

EXPERT: An Explainable Image Captioning Evaluation Metric with Structured Explanations

Hyunjong Kim¹ Sangyeop Kim^{1,2} Jongheon Jeong¹
Yeongjae Cho¹ Sungzoon Cho^{1†}

¹Seoul National University ²Coxwave

{hjkim0811, pdg01117, ww0713, zxc5932, zoon}@snu.ac.kr

Abstract

Recent advances in large language models and vision-language models have led to growing interest in explainable evaluation metrics for image captioning. However, these metrics generate explanations without standardized criteria, and the overall quality of the generated explanations remains unverified. In this paper, we propose EXPERT, a reference-free evaluation metric that provides structured explanations based on three fundamental criteria: fluency, relevance, and descriptiveness. By constructing large-scale datasets of high-quality structured explanations, we develop a two-stage evaluation template to effectively supervise a vision-language model for both scoring and explanation generation. EXPERT achieves state-of-the-art results on benchmark datasets while providing significantly higher-quality explanations than existing metrics, as validated through comprehensive human evaluation. Our code and datasets are available at <https://github.com/hjkim811/EXPERT>.

1 Introduction

Automatic evaluation of image captions is crucial for measuring and improving image captioning models without the significant cost and time required for human evaluation (Ghandi et al., 2023). Along with the advancement of large language models (LLMs) (Jiang et al., 2023; Touvron et al., 2023; Dubey et al., 2024; Yang et al., 2024; Abdin et al., 2024) and vision-language models (VLMs) (Radford et al., 2021; Li et al., 2022; Liu et al., 2023, 2024), explainable evaluation metrics have recently drawn increasing attention in image captioning (Chan et al., 2023; Lee et al., 2024).

Explainable metrics not only provide numerical scores but also textual explanations, enhancing the interpretability and transparency of the evaluation. However, existing research on explainable metrics

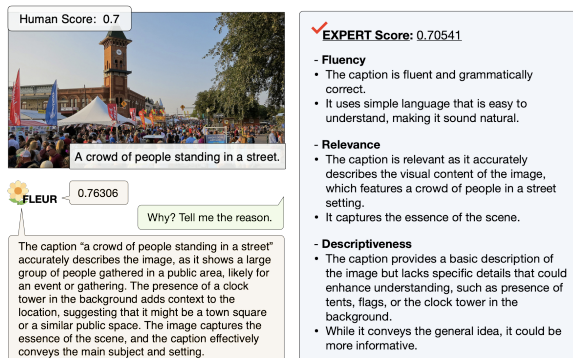


Figure 1: Example of a structured explanation.

has two limitations. First, the explanations provided by existing metrics are not generated based on standardized criteria or guidelines, which may lead to inconsistency in content and structure. Second, previous studies lack a thorough evaluation of the quality of generated explanations, leaving the overall quality unverified.

To address these issues, we propose EXPERT¹, a VLM-based reference-free evaluation metric for image captioning that provides structured explanations. We first construct large-scale datasets containing over 42,000 structured explanations by extending existing human judgment datasets. Each explanation is structured according to three fundamental criteria—fluency, relevance, and descriptiveness—and its quality is validated through human evaluation. We then train EXPERT using a two-stage evaluation template that effectively guides both scoring and explanation generation. Unlike existing explainable metrics, EXPERT generates structured, criterion-specific explanations, as illustrated in Figure 1.

Experimental results on benchmark human judgment datasets show that EXPERT establishes new state-of-the-art performance. Through comprehensive human evaluation, we further demonstrate that

[†]Corresponding author.

¹An EXPLAINABLE Image Captioning Evaluation Metric with STRUCTURED Explanations.

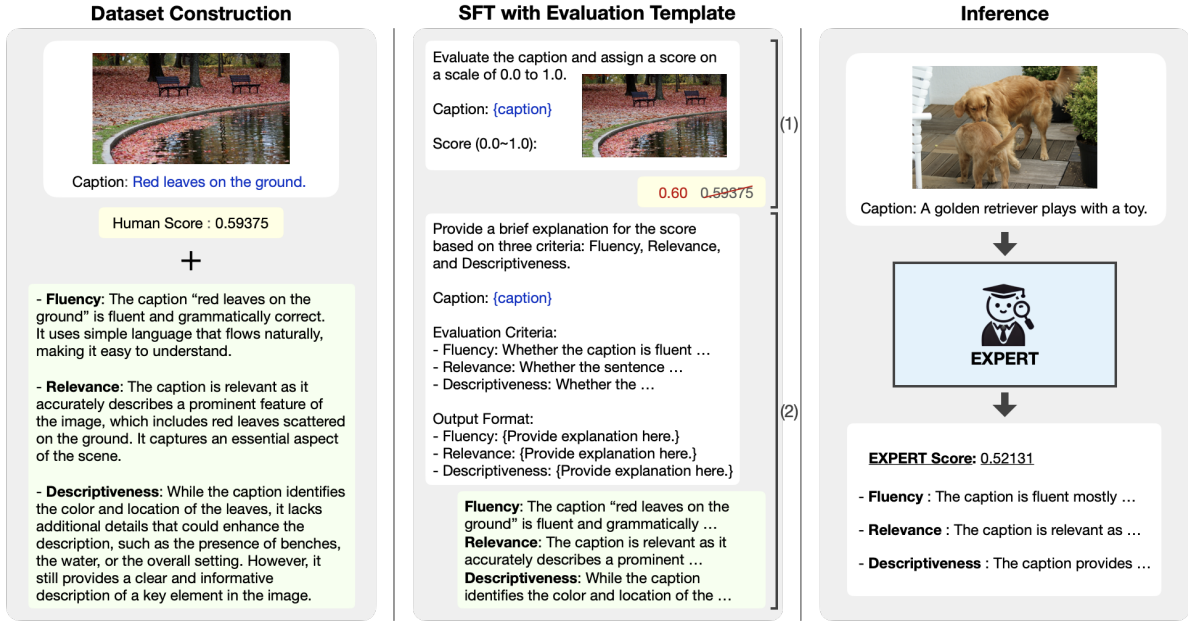


Figure 2: Overall framework of EXPERT.

EXPERT generates significantly higher-quality explanations than existing explainable metrics. To the best of our knowledge, we are the first to systematically assess the explanation capabilities of explainable image captioning evaluation metrics.

2 EXPERT

2.1 Dataset Construction

Generating Explanations Existing image captioning datasets consist of image-caption pairs with the corresponding human judgment scores. However, these datasets lack explanations for the scores, limiting research on the explainability of image captioning evaluation metrics. To overcome this issue, we construct *Polaris-exp* and *Nebula-exp* datasets by extending the *Polaris* dataset (Wada et al., 2024) and *Nebula* dataset (Matsuda et al., 2024), respectively. We add an *explanation* to each image-caption pair, with each explanation structured according to three dimensions: (1) **Fluency** evaluates whether the caption is fluent, natural, and grammatically correct. (2) **Relevance** evaluates whether the sentence correctly describes the visual content and is closely relevant to the image. (3) **Descriptiveness** evaluates whether the sentence is a precise, informative caption that describes important details of the image. These dimensions are the common criteria used for human annotation in the development of recent human judgment datasets (Lee et al., 2021; Wada et al., 2024; Matsuda et al., 2024). We generate explanations by prompting

GPT-4o (OpenAI, 2024) with the prompt presented in Appendix A and the corresponding image. Consequently, *Polaris-exp* and *Nebula-exp* contain 16,014 explanations and 26,152 explanations for unique image-caption pairs, respectively.

Dataset Evaluation To assess the quality of the generated explanations, we conduct human evaluation on 100 sampled instances from the combined set of *Polaris-exp* and *Nebula-exp*. To ensure balanced representation across different quality levels, we sample uniformly from three score intervals—[0, 0.33), [0.33, 0.66), and [0.66, 1]—with random selection within each interval. We recruited four annotators native in English to evaluate the explanation quality, particularly based on three criteria: (1) **Consistency** measures how logically consistent the explanation is with the given score. (2) **Factuality** measures how factually accurate the explanation is in describing the given image-caption pair. (3) **Informativeness** measures how much relevant detail and meaningful information the explanation provides to support and justify the given score. Each annotator was instructed to evaluate all 100 explanations based on these criteria using a 4-point Likert scale. Scoring guidelines and the evaluation form are provided in Appendix C.

The evaluation results in Table 1 indicate that the explanations in *Polaris-exp* and *Nebula-exp* exhibit high quality while maintaining strong alignment with the original dataset components.

Criteria	Average	Std. Dev.
Consistency	3.72	0.52
Factuality	3.84	0.39
Informativeness	3.72	0.45

Table 1: Human evaluation results for generated explanations in Polaris-exp and Nebula-exp. Std. Dev. represents the overall standard deviation across all ratings for each criterion.

2.2 Two-stage Evaluation Template

Based on the constructed Polaris-exp and Nebula-exp datasets, we design a *two-stage evaluation template* to effectively supervise a general-purpose VLM with human preferences and high-quality structured explanations. As illustrated in Figure 2, the template follows a scoring-explanation order, which has been shown to be effective in previous studies (Chan et al., 2023; Lee et al., 2024).

The first *scoring stage* contains a query to assign a score to the given image-caption pair, followed by a response which is a human score from the constructed datasets. At this stage, we apply *score binning*, an operation that rounds scores to the nearest multiple of a specified bin size, for simplified numerical representation (see Appendix D for details). We use a bin size of 0.10 for our experiments.

The second *explanation stage* contains a query to provide a brief explanation for the score based on three criteria—*fluency*, *relevance*, and *descriptiveness*—along with a description for each criterion and a predefined output format. The following response is the corresponding explanation from the constructed datasets. For each criterion, we adopt the same descriptions used in Section 2.1 to maintain consistency. The full text of the two-stage evaluation template is available in Appendix B.

2.3 Supervised Fine-tuning

We use LLaVA-1.5 (13B) (Liu et al., 2024) as our base model for supervised fine-tuning (SFT). The combined training splits of Polaris-exp and Nebula-exp are converted into two-stage evaluation templates for training. Since Polaris contains duplicate image-caption pairs with scores from different annotators, we group these into single instances by averaging the human scores. Additionally, as Polaris and Nebula share overlapping image-caption pairs across the two datasets, we merge these cross-dataset duplicates into single instances by averaging the human scores. Implementation details for SFT are provided in Appendix F.

2.4 Inference

Given an image-caption pair, we sequentially prompt the trained model with the queries of the two-stage evaluation template. We use greedy decoding to ensure deterministic and reproducible results, and apply score smoothing (Lee et al., 2024) in the scoring stage to obtain more detailed scores. Score smoothing involves computing the probability $p(i, j)$ for each digit i ($0 \leq i \leq 9$) being generated at the j -th decimal place ($j = 1, 2$), which is then used to produce the final score s :

$$s = \sum_{j=1}^2 10^{-j} \sum_{i=0}^9 i \times p(i, j).$$

3 Experiments

3.1 Evaluation Settings

We evaluate EXPERT on benchmark human judgment datasets: Flickr8k-EX, Flickr8k-CF (Hodosh et al., 2013), COMPOSITE (Aditya et al., 2015), Polaris (Wada et al., 2024), Nebula (Matsuda et al., 2024), and Pascal-50S (Vedantam et al., 2015). For Polaris and Nebula, we use the test split to ensure no overlap with the training data. Following previous studies, we use Kendall-B (τ_b) for Flickr8k-CF, accuracy for Pascal-50S, and Kendall-C (τ_c) for the other datasets. Further details of these datasets are provided in Appendix I.

We compare the performance of EXPERT with various reference-free and reference-based evaluation metrics, including recently proposed methods such as FLEUR (Lee et al., 2024), HICE-S (Zeng et al., 2024), and DENEb (Matsuda et al., 2024). Detailed descriptions of the baseline metrics are provided in Appendix J. We also include a comparison with an LLM-as-a-judge approach by directly prompting GPT-4o (OpenAI, 2024), a representative proprietary LLM. Experimental details for GPT-4o are provided in Appendix E.1.

3.2 Correlation with Human Judgments

The evaluation results of our experiments are shown in Table 2. EXPERT achieves state-of-the-art results on all benchmark datasets among reference-free metrics, except for Pascal-50S. On Pascal-50S HC, MM, and Avg, HICE-S attains higher accuracy, while EXPERT still achieves the second-best results. Notably, EXPERT even outperforms existing reference-based metrics on Flickr8k-CF, COMPOSITE, Polaris, and Nebula.

Type	Metric	Flickr8k		COM	Polaris	Nebula	Pascal-50S (Accuracy \uparrow)					
		EX ($\tau_c \uparrow$)	CF ($\tau_b \uparrow$)	($\tau_c \uparrow$)	($\tau_c \uparrow$)	($\tau_c \uparrow$)	HC	HI	HM	MM	Avg	
Reference-based	BLEU-4	30.8	16.9	30.6	46.3	40.4	53.0	92.4	86.7	59.4	72.9	
	ROUGE-L	32.3	19.9	32.4	46.3	42.6	51.5	94.5	92.5	57.7	74.1	
	METEOR	41.8	22.2	38.9	51.2	46.8	56.7	97.6	94.2	63.4	78.0	
	CIDEr	43.9	24.6	37.7	52.1	48.1	53.0	98.0	91.5	64.5	76.8	
	SPICE	44.9	24.4	40.3	51.0	44.0	52.6	93.9	83.6	48.1	69.6	
	BERTScore	39.2	22.8	30.1	51.6	47.0	65.4	96.2	93.3	61.4	79.1	
	CLAIR	48.3	34.4	<u>61.0</u>	53.3	52.7	52.4	99.5	89.8	73.0	78.7	
	TIGer	49.3	–	45.4	–	–	56.0	99.8	92.8	74.2	80.7	
	ViBERTScore-F	50.1	–	52.4	–	–	49.9	99.6	93.1	75.8	79.6	
	RefCLIPScore	53.0	36.4	55.4	52.3	46.9	64.5	99.6	95.4	72.8	83.1	
	RefPAC-S	55.9	37.6	57.3	56.0	50.6	67.7	99.6	96.0	75.6	84.7	
	Polos	56.4	37.8	57.6	57.8	53.9	70.0	99.6	97.4	<u>79.0</u>	86.5	
	RefFLEUR	51.9	38.8	64.2	58.8	54.4	68.0	99.8	98.0	76.1	85.5	
	RefHICE-S	57.7	<u>38.2</u>	58.7	–	–	<u>71.4</u>	<u>99.7</u>	<u>97.7</u>	79.7	87.3	
	DENEb	<u>56.5</u>	38.0	57.9	<u>58.2</u>	<u>54.1</u>	74.4	99.8	97.3	76.5	<u>87.0</u>	
	Reference-free	CLIPScore	51.2	34.4	53.8	52.3	46.9	56.5	99.3	96.4	70.4	80.7
		PAC-S	54.3	36.0	55.7	52.5	47.2	60.6	99.3	96.9	72.9	82.4
InfoMetIC+ ²		55.5	36.6	59.3	–	–	–	–	–	–	–	
FLEUR		53.0	<u>38.6</u>	<u>63.5</u>	<u>58.3</u>	<u>51.7</u>	61.3	<u>99.7</u>	<u>97.6</u>	74.2	83.2	
BRIDGE		54.8	36.1	55.0	–	–	59.4	99.4	97.5	74.0	82.6	
HICE-S		<u>56.4</u>	37.2	57.9	–	–	68.6	<u>99.7</u>	96.5	79.5	86.1	
EXPERT		56.7	39.3	65.0	61.1	54.9	<u>62.8</u>	99.8	97.8	<u>78.4</u>	<u>84.7</u>	
LLM-as-a-judge	GPT-4o	54.3	39.3	65.9	58.2	54.3	60.8	99.7	97.3	73.7	82.9	

Table 2: Comparison of correlation coefficients and accuracy on various human judgment datasets. **Bold** and underlined values indicate the best and second-best results for each type, respectively. *Italicized* values indicate results reproduced or implemented in this work, while the others are reported from previous works. The ‘–’ symbol indicates unreproducible values due to non-executable code or unavailable data. In the reference-based section, the upper block contains metrics that rely solely on reference captions, while the lower block contains metrics that use both the image and the reference caption.

When compared to GPT-4o, EXPERT achieves equal or superior results across all datasets except for COMPOSITE. GPT-4o exhibits relatively lower performance despite its substantially larger model size, which we attribute to two main factors: tokenizer differences and limited access to token probabilities. A detailed discussion of these issues is provided in Appendix E.2.

These results demonstrate the effectiveness of SFT with the proposed evaluation template. The superior performance of EXPERT on datasets other than Polaris and Nebula suggests that human preferences across different datasets and annotators exhibit a certain degree of consistency, making it a robust and generalizable evaluation metric.

3.3 Evaluation of Explanation Quality

Qualitative Example Figure 3 presents examples of explanations generated by EXPERT and FLEUR. In the first example, both metrics correctly identify that the image contains only one

²We do not include the Pascal-50S results of InfoMetIC+ as Lee et al. (2024) cast doubt on the experimental settings. Despite our efforts to reproduce the results, the necessary resources for reproduction were inaccessible.

dog, not three. However, while FLEUR overlooks the omission of the frisbee in the caption, EXPERT explicitly points out the lack of detail regarding the dog’s action—chasing a frisbee. The second example demonstrates a caption with an incomplete sentence structure. While FLEUR misinterprets the caption as mentioning “a blue bed”, EXPERT accurately recognizes the grammatical incompleteness and provides appropriate explanations. Additional examples are provided in Appendix K.

Human Evaluation Through human evaluation, we compare the overall quality of explanations generated by EXPERT and FLEUR, the only two metrics that are both explainable and reference-free. We introduce an additional baseline, EXPERT_{w/o SFT}, which prompts LLaVA-1.5 (13B) with the queries of the two-stage evaluation template without SFT. For each metric, we evaluate explanations generated for the same 100 image-caption pairs sampled from Flickr8k-EX. To ensure diversity across different quality levels, we sample uniformly for each human score (1, 2, 3, and 4), with random selection within each score. Four native English-speaking annotators each evaluated

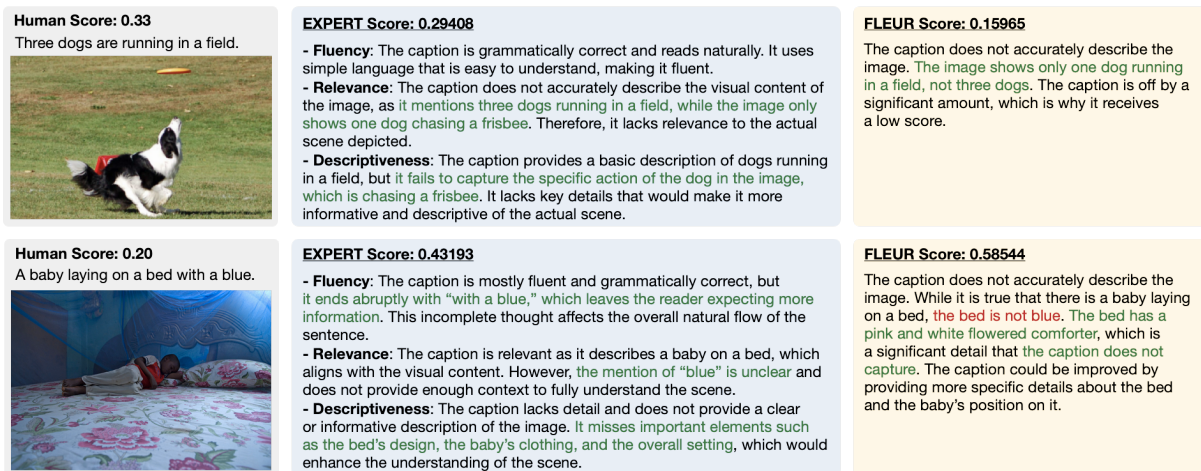


Figure 3: Examples of explanations generated by EXPERT and FLEUR. Green text indicates valid justifications for the given score, while red text indicates incorrect or misleading explanations.

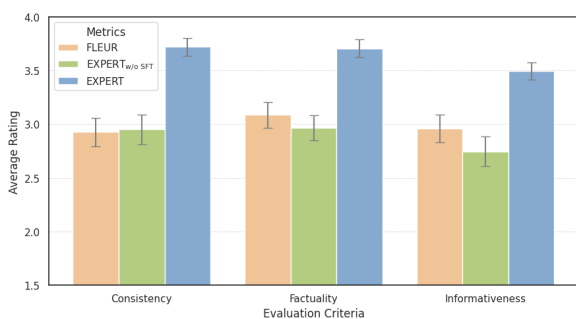


Figure 4: Human evaluation results on explanation quality. Error bars represent the 95% confidence intervals. Differences between EXPERT and the other metrics are statistically significant for all criteria at the 0.01 significance level.

the explanations for all three metrics across all 100 samples, based on the same criteria and scoring guidelines used in Section 2.1. Unlike the previous evaluation, the given scores in this evaluation are also generated by the metrics themselves. The evaluation form is provided in Appendix C.2.

The evaluation results are illustrated in Figure 4. EXPERT achieves the highest average rating across all criteria by a large margin, demonstrating that it provides the most logically consistent, factually accurate, and informative explanations for the scores. EXPERT_{w/o SFT}, however, demonstrates comparable or slightly lower performance than FLEUR. These results suggest that while the use of standardized and detailed criteria alone is insufficient to enhance explanation quality, it can lead to significant improvements when combined with supervision using high-quality explanations.

4 Conclusion

In this work, we propose EXPERT, a state-of-the-art reference-free evaluation metric that provides structured explanations based on standardized criteria. Along with our novel explanation datasets and analysis of explanation quality, we hope this work offers a valuable contribution to advancing the explainability of image captioning evaluation.

Limitations

Error Analysis Although EXPERT generally performs well in scoring and generating explanations, it sometimes fails to provide accurate assessments or explanations. We analyzed the 100 samples with the largest absolute difference between the EXPERT score and human score. The errors could be grouped into six main categories, where the most common type of error was the *overpenalization of captions lacking details*. Further details of the error analysis are provided in Appendix H.

Inference Time The inference times for EXPERT and FLEUR are shown in Table 8 of Appendix G. Since both metrics adopt a scoring-explanation order, it is possible to obtain only the scores for faster inference, represented as EXPERT_{score} and FLEUR_{score}. As the inference time of VLMs heavily depends on the number of output tokens, generating both the score and textual explanations requires significantly more time than generating only the score. Additionally, EXPERT takes longer than FLEUR due to its longer explanations. The issue of inference time in explainable metrics should be resolved for practical usability.

Ethics Statement

Human Evaluation In this work, we conduct two types of human evaluation: evaluation of Polaris-exp and Nebula-exp datasets, and evaluation of the explanations generated by explainable metrics. The same group of four annotators participated in both evaluations and were compensated at a rate of USD \$15 per hour. Our human evaluation qualifies as exempt from IRB approval according to U.S. federal regulation 45 CFR 46.104(d)(2)(i), as we did not collect any personally identifiable information and recorded the data in such a manner that the identity of participants could not be readily ascertained.

Acknowledgments

This work was supported by the BK21 FOUR Program (Education and Research Center for Industrial Innovation Analytics) funded by the Ministry of Education, Korea (No. 4120240214912) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2093785).

References

- Marah Abidin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermüller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.
- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.
- David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. 2023. **CLAIR: Evaluating image captions with large language models**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13638–13646, Singapore. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. **Deep learning approaches on image captioning: A review**. *ACM Comput. Surv.*, 56(3).
- Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. 2024. Regionopt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13806.
- 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.
- Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.
- Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. 2023. **InfoMetIC: An informative metric for reference-free image caption evaluation**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3171–3185, Toronto, Canada. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo 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.
- Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. **TIGER: Text-to-image grounding for image caption evaluation**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2141–2152, Hong Kong, China. Association for Computational Linguistics.
- Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation, StatMT '07*, page 228–231, USA. Association for Computational Linguistics.
- Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021. **UMIC: An**

- unreferenced metric for image captioning via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 220–226, Online. Association for Computational Linguistics.
- Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. **ViLBERTScore: Evaluating image caption using vision-and-language BERT**. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 34–39, Online. Association for Computational Linguistics.
- Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. **FLEUR: An explainable reference-free evaluation metric for image captioning using a large multimodal model**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3732–3746, Bangkok, Thailand. Association for Computational Linguistics.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- 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.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Kazuki Matsuda, Yuiga Wada, and Komei Sugiura. 2024. Deneb: A hallucination-robust automatic evaluation metric for image captioning. In *Proceedings of the Asian Conference on Computer Vision*, pages 3570–3586.
- OpenAI. 2024. **Gpt-4o system card**.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. **Collecting image annotations using Amazon’s Mechanical Turk**. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 139–147, Los Angeles. Association for Computational Linguistics.
- Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6914–6924.
- Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Bridge: Bridging gaps in image captioning evaluation with stronger visual cues. In *European Conference on Computer Vision*, pages 70–87. Springer.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13559–13568.
- Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim M Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. 2024. Locca: Visual pretraining with location-aware captioners. *Advances in Neural Information Processing Systems*, 37:116355–116387.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Zejun Zeng, Jianqiao Sun, Hao Zhang, Tiansheng Wen, Yudi Su, Yan Xie, Zhengjue Wang, and Bo Chen. 2024. Hicescore: A hierarchical metric for image captioning evaluation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 866–875.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. **Bertscore: Evaluating text generation with bert**. In *International Conference on Learning Representations*.

A Prompt for Generating Explanations

Your task is to evaluate the given caption based on three criteria: Fluency, Relevance, and Descriptiveness. Provide a brief explanation for each criterion without assigning a numerical score.

Caption: **{caption}**

Criterion 1: Fluency

- Whether the caption is fluent, natural, and grammatically correct.

Criterion 2: Relevance

- Whether the sentence correctly describes the visual content and is closely relevant to the image.

Criterion 3: Descriptiveness

- Whether the sentence is a precise, informative caption that describes important details of the image.
- If the caption includes the necessary key details to provide a clear and informative description of the image, it should be considered descriptive enough. In this case, do not say it lacks detail or it could be more informative.

Output format:

Fluency: {Provide explanation here.}

Relevance: {Provide explanation here.}

Descriptiveness: {Provide explanation here.}

In preliminary experiments, we found that GPT-4o is overly rigorous in assessing descriptiveness, generating critical feedback such as “it could be more informative” even for captions that received high human scores. To address this challenge, we enhanced the description for descriptiveness with additional specifications.

We also tried providing the human score in the prompt, but this often led to biased explanations, such as generating negative feedback across all three criteria for captions with low human scores, even when some criteria demonstrated high quality. Therefore, we excluded human scores from the prompt to ensure objective explanations.

B Two-stage Evaluation Template

B.1 Scoring Stage: Query

Evaluate the caption and assign a score on a scale of 0.0 to 1.0.

Caption: **{caption}**

Score (0.0~1.0):

B.2 Scoring Stage: Response

{score}

B.3 Explanation Stage: Query

Provide a brief explanation for the score based on three criteria: Fluency, Relevance, and Descriptiveness.

Caption: **{caption}**

Evaluation Criteria:

- Fluency: Whether the caption is fluent, natural, and grammatically correct.
- Relevance: Whether the sentence correctly describes the visual content and is closely relevant to the image.
- Descriptiveness: Whether the sentence is a precise, informative caption that describes important details of the image.

Output Format:

Fluency: {Provide explanation here.}

Relevance: {Provide explanation here.}

Descriptiveness: {Provide explanation here.}

B.4 Explanation Stage: Response

Fluency: **{explanation for fluency}**

Relevance: **{explanation for relevance}**

Descriptiveness: **{explanation for descriptiveness}**

C Human Evaluation

C.1 Evaluation Criteria and Scoring Guidelines

Table 3 shows the evaluation criteria and detailed scoring guidelines for each criterion. A 4-point scale was chosen over a 5-point scale to avoid neutral responses and encourage more decisive judgments. For informativeness, we added a ‘Disagreement Option’ to address cases where the score is severely misaligned with the image-caption pair, making it difficult to evaluate the explanation meaningfully. We intended to exclude these cases from result aggregation, but this option was not selected by any evaluator for any data sample.

Consistency

How logically consistent is the explanation with the score?

Consider only the explanation text and the numerical score, without referencing the image.

Scoring Guidelines:

- 1: The nuance and tone of the explanation are **completely misaligned** with the score, offering contradictory reasoning that fails to justify the score in any meaningful way.
- 2: The nuance and tone of the explanation are **partially aligned** with the score, with noticeable mismatches or gaps in logic that make the connection unclear.
- 3: The nuance and tone of the explanation are **mostly aligned** with the score, but minor inconsistencies or subtle mismatches slightly reduce its overall coherence.
- 4: The nuance and tone of the explanation are **fully aligned** with the score, leaving no doubt about the connection between the explanation and the score.

Factuality

How factually accurate is the explanation in describing the image-caption pair?

Scoring Guidelines:

- 1: The explanation contains **critical factual errors** or describes something **entirely unrelated or contradictory** to the image-caption pair.
- 2: The explanation is **partially factual**, with some accurate elements but significant incorrect details.
- 3: The explanation is **mostly factual**, with only minor factual inaccuracies that do not affect the overall context.
- 4: The explanation is **entirely factual**, containing only accurate and verifiable information consistent with the image-caption pair.

Informativeness

How much relevant detail and meaningful information does the explanation provide to support and justify the score?

Scoring Guidelines:

- **Disagreement Option:** I cannot understand or justify the given score for this image-caption pair. The score is severely misaligned, and I am unable to evaluate the explanation in a meaningful way.
- 1: The explanation is **not informative at all**, providing little to no relevant information to justify the score, or presenting incorrect or misleading justifications that confuse the evaluator.
- 2: The explanation is **partially informative**, offering some relevant information but lacking sufficient detail or depth, or providing unclear or slightly inaccurate justifications.
- 3: The explanation is **fairly informative**, providing useful information but missing important aspects or including some redundant or unnecessary information.
- 4: The explanation is **highly informative**, offering rich, relevant, and detailed reasoning that thoroughly supports the score.

Table 3: Scoring guidelines for consistency, factuality, and informativeness.

C.2 Evaluation Form

Human Evaluation Form

Hello and Welcome!

This evaluation consists of 100 data samples. Each data sample includes an **image** and a **caption**, along with a **score** and an **explanation** justifying the score. The explanations are divided into three aspects: **fluency**, **relevance**, and **descriptiveness**. Please evaluate them holistically, considering all three together.

Your task is **NOT to evaluate the given score itself**. Instead, you should assess the **quality of the explanation that justifies the score**. Specifically, you will evaluate each explanation based on the following three criteria:

Evaluation Criteria

1. Consistency
How logically consistent is the explanation with the given score? Consider only the explanation text and the numerical score, without referencing the image.

Score Descriptions:

- 4: The nuance and tone of the explanation are **fully aligned** with the score, leaving no doubt about the connection between the explanation and the score.
- 3: The nuance and tone of the explanation are **mostly aligned** with the score, but minor inconsistencies or subtle mismatches slightly reduce its overall coherence.
- 2: The nuance and tone of the explanation are **partially aligned** with the score, with noticeable mismatches or gaps in logic that make the connection unclear.
- 1: The nuance and tone of the explanation are **completely misaligned** with the score, offering contradictory reasoning that fails to justify the score in any meaningful way.

2. Factuality
How factually accurate is the explanation in describing the image-caption pair?

Score Descriptions:


- 4: The explanation is **entirely factual**, containing only accurate and verifiable information consistent with the image-caption pair.
- 3: The explanation is **mostly factual**, with only minor factual inaccuracies that do not affect the overall context.
- 2: The explanation is **partially factual**, with some accurate elements but significant incorrect details.
- 1: The explanation contains **critical factual errors** or describes something **entirely unrelated or contradictory** to the image-caption pair.

3. Informativeness
How detailed and meaningful is the explanation in justifying the score?

Score Descriptions:

- 4: The explanation is **highly informative**, offering rich, relevant, and detailed reasoning that thoroughly supports the score.
- 3: The explanation is **fairly informative**, providing useful information but missing important aspects or including some redundant or unnecessary information.
- 2: The explanation is **partially informative**, offering some relevant information but lacking sufficient detail or depth, or providing unclear or slightly inaccurate justifications.
- 1: The explanation is **not informative at all**, providing little to no relevant information to justify the score, or presenting incorrect or misleading justifications that confuse the evaluator.
- 0 (**Disagreement Option**): I cannot understand or justify the given score for this

Image 1.



Caption 1: A wall on the side of a building.

Score: 0.57143

Explanation:

- Fluency: The caption is fluent and grammatically correct. It uses simple language that is easy to understand, making it sound natural.
- Relevance: The caption is relevant as it accurately describes a key element of the image, which is indeed a wall on the side of a building. It captures the main focus of the visual content.
- Descriptiveness: The caption is somewhat descriptive but lacks detail about the wall's characteristics, such as its color, texture, or any surrounding context. While it identifies the wall, it does not provide additional information that could enhance the viewer's understanding of the scene.

Consistency *

	1	2	3	4
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Factuality *

	1	2	3	4
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Informativeness *

	0	1	2	3	4
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 5: Evaluation form for human evaluation of Polaris-exp and Nebula-exp datasets.

Human Evaluation Form

Hello and Welcome!

This evaluation consists of 100 data samples. Each data sample includes an **image** and a **caption**, along with a **score** and an **explanation** justifying the score. For each data sample, you will see three sets of **score-explanation** (from Evaluator 1, Evaluator 2, and Evaluator 3).

Some explanations are written as a single paragraph, while others are divided into three aspects: **fluency**, **relevance**, and **descriptiveness**. For explanations that are broken down into these aspects, please evaluate them holistically, considering all three together.

Your task is **NOT to evaluate the given score itself**. Instead, you should assess the **quality of the explanation that justifies the score**. Specifically, you will evaluate each explanation based on the following three criteria:

Evaluation Criteria

1. Consistency
How logically consistent is the explanation with the given score? Consider only the explanation text and the numerical score, without referencing the image.

Score Descriptions:

- 4: The nuance and tone of the explanation are **fully aligned** with the score, leaving no doubt about the connection between the explanation and the score.
- 3: The nuance and tone of the explanation are **mostly aligned** with the score, but minor inconsistencies or subtle mismatches slightly reduce its overall coherence.
- 2: The nuance and tone of the explanation are **partially aligned** with the score, with noticeable mismatches or gaps in logic that make the connection unclear.
- 1: The nuance and tone of the explanation are **completely misaligned** with the score, offering contradictory reasoning that fails to justify the score in any meaningful way.

2. Factuality
How factually accurate is the explanation in describing the image-caption pair?

Score Descriptions:


- 4: The explanation is **entirely factual**, containing only accurate and verifiable information consistent with the image-caption pair.
- 3: The explanation is **mostly factual**, with only minor factual inaccuracies that do not affect the overall context.
- 2: The explanation is **partially factual**, with some accurate elements but significant incorrect details.
- 1: The explanation contains **critical factual errors** or describes something **entirely unrelated or contradictory** to the image-caption pair.

3. Informativeness
How detailed and meaningful is the explanation in justifying the score?

Score Descriptions:

- 4: The explanation is **highly informative**, offering rich, relevant, and detailed reasoning that thoroughly supports the score.
- 3: The explanation is **fairly informative**, providing useful information but missing important aspects or including some redundant or unnecessary information.
- 2: The explanation is **partially informative**, offering some relevant information but lacking sufficient detail or depth, or providing unclear or slightly inaccurate justifications.
- 1: The explanation is **not informative at all**, providing little to no relevant information to justify the score, or presenting incorrect or misleading justifications that confuse the evaluator.
- 0 (**Disagree**): I cannot understand or justify the given score for this

Image 1.



Caption 1: A dog running with a blue toy in its mouth.

[Evaluator 1] Score: 0.04079
Explanation:
The caption does not accurately describe the image. The image shows a young boy wearing a Hawaiian shirt and flip-flops, jumping in the air with his arms up. There is no dog or blue toy in the image.

[Evaluator 2] Score: 0.07280
Explanation:
- Fluency: The caption is fluent and grammatically correct.
- Relevance: The caption is not relevant to the image, as it describes a dog running with a toy in its mouth, while the image shows a young boy in a Hawaiian shirt jumping in the air.
- Descriptiveness: The caption is not descriptive, as it does not provide any specific details about the boy's actions or the surrounding environment.

[Evaluator 3] Score: 0.00392
Explanation:
- Fluency: The caption is fluent and grammatically correct. It uses simple language that flows naturally, making it easy to read and understand.
- Relevance: The caption is not relevant to the image provided. It describes a dog running with a toy, while the image depicts a child in a Hawaiian shirt and flip-flops. There is a clear disconnect between the caption and the visual content.
- Descriptiveness: The caption lacks descriptiveness in relation to the image. It does not provide any details about the child, their clothing, or the setting, which are key elements that would help convey a clearer understanding of the scene.

Consistency *

	1	2	3	4
Evaluator 1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Evaluator 2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Evaluator 3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Factuality *

	1	2	3	4
Evaluator 1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Evaluator 2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Evaluator 3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Informativeness *

	0 (Disagree)	1	2	3	4
Evaluator 1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Evaluator 2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Evaluator 3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 6: Evaluation form for human evaluation of explanations of FLEUR, EXPERT_{w/o SFT}, and EXPERT.

D Score Binning

As VLMs recognize each digit of a score as a separate token, not a continuous value as a whole, trivial differences in numerical values (e.g., 0.59375 vs. 0.60) may introduce unnecessary complexity, thereby degrading learning efficiency. To address this issue, we introduce score binning to guide the model to learn human preferences more effectively. Human scores in human judgment datasets often include lengthy values due to averaging the scores from multiple annotators. Therefore, we intentionally reduce the granularity of these scores by applying score binning.

Table 4 presents the evaluation results on benchmark datasets with and without score binning, demonstrating that score binning consistently improves performance across datasets.

Metric	Flickr8k		COM	Polaris	Nebula	Pascal-50S (Accuracy \uparrow)				
	EX ($\tau_c \uparrow$)	CF ($\tau_b \uparrow$)	($\tau_c \uparrow$)	($\tau_c \uparrow$)	($\tau_c \uparrow$)	HC	HI	HM	MM	Avg
EXPERT	56.7	39.3	65.0	61.1	54.9	62.8	99.8	97.8	78.4	84.7
w/o Score Binning	56.2	39.1	65.0	60.8	54.7	61.1	99.7	97.7	77.5	84.0

Table 4: Effect of score binning on performance in human judgment datasets.

E LLM-as-a-Judge with GPT-4o

E.1 Experimental Setup

In our experiments, we include an LLM-as-a-judge baseline using GPT-4o. We directly prompt GPT-4o with the prompt shown below and the corresponding image, using the `gpt-4o-2024-08-06` version with the temperature set to 0. While we closely follow the prompt format used in the scoring stage of EXPERT, we adopt a scale of 0 to 10 instead of 0 to 1 due to output instability. Because of the issues described below, using a scale of 0 to 1 often leads to inappropriate outputs for score smoothing. We also provide grading criteria for the scores in the prompt, which have been shown to be effective in Lee et al. (2024).

Evaluate the caption and assign a score on a scale of 0 to 10.

Grading Criteria:

- 0: The caption does not describe the image at all or is completely incorrect.
- 2: The caption contains minimal accurate details but misses most key elements.
- 4: The caption describes some elements correctly but has significant missing or incorrect information.
- 6: The caption describes the main elements correctly but lacks some details or clarity.
- 8: The caption describes the image well with minor omissions or imprecisions.
- 10: The caption accurately and clearly describes all key elements of the image.

Caption: **{caption}**

Please only provide the score, nothing else.

Score:

E.2 Experimental Results and Analysis

Table 5 presents the experimental results of FLEUR, EXPERT, and GPT-4o *before score smoothing* and *after score smoothing*. Note that the results after score smoothing are the final results reported in Table 2.

Setting	Metric	Flickr8k		COM	Polaris	Nebula	Pascal-50S (Accuracy \uparrow)				
		EX ($\tau_c \uparrow$)	CF ($\tau_b \uparrow$)	($\tau_c \uparrow$)	($\tau_c \uparrow$)	($\tau_c \uparrow$)	HC	HI	HM	MM	Avg
Before score smoothing	FLEUR	34.9	51.6	58.9	53.5	52.3	26.0	99.3	92.4	42.7	65.1
	EXPERT	35.7	52.0	58.4	54.0	57.2	26.8	99.3	93.4	45.3	66.2
	GPT-4o	40.4	54.4	62.9	56.1	61.5	31.5	99.4	95.1	48.0	68.5
After score smoothing	FLEUR	53.0	38.6	63.5	58.3	51.7	61.3	99.7	97.6	74.2	83.2
	EXPERT	56.7	39.3	65.0	61.1	54.9	62.8	99.8	97.8	78.4	84.7
	GPT-4o	54.3	39.3	65.9	58.2	54.3	60.8	99.7	97.3	73.7	82.9

Table 5: Experimental results of FLEUR, EXPERT, and GPT-4o before score smoothing and after score smoothing. **Bold** values indicate the best results for each setting.

Before score smoothing, GPT-4o achieves the highest performance across all datasets. After score smoothing, however, its performance gains are relatively modest compared to the other models, ultimately leading to inferior results compared to EXPERT. We attribute this to two main factors:

- **Tokenizer Differences:** Since the tokenizer of GPT-4o often outputs multi-digit numbers as a single token, values such as ‘1’, ‘5’, ‘15’, and ‘02’ can all appear as candidate tokens for the same position. This may lead to a more complex and potentially distorted probability distribution for score smoothing, which expects a single-digit number for one token.
- **Limited Access to Token Probabilities:** GPT-4o currently provides log probabilities for only the top 20 candidates for each token. We frequently encountered cases where not all digits (0-9) were included, making it impossible to apply score smoothing properly.

Due to these constraints, our implementation had to rely only on the available and valid digits, leading to suboptimal performance. Moreover, using closed models like GPT-4o as evaluation metrics involves the following limitations:

- **Substantial Cost:** The token-based pricing of proprietary models incurs substantial costs. Table 6 presents the actual expenses from our experiments, where approximately \$106 was spent on the evaluation across all benchmark datasets.
- **Consistency Over Time:** Updates to closed models can pose challenges to consistency and reproducibility, both of which are essential qualities for a reliable evaluation metric.

In summary, our findings suggest that GPT-4o, despite its strong capabilities across various tasks, does not provide clear benefits over EXPERT as an image captioning evaluation metric.

Metric	Flickr8k		COM	Polaris	Nebula	Pascal-50S				Total
	EX	CF				HC	HI	HM	MM	
GPT-4o	\$6.64	\$56.10	\$15.69	\$14.05	\$5.17	\$2.11	\$2.11	\$2.12	\$2.11	\$106.10

Table 6: Summary of GPT-4o evaluation costs across benchmark datasets.

F Implementation Details

The configurations for training EXPERT are shown in Table 7. We apply LoRA (Hu et al., 2021) for SFT. Two NVIDIA A100 GPUs with 40GB memory were used for SFT, which took approximately 2 hours.

Configuration	Setting
Epochs	1
Batch Size	8
Optimizer	AdamW
Learning Rate	2e-05
Weight Decay	0.0
Warmup Ratio	0.03
Learning Rate Scheduler	Cosine
LoRA Rank	128
LoRA Alpha	256

Table 7: Training configurations for SFT.

G Inference Time

Table 8 presents the inference times for EXPERT and FLEUR on a single NVIDIA A100 GPU.

Metric	Inference Time (sec)
FLEUR _{score}	0.32
FLEUR	2.76
EXPERT _{score}	0.36
EXPERT	3.80

Table 8: Inference time of explainable metrics.

H Error Analysis

We conducted a detailed error analysis on the 100 samples with the largest absolute difference between the EXPERT score and human score. The errors could be grouped into six main categories:

Error Type	# Error
Overpenalization of Captions Lacking Details	45
Overestimation of Captions with Incorrect Details	17
Overestimation of Captions Lacking Details	13
Misjudgment of Captions with Grammatical Errors	10
Human Annotation Errors	8
Others	7
Total	100

Table 9: Categorization of error cases.

- 1. Overpenalization of Captions Lacking Details:** Instances where EXPERT assigned excessively low scores to captions that, while generally accurate, lacked sufficient details.
- 2. Overestimation of Captions with Incorrect Details:** Instances where EXPERT assigned high scores to captions containing incorrect or misleading content.
- 3. Overestimation of Captions Lacking Details:** Instances where EXPERT assigned high scores to captions that omitted important details.
- 4. Misjudgment of Captions with Grammatical Errors:** Instances where captions contained grammatical errors, and EXPERT failed to assign appropriate scores.
- 5. Human Annotation Errors:** Instances where human scores were found to be inaccurate—either higher or lower than what could reasonably be expected.
- 6. Others:** Errors that do not fall into any of the categories above.

As shown in Table 9, the most common type of error was the overpenalization of captions that lacked sufficient details. A possible solution for this could be to oversample concise captions that received high human scores during training, helping the model better align with human judgment and reduce its bias toward detail-heavy captions.

Furthermore, we analyzed the instances used in the human evaluation in Section 3.3 that received a score of 2 or below in at least one criterion. Approximately 35% of these were associated with the *overpenalization of captions lacking details*, where the explanations failed to provide a valid justification for the overpenalization. In the remaining cases, the explanations contained factual inaccuracies regarding elements mentioned in the captions that occupy only a small area in the image, which in turn diminished the informativeness of the explanations. Such difficulty in capturing fine-grained, region-specific details in images is a well-known and fundamental challenge for vision transformers and vision-language models, where active research is underway to address this limitation (Guo et al., 2024; Wan et al., 2024).

I Benchmark Datasets

- **Flickr8k** (Hodosh et al., 2013) contains 8,092 images, each paired with approximately five reference captions. The dataset includes two evaluation sets: Flickr8k-Expert (5,664 image-caption pairs evaluated by three experts on a 1-4 scale) and Flickr8k-CF (47,830 pairs assessed by three crowdworkers with binary yes/no judgments, with final scores calculated as the proportion of ‘yes’ responses). Both evaluation sets use 1,000 unique images, with 158 direct reference candidates excluded from the analysis to maintain evaluation integrity.
- **COMPOSITE** (Aditya et al., 2015) is a multi-source evaluation dataset containing 3,995 images drawn from three collections: MSCOCO (2,007), Flickr8k (997), and Flickr30k (991). It features 11,985 human-evaluated caption pairs, where each image has three candidate captions (one human-authored and two machine-generated) alongside five reference captions. Evaluators rated the caption-image correspondence on a 1-5 scale, from completely irrelevant to perfectly matching. This combination of human and machine-generated captions provides a robust benchmark for comparing automated captioning systems against human performance.
- **Polaris** (Wada et al., 2024) represents a significant advancement in image-caption evaluation, incorporating 131,020 human judgments. It features outputs from ten contemporary captioning models, including SAT, M2-Transformer, VinVL, GRIT, BLIPbase, BLIPLarge, GIT, OFA, and two BLIP-2 variants (flan and opt). The dataset draws from both MSCOCO and nocaps collections, ensuring caption diversity. Evaluators rated captions on a five-point scale, considering descriptiveness, fluency, and relevance. The final scores were normalized to [0,1], with careful filtering of unreliable evaluations based on response patterns and timing. Polaris is licensed under the Clear BSD License.
- **Nebula** (Matsuda et al., 2024) expands upon Polaris’s foundation, tripling the image count to 32,978. It contains 183,472 reference captions and gathered judgments from 805 evaluators. The dataset’s vocabulary spans 32,870 unique words across reference captions (totaling 1,945,956 words) and 3,695 words in candidate captions (288,922 total words). Reference captions average 10.61 words in length, while candidate captions average 8.76 words. To prevent training data leakage, images were sourced from MSCOCO and nocaps validation sets. Evaluators assessed captions using a five-point scale, considering the same criteria as Polaris: fluency, relevance, and descriptiveness. Nebula is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License.
- **Pascal-50S** (Vedantam et al., 2015) draws from the UIUC PASCAL Sentence Dataset (Rashtchian et al., 2010), comprising 1,000 images - 950 with 50 reference captions each and 50 with 120 captions. The dataset features 4,000 evaluation triplets where 48 evaluators compare two candidate captions against the references to determine which better describes the image. The candidate pairs follow four patterns: both correct human-written captions (HC), correct and incorrect human captions (HI), human versus machine-generated captions (HM), and both correct machine-generated captions (MM).

J Baseline Metrics

- **BLEU** (Papineni et al., 2002) measures translation quality by comparing n-grams between candidate and reference sentences using modified precision with a brevity penalty.
- **ROUGE-L** (Lin, 2004) evaluates text similarity by finding the Longest Common Subsequence between a candidate and reference sentences, computing recall and precision.
- **METEOR** (Lavie and Agarwal, 2007) evaluates text similarity by aligning words using exact matches, stems, synonyms, and paraphrases, then computes a recall-weighted F-score.
- **CIDEr** (Vedantam et al., 2015) assesses text similarity by using the TF-IDF weighted n-gram similarity between candidate and reference texts, focusing on agreement across multi-references.
- **SPICE** (Anderson et al., 2016) evaluates image captions by converting both candidate and reference captions into scene graphs and comparing their semantic content through scene graph overlap comparison.
- **BERTScore** (Zhang et al., 2020) computes token-level similarity between candidate and reference texts using contextual embeddings from BERT, capturing semantic meaning more effectively than n-gram based metrics.
- **CLAIR** (Chan et al., 2023) evaluates image captions using LLMs to score and explain caption quality by comparing candidate and reference captions that match human judgments.
- **TIGER** (Jiang et al., 2019) evaluates image captions using a stacked cross-attention network (SCAN), measuring semantic alignment with image regions and similarity to human captions via ranking and grounding distributions.
- **ViLBERTScore-F** (Lee et al., 2020) evaluates image captions by incorporating visual context into token embeddings, using ViLBERT to generate image-conditioned representations and compute similarity via cosine similarity.
- **CLIPScore** (Hessel et al., 2021) leverages CLIP embeddings to evaluate image-caption alignment without requiring reference captions, while its variant **RefCLIPScore** incorporates reference captions by computing the harmonic mean between CLIPScore and reference similarity.
- **PAC-S** (Sarto et al., 2023) enhances CLIP-based evaluation through positive-augmented contrastive learning using generated image-text pairs, with **RefPAC-S** incorporating reference captions similar to RefCLIPScore.
- **Polos** (Wada et al., 2024) is a multi-modal metric for image captioning that integrates reference-based assessment and learning-based modeling using CLIP and SimCSE-trained RoBERTa.
- **FLEUR** (Lee et al., 2024) utilizes LLaVA to directly evaluate captions against images using structured prompts and score smoothing, while **RefFLEUR** extends this by incorporating reference captions into the evaluation process.
- **HICE-S** (Zeng et al., 2024) evaluates image captions using global and local image-text similarities, improving hallucination detection. **RefHICE-S** extends it by adding reference captions via text-text compatibility analysis.
- **DENEB** (Matsuda et al., 2024) evaluates image captions with Sim-Vec Transformer, processing multiple references simultaneously to improve robustness against hallucinations using Sim-Vec Extraction for feature similarity.
- **InfoMetIC+** (Hu et al., 2023) is a reference-free metric for image captioning that evaluates captions at both coarse- and fine-grained levels by assessing text precision, vision recall, and token-level alignment with image regions.
- **BRIDGE** (Sarto et al., 2024) employs a dual-encoder architecture with a mapping module that fills masked caption templates using visually enriched pseudo tokens, enabling reference-free caption evaluation.

K Additional Examples

<p>Human Score: 0.70 A man standing on a field with a soccer ball.</p> 	<p>EXPERT Score: 0.67034</p> <ul style="list-style-type: none"> - Fluency: The caption is fluent and grammatically correct. It uses simple language that flows naturally, making it easy to read and understand. - Relevance: The caption is relevant as it accurately describes the visual content of the image. It identifies the main subject (a man) and his activity (standing on a field with a soccer ball), which aligns well with what is depicted. - Descriptiveness: The caption provides a basic description of the scene but lacks specific details that could enhance understanding, such as the man's attire the setting or the context of the activity. However, it still conveys the essential elements of the image. 	<p>FLEUR Score: 0.77070</p> <p>The caption accurately describes the image, as it shows a man standing on a field with a soccer ball. However, the caption could be more specific by mentioning that the man is a soccer player and that he is holding the soccer ball. The score of 0.7706974744796753 indicates that the caption is close to being perfect, but it could still be improved for more accuracy.</p>
<p>Human Score: 0.33 A person is in the air while riding his bicycle on top of the train on a clear blue sky day.</p> 	<p>EXPERT Score: 0.29035</p> <ul style="list-style-type: none"> - Fluency: The caption is grammatically correct and flows naturally, making it easy to read. However, it contains a few awkward phrases that could be improved for better clarity. - Relevance: The caption does not accurately describe the visual content of the image. It mentions a person riding a bicycle on a train, which is not present in the image. Instead, the image depicts a person on a power line. - Descriptiveness: The caption lacks key details that would provide a clear understanding of the image. It does not mention the power line or the surrounding landscape, which are important elements of the scene. 	<p>FLEUR Score: 0.11164</p> <p>The caption does not accurately describe the image. The image shows a person riding a bicycle on a wire, not on top of a train. The caption is not clear and does not provide enough information to accurately describe the scene.</p>
<p>Human Score: 0.65 A street sign on the side of a traffic light.</p> 	<p>EXPERT Score: 0.69724</p> <ul style="list-style-type: none"> - Fluency: The caption is fluent and grammatically correct. It uses simple language that flows naturally, making it easy to read and understand. - Relevance: The caption is relevant as it accurately describes the visual content of the image, which features a street sign and a traffic light. It captures the main elements present in the scene. - Descriptiveness: The caption provides a basic description of the image, mentioning both the street sign and the traffic light. However, it could be more informative by including details such as the sign's message or the specific location (e.g., 9th Street), but it still conveys the essential elements of the scene. 	<p>FLEUR Score: 0.72840</p> <p>The caption "a street sign on the side of a traffic light" does not accurately describe the image. The image shows a detour sign on a pole, not a street sign on the side of a traffic light. The caption does not provide enough information about the content of the image, and it does not accurately capture the essence of the scene. Therefore, the score is 0.7284017205238342.</p>
<p>Human Score: 0.15 A group of people playing frisbee in a city.</p> 	<p>EXPERT Score: 0.40036</p> <ul style="list-style-type: none"> - Fluency: The caption is fluent and grammatically correct. It uses simple language that flows naturally, making it easy to read and understand. - Relevance: The caption is relevant as it describes a group of people engaged in an activity, which aligns with the visual content of the image. However, it inaccurately specifies that they are playing frisbee, while the image depicts a different activity. - Descriptiveness: The caption provides a basic description of the scene but lacks specific details about the activity being performed or the setting. It does not capture the unique elements of the image, such as the blue structure or the presence of children, which would enhance the overall understanding of the scene. 	<p>FLEUR Score: 0.61829</p> <p>The caption does not accurately describe the image. While there is a group of people playing frisbee in the city, the image also shows a woman and two children walking down the street, and a woman and a child playing with a blue sculpture. The caption should be more specific and include these details to accurately describe the scene.</p>
<p>Human Score: 0.67 A person is standing on jagged rocks above the water below.</p> 	<p>EXPERT Score: 0.62650</p> <ul style="list-style-type: none"> - Fluency: The caption is fluent and grammatically correct. It uses clear and straightforward language, making it easy to read and understand. - Relevance: The caption is relevant as it describes a person standing on rocks, which aligns with the visual content of the image. However, it inaccurately states that the person is above water, which does not match the actual scene. - Descriptiveness: The caption provides a basic description of the scene, mentioning the person and the jagged rocks. However, it lacks specific details about the environment, such as the presence of clouds or the overall setting, which could enhance the informative quality of the description. 	<p>FLEUR Score: 0.79170</p> <p>The caption accurately describes the image, as it shows a person standing on jagged rocks above the water below. The image captures the essence of the scene, and the caption provides a clear and concise description of the situation. The score of 0.7916955947875977 indicates that the caption is a good representation of the image.</p>
<p>Human Score: 0.40 A group of people standing on top of a train.</p> 	<p>EXPERT Score: 0.54064</p> <ul style="list-style-type: none"> - Fluency: The caption is fluent and grammatically correct. It uses simple language that is easy to understand, making it sound natural. - Relevance: The caption is relevant as it describes a group of people on a train, which aligns with the visual content. However, it inaccurately states that they are standing on top of the train, which may not be the case based on the image. - Descriptiveness: The caption provides a basic description of the scene but lacks specific details about the setting, the train's appearance, or the activities of the people. While it captures the main elements, it could be more informative to enhance the viewer's understanding of the image. 	<p>FLEUR Score: 0.77680</p> <p>The caption accurately describes the image, as it shows a group of people standing on top of a train. However, the caption could be improved by providing more specific details about the train, such as its type or location.</p>

Figure 7: Additional examples of EXPERT.