IRR: Image Review Ranking Framework for Evaluating Vision-Language Models

Kazuki Hayashi[†], **Kazuma Onishi**[‡], **Toma Suzuki**[†], **Yusuke Ide**[†], **Seiji Gobara**[†], **Shigeki Saito**[‡], **Yusuke Sakai**[†], **Hidetaka Kamigaito**[†], **Katsuhiko Hayashi**[§], **Taro Watanabe**[†]

[†]Nara Institute of Science and Technology
[‡]Hokkaido University [§]The University of Tokyo
kazuki.hayashi.hl4@naist.ac.jp
{sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp
katsuhiko-hayashi@g.ecc.u-tokyo.ac.jp

Abstract

Large-scale Vision-Language Models (LVLMs) process both images and text, excelling in multimodal tasks such as image captioning and description generation. However, while these models excel at generating factual content, their ability to generate and evaluate texts reflecting perspectives on the same image, depending on the context, has not been sufficiently explored. To address this, we propose IRR: Image Review Rank, a novel evaluation framework designed to assess critic review texts from multiple perspectives. IRR evaluates LVLMs by measuring how closely their judgments align with human interpretations. We validate it using a dataset of images from 15 categories, each with five critic review texts and annotated rankings in both English and Japanese, totaling over 2,000 data instances [1]. Our results indicate that, although LVLMs exhibited consistent performance across languages, their correlation with human annotations was insufficient, highlighting the need for further advancements. These findings highlight the limitations of current evaluation methods and the need for approaches that better capture human reasoning in Vision & Language tasks.

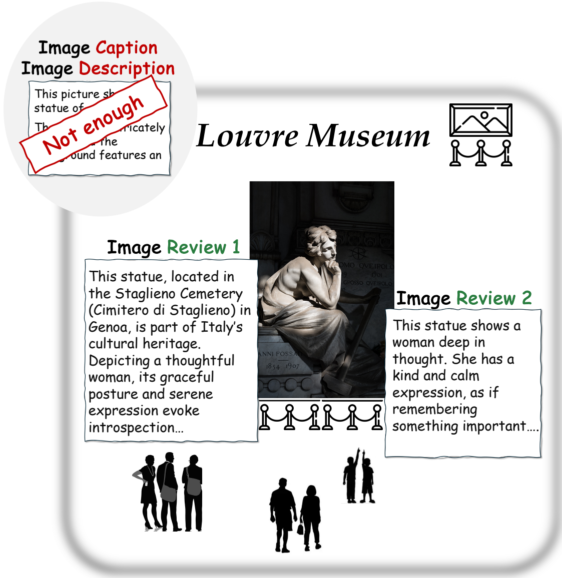

Figure 1: Different three image-to-text generation tasks and their corresponding metrics.

1 Introduction

Large language models (LLMs) (Touvron et al., 2023; OpenAI, 2023a; Chiang et al., 2023) have achieved significant success in NLP tasks. Recently, leveraging these developments, several large-scale vision language models (LVLMs) have been proposed (Liu et al., 2023b, 2024; Ye et al., 2023, 2024; Bai et al., 2023b), demonstrating strong abilities in visual information processing. One such application of LVLMs involves generating textual representations of image content, which can fulfill various practical purposes. While these models are capable of producing rich, detailed linguistic depictions of visual scenes, such representations are not constrained to a single "correct" interpretation.

In practice, image-related sentences often depend on contextual factors and the viewer's perspective, with their varying based on according to specific context and focus. For instance, as shown in Figure 1, adults may emphasize the historical and cultural significance of a statue, while children might prefer to imagine the story behind it or discuss the emotions it evokes. These differences highlight the importance of context-specific explanations that can accommodate various perspectives and interests (Biten et al., 2019; Naik et al., 2023).

This inherent diversity in image interpretation introduces significant challenges for evaluating LVLMs. Existing evaluation approaches for image captioning (Lin et al., 2014; Desai and Johnson, 2020; Li et al., 2020), image description (Kreiss

[1]The datasets are available at https://hf.co/datasets/naist-nlp/Wiki-ImageReview1.0

Proceedings of the 31st International Conference on Computational Linguistics, pages 9939–9956
January 19–24, 2025.

et al., 2022; Stangl et al., 2021), and image explanation (Hayashi et al., 2024; Ozaki et al., 2024) primarily rely on comparing generated texts with reference texts, assuming a single "correct" answer based on factual elements of the image (Vedantam et al., 2015; Lin, 2004; Papineni et al., 2002). Such reference-based evaluations are unsuitable for assessing review texts, as they cannot comprehensively capture all valid interpretations, even with multiple references. Recent studies have explored methods to align model evaluations with human judgments. For instance, pairwise comparisons have been proposed to better reflect human preferences. However, these approaches often fail to account for diverse perspectives. Similarly, reference-free evaluation frameworks have been introduced (Chen et al., 2024a; Jiang et al., 2024; Xiong et al., 2024). Nevertheless, further validation is needed to determine whether these models (Radford et al., 2021; Ye et al., 2023, 2024; Liu et al., 2024; Bai et al., 2023b) can effectively capture context and evaluate texts from a human perspective.

To address this, we propose IRR: IMAGE REVIEW RANK, a new evaluation framework designed to evaluate critic review texts from multiple perspectives. IRR recognizes various interpretations and evaluates the understanding of LVLMs based on how closely their judgments align with those of humans. Our method has the model rank five review texts for a given image, comparing the rankings with human rankings to measure correlation. This approach assesses whether the model can go beyond factual recognition to identify the most contextually appropriate review. We constructed a dataset using images from Wikipedia, covering 15 categories. Each image has five review texts generated by GPT-4V (OpenAI, 2023b), manually annotated in English and Japanese, totaling over 2,000 data instances. Both the English and Japanese datasets were created using the same methodology.

Our results indicate that, although LVLMs performed consistently across languages, their correlation with human annotations reveals room for improvement. Additionally, our framework demonstrates that methods like CLIP (Radford et al., 2021) are insufficient for evaluating texts in the context of image review. By integrating LLM inferential capabilities with visual information, we better align with human reasoning, further highlighting the limitations of relying solely on CLIP Score (Hessel et al., 2021).

2 Evaluation framework

2.1 Task: Ranking Review Texts

We design a task where the model ranks pre-generated texts for a given image based on their relevance within the context of that image. Perplexity is used as the evaluation metric, as it correlates with human judgments of sentence quality (Lau et al., 2020; Muñoz Sánchez et al., 2024). Lower perplexity values indicate the model's ability to predict the next word based on context, making it a reliable metric for assessing how well a text aligns with the image context. To evaluate relevance, perplexity is computed by providing the LVLM with a prefix instruction (see Fig. A.2) indicating it is processing a review for an image. The image and each review text are then input into the model one at a time. Among the five review texts, those with lower perplexity values are considered better aligned with the image context. The review texts are ranked in ascending order of perplexity.

2.2 Metric: Measuring Rank Correlation

To assess how well the LVLM's rankings align with human judgments, we calculate Spearman's rank correlation coefficient (Spearman, 1904). This coefficient ranges from -1 (perfect inverse order) to 1 (perfect alignment) and is computed between the model's rankings and the human rankings for the five texts associated with each image. Among the three annotators, we first identify the pair with the highest agreement and then calculate the correlation between the LVLM's rankings and each of these two annotators. The average of these two correlations represents the degree of alignment between the model's rankings and human evaluations. A higher value indicates a stronger correlation, reflecting the model's ability to identify high-quality review texts similarly to humans.

3 Dataset construction

The dataset construction process, shown in Figure 2, was applied to create both the English and Japanese datasets following the same methodology.

STEP 1: Collecting images We collected images from the "Featured pictures" section of English Wikipedia[2]. This section comprises high-quality images such as photographs, illustrations,

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

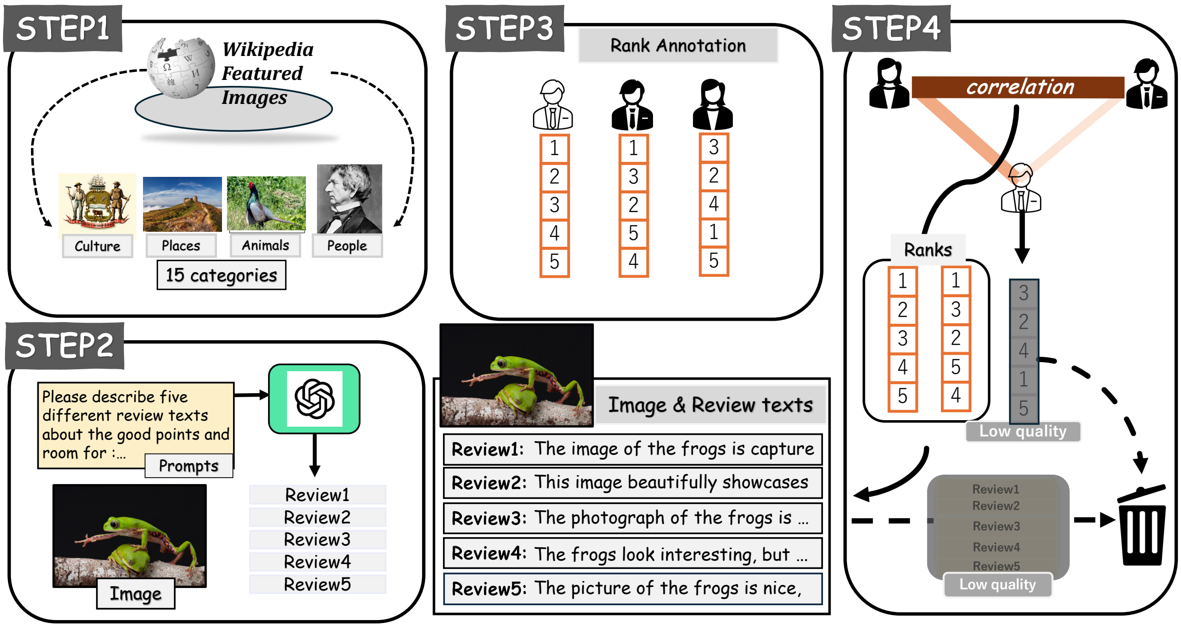

Figure 2: Dataset Construction Process.

and diagrams selected by user votes, covering diverse genres such as artwork, natural landscapes, historical events, and science. We therefore selected it as the image source. To achieve a diverse selection, we included images from 15 categories, as outlined in Table 1.

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 1: Categories and Number of Items.

STEP 2: Generating Review Texts We used GPT-4V to generate five distinct review texts for each image, reflecting different levels of "Reasonableness" and "Objectivity." Given the diverse range of image genres, gathering experts or creating texts using external references would have been extremely time-consuming and impractical. Therefore, we opted to use GPT-4V for text generation. To ensure that the review texts were of varying quality, we designed a prompt to generate five review texts with different levels of reasonableness (see Appendix A.1 for details).

STEP 3: Ranking review texts manually The five review texts of each image are manually ranked by X (≥ 3) annotators. English texts were ranked by native or near-native English speakers, and Japanese texts by native Japanese speakers. To avoid potential biases, the five review texts were randomized before presentation to the annotators. Annotators followed detailed instructions (Appendix B) that emphasized two key criteria: "Reasonableness" and "Objectivity." To guide annotators in selecting the most appropriate text within the context of the image review, "Reasonableness" was further divided into three subcategories: truthfulness, consistency, and informativeness.

STEP 4: Filtering low-quality data During annotation, errors caused by misinterpretation, fatigue, or inattention can reduce data quality. To mitigate this, we measure the rank correlations among annotators and filter the data using the correlation scores of the annotator pair with the highest agreement. Specifically, we used the correlation coefficient scores described in Section 2.2 to filter the data. We set the correlation coefficient threshold at 0.6, retaining only high-quality data with strong

LVLM	Size	EN	JP
mPLUG_Owl	7B	0.310	0.065
mPLUG_Owl 2	7B	0.365	0.369
InstructBLIP (Vicuna-7B)	7B	0.466	0.495
InstructBLIP (Vicuna-13B)	13B	0.496	0.520
LLaVA-1.5 (Vicuna-7B)	7B	0.516	**0.595**
LLaVA-1.5 (Vicuna-13B)	13B	0.529	0.591
LLaVA-NeXT (Vicuna-7B)	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., 2023b)	7B	0.432	0.487
GPT-4V (Reference)	-	0.399	0.506
CLIP Score (Reference)	-	-0.437	-
Human (Reference)	-	0.795	0.846

Table 2: Correlation comparison of LVLMs in English and Japanese. The bold font indicates the best score.

inter-annotator agreement (see Appendix D.3).

4 Experiments

4.1 Setup

We evaluated 12 models spanning 7 types of LVLMs: mPLUG-Owl (Ye et al., 2023), mPLUG-Owl2 (Ye et al., 2024), InstructBLIP (Dai et al., 2023), LLaVA-1.5 (Liu et al., 2023a), LLaVA-NeXT (Liu et al., 2024), Qwen-VL-Chat (Bai et al., 2023b), and GPT-4 Vision. Additionally, we evaluated six underlying LLMs as the foundation for these LVLMs: Llama 2 (Touvron et al., 2023), 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. These models were evaluated to compare their perplexity-based ranking in both English and Japanese. For further details, refer to Appendix C. As GPT-4 and GPT-4V lack perplexity measurement for input tokens, we addressed this by ranking five review texts using the same instructions as human annotations based on context-specific relevance (see Appendix A.3).

4.2 Results

LVLMs Table 2 presents the evaluation results for LVLMs. In English, LLaVA-NeXT's Mistral-7B achieved the highest performance, with all models scoring between 0.3 and 0.6, demonstrating a moderate correlation with human judgment (Spearman, 1904). 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. These models suggest that, despite being trained only in English (§5), they have inherited the multilingual understanding abilities of LLMs and can effectively handle Japanese, even in vision-based reasoning tasks (Briakou et al., 2023)." In Table 2, the 'Human Reference' reveals the correlation among human annotations. Comparing this with the correlation obtained from LVLMs indicates that there is potential for improvement in LVLM performance in the conducted evaluation.

LLM	Size	EN	JP
Llama 2	7B	0.319	0.413
Vicuna-7B	7B	0.362	**0.422**
Vicuna-13B	13B	0.358	0.365
Mistral-7B	7B	0.342	0.194
Yi-34B-Chat	34B	**0.405**	0.132
Qwen-Chat	7B	0.386	0.386
GPT-4 (Reference)	-	0.384	0.478
CLIP Score (Reference)	-	-0.437	-

Table 3: Correlation comparison of LLMs in English and Japanese. The notations are the same as Table 2.

Threshold	<0	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 4: Correlation between Human and Human Including GPT-4 Evaluations.

LVLMs vs. LLMs To examine whether the models rely solely on text quality or also utilize image information for ranking, we conducted the same evaluation using the underlying LLMs without image inputs. As shown in Table 3, in English, LVLMs slightly outperformed their corresponding LLMs, indicating that image information improves alignment with human judgments. The weak correlation observed with LLMs suggests that while text quality contributes to the rankings, it has limitations. Notably, the performance gap between LVLMs and LLMs was larger in Japanese; for example, LLaVA-NeXT (Mistral-7B) achieved a correlation of 0.450 in Japanese, while Mistral-7B alone scored 0.194. This suggests that LVLMs may have the potential to utilize image information more effectively in Japanese. These findings indicate that multimodal inputs could potentially enhance model performance, particularly in Japanese.

Comparison to CLIP Score Instead of using perplexity for LVLMs, we employed CLIP Score (Hessel et al., 2021) to measure the alignment between images and text, and used it for scoring and ranking. However, as shown in the table, the CLIP alignment score resulted in negative correlation values. Therefore, it is clearly shown in our framework that relying solely on image-text alignment methods like CLIP is insufficient for evaluating the appropriateness of texts in the context of image review. We can only achieve alignment with human reasoning by integrating the inferential capabilities of LLMs with visual information. This empirically confirms the limitations of existing studies that use CLIP Score as an evaluation metric.

Comparison between GPT-4V and Humans Table 4 shows the correlation between human annotations and GPT-4V's rankings when GPT-4V is included alongside human annotators. GPT-4V was evaluated using the same prompt as human annotators for direct comparison. To focus on reliably annotated data, we set a threshold on inter-annotator agreement (see Appendix D.1), retaining only instances where human annotators exhibited high agreement. While the agreement among human annotators increased with higher thresholds, the correlation between GPT-4V's rankings and human annotations showed only a slight improvement in both English and Japanese, remaining approximately 0.5 even at a threshold of 0.8. These results suggest that although GPT-4V generally captures human perspectives, it diverges in some areas, indicating that its ability to rank reviews in alignment with human judgments is still limited.

5 Related Work

LVLMs LVLMs (Li et al., 2023; Liu et al., 2024; Bai et al., 2023b; Ye et al., 2024) integrate a Vision Encoder (Li et al., 2023), trained through contrastive learning for visual information processing, with Large Language Models (LLMs) (Touvron et al., 2023; Chiang et al., 2023; Bai et al., 2023a; Jiang et al., 2023). This integration requires additional training to effectively combine vision and language capabilities. Consequently, these LVLMs outperform conventional pre-trained models, even those with over ten times more parameters (Alayrac et al., 2022; Driess et al., 2023). However, CLIP (Radford et al., 2021), a prominent approach in this domain, primarily aligns images with concise and factual descriptions through contrastive learning. Furthermore, when integrating visual processing capabilities into LLMs via CLIP to construct LVLMs, this method remains limited to factual alignment and lacks the ability to handle diverse perspectives. As a result, LVLMs may be insufficient for generating and interpreting texts that reflect diverse perspectives or identifying the most contextually appropriate ones.

Existing Evaluation Frameworks Recent V&L evaluation frameworks, like Chatbot Arena (Chiang et al., 2024) and WildVision Arena (Lu et al., 2024), use pairwise comparisons to align more closely model outputs with human preferences. While these frameworks enhance subjective evaluation, they are primarily designed for text-based dialogue or general multimodal tasks and lack multi-perspective consideration. LLaVA-Critic (Xiong et al., 2024), MM-Vet (Yu et al., 2024a,b), and MLLM as a Judge (Chen et al., 2024a) also utilize LLMs to evaluate LVLMs across various visual tasks focused on factual accuracy and reasoning. However, these frameworks are inadequate for critic review texts requiring various viewpoints. Additionally, their validation for variety and reliability with LLM evaluators is limited, limiting reliable multi-perspective assessments. In contrast, our Image-Review framework provides thorough validation for models like LLaVA-Critic and MLLM-as-a-Judge.

6 Conclusion

In this study, we proposed IRR: IMAGE REVIEW RANK, a novel evaluation framework to assess the ability of LVLMs to rank image texts from multiple perspectives and created a corresponding critic review dataset. Our results showed that while LVLMs demonstrated consistent performance across languages, their correlation with human annotations shows room for improvement, highlighting areas for further improvement. Additionally, our framework reveals that methods like CLIP (Radford et al., 2021) are not adequate for evaluating texts in the context of image reviews. By integrating the inferential capabilities of LLMs with visual information, we improved alignment with human reasoning, further highlighting the challenges of using CLIP Score alone (Hessel et al., 2021). Furthermore, our evaluation framework enables testing model adaptability across contexts by altering the dataset domain. For instance, museum guides require detailed explanations of history and art, while advertising prioritizes concise, attention-grabbing text. Adapting to such diverse contexts remains a key challenge for future research.

7 Limitations

Languages. In this study, we focused only on English and Japanese, which allowed us to explore the multilingual capabilities of LVLMs. However, it has not been thoroughly investigated whether these findings apply to other languages. One potential limitation is the difference in token length: English had an average token length of 94.77, while Japanese had 50.20, as measured using a multilingual tokenizer. Although Japanese is often considered more compact, direct comparisons are difficult due to structural differences between the languages. This issue reflects the difficulty of controling output lengths by LLMs (Juseon-Do et al., 2024). Additionally, the quality of annotations may have varied, as English annotations were provided by non-native speakers, while Japanese annotations were done by native speakers. These factors make it challenging to directly compare accuracy across languages. Note that machine translation may diversify our dataset into multiple languages, whereas it ignores cultural aspects covered by human annotators (Sakai et al., 2024b).

Number of the images. The dataset used in this study is domain-specific to Image Review, and it is unclear whether the results can be generalized to other domains. Additionally, the number of images is relatively small, and only test data is provided. Therefore, evaluation using training data or testing generalization capabilities through comparisons with larger datasets remains insufficient.

Prompt for annotations. We generated five review texts using GPT-4V with a single prompt, which raises concerns about potential biases introduced by the model (Sakai et al., 2024c). Specifically, GPT-4V is known to exhibit a positivity bias (Bender et al., 2021), where certain environmental elements tend to lead to overly positive descriptions. Such biases may result in an imbalanced dataset that does not adequately reflect the diversity of real-world reviews. Additionally, since all reviews were generated solely by GPT-4V, there is a possibility that fundamental biases inherent to the model are embedded in the dataset. While human ranking and filtering of reviews were done manually and through correlation coefficients, these biases still raise concerns regarding the generalizability and neutrality of the dataset.

Entity information. When LVLMs cannot understand entities in given image reviews, their quality of image review evaluation decreases. Furthermore, image review evaluation requires the alignment of entity knowledge between LLMs and vision encoders similar to other vision and language tasks (Kamigaito et al., 2023). Currently, a decisive approach to this problem does not exist. Vision RAG (Faysse et al., 2024) requires an image review specific datastore. Multimodal Knowledge Graphs (KGs) (Chen et al., 2024b) are sparse. KG completion (KGC) based on the pre-trained language model (PLM) (Yao et al., 2019) suffers from data leakage (Sakai et al., 2024a). Traditional KGC models (Nickel et al., 2011) are reliable by theoretical studies like Kamigaito and Hayashi (2021, 2022a,b); Feng et al. (2023, 2024), while their performance is lower than PLM-based ones.

8 Ethics Considerations

Licenses. We used Wikipedia materials in the dataset curation process. While Wikipedia text content is available under fair use and the CC-BY-SA 4.0 license[3], we recognize that images have individual licenses. We verified that the images used, particularly from the "Featured Images" section, are covered by the appropriate licenses, including CC-BY-SA when applicable. For transparency, we provide the URL to the source for each image. Additionally, our dataset includes outputs from GPT-4V under OpenAI's license terms[4], granting us full ownership of the generated content.

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[5], and we also excluded images involving political issues and other sensitive topics from our dataset. However, as acknowledged on its official pages[6], 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.

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

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

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

[6] 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

References

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, 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. Preprint, arXiv:2403.04652.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In Advances in Neural Information Processing Systems.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. Preprint, arXiv:2309.16609.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. Preprint, arXiv:2308.12966.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 610–623.

Ali Furkan Biten, Lluis Gomez, Marcal Rusinol, and DImosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In Proceedings - 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 12458–12467. Funding Information: This work has been supported by projects TIN2017-89779-P, Marie-Curie (712949 TECNIOspring PLUS), aBSINTHE (Fundación BBVA 2017), the CERCA Programme / Generalitat de Catalunya, NVIDIA Corporation and a UAB PhD scholarship. Publisher Copyright: © 2019 IEEE.

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.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pages 6562–6595. PMLR.

Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, Jiaqi Li, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, and Huajun Chen. 2024b. Knowledge graphs meet multi-modal learning: A comprehensive survey. Preprint, arXiv:2402.05391.

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.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. Preprint, arXiv:2403.04132.

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.

Karan Desai and Justin Johnson. 2020. Virtex: Learning visual representations from textual annotations. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11157–11168.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-e: An embodied multimodal language model. In Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pages 8469–8488. PMLR.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. Preprint, arXiv:2407.01449.

Xincan Feng, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2023. Model-based subsampling for knowledge graph completion. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 910–920, Nusa Dua, Bali. Association for Computational Linguistics.

Xincan Feng, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Unified interpretation of smoothing methods for negative sampling loss functions in knowledge graph embedding. In Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024), pages 78–98, Bangkok, Thailand. Association for Computational Linguistics.

Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Towards artwork explanation in large-scale vision language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 705–729, Bangkok, Thailand. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. Preprint, arXiv:2310.06825.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. 2024. Genai arena: An open evaluation platform for generative models. Preprint, arXiv:2406.04485.

Juseon-Do, Jingun Kwon, Hidetaka Kamigaito, and Manabu Okumura. 2024. Instructcmp: Length control in sentence compression through instruction-based large language models. Preprint, arXiv:2406.11097.

Hidetaka Kamigaito and Katsuhiko Hayashi. 2021. Unified interpretation of softmax cross-entropy and negative sampling: With case study for knowledge graph embedding. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5517–5531, Online. Association for Computational Linguistics.

Hidetaka Kamigaito and Katsuhiko Hayashi. 2022a. Comprehensive analysis of negative sampling in knowledge graph representation learning. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 10661–10675. PMLR.

Hidetaka Kamigaito and Katsuhiko Hayashi. 2022b. Erratum to: Comprehensive analysis of negative sampling in knowledge graph representation learning.

Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2023. Table and image generation for investigating knowledge of entities in pre-trained vision and language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1904–1917, Toronto, Canada. Association for Computational Linguistics.

Elisa Kreiss, Fei Fang, Noah Goodman, and Christopher Potts. 2022. Concadia: Towards image-based text generation with a purpose. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4667–4684, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? sentence acceptability in context. Transactions of the Association for Computational Linguistics, 8:296–310.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. 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, volume 202 of Proceedings of Machine Learning Research, pages 19730–19742. PMLR.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Computer Vision – ECCV 2020, pages 121–137, Cham. Springer International Publishing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization

Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll¥'ar, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In Computer Vision – ECCV 2014, pages 740–755, Cham. Springer International Publishing.

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In Thirty-seventh Conference on Neural Information Processing Systems.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. Wildvision: Evaluating vision-language models in the wild with human preferences. Preprint, arXiv:2406.11069.

Ricardo Muñoz Sánchez, Simon Dobnik, and Elena Volodina. 2024. Harnessing GPT to study second language learner essays: Can we use perplexity to determine linguistic competence? In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), pages 414–427, Mexico City, Mexico. Association for Computational Linguistics.

Nandita Shankar Naik, Christopher Potts, and Elisa Kreiss. 2023. Context-vqa: Towards context-aware and purposeful visual question answering. 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pages 2813–2817.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, page 809–816, Madison, WI, USA. Omnipress.

OpenAI. 2023a. Gpt-4 technical report. ArXiv, abs/2303.08774.

OpenAI. 2023b. Gpt-4v(ision) system card.

Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Towards cross-lingual explanation of artwork in large-scale vision language models. Preprint, arXiv:2409.01584.

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.

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. 2021. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning.

Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024a. Does pre-trained language model actually infer unseen links in knowledge graph completion? In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8091–8106, Mexico City, Mexico. Association for Computational Linguistics.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024b. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In Findings of the Association for Computational Linguistics: ACL 2024, pages 14182–14214, Bangkok, Thailand. Association for Computational Linguistics.

Yusuke Sakai, Adam Nohejl, Jiangnan Hang, Hidetaka Kamigaito, and Taro Watanabe. 2024c. Toward the evaluation of large language models considering score variance across instruction templates. In Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 499–529, Miami, Florida, US. Association for Computational Linguistics.

C. Spearman. 1904. The proof and measurement of association between two things. American Journal of Psychology, 15:88–103.

Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem 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, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. Preprint, arXiv:2307.09288.

R. Vedantam, C. Zitnick, and D. Parikh. 2015. Cider: Consensus-based image description evaluation. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4566–4575, Los Alamitos, CA, USA. IEEE Computer Society.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. LLaVA-critic: Learning to evaluate multimodal models.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. Preprint, arXiv:1909.03193.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. 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.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024a. MM-vet: Evaluating large multimodal models for integrated capabilities. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pages 57730–57754. PMLR.

Weihao Yu, Zhengyuan Yang, Linfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. 2024b. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. Preprint, arXiv:2408.00765.

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.

Prompt for Generating Reviews

```
You are a perceptive and insightful reviewer.  Your task is to
write five distinct review texts that discuss the strengths and
areas for improvement of the given image, while following the
constraints below:

Guidelines:
1.  Each review text should present unique content.
2.  Ensure that the length of each review is approximately
equal.
3.  Do not use bullet points or lists; maintain a cohesive
narrative.
4.  Write reviews 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 should address both the strengths and potential
areas for improvement of the image.
6.  If no improvements are necessary, explicitly state this
within the review.

Your reviews will contribute to research purposes only and
should reflect careful thought and analysis.
```

A.2 Prompt for Measuring Perplexity

We measured perplexity for each data using the following prompt.

Prefix Sentence

```
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., 2024), Qwen-VL-Chat (Bai et al., 2023a), and GPT-4 API ver. 0.28.0 (OpenAI, 2023b), using their default hyperparameters. Additionally, our dataset and code are available at `https://github.com/naist-nlp/Hackathon-2023-Summer`. For LLMs, we used models such as Llama2 (Touvron et al., 2023), 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

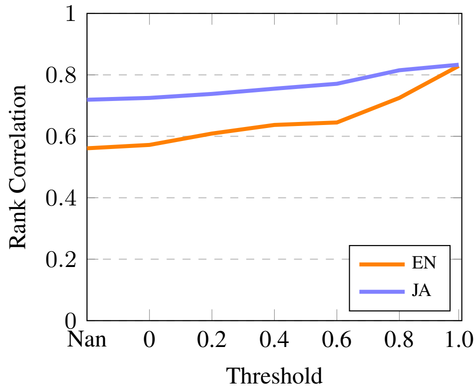

Figure 3: Correlation between prompt and human ranks.

D Details of Dataset

D.1 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. Figure 3 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

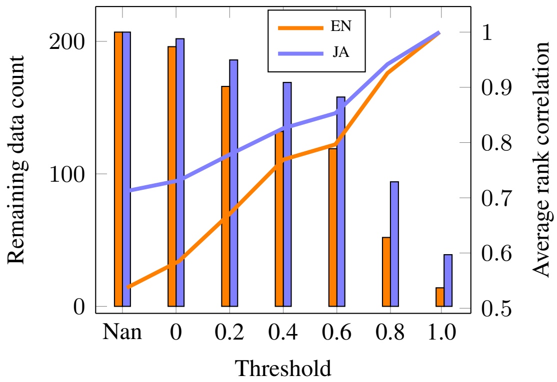

Figure 4: Changes for remaining data count and average rank correlation when varying threshold. The bar graphs represent the remaining data count and the line graphs denote average rank correlation. Nan means no threshold.

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.

D.2 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.3 Dataset Construction & Filtering

We collected 207 images from 15 different genres, including Animals and Artwork (see Appendix 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, 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). The results are shown in Figure 4. 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 4) 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**: The 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.