Reference-Based Metrics Are Biased Against Blind and Low-Vision Users' Image Description Preferences

Rhea Kapur
Department of Computer Science
Stanford University
rheak@stanford.edu

Elisa Kreiss
Department of Communication
University of California, Los Angeles
ekreiss@ucla.edu

Abstract

Image description generation models are sophisticated Vision-Language Models which promise to make visual content, such as images, non-visually accessible through linguistic descriptions. While these systems can benefit all, their primary motivation tends to lie in allowing blind and low-vision (BLV) users access to increasingly visual (online) discourse. Well-defined evaluation methods are crucial for steering model development into socially useful directions. In this work, we show that the most popular evaluation metrics (*reference-based* metrics) are biased against BLV users and therefore potentially stifle useful model development. Reference-based metrics assign quality scores based on the similarity to human-generated ground-truth descriptions and are widely accepted as neutrally representing the needs of all users. However, we find that these metrics are more strongly correlated with sighted participant ratings than BLV ratings, and we explore factors which appear to mediate this finding: description length, the image's context of appearance, and the number of reference descriptions available. These findings suggest that there is a need for developing evaluation methods that are established based on specific downstream user groups, and they highlight the importance of reflecting on emerging biases against minorities in the development of general-purpose automatic metrics.

1 Introduction

As the internet becomes increasingly visual, long-standing accessibility issues blind and low-vision (BLV) users face remain largely unresolved (Gleason et al., 2019; Kreiss et al., 2022b). Vision-language models have enabled the automation of image-to-text description generation, which can be used to generate alt-text descriptions; this could make visual content accessible to BLV users through, e.g., screen readers (Morris et al., 2016; Gleason et al., 2019). However, these systems often do not directly address the needs of BLV users (MacLeod et al., 2017), which recent work has started to partially attribute to the evaluation methods used (Kreiss et al., 2022a).

Originating from machine translation and summarization literature, reference-based metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), CIDEr (Anderson et al., 2016), and SPICE (Vedantam et al., 2015) are the most common method for evaluating descriptions. These metrics require human-generated ground-truth descriptions (i.e., *references*) for scoring. With these, reference-based metrics quantify the similarity of the proposed description (i.e., *hypothesis*) to the ground-truth reference descriptions. The more similar the proposed hypothesis description is to the presumed "ideal" references (relevant content, sufficiently detailed, aligned with user preferences, etc.), the higher the assigned score for the description.

To calculate this similarity, reference-based metrics make decisions on how to quantify semantic similarity, trade off the signal from multiple reference descriptions, and treat variation in description length. Prior work suggests that BLV users have strong preferences about description length (McCall and Chagnon, 2022) and care greatly that they make sense in the context of where images appear (Kreiss et al., 2022a). Reference-based metrics do not necessarily make decisions on these variables, but their implications have not been explicitly evaluated for their potential downstream effects. With all of this in mind, we ask: when scoring descriptions, do reference-based metrics reflect the preferences of BLV users?

Overall, we find converging evidence that metric design and common use actively favor sighted over BLV user preferences, highlighting the need for automatic metric development that's grounded in downstream user needs.

Proceedings of the Third Workshop on NLP for Positive Impact, pages 308–314

2 Related Work

2.1 Reference-Based Metrics

While many reference-based metrics have been proposed, we focus in this work on the three most commonly used metrics for image captioning: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004). Like most reference-based metrics, BLEU was originally proposed for evaluating machine translation task performance. BLEU draws from a corpus of quality human reference "translations" (in our case, descriptions) and compares hypothesis and reference descriptions using their n-gram overlap match numbers (ranges from 1-4 ngrams; BLEU-1, BLEU-2, BLEU-3, and BLEU-4). Specifically, BLEU uses the maximum number of n-gram matches across *all* references for scoring. Additionally, BLEU applies a brevity penalty based on the hypothesis description and effective reference corpus lengths. BLEU prioritizes hypotheses that roughly match the length of reference descriptions.

METEOR finds generalized unigram matches between hypothesis and reference texts while also accounting for synonyms and morphological variants (unlike BLEU). It scores based on unigram precision and recall for each hypothesis-references pairing, implements a fragmentation measure to account for the relationship between the ordering of words in the hypothesis and reference texts, and then reports the maximum as the METEOR score. ROUGE-L, which was originally developed for text summarization, uses the longest common subsequence length rather than explicit n-gram overlap to score hypothesis-reference pairs, outputting the maximum score across pairs as the ROUGE-L score. Both METEOR and ROUGE-L do not reward nor penalize hypothesis length.

While these metrics have been repeatedly tested for their alignment with sighted participant judgments on machine translation or image description-like tasks, this was never extended to BLV user groups. In this work, we put the implicit assumption that these similarity measures are application-agnostic to the test, specifically focusing on their fit for guiding image accessibility efforts.

2.2 Comparing BLV and sighted image description preferences

While BLV user preferences for image descriptions are usually studied in isolation (Das et al., 2024; Muehlbradt and Kane, 2022; Stangl et al., 2020), recent work has started to investigate how those preferences compare to sighted user judgments (Kreiss et al., 2022a; Lundgard and Satyanarayan, 2021). Most significant to our work, Kreiss et al. (2022a) sampled images from Wikipedia and paired them with distinct article contexts. For example, the same image was shown in the article for Hairstyle, Advertisement, and Cooperation. The authors then crowdsourced descriptions for image-article pairs and had BLV and sighted participants rate the description quality. While BLV and sighted participant ratings were largely correlated, they also significantly diverged. Specifically, in contrast to sighted participants, BLV participants showed a strong preference for longer descriptions. Their data further indicates that while the context an image appeared in generally affected ratings, BLV participants were even more sensitive to it (Kreiss, 2023). These findings directly suggest that there are potential misalignments between BLV and sighted user preferences in image descriptions, and we investigate whether this leads to a potential misalignment with reference-based evaluation metrics.

3 Investigating Sighted and BLV User Alignment

Reference-based metrics are intended to be neutral measures of similarity. Since gold-standard reference descriptions are given, there is a priori no reason for the metrics to correspond more with the preferences of sighted vs. BLV users. However, in practice, all metrics are based on implicit assumptions that might induce biases in their scoring. We test this explicitly by correlating assigned metric scores with sighted and BLV participant judgments.

While recent work suggests that image context shapes sighted and BLV participants' preference ratings (Kreiss et al., 2022a; Stangl et al., 2021; Muehlbradt and Kane, 2022), these metrics tend to be used in context-agnostic settings. We therefore test these metrics in both context-sensitive and context-insensitive environments.

3.1 Datasets

We contrast two dataset conditions. For the *context-sensitive* dataset, we extract all image-article pairings, crowdsourced descriptions, and anonymized BLV/sighted participant ratings of those descriptions from Kreiss et al. (2022a). Then, for all descriptions *within* an image-article pair, we sample a description as hypothesis and the rest as ref-

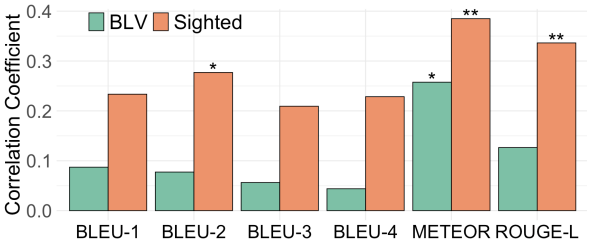

Figure 1: Correlations between BLV/sighted participant ratings and metric scores for each hypothesis-references pairing within the context-sensitive dataset. Asterisks denote statistically significant correlations.

erences, rotating until each description has been a hypothesis (68 examples). Using the nlg-eval Python package (Sharma et al., 2017), we calculate BLEU-1/2/3/4, METEOR, and ROUGE-L scores for each hypothesis-references pairing across all image-article contexts.

The *context-**in**sensitive* dataset is constructed in the same way but we collapse over different contexts. Suppose, we want to score the quality of a description written for an image that appeared in the Wikipedia article on Advertisement. In the context-sensitive condition, we compute the similarity of this description to the other descriptions for this image in this particular article. However, this same image also appeared in the Wikipedia article on Hairstyles. In the context-insensitive dataset, we compare the hypothesis description to all other available descriptions for this image, independent of the context in which they were written. In other words, all possible descriptions for the same image across *different* contexts are sampled as references (250 examples).

3.2 Methods and Correlation Results

For each hypothesis-references pairing, we calculate the Pearson correlation between metric scores and ratings from both BLV and sighted participant groups. In the context-sensitive condition, we find that across all metrics, correlations with sighted participant ratings are not only always higher but also more often *significant* in comparison to those with BLV participant ratings (see Figure 1). Contrary to the alleged neutrality of reference-based metrics, there is clear bias toward sighted user preferences.

Figure 2 shows the difference in correlations from the context-sensitive and context-insensitive datasets. A positive difference suggests that the correlation with participant ratings increased in the context-insensitive condition, i.e., when references were pooled across contexts. Prior evidence suggests that context is an important signal for evaluating the usefulness of a description (Stangl et al., 2021; Muehlbradt and Kane, 2022; Kreiss et al., 2022a) and we should therefore expect the correlations to decrease due to the noisier reference signal.

Overall, the results are very mixed. Firstly, while BLV ratings largely decrease in correlation, sighted ratings increase. This is aligned with previously observed patterns suggesting that BLV participants are more sensitive to contextual variation compared to sighted participants (Kreiss, 2023). Additionally, there is significant variation between metrics in their context-sensitivity. With METEOR and ROUGE-L, the correlation with ratings from both participant groups are robustly higher when the context condition is respected. For BLEU-1 and BLEU-2, this pattern is reversed, suggesting that they are insensitive to the lost contextual signal. Interestingly, for BLEU-3 and BLEU-4, we see divergent behavior, where the expected pattern arises with the BLV data but not for the sighted data.

The results suggest that implicit metric decisions have significant impact on their alignment with participant ratings, and they underscore the importance of explicitly considering context relevance in evaluation of model performance.

4 Understanding the Misalignment

To contextualize the misalignment with BLV participant ratings, we analyze specific factors that might mediate the computed similarity.

4.1 Description Length

Hypothesis description length is a parameter over which metrics make decisions (for example, BLEU enforces a brevity penalty). Strikingly, BLV users tend to have strong preferences for description length (McCall and Chagnon, 2022). In the dataset

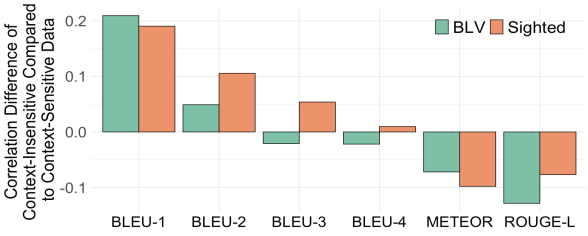

Figure 2: Comparison of correlations between BLV/sighted participant ratings and metric scores for each hypothesis-references pairing in the context-sensitive vs. context-insensitive datasets.

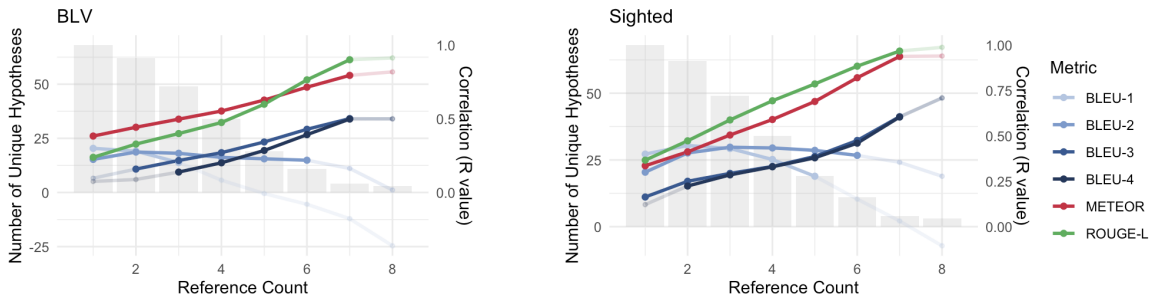

Figure 3: Reference count vs. correlation between metrics and human ratings in the context-sensitive data. Bars indicate number of unique hypotheses. Lines and points are faded when correlations are not significant ($p > 0.05$).

we investigate here, BLV participants had a strong preference for the longer descriptions; a core point of contrast to the sighted participant ratings (Kreiss et al., 2022a). It is therefore plausible that the distinct treatment of description length could be one cause for a potential misalignment between metric scores and BLV participant judgments. Using the context-sensitive dataset, we therefore calculated Pearson correlations between the metric scores and hypothesis length. If they reflect BLV participant behavior, metric scores should significantly correlate with description length. Otherwise, they rather reflect sighted participant preference trends.

Across all metrics, there were no significant correlations with hypothesis length (see Appendix A). This suggests a lack of adequate consideration of description length in reference-based metric design, which may account for their current bias toward sighted and against BLV user judgments.

4.2 Robustness: Reference Count

A widely attested variable shown to negotiate the reliability of reference-based metric scores is the number of references that the hypothesis description is compared to. Prior work suggests that approximately 5 references strike a balance between being reasonable to obtain labels for and converging to stable metric scores (Vedantam et al., 2015).

In this analysis, we investigate how stable the correlation results are based on the number of references available for evaluation. Some reference-based metrics evaluate each available hypothesis-references pairing and then take the maximum score (METEOR, ROUGE-L), while others consider all available references at once during evaluation (BLEU). To investigate this question, we use the context-sensitive data and construct dataset variants by sampling references ranging from 1 to the maximum amount of descriptions available for an image-article pairing (1956 examples).

Figure 3 shows the correlation between metrics and BLV/sighted participant ratings against reference count. The light gray bars indicate the number of unique hypothesis descriptions that the correlations are computed over. Firstly, the overall changes in correlations pattern similarly for BLV and sighted participant judgments. METEOR and ROUGE-L produce higher correlations as the reference count increases, while there is divergence within BLEU. Correlations involving BLEU-1 and BLEU-2 appear to decrease with increasing reference counts, while they increase with BLEU-3 and BLEU-4. Interestingly, this divergence resembles the context-sensitivity analysis pattern in Figure 2, further suggesting a qualitative shift in metric behavior when comparing BLEU-1/2 to BLEU-3/4.

Overall, these results suggest a high degree of variation depending on the underlying reference count. For most metrics, increasing the number of references results in increased correlation with BLV and sighted participant ratings, suggesting that it is essential for reliable downstream estimates.

5 Conclusion

In contrast to the general perception of reference-based metrics as universally applicable, we find that they're significantly biased toward sighted and against BLV user preferences. In an effort to understand this pattern, we find that these metrics do not correlate well with description length (which BLV users care for greatly), and performance varies with regard to context and reference count for certain metrics more than others. Our results highlight the necessity for developing reference-based metrics which put BLV user needs at the center of their design and evaluation pipeline in order to mitigate the current bias.

6 Limitations

The findings in this work indicate that reference-based metrics are likely biased toward sighted user preferences, and our ablation studies suggest that this may originate from their treatment of factors important to BLV users, such as context and length, as well as variables which implicitly affect scoring, such as reference count. However, specifically in the robustness analysis, the limited size and sourcing of the underlying dataset restricts the scope of the findings. Only few hypotheses have as many as eight references and further data efforts are needed to robustly quantify the benefits of increased reference count. Additionally, pairings and descriptions are solely scraped from Wikipedia, which may introduce platform-specific bias in the results.

While we analyzed context and length, there are a number of other factors important to BLV users that appear to be fundamental limitations for reference-based metrics. For example, models should be able to indicate uncertainty over generated content (MacLeod et al., 2017), optimize for identity-respecting language (Bennett et al., 2021), and be severely sensitive to hallucinations (MacLeod et al., 2017). In future analyses and development of accessibility-first metrics, we need to holistically evaluate and document these dimensions of quality assessment to promote evaluation metrics that can more easily translate to lasting social impact.

Acknowledgments

This research is supported in part by a grant from Google through the Stanford Institute for Human-Centered AI. We would like to thank the anonymous reviewers for their thoughtful feedback. We also thank Carmen Gutierrez for her contributions to early iterations of this work as part of the Spring 2023 offering of the CS224U: Natural Language Understanding class at Stanford University.

References

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.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Cynthia L Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P Bigham, Anhong Guo, and Alexandra To. 2021. "It's complicated": Negotiating accessibility and (mis) representation in image descriptions of race, gender, and disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Maitraye Das, Alexander J. Fiannaca, Meredith Ringel Morris, Shaun K. Kane, and Cynthia L. Bennett. 2024. From Provenance to Aberrations: Image Creator and Screen Reader User Perspectives on Alt Text for AI-Generated Images. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference*, WWW '19, page 549–559, New York, NY, USA. Association for Computing Machinery.

Elisa Kreiss. 2023. *The Pragmatics of Image Description Generation*. Ph.D. thesis, Stanford University.

Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022a. Context Matters for Image Descriptions for Accessibility: Challenges for Referenceless Evaluation Metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4685–4697, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elisa Kreiss, Fei Fang, Noah Goodman, and Christopher Potts. 2022b. Concadia: Towards Image-Based Text Generation with a Purpose. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4667–4684.

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.

Alan Lundgard and Arvind Satyanarayan. 2021. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE transactions on visualization and computer graphics*, 28(1):1073–1083.

Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. In

Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, page 5988–5999, New York, NY, USA. Association for Computing Machinery.

Karen McCall and Beverly Chagnon. 2022. Rethinking Alt text to improve its effectiveness. In *International Conference on Computers Helping People with Special Needs*, pages 26–33. Springer.

Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P. Bigham, and Shaun K. Kane. 2016. "With most of it being pictures now, I rarely use it": Understanding Twitter's Evolving Accessibility to Blind Users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5506–5516, New York, NY, USA. Association for Computing Machinery.

Annika Muehlbradt and Shaun K. Kane. 2022. What's in an ALT Tag? Exploring Caption Content Priorities through Collaborative Captioning. *ACM Trans. Access. Comput.*, 15(1).

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.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who are Blind or Have Low Vision. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '21, New York, NY, USA. Association for Computing Machinery.

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.

A Length Correlations

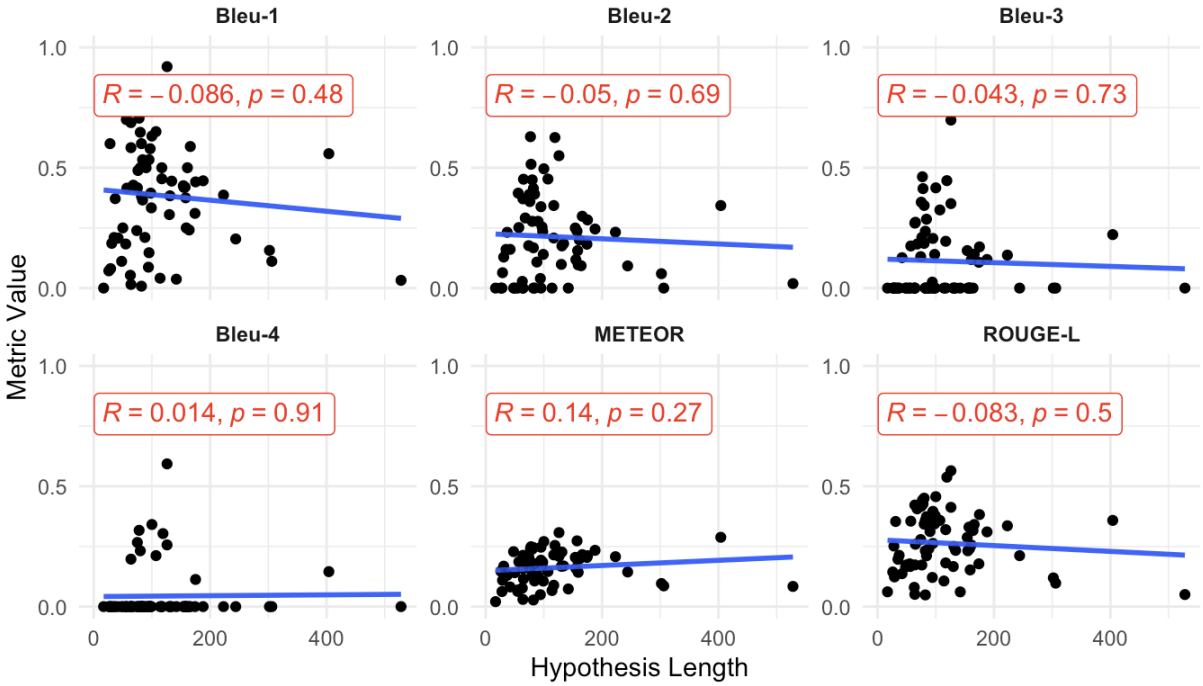

Figure 4: Correlation between reference-based metrics and hypothesis description length for the context-sensitive dataset.