Evaluating Image Review Ability of Vision Language Models

Anonymous ACL submission

Abstract

Large-scale vision language models (LVLMs) are language models that are capable of processing images and text inputs by a single model. This paper explores the use of LVLMs to generate review texts for images. The ability of LVLMs to review images is not fully understood, highlighting the need for a methodical evaluation of their review abilities. Unlike image captions, review texts can be written from various perspectives such as image composition and exposure. This diversity of review perspectives makes it difficult to uniquely determine a single correct review for an image. To address this challenge, we introduce an evaluation method based on rank correlation analysis, in which review texts are ranked by humans and LVLMs, then, measures the correlation between these rankings. We further validate this approach by creating a benchmark dataset aimed at assessing the image review ability of recent LVLMs. Our experiments with the dataset reveal that LVLMs, particularly those with proven superiority in other evaluative contexts, excel at distinguishing between high-quality and substandard image reviews.

1 Introduction

Large language models (LLMs), such as LLaMA (Touvron et al., 2023), GPT-3 (Brown et al., 2020), and Vicuna (Chiang et al., 2023), have achieved remarkable success in diverse natural language processing (NLP) tasks. Recently, leveraging the development of LLMs, several large-scale vision language models (LVLMs) (Alayrac et al., 2022; Liu et al., 2023b,a; Ye et al., 2023a,b; Bai et al., 2023a) have also been proposed with the objective of multimodal language comprehension and generation. LVLMs exhibit a remarkable capacity for perceiving, understanding, and processing high-level visual information. Currently, numerous researchers are exploring various tasks to broaden the application areas of LVLMs (Xu et al., 2023; Liu et al., 2023c; Bitton et al., 2023; Li et al., 2023; Bai et al., 2023b).

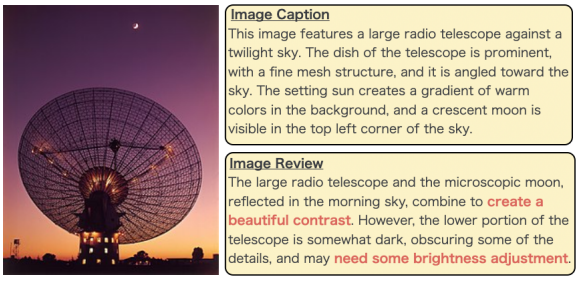

Figure 1: Image caption and review examples. The red-colored parts are reviewed from the perspectives of composition and exposure in the image.

In this paper, we introduce a novel task for LVLMs, which involves reviewing the good and bad points of a given image (see Fig.1). If LVLMs can properly perform image review tasks, it becomes possible to use them as substitutes for judges in image assessment contests, and to provide users with detailed and objective feedback to enhance the quality of images. However, the extent to which LVLMs can generate high-quality review texts remains unclear, necessitating further evaluation of their abilities. As shown in Fig.1, review texts, unlike image captions, can be written from various perspectives, such as image composition and exposure, making it challenging to identify a single correct review for an image. Therefore, evaluating the image review abilities is not straightforward.

To address this challenge, we introduce an evaluation method based on rank correlation analysis. The proposed method ranks review texts by humans and LVLMs and then measures the correlation between these rankings. It does not require uniquely determining a correct review text. Instead, manual annotation only needs to establish rankings for multiple review texts generated by a powerful LVLM, such as GPT-4V (OpenAI et al., 2023), thereby drastically reducing annotation costs. To validate

this approach, we construct a benchmark dataset containing 207 images selected from Wikipedia. Each image is accompanied by five review texts and a manually annotated ranking of these texts in both English and Japanese. Our experiments with the dataset reveal that LVLMs, particularly those with proven superiority in other evaluative contexts, excel at distinguishing between high-quality and substandard image reviews.

The dataset and source code for the proposed method are available from `https://[innominated]`. We released English and Japanese versions.

2 Related Work

2.1 Vision language models

Flamingo (Alayrac et al., 2022) is an early study on LVLM, which has demonstrated the potential of combining a visual encoder with an LLM where the authors show the performance of zero-shot or few-shot visual question answering (VQA). In subsequent studies, various models have been proposed. For instance, LLaVA (Liu et al., 2023b,a) explored instruction tuning, and mPLUG-Owl (Ye et al., 2023a,b) improved the alignment between texts and images by retraining the visual encoder. Moreover, Qwen-VL (Bai et al., 2023a) achieved impressive performance improvements in a variety of vision and language (V&L) tasks, which incorporates a vision language adapter for improving the ability of object recognition.

2.2 Evaluation for LVLMs

Early studies on LVLMs mainly focused on evaluations in simple VQA tasks (Antol et al., 2015; Lin et al., 2014). However, with the rapid development of LVLMs, some studies presented several benchmarks for evaluating higher-level perceptual and cognitive abilities of LVLMs (Fu et al., 2023; Li et al., 2023; Bai et al., 2023b). Recently, the methods utilizing LVLMs for evaluation have become more prevalent, such as LVLMs evaluating each other and using powerful LVLMs (e.g., GPT-4V (OpenAI et al., 2023)) as evaluators or for evaluation support. As in this study, methods for dataset construction using GPT-4V have also been explored (Xu et al., 2023; Liu et al., 2023c; Bitton et al., 2023; Li et al., 2023; Bai et al., 2023b).

3 Dataset Construction

Our dataset construction process consists of the following four steps; (1) Collecting images, (2) Generating five review texts, (3) Ranking review texts manually and (4) Filtering low-quality data.

3.1 Collecting images

The images are collected from the "Featured pictures" section of English Wikipedia. This section is composed of images, such as photographs, illustrations, and diagrams selected by user votes. The image data contained in this section is of very high quality and covers a diverse range of genres including artwork, natural landscapes, historical events, and science. We therefore select it as the image source.

3.2 Generating five review texts

Five review texts are generated for each image by using GPT-4V (OpenAI et al., 2023). Initially, we consider asking experts to create review texts or creating them ourselves, with reference to external knowledge. However, the diverse image genres make it difficult to gather experts. In addition, creation with reference to external knowledge takes a great deal of time. Thus, we decide to use GPT-4V to generate review texts.

In cases where five review texts are generated by merely adjusting the temperature parameters in GPT-4V, there tends to be no significant difference in the quality of the texts. That makes ranking annotations difficult. Hence, we formulate a prompt specifically designed to underscore distinctions. This prompt is tailored to generate five distinct review texts, each uniquely characterized by their degree of reasonableness and objectivity. See Appendix A.1 for detailed information about the prompt.

In addition, a generated text sometimes ends with contradictory expressions that negate itself, such as **"Note: the review contains an error as the stars are not blurred in the image provided."** This would happen because GPT-4V shapes its responses by Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). We check these phrases and remove them manually.

3.3 Ranking review texts manually

The five review texts of each image are manually ranked by X (≥ 3) annotators. To avoid potential biases from the generation order, the five review

texts are randomized before being shown to the annotators. To ensure consistency, detailed explanation and instruction described in Appendix B.1 are provided before the ranking process. The instruction is designed to rank the five review texts with reasonableness and objectivity. We break down "reasonable" into three core elements: Truthfulness, Consistency, and Informativeness, with a detailed explanation for each.

3.4 Filtering low-quality data

During annotation, errors may occur due to factors such as misinterpretation of data or annotator's fatigue and inattention. Such errors can adversely affect data quality. To alleviate these errors, we measure rank correlations among annotators and conduct filtering by setting a threshold on the rank correlation of the pair of annotators with the highest correlation. We define this pair with the highest rank correlation as **"top-correlated annotators"**. We here use Spearman's rank correlation coefficient. The coefficients take values ranging from -1 to 1, where, -1 indicates perfect inverse order and 1 indicates perfect equal order.

4 Evaluation Method

In this section, we describe the method for evaluating LVLM's review abilities on the dataset described in Sec.3. The evaluation method consists of the following two steps; (1) Ranking review texts by LVLM and (2) Measuring rank correlation between LVLM and humans.

4.1 Ranking review texts by LVLM

We employ perplexity as the evaluation metric for ranking review texts by LVLM. Perplexity is a measurement of the uncertainty of a language model's predictions for a given text. We compute perplexity by inputting both the image and its corresponding review text, along with a prompt described in Appendix A.2). Lower perplexity means that review texts are more confidently generated by the LVLM. Thus, we ranked the review texts in ascending order of perplexity. We call this method **perplexity-based ranking**.

In some LVLMs like GPT-4V (OpenAI et al., 2023), calculating perplexity is not straightforward. Therefore, we also consider a method of directly ranking with a Prompt (see Appendix A.3). We call this method **response-based ranking**. This experiment will be conducted in conjunction with the perplexity-based method in Sec.5.2.

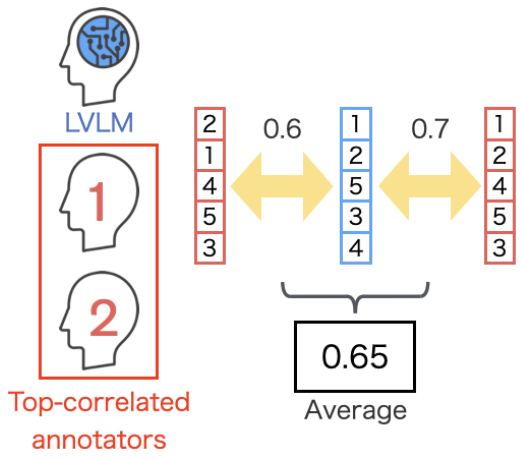

Figure 2: Measuring rank correlation between LVLM and human.

4.2 Measuring rank correlation between LVLM and human

The rank correlation between top-correlated annotators and an LVLM is measured using the procedure shown in Fig.2. Here, we employ Spearman's rank correlation coefficient.

After performing the filtering described in Sec.3.4, top-correlated annotators for each data sample are identified. Then, the correlation coefficient is calculated between the rankings by each of these two annotators and an LVLM, and the average of the coefficients is used to determine the correlation between the LVLM and humans. If the average correlation is high, the result indicates that review texts judged "excellent" by human annotators are likely to be more predictable for the LVLM while texts judged "inferior" are considered less predictable.

5 Experiments

5.1 Dataset construction & Filtering

We collected 207 images from 15 different genres, including Animals and Artwork (see Appendix C.1 for details). We generated five review texts for each image in English and Japanese. The English data were ranked by three native and near-native English speakers, whereas the Japanese data were ranked by three native Japanese speakers.

As present in Sec.3.4, errors may occur in annotations and there is a need for filtering. Thus, to identify an optimal threshold for balancing data count and correlation, we varied thresholds

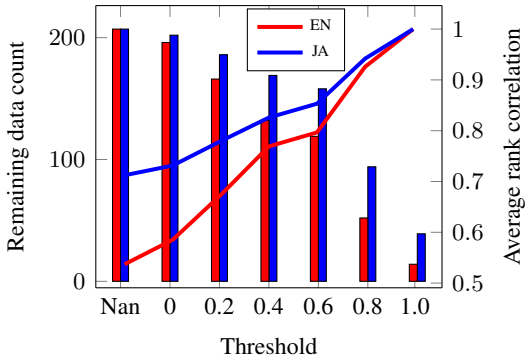

Figure 3: Changes for remaining data count and average rank correlation when varying threshold. The bar graphs represent the remaining data count and the line graphs denote average rank correlation. Nan means no threshold.

	mPLUG-Owl	mPLUG-Owl2	Qwen-VL
EN	0.310	0.365	**0.432**
JA	0.065	0.369	**0.487**

Table 1: Rank correlations between LVLMs and top-correlated annotators.

over positive values and analyzed changes in data count and average rank correlation between top-correlated annotators (refer to Sec.3.4). The results are shown in Fig.3. Setting the threshold to 0.8 resulted in a rapid decrease in the remaining data count. Alternatively, setting the threshold to 0.6 not only maintained over 100 data count but also showed an average correlation close to 0.8, a very strong correlation. Based on these results, we set the optimal threshold to 0.6. Furthermore, even when no threshold (denoted as Nan in Fig.3) was set, the average rank correlation was higher than 0.5. The high correlation suggests that annotations within our proposed data construction scheme can be conducted without the need for specialized training of the annotators.

5.2 Results on evaluation method

Perplexity-based ranking As presented in Sec.4, we evaluated LVLMs through rank correlation analysis. We selected three LVLMs for evaluation; mPLUG-Owl (Ye et al., 2023a), mPLUG-Owl2 (Ye et al., 2023b), and Qwen-VL (Bai et al., 2023a). The filtering threshold was set to 0.6 based on the results in Sec.5.1. Tab.1 shows the evaluation results for LVLMs. Qwen-VL, mPLUG-Owl2, and mPLUG-Owl showed stronger correlations with annotators in both Japanese and English, in that order. These three LVLMs have also demonstrated better performances on the existing evaluation tasks in the same order. Therefore, our method has a certain validity as an evaluation method for LVLMs. Moreover, the early LVLM, mPLUG-Owl, shows a significant decrease in performance with Japanese data compared to English data. This result suggests that newer LVLMs are more likely to support multiple languages, including Japanese.

Response-based ranking In some LVLMs like GPT-4V, calculating perplexity is challenging. Thus, we devised response-based ranking as an alternative to perplexity-based ranking. Specifically, we input images, five review texts, and a prompt described in Appendix A.3 into the LVLMs, and then conducted ranking based on the responses, similar to annotator ranking. After ranking, we evaluated them using the same procedure described in Fig.2. We evaluated GPT-4V using prompt-based ranking. We also attempted to evaluate Qwen-VL, but abandoned the effort because we could not control the output format.

As a result, English data showed a value of **0.399**, while the Japanese data showed a value of **0.506**. Based on these results, we found that GPT-4V would judge the quality of review texts with high correlation to humans. However, as demonstrated in the former paragraph, in the perplexity-based ranking experiment, Qwen-VL showed the same level of correlation as GPT-4V for Japanese data and achieved higher performance for English data. Although GPT-4V is generally considered to have higher performance than Qwen-VL, in this experiment, no significant performance difference was observed.

6 Summary

In this study, we proposed a method to assess the ability of LLMs to generate review text, which has been considered challenging. Furthermore, we created a benchmark dataset and conducted experiments to validate the proposed method.

Experiments showed that while LVLMs are still developing the ability to distinguish the quality of review texts at a level comparable to human sensibility, higher-performing models like Qwen-VL (Bai et al., 2023a) and GPT-4V (OpenAI et al., 2023) are increasingly demonstrating the capacity to generate superior review texts at a level akin to human judgment.

Limitations

While the proposed method emphasizes consistency and objectivity in assessing image review capabilities of LVLM, it does not evaluate from the perspective of domain knowledge, which remains a challenge for future work.

Ethics Considerations

In this study, our dataset was created using images obtained from English Wikipedia. The editors of English Wikipedia remove unnecessarily aggressive content[1], and we also excluded images involving political issues and other sensitive topics from our dataset. However, as acknowledged on its official pages[2], the present English Wikipedia allows the inclusion of information from sources that may be biased. Consequently, the dataset we developed might also reflect the inherent biases of the English Wikipedia.

Reproducibility Statement

In the experiments conducted in Sec.5.2 , we utilized publicly available models of mPLUG-Owl (Ye et al., 2023a), mPLUG-Owl2 (Ye et al., 2023b), Qwen-VL (Bai et al., 2023a), and GPT-4 API ver. 0.28.0 (OpenAI et al., 2023), using their default hyperparameters. Furthermore, our dataset and code are available at `https://[innominated]`.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023a. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

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.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

[1] `https://en.wikipedia.org/wiki/Wikipedia:Offensive_material`

[2] `https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view#Bias_in_sources`,`https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources#Biased_or_opinionated_sources`

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

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, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

A Details of Prompts

A.1 Prompt for generating five review texts

We generated five review texts for a certain image using the following prompt, which was designed to create differences among the five review texts.

Generation prompt

Please describe five different review texts about the good points and room for improvement of the image, following the constraints below:
1. Each review text should have different content.
2. The length of each review text should be almost the same.
3. Do not include bullet points within the review texts.
4. The review texts should be described in the following order: "Objective and reasonable," "Subjective but reasonable," "Objective but unreasonable," "Subjective and unreasonable," and "Subjective and containing an error".
5. Each review text should describe both the good points and room for improvement of the image.
6. If the image has no room for improvement, explicitly state that within the review text.

A.2 Prompt for measuring perplexity

We measured perplexity for each data using following prompt.

Perplexity prompt

Please describe a review text about the good points and room for improvement of the image

A.3 Prompt for ranking review texts

We input the following prompt into GPT-4V for response-base ranking. The content of this prompt is based on the instruction for human annotators in Appendix B.1.

GPT-4V ranking prompt

Below are the images and their review texts. Please rank the review text of each image from 1 to 5, in order of appropriateness. Please note that the numbers from 1 to 5 are not scores but rankings, and the smaller the number, the more appropriate it is. There should be no ties, and each rank from 1 to 5 should always appear once.
Please judge the appropriateness by the following aspects in the following order. That is, first, rank the texts by truthfulness. If there are equally truthful texts, rank them by consistency. Similarly, if they are equal also in consistency, rank them by informativeness; if they are equal also in it, rank them by objectivity; if they are equal also in it, rank them by fluency.
1. Truthfulness: Is it free of false information?
2. Consistency: Does it correspond to the image? 3. Informativeness: Does it describe detailed information or features of the image?
4. Objectivity: Is it an objective description?
5. Fluency: Is it grammatically correct?
If the text contains unfamiliar information, you may use a dictionary or search engine. However, please do not use a generative AI such as ChatGPT or image search.
Do not include the reason for ranking Absolutely respond in the following format. text1:2nd place, text2:3rd place, text3:1st place, text4:5th place, text5:4th place

B Details of Instruction

B.1 Instruction to annotators

The annotators ranked the review texts according to the following instructions.

Instruction

Below are the images and their review texts. Please rank the review text of each image from 1 to 5, in order of appropriateness. Please note that the numbers from 1 to 5 are not scores but rankings, and the smaller the number, the more appropriate it is. There should be no ties, and each rank from 1 to 5 should always appear once.
Please judge the appropriateness by the following aspects in the following order. That is, first, rank the texts by truthfulness. If there are equally truthful texts, rank them by consistency. Similarly, if they are equal also in consistency, rank them by informativeness; if they are equal also in it, rank them by objectivity; if they are equal also in it, rank them by fluency.

1. Truthfulness: Is it free of false information?
2. Consistency: Does it correspond to the image?
3. Informativeness: Does it describe detailed information or features of the image?
4. Objectivity: Is it an objective description?
5. Fluency: Is it grammatically correct?

If the text contains unfamiliar information, you may use a dictionary or search engine. However, please do not use a generative AI such as ChatGPT or image search.

C Details of Dataset

C.1 Image genres

The genre breakdown for the 207 collected image data is as follows. The numbers in () represent the number of images in each genre.

Animals (17) / Artwork (17) / Culture, entertainment, and lifestyle (16) / Currency (15) / Diagrams, drawings, and maps (15) / Engineering and technology (17) / Natural phenomena (15) / People (14) / Places (17) / Plants (16) / Sciences (15) / Space (15) / Vehicles (5) / Other lifeforms (3) / Other (10)

C.2 Correlation between prompt rank and human rank

The prompt given to GPT-4V (see Appendix A.1) instructs it to generate the following five types of review texts; **"Objective and reasonable," "Subjective but reasonable," "Objective but unreasonable," "Subjective and unreasonable," "Subjective and containing an error"**. This order of instructions is defined as **"prompt rank"**. In ranking, human annotators emphasized being reasonable and objective. Consequently, if GPT-4V generates review texts precisely following the prompt, we expect a match between the prompt rank and human rank.

Here, we analyzed the correlation between prompt rank and human rank, and investigated the extent to which GPT-4V can generate review texts following the prompt. Specifically, we measured the correlation between the prompt rank and top-correlated annotators rank as the threshold was changed. Fig.4 shows the results. Based on these

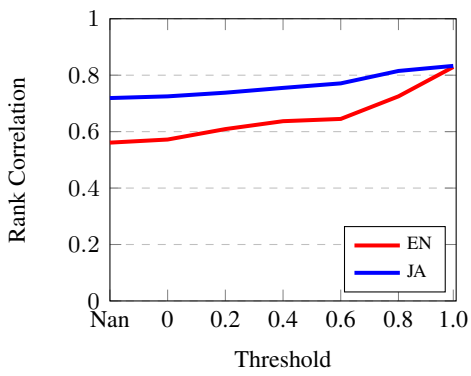

Figure 4: Correlation between prompt and human ranks.

results, the correlation between prompt rank and human rank showed a strong correlation close to 0.6 even without setting a threshold. These findings suggest that there is some validity in the assumption that the 5 review - "objective and consistent," "subjective but consistent," "objective but inconsistent," "subjective and inconsistent," and "subjective and containing errors" - are higher quality in the order of generation in this study's ranking instruction, which emphasizes being reasonable and objective. We also found that GPT-4V was able to generate review texts following the prompt to some extent.

Image	Five review texts
	1. Objective and reasonable: The image captures the exquisite silhouette of a large satellite dish against a vibrant sunset, creating a striking contrast that is visually appealing. The juxtaposition of technology and nature is well executed, showing a blend of human innovation with the beauty of a natural landscape. Room for improvement would include ensuring that the satellite dish is completely in focus to enhance the sharpness of its structure, which is slightly soft around the edges. **2. Subjective but reasonable**: This photograph exudes a profound sense of tranquility with the warm hues of the sky complementing the majestic presence of the satellite dish. The composition evokes a sense of wonder and could inspire contemplation about the universe beyond. However, personally, I feel that a wider angle could have included more of the surrounding landscape to further contextualize the dish within its environment. **3. Objective but unreasonable**: The technical quality of the image is commendable as it successfully captures the moon in a crescent phase above the dish, which adds an element of scale. Nevertheless, from a strictly compositional standpoint, the satellite appears disproportionately large compared to the moon, which might mislead an uninformed viewer about the actual size differences between the two objects. **4. Subjective and unreasonable**: The choice to photograph the satellite dish at dusk is enchanting, as it gives the whole scene a surreal, otherworldly vibe. On the downside, the moon looks disproportionately small, which personally offends my sense of cosmic balance. The photographer should have waited for a night when the moon was bigger or somehow made the satellite dish look smaller to better suit my taste. **5. Subjective and containing an error**: The image is enchanting with its blend of purples and blues creating a serene backdrop for the impressive structure of the satellite dish. The way the moon illuminates the scene adds a dreamy quality. However, the image would be more accurate if it showed the dish actively receiving signals, denoted by an unrealistic visual representation like visible beams of light coming down to the dish.

Table 2: Example of five review texts generated by GPT-4V.

C.3 Review texts example

Tab.2 displays an example of five review texts generated by GPT-4V.