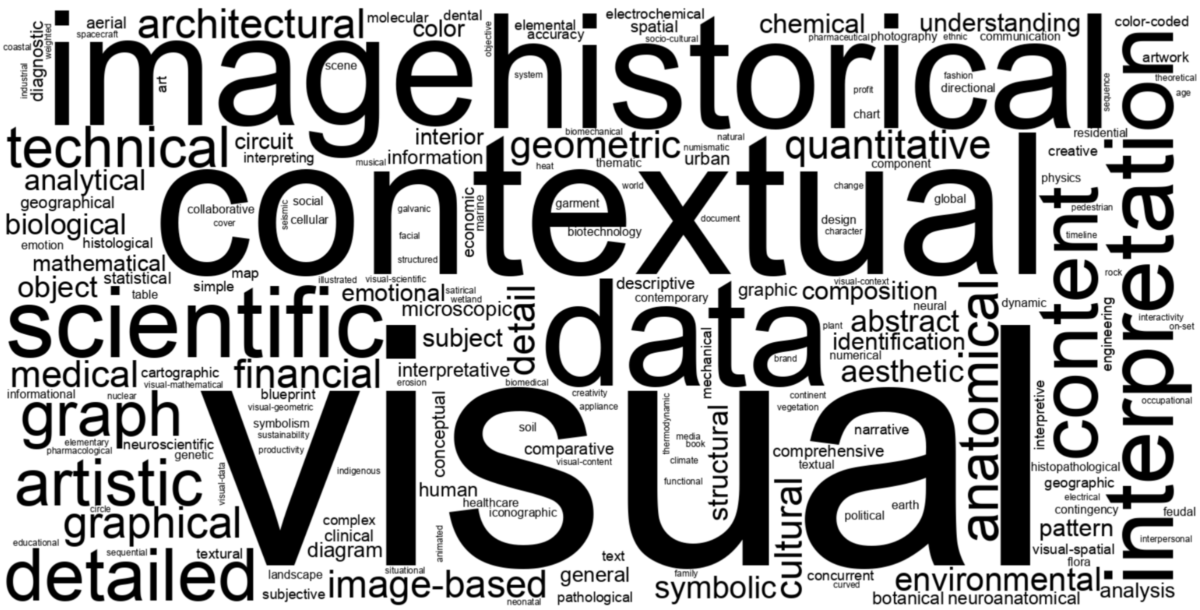
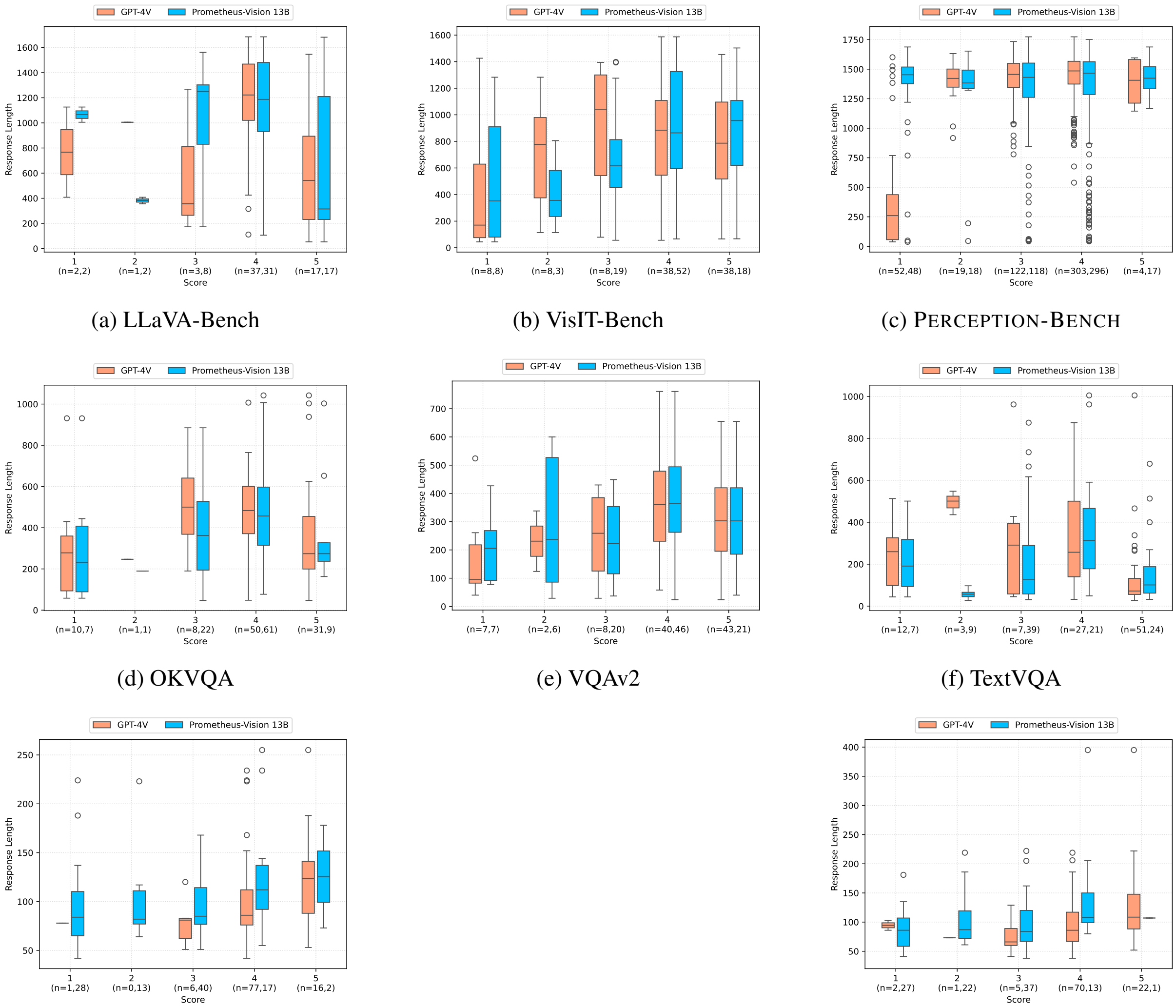


Figure 12: Word cloud of keywords in domain-specific score rubrics within PERCEPTION-COLLECTION



(a) LLaVA-Bench (b) VisIT-Bench (c) PERCEPTION-BENCH

(d) OKVQA (e) VQAv2 (f) TextVQA

(g) COCO-Captions (h) NoCaps

Figure 13: Full distribution of length of responses by GPT-4V across different scores, as evaluated by GPT-4V and PROMETHEUS-VISION 13B, in each test set. Each scoring category on the x-axis is annotated with the number of responses that received that particular score from each Evaluator VLM.

wise feedback quality comparison, we utilize feedback from GPT-4V, GPT-4, and PROMETHEUS-VISION 13B. Each of the 45 problems is structured to compare two out of the three feedbacks. Consequently, 3 sets of the same 45 problems are prepared, and the 9 participants are divided into 3 groups, with each group evaluating the same set of problems. We use Label Studio as the evaluation platform[1]. The annotation interface is shown in Figure 18.

I Length Bias during Evaluation

We report GPT-4V response length distribution scored by GPT-4V and PROMETHEUS-VISION 13B on individual test sets in Figure 13. Overall trends show that both evaluator VLMs do not display bias towards lengths in responses during inference.

[1] https://labelstud.io

J List of Prompts

J.1 Prompts for PERCEPTION COLLECTION Creation

We include the prompts used in the creation of our training dataset, PERCEPTION COLLECTION. The Example Criteria include hand-crafted seed rubrics that were sampled and inserted beforehand. Additionally, for fine-grained rubric augmentation, the same prompt is used, but general-purpose rubrics and domain-specific rubrics are augmented separately, ensuring the seed rubrics are also individually incorporated without mixing. Notably, although the prompt does not feature an image insertion, in practice, images are included when calling the GPT-4V API. Detailed information is in the OpenAI API document[2].

[2] https://platform.openai.com/docs/guides/vision

Prompt for rubric augmentation

You are helpful and creative rubric generator. You should brainstorm creative and impressive three rubrics used to evaluate the ability of a vision-language model to generate text when given an image.

The rubric must be structured to assess areas that can be answered by viewing the image. It consists of a description explaining specific tasks and criteria for scoring. Here you will see 4 examples of 'criteria', and their scoring rubrics, formatted as JSON.

Criteria 1:
{Example Criteria 1}

Criteria 2:
{Example Criteria 2}

Criteria 3:
{Example Criteria 3}

Criteria 4:
{Example Criteria 4}

Please brainstorm new three criterias and scoring rubrics.
Be creative and create new but useful criteria that people in different settings or industries might find practical.
Please format the output as same as the above examples with no extra or surrounding text. And you should not mention the term like 'Criteria X:' and "'json"'.
In JSON, all keys and string values must be enclosed in double quotes (""). For example, "key": "value" is a valid format, but key: "value" or 'key': 'value' are not valid.
You should create a diverse rubrics suitable for the given image

Generated criteria:

Prompt for checking alignment

You are helpful and creative rubric evaluator. You will be given one image and a rubric used to evaluate the capabilities of a vision-language model based on that image. If the rubric is well-aligned with the given image, you should answer 'align'. However, if the rubric does not fit the given image and there are areas for improvement, you should answer 'misalign'.

The rubric must be structured to assess areas that can be answered by viewing the image. It consists of a description explaining specific tasks and criteria for scoring. Here you will see the rubric, and their scoring rubrics, formatted as JSON.

Rubric:
{Rubric}

Please answer 'align' or 'misalign'. You should generate the output in lowercase.

Alignment:

Prompt for refining rubric

You are helpful and creative rubric creator. You will be given one image and a rubric used to evaluate the capabilities of a vision-language model based on that image. If the rubric does not fit the given image and there are areas for improvement, you should make improvements to create a better rubric.

The rubric must be structured to assess areas that can be answered by viewing the image. It consists of a description explaining specific tasks and criteria for scoring. Here you will see the rubric, and their scoring rubrics, formatted as JSON.

Rubric:
{Rubric}

If there are areas that need improvement in the given rubric, improve the rubric that better fits the given image. Maximize your creativity to ensure that the rubric you refine is not too similar to the already existing one.

Please format the output as same as the above examples with no extra or surrounding text. You should generate only one rubric. And you should not mention the term like 'Criteria X:' and "'json"'. In JSON, all keys and string values must be enclosed in double quotes (""). For example, "key": "value" is a valid format, but key: "value" or 'key': 'value' are not valid.

Generated rubric:

Prompt for generating instruction (1)

Your job is to generate a new novel problem and a response that is related to the given score rubric and image.

The score rubric:
{Rubric}

* Problem
- The problem should inherently be related to the score criteria, score rubric and image given above. Specifically, the score criteria should be the core attributes required to solve the problem.
- The problem itself should not be too generic or easy to solve.
- Try to make the person who might solve the problem not notice the existence of the score rubric by not explicitly mentioning it, and also provide additional inputs and options if needed.
- Assume a situation where a user is interacting with an AI model. The user would try to ask in a first-person point of view, but not using terms like "I", "A User" or "You" in the first sentence.
- Do not give a role to the AI, assume that the user is asking a question from his point of view.
- Do not include any phrase related to AI model in the problem.
- The problem should only be answered by looking at an image, not just by reading the problem.

Prompt for generating instruction (2)

* Response
- The response should be a response that would get a score of 5 from the score rubric.
- The response should be as detailed as possible unless the score rubric is related to conciseness or brevity. It should consist of multiple paragraphs, a list of items, or a step-by-step reasoning process.
- The response should look like how a well-prompted GPT-4 would normally answer your problem.

* Format
- DO NOT WRITE ANY GREETING MESSAGES, just write the problem and response only.
- In front of the problem, append the phrase 'Problem:' and in front of the response, append the phrase 'Response:'.
- Write in the order of 'Problem' - 'Response', where the two items are separated by the phrase '[NEXT]'.
- Write [END] after you are done.

Data Generation:

Prompt for response and feedback (1)

Your job is to generate a response that would get a score of {score} and corresponding feedback based on the given score rubric and image. For reference, a reference response that would get a score of 5 is also given.

Instruction:
{instruction}

The score rubric:
{rubric}

Reference response (Score 5):
{response}

* Response
- The quality of the score {score} response should be determined based on the score rubric and image, not by its length.
- The score {score} response should have the same length as the reference response, composed of {number of sentences} sentences.
- Do not explicitly state the keywords of the score rubric inside the response.

Prompt for response and feedback (2)

* Feedback
- The score {score} feedback should each be an explanation of why the response would get a score of {score}. It should be written based on the generated response, score rubric and image.
- The score {score} feedback shouldn't just copy and paste the score rubric, but it should also give very detailed feedback on the content of the corresponding response.
- The score {score} feedback should include the phrase 'So the overall score is {score}' in the last sentence.

* Format
- DO NOT WRITE ANY GREETING MESSAGES, just write the problem and response only.
- In front of the response, append the phrase 'Response:' and in front of the feedback, append the phrase 'Feedback:'.
- Write in the order of 'Response' - 'Feedback', where the two items are separated by the phrase '[NEXT]'.
- Write [END] after you are done.

Data Generation:

J.2 Prompts for PROMETHEUS-VISION

Prompt for evaluation

###Task Description:
An instruction (might include an Input inside it), a response to evaluate, a reference answer that gets a score of 5, image and a score rubric representing an evaluation criterion is given.
1. Write a detailed feedback that assesses the quality of the response strictly based on the given score rubric, not evaluating in general.
2. After writing a feedback, write a score that is an integer between 1 and 5. You should refer to the score rubric.
3. The output format should look as follows: Feedback: (write a feedback for criteria) [RESULT] (an integer number between 1 and 5)
4. Please do not generate any other opening, closing, and explanations.

###The instruction to evaluate:
{Instruction}

###Response to evaluate:
{Response}

###Reference Answer (Score 5):
{Reference answer}

###Score Rubrics:
{Description}
Score 1: {Criteria of score 1}
Score 2: {Criteria of score 2}
Score 3: {Criteria of score 3}
Score 4: {Criteria of score 4}
Score 5: {Criteria of score 5}

###Feedback:

K Qualitative Examples

In, Figure 14 and Figure 15, we show instances where PROMETHEUS-VISION 13B outperforms GPT-4 and GPT-4V. In Figure 16 and Figure 17, we compare the effectiveness of the evaluation method used in LLaVA-Bench and PERCEPTION-BENCH. Specifically, we compare how responses that are long and plausible but with little substance are evaluated in the benchmarks. We show that such a response can fool the LLM evaluator (GPT-4) in LLaVA-Bench but is scrutinized by PROMETHEUS-VISION in PERCEPTION-BENCH.

Score Rubric

This criterion evaluates the model's ability to identify various objects within the image and understand their functionality or use-case in a real-world context. It requires the model not only to name the items but also to describe their purpose or common use.

Score 1: The model fails to correctly identify any of the objects in the image and does not recognize their functionality or use-case.
Score 2: The model identifies very few objects in the image and makes significant errors in explaining their functionality or use-case.
Score 3: The model recognizes some objects correctly but provides limited or only partially accurate information regarding their functionality or use-case.
Score 4: The model correctly identifies most objects and provides mostly accurate information about their functionality or use-case, with minor mistakes or omissions.
Score 5: The model demonstrates an excellent ability to recognize all objects in the image and accurately describes their functionality or use-case in detail.

Instruction

Explaining the objects on the desk, how might one use them in daily life and what purpose they serve? Describe each item in the context of its typical functionality.

Response from GPT-4V

In the image, there are several items typically found in an office or home environment that serve various purposes in daily life:

1. Envelope - This appears to be a standard mailing envelope, which is used to send letters or documents through the postal service. In daily life, one would enclose correspondence or small items inside, seal it, write the recipient's address on the front, and post it.

2. Printed Papers - There are various printed papers, possibly instructions, informational content, bills, or receipts. They are typically used to communicate information, instructions for a product, or record transactions. One would read these to gain important information regarding whatever they accompany or pertain to.

3. Cordless Home Telephone - The Siemens branded phone is a cordless landline telephone used to make and receive calls. Such telephones communicate with a base station that is connected to the telephone network and can be used anywhere within range of the base station. It provides a convenient way to have voice communications at home or within an office without being tethered to a wire.

4. Power Adapter - This is used to provide power to electronic devices, likely for charging the battery of the cordless phone. It converts mains AC to low-voltage regulated DC power for the operation of an electronic device. The adapter would be plugged into an electrical outlet, and the other end would connect to the device that requires power or charging.

These items collectively represent how one may communicate, handle paperwork or administrative tasks, and maintain electronic (...)

Image Caption from LLaVA-1.5 ***(input for LLM evaluator only)***

A Samsung cell phone sits on a white sheet, surrounded by a few papers and a charger.

Feedback from GPT-4

Although the response identifies several objects and provides detailed explanations of their functionality, it refers to objects that are not present in the image. The image description indicates a phone and books on a bed, not an envelope, printed papers, a cordless home telephone, or a power adapter. The response does not clearly conform to the instructions or the provided image description, meaning it does not accurately identify or explain the functionality and use-case for the objects in the image. It is significant that the response interprets the image quite differently from the given image description.
[RESULT] 1

Feedback from Prometheus-Vision 13B

The response correctly identifies most of the objects in the image and provides a mostly accurate description of their functionality or use-case. However, there are minor mistakes or omissions, such as not explicitly mentioning the phone as a mobile phone and not detailing the specifics of the printed papers, which are likely manuals or instructions for the phone. The response also does not mention the box with the label 'PAKKENS INHOLD' and its significance in the context of the image. These minor shortcomings prevent the response from achieving a perfect score. So the overall score is 4.

Figure 14: An example of comparison between GPT-4 feedback and PROMETHEUS-VISION feedback on a PERCEPTION-BENCH instance. In here, GPT-4 shows its limitation of as a LM judge for this task. As GPT-4 is not trained to process images, a description of the image produced by LLaVA-1.5 is provided to GPT-4 as a proxy for the actual image. GPT-4 cannot detect objects existing in the image that are *not* mentioned in the image caption but are correctly mentioned in the response being evaluated. The incorrect parts of the feedback are in red, and for comparison, the correct parts are in blue. Note that the reference answer is included in the input for both evaluators but is omitted in this figure for brevity.

Score Rubric

Does the model's response correctly identify the activity that now enables differently-abled people to participate in?

Score 1: The response does not correctly identify any activity that the modified store item enables for differently-abled people.
Score 2: The response hints at the correct activity but is either incorrect or too vague to be certain.
Score 3: The response identifies a correct activity but may not capture the full scope of how the modification aids in participation in that activity.
Score 4: The response largely identifies the correct activity and provides a good understanding of how the modification aids participation, with minor details lacking.
Score 5: The response fully identifies the correct activity that the modified store item enables and provides a thorough explanation of how it assists differently-abled people in participating.

Instruction

How has this store item been modified to make it enabling for people with mobility issues? What activity does it allow differently-abled people to participate in?

Response from GPT-4V

The store item in the image is a motorized shopping cart (also known as a mobility scooter), which has been specifically designed to assist people with mobility issues. It enables individuals who have difficulty walking or standing for extended periods to participate in the activity of shopping within a store.

The modifications that make it enabling include:

1. A comfortable, seated position for the user, which reduces the need for standing or walking.
2. Motorized wheels and an easy-to-operate steering handle, which allow the user to move around the store effortlessly.
3. A shopping basket is attached to the scooter, enabling the user to collect and carry items as they shop without the need for a separate cart.
4. A simple interface for acceleration and braking, which usually involves just a button or a lever, making it user-friendly for individuals with limited hand strength or dexterity.

These features allow differently-abled people to have a degree of independence while shopping, making the experience more accessible and inclusive.

Reference Answer

The store item in the image has been modified by integrating it with a motorized scooter, which includes a seat and controls accessible from a seated position. This modification allows people with mobility issues who find it difficult or impossible to walk through a store to participate in the activity of shopping independently. The scooter includes a basket in front, which ensures that the user can carry items just like they would with a traditional shopping cart.

Feedback from GPT-4V

The response effectively recognizes and describes the modifications made to the store item. It mentions the seated position for comfort, motorized wheels, a steering handle for effortless movement, and a shopping basket integrated into the scooter. Additionally, the response discusses the simple interface for acceleration and braking designed for user-friendliness. These aspects are clearly linked to improving accessibility for individuals with mobility issues, consistent with the highest level of the score rubric. However, although the response is comprehensive, there could be a mention of the height of the device being suitable for aisle navigation, allowing users to view and reach products on shelves if such a feature exists. It could also discuss the potential adjustability of the seat or steering column if visible in the image. The response could only score less than five if there are such visible features that are not discussed, but without that evidence, the response identifies most modifications and explains their relevance to enhancing accessibility, fitting well into score 4 of the rubric. [RESULT] 4

Feedback from Prometheus-Vision 13B

The response accurately identifies all visible modifications to the store item, such as the seated position, motorized wheels, and the attached shopping basket, and expertly explains their relevance to enhancing accessibility for people with mobility issues. It also correctly notes the ease of operation and the user-friendly interface, which are essential for individuals with limited mobility. The response is comprehensive and aligns with the score rubric's highest criteria by providing a detailed explanation of how the modifications enable differently-abled people to participate in the activity of shopping. So the overall score is 5.

Figure 15: An example of comparison between GPT-4V feedback and PROMETHEUS-VISION feedback on a VisIT-Bench instance. PROMETHEUS-VISION follows the score rubric and reference answer well while GPT-4V gives an overly critical feedback (orange).

Instruction

If someone were to recommend a new flavor or topping to the dish, describe the reason for this change and how it might alter the overall taste.

Response from LLaVA-RLHF-v1.5 13B *(Assistant 2)*

Introducing a new flavor or topping to the dish can enhance the overall taste and provide a unique experience for those who enjoy the dish. Here are some reasons for considering a new flavor or topping and how it might alter the overall taste:

1. Enhance the taste: Introducing a new flavor or topping can add complexity to the dish, making it more interesting and enjoyable. This can be achieved by incorporating a variety of ingredients, such as herbs, spices, or sauces, which can complement the existing flavors in the dish.

2. Add variety: Introducing new flavors or toppings can help to keep the dish fresh and exciting, preventing it from becoming monotonous or boring. This can be particularly useful for those who enjoy experimenting with different tastes and textures, as it allows them to explore new combinations and flavors.

3. Create a new dish: Introducing a new flavor or topping can transform the dish into a completely new creation, with a unique taste and presentation. This can be an excellent way to showcase creativity and experimentation, as well as to cater to different preferences and dietary requirements.

4. Enhance the visual appeal: Introducing new flavors or toppings can also improve the visual appeal of the dish, making it more colorful and visually appealing. This can be achieved by incorporating a variety of colors and textures, which can make the dish more attractive and enticing.

5. Improve the nutritional value: Introducing new flavors or toppings can also improve the nutritional value of the dish, by incorporating additional vitamins, minerals, or other nutrients. This can be particularly useful for those who are conscious about their diet and health, as it allows them to enjoy a delicious and satisfying meal while also benefiting from a variety of nutrients.

In conclusion, introducing a new flavor or topping to the dish can enhance the overall taste, add variety, create a new dish, improve the visual appeal, and increase the nutritional value. This can make the dish more enjoyable and satisfying for those who enjoy it, while also providing a unique and exciting experience for those who are open to experimenting with new flavors and combinations.

Reference Answer *(Assistant 1)*

If someone were to recommend a new flavor or topping to the chashu ramen dish, the reason for this change could be to enhance or complement the existing flavors. For example, adding a soft-boiled egg would enrich the dish with its silky yolk, providing an added creaminess that would complement the spicy chili sauce and the tender chashu. This alteration would make the overall taste more complex, creating a satisfying balance of flavors and textures in the meal.

Image Caption Provided in LLaVA-Bench

A close-up photo of a meal at ICHIRAN. The chashu ramen bowl with a spoon is placed in the center. The ramen is seasoned with chili sauce, chopped scallions, and served with two pieces of chashu. Chopsticks are placed to the right of the bowl, still in their paper wrap, not yet opened. The ramen is also served with nori on the left. On top, from left to right, the following sides are served: a bowl of orange spice (possibly garlic sauce), a plate of smoke-flavored stewed pork with chopped scallions, and a cup of matcha green tea.

LLaVA-Bench Style Evaluation by GPT-4

8 9
Both Assistant 1 and Assistant 2 provided helpful and relevant answers to the user's question. Assistant 1 focused on a specific example of adding a soft-boiled egg to the dish and explained how it would enhance the dish's flavor and texture. This response was accurate and detailed but limited to one suggestion. Assistant 2, on the other hand, provided a more comprehensive answer, discussing various reasons for introducing a new flavor or topping and how it might alter the overall taste. The response was well-structured, detailed, and covered multiple aspects of the question. Therefore, Assistant 2 receives a slightly higher score due to the broader coverage of the topic.

Figure 16: An example of LLaVA-Bench style GPT-4 evaluation of LLaVA-RLHF-v1.5 13B response on a LLaVA-Bench instance. In the original LLaVA-Bench-style evaluation, GPT-4 is the evaluator and compares the model response against the reference answer, where it rates the reference answer (Assistant 1) first and then rates the model response (Assistant 2) second. In this example, the LLaVA-RLHF-v1.5 13B response scored higher than the reference answer, given credit for its comprehensiveness. However, the response is too generic to the extent that it could have been generated *without* the image in context.

Instruction

Analyze the cartoon and elaborate on the dualistic representation of the character depicted, focusing on how the image comments on the contrast between his economic policies and philanthropic ventures, and what this suggests about the societal and historical implications of his actions during the period portrayed.

—"Forty-Millionaire Carnegie in his Great Double Role," *The Saturday*

Response from LLaVA-RLHF-v1.5 13B

The cartoon depicts a caricature of Mark Twain, the famous author and humorist, standing on a platform with a suitcase. The image presents a dualistic representation of the character, as it contrasts his economic policies and philanthropic ventures.

On one hand, Mark Twain is known for his economic policies that favored the wealthy and supported the growth of industry, which can be seen in the image as he is holding a suitcase, symbolizing wealth and prosperity. This aspect of the caricature highlights the economic success and the role of wealth in society during the period portrayed.

On the other hand, Mark Twain is also known for his philanthropic ventures, such as his support for the poor and his efforts to improve education and healthcare for the less fortunate. The image shows him standing on a platform, which could symbolize his desire to raise awareness and advocate for social justice and equality. This aspect of the caricature emphasizes the importance of giving back to the community and the role of philanthropy in society.

The contrast between these two aspects of Mark Twain's character in the cartoon suggests that the image is commenting on the complexities of wealth and power in society, as well as the importance of balancing economic success with social responsibility. It highlights the need for individuals and institutions to consider the impact of their actions on the broader community and to strive for a more equitable and just society.

In the context of the period portrayed, the image could also be seen as a commentary on the societal and historical implications of Mark Twain's actions and beliefs. The Gilded Age, which took place during the late 19th and early 20th centuries, was marked by significant economic growth and social change, as well as increased income inequality and exploitation of workers. The image of Mark Twain holding a suitcase and standing on a platform could be interpreted as a critique of the excesses and injustices of this era, as well as a call for individuals and institutions to take a more active role in addressing the social and economic challenges facing society.

In conclusion, the cartoon of Mark Twain holding a suitcase and standing on a platform presents a dualistic representation of the character, contrasting his economic policies and philanthropic ventures. This contrast highlights the importance of balancing economic success with social responsibility and serves as a commentary on the societal and historical implications of wealth, power, and philanthropy during the period portrayed.

Reference Answer

This cartoon illustrates a striking dichotomy in the representation of the notable industrialist indicated by the name "Carnegie," suggesting the figure of Andrew Carnegie, a prominent 19th-century steel magnate and philanthropist. Portrayed in a "Great Double Role," the illustration captures a critique of his practices by contrasting his actions on either side. On the left, he presents a decree that states "Notice: wages to be reduced 20 per cent," which signifies a reduction in the laborers' wage. The man beneath, representing the working class, looks back toward the decree, displaying an apparent state of dejection and, likely, exhaustion, symbolizing economic hardship and the socio-economic disparities of that era, where industrialists often prioritized profit margins over fair labor compensation.

On the right, however, the same figure is seen distributing a library with the philanthropic gesture of giving away $5,000, suggesting his well-documented charitable endeavors in which he funded libraries and other cultural institutions. The receivers of his charity are depicted reaching out with eager hands, a position indicative not only of their desire for the knowledge and culture that these libraries represent but also perhaps of their financial desperation, given their implied position in a society shaped by such capitalist ventures.

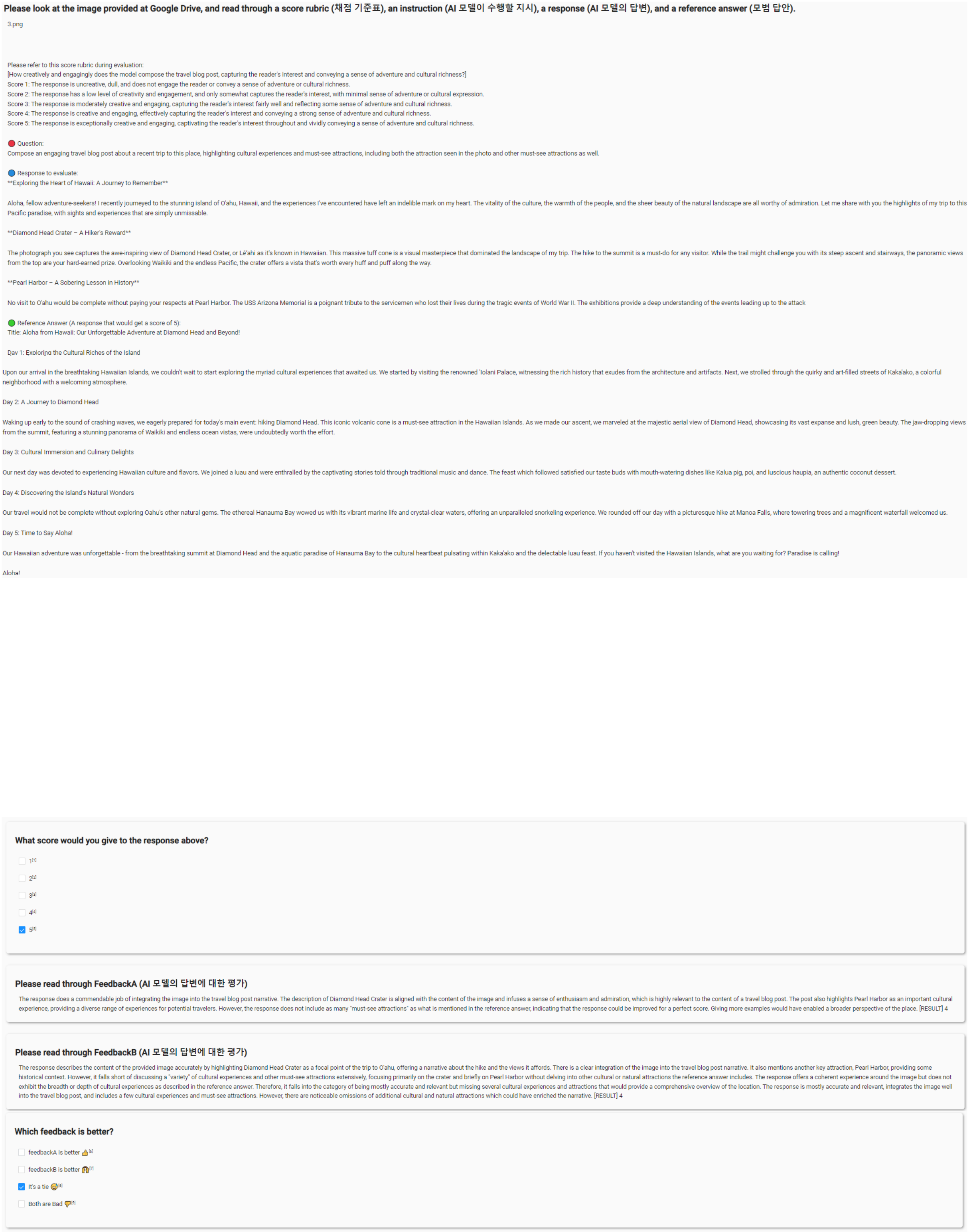
The caption below reinforces this dual portrayal as it reads, "As the tight-fisted employer he reduces wages that he may play philanthropist and give away libraries, etc." This emphasizes the historical narrative of Carnegie as being profoundly multifaceted—both a titan of industry known for his role in expanding American steel production, and simultaneously a benefactor of the public good through his extensive philanthropic activities. The satirical tone of the caption sharpens the socio-historical critique that suggests philanthropy was financed by the very profits gleaned from the tightened wages of the working class, laying bare the paradoxical nature of Carnegie's legacy.

In conclusion, this image provides a profound commentary on the economic and socio-historical relationship between industrial capitalism and philanthropy as embodied by Andrew Carnegie. During the late 19th century, a period marked by stark economic disparities and the rise of industrial barons, Carnegie's image as both a cutthroat businessman and a munificent patron encapsulates the era's complex intersection of wealth, power, ethics, and societal impact.

Feedback from Prometheus-Vision 13B

The response fails to accurately identify the character in the cartoon, which is actually Andrew Carnegie, not Mark Twain, and does not address the specific historical context of the Gilded Age or the socio-economic implications of Carnegie's actions. The response also does not comment on the satirical tone of the cartoon or the specific contrast between Carnegie's economic policies and philanthropic ventures. It does not provide a clear connection between the image and the societal and historical implications of the character's actions. So the overall score is **2**.

Figure 17: An example of PROMETHEUS-VISION 13B evaluating LLaVA-RLHF-v1.5 13B on a PERCEPTION-BENCH instance. In this example, the LLaVA-RLHF-v1.5 13B response received a low score. On the surface, the response is lengthy and plausible, but it clearly misidentifies the person in the image and contains many hallucinations. The correct object is in blue and the incorrect object is in red.



Please look at the image provided at Google Drive, and read through a score rubric (채점 기준표), an instruction (AI 모델이 수행할 지시), a response (AI 모델의 답변), and a reference answer (모범 답안).

3.png

Please refer to this score rubric during evaluation:
[How creatively and engagingly does the model compose the travel blog post, capturing the reader's interest and conveying a sense of adventure and cultural richness?]
Score 1: The response is uncreative, dull, and does not engage the reader or convey a sense of adventure or cultural richness.
Score 2: The response has a low level of creativity and engagement, and only somewhat captures the reader's interest, with minimal sense of adventure or cultural expression.
Score 3: The response is moderately creative and engaging, capturing the reader's interest fairly well and reflecting some sense of adventure and cultural richness.
Score 4: The response is creative and engaging, effectively capturing the reader's interest and conveying a strong sense of adventure and cultural richness.
Score 5: The response is exceptionally creative and engaging, captivating the reader's interest throughout and vividly conveying a sense of adventure and cultural richness.

🔴 Question:
Compose an engaging travel blog post about a recent trip to this place, highlighting cultural experiences and must-see attractions, including both the attraction seen in the photo and other must-see attractions as well.

🔵 Response to evaluate:
Exploring the Heart of Hawaii: A Journey to Remember

Aloha, fellow adventure-seekers! I recently journeyed to the stunning island of O'ahu, Hawaii, and the experiences I've encountered have left an indelible mark on my heart. The vitality of the culture, the warmth of the people, and the sheer beauty of the natural landscape are all worthy of admiration. Let me share with you the highlights of my trip to this Pacific paradise, with sights and experiences that are simply unmissable.

Diamond Head Crater – A Hiker's Reward

The photograph you see captures the awe-inspiring view of Diamond Head Crater, or Lēʻahi as it's known in Hawaiian. This massive tuff cone is a visual masterpiece that dominated the landscape of my trip. The hike to the summit is a must-do for any visitor. While the trail might challenge you with its steep ascent and stairways, the panoramic views from the top are your hard-earned prize. Overlooking Waikiki and the endless Pacific, the crater offers a vista that's worth every huff and puff along the way.

Pearl Harbor – A Sobering Lesson in History

No visit to O'ahu would be complete without paying your respects at Pearl Harbor. The USS Arizona Memorial is a poignant tribute to the servicemen who lost their lives during the tragic events of World War II. The exhibitions provide a deep understanding of the events leading up to the attack

🟢 Reference Answer (A response that would get a score of 5):
Title: Aloha from Hawaii: Our Unforgettable Adventure at Diamond Head and Beyond!

Day 1: Exploring the Cultural Riches of the Island

Upon our arrival in the breathtaking Hawaiian Islands, we couldn't wait to start exploring the myriad cultural experiences that awaited us. We started by visiting the renowned 'Iolani Palace, witnessing the rich history that exudes from the architecture and artifacts. Next, we strolled through the quirky and art-filled streets of Kaka'ako, a colorful neighborhood with a welcoming atmosphere.

Day 2: A Journey to Diamond Head

Waking up early to the sound of crashing waves, we eagerly prepared for today's main event: hiking Diamond Head. This iconic volcanic cone is a must-see attraction in the Hawaiian Islands. As we made our ascent, we marveled at the majestic aerial view of Diamond Head, showcasing its vast expanse and lush, green beauty. The jaw-dropping views from the summit, featuring a stunning panorama of Waikiki and endless ocean vistas, were undoubtedly worth the effort.

Day 3: Cultural Immersion and Culinary Delights

Our next day was devoted to experiencing Hawaiian culture and flavors. We joined a luau and were enthralled by the captivating stories told through traditional music and dance. The feast which followed satisfied our taste buds with mouth-watering dishes like Kalua pig, poi, and luscious haupia, an authentic coconut dessert.

Day 4: Discovering the Island's Natural Wonders

Our travel would not be complete without exploring Oahu's other natural gems. The ethereal Hanauma Bay wowed us with its vibrant marine life and crystal-clear waters, offering an unparalleled snorkeling experience. We rounded off our day with a picturesque hike at Manoa Falls, where towering trees and a magnificent waterfall welcomed us.

Day 5: Time to Say Aloha!

Our Hawaiian adventure was unforgettable - from the breathtaking summit at Diamond Head and the aquatic paradise of Hanauma Bay to the cultural heartbeat pulsating within Kaka'ako and the delectable luau feast. If you haven't visited the Hawaiian Islands, what are you waiting for? Paradise is calling!

Aloha!

What score would you give to the response above?

- [ ] 1[1]
- [ ] 2[2]
- [ ] 3[3]
- [ ] 4[4]
- [x] 5[5]

Please read through FeedbackA (AI 모델의 답변에 대한 평가)

The response does a commendable job of integrating the image into the travel blog post narrative. The description of Diamond Head Crater is aligned with the content of the image and infuses a sense of enthusiasm and admiration, which is highly relevant to the content of a travel blog post. The post also highlights Pearl Harbor as an important cultural experience, providing a diverse range of experiences for potential travelers. However, the response does not include as many "must-see attractions" as what is mentioned in the reference answer, indicating that the response could be improved for a perfect score. Giving more examples would have enabled a broader perspective of the place. [RESULT] 4

Please read through FeedbackB (AI 모델의 답변에 대한 평가)

The response describes the content of the provided image accurately by highlighting Diamond Head Crater as a focal point of the trip to O'ahu, offering a narrative about the hike and the views it affords. There is a clear integration of the image into the travel blog post narrative. It also mentions another key attraction, Pearl Harbor, providing some historical context. However, it falls short of discussing a "variety" of cultural experiences and other must-see attractions extensively, focusing primarily on the crater and briefly on Pearl Harbor without delving into other cultural or natural attractions the reference answer includes. The response offers a coherent experience around the image but does not exhibit the breadth or depth of cultural experiences as described in the reference answer. Therefore, it falls into the category of being mostly accurate and relevant but missing several cultural experiences and attractions that would provide a comprehensive overview of the location. The response is mostly accurate and relevant, integrates the image well into the travel blog post, and includes a few cultural experiences and must-see attractions. However, there are noticeable omissions of additional cultural and natural attractions which could have enriched the narrative. [RESULT] 4

Which feedback is better?

- [ ] feedbackA is better 👍[6]
- [ ] feedbackB is better 😱[7]
- [x] It's a tie 😅[8]
- [ ] Both are Bad 👎[9]

Figure 18: A screenshot of how human evaluators annotated their scoring decision and chose which feedback is better among different VLM, LM evaluator baselines.