Evaluating Image Review Ability of Vision Language Models

Anonymous ACL submission

Abstract

Large-scale Vision-Language Models (LVLMs) can process both images and text, demonstrating advanced capabilities in multimodal tasks like image captioning and visual question answering (VQA). However, it remains unclear whether they have an ability to understand and evaluate images, particularly in capturing the nuanced impressions and evaluations. To address this, we propose an image review evaluation method using rank correlation analysis. Our method asks a model to rank five review texts for an image. We then compare the model's rankings with human rankings to measure correlation. This enables effective evaluation of review texts that do not have a single correct answer. We validate this approach with a benchmark dataset of images from 15 categories, each with five review texts and annotated rankings in English and Japanese, resulting in over 2,000 data instances. Our experiments show that LVLMs excel at distinguishing between high-quality and low-quality reviews.

1 Introduction

Large language models (LLMs) like LLaMA (Touvron et al., 2023a), GPT-3 (Brown et al., 2020), and Vicuna (Chiang et al., 2023) have achieved remarkable success in NLP tasks. Recently, leveraging these developments, several large-scale vision language models (LVLMs) have been proposed (Alayrac et al., 2022; Liu et al., 2023b, 2024a; Ye et al., 2023, 2024; Bai et al., 2023a), exhibiting strong capabilities in visual information processing. Researchers are now exploring various tasks (Xu et al., 2023; Liu et al., 2023c; Bitton et al., 2023; Li et al., 2023a; Bai et al., 2023b) to broaden LVLM applications.

Despite these advancements, tasks like Visual Question Answering (VQA) (Zhang et al., 2022; Yue et al., 2024) and Image Captioning (Agrawal et al., 2019; Lin et al., 2014) primarily focus on factual understanding based on images. As shown in Figure 1, these tasks emphasize the "What" aspect by looking at the image and considering its description. However, these tasks do not take into account the "How" aspect of understanding the impressions and evaluations that people have when they view an image. Especially in real-world applications, models need the ability to integrate visual elements with viewers' impressions and reactions, and generate text based on how humans perceive the image. This need has not been adequately addressed by existing tasks, highlighting the necessity for a new metric to measure how well models convey visual information to people.

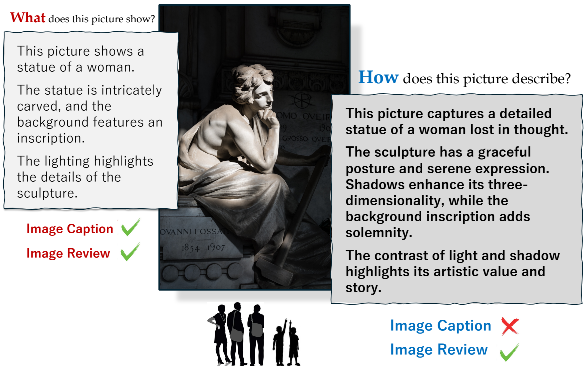

Figure 1: Exploring How Vision-Language Models Communicate Images: From What to How

A typical example where LVLMs need to excel is in creative support for art exhibitions and advertising. In art exhibitions, as shown in Figure 1, models must provide reviews considering the composition, color usage, and overall impact of the artworks. The review text can significantly influence the viewer's experience and understanding of the art. Similarly, in advertising, models must generate reviews that consider the visual elements of the advertisement and the responses from viewers. These tasks require LVLMs to go beyond factual recognition and understand how to convey visual content in a way that influences and engages people.

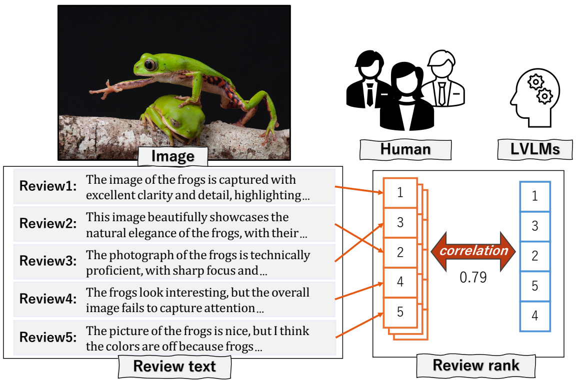

Figure 2: An evaluation metric using rank correlation analysis with an image and five review texts.

To address this challenge, we introduce a new benchmark based on rank correlation analysis. This method involves having both humans and LVLMs rank five review texts generated by GPT-4V (OpenAI et al., 2023) for a given image, then measuring the correlation between these rankings. In this way, we evaluate whether the model can go beyond mere factual recognition to understand visual content and accurately identify reviews that convey it appropriately. This approach allows for the effective evaluation of reviews, which do not have a single correct answer, by measuring rank correlation. To validate this approach, we constructed a benchmark dataset from Wikipedia consisting of images across 15 categories. Each image is accompanied by five review texts, which were manually annotated in both English and Japanese, resulting in over 2,000 reviews. Our experiments reveal that LVLMs, particularly those with strong evaluative capabilities, excel at distinguishing high-quality reviews from substandard ones. The dataset and source code are available at `https://anonymized`, with both English and Japanese versions provided.

2 LVLMs

LVLMs (Li et al., 2023b; Liu et al., 2024b; Bai et al., 2023a; Ye et al., 2024) integrate a Vision Encoder (Radford et al., 2021b) to process visual information with Large Language Models (LLMs) (Touvron et al., 2023a; Chiang et al., 2023; Bai et al., 2023a; Jiang et al., 2023). This integration requires further training to effectively combine vision and language capabilities. As a result, these LVLMs outperform conventional pre-trained models, even those with over ten times more parameters However, CLIP (Radford et al., 2021a) is primarily trained for image classification through contrastive learning. Additionally, when integrating visual processing capabilities with LLMs via CLIP, the training typically aims to align images with their short descriptions, focusing on factual content. This approach emphasizes the alignment between images and texts, but it is considered insufficient for generating reviews that understand and take into account the impressions and sensations conveyed to viewers.

3 Image Review Evaluation

3.1 Evaluation Method

Ranking Review Texts We use perplexity as the evaluation metric for ranking review texts generated by LVLM. Perplexity measures how confidently a language model predicts a given text. To compute perplexity, we provide the LVLM with a prefix instruction described in Appendix A.2 indicating that it is processing a review for an image, along with the image and its corresponding review text. We then rank the review texts in ascending order of perplexity, with lower perplexity indicating better review texts according to the model.

Measuring Rank Correlation For the five review texts, we compare the rankings annotated by humans with the rankings generated by the model. As shown in Figure 2, we calculate the correlation coefficients between the human and model outputs using Spearman's rank correlation coefficient, and use the average of these coefficients to determine the correlation between the LVLM and humans. If the average correlation is high, it indicates that the LVLM has a strong ability to distinguish high-quality reviews, understanding and evaluating review texts like humans.

3.2 Dataset Construction

The dataset construction process, illustrated in Figure 3, involves the following four steps:

STEP 1: Collecting images The images are collected from the "Featured pictures" section of English Wikipedia [1]. This section is composed of images, such as photographs, illustrations, and diagrams selected by user votes. The image data found in this section is of very high quality and covers a diverse range of genres including artwork,

[1] Wikipedia:Featured pictures `https://en.wikipedia.org/wiki/Wikipedia:Featured_pictures`

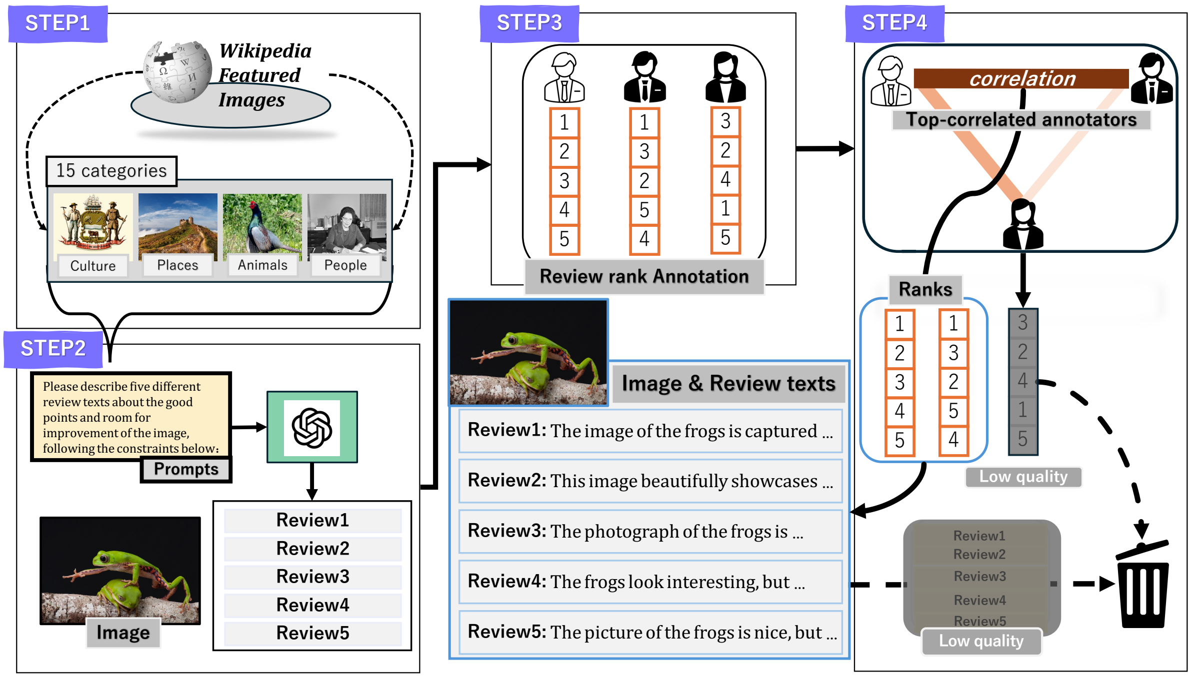

Figure 3: Dataset Construction Process.

natural landscapes, historical events, and science. We therefore select it as the image source.

STEP 2: Generating five review texts We use GPT-4V (OpenAI et al., 2023) to generate five review texts for each image. Due to the diverse image genres, it was difficult to gather experts and time-consuming to create texts using external references. Therefore, we chose to use GPT-4V. Simply adjusting GPT-4V's temperature parameters often results in similar quality texts, making ranking difficult. To address this, we designed a prompt to generate five distinct review texts with different levels of reasonableness (see Appendix A.1 for details). Additionally, the generated texts sometimes include self-contradictory expressions due to Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). We manually checked and removed such phrases.

STEP 3: Ranking review texts manually The five review texts of each image are manually ranked by X (≥ 3) annotators. The English texts were ranked by three native or near-native English speakers, and the Japanese texts were ranked by three native Japanese speakers. To avoid potential biases, the five review texts were randomized before being shown to the annotators. Detailed instructions (Appendix B) were provided to ensure consistency, focusing on reasonableness and objectivity. "Reasonable" was broken down into three core elements: Truthfulness, Consistency, and Informativeness, with detailed explanations for each.

STEP 4: Filtering low-quality data During annotation, errors from misinterpretation, fatigue, or inattention can affect data quality. To mitigate these, we measure rank correlations among annotators and filter by setting a threshold on the highest rank correlation pair. We define this pair as "top-correlated annotators." Spearman's rank correlation coefficient is used, ranging from -1 (perfect inverse order) to 1 (perfect equal order). We set the correlation coefficient threshold to 0.6, retaining only high-quality data with strong inter-annotator agreement (see Appendix D.4).

4 Experiments

4.1 Setup

We evaluated seven types of LVLMs and twelve models, along with six underlying LLMs based on these LVLMs to compare their perplexity in both English and Japanese. For details Appendix C. Unlike Local models, both GPT-4 and GPT-4V cannot measure the Perplexity of input tokens. Therefore, we provided five review texts with the same instructions as human annotations to generate rankings (see Appendix A.3).

LVLM	Size	EN	JP
mPLUG_Owl (Ye et al., 2023)	7B	0.310	0.065
mPLUG_Owl 2 (Ye et al., 2024)	7B	0.365	0.369
InstructBLIP (Vicuna-7B) (Dai et al., 2023)	7B	0.466	0.495
InstructBLIP (Vicuna-13B)	13B	0.496	0.520
LLaVA-1.5 (Vicuna-7B) (Liu et al., 2023a)	7B	0.516	**0.595**
LLaVA-1.5 (Vicuna-13B)	13B	0.529	0.591
Llava-Next (Vicuna-7B) (Liu et al., 2024b)	7B	0.510	**0.595**
Llava-Next (Vicuna-13B)	13B	0.535	0.553
Llava-Next (Mistral-7B)	7B	**0.543**	0.450
Llava-Next (Yi-34B)	34B	0.471	0.347
Qwen-VL-Chat (Bai et al., 2023a)	7B	0.432	0.487
GPT-4V (Reference)	-	0.399	0.506
Human (Reference)	-	0.795	0.846

Table 1: Correlation comparison of LVLMs in English and Japanese. The bold font indicates the best score.

4.2 Results

LVLMs Table 1 shows the evaluation results for LVLMs. In English, Llava-Next's Mistral-7B showed the highest performance, with all models achieving scores above 0.3, indicating a certain correlation with human judgment. In Japanese, all LVLMs based on Vicuna performed better in Japanese than in English, with Llava-Next and Llava-1.5's Vicuna-7B showing the highest score in Japanese. This suggests that despite being trained only in English (§2), the models can handle Japanese by inheriting LLMs' multilingual understanding abilities (Briakou et al., 2023). Additionally, comparing the performance of LVLMs in the same framework, such as those from Llava-Next, indicates that the underlying LLM is an important factor for LVLMs. Local models couldn't be evaluated with the same prompts as GPT-4V due to output format control issues. However, comparing with GPT-4V shows that ranking by perplexity doesn't degrade performance, validating our approach. Comparing the correlation with human annotations and the notable performance differences between models in our evaluation and other tasks (Liu et al., 2024c) suggests that LVLMs have room for improvement in image review evaluation.

LVLMs vs. LLMs We examined whether the models rank based solely on text quality. In English, LVLMs performed better than LLMs, indicating that image information slightly influences the results. However, the significant correlation suggests that text quality also plays an important role. The performance gap between LLMs and LVLMs in Japanese is larger than in English. For example, Llava-Next (Mistral-7B) scored 0.450 in Japanese, while Mistral-7B scored a low 0.194. This suggests that when solving tasks in Japanese, LVLMs are more likely to rely on image information.

LLM	Size	EN	JP
Llama 2 (Touvron et al., 2023b)	7B	0.319	0.413
Vicuna-7B (Chiang et al., 2023)	7B	0.362	**0.422**
Vicuna-13B	13B	0.358	0.365
Mistral-7B (Jiang et al., 2023)	7B	0.342	0.194
Yi-34B-Chat (AI et al., 2024)	34B	**0.405**	0.132
Qwen-Chat (Bai et al., 2023a)	7B	0.386	0.386
GPT-4 (Reference)	-	0.384	0.478

Table 2: Correlation comparison of LLMs in English and Japanese. The notations are the same as Table 1.

Threshold	-	0	0.2	0.4	0.6	0.8
Human Correlation (EN)	0.539	0.588	0.677	0.766	0.795	0.927
GPT-4V Included (EN)	0.344	0.354	0.390	0.387	0.399	0.464
Human Correlation (JP)	0.712	0.728	0.780	0.824	0.846	0.942
GPT-4V Included (JP)	0.478	0.485	0.494	0.511	0.506	0.543

Table 3: Correlation between Human and Human Including GPT-4 Evaluations.

Comparison between GPT-4V and Humans Table 3 shows the correlation between human annotations and the results when GPT-4V is included in the human annotations. GPT-4V was evaluated using the same prompt as human annotations, allowing for a direct comparison with human results. When setting a threshold to narrow down the dataset (see Appendix D.2), the agreement among human annotators increases, but the evaluation of GPT-4 shows only a slight increase in both English and Japanese. Specifically, when the threshold is 0.8, the correlation is in the 0.5 range for both languages, which is lower than the agreement among human annotators. These results indicate that while GPT-4V generally understands the human perspective, it does not align in certain aspects. It also shows that the ability to rank written reviews from a human perspective is still not very high, confirming the effectiveness of our method.

5 Conclusion

In this study, we proposed a method to evaluate LVLMs' ability to assess review texts and created a benchmark dataset. The results showed that most models demonstrated consistent and high accuracy across different languages. On the other hand, the correlation with human annotations and the notable performance differences between models in our evaluation and other tasks (Liu et al., 2024c) suggest that LVLMs have room for improvement in image review evaluation.

6 Limitations

Languages. In this study, we focused only on English and Japanese. It has not been comprehensively investigated whether the same applies to other languages. However, we aim to verify the multilingual capability of LVLMs. Since GPT-4 is primarily trained on English data, there is a possibility that it may generate biased outputs towards English-centric contexts. Moreover, English and Japanese are different languages in various aspects, such as language families, word order, and scripts. Therefore, we believe that using only English and Japanese is sufficient for the first step in experimenting with a multilingual setting. One of our research extension directions will be exploring the potential applications to other languages, such as Chinese, Spanish, or low-resource languages.

Number of the images. Our dataset consists of a relatively small number of images. We provide only the test data, so it will not be used for training. Moreover, recent datasets for LLM and LVLM are often evaluated using only a small number of test sets (Suzgun et al., 2023; Hendrycks et al., 2021; Yue et al., 2024). Furthermore, our dataset has undergone fine-grained manual checks, ensuring that it is clean. For these reasons, our dataset is sufficiently sized for evaluating LVLM. However, in the future, our dataset will be expanded to larger sizes, such as those used for instruction-tuning.

Prompt for annotations. We generated five review texts with GPT-4V using a single prompt. This might raise doubts about the variance across these lists of reviews. However, in dataset creation by LLMs, single-prompt attempts are made for cost efficiency and other reasons (Putri et al., 2024; Sakai et al., 2024). Additionally, the generated reviews are not the final output; they serve as the basis for further human ranking annotations. Therefore, GPT-4V was used to create seed reviews, while the actual annotations were done manually. Furthermore, despite variations in the quality of human annotations, the high correlation indicates that diverse review texts with inherent trends were successfully generated using a single prompt. Therefore, we believe that such concerns are unfounded. However, there is room for improvement in the prompts used for review generation, so prompt engineering will be left for future research.

7 Ethics Considerations

Licenses. We used Wikipedia materials in the dataset curation process. Wikipedia is available under fair use and the CC-BY-SA 4.0 license[2]. To clarify the data source information, we include the URL of the source images for each instance in our dataset. Additionally, our dataset includes outputs from GPT-4V under OpenAI's license terms[3]. OpenAI assigns to us all rights, titles, and interests in and to the output. As a result, we retain ownership rights. There are no restrictions on distributing the datasets, but using OpenAI's model output to develop models that compete with OpenAI is prohibited. Furthermore, we paid all recruited annotators above the minimum wage and obtained their consent for the transfer of rights to the annotated materials. Therefore, our dataset does not present any licensing concerns.

Moderations and biases. In this study, our dataset was created using images obtained from English Wikipedia. The editors of English Wikipedia remove unnecessarily aggressive content[4], and we also excluded images involving political issues and other sensitive topics from our dataset. However, as acknowledged on its official pages[5], 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.

References

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 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 8947–8956.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang,

[2] https://en.wikipedia.org/wiki/Wikipedia:Copyrights

[3] https://openai.com/policies/terms-of-use

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

[5] 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

Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

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.

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 Schmidt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. In Advances in Neural Information Processing Systems, volume 36, pages 26898–26922. Curran Associates, Inc.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

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.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. ArXiv, abs/2305.06500.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In International Conference on Learning Representations.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning, ICML'23. JMLR.org.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In European Conference on Computer Vision.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 26296–26306.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In Advances in Neural Information Processing Systems, volume 36, pages 34892–34916. Curran Associates, Inc.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. Mmbench: Is your multi-modal model an all-around player?

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.

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.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can llm generate culturally relevant commonsense qa data? case study in indonesian and sundanese.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021b. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR.

Yusuke Sakai, Mana Makinae, Hidetaka Kamigaito, and Taro Watanabe. 2024. Simultaneous interpretation corpus construction by large language models in distant language pair.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Findings of the Association for

Computational Linguistics: ACL 2023, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

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

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, D. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, A. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

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. 2023. 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, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13040–13051.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of CVPR.

Yao Zhang, Haokun Chen, Ahmed Frikha, Yezi Yang, Denis Krompass, Gengyuan Zhang, Jindong Gu, and Volker Tresp. 2022. Cl-crossvqa: A continual learning benchmark for cross-domain visual question answering.

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 point and room for improvement of the image, following the constraints below:

1. Each review texts 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 the following prompt.

Perplexity LVLM's 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-based ranking. The content of this prompt is based on the instruction for human annotators in Appendix B.

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

We input the following prompt into GPT-4 for responsed-base ranking without using an image.

GPT-4 ranking prompt

Please rank the review text by quality.

text1:review text1
text2:review text2
text3:review text3
text4:review text4
text5:review text5

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

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 Experimental setting

C.1 Reproduction statements

In the experiments conducted in Section 4.2, we utilized publicly available models for both LVLM and LLM, including mPLUG-Owl (Ye et al., 2023), mPLUG-Owl2 (Ye et al., 2024), InstructBLIP (Dai et al., 2023), LLava1.5 (Liu et al., 2023a), LLava-Next (Liu et al., 2024b), Qwen-VL-Chat (Bai et al., 2023a), and GPT-4 API ver. 0.28.0 (OpenAI et al., 2023), using their default hyperparameters. Additionally, our dataset and code are available at `https://[inominatenew]`. For LLMs, we used models such as Llama2 (Touvron et al., 2023b), Vicuna (Chiang et al., 2023), Mistral (Jiang et al., 2023), Yi-34B-Chat (AI et al., 2024), Qwen-Chat (Bai et al., 2023a), and GPT-4. To ensure a fair comparison of performance across multiple models, all experiments were conducted on an NVIDIA RTX 6000 Ada GPU, using 16-bit quantization to measure Perplexity. However, due to resource constraints, the LLaVA-NeXT (Yi-34B-Chat) model was loaded and inferred using an NVIDIA A100 80GB PCIe in 16-bit quantization. The same settings were applied to each model for performance comparison purposes.

C.2 LVLM details

Model	Base Model	HuggingFace Name/OpenAI API
mPLUG-Owl	LLaMA	MAGAer13/mplug-owl-llama-7b
mPLUG-Owl2	LLaMA2-7B	MAGAer13/mplug-owl2-llama2-7b
InstructBLIP (Vicuna-7B)	Vicuna-7B	Salesforce/instructblip-vicuna-7b
InstructBLIP (Vicuna-13B)	Vicuna-13B	Salesforce/instructblip-vicuna-13b
LLaVA-1.5	Vicuna-7B	liuhaotian/llava-v1.5-7b
LLaVA-1.5	Vicuna-13B	liuhaotian/llava-v1.5-13b
LLaVA-NeXT (Vicuna-7B)	Vicuna-7B	liuhaotian/llava-v1.6-vicuna-7b
LLaVA-NeXT (Vicuna-13B)	Vicuna-13B	liuhaotian/llava-v1.6-vicuna-13b
LLaVA-Next (Mistral)	Mistral	liuhaotian/llava-v1.6-mistral-7b
LLaVA-NeXT (Yi-34B)	Yi-34B	liuhaotian/llava-v1.6-34b
Qwen-VL-Chat	Qwen	Qwen/Qwen-VL-Chat
GPT-4-Vision	-	gpt-4-1106-vision-preview

C.3 LLM details

Model	HuggingFace Name
Llama2	meta-llama/Llama-2-7b
Vicuna-7B	lmsys/vicuna-7b-v1.5
Vicuna-13B	lmsys/vicuna-13b-v1.5
Mistral	mistralai/Mistral-7B-Instruct-v0.2
Yi-34B	01-ai/Yi-34B
Qwen-Chat	Qwen/Qwen-7B-Chat
GPT-4	gpt-4-1106-preview

D Details of Dataset

D.1 Image genres

The genre breakdown for the 207 collected image data is shown in Table 4.

D.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".

Category	Number of Items
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

Table 4: Categories and Number of Items.

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. Figure 4 shows the results.

Based on these 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

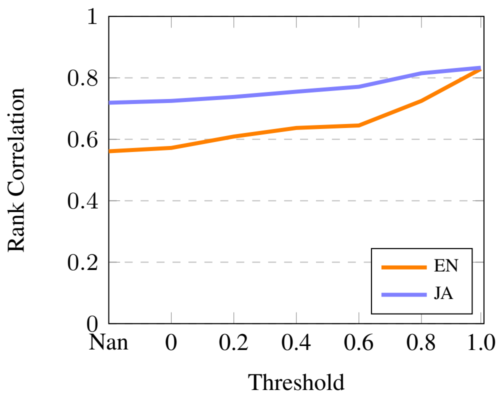

Figure 4: Correlation between prompt and human ranks.

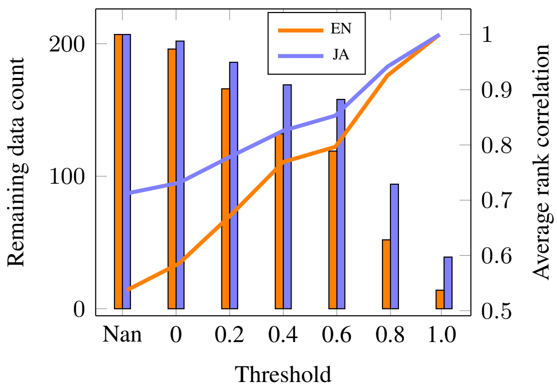

Figure 5: Changes for remaining data count and average rank correlation when varying threshold. The bar graphs represent the remaining data count and the line graphs denote average rank correlation. Nan means no threshold.

and objective. We also found that GPT-4V was able to generate review texts following the prompt to some extent.

D.3 Review texts example

Table 5, in English, and Table 6, in Japanese, each displays two examples of five review texts generated by GPT-4V.

D.4 Dataset construction & Filtering

We collected 207 images from 15 different genres, including Animals and Artwork (see Appendix D.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 Section 3.2, 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 over positive values and analyzed changes in data count and average rank correlation between top-correlated annotators (refer to Section 3.2). The results are shown in Figure 5. 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 Figure 5) 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.

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.
	**1. Objective and reasonable**: The image effectively captures the formidable structure of a hurricane, showcasing the spiraling clouds and the clear eye at the center, which is indicative of a well-developed storm system. The resolution is high, allowing meteorologists to analyze the storm's features in detail. However, the image is limited in conveying the scale of devastation potential; incorporating overlays indicating size relative to landmasses could enhance its informative value. **2. Subjective but reasonable**: From an aesthetic standpoint, the image is breathtaking, with the swirling clouds creating a mesmerizing pattern that is both beautiful and terrifying. The contrast between the storm and the surrounding environment is stark and evocative. On a more subjective note, the image could benefit from a different angle that provides a unique perspective, perhaps highlighting the juxtaposition of the hurricane's might against the vulnerability of the affected areas. **3. Objective but unreasonable**: he satellite image is of impeccable quality, presenting the hurricane with precision and clarity, which is crucial for tracking and prediction purposes. However, one might argue that the image should capture more of the surrounding environment to provide a comprehensive view of the storm's reach. This critique, while objective, may be unreasonable considering the limitations of satellite imaging technology and the focus required on the storm itself. **4. Subjective and unreasonable**: While the image captures the hurricane's intensity, one might find it lacks dramatic effect. The colors could be more vivid to convey the true ferocity of the storm. This perspective, though subjective, is unreasonable because the purpose of satellite imagery is to provide an accurate representation of reality, not to entertain or exaggerate features for dramatic impact. **5. Subjective and containing an error**: The photograph captures what appears to be a tranquil scene, misrepresenting the chaos usually associated with hurricanes. It's serene and misleadingly calm, which could be improved by adding elements that suggest movement or distress.

Table 5: Examples of five review texts generated by GPT-4V in the English version.

Image	Five review texts
	**1. Objective and reasonable**: この画像は明瞭な空の下での風車の構造を鮮明に捉えています。青空を背景にした風車の姿は、建築と自然の調和を表しているようです。また、風車の羽が完全に展開されていることで、その機能性を強調しています。ただし、画像の構図は風車を中心に据えすぎており、周囲の環境にもう少し焦点を当てることで、その場所の広がりや文化的背景に対する理解を深めることができたでしょう。 **2. Subjective but reasonable**:この風車の画像を見ると、私は開放感と歴史の重みを同時に感じます。深い青の空が風車の白い羽と石造りの体を引き立てており、そこには一種の詩的な美しさがあります。もし私が撮影者であったならば、風車の足元にもう少し焦点を当て、古い石の質感を表現してみたいと思いますが、この視点もまた、風車の威厳を感じさせるものです。 **3. Objective but unreasonable**: この風車の画像は、日中の明るい光の中で撮影されており、細部までくっきりと見えます。しかし、風車のデザインが古典的であるかのように感じる一方で、このような建造物が現代においても依然として機能しているかは疑問です。また、この風車が稼働している場面を捉えることができれば、よりダイナミックな画像になったかもしれません。 **4. Subjective and unreasonable**: この画像は、青と緑のコントラストが心地よく、風車の古風な魅力を引き出しています。風車の周囲には花や植物がなく、ある種の孤独感を与えていますが、それがまたこの場所の魔法のように感じます。風車の周囲にもっと生命があれば、画像に暖かみが出て、視覚的にも心にも温かい影響を与えたかもしれません。 **5. Subjective and containing an error**: この風車は非常に珍しいデザインで、まるで童話の中から飛び出してきたような不思議な存在感があります。背後に広がる森林が風車の孤立を際立たせています。この写真は、夜に撮影することで神秘的な雰囲気をより高めることができたかもしれませんが、風車の存在そのものがすでに完璧なため、改善の余地はないとも言えます。
	**1. Objective and reasonable**: 画像はクリアなブルースカイの背景に、奇妙な形状の樹木を特色としています。樹木の詳細が鮮明にキャッチされており、色彩が鮮やかです。ただし、画像の下部には乱雑な石が多く、これが全体のシンプルさを乱しているように見えます。 **2. Subjective but reasonable**: この画像は自然の美しさとユニークな景色を魅力的に捉えています。特に樹木の形状と色が印象的です。しかし、私の感覚では、石の配置が少し散らかって見えるため、それが画像の主要な焦点から視線を逸らしているように感じます。 **3. Objective but unreasonable**:画像には大きな樹木があり、背景には青空が広がっています。空は綺麗だが、樹木は他の植物と競合しているように見えます。また、画像の下部は不明瞭で何を示しているのか分かりにくい。 **4. Subjective and unreasonable**: 私はこの樹木がとても幻想的に見えると感じます。それはまるで別の惑星から来たようです。しかし、背景の青空はちょっと普通過ぎる気がします。そして、地面の石は何か秘密のメッセージを持っているのかもしれません。 **5. Subjective and containing an error**: この画像は熱帯のジャングルで撮影されたようです。大きな樹木はジャングルの中心部に位置しており、背景の青空は雨季の始まりを示しているかもしれません。ただ、ジャングルの地面にこんなに多くの石があることは稀ですので、もう少し自然な見た目にすることができるでしょう。

Table 6: Examples of five review texts generated by GPT-4V in the Japanese version.