CLAIR: Evaluating Image Captions with Large Language Models

David M. Chan, Suzanne Petryk, Joseph E. Gonzalez, Trevor Darrell, John Canny
University of California, Berkeley
{davidchan,spetryk,jegonzal,trevordarrell,canny}@berkeley.edu

Abstract

The evaluation of machine-generated image captions poses an interesting yet persistent challenge. Effective evaluation measures must consider numerous dimensions of similarity, including semantic relevance, visual structure, object interactions, caption diversity, and specificity. Existing highly-engineered measures attempt to capture specific aspects, but fall short in providing a holistic score that aligns closely with human judgments. Here, we propose CLAIR[1], a novel method that leverages the zero-shot language modeling capabilities of large language models (LLMs) to evaluate candidate captions. In our evaluations, CLAIR demonstrates a stronger correlation with human judgments of caption quality compared to existing measures. Notably, on Flickr8K-Expert, CLAIR achieves relative correlation improvements over SPICE of 39.6% and over image-augmented methods such as RefCLIP-S of 18.3%. Moreover, CLAIR provides noisily interpretable results by allowing the language model to identify the underlying reasoning behind its assigned score. Code is available at https://davidmchan.github.io/clair/.

1 Introduction & Background

Automatically evaluating the quality of image captions is challenging. There are many dimensions to consider, such as grammatical quality, semantic relevance, correctness, and specificity, among others. To ensure fair evaluations, most image captioning works employ a suite of measures, each capturing different aspects. For instance, n-gram-based measures like BLEU (Papineni et al., 2002) or CIDEr (Vedantam et al., 2015) broadly measure content overlap, SPICE (Anderson et al., 2016) compares scene graph structures, and CLIPScore, TIFA, SeeTrue and VPEval (Hessel et al., 2021; Hu et al., 2023; Yarom et al., 2023; Cho et al., 2023) directly incorporate visual information. Unfortunately, while significant strides have been made in automated evaluation, human preference studies remain the most reliable (yet costly) source of caption evaluation.

```
You are trying to tell if a candidate set of captions is describing the same image as a reference set of captions.
Candidate set:
{candidate captions}
Reference set:
{reference captions}
On a precise scale from 0 to 100, how likely is it that the candidate set is describing the same image as the reference set? (JSON format, with a key "score", value between 0 and 100, and a key "reason" with a string value.)
```

Figure 1: CLAIR: a (surprisingly simple) large language model-based measure for image caption evaluation. We find that CLAIR not only correlates strongly with human judgments of caption quality but can also generate interpretable reasons for the generated scores.

Fortunately, recent advances in large language models (LLMs) have opened new avenues for automatic evaluation. Models trained with reinforcement learning from human feedback (RLHF, Christiano et al. (2017)) or similar methods are particularly useful for open-ended evaluation tasks, including image captioning, due to their explicit training to align with human preferences.

In our work, paralleling several recent works which find that LLMs can act as effective "judges" for selecting the better answer from two candidates (Bubeck et al., 2023; Dettmers et al., 2023; Chiang et al., 2023), we explore the ability of LLMs to evaluate caption quality in the multimodal setting.

[1]Meaning "clear" in French, in line with other colorful metric names (Papineni et al., 2002; Lin, 2004; Lita et al., 2005).

We introduce CLAIR (Criterion using LAnguage models for Image caption Rating), a measure which scores a candidate caption based on the likelihood that it describes the same image as a set of references by directly asking an LLM to produce a numeric rating. We further query the LLM to provide a *reason* for its score, providing a level of interpretability to the scalar rating. As far as we are aware, this is the first paper to explore replacing measures of *semantic text quality* with directly obtained LLM judgments, however concurrently, Zheng et al. (2023) have shown that directly providing an answer rating can align highly with human preferences on a range of standard language-based tasks, such as conversational instruction following.

Through several experiments on captioning datasets such as MS-COCO (Xu et al., 2016), Flickr8k (Mao et al., 2014), and PASCAL-50S (Vedantam et al., 2015), we find that CLAIR correlates surprisingly well with human preferences, outperforming prior captioning measures. We additionally propose CLAIR$_E$, where we Ensemble the outputs of several LLMs by taking the average score, leading to further improvements.

Despite a simple pipeline using an LLM prompt with minimal output parsing, CLAIR's strong correlation with human preferences suggests that it captures multiple dimensions of caption similarity at once – a feature that prior measures struggle to achieve alone. More generally, CLAIR demonstrates how language-only models can evaluate vision-language tasks. We show LLMs can provide not only reliable scalar ratings but also corresponding reasoning for a given rating, offering a valuable combination of accuracy and interpretability.

2 CLAIR: LLMs for Caption Evaluation

In CLAIR, we adapt the zero-shot in-context learning approach described in Brown et al. (2020) to score candidate captions with large language models (LLMs). This involves converting the caption evaluation problem into a human-readable text completion task which is solved by the LLM. Using the prompt in Figure 1, CLAIR first generates completions from the LLM and then parses those completions into both candidate scores and an explainable reason for the score. We use a greedy sampling method ($t=0$) to encourage reproducibility in the results, while acknowledging the inherent nondeterminism in LLMs (see section 4). CLAIR's experimental implementation is surprisingly simple: it uses no in-context examples (is entirely zero-shot), and default inference parameters for the APIs. See Appendix B for further implementation details.

The choice of language model directly affects the quality of the CLAIR measure – more accurate models should produce evaluations that align better with human judgment. We explore three language models: GPT-3.5 (ChatGPT) (OpenAI, 2022), Claude (Instant) (Bai et al., 2022), and PaLM (Chowdhery et al., 2022). Unfortunately, we found for several open-weight language models including Koala (Geng et al., 2023) and Vicuna (Chiang et al., 2023) that CLAIR aligned poorly with human judgment.

As CLAIR is language model-agnostic, we can leverage the different distributions learned by each model and combine their decisions in an ensemble approach we term CLAIR$_E$. We calculate individual CLAIR scores for each model and compute an unweighted average to obtain the ensemble score.

Benchmark measures: We benchmark against several existing measure of caption similarity. BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Agarwal and Lavie, 2008) and CIDEr (Vedantam et al., 2015) all primarily measure n-gram overlap (however have different weighting schemes between n-grams, and across precision/recall). We also compare against SPICE (Anderson et al., 2016), which compares caption parse trees and focuses on matching perceived action and object relationships in addition to n-grams. While the aforementioned measures are commonly reported in image captioning works, we also compare against several more modern measures, including BERT-Score (Zhang et al., 2020) (which measures the distance between BERT embeddings in the text), BERT-Score++ (Yi et al., 2020) (which fine-tunes BERT for image captioning), LEIC (Cui et al., 2018) and NUBIA (Kane et al., 2020) (which are custom trained models for image caption evaluation), TIGEr (Jiang et al., 2019) (which is a model trained for caption evaluation which takes into account the original image context), and CLIP-Score (Hessel et al., 2021) which uses the recent CLIP (Radford et al., 2021) model for reference-free evaluation.

3 Evaluation & Discussion

To evaluate the quality of the measure, we run several evaluations that compare scores generated

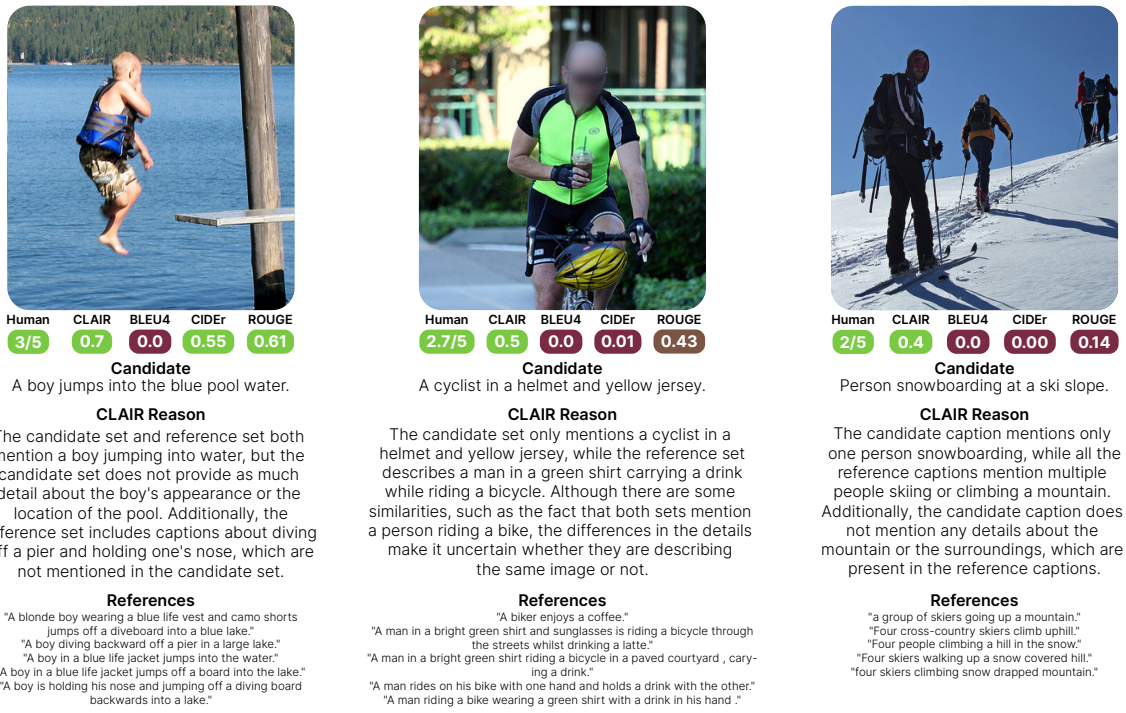

Figure 2: Several qualitative examples of CLAIR from the Flickr8K-Expert dataset. CLAIR not only correlates better with human judgments of caption quality but also provides detailed explanations for its score. CLAIR scores normalized by 100.

by CLAIR to both human judgments of caption quality and other image captioning evaluation measures. We additionally provide several qualitative examples in Figure 2. A unique benefit of CLAIR is that it provides not only numeric scores but is also introspectable, as it can identify which details in the candidate caption set match the reference set.

Sample-level human correlation: We first ask the question, how well does CLAIR correlate with human judgments of caption quality at a sample level? We do so by exploring the performance on three datasets, COMPOSITE, Flickr8K-Expert, and MS-COCO (See Appendix B for details).

The results of our sample-level correlation experiments are shown in Table 1. We can see that CLAIR outperforms language-only measures (e.g., 0.604 to 0.449 for BERT-S++), and in most cases, outperforms vision-augmented measures. CLAIR$_E$ achieves strong sample-level correlation on all datasets; for instance, CLAIR$_E$ closes the gap to inter-human agreement by 0.097 over vision-based measures and 0.132 over language-based measures. The improvements of CLAIR$_E$ over CLAIR suggest that each language model may have some bias (similar to each human), yet the ensemble of models correlates more strongly with human judgments. A reasonable concern might be that the models underlying existing approaches are significantly smaller than those in CLAIR, and trained on less data. To address this, we introduce and compare against RefCLIP-X, which replaces the CLIP model in RefCLIP with a CLIP ViT-bigG/14 model trained on LAION 2B (Ilharco et al., 2021). Even in this case, CLAIR demonstrates significantly improved performance.

Table 1: Sample-level correlation (Kendall's τ) with human judgments. All p-values < 0.001. *: Model has access to additional visual context. Results for LEIC, BERT-S++, TIGEr, and NUBIA are drawn from their original work.

	Dataset		
Measure	COMPOSITE	Flickr8K	MS-COCO
BLEU@1	0.313	0.323	0.265
BLEU@4	0.306	0.308	0.215
ROUGE-L	0.324	0.323	0.221
BERT-S	0.301	0.392	0.163
METEOR	0.389	0.418	0.239
CIDEr	0.377	0.439	0.262
SPICE	0.403	0.449	0.257
BERT-S++	0.449	0.467	-
NUBIA	-	0.495	-
LEIC*	-	0.466	-
TIGEr*	0.454	0.493	-
CLIP-S*	0.538	0.512	0.217
RefCLIP-S*	0.554	0.530	0.305
RefCLIP-X*	0.523	0.549	0.274
CLAIR			
+ GPT3.5	**0.604**	0.616	0.296
+ Claude	0.542	0.563	0.320
+ PaLM	0.580	0.546	0.355
CLAIR$_E$	0.592	**0.627**	**0.374**
Inter-Human	-	0.736	-

Table 2: System-level correlation between the average CLAIR score and human model evaluation for 5 models trained and evaluated on MS-COCO. All p-values < 0.05.

Measure	Kendall's τ	Spearman's ρ	Pearson r
BLEU@1	0.399	0.600	0.706
BLEU@4	0.799	0.899	0.910
ROUGE-L	0.600	0.700	0.792
METEOR	0.600	0.700	0.666
CIDEr	0.399	0.600	0.856
SPICE	0.399	0.600	0.690
CLAIR			
+ GPT3.5	0.799	0.899	0.869
+ Claude	**1.000**	**1.000**	0.868
+ PaLM	**1.000**	**1.000**	**0.954**
CLAIR$_E$	**1.000**	**1.000**	0.903

Table 3: Accuracy of measures when matching human decisions for PASCAL-50S (5 reference captions). *: Model has access to additional visual context.

Measure	HC	HI	HM	MM	All
BLEU@1	51.20	95.70	91.20	58.20	74.08
BLEU@4	53.00	92.40	86.70	59.40	72.88
ROUGE-L	51.50	94.50	92.50	57.70	74.05
METEOR	56.70	97.60	94.20	63.40	77.98
CIDEr	53.00	98.00	91.50	64.50	76.75
SPICE	52.60	93.90	83.60	48.10	69.55
TIGEr*	56.00	99.80	92.80	74.20	80.70
CLIP-S*	56.50	99.30	96.40	70.40	80.70
RefCLIP-S*	64.50	99.60	95.40	72.80	83.10
CLAIR					
+ GPT3.5	52.40	99.50	89.80	73.00	78.67
+ Claude	**57.90**	98.50	91.30	62.90	77.65
+ PaLM	54.70	98.30	87.30	64.00	76.08
CLAIR$_E$	57.70	**99.80**	**94.60**	**75.60**	**81.93**

System-level human correlation: In addition to computing the sample-level correlation on the MS-COCO dataset, we use the annotations from the five models considered by Rohrbach et al. (2018) to compute the system-level correlation. For each of the methods, we compute the mean human score on the test samples, and mean metric score on the test samples, followed by the Kendall's rank correlation coefficient (Kendall's tau, strength of ordinal association) between these values (the set of five mean human scores, and the set of five metric scores). Our results, given in Table 2, demonstrate that CLAIR ranks the five methods in a novel way that is more accordant with human rankings of the methods. These results further suggest that CLAIR has the potential to redefine which methods are preferable to humans compared to existing n-gram approaches.

Decision Making: In addition to evaluating the correlation with human judgments, we also evaluate the capability of the measure to perform discriminative analysis. The PASCAL-50S dataset (Vedantam et al., 2015) contains a set of 4000 human-annotated caption pairs. For each pair of captions, humans label which caption in the pair is closest to the reference set for the image. The caption pairs fall into four groups: "HC:" two human-written captions matching the image, "HI:" one human caption, and one machine-generated caption, with only one matching the image, "HM:" a matching human caption and a matching machine-generated caption and "MM:" two matching machine-generated captions. See Appendix B for more dataset information.

Table 4: Pearson correlation with human judgments when evaluating sets of captions on MS-COCO ($N = 794$).

Measure	Coverage$_{\text{p-value}}$	Correctness$_{\text{p-value}}$
BLEU@4	0.004 $_{0.816}$	0.003 $_{0.888}$
ROUGE-L	0.011 $_{0.563}$	0.038 $_{0.184}$
METEOR	0.016 $_{0.398}$	0.006 $_{0.765}$
CIDEr	0.004 $_{0.844}$	0.026 $_{0.173}$
TRM-METEOR	0.128$_{<0.001}$	0.108$_{<0.001}$
TRM-BLEU	0.127$_{<0.001}$	0.151$_{<0.001}$
MMD-BERT	0.129$_{<0.001}$	0.124$_{<0.001}$
FID-BERT	0.081 $_{0.011}$	0.098$_{<0.001}$
CLAIR		
+ GPT3.5	**0.195** $_{0.011}$	**0.187** $_{0.014}$
+ Claude	0.110 $_{0.099}$	0.124 $_{0.145}$
+ PaLM	0.129 $_{0.081}$	0.085 $_{0.172}$
CLAIR$_E$	0.183 $_{0.027}$	0.156 $_{0.018}$
Inter-Human	0.225$_{<0.001}$	0.274$_{<0.001}$

The performance on PASCAL-50S is given in Table 3. We can see that CLAIR$_E$ outperforms all existing text-only measures (e.g., by 5.18% overall score over CIDEr), and in many cases, even outperforms measures that have access to the image at test time. Note that it is relatively weaker than image-augmented models in the HC setting; however, since both captions are correct, the model often cannot judge which is better purely the text. Models such as RefCLIP-S that have access to the image are naturally better discriminators in this case. We suspect that CLAIR's discriminative performance could be further improved by giving the LLM a choice between the two captions; however, we leave this optimization to future work.

Groups of Captions: While CLAIR is capable of comparing a single candidate caption to a set of reference captions, it is also capable of comparing *sets* of candidate captions to sets of reference captions. This task is necessary when evaluating the ability of a model to generate captions that are diverse and that fully describe the conditional text

distribution. We evaluate on the COCO-Sets dataset (Chan et al., 2022), 794 caption sets rated by AMT workers on two scales: how closely a candidate set matches the reference set in terms of both correctness and content coverage (See Appendix B for details). The results of this experiment are given in Table 4. We can see that CLAIR outperforms well when measuring the quality of a group of captions, and approaches the inter-human correlation on the (very) challenging task. CLAIR also outperforms TRM-METEOR and TRM-BLEU (Chan et al., 2022), suggesting that LLMs can judge both the content and diversity of the caption sets.

4 Limitations

While CLAIR correlates well with human judgments of caption quality, it has several limitations:

Non-Determinism and Parsing Errors: Because CLAIR depends on the output of a language model, the measure can be non-deterministic and noisy. For instance, it may fail to elicit a judgment (e.g., "As an AI language model, I cannot see, and thus, cannot determine if the image captions match the references"), or rarely, generate malformed JSON output. To address these issues, we perform multiple queries to the LLM, sometimes at higher temperatures if necessary. As a consequence, the measure may differ between runs, although we found the variance to be relatively insignificant (< 0.01 in many of the experiments). Additionally, since the language models used are not open-source, the models are subject to arbitrary change, replacement, or removal, which limits the efficacy of the measure as a long-term comparable measurement. We hope that increasing open access to language models with efforts such as Koala (Geng et al., 2023) and Vicuna (Chiang et al., 2023), will help to alleviate these challenges in the future.

Increased Cost: CLAIR relies on language models which contain many billions of parameters. These language models have not only monetary cost but also human and environmental costs (Bender et al., 2021) which can reduce its utility as a target during training, such as for self-critical sequence training (Rennie et al., 2017). While API-based LLMs may be considered costly, even open-source LLMs have a cost (which can often be hard to quantify). CLAIR on the MS-COCO dataset uses an average of 226.148 tokens per sample (on OpenAI's API), representing a cost of $0.0067 per sample (GPT-4), or $0.00033 per sample (GPT 3.5). For PALM, this drops to $0.000113 per sample. We hope that over time, advances in LLM inference (such as quantization and distillation), coupled with improvements in architecture will continue to yield lower-cost alternatives with strong performance on the caption evaluation task.

Hallucination: While CLAIR does suffer from potential hallucination, we strongly believe that this weakness does not diminish the fact that CLAIR still correlates strongly with human judgment. In CLAIR, hallucinations in the score manifest as "incorrect" judgements of similarity, while hallucinations in the explanations manifest as poorly grounded explanations of the score/quality. Hallucinations in the score should be considered false negatives (blind spots instead of hallucinations). In the case of hallucinations in the explanations, such hallucinations may lead to misinterpretation, but arguably less misinterpretation than a black box method, and may even indicate misunderstandings in the model. Hallucination is a well-known challenge of current LLMs and is the subject of a great amount of research on instruction-tuning, RLHF, RLAIF, and other methods. As hallucination and instruction-following performance of the base models improves, CLAIR inherit similar improvements.

Explainability: While CLAIR generates explanations for each rating, CLAIR has no strict scoring rubric. Much like human judgments, there is no direct way of attributing changes in score to changes in caption quality. For similar reasons, it is difficult to evaluate the quality of the generated explanations. Qualitatively, the explanations are generally reasonable and consider multiple axes of judgment.

5 Conclusion

This work introduces CLAIR, an LLM-based evaluation measure for image captioning. CLAIR's superior performance compared to highly-engineered measures indicates a remarkable fact: LLMs are well aligned with human judgments of caption quality, even more so than some measures designed specifically for semantic similarity. CLAIR is only a glimpse into how LLMs can be used for evaluation tasks, and image captioning is only the beginning. We hope that our work will inspire further exploration of similar measures in other vision and language domains, such as visual storytelling (Huang et al., 2016), where human evaluation of generated text remains a challenging task.

References

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *ArXiv preprint*, abs/1511.03292.

Abhaya Agarwal and Alon Lavie. 2008. Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118. Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv preprint*, abs/2204.05862.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv preprint*, abs/2303.12712.

David M Chan, Yiming Ni, Austin Myers, Sudheendra Vijayanarasimhan, David A Ross, and John Canny. 2022. Distribution aware metrics for conditional natural language generation. *ArXiv preprint*, abs/2209.07518.

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.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Visual programming for text-to-image generation and evaluation. *arXiv preprint arXiv:2305.15328*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge J. Belongie. 2018. Learning to evaluate image captioning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5804–5812. IEEE Computer Society.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *ArXiv preprint*, abs/2305.14314.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

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. 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.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239,

San Diego, California. Association for Computational Linguistics.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

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. Association for Computational Linguistics.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Lucian Vlad Lita, Monica Rogati, and Alon Lavie. 2005. Blanc: Learning evaluation metrics for mt. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 740–747.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. 2014. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.

OpenAI. 2022. Introducing chatgpt.

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. 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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195. IEEE Computer Society.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. 2023. What you see is what you read? improving text-image alignment evaluation. *arXiv preprint arXiv:2305.10400*.

Yanzhi Yi, Hangyu Deng, and Jinglu Hu. 2020. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International*

Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv preprint*, abs/2306.05685.

Appendix

A Acknowledgements

We thank Suhong Moon and Kate Saenko for their helpful comments on the work. Authors, as part of their affiliation with UC Berkeley, were supported in part by the NSF, DoD, and/or the Berkeley Artificial Intelligence Research (BAIR) industrial alliance program, as well as gifts from Anyscale, Astronomer, Google, IBM, Intel, Lacework, Microsoft, Mohamed Bin Zayed University of Artificial Intelligence, Samsung SDS, Uber, and VMware.

B Additional Experimental Details

In this section, we provide several additional details for the experiments in section 3 run with the CLAIR measure.

B.1 Input Prompt Formatting

The CLAIR prompt is given in its entirety in Figure 1. During run-time, candidate and reference captions are prefixed with a "- " and inserted into the prompt, one per line. The resulting query is passed to the large language model. In addition, for models which were not RLHF-tuned to perform conversation (such as PaLM), we found that it was helpful to append an additional prefix `{"score":` to the beginning of the output, to encourage the correct output formatting. CLAIR is surprisingly simple: it uses no in-context examples (is entirely zero-shot), and default inference parameters for the APIs. The model checkpoint metadata is generally unknown (as the APIs are somewhat fluid and evolving).

B.2 LLM Output Post-Processing

Because CLAIR relies on an LLM to produce output, there is no guarantee that the output will be in the format we expect (i.e. valid, parsable JSON). To extract both the score and the reason, we first extract the first set of paired braces from the output of the LLM and attempt to parse the result as JSON. In most cases (99.997% for GPT-3, 99.991% for Claude, and 99.94% for PaLM during the course of our experiments), this is successful, and the score and reason are returned. In the case that the JSON output is malformed, we attempt to extract any sequence of digits from the LLM to use as a score, and set the reason to "Unknown." When this fails, as can be the case when the models produce an output such as "As an AI language model, I cannot see, and thus, cannot determine if the image captions match the references", we retry the prompt at a higher temperature ($t = 1.0$) several times. Failing this (which occurred only three times in the entire evaluation of this paper, across several hundred thousand calls), we set the score to 0 for the caption.

B.3 Datasets

In this section, we provide additional detail regarding the datasets used in the evaluations in section 3.

COMPOSITE: The COMPOSITE dataset (Aditya et al., 2015) contains machine-generated test captions for 3995 images spread across the MS-COCO (Xu et al., 2016), Flickr8K (Mao et al., 2014) and Flickr30k (Young et al., 2014) datasets. Each image has three test captions, one written by a human, and two that are model generated. The candidate captions are graded by annotators on Amazon Mechanical Turk (AMT) on a scale of 1 (not relevant) to 5 (very relevant). Inter-human correlations are not available for this dataset.

Flickr8K-Expert: The Flickr8K-Expert dataset (Hodosh et al., 2013) contains 5822 captions associated with 1000 images. The dataset is annotated with expert human judgments of quality, where images are rated from 1 (caption is unrelated to the image) to 4 (caption describes the image without errors). Unlike the composite and MS-COCO datasets, the captions here are selected using an image retrieval system, instead of generated using a learned image captioning model. Following Jiang et al. (2019), we exclude any candidate captions that overlap the reference set.

MS-COCO: Following experiments by Rohrbach et al. (2018), we compute the sample-level correlation between our method and human ratings on a 500-image subset of the MS-COCO Karpathy test set. Each image in the subset contains candidate captions generated by 5 models, and each caption is labeled with the average three human

ratings generated by AMT workers which range from 1 (very bad) to 5 (very good). Inter-human correlations are not available for this dataset.

PASCAL-50S: PASCAL-50S contains 1000 images drawn from the PASCAL sentence dataset. Each image is associated with at least 50 (and as many as 120) reference captions. In addition to the reference captions, PASCAL-50S contains a set of 4000 human annotated image/caption pairs containing an image, and two candidate captions. The caption pairs fall into four groups:

1. HC: In the HC group, both captions in the pair are human written, and describe the content of the target image correctly.
2. HI: In the HI group, both captions in the pair are human written, but one caption correctly describes the content of the image, and the other caption describes the content of a different image.
3. HM: In the HM group, one caption is written by a human, and one caption is written by a machine, but both correctly describe the content of the image.
4. MM: In the MM group, both captions are written by a machine, and both correctly describe the image content.

In PASCAL-50S, the task is to decide which caption in the pair humans prefer more (a subjective task, hopefully indicating caption quality). Following previous work (Jiang et al., 2019; Hessel et al., 2021), we limit the number of reference sentences to five during evaluation.

COCO-Sets: The COCO-Sets dataset (Chan et al., 2022) is a set of samples that are designed to evaluate the correlation of distribution-aware image captioning measures with human judgments of distributional distance. In this dataset, humans were presented with two candidate caption sets (two image captioning models, OFA (Wang et al., 2022) and BLIP (Li et al., 2022) using different temperatures), and asked which candidate caption set correlated better with a reference caption set on two measures: how much they overlapped factually (correctness), and how much information they provided about the references (coverage). It consists of 794 AMT worker-generated judgments of caption quality for images in the MS-COCO dataset.