Context Matters for Image Descriptions for Accessibility: Challenges for Referenceless Evaluation Metrics

Elisa Kreiss[*]
Stanford University

Cynthia Bennett
Google Research

Shayan Hooshmand
Columbia University

Eric Zelikman
Stanford University

Meredith Ringel Morris
Google Research

Christopher Potts
Stanford University

Abstract

Few images on the Web receive alt-text descriptions that would make them accessible to blind and low vision (BLV) users. Image-based NLG systems have progressed to the point where they can begin to address this persistent societal problem, but these systems will not be fully successful unless we evaluate them on metrics that guide their development correctly. Here, we argue against current *referenceless* metrics – those that don't rely on human-generated ground-truth descriptions – on the grounds that they do not align with the needs of BLV users. The fundamental shortcoming of these metrics is that they do not take context into account, whereas contextual information is highly valued by BLV users. To substantiate these claims, we present a study with BLV participants who rated descriptions along a variety of dimensions. An in-depth analysis reveals that the lack of context-awareness makes current referenceless metrics inadequate for advancing image accessibility. As a proof-of-concept, we provide a contextual version of the referenceless metric CLIPScore which begins to address the disconnect to the BLV data. An accessible HTML version of this paper is available at https://elisakreiss.github.io/contextual-description-evaluation/paper/reflessmetrics.html

1 Introduction

In the pursuit of ever more powerful image description systems, we need evaluation metrics that provide a clear window into model capabilities. At present, we are seeing a rise in *referenceless* (or reference-free) metrics (Hessel et al., 2021; Lee et al., 2021a,b; Feinglass and Yang, 2021), building on prior work in domains such as machine translation (Lo, 2019; Zhao et al., 2020) and summarization (Louis and Nenkova, 2013; Peyrard and Gurevych, 2018; Gao et al., 2020; Deutsch et al., 2021). These metrics seek to estimate the quality of a text corresponding to an image without requiring ground-truth labels (i.e., reference descriptions), or crowd worker judgments. Thus, they offer the promise of quick and efficient evaluation of image description models, and are even suggested to be more reliable than existing reference-based metrics (Kasai et al., 2022b,a). Here, we investigate the value of such metrics for a high social-impact domain: assessing the usefulness of image descriptions for blind and low vision (BLV) users.

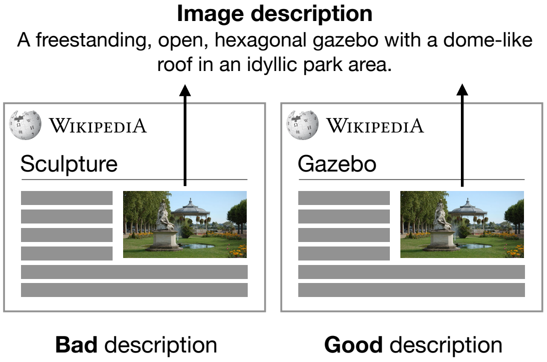

Figure 1: Whether an image description makes an image accessible depends on the context in which the image appears. Referenceless metrics like CLIPScore can't capture such context-sensitivity. We provide experimental evidence with blind and low vision (BLV) participants that this makes current referenceless metrics insufficient for evaluating image description quality.

Automatically generating descriptions to make images accessible is an important goal: though images are omnipresent in digital communication (Hackett et al., 2003; Bigham et al., 2006; Buzzi et al., 2011; Voykinska et al., 2016; Gleason et al., 2019), user-generated descriptions are rare (Gleason et al., 2020), which has serious implications for BLV users (Morris et al., 2016). Can referenceless metrics help guide models to generate descriptions that align with what BLV users value?

Studies with BLV users emphasize the impor-

[*]Corresponding author: ekreiss@stanford.edu

tance of the *context* in which an image appears. For example, while people's clothing is highly relevant when browsing shopping websites, their identities become central when reading the news (Stangl et al., 2021; Muehlbradt and Kane, 2022; Stangl et al., 2020). Not only the domain but also the immediate context matters for selecting what is relevant. Consider the image in Figure 1, showing a park with a gazebo in the center and a sculpture on a pedestal in the foreground. A description written for the image's occurrence in the Wikipedia article of gazebos ("A freestanding, open, hexagonal gazebo with a dome-like roof in an idyllic park area.") is unhelpful if the image instead appears in the article on sculpture. This simple example illustrates that context could play a central role in the assessment of description quality.

In this work, we report on studies with sighted and BLV participants that seek to provide rich, multidimensional information about what people value in image descriptions for accessibility. In contrast to current practices, we elicit and evaluate image descriptions within contexts the images could appear in, here Wikipedia articles. We find that, for both sighted and BLV participants, the description's relevance to the context is a major driver of their overall assessments.

We then use this experimental data to evaluate two very different referenceless metrics: CLIPScore (Hessel et al., 2021), which assesses a description's quality relative to its associated image, and SPURTS (Feinglass and Yang, 2021), which relies only on linguistic properties of the text. By their very design, these metrics don't capture the effects of context seen in our user studies, since they treat description evaluation as a context-less problem. This shortcoming goes undetected on most existing datasets and previously conducted human evaluations, which presume that image descriptions are context-independent, but it is immediately apparent in our evaluations.

These results suggest that current referenceless metrics are not reliable guides due to their lack of context integration, but offers a path forward: perhaps referenceless metrics can be modified to include this missing context. As a proof-of-concept, we show that a context-sensitive adaptation of CLIPScore results in improved correlations with human judgments – a promising signal for the development of future context-sensitive referenceless metrics.

2 Background

2.1 Image Accessibility

Screen readers provide auditory and braille access to Web content. To make images accessible in this way, screen readers use image descriptions embedded in HTML `alt` tags. However, such descriptions are rare. While frequently visited websites are estimated to have about 72% coverage (Guinness et al., 2018), this drops to less than 6% on English-language Wikipedia (Kreiss et al., 2022) and to 0.1% on English-language Twitter (Gleason et al., 2019). This has severe implications especially for BLV users who have to rely on such descriptions to engage socially (Morris et al., 2016; MacLeod et al., 2017; Buzzi et al., 2011; Voykinska et al., 2016) and stay informed (Gleason et al., 2019; Morris et al., 2016).

Moreover, these coverage estimates are based on any description being available, without regard for whether the descriptions are useful. Precisely what constitutes a useful description is still an underexplored question. A central finding from work with BLV users is that one-size-fits-all image descriptions don't address image accessibility needs (Stangl et al., 2021; Muehlbradt and Kane, 2022; Stangl et al., 2020). Stangl et al. (2021) specifically tested the importance of the *scenario* – the source of the image and the informational goal of the user – by placing each image within different source domains (e.g., news or shopping website) which were associated with specific goals (e.g., learning or browsing for a gift). They find that BLV users have certain description preferences that are stable across scenarios (e.g., people's identity and facial expressions, or the type of location depicted), whereas others are scenario-dependent (e.g., hair color). We extend this previous work by keeping the scenario stable but varying the immediate context the image is embedded in.

Current referenceless metrics take the one-size-fits-all approach. We explicitly test whether this is sufficient to capture the ratings provided by BLV users when they have access to the broader context.

2.2 Image-based Text Evaluation Metrics

There are two evaluation strategies for automatically assessing the quality of a model's generated text from images: *reference-based* and *referenceless* (or reference-free) metrics.

Reference-based metrics rely on human-created ground-truth texts associated with each image. The

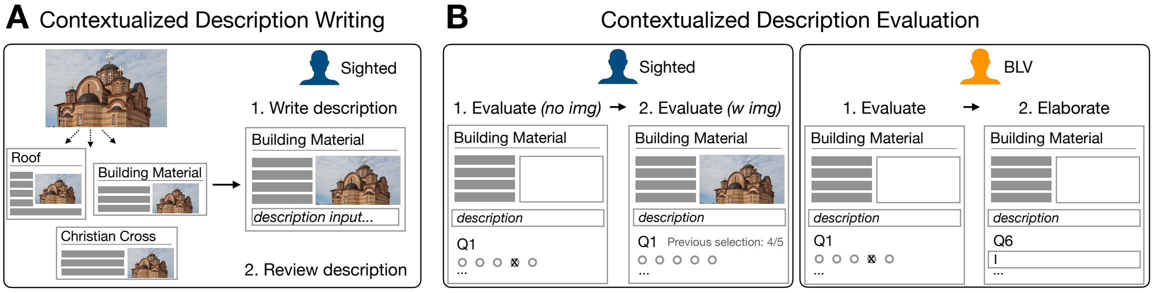

Figure 2: Experimental design overview consisting of two main phases: (A) eliciting descriptions written for images occurring within varying contexts, (B) obtaining detailed evaluations of those descriptions from sighted and BLV participants. These evaluations give insights into the role that context needs to play for providing useful descriptions, and function as the gold standard that the results from referenceless metrics are then compared to.

candidate text generated by the model is then compared with those ground-truth references, returning a similarity score. A wide variety of scoring techniques have been explored. Examples are BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016), ROUGE (Lin, 2004), and BERTscore (Zhang et al., 2019). The more references are provided, the more reliable the scores, which requires datasets with multiple high-quality annotations for each image. Such datasets are expensive and difficult to obtain.

As discussed above, referenceless metrics dispense with the need for ground-truth reference texts. Instead, text quality is assessed based either on how the text relates to the image content (Hessel et al., 2021; Lee et al., 2021b,a) or on text quality alone (Feinglass and Yang, 2021). As a result, these metrics can in principle be used anywhere without the need for an expensive annotation effort.

How the score is computed varies between metrics. CLIPScore (Hessel et al., 2021) and UMIC (Lee et al., 2021b) pose a classification problem where models are trained contrastively on compatible and incompatible image–text pairs. A higher score for a given image and text as input then corresponds to a high compatibility between them. QACE provides a high score if descriptions and images give similar answers to the same questions (Lee et al., 2021a). SPURTS is a referenceless metric which judges text quality solely based on text-internal properties that can be conceptualized as maximizing unexpected content (Feinglass and Yang, 2021). SPURTS was originally proposed as part of the metric SMURF, which additionally contains a reference-based component, specifically designed to capture the semantics of the description. However, Feinglass and Yang find that SPURTS alone already seems to approximate human judgments well, which makes it a relevant referenceless metric to consider. While varying in their approach, all current referenceless metrics share that they treat image-based text generation as a context-independent problem.

Reference-based metrics have the *potential* to reflect context-dependence, assuming the reference texts are created in ways that engage with the image's context. Referenceless methods are much more limited in this regard: if a single image–description pair should receive different scores in different contexts, but the metric operates only on image–description pairs, then the metric will be intrinsically unable to provide the desired scores.

3 Experiment: The Effect of Context on Human Image Description Evaluation

Efforts to obtain and evaluate image descriptions through crowdsourcing are mainly conducted out-of-context: images that might have originally been part of a tweet or news article are presented in isolation to obtain a description or evaluation thereof. Following recent insights on the importance of the domains an image appeared in (Stangl et al., 2021; Muehlbradt and Kane, 2022; Stangl et al., 2020), we seek to understand the role of context in shaping how people evaluate descriptions. Figure 2 provides an overview of the two main phases. Firstly, we obtained contextual descriptions by explicitly varying the context each image could occur in (Figure 2A). We then explored how context affects sighted and BLV users' assessments of descriptions along a number of dimensions (Figure 2B). Finally, in Section 4, we compare these contextual evaluations with the results from the referenceless metrics CLIPScore (Hessel et al., 2021) and SPURTS (Feinglass and Yang, 2021).

3.1 Data

To investigate the effect of context on image descriptions, we designed a dataset where each image was paired with three distinct contexts, here Wikipedia articles. For instance, an image of a church was paired with the first paragraphs of the Wikipedia articles on *Building material*, *Roof*, and *Christian cross*. Similarly, each article appeared with three distinct images. The images were made publicly available through Wikimedia Commons. Overall, we obtained 54 unique image–context pairs, consisting of 18 unique images and 17 unique articles. The dataset, experiments used for data collection, and analyses are made available.[1]

3.2 Contextual Description Writing

In our first experiment, participants sought to write descriptions that could make images accessible to users who can't see them.

Task Each participant went through a brief introduction explaining the challenge and purpose of image descriptions and was then shown six distinct articles, each of them containing a different image they were asked to describe. To enable participants to judge their descriptions, the description then replaced the image and participants could choose to edit their response before continuing. The task did not contain any guidance on which information should or should not be included in the description. Consequently, any context-dependence is simply induced by presenting the images within contexts (Wikipedia articles) instead of in isolation.

Participants and Exclusions We recruited 74 participants on Amazon's Mechanical Turk. We excluded six participants who indicated confusion about the task in the post-questionnaire and one for whom the experiment didn't display properly. Overall, each image–article pair received on average five descriptions.

Results After exclusions, we obtained 272 descriptions that varied in length between 13 and 541 characters, with an average of 24.9 words. We evaluate to what extent the description content was affected by the image context based on the following human subject evaluation experiment.

3.3 Contextual Description Evaluation

After obtaining contextual image descriptions, we designed a description evaluation study which we conducted with BLV as well as sighted participants. Both groups can provide important insights. We consider the ratings of BLV participants as the primary window into accessibility needs. However, sighted participant judgments can complement these results, in particular by helping us determine whether a description is true for an image. Furthermore, the sighted participants' intuitions about what makes a good description are potentially informative since sighted users are usually the ones providing image descriptions.

Task Sighted as well as BLV participants rated each image description as it occurred within the respective Wikipedia article. To get a better understanding of the kinds of content that might affect description quality, each description was evaluated according to five dimensions: *Overall* quality, *Imaginability* of the image from the description, *Relevance* and *Irrelevance* of the mentioned details, and image *Fit* to the article.

The *Imaginability* and *(Ir)Relevance* questions are designed to capture two central aspects of description content. While *Imaginability* has no direct contextual component, *Relevance* and *Irrelevance* specifically ask about the contextually determined aspects of the description. These dimensions give us insights into the importance of context in the *Overall* description quality ratings.

Responses were provided on 5-point Likert scales. In addition to 17 critical trials each participant completed, we further included two trials with descriptions carefully constructed to exhibit for instance low vs. high context sensitivity. These trials allowed us to ensure that the questions and scales were interpreted as intended by the participants. Overall, each participant completed 19 trials, where each trial consisted of a different article and image. Trial order and question order were randomized between participants to avoid potential ordering biases.

3.3.1 Sighted Participants

Task To ensure high data quality, sighted participants were asked a reading comprehension question before starting the experiment, which also familiarized them with the overall topic of image accessibility. If they passed, they could choose to enter the main study, otherwise they exited the study and were only compensated for completing the comprehension task.

In each trial, participants first saw the Wikipedia article, followed by an image description. This *no*

[1] https://github.com/elisakreiss/contextual-description-evaluation

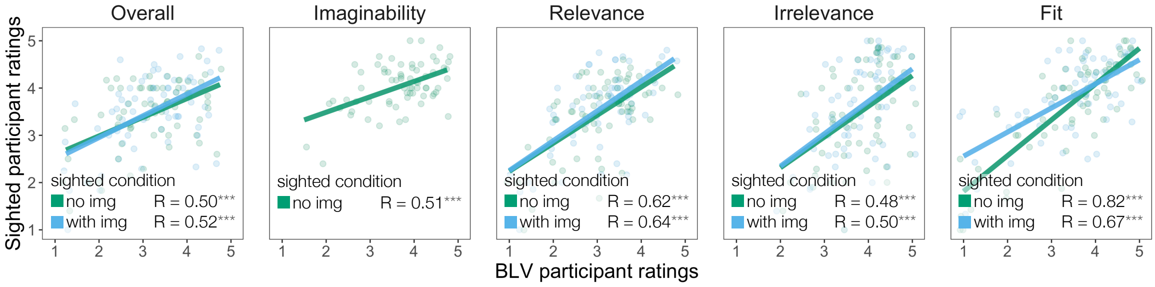

Figure 3: Correlation of BLV and sighted participant ratings across questions. Sighted participants provided ratings twice – before seeing the image (in green) and after (in blue). Each point denotes the average rating for a description. The Pearson correlations (R) are all statistically significant, as indicated by the asterisks. For all questions, higher ratings are associated with higher quality descriptions.

image condition can be conceptualized as providing sighted participants with the same information as a BLV user. They then responded to the five questions and were asked to indicate if the description contained false statements or discriminatory language. After submitting the response, the image was revealed and participants responded again to four of the five questions. The *Imaginability* question was omitted since it isn't clearly interpretable once the image is visible. Their previous rating for each question was made available to them so that they could reason about whether they wanted to keep or change their response.

Participants and Exclusions 79 participants were recruited over Amazon's Mechanical Turk, 68 of whom continued past the reading comprehension question. We excluded eight participants since they spent less than 19 minutes on the task, and one participant whose logged data was incomplete. This resulted in 59 submissions for further analysis.

3.3.2 BLV Participants

The 68 most-rated descriptions across the 17 Wikipedia articles and 18 images were then selected to be further evaluated by BLV participants.

Task To provide BLV participants with the same information as sighted participants, they similarly started with the reading comprehension question before continuing to the main trials. After reading the Wikipedia article and the image description, participants first responded to the five evaluation dimensions. Afterwards, they provided answers to five open-ended questions about the description content. The main focus of the analysis presented here is on the Likert scale responses, but the open-ended explanations allow more detailed insights into description preferences. Each description was rated by exactly four participants.

Participants 16 participants were recruited via email lists for BLV users, and participants were unknowing about the purpose of the study. Participants self-described their level of vision as totally blind (7), nearly blind (3), light perception only (5), and low vision (1). 15 participants reported relying on screen readers (almost) always when browsing the Web, and one reported using them often.

We enrolled fewer blind participants than sighted participants, as they are a low-incidence population, requiring targeted and time-consuming recruitment. For example, crowd platforms that enable large sample recruitment are inaccessible to blind crowd workers (Vashistha et al., 2018).

3.3.3 Evaluation Results

The following analyses are based on the 68 descriptions, comprising 18 images and 17 Wikipedia articles. Each description is evaluated according to multiple dimensions by sighted as well as BLV participants for how well the description serves an accessibility goal.

Figure 3 shows the correlation of BLV and sighted participant ratings across questions. We find that the judgments of the two groups are significantly correlated for all questions. The correlation is encouraging since it shows an alignment between the BLV participants' reported preferences and the sighted participants' intuitions. Whether sighted participants could see the image when responding didn't make a qualitative difference. The results further show that the dataset provides very poor to very good descriptions, covering the whole range of possible responses. This range is important for insights into whether a proposed evaluation metric can detect what makes a description useful.

We conducted a mixed effects linear regression analysis of the BLV participant judgments to in-

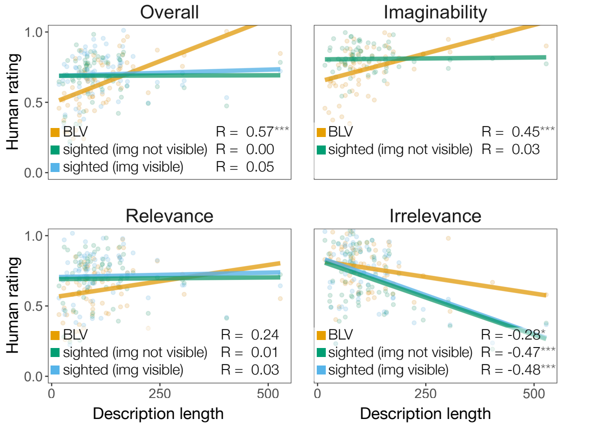

Figure 4: Correlation of BLV and sighted participant judgments with description length (in characters). Human judgments are rescaled to the zero to one range.

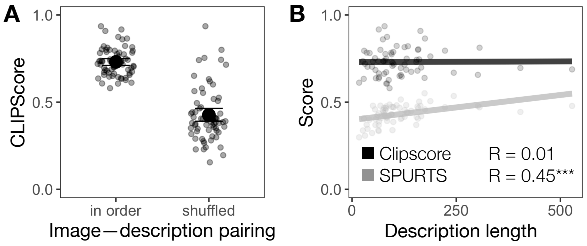

Figure 5: Analyses of the capabilities of referenceless metrics. (A) CLIPScore can pick out whether a written description is compatible with the image. When shuffling image–description pairs, the average CLIPScore drops from 0.73 to 0.43. SPURTS can't make this distinction due to its image-independence. (B) Longer descriptions are associated with higher scores of SPURTS but not CLIPScore.

vestigate which responses are significant predictors of the overall ratings. We used normalized and centered fixed effects of the three content questions (*Imaginability*, *Relevance* and *Irrelevance*), and random by-participant and by-description intercepts. If context doesn't affect the quality of a description, *Imaginability* should be a sufficient predictor of the overall ratings. However, in addition to an effect of *Imaginability* ($\beta = .42$, SE $= .06$, $p < .001$), we find a significant effect of *Relevance* as well ($\beta = .44$, SE $= .05$, $p < .001$), suggesting that context plays an essential role in guiding what makes a description useful. This finding replicates with the sighted participant judgments.

A case where BLV and sighted participant ratings diverge is in the effect of description length (Figure 4). While longer descriptions tend to be judged overall more highly by BLV participants, there is no such correlation for sighted participants. This finding contrasts with popular image description guidelines, which often advocate for shorter descriptions.[2] The lack of correlation between sighted participant ratings and description length might be linked to this potential misconception.

4 Referenceless Metrics for Image Accessibility

Referenceless metrics have been shown to correlate well with how sighted participants judge description quality when descriptions are written and presented out-of-context (Hessel et al., 2021; Feinglass and Yang, 2021; Lee et al., 2021b; Kasai et al., 2022a). While image accessibility is one of the main goals referenceless metrics are intended to facilitate (Kasai et al., 2022b; Hessel et al., 2021; Kasai et al., 2022a), it remains unclear whether they can approximate the usefulness of a description for BLV users. Inspired by recent insights into what makes a description useful, we argue that the inherently decontextualized nature of current referenceless metrics makes them inadequate as a measure of image accessibility. We focus on two referenceless metrics to support these claims: CLIPScore (Hessel et al., 2021) and SPURTS (Feinglass and Yang, 2021). Appendix B briefly considers other referenceless metrics.

CLIPScore uses the similarity of CLIP's image and description embeddings as the predictor for description quality, as formulated in (1). Denoting $\frac{x}{|x|}$ as $\overline{x}$, we can express CLIPScore as

$$\max\left(\overline{\text{image}} \cdot \overline{\text{description}}, 0\right) \tag{1}$$

SPURTS is different from CLIPScore since it only considers the description itself, without taking image information into account. The main goal of SPURTS is to detect fluency and style, and it can be written as

$$\text{median}_{\text{layer}}\text{max}_{\text{head}}\ I_{\text{flow}}(y_{w/o}, \theta), \tag{2}$$

where I_{flow}, which Feinglass and Yang (2021) refer to as information flow, is normalized mutual information as defined in Witten et al. 2005. For an input text without stop words, $y_{w/o}$, and a Transformer with parameters θ (Vaswani et al., 2017), SPURTS computes the information flow for each Transformer head at each layer, and then returns the layer-wise median of the head-wise maxima.

[2] E.g., https://webaim.org/techniques/alttext/

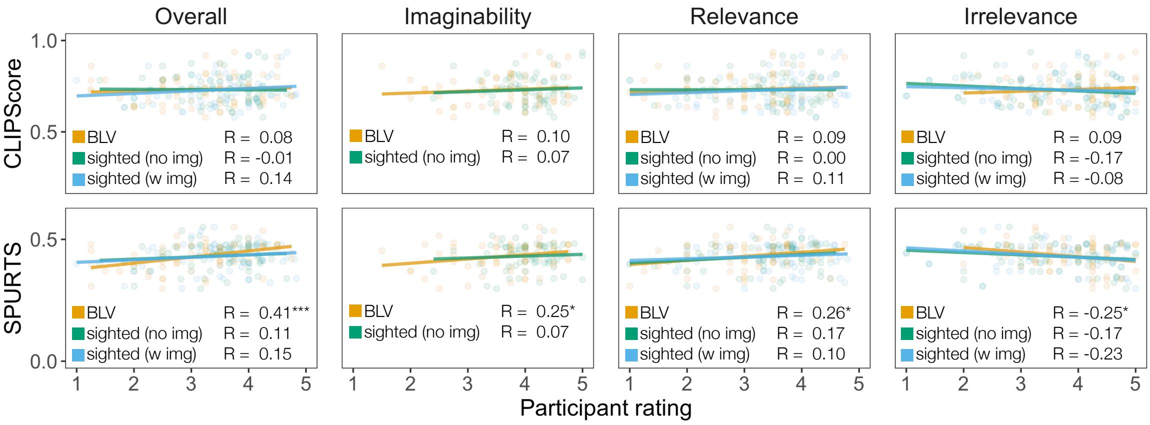

Figure 6: Correlations of CLIPScore (top row) and SPURTS (bottom row) with ratings of sighted and BLV participants. The Pearson correlations are computed over the human evaluators' average per-description rating. Sighted participants responded to the questions twice; once without (in green) and after seeing the image (in blue). There are no significant correlations between CLIPScore and human ratings. SPURTS correlates significantly with all responses provided by BLV participants but negatively with *Irrelevance* and not at all with sighted participant ratings, indicating a fundamental mismatch.

4.1 Compatibility

We first inspect the extent to which current referenceless metrics can capture whether a description is true for an image. SPURTS provides scores independent of the image and therefore inherently can't capture any notion of truthfulness. In contrast, CLIPScore is trained to distinguish between fitting and non-fitting image–text pairs, returning a compatibility score. We test whether this generalizes to our experimental data by providing CLIPScore with the true descriptions written for each image and a shuffled variant where images and descriptions were randomly paired. As Figure 5A demonstrates, CLIPScore rates the ordered pairs significantly higher compared to the shuffled counterparts ($\beta = 2.02$, SE $= .14$, $p < .001$),[3] suggesting that it captures image–text compatibility.

4.2 Description Length Correlation

Since the length of the description can already account for some of the variance of the BLV ratings, we further investigate whether description length is a general predictor for CLIPScore and SPURTS (see Figure 5B). For CLIPScore, description length doesn't correlate with predicted quality of the description. This is likely a consequence of the contrastive learning objective, which only optimizes for compatibility but not quality. SPURTS scores, in contrast, significantly correlate with description length, which is aligned with the BLV ratings.

4.3 Context Sensitivity

Crucially, the descriptions were written and evaluated within contexts, i.e., their respective Wikipedia article, and previous work suggests that the availability of context should affect what constitutes a good and useful description. Since current referenceless metrics can't integrate context, we expect that they shouldn't be able to capture the variation in the human description evaluations, and this is indeed what we find.

To investigate this hypothesis, we correlated sighted and BLV description evaluations with the CLIPScore and SPURTS ratings. As shown in Figure 6, CLIPScore fails to capture any variation observed in the human judgments across questions. This suggests that, while CLIPScore can add a perspective on the compatibility of a text for an image, it can't get beyond that as an indication of how useful a description is if it's true for the image.

Like CLIPScore, SPURTS scores don't correlate with the sighted participant judgments (Figure 6, bottom). However, specifically with respect to the overall rating, SPURTS scores show a significant correlation with the BLV participant ratings. While this seems encouraging, further analysis revealed that this correlation is primarily driven by the fact that both BLV and SPURTS ratings correlate with description length. The explained variance of the BLV ratings from description length alone is 0.152 and SPURTS score alone explains 0.08 of the variance. In conjunction, however, they only explain 0.166 of the variance, which means that most of

[3] Result from a linear effects analyses where the shuffled condition is coded as 0, and the ordered condition as 1.

		Overall	Imagin.	Relev.	Irrelev.
BLV	CLIPScore	0.075	0.104	0.086	0.090
	+Context	0.201	0.182	0.202	0.142
Sighted,	CLIPScore	−0.013	0.064	0.000	−0.166
no img	+Context	0.238	0.315	0.190	−0.019
Sighted,	CLIPScore	0.139		0.106	−0.079
w img	+Context	0.331		0.240	0.052

Table 1: Comparison of the human rating correlations with the original context-independent CLIPScore and the context-sensitive adaptation, using the same CLIP embeddings. Missing cells were not experimentally measured by design (Section 3.3.1). Across questions and participant groups, correlations improve. The CLIPScore correlations are a replication of Figure 6.

the predictability of SPURTS is due to the length correlation. This is further supported by a mixed effects linear regression analysis in which we fail to find a significant effect of SPURTS ($\beta = .80$, SE $= .44$, $p > .05$) once we include length as a predictor ($\beta = .64$, SE $= .15$, $p < .001$).[4]

A further indication that SPURTS isn't capturing essential variance in BLV judgments is apparent from the negative correlation in the *Irrelevance* question ($R = -0.25$). This suggests that SPURTS scores tend to be higher for descriptions that are judged to contain too much irrelevant information and low when participants assess the level of information to be appropriate. In the BLV responses, *Irrelevance* is positively correlated with the *Overall* ratings ($R = 0.33$), posing a clear qualitative mismatch to SPURTS. Since what is considered extra information is dependent on the context, this is a concrete case where the metric's lack of context integration results in undesired behavior.

Finally, SPURTS' complete lack of correlation with sighted participant judgments further suggests that SPURTS is insufficient for picking up the semantic components of the descriptions. This aligns with the original conception of the metric, where a reference-based metric (SPARCS) is used to estimate semantic quality.

Overall, our results highlight that SPURTS captures the BLV participants' preferences for longer descriptions but falls short in capturing additional semantic preferences, and is inherently inadequate for judging the truthfulness of a description more generally. CLIPScore can't capture any of the variation in BLV or sighted participant ratings, uncovering clear limitations.

5 The Potential for Integrating Context into CLIPScore

Can referenceless metrics like CLIPScore be made context sensitive? To begin exploring this question, as a proof of concept, we amend (1) as follows:

$$\overline{\text{description}} \cdot \text{context} + \text{description} \cdot \left(\overline{\text{image}} - \overline{\text{context}}\right) \quad (3)$$

Here, quality is a function of (a) the description's similarity to the context (first addend) and (b) whether the description captures the information that the image adds to the context (second addend). These two addends can be seen as capturing aspects of (ir)relevance and imaginability, respectively, though we anticipate many alternative ways to quantify these dimensions.

Table 1 reports correlations between this augmented version of CLIPScore and our sighted and BLV participant judgments. We find it encouraging that even this simple approach to incorporating context boosts correlations with human ratings for all the questions in our experiment. For the *Irrelevance* question, it even clearly captures the positive correlation with BLV ratings, which is negative for both CLIPScore and SPURTS, indicating a promising shift. We consider this an encouraging signal that large pretrained models such as CLIP might still constitute a resource for developing future referenceless metrics.

However, despite these promising signs, there are also reasons to believe that CLIP-based metrics have other restrictive limitations. Due to CLIP's training, images are cropped at the center region and texts need to be truncated at 77 tokens (Radford et al., 2021). CLIP relies on embeddings learned for each absolute token position in the text window and each patch position in the image. These can therefore not be easily extended to avoid any context or image cropping that is currently limiting CLIPScore. Specifically for the purpose of accessibility, the information this removes can be crucial for determining whether a description is useful or not. For instance, our experiments show that the length of a description is an important indicator for description quality – information lost in CLIP-based metrics. Moreover, this disproportionately affects the ability to encode the context paragraphs,

[4] We assume random intercepts by participant and description, and we rescaled description length into $[0, 1]$.

which are often longer than a typical description. These decisions are therefore likely reflected in any resulting metric and should therefore be reconsidered when devising a new metric.

6 Conclusion

The context an image appears in shapes the way high-quality accessibility descriptions are written. In this work, we reported on experiments in which we explicitly varied the contexts images were presented in and investigated the effects of this contextual evaluation on participant ratings. These experiments reveal strong contextual effects for sighted and BLV participants. We showed that this poses a serious obstacle for current referenceless metrics, but we also see promise for future efforts since the inclusion of context to a prominent metric such as CLIPScore begins to address the disconnect from BLV needs.

Limitations and Ethics

Our investigation focuses on whether a description is judged as fulfilling the purpose of accessibility and is therefore entirely based on utility considerations. However, generated image-based texts can also differ in stylistic qualities such as grammaticality. Kasai et al. (2022a) suggest a human annotation scheme that focuses on such dimensions, which together with our work provides a broad assessment of description quality.

To investigate the effect of context, we chose an experimental design where the same image can be placed in a variety of contexts and vice versa. Consequently, the images were complementary to the text, but the text could easily be understood without the image as well. Web accessibility guides suggest that only important images should receive alt descriptions and purely decorative images shouldn't receive any. Our image–context pairs are in between these extremes, and future work should explore how these effects vary depending on image–context relations. Similarly, we looked only at Wikipedia articles for providing context, but previous work has argued for paying attention to intricate differences between domains. While we expect the observed context effects to carry over to other domains such as social media, this is a matter for future investigation.

As our primary human evaluation method, we used 5-point Likert scales where a higher rating corresponded to a higher quality description across dimensions. Recently, Ethayarajh and Jurafsky (2022) argued against using Likert scales for comparing the performance of two systems on natural language generation tasks. Likert scales indeed come with challenges since the interpretation of the intervals is likely variable and asymmetric, posing a challenge for analysis (Jamieson 2004, but see Carifio and Perla 2008, Norman 2010). However, Likert scales are better supported in BLV optimized interfaces such as Google Forms, and were therefore chosen to allow a direct comparison between BLV and sighted participant judgments. To minimize potential artifacts due to the scale, we obtained multiple ratings from each participant and included by-participant random effects in the statistical analyses. Though not without challenges, Likert scales still provide the best method for quality assessments in accessibility-oriented comparisons.

All of our human subject experiments were conducted under IRB protocols. Most sighted participants spent between 20 and 30 minutes on the study and were paid \$6.15 (\$12.30–18.45/hr) over Amazon's Mechanical Turk. Most BLV participants completed the experiment between 1.5 and 2.5 hours (based on self reporting) and were paid \$75 in Amazon gift cards (\$30–50/hr, other gift cards being available upon request). The BLV study was thoroughly tested for its accessibility before it was distributed. All data were completely anonymized before analysis.

Acknowledgements

This work is supported in part by a grant from Google through the Stanford Institute for Human-Centered AI and by the NSF under project REU-1950223. We thank our experiment participants for their invaluable input. We further thank Gabriel Poesia, Rachit Dubey, and Zhengxuan Wu for helpful discussions and comments.

References

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. ZSCC: 0000514.

Jeffrey P. Bigham, Ryan S. Kaminsky, Richard E. Ladner, Oscar M. Danielsson, and Gordon L. Hempton. 2006. WebInSight:: making web images accessible. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility -*

Assets '06, page 181, Portland, Oregon, USA. ACM Press.

Ali Furkan Biten, Lluis Gomez, Marcal Rusinol, and Dimosthenis Karatzas. 2019. Good News, everyone! Context driven entity-aware captioning for news images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12458–12467, Long Beach, CA, USA. IEEE.

Maria Claudia Buzzi, Marina Buzzi, and Barbara Leporini. 2011. Web 2.0: Twitter and the blind. In *Proceedings of the 9th ACM SIGCHI Italian Chapter International Conference on Computer-Human Interaction: Facing Complexity*, CHItaly, pages 151–156, New York, NY, USA. Association for Computing Machinery.

James Carifio and Rocco Perla. 2008. Resolving the 50-year debate around using and misusing Likert scales. *Medical education*, 42(12):1150–1152.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Kawin Ethayarajh and Dan Jurafsky. 2022. How human is human evaluation? Improving the gold standard for NLG with utility theory. *arXiv preprint arXiv:2205.11930*.

Joshua Feinglass and Yezhou Yang. 2021. SMURF: SeMantic and linguistic UndeRstanding Fusion for caption evaluation via typicality analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2250–2260.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

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 on - WWW '19*, pages 549–559, San Francisco, CA, USA. ACM Press.

Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter A11y: A browser extension to make Twitter images accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. 2018. Caption Crawler: Enabling reusable alternative text descriptions using reverse image search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–11, Montreal QC, Canada. ACM Press.

Stephanie Hackett, Bambang Parmanto, and Xiaoming Zeng. 2003. Accessibility of Internet websites through time. In *Proceedings of the 6th international ACM SIGACCESS conference on Computers and accessibility*, Assets '04, pages 32–39, New York, NY, USA. Association for Computing Machinery.

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.

Susan Jamieson. 2004. Likert scales: How to (ab) use them? *Medical education*, 38(12):1217–1218.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. 2022a. Transparent human evaluation for image captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3464–3478, Seattle, United States. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander Fabbri, Yejin Choi, and Noah A. Smith. 2022b. Bidimensional leaderboards: Generate and evaluate language hand in hand. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3540–3557, Seattle, United States. Association for Computational Linguistics.

Elisa Kreiss, Fei Fang, Noah D. Goodman, and Christopher Potts. 2022. Concadia: Towards image-based text generation with a purpose. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Hwanhee Lee, Thomas Scialom, Seunghyun Yoon, Franck Dernoncourt, and Kyomin Jung. 2021a. QACE: Asking questions to evaluate an image caption. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4631–4638.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021b. 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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chi-kiu Lo. 2019. YiSi - A unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300. Place: Cambridge, MA Publisher: MIT Press.

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*, pages 5988–5999, Denver Colorado USA. ACM.

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*, pages 5506–5516.

Annika Muehlbradt and Shaun K. Kane. 2022. What's in an ALT tag? Exploring caption content priorities through collaborative captioning. *ACM Transactions on Accessible Computing*, 15(1):6:1–6:32.

Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5):625–632.

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.

Maxime Peyrard and Iryna Gurevych. 2018. Objective function learning to match human judgements for optimization-based summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 654–660, New Orleans, Louisiana. 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.

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*, pages 1–13, Honolulu HI USA. ACM.

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 *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–15.

Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2018. BSpeak: An accessible voice-based crowdsourcing marketplace for low-income blind people. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA. Association for Computing Machinery.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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. ZSCC: 0001489.

Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How blind people interact with visual content on social networking services. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 1584–1595.

Ian H Witten, Eibe Frank, Mark A Hall, Christopher J Pal, and MINING DATA. 2005. Practical machine learning tools and techniques. In *DATA MINING*, volume 2, page 4.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

Appendix

A Guide to Supplementary Materials

All data and code needed to replicate our results, along with additional analyses on the human subject data, are made available.[5] Specifically, all studies can be completed the exact way they appeared to participants, and our data and code enable replications of our graphs and statistical analyses, and provide further insights into the participant distributions, comments, and other additional analyses. Our materials also contain the code for the presented referenceless metrics adapted to our specific data. The README provides further details and points to the folders and files necessary to replicate our results.

B Implications for Other Referenceless Metrics

In the previous experiments, we established that the referenceless metrics CLIPScore and SPURTS can't get traction on what makes a good description when the images and descriptions are contextualized. Other referenceless metrics such as UMIC (Lee et al., 2021b) and QACE (Lee et al., 2021a) face the same fundamental issue as CLIPScore and SPURTS due to their contextless nature. Like CLIPScore, UMIC is based on an image–text model (UNITER; Chen et al. 2020) trained under a contrastive learning objective. Similarly, it produces an image–text compatibility score solely by receiving a decontextualized image and text as input. QACE uses the candidate description to derive potential questions that should be answerable based on the image. The evaluation is therefore whether the description mentions aspects that are true of the image and not about which aspects of the image are relevant to describe. This again only provides insights into image–text compatibility but not contextual relevance. Unfortunately, we are unable to provide quantitative results for these referenceless metrics since the authors haven't provided the code necessary (QACE), or the code relies on image features that can't be created for novel datasets with currently available hardware (UMIC, QACE).

In summary, the current context-independence of all existing referenceless metrics is a major limitation for their usefulness. This is a challenge that needs to be addressed to make these metrics a useful tool for advancing image-based NLG systems.

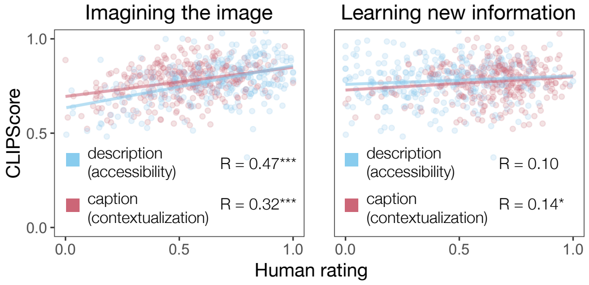

Figure 7: CLIPScore provides higher ratings for image-based texts that capture image content well. Whether the descriptions and captions provide additional information to the image content doesn't affect the ratings.

C Referenceless Metrics for Image-Based NLG Beyond Accessibility

While we have specifically focused on the usefulness of referenceless metrics for image accessibility, this isn't the only potential purpose an image-based text might address. Kreiss et al. (2022) distinguish *descriptions*, i.e., image-based texts that are written to replace the image, and *captions*, i.e., texts that are intended to appear alongside images, such as tweets or newspaper captions. This suggests that the same text can be very useful for contextualizing an image but fail at providing image accessibility, and vice versa. To investigate this distinction, they asked participants to rate `alt` descriptions as well as image captions from Wikipedia according to (1) how much the text helped them imagine the image, and (2) how much they learned from the text that they couldn't have learned from the image. Descriptions were rated more useful for imagining the image, whereas captions were rated more useful for learning additional information. Captions used for contextualizing an image might therefore be another potential use domain for a referenceless metric such as CLIPScore.

To see whether CLIPScore might be a promising resource for evaluating captions, we obtained CLIPScore ratings for the descriptions and captions in Kreiss et al. (2022). CLIPScore ratings correlate with the reconstruction as opposed to the contextualization goal (see Figure 7), suggesting that CLIPScore is inherently less appropriate to be used for assessing caption datasets. This aligns with the original observation in Hessel et al. (2021) that CLIPScore performs less well on the news caption dataset GoodNews (Biten et al., 2019) compared

[5] https://github.com/elisakreiss/contextual-description-evaluation

to MSCOCO (Hessel et al., 2021), a contextless description dataset.

Taken together, this is further evidence that the "one-size-fits-all" approach to referenceless image-based text evaluation is not sufficient for adequately assessing text quality for the contextualization or the accessibility domain.