CLIPScore: A Reference-free Evaluation Metric for Image Captioning

Jack Hessel[†] **Ari Holtzman**[‡] **Maxwell Forbes**[‡] **Ronan Le Bras**[†] **Yejin Choi**[†‡]
[†]Allen Institute for AI
[‡]Paul G. Allen School of Computer Science & Engineering, University of Washington
{jackh,ronanlb}@allenai.org {ahai,mbforbes,yejin}@cs.washington.edu

Abstract

Image captioning has conventionally relied on *reference-based* automatic evaluations, where machine captions are compared against captions written by humans. This is in contrast to the *reference-free* manner in which humans assess caption quality.

In this paper, we report the surprising empirical finding that CLIP (Radford et al., 2021), a cross-modal model pretrained on 400M image+caption pairs from the web, can be used for robust automatic evaluation of image captioning without the need for references. Experiments spanning several corpora demonstrate that our new reference-free metric, CLIPScore, achieves the highest correlation with human judgements, outperforming existing reference-based metrics like CIDEr and SPICE. Information gain experiments demonstrate that CLIPScore, with its tight focus on *image–text* compatibility, is complementary to existing reference-based metrics that emphasize *text–text* similarities. Thus, we also present a reference-augmented version, RefCLIPScore, which achieves even higher correlation. Beyond literal description tasks, several case studies reveal domains where CLIPScore performs well (clip-art images, alt-text rating), but also where it is relatively weaker in comparison to reference-based metrics, e.g., news captions that require richer contextual knowledge.

1 Introduction

For most text generation tasks, reference-based n-gram overlap methods are still the dominant means of automatic evaluation. For image caption generation, recent reference-based metrics have sought to transcend overlap by considering richer models of reference-candidate similarity: e.g., approximate scene graphs (Anderson et al., 2016), allowing reference-based methods to incorporate the image (Jiang et al., 2019; Lee et al., 2020). But, references can be expensive to collect and comparing against even *multiple* human-authored captions for each image is often insufficient (see Figure 1). As a result, for many corpora, a significant gap remains between reference-based scoring and human quality judgments.[1]

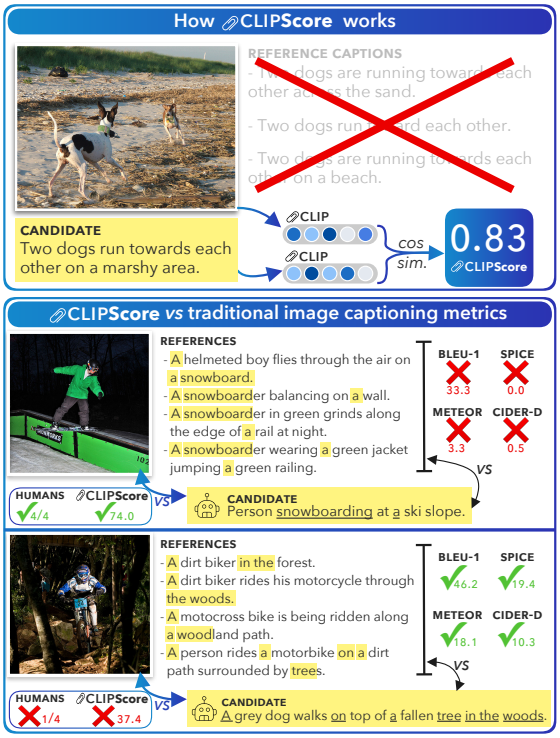


Figure 1: **Top:** CLIPScore uses CLIP to assess image-caption compatibility *without* using references, just like humans. **Bottom:** This frees CLIPScore from the well-known shortcomings of n-gram matching metrics, which disfavor good captions with new words (top) and favor *any* captions with familiar words (bottom). Attribution: Paperclip, robot icons by Hasanudin, Adiyogi (resp.) from the Noun Project.

Should we need references for the evaluation of image captions? After all, when humans assess the appropriateness of an image caption, we do so just by looking at the image and reading the candidate's text.

[1] See Elliott and Keller (2014) and Kilickaya et al. (2017) for thorough comparisons of caption generation metrics.

A recent trend in machine translation serves as inspiration: there, a key hurdle for reference-free evaluation (sometimes called *quality estimation*) has been estimating cross-lingual similarity between source+candidate pairs (Blatz et al., 2004; Specia et al., 2010; Mehdad et al., 2012; Specia and Shah, 2018). But recent work (Lo, 2019; Yankovskaya et al., 2019; Zhao et al., 2020) has improved correlation with human judgment not by gathering more monolingual references, but instead by utilizing cross-lingual representations learned by large-scale, pre-trained, multilingual models e.g., LASER (Artetxe and Schwenk, 2019) or M-BERT (Devlin et al., 2019). [2]

We hypothesize that the relationships learned by pretrained vision+language models (e.g., ALIGN (Jia et al., 2021) and CLIP (Radford et al., 2021)) could similarly support reference-free evaluation in the image captioning case. Indeed, they can: we show that a relatively direct application of CLIP to (image, generated caption) pairs results in surprisingly high correlation with human judgments on a suite of standard image description benchmarks (e.g., MSCOCO (Lin et al., 2014)). We call this process `CLIPScore` (abbreviated to CLIP-S). Beyond direct correlation with human judgments, an information gain analysis reveals that CLIP-S is complementary both to commonly reported metrics (like BLEU-4, SPICE, and CIDEr) and to newly proposed reference-based metrics (e.g., ViLBERTScore-F (Lee et al., 2020)).

We additionally (1) propose a reference-augmented version of `CLIPScore`, `RefCLIPScore`, that achieves even higher human correlation, (2) verify that CLIP-S is sensitive to adversarially constructed image captions, where one noun-phrase has been swapped for a plausible (but incorrect) distractor; and (3) construct a corpus of images that have never been posted publicly online to verify that CLIP-S is able to reconstruct human judgments on never-before-seen images.

Finally, we assess CLIP-S in the context of four case studies that diverge from context-free, literal photograph description. In two cases, CLIP-S works well: it achieves high correlation with alt-text quality rating on Twitter, and demonstrates surprising capacity to reason about clipart images+captions. For news caption generation, reference-based methods correlate best with human judgments. And, for emotive captions inspired by language use on social media, even reference-based metrics fall short.

2 Related Work

Reference-only image caption evaluation In general, image caption generation models are evaluated by a suite of 5 reference based metrics: BLEU-4 (Papineni et al., 2002) (which measures a version of precision between a candidate and the references), ROUGE-L (Lin, 2004) (which measures a version of recall), METEOR (Banerjee and Lavie, 2005) (which computes a word-level alignment), CIDEr (Vedantam et al., 2015) (which combines n-gram tf-idf weighting and stemming) and SPICE (Anderson et al., 2016) (which applies a semantic parser to a set of references, and computes similarity using the predicted scene graph).[3] Yi et al. (2020) give a method for re-weighting BERTScore (Zhang et al., 2020) specifically tuned to the image caption generation domain (we refer to their method as BERT-S++).

Reference+image caption evaluation Recent metrics incorporate image-text grounding features in addition to references: TIGEr (Jiang et al., 2019) uses a pretrained SCAN model (Lee et al., 2018), and ViLBERTScore-F (Lee et al., 2020) uses a pretrained ViLBERT model (Lu et al., 2019) that is also fine-tuned on 12 downstream vision and language tasks (Lu et al., 2020). Our work provides perspective on the next logical extension: instead of incorporating visual-textual interactions in addition to references, can we ignore references entirely?

Self-retrieval for image captioning Prior works have proposed incorporating a *self-retrieval* loss into caption generation, with the intuition that good captions should be able to uniquely identify their images with high accuracy (Dai and Lin, 2017; Luo et al., 2018; Liu et al., 2018); monitoring this type of loss can provide insight into how distinctive the captions are according to the model itself. CLIP-S is similar in spirit, but distinct for its utility as an extrinsic evaluation metric like BLEU-4 or CIDEr.

Reference-free evaluation In addition to the machine translation cases highlighted in the introduction, reference-free evaluations have been proposed for other generation tasks, including summarization

[2] K et al. (2020), Pires et al. (2019), and Wu and Dredze (2019) explore how M-BERT learns and utilizes cross-lingual information.

[3] For comparison with these metrics, we use the standard COCO evaluation tools available at `https://github.com/tylin/coco-caption`.

(Louis and Nenkova, 2013; Peyrard and Gurevych, 2018; Sun and Nenkova, 2019) and dialogue (Tao et al., 2018; Mehri and Eskenazi, 2020). These metrics can be supervised, relying on human judgments for quality estimation, or less-supervised, relying on pre-trained model representations. For image captioning, a version of VIFIDEL (Madhyastha et al., 2019) was proposed for reference-free evaluation; however, VIFIDEL, computed based on a list of detected objects in the image from a fixed object vocabulary, generally produces less correlation with human ratings vs. reference-based metrics.

3 CLIPScore

Model Details. CLIP (Radford et al., 2021) is a cross-modal retrieval model trained on 400M (image, caption) pairs gathered from the web. 500K search queries, consisting of common unigram/bigrams, named entities, etc., were executed on a search engine. For each query, up to 20K (image, caption) pairs were collected.

The model we use is the `ViT-B/32` version.[4] It represents images via a Vision Transformer (Vaswani et al., 2017; Dosovitskiy et al., 2021), which forgoes convolutional filters in favor of self-attention maps computed between a 7 by 7 grid of image patches, which evenly divides a 224 by 224 pixel input image. This model has 12 transformer layers and 86M parameters. The text is similarly represented by a 12-layer transformer trained over a vocab of 49K BPE token types (Sennrich et al., 2016) (and is more fully described in Radford et al. (2019)). Both the text and image networks output a single vector; these vectors aim to represent the content of an input caption or an image, respectively. In the case of `ViT-B/32`, these vectors are 512-D. The model's weights are trained to maximize the scaled cosine similarity of truly corresponding image/caption pairs while simultaneously minimizing the similarity of mismatched image/caption pairs using InfoNCE (Sohn, 2016; Oord et al., 2018). We hold fixed this set of weights for our experiments.

Evaluating Caption Generations with CLIP. To assess the quality of a candidate generation, we pass both the image and the candidate caption through their respective feature extractors. Then, we compute the cosine similarity of the resultant embeddings.[5] We found that prefixing candidates with the prompt: "A photo depicts" improved correlations slightly (and is our recommended/standard configuration), though "A photo of", the recommended prompt from Radford et al. (2021), worked well too. Following Zhang et al. (2020), we perform a re-scaling operation.[6] For an image with visual CLIP embedding v and a candidate caption with textual CLIP embedding c, we set $w = 2.5$ and compute CLIP-S as:

$$\texttt{CLIP-S}(\mathbf{c}, \mathbf{v}) = w * \max(cos(\mathbf{c}, \mathbf{v}), 0)$$

To compute corpus-level CLIP-S, we simply average over (candidate, image) pairs. Note that this evaluation *does not depend on underlying references.* The runtime of CLIP-S with the `ViT-B/32` backbone is fast: on our single consumer GPU and hard drive, roughly 4K image-candidate pairings can be processed per minute.

RefCLIPScore CLIP-S can additionally be extended to incorporate references, if they are available. We extract vector representations of each available reference by passing them through CLIP's text transformer; the result is the set of vector representation of all references, R. Then, `RefCLIPScore` is computed as a harmonic mean of CLIP-S, and the maximal reference cosine similarity, i.e.,

$$\texttt{RefCLIP-S}(\mathbf{c}, \mathbf{R}, \mathbf{v}) =$$
$$\text{H-Mean}(\texttt{CLIP-S}(\mathbf{c}, \mathbf{v}), \max(\max_{\mathbf{r} \in \mathbf{R}} cos(\mathbf{c}, \mathbf{r}), 0))$$

4 Benchmark Captioning Evaluations

We first evaluate on a set of literal description corpora. Broadly, the captions in these corpora aim to identify and highlight the literal, salient objects/actions in a photographic image, presented without additional context.[7]

[4] We expect that more powerful, larger versions of the model, if released at a later date, could perform better.

[5] More sophisticated CLIP configurations, e.g., region-level/token-level correspondence models, did not achieve better performance.

[6] While the cosine similarity, in theory, can range from $[-1, 1]$ (1) we never observed a negative cosine similarity; and (2) we generally observe values ranging from roughly zero to roughly .4. The particular value of w we advocate for, $w = 2.5$, attempts to stretch the range of the score distribution to $[0, 1]$. For more details and justification for our re-scaling, including a demonstration of generality across several corpora, see Appendix B).

[7] See Berg et al. (2012) for a statistical exploration of salience in a such a corpus.

	τ_c
BLEU-1	32.3
BLEU-4	30.8
ROUGE-L	32.3
BERT-S (RoBERTa-F)	39.2
METEOR	41.8
CIDEr	43.9
SPICE	44.9
LEIC (τ_b)* (Cui et al., 2018)	46.6
BERT-S++ (Yi et al., 2020)	46.7
TIGEr (Jiang et al., 2019)	49.3
NUBIA * (Kane et al., 2020)	49.5
ViLBERTScore-F (Lee et al., 2020)	50.1
CLIP-S (no refs)	51.2
RefCLIP-S	**53.0**

Table 1: Flickr8K-Expert correlations with human judgment. All metrics use 4-5 ground truth references, except for CLIP-S (which uses none). * indicates a result reported in prior work.

	τ_b
BLEU-4	16.9
CIDEr	24.6
METEOR	22.2
ROUGE-L	19.9
SPICE	24.4
BERT-S (RoBERTa-F)	22.8
LEIC *	29.5
CLIP-S (no refs)	34.4
RefCLIP-S	**36.4**

Table 2: Flickr8K-CF correlations with human judgment. * indicates a result reported in prior work.

4.1 Caption-level likert judgments

We first explore three corpora consisting of human likert-scale judgments at the level of individual image/caption pairs. Flickr8K-Expert (Hodosh et al., 2013) contains 17K "expert" human judgments between 5664 images: humans graded captions on a scale of 1 to 4 (4="caption describes the image without any errors"; 1="caption is unrelated to the image"). Flickr8K-CF is a set of 145K binary quality judgments gathered from CrowdFlower over 48K (image, caption) pairs (1K unique images). Each pair has at least 3 binary judgments, and we take the mean proportion of "yes" annotations as a score for each pair to compute correlations.

Composite (Aditya et al., 2015) contains 12K human judgments between images from MSCOCO (2007 images), Flickr8k (997 images), and Flickr30k (Young et al., 2014) (991 images). Each image originally has five references, but one of the references was selected to be rated by humans in the set (and so we remove it from the reference set when computing metrics; this differs from some prior work, see Appendix A for why we consider the more difficult setting). For Composite and Flickr8K judgments, we compute correlation between each metric and the human ratings using Kendall τ.

Results The results for Flickr8K-Expert are given in Table 1, for Flickr8K-CF are given in Table 2 (in τ_b, following Cui et al. (2018)), and for Composite are given in Table 3. For the caption-level corpora we consider, CLIP-S without references achieves higher correlation with human judgment compared to previously proposed metrics that rely on references. Additionally, in all cases, RefCLIP-S improves correlation even further. This provides strong evidence that, in terms of correlating with human judgment at the caption-level for these literal photographic image description tasks, a relatively direct application of CLIP can serve as a strong automatic evaluation metric.

4.2 Pairwise ranking on Pascal-50S

In Pascal-50S (Vedantam et al., 2015), raters made pairwise preference judgments between pairs of sentences. There are 4K sentence pairs total, split evenly across four categories, e.g., two human captions, two machine captions, etc. For each pair, 48 human pairwise judgments were gathered.[8] Following prior work, instead of computing correlation coefficients, we compute accuracy, i.e., we consider the caption preferred by a majority of annotators to be correct, and measure how often the evaluation metric assigns a higher score to that member of the pair. Ties are broken randomly. Due to random selection of 5 references among the 48 candidates to serve as ground-truth for the reference-based metrics, the results may differ slightly from prior work (we average over 5 random draws of references).

The results are given in Table 4. Evaluation is split across four categories of caption pairs (detailed in the table caption). CLIP-S and RefCLIP-S generally achieve high performance in all categories.

[8]Instead of being presented with the image, annotators were presented only with a reference (and the two candidates to rank).

	τ_c
BLEU-1	31.3
BLEU-4	30.6
ROUGE-L	32.4
BERT-S (RoBERTa-F)	30.1
METEOR	38.9
CIDEr	37.7
SPICE	40.3
BERT-S++ *	44.9
TIGEr	45.4
ViLBERTScore-F	52.4
CLIP-S (no refs)	53.8
RefCLIP-S	**55.4**

Table 3: Composite correlations with human judgment. All metrics use between 4 and 5 ground truth references, except for CLIP-S (which uses none). In contrast to some prior work, we consider a harder setting, and remove the candidate from the reference set (see Appendix A for details; for comparison purposes, RefCLIP-S achieves $\tau_c = 60.0$ in the easier setting). * indicates a result reported in prior work.

	HC	HI	HM	MM	Mean
length	51.7	52.3	63.6	49.6	54.3
BLEU-4	60.4	90.6	84.9	54.7	72.6
SPICE	63.6	96.3	86.7	68.3	78.7
METEOR	63.8	97.7	93.7	65.4	80.1
ROUGE-L	63.7	95.3	92.3	61.2	78.1
CIDEr	65.1	98.1	90.5	64.8	79.6
BERT-S (RoBERTa-F)	65.4	96.2	93.3	61.4	79.1
TIGEr *	56.0	**99.8**	92.8	74.2	80.7
ViLBERTScore-F *	49.9	99.6	93.1	**75.8**	79.6
BERT-S++ *	**65.4**	98.1	**96.4**	60.3	80.1
CLIP-S (no refs)	56.5	99.3	**96.4**	70.4	80.7
RefCLIP-S	64.5	99.6	95.4	72.8	**83.1**

Table 4: Pascal50S accuracy results (5 references). HC = two human correct captions; HI = both captions are human written, but one is wrong; HM = both captions are for the image, but one is written by a human, one by an algorithm; MM = both captions are for the image, and both are written by an algorithm. * indicates a result reported in prior work: the comparability of our results to *-rows is subject to the (arbitrary) sample of references. We average our results over 5 random samples (but CLIP-S doesn't change because it doesn't use references).

4.3 System-level correlation for MSCOCO

CLIP-S achieves high correlation with human judgments at the system-level as well: we evaluate the outputs of systems submitted to the 2015 MSCOCO Image Captioning Challenge (Vinyals et al., 2016). We have some concerns with standard evaluation setup on this corpus, mostly related to the fact that it consists of only 12 datapoints (see supplementary for more discussion). Nonetheless, following the standard procedure, we correlate CLIP-S and RefCLIP-S with two metrics: "the percentage of captions that are evaluated as better or equal to a human caption (M1)" and percentage of captions that pass the "Turing Test" (M2), respectively. CLIP-S achieves Spearman $\rho_{M1}/\rho_{M2} = .59/.63$ and RefCLIP-S achieves $\rho_{M1}/\rho_{M2} = .69/.74$ (all $p < .05$) with these system-level metrics.

4.4 Sensitivity of CLIP-S to hallucination

Prior work has demonstrated that, for many literal description tasks, humans often prefer *correctness* in captions over specificity (Rohrbach et al., 2018, 2017).[9] Thus, understanding if and how evaluation metrics handle image captions that contain incorrect "hallucinations," e.g., references to objects that are not depicted, is important. We use a sample of image captions from the FOIL dataset, constructed by Shekhar et al. (2017), to test how sensitive CLIP-S is to detecting potentially subtle inaccurate details in descriptions. This corpus consists of modified reference captions from MSCOCO that have a single noun-phrase adversarially swapped out to make the FOIL caption incorrect, e.g., switching "motorcycle" for "bicycle".

To adapt the corpus to our setting, for each of the 32K test images, we sample a (FOIL, true) pair, and compute the accuracy of each evaluation metric in their capacity to assign a higher score to the true candidate versus the FOIL. To compute reference-based metrics, we give access to the MSCOCO reference captions for the image (excluding the the true candidate being assessed against the FOIL). While the paired setting we consider isn't identical, Shekhar et al. (2017) estimate roughly 92% human agreement on the unpaired version of the task, relative to a 50/50 random guessing baseline.

Table 5 contains the results. In this setting, having access to more annotation is quite helpful for reference based metrics, e.g., the accuracy of SPICE and BLEU-4 increase by over ten points when shifting from one to four references. But in the reference-limited setting, CLIP-S, without *any* ref-

[9] This is not always the case: MacLeod et al. (2017) show there is a range of opinion among a sample of low vision and blind users of social media.

	1-ref	4-ref
length	50.2	50.2
BLEU-4	66.5	82.6
METEOR	78.8	85.4
ROUGE-L	71.7	79.3
CIDEr	82.5	90.6
SPICE	75.5	86.1
BERT-S (RoBERTa-F)	88.6	92.1
CLIP-S (no refs)	87.2	87.2
RefCLIP-S	**91.0**	**92.6**

Table 5: Accuracy of evaluation metrics in the pairwise FOIL hallucination detection setting. All reference-based metrics are given access to either one or four references.

erence outperforms all metrics except for BERT-S (RoBERTa-F). And, RefCLIP-S works best in all cases.

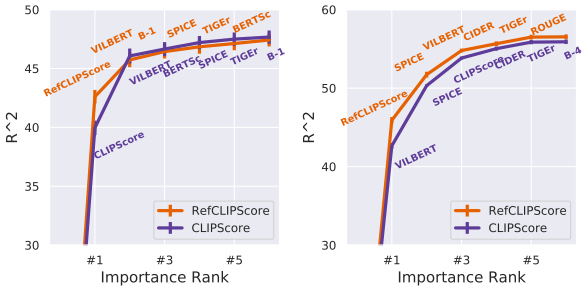
Overall, we corroborate Rohrbach et al. (2018)'s finding that "object hallucination can not be always predicted based on the traditional sentence metrics" using a corpus derived from Shekhar et al. (2017), particularly in the case where there are few references available. However, CLIP-S and RefCLIP-S offer a performance improvement in the pairwise setting.

4.5 Sensitivity of CLIP-S to memorization

One concern with model-based scoring methods is memorization, i.e., if a model's weights are pretrained using a large corpus, there's a risk that data used at evaluation time have already been seen at pretraining time. While Radford et al. (2021) conduct a train-test overlap analysis and find that CLIP is unlikely to succeed because of memorization, we nonetheless conduct an experiment with images CLIP has never seen before.

The authors of this work created a set of 250 images that have never been posted to the Internet by aggregating personal photographs. The set contains a variety of Flickr-like situations, e.g., nature scenes, animals, city streets, objects, etc. For each image, we collect two automatically generated captions: one from a commercial API, Microsoft Azure Cognitive Services (v 3.1)[10] and one from Luo et al. (2018)'s pretrained model, which is trained to maximize CIDEr score with a self-critical

[10] https://azure.microsoft.com/en-us/services/cognitive-services/



(a) Composite (b) Flickr8k-Expert

Figure 2: R^2 for the forward-selection regression of metrics on human Likert ratings for two corpora. Foward-selection tends to identify both **CLIP-S** and **RefCLIP-S** early-on: other informative and complementary metrics include ViLBERTScore-F and SPICE.

baseline.[11] Then, for each image, three authors of this work independently selected which caption described the image content more accurately. Relative to a 50% random baseline (and a 72% length baseline of selecting the *shorter* caption) CLIP-S correctly recovers majority human preference in 86% of cases. Human agreement for this corpus is 93%.[12]

While this setup cannot definitively refute the notion that CLIP works well because it has memorized images, we hope the results here contribute to the evolving discussion about the nature of generalization for web-scale pretrained models.

4.6 Which metrics should I report?

Most caption generation works report multiple metrics, each of which (presumably) correlates with human judgment to different degrees. But it's not always clear if individual metrics capture distinct or redundant dimensions of human judgment. For example, while CLIP-S and ViLBERTScore-F both produce high correlations, are they redundant or complementary?

We seek a (minimal) set of metrics that explains the most variance in human judgment. To find this set, we undertake a forward selection on a set of ten candidate metrics comprising six widely-reported metrics,[13] and four newer metrics, BERT-S (RoBERTa-F), TIGEr, ViLBERTScore-F, and CLIP-S (we also include experiments starting with RefCLIP-S instead of CLIP-S, too). Starting from an empty set, we perform an iterative greedy selection by picking

[11] We use the ResNet101 pretrained version, which achieves 1.05 CIDEr and 0.19 SPICE on the COCO validation set.

[12] Raters preferred the Microsoft captions to the ResNet101 model 81% of the time.

[13] BLEU-1, BLEU-4, METEOR, CIDEr, ROUGE-L, SPICE

the most informative additional metric to add.[14] To estimate variance, we repeat the forward-selection process 10 times with bootstrap re-sampled versions of the corpus.

Figure 2 shows the information gain that results from running this experiment on the Composite and Flickr8K-Expert corpora; we also show which metric is most commonly selected at each iteration (earlier = more information gain). For Composite, CLIP-S (or RefCLIP-S) is always selected first, followed by ViLBERTScore-F, and then (most commonly) BERT-S (RoBERTa-F). For Flickr8k-Expert, the top three choices are always CLIP-S (or RefCLIP-S), ViLBERTScore-F, and SPICE. While CLIP-S and ViLBERTScore-F tend to be the most informative metrics, (1) while they are correlated, they are not purely redundant; and (2) image-unaware, reference-based metrics like SPICE can still be useful.

In summary, these results suggest that evaluation metrics like CLIP-S, which take into account visual content, indeed capture axes of human judgment not currently covered by text-only reference-based metrics. *For the literal image description evaluation settings we consider, a reasonable mix of metrics to report is at least one image-aware metric (e.g., CLIP-S) plus a strong reference-only metric (e.g., SPICE).*

5 Case Studies Using `CLIPScore`

Our results thus far have demonstrated that CLIP encodes information useful for evaluating literal image description tasks. But, reference-based metrics may *a priori* seem more adaptable versus CLIP-S. Does CLIP-S correlate with human judgment beyond cases like MSCOCO and Flickr8K?

To address this question, we consider four case studies, exploring the correlation between CLIP-S and human judgment across "divergent" image description datasets. These corpora qualitatively differ from the more popular domains explored in §4, either because the images are not "everyday" images from Flickr, or because the captions are not literal description (Figure 3 illustrates).

5.1 Alt-Text ratings from Twitter

When uploading an image alongside a tweet, users of Twitter have the option of providing alternative text: while few use this feature (Gleason et al. (2019) find that fewer than .1% of image tweets have alt-text), its broader adoption might someday make social media more accessible for low vision and blind users. We measure CLIP-S's capacity to reconstruct a set of 2.8K human judgments of alt-text quality. This corpus was collected and rated by the authors of Gleason et al. (2019, 2020). Each alt-text was rated on a scale of 0 to 3 in terms of its probable utility as an alt-text. While the human-raters raters themselves are sighted thus cannot directly assess the utility of a given alt-text to a low vision or blind user, they are experts in designing and evaluating alt-text systems. Tweets were sampled from a mix of the Twitter FireHose API, and the timelines of low vision and blind users of the site. The images, qualitatively, are a broader mix of web content in comparison to Flickr-like domains, e.g., screenshots, memes, etc. Alt-text candidates are a mix of user-uploaded and machine-generated. The corpus contains no references, but for the purposes of comparison to reference-based metrics, we (programmatically) treat any textual context of the tweet as a reference.

CLIP-S achieves 48.4 τ_c correlation with the human judgements. In contrast, likely due to the unreliability of Tweet texts as viable alt-texts, reference-based methods struggle: the best performing purely-reference based metric, BERT-S (RoBERTa-F) (which achieves 15 τ_c) under-performs relative to length baseline (which achieves 25 τ_c). While gathering high-quality, contextual reference alt-texts is a promising avenue for future work,[15] CLIP-S offers a promising evaluation metric candidate in this domain.

Figure 3: Instances from our four case-study corpora.

5.2 Abstract-50S

We assess CLIP-S's capacity to generalize to abstract, non-photographic clip-art images using Abstract-50S (Vedantam et al., 2015). This dataset

[14] Our criteria is how much additional R^2 correlation with human judgment a metric adds according to a linear regression. We use `sklearn` (Pedregosa et al., 2011)'s forward selection, which applies 5-fold cross-validation at each step.

[15] See Stangl et al. (2020), who conducted user-studies across six domains.

pairs clip-art images (originally constructed by Zitnick and Parikh (2013)) with 48 human-written reference captions. These images depict two cartoon characters, Mike and Jenny, in various outdoor situations, e.g., playing sports, having a picnic, etc. For 400 human-written candidate caption pairs (200 pairs are from the same image, 200 are from different images), human judgments were collected: annotators were instructed to choose which of the paired captions were more similar to each reference caption, so 48 judgments were collected for each candidate pair (for a total of 19200).

We compare CLIP-S to several reference-based metrics when given access to a random sample of five reference captions. Following our procedure for Pascal-50S, we randomly re-sample 5 times, and report average pairwise accuracy. Two baselines (BL) both achieve 53: length-only (i.e., saying the longer caption is better); and randomly shuffling images as input to CLIP-S (so that it cannot rely on meaningful visual-textual interactions).

BL	BLEU-4	CIDEr	METEOR	BERT-S	CLIP-S (no refs)
53	71	79	79	73	68

Overall, while CLIP-S underperforms relative to the reference-based metrics, it outperforms the baselines by a wide margin. This result suggests that CLIP-S is capable of reasoning about visual-textual interactions, even in non-photographic images.

5.3 Personality Captions

Inspired by language use on social media, Shuster et al. (2019) collected image captions by prompting annotators with a "personality" (e.g., dramatic, sympathetic, sad, etc.) and asking them to "write a comment in the context of [a] given personality trait... about an image that someone else would find engaging." To evaluate their models, the authors collected pairwise human judgments, where evaluators were instructed to "to pick which comment is the most engaging". We assess CLIP-S in two capacities: (1) does it prefer literal descriptions, or the less-literal, more engaging, personality captions?; and (2) if it is given two personality captions, can it predict which humans judge to be more engaging?

For (1): Over a set of 2.4K "traditional" vs. personality captions pairwise ratings, humans rate the personality captions to be more engaging 65% of the time, whereas CLIP-S prefers the traditional 80% of the time.[16] Our takeaway: when given a direct description and a more engaging, non-literal caption, CLIP-S will generally prefer the literal.

For (2): CLIP-S performs slightly better than random, e.g., 57% over 2.5K human pairwise judgments comparing two neural generator models: TransResNet (ResNeXt-IG-3.5B) vs. TransResNet (ResNet-152) (see Shuster et al. (2019) Table 7, Row 5), but no better than a length-only baseline (also 57%). Notably, even reference-based metrics fail to provide correlation with pairwise human judgment of engagingness on this corpus: e.g., BLEU-4, CIDEr, and SPICE agree with human judgment 52%/53%/51% when provided with one personality-primed reference. Our takeaway: when given two engaging, non-literal descriptions, both CLIP-S and traditional reference-based metrics fail to predict which humans will judge to be more engaging.

5.4 News image captioning

Biten et al. (2019) consider caption generation for images from New York Times articles; their task differs from MSCOCO because 1) 95% of captions contain at least one named entity, e.g., a politician, celebrity, or place; and 2) captions generally "do not describe scene objects, but rather offer a contextualized interpretation of the scene." They collected 2.1K pairwise human judgments over 106 images that compare the performance of two news image captioning models. For each image, 20 annotators were instructed to pick which of two model generations was closer to the ground-truth caption (they were also presented with the image itself). We compare metrics in terms of their accuracy in matching human judgment between the two candidates.

Reference-based metrics dominate: METEOR and BLEU-4 achieve the highest accuracies of 93 and 91 respectively, whereas CLIP-S achieves only slightly above random at 65. Qualitatively, CLIP-S succeeds when there are visually-verifiable content, e.g., matching black-and-white photos to older dates (e.g., picking 1933 vs. 1977, in one case), and matching particularly iconic celebrities (e.g., it confidently identifies Muhammad Ali boxing).[17] But, its most common failure case are captions that may

[16] Preliminary prompt-engineering experiments (e.g., "when I look at this photo, I feel [PERSONALITY] and think [CAPTION]") could not overcome this.

[17] Luo et al. (2021)'s recent experiments quantitatively demonstrate that CLIP is capable of reasoning about real-world entities within news images.

simply be unverifiable given only the image content. For example: CLIP-S selects “The dining room at Elle Decor" for an image of a room, but annotators preferred a caption that mentioned “the Junior League of New York;" the ground truth caption reveals why the image was pictured in the first place: “A Manhattan home on a May 7 tour by the Junior League of New York."

Overall, we do not advocate for reference-free evaluation in this case, especially because our results suggest that (at least for this particular set of annotations) reference-based n-gram overlap metrics achieve high correlation with human judgment.

6 Conclusion

For literal image description tasks, `CLIPScore` achieves high correlation with human judgments of caption quality *without* references when used in an off-the-shelf fashion. Additional experiments in divergent domains suggest that CLIP can also reason about non-photographic clip-art, and serves as a reasonable option for reference-free evaluation in the alt-text case. Promising future work includes exploring 1) CLIP-S as a reinforcement learning reward for literal caption generators; and 2) whether a small amount of labelled human rating data could help CLIP-S adapt to domains where it struggles, e.g., engagingness prediction. We hope our work can contribute to the ongoing discussion about the role of pretrained models in generation evaluation.

Reference-free evaluation runs some risks. Much like BERTScore, model-based metrics like CLIP-S reflect the biases of the pre-training data. While we believe that using CLIP-S as an offline evaluation metric for literal caption quality accords with the recommendations of CLIP's model card[18] (Mitchell et al., 2019), Agarwal et al. (2021)'s study demonstrates that CLIP can make disproportionate incorrect classifications of people, e.g., “male images were misclassified into classes related to crime.” Exploring potential social biases of candidate generations (as in, e.g., Hendricks et al. (2018)) remains paramount, particularly if a system is to be deployed.

Contemporaneous work While this work was under submission, two alternate reference-free evaluation metrics for image caption generation were introduced: FAIEr (Wang et al., 2021) (based on a pretrained object detector, and fine-tuned on MSCOCO) and UMIC (Lee et al., 2021) (based on UNITER (Chen et al., 2020)). UMIC, in particular, produces similar correlations with human judgment on the literal image description tasks (§4) compared to CLIP-S, but with the complementary approach of fine-tuning on synthetic negative captions. Future work would be well-suited to explore if the textual data augmentations proposed by Lee et al. (2021) (1) result in a metric that complements or overlaps with the non-finetuned CLIP-S (§4.6); and (2) could be extended beyond cases of literal description (§5).

[18] https://github.com/openai/CLIP/blob/main/model-card.md

Acknowledgements

This research is supported in part by DARPA MCS program through NIWC Pacific (N66001-19-2-4031), DARPA SemaFor program, and the Allen Institute for AI. We additionally thank Ximing Lu, Swabha Swayamdipta, Youngjae Yu, and the anonymous EMNLP reviewers for the helpful comments, thoughts, and discussions. Finally, we thank Jin-Hwa Kim, who in March 2022, helped discover a now fixed discrepancy for the Pascal-50S results, see Appendix A.

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 arXiv:1511.03292*.

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating CLIP: Towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*. Springer.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, 7:597–610.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on Evaluation Measures for MT and Summarization*.

Alexander C. Berg, Tamara L. Berg, Hal Daumé III, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, and Kota Yamaguchi. 2012. Understanding and predicting importance in images. In *CVPR*.

Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *CVPR*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *COLING*.

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 *ECCV*.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *CVPR*.

Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. In *NeurIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *ACL*.

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 *WWW*.

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 *CHI*.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 47:853–899.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

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 *EMNLP*.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: An empirical study. In *ICLR*.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In *1st Workshop on Evaluating NLG Evaluation*.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *EACL*.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021. UMIC: an unreferenced metric for image captioning via contrastive learning. In *ACL*.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. Vilbertscore: Evaluating image caption using vision-and-language bert. In *First Workshop on Evaluation and Comparison of NLP Systems*.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *ECCV*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*. Springer.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *ECCV*.

Chi-kiu Lo. 2019. Yisi-a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In *Fourth Conference on Machine Translation*.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *CVPR*.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLIPpings: automatic generation of out-of-context multimodal media. *arXiv preprint arXiv:2104.05893*.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *CVPR*.

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 *CHI*.

Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2019. VIFIDEL: Evaluating the visual fidelity of image descriptions. In *ACL*.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012. Match without a referee: evaluating mt adequacy without reference translations. In *Seventh Workshop on Statistical Machine Translation*.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *ACL*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *FAccT*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *JMLR*, 12.

Maxime Peyrard and Iryna Gurevych. 2018. Objective function learning to match human judgements for optimization-based summarization. In *NAACL*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *ACL*.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *EMNLP*.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *IJCV*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! find one mismatch between image and language caption. In *ACL*.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *CVPR*.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Lucia Specia and Kashif Shah. 2018. Machine translation quality estimation: Applications and future perspectives. In *Translation Quality Assessment*, pages 201–235. Springer.

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 *CHI*.

Simeng Sun and Ani Nenkova. 2019. The feasibility of embedding based automatic evaluation for single document summarization. In *EMNLP*.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *TPAMI*, 39(4):652–663.

Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. 2021. FAIEr: Fidelity and adequacy ensured image caption evaluation. In *CVPR*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *EMNLP*.

Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2019. Quality estimation and translation metrics via pre-trained word and sentence embeddings. In *Fourth Conference on Machine Translation*.

Yanzhi Yi, Hangyu Deng, and Jinglu Hu. 2020. Improving image captioning evaluation by considering inter references variance. In *ACL*.

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. *TACL*, 2:67–78.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *ICLR*.

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 *ACL*.

C Lawrence Zitnick and Devi Parikh. 2013. Bringing semantics into focus using visual abstraction. In *CVPR*.

A Evaluation and Replication Details

Anderson et al. (2016) introduced a set of corpora, metrics, and experimental settings for comparing image caption generation evaluation metrics. Perhaps unwittingly, their introduced protocols have become the accepted standard for evaluation of new caption generation metrics. However, seemingly innocuous preprocessing+reporting choices can significantly impact correlations with human judgment on these corpora. In what follows, we detail our replication efforts. Our goal was to make the experimental comparisons involving `CLIPScore` reported in the main paper as fair as possible. We hope it can be useful for researchers reporting metrics on this setup going forward.

Flickr8K details

We contacted the authors of some prior work, and did our best to re-create their evaluation settings. We uncovered two types of discrepancies when reporting on this corpus. The first discrepancy is that prior work has mixed evaluating rank correlations with kendall-C and kendall-B. These metrics handle ties differently, and ties are frequent because human Likert judgements are discretized. The second discrepancy is the method of aggregation of human ratings. Three human ratings were gathered for 5664 (image, candidate) pairs. The majority of prior works flatten all human judgments to a single list, and report rank correlation over 5664 * 3 = 16992 instances (method A). However, another (possibly more defensible) evaluation choice is to average human ratings for each pair, and report rank correlation instead over 5664 instances (method B). The choice of aggregation method has a significant impact on correlations. For example, when we used aggregation method A and τ_c for SPICE, we can exactly replicate the correlation, 44.9, originally reported in (Anderson et al., 2016). But, if we use τ_c and instead use aggregation method B, the correlation increases to 52.9: this inflation occurs with other metrics, too.

For our results, we do our best to report all results for the most common setting: using τ_c correlation, and using aggregation method A. Thus, the results we report may differ slightly than the results reported in prior work.

Composite details

For this corpus too, prior work has mixed evaluating with kendall-C and kendall-B correlations, which can have an impact, e.g., for CIDEr in our setting, switching from τ_b to τ_c results in an increase from 35 to 38 rank correlation. But perhaps the most impactful decision for this corpus relates to the references: each image originally has (roughly) five references. But when gathering human judgments, one of the candidate captions that was rated by humans was sampled from the references. For Flickr8k, Anderson et al. (2016) "exclude 158 correct image-caption pairs where the candidate caption appears in the reference set;" this curation choice has become standard for Flickr8k. But for Composite, it's not clear if they repeated this curation choice, or not. And because of this ambiguity, it's not obvious which standard each prior work followed, either. For fair comparison, in an effort to reconstruct Anderson et al. (2016), we tried both ways: removing the ground truth candidate reference, and not.

	Original	τ_b no GT	τ_b w/ GT	τ_c no GT	τ_c w/ GT
BLEU-1	26	29	45	31	49
BLEU-4	18	31	46	31	50
ROUGE-L	28	30	48	32	49
METEOR	35	36	49	39	50
CIDEr	36	35	48	38	52
SPICE	39	39	51	40	53

Table 6: Attempts at replicating Anderson et al. (2016)'s results on the composite corpus.

Our efforts to replicate the exact values of Anderson et al. (2016) are in Table 6. We suspect the discrepancy in BLEU-4 likely results from a smoothing issue related to the application of BLEU-4 to individual captions vs. the whole corpus (as mentioned in Kane et al. (2020)). Based on these replication efforts, it's likely that the original evaluations for this corpus were computed using τ_c with GT references removed. We agree that the fairest analysis on this corpus should not include a reference that is also a candidate. And while we didn't go through all prior works and recompute their metrics with this change, we did compute ViLBERTScore-F in this setting, because it was, before `CLIPScore`, the state-of-the-art for this corpus. If it's helpful for future reporting: in the setting where all references (including the GT reference) are used, RefCLIP-S gets $\tau_c = 60.0$.

MSCOCO system-level details

The MSCOCO 2015 image captioning challenge is a standard corpus for evaluation the system-level correlation between new evaluation metrics and hu-

man judgments on the MSCOCO test set. To our knowledge, this evaluation was first conducted by Anderson et al. (2016) using a random sample of 1K test set submissions from 15 teams. But because the test set predictions are not public, more recent work (e.g., Cui et al. (2018); Zhang et al. (2020)) has evaluated using dev set predictions from systems, and assuming dev set results correlate with test set results (12 teams submitted dev predictions). However, there are some potential problems with this setup:

1. There's reason to believe that some teams give dev set predictions with different models vs. test set predictions. For example, the dev set predictions are identical between the two submissions: `m-RNN` and `m-RNN (Baidu/ UCLA)`, but the test set predictions differ (and achieve significantly different scores).
2. Correlations are reported over 12 (or possibly only 11, given the duplicate predictions) systems. But spearman/pearson correlation over only 12 observations is unfortunately simple to (accidentally) "game" due to the low statistical power of the comparison (see Card et al. (2020) for an overview of statistical power in NLP). Consider a (nonsense) evaluation metric that assigns a random uniform $[0, 1)$ "score" to systems without examining outputs, and consider applying this metric, e.g., $N = 10$ times to the 12 systems and taking the best performing run as the final metric (simulating either a single researcher developing a new evaluation metric and/or the community's collective trials). We ran a simulation of this process 1000 times: the average spearman/pearson correlation between human judgments and our bogus metric were $r/\rho = .91$, due to repeated evaluation and low sample size.

Thus, while the intent of this evaluation is understandable, and it may be possible to garner some insight if relatively few evaluations are conducted, this specific setup as a fine-grained comparison between new evaluation metrics for caption generation has likely outlived its utility.

Pascal-50S Setup Erratum

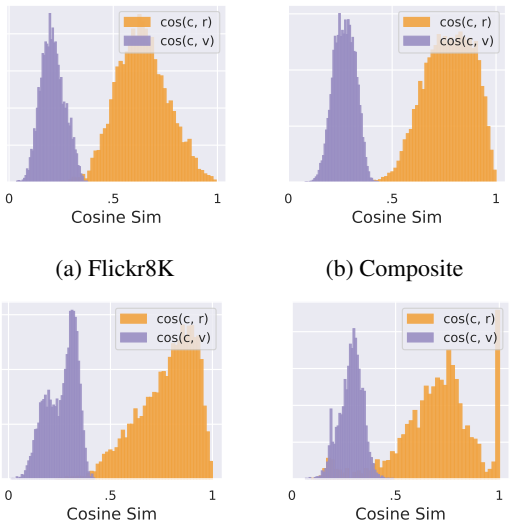
In March 2022, Jin-Hwa Kim reported some small discrepancies in a replication effort for the Pascal-50S corpus. Upon further investigation, it was discovered that the original version of this work was using a different set of human judgments than the usual setup. In particular, the Pascal-50S corpus contains two types of human judgments: 11 human judgments per pair (located in a file named `pair_pascal.mat`); and 48 human judgments per pair (located in a file named `consensus_pascal.mat`). The 48 judgments are intended to be used, and the results in the main paper have been updated accordingly. For reproducibility sake, in case future work utilizes the 11 judgments, we have included those results in Table 7.

	HC	HI	HM	MM	Mean
length	**65.4**	52.4	63.0	42.3	55.8
BLEU-4	52.5	90.4	84.9	55.3	70.8
SPICE	56.9	96.3	87.1	66.4	76.7
METEOR	59.0	97.7	93.9	62.0	78.2
ROUGE-L	55.0	95.3	93.1	58.7	75.5
CIDEr	53.7	98.1	90.8	63.7	76.6
BERT-S (RoBERTa-F)	54.4	96.1	94.3	56.4	75.3
CLIP-S (no refs)	60.3	99.4	**97.9**	77.3	**83.7**
RefCLIP-S	57.9	**99.5**	96.1	**80.8**	83.6

Table 7: Pascal50S-11-judgment accuracy results (5 references, non-standard 11 human judgment version). HC = two human correct captions; HI = both captions are human written, but one is wrong; HM = both captions are for the image, but one is written by a human, one by an algorithm; MM = both captions are for the image, and both are written by an algorithm. We average our results over 5 random samples (but CLIP-S doesn't change because it doesn't use references).

B Rescaling `CLIPScore`

For readability purposes, as in Zhang et al. (2020), we sought to re-scale the raw cosine similarities computed by CLIP `ViT-B/32`. While such a monotonic rescaling operation doesn't affect ranking results, for reporting purposes, it can be easier to compare raw values if they are on a scale more closely-aligned with other evaluation metrics (e.g., from roughly zero to roughly one). Figure 4 shows the raw candidate-reference and candidate-image cosine similarities for four corpora. (Many "reference"-candidate similarities for the Twitter corpus are 1.0 because users frequently use the text of their tweet as the AltText.) Across all of these cases, we never observed a negative negative cosine similarity. But, to be safe, we take a maximum between the cosine similarity and zero because the harmonic mean used to compute `RefCLIPScore` would be undefined for negative values. Multi-



(a) Flickr8K (b) Composite

(c) Pascal50S (d) Twitter AltText

Figure 4: Distributions of raw cosine similarities between candidate and references and candidate and visual content from CLIP `ViT-B/32`.

plying by 2.5 has the effect of "stretching" the `CLIPScore` distribution to more uniformly span between zero and one, though, `CLIPScore` can be greater than 1. Furthermore, when computing `RefCLIPScore`, we maintain this weighting, because it has the effect of mapping the visual-textual cosine similarity distribution to more closely match the reference-candidate distribution: this provides a roughly equal importance weighting between the image-candidate and reference-candidate similarity factors.

We note that the exact parameters of our rescaling method only apply to CLIP `ViT-B/32`. If future, bigger models are released, e.g., the presently unreleased `ViT-L/14` CLIP variant, they could exhibit a different cosine similarity distribution.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*. Springer.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *EMNLP*.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *CVPR*.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In *1st Workshop on Evaluating NLG Evaluation*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *ICLR*.