Distribution Aware Metrics for Conditional Natural Language Generation

David M. Chan[1], **Yiming Ni**[1], **David A. Ross**[2], **Sudheendra Vijayanarasimhan**[2], **Austin Myers**[2], **John F. Canny**[1]
[1]University of California, Berkeley, [2]Google Research
{davidchan, nym_claire, canny}@berkeley.edu
{aom, svnaras, dross}@google.com

Abstract

Traditional automated metrics for evaluating conditional natural language generation rely on pairwise comparisons between a single generated text and the best-matching gold-standard reference. This method is effective when ground truth data diversity can be attributed to noise, however, it falls short when diversity in references holds valuable contextual information, as in visual description or summarization, as it does not evaluate the ability of a model to generate text matching the diversity of the ground truth samples. In this paper, we challenge the adequacy of existing metrics in such semantically diverse contexts and introduce a novel approach for evaluating conditional language generation models, leveraging a family of meta-metrics that build on existing pairwise distance functions. These meta-metrics assess not just single-samples, but distributions of reference and model-generated captions using small sample sets. We demonstrate our approach through a case study of visual description in the English language which reveals not only how current models prioritize single-description quality over diversity, but further sheds light on the impact of sampling methods and temperature settings on description quality and diversity.

Keywords: Evaluation Methodologies, Language Modeling, Natural Language Generation

1. Introduction

Recent models for conditional language generation, particularly in the field of visual description, have shown dramatic improvements in both fluency and the ability to ground generated language in context (Liu et al., 2021; Zhou et al., 2020; Mokady et al., 2021; Chen et al., 2018). Standard metrics for these tasks such as BLEU, ROUGE, METEOR, and CIDEr, compare a generated text with a reference set of texts and compute some measure of quality for the generated text. By construction of these metrics, a model will achieve the best performance by generating a single high-scoring text. In contrast, it has been widely observed that large language models such as GPT-3 (Brown et al., 2020) or LAMDA (Thoppilan et al., 2022) generate the most realistic texts at temperatures close to one, where the set of potential texts generated is often very diverse. More significantly, if we look at an example of an image from MS-COCO and its set of reference captions (Figure 1), we notice that each (human-generated) reference contains a unique subset of the overall information in the image:

> "A woman in a red robe is sitting at a dining table."
> "A woman in a red flowered shawl sits at a table while a man wearing jeans is in the kitchen looking at her."
> "A person sits at a table and another person stands in the kitchen."
> "A woman is sitting at a table wearing a robe while a man is cooking."
> "Man and woman in a kitchen looking in the same direction."

Important features like the red robe, the man, the gaze of the two people etc, are mentioned only in one or a few captions. Metrics that encourage gen-

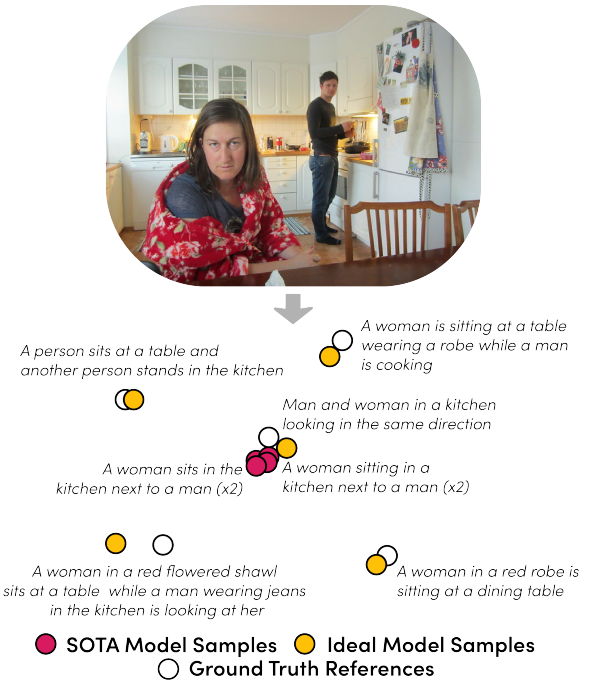


Figure 1: Samples from these two models achieve similar BLEU scores, however, the samples from a SOTA model (VLP) lie near a center of the distribution, and fail to capture the dispersion of natural language in the ground truths, while the samples from an ideal model better match the ground truth distribution. In this work, we introduce metrics which better measure deviations between samples from candidate and reference distributions, compared to single-sample pairwise metrics.

LREC-COLING 2024, pages 5064–5095
20-25 May, 2024.

erating information from *only one* of these captions will generally fail to capture much of the important detail in the image. This holds for more than just image description. For many conditional language generation tasks such as video captioning, abstractive summarization, translation, and open-ended question-answering, it is often beneficial to be able to sample from a diverse distribution of generated outputs. If we compute a maximum-likelihood generated caption from a state-of-the-art model (Zhou et al., 2020) we get:

"A woman sitting in a kitchen next to a man."

In this description, we see that only information common to most or all of the reference captions is preserved. This is intuitive, since including more information runs the risk that no reference caption contains that information, leading to a low score. It seems the designers of metrics such as BLEU are already aware that direct use of shortest distance to a reference caption favors generated captions which are even shorter and more impoverished, and thus, the BLEU score, and many others, also include a term encouraging longer texts. However, the (log-) text length heuristic in standard metrics is intuitively a poor proxy for actual diversity. Thus, since models optimize for standard measures, drawing multiple maximum-likelihood samples using beam search from SOTA models only produce repetitions, or slight variations of the above caption.

Thus, we encounter an issue in the evaluation of conditional text generation models with multiple available references. With multiple references, typically the metric score is based on the maximum score over a set of ground truths (e.g. max pairwise score for a particular n-gram as in BLEU), leading measures to erroneously incentivize the production of text minimizing the expected pairwise distance to the reference set, i.e. near a strong mode in the training text distribution, causing the issues discussed above. Changing the metric aggregation method (e.g. sum as in ROUGE) does not substantially alter this situation, as the model still strives to produce a high-scoring output that is close to nearby references which will be maximized at a smoothed mode in the training text distribution (Caglayan et al., 2020; Yeh et al., 2021).

An over-reliance on simple aggregations for multiple candidates and references has, over time, compounded into several issues: The first, discussed further in section 3, is that, as observed in visual description by Chan et al. (2022) and dialog generation by Caglayan et al. (2020), human-generated captions tend to receive lower scores than model-generated captions using automated measures, even though they actually receive higher scores under human evaluation. The second, discussed in section 2, is that diversity of candidate texts is largely relegated to reference-unaware measures, encouraging models to diverge from ground truth distributions to hit diversity targets.

In this work, we aim to solve these problems by introducing several novel automated ways of measuring the performance of conditional text generation models. Our measures encourages models to not only to generate samples at the locus of a distribution but also with sufficient variance, since they are designed computing the divergence between candidate and reference *distributions*. While some recent methods have been designed to closely measure the divergence between full distributions of text data in the unconditional case (Pillutla et al., 2021), no such methods exist for conditional generation, which often operates on the level of 10s of reference samples and candidates. Our contributions are summarized as follows:

1. We demonstrate that existing automatic metrics that use simple aggregations of candidate and reference distributions are insufficient, and we introduce a new paradigm that instead involves sampling from these distributions, and comparing the samples.
2. We introduce two new families of metrics which *extend* existing semantic distances: triangle-rank metrics, and kernel-based metrics, designed to measure the divergence between small text samples from candidate and reference distributions.
3. We explore how our new metrics behave in the context of visual description (both image and video description) and show that by measuring distributional effects, we can capture nuances in the data that existing metrics cannot explore.

2. Related Work

This work is not the first to notice the shortcomings of traditional metrics for the automated evaluation of conditional language generation models. In visual dialog, Caglayan et al. (2020) find that a number of the automated metrics proposed for visual dialog do not match well with human judgment, while in visual description, Chan et al. (2022) find that current automated metrics do not assign high scores to human-generated descriptions. This work not only quantifies such issues but proposes a method for addressing these cases without developing novel metrics for measuring text semantic distance. In this section, we review related works, roughly divided into three groups; methods for evaluating text quality, text diversity and distribution aware metrics.

Measuring the Quality of Generated Text The evaluation of machine-generated text has long been an active area of research, which has continuously evolved to keep pace with accelerating advances in text generation. As a consequence of the tools

available and the state of early text generation approaches, classical measures have primarily focused on evaluating the quality of generated text with respect to ground truth references using surface-level text statistics. Most notably, these include n-gram matching based metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015). More recently, the rapid progress enabled by large-scale language models has motivated new evaluation techniques which go beyond superficial n-gram statistics and toward measures that aim to capture the underlying semantics of language (Shimanaka et al., 2018; Clark et al., 2019; Zhang et al., 2020b; Sellam et al., 2020). These approaches leverage high dimensional representations of generated and reference text provided by a state-of-the-art language model, such as BERT (Devlin et al., 2019) in the case of BERTScore (Zhang et al., 2020b) and BLEURT (Sellam et al., 2020). While such methods are focused on measuring the semantic distance between two pairs of natural language texts, the evaluation of the diversity of the generated captions has largely been done independently of quality.

Measuring the Diversity of Generated Text Until recently, measures of diversity for generated text have been largely secondary to measures of quality, since the pursuit of human-like generated text has been the primary focus of the field. In fact, many diversity measures quantify surface-level statistics of the generated text (van Miltenburg et al., 2018), such as metrics based on the number of unique tokens, unique sentences, or unigram frequency statistics, such as Zipf coefficients (Holtzman et al., 2020). Similarly, n-gram-based diversity measures such as self-BLEU (Zhu et al., 2018), compute scores between samples from a model. Unfortunately, these approaches do not consider the diversity of a model's outputs with respect to the diversity of human references, and are primarily focused on the diversity of the vocabulary, rather than the aggregate semantic diversity, factors that our proposed work aims to address.

Distribution Aware Measures of Generated Text MAUVE, proposed by Pillutla et al. (2021), measures the divergence between multi-candidate samples and multiple ground truths using density estimates in a text embedding space. This approach measures both text dispersion and quality simultaneously, however, MAUVE is designed for unconditional text generation with many thousands of candidate and ground truth samples available. While MAUVE works well in these scenarios, it does not work well when only a few references are available (due to the K-means approximation) (see appendix B.4). Such a low-reference scenario is common in conditional NLG, making MAUVE unsuitable for many potential applications, and motivating the need for more sensitive measures.

3. Methods

In this section, we introduce our two primary contributions. First, we introduce and demonstrate the need for a paradigm for multiple candidate evaluation for conditional language generation, and second, we introduce several simple augmentations to existing pairwise metrics, designed to alleviate the sensitivity issues induced by evaluating conditional language generation models with only a single candidate text. Our family of augmented metrics, which we call Triangle-Rank Metrics (TRMs), represents the first step towards optimizing metrics that force models not only to generate samples at the locus of a distribution but also with sufficient variance, hopefully alleviating the field-wide issues that optimizing standard pairwise-metrics can induce.

3.1. Multi Candidate/Reference Evaluation

Traditionally, most methods for conditional language generation have been designed to sample a single candidate example using beam search, designed to be a maximum likelihood sample of the data. This single candidate is compared against the reference data. Unfortunately, as discussed in section 1, models can easily exploit such aggregations. For example, when the best score amongst the ground truths is chosen (the "min-distance" aggregate), models generate texts optimizing the *expected minimum distance to the reference distribution*. Such a text is, by definition, the mode of the distribution. This mode likely represents some amount of central tendency, as we observe such captions to be bland and uninformative (See B.5, (Chan et al., 2022; Yang et al., 2019)).

Thus, a single candidate may not be sufficient to understand if the model has learned to approximate the reference distribution. Consequently, we aim to develop methods that can sample several suitable candidate texts, each with high accuracy, while matching the diversity of the ground truth distribution. In this work, to extend methods to multiple candidate generation, we leverage temperature-based sampling or nucleus sampling (as indicated) to produce multiple candidates from each model's distribution. While beam search can generate multiple candidates, Vijayakumar et al. (2016) showed diversity among beams is relatively poor, leading to samples that diverge from the model distribution. This gives us a model which *generates multiple candidate samples*, and requires an evaluation metric which *compares multiple candidate samples to multiple reference samples*.

Extending Existing Metrics for Multi-Candidate

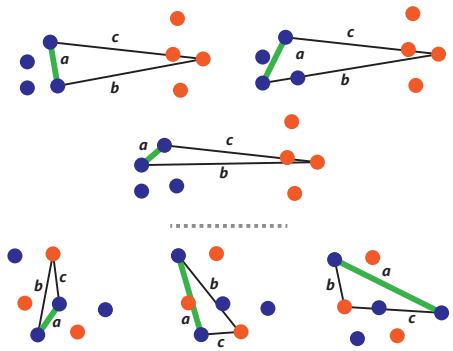


Figure 2: Intuition for TRMs. For samples from different distributions (top), in-distribution edges will often be short, but for identical distributions (bottom), edge rank-distributions will be more uniform.

Evaluation Currently, no standard pairwise metrics (Papineni et al., 2002; Agarwal and Lavie, 2008; Lin, 2004; Vedantam et al., 2015; Zhang et al., 2020b) support a comparison between multiple candidates and multiple references, and the most efficient extension of existing metrics to multi-candidate, multi-reference situations is a non-trivial task. In this work, we naively extend the existing pairwise metrics to multiple candidates through the use of mean aggregation. Thus, for a standard pairwise score S, set of candidates $(c_1,...,c_n)=C$ and a set of references $(r_1,...,r_m)=R$, we assign the output score S_{agg} as:

$$S_{\text{agg}}=\frac{1}{N}\sum_{i=1}^{N}S(c_i,R) \tag{1}$$

3.2. Triangle-Rank Metrics (TRMs)

While existing metrics for semantic similarity are powerful for determining the pairwise semantic distances between two utterances (Papineni et al., 2002; Agarwal and Lavie, 2008; Lin, 2004; Vedantam et al., 2015; Anderson et al., 2016), these measures cannot accurately measure the distance between distributions. How, then, can we leverage already strong pairwise tools in a multiple candidate scenario? Unfortunately, many statistical techniques for measuring the distances between samples require points to lie in a metric space (Basseville, 2013) - however, most text distances neither respect symmetry nor triangle inequality.

We propose a novel answer based on an application of the triangle-rank statistic for statistical testing proposed by Liu and Modarres (2011). The triangle-rank statistic has several promising properties: it neither requires symmetry nor the triangle inequality in the metric space (it only requires $d(x,x)=0$), and it is computed using only pairwise distances, meaning that we can easily reuse existing text semantic distance functions when computing the statistic.

For the purpose of explanation, it can be helpful to think of texts as points on an arbitrary manifold (based on the selected text distance function). To compute the triangle-rank statistic for a given distance S, a set of candidates $(c_1,...,c_n)=C$ and a set of references $(r_1,...,r_m)=R$, we first extract all directed triangles $(t_1,...)=T$, such that one point lies in C and two points lie in R. We refer to the edge between points from the same distribution as $e_{t_i}^{\text{IN}}$ and the other two edges as $e_{t_i}^{E_0}$ and $e_{t_i}^{E_1}$. We then compute the score for each of the edges. For $(a,b)=e_{t_i}^{\cdot\cdot}$, let

$$d(e_{t_i}^{\cdot\cdot})=S(a,b) \tag{2}$$

We then compute indicators I_0,I_1,I_2 for each triangle t_i as follows:

$$\begin{aligned}
I_0(t_i)&=1 \text{ if } d(e_{t_i}^{\text{IN}})\le d(e_{t_i}^{E_0}),d(e_{t_i}^{E_1}) \text{ else } 0\\
I_1(t_i)&=1 \text{ if } d(e_{t_i}^{E_0})\le d(e_{t_i}^{\text{IN}})\le d(e_{t_i}^{E_1}) \text{ or}\\
&\quad\ d(e_{t_i}^{E_1})\le d(e_{t_i}^{\text{IN}})\le d(e_{t_i}^{E_0}) \text{ else } 0\\
I_2(t_i)&=1 \text{ if } d(e_{t_i}^{E_0}),d(e_{t_i}^{E_1})\le d(e_{t_i}^{\text{IN}}) \text{ else } 0
\end{aligned} \tag{3}$$

These indicators represent the rank of the same-sample edge (if it is the smallest, largest, or middle-sized edge). The directed statistic for the sample (C,R), $Q(C,R)$ is then computed as:

$$\begin{aligned}
Q(C,R)=&\left|\frac{\sum_{t_i\in T}I_0(t_i)}{|T|}-\frac{1}{3}\right|+\\
&\left|\frac{\sum_{t_i\in T}I_1(t_i)}{|T|}-\frac{1}{3}\right|+\left|\frac{\sum_{t_i\in T}I_2(t_i)}{|T|}-\frac{1}{3}\right|
\end{aligned} \tag{4}$$

For the experiments in this paper, we use an extension of the directed statistic, the undirected statistic, $TRM(C,R)=Q(C,R)+Q(R,C)$, which increases the sensitivity of the metric by taking into account rank statistics of both within-candidate and within-reference edges.

An intuition for how this statistic measures divergence between distributions is given in Figure 2. If the in-distribution edges are always short compared to the cross-distribution edges, this suggests that either the distance between the candidate and reference distributions is high (different locus), or the spread of the candidates in the semantic space is significantly less than that of the references (different spread). If the in-distribution edge is always the longest edge, it suggests that the spread or dispersion of the candidate samples is higher than the dispersion of the reference samples. Because this statistic takes into account the full distribution through triplets of samples, it does not suffer from the issues with aggregation discussed in section 1 and earlier in this section. Not only does it solve these issues, but TRMs build on existing pairwise metrics, allowing us to increase sensitivity while retaining existing semantic distance measure and intuitions.

Notably, $Q(C, R)$ does not distinguish between situations where $I_0 = 1$ and $I_2 = 1$. Intuitively, a model that can generate a candidate that is closer to two references than the references are to each other ($I_0 = 1$) seems to be better than another model where the candidate is far apart from one (or both) of the references ($I_2 = 1$), however this is not always a desirable situation (in fact, it is often a situation we wish to avoid). Consider the situation where the "mean" of all reference captions is generated by the candidate set. This caption is closer to any individual caption than any reference caption may be to other reference captions, however as seen in Figure 1, and discussed in prior work (Caglayan et al., 2020; Yeh et al., 2021; Chan et al., 2022), such captions capture only mutual information in the references, and fail to match the full distribution.

It is worth mentioning that the axes of diversity and locality are not separated numerically: a low score could indicate that either the scores are not diverse enough *or* the captions are factually incorrect. This is both a strength, in that it gives a single omnibus measure with which both axes can be measured, but can also be less directly interpretable, as it could be unclear how to improve any specific sample. To that end, it still remains a valuable approach to augment the proposed measures with existing pairwise measures. By doing so, it becomes easier to determine when the correctness of the generated candidates is poor (i.e. the content of the generated captions is different from the content of the reference captions) vs. when the coverage is poor. For example, one could consider the minimum/maximum of the pairwise distances across the candidate set to bound the content distance.

3.3. Kernel-Based Metrics

While TRMs represent one method of augmenting existing pairwise metrics, a second possible approach relies on representing utterances as points in the embedding space of a model, particularly a large pre-trained model such as BERT (Devlin et al., 2019) or GPT (Brown et al., 2020). Evaluating the distance between two distributions based on representative samples on a Euclidean manifold is relatively well studied in GAN literature. One option, MAUVE, introduced by Pillutla et al. (2021), uses a K-Means density estimator to estimate the distribution of the points on this manifold and then computes a fixed divergence (such as Kullbeck-Libeller) between the two density estimates. Unfortuantely, MAUVE cannot correctly estimate the density when there are few samples, such as in the case of conditional language generation, as the K-means density estimator requires at least K (usually at least 50) samples. In this work, we introduce several possible extensions to MAUVE as an alternative family of distribution-aware metrics, which we dub "Kernel-Based Metrics" (KBMs):

- **FID-BERT (A.6):** The Frechet Inception Distance (Salimans et al., 2016) represents the squared Wasserstein distance between multidimensional Gaussian distributions fitted to the components of the input. In the FID-BERT metric, we replace Inception embeddings with those from a pre-trained BERT model (Devlin et al., 2019).
- **MMD-BERT (A.7)**: A related metric is the maximum mean discrepancy distance function (Li et al., 2017), which leverages a density estimate of the data, and computes the maximum mean discrepancy between the density estimates for each sample. In our case, we leverage a Gaussian kernel estimate over the embeddings generated by a pre-trained BERT model (Devlin et al., 2019).

While we primarily explore BERT-based embeddings for KBMs, we explore additional text embedding methods in Appendix B.1.

4. Case Study: Visual Description

Visual description is a challenging task where a model must generate natural language descriptions of visual scenes. Datasets for visual description often set themselves apart from other datasets for conditional natural language generation (such as those for translation and summarization), as they contain more than one ground truth sample, making it possible to evaluate multi-reference measures. In this set of experiments, we look at two datasets for visual description: MSCOCO (image description) (Lin et al., 2014) and MSR-VTT (Xu et al., 2016) (video-description) (full dataset details in appendix A.2). We demonstrate first that current metrics are not sensitive enough to evaluate the performance of existing approaches, and then show quantitatively how a multi-candidate evaluation paradigm can close this gap, and how a distributionally sensitive metric, such as TRMs, can provide new insights.

Single caption evaluation is insufficient A natural first question to ask when evaluating the performance of a metric is, "given the data, is the metric sensitive enough to distinguish between captions from a model and caption from a reference distribution?" To answer this question, we evaluated the p-values using a permutation-test for each measure under the null hypothesis that the candidate and reference samples come from the same caption distribution. The p-values represent the probability of obtaining the observed result under the null hypothesis: a higher p-value means that it is immanently possible the results obtained are due to

Table 1: The p-value (lower is better) produced by measuring standard metrics under the null hypothesis that the candidate distribution is the same as the reference distribution (using single-image/video tests aggregated with HMP (Wilson, 2019)). With a single candidate text, the metrics are unable to make a statistically significant distinction ($p < 0.05$) between ground truth and candidate samples, motivating the need for multi-candidate evaluation. BERT refers to the BERT-Score (Zhang et al., 2020b). Additional experimental detail in A.5.

Model	BERT	CIDEr	BLEU	METEOR	ROUGE
(Video) MSR-VTT Test Set p-values					
TVT	0.658	0.409	0.781	0.457	0.477
O2NA	0.645	0.457	0.795	0.564	0.593
Human	0.515	0.531	0.829	0.530	0.566
(Images) MS-COCO Karpathy Test Set p-values					
CLIPCap	0.558	0.822	0.878	0.748	0.798
VLP	0.592	0.742	0.859	0.664	0.770
Human	0.640	0.668	0.874	0.635	0.684

chance rather than any signal in the underlying experiment. It is important to highlight that in this paper, when we compare p-values, we are evaluating the *sensitivity* of the measures on a *single experiment* and *not comparing p-values between experiments*. It is generally not the case that lower p-values correspond to better captions, rather, lower p-values when comparing two differing distributions indicate a more sensitive measure.

The results, shown in Table 1 demonstrate that under all existing measures, using a single description for the candidate dataset does not have sufficient sensitivity ($p < 0.05$) to tell different distributions apart, motivating a transition to a paradigm with significantly more sensitivity. This result confirms observations made in Yeh et al. (2021) and Liu et al. (2016): most metrics are unable to produce statistically significant results. Thus, even for standard metrics, it makes sense to sample more than one ideal candidate description and aggregate the metric score across these candidate descriptions. Such a sampling approach for evaluation does not preclude efforts toward generating single "omnibus" captions capturing details from several diverse captions. However, such captions will be much longer than typical human captions, and will score poorly under the standard metrics, as they would differ greatly from individual reference captions.

TRM and KBM metrics are more sensitive than naive aggregation In section 3, we proposed several new metrics which can be leveraged by switching to multi-candidate evaluation. Figure 3 shows the sensitivity of both the newly introduced metrics and existing metrics using the naive aggregation schemes discussed in section 3, as

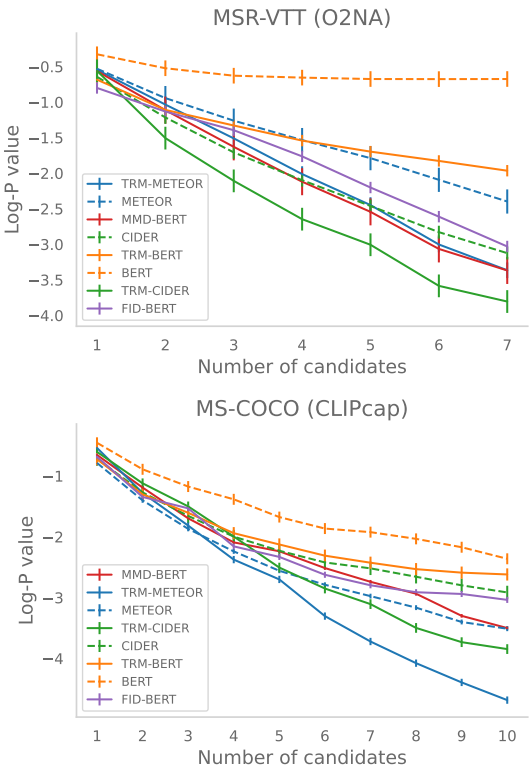


Figure 3: Plots showing the log p-values for the existing and proposed metrics as we increase the number of sampled candidate descriptions from the models. $\text{TRM}_{\text{METEOR}}$ achieves a 162% increase in sensitivity over METEOR, while $\text{TRM}_{\text{CIDEr}}$ represents a 49.3% increase over CIDEr-D for O2NA evaluated on the MSR-VTT dataset. Additional experimental details are given in A.5.

we increase the number of candidate samples from the model. While the sensitivity increases for all models to significance, our proposed metrics are much more sensitive with fewer candidate and reference descriptions. As an additional check, when tested on human captions, our metrics do not consider the two distributions significantly different ($p > 0.05$, see B.3). Our proposed metrics do not alter the manifold: so, for example, $\text{TRM}_{\text{METEOR}}$ and METEOR measure the same underlying intuitive divergences (n-gram recall with some additional synonym matching), however, our TRM method increases the sensitivity of the test, allowing us to measure the full distribution divergence, instead of using naive aggregates. For a practitioner, computing the full p-value of the data is unnecessary; we need only sample enough candidates to be sure of the statistical significance.

Multi-candidate evaluation illustrates a diversity vs. likelihood trade-off A metric's sensitivity to the full distribution can give us novel insights into the visual description task. Consider the two models, VLP (Zhou et al., 2020), a standard transformer-based model pre-trained on large-scale vision and language data, and CLIPCap (Mokady

Candidate Set 1
METEOR (↑): 1.0
TRM-METEOR (↓): 0.574
The cows are grazing in a field.
The cows are grazing in a field.
The cows are grazing in a field.
The cows are grazing in a field.
...

Candidate Set 2
METEOR (↑): 0.393
TRM-METEOR (↓): 0.069
Animals grazing on grass in an enclosed area.
Several cows grazing in a field with trees in the background.
Cows grazing in a large green pasture in a distant scene.
A grassy field overlooking cows in a pasture.
...

References
Cows grazing in a pasture ringed with trees.
Polaroid-looking photograph of cows in a green pasture.
A herd of animals grazing on a lush green field.
The cows are grazing in a field.

Figure 4: A qualitative sample from CLIPcap. Candidate set one uses beam search (8 beams), while candidate set two uses nucleus sampling (with temperature one, top-k of 20 and top-p of 0.9). As the diversity increases, the $\text{TRM}_{\text{METEOR}}$ divergence decreases, but METEOR fails to correctly capture the diversity/correctness trade-off, leading to decreased scores for more complete caption sets that are still relatively high quality. Additional qualitative examples are provided in B.6.

et al., 2021), a transformer-based model which is initialized with a large language model, and uses prefix-tuning with CLIP (Radford et al., 2021a) embeddings (Additional details in A.3). Figure 5 illustrates that $\text{TRM}_{\text{METEOR}}$ captures a subtlety in the model comparisons that METEOR does not capture alone: while VLP produces better descriptions at low temperatures, it becomes less fluent (likelihood) on average as we introduce diversity, leading to worse captions when sampling at high diversities. CLIPcap retains better fluency at high sampling temperatures, leading to improved performance in diverse captioning tasks. While $\text{TRM}_{\text{METEOR}}$ demonstrates this, METEOR monotonically decreases, giving little insight into this problem. The sensitivity of the TRM measure is also visible in qualitative samples, given in Figure 4, where we see TRM metrics are sensitive to both diversity and likelihood. These results confirm observations made by Zhang et al. (2021a) for open-ended language generation tasks such as storytelling and dialogue: a fair comparison of approaches must not only compare at the same level of entropy but at a range of entropy levels.

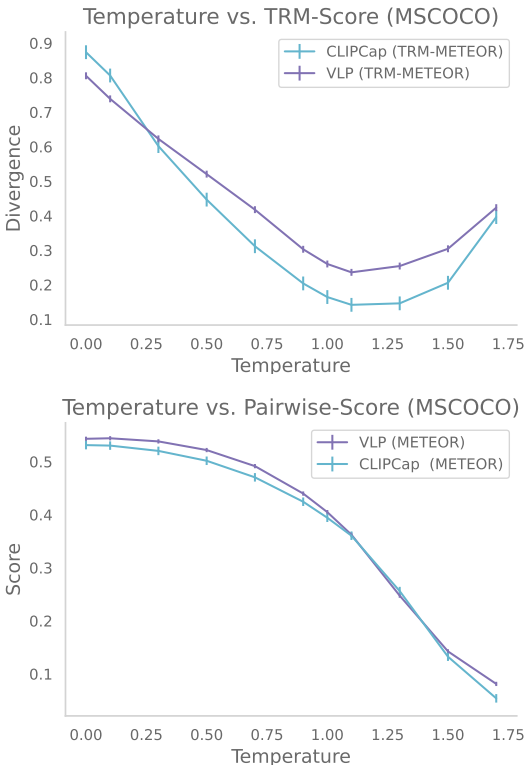


Figure 5: Plots indicating the impact of temperature on the metric scores. Top: $\text{TRM}_{\text{METEOR}}$ (↓) for CLIPcap and VLP. Bottom: Standard METEOR Score (↑) for CLIPcap and VLP.

Sampling algorithms matter Not only does the temperature of the generation process matter when correctly trading off between diversity and description correctness (as seen in the previous discussion), but the sampling process itself matters. Figure 6 shows the performance at different temperatures of the Nucleus sampling method (Holtzman et al., 2020) vs. standard sampling, beam search, and greedy, approaches. While maximum-likelihood methods achieve the best METEOR scores, they have relatively high divergence, as they sample only a single description. Further, Figure 6 shows that $\text{TRM}_{\text{METEOR}}$ illustrates how Nucleus sampling allows models to achieve higher temperatures than standard sampling without diverging significantly from the distribution. METEOR alone does not indicate such an effect and only monotonically decreases.

TRM Measures correlate with human judgements It has long been known that humans are relatively poor at measuring the semantic distance between two sets of objects, particularly in the pres-

Table 2: Method evaluation efficiency on the MS-COCO dataset with 5 references and 10 candidates.

	METEOR	$\text{TRM}_{\text{METEOR}}$	CIDEr	$\text{TRM}_{\text{CIDEr}}$	MMD-BERT	FID-BERT	MAUVE
Samples/Sec	298.4 ± 18.3	161.18 ± 21.2	131.23 ± 12.6	97.54 ± 9.1	53.76 ± 38.7	17.45 ± 4.6	2.29 ± 0.78
Wall Time (Min)	2.26	4.18	5.14	6.92	12.55	38.68	294.78

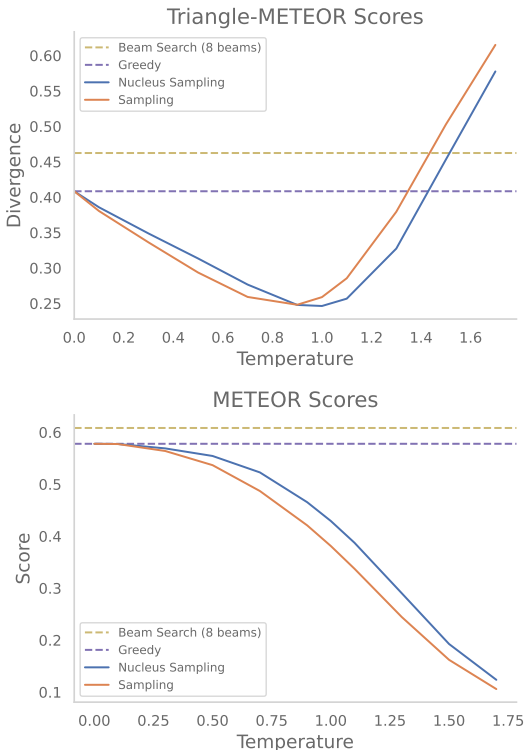


Figure 6: Plots indicating the impact of search technique on divergences. Top: $\text{TRM}_{\text{METEOR}}$ (↓) for TVT on MSR-VTT. Bottom: METEOR Score (↑). See A.8 for experimental details.

Table 3: Pearson Correlation with human judgement, $N = 794$.

Method	Coverage	Correctness
Human	0.2247 ($p < 0.001$)	0.2247 ($p < 0.001$)
TRM-Meteor	0.1278 ($p < 0.001$)	0.1082 ($p < 0.001$)
TRM-BLEU	0.1271 ($p < 0.001$)	0.1510 ($p < 0.001$)
MMD-BERT	0.1288 ($p < 0.001$)	0.1243 ($p < 0.001$)
FID-BERT	0.0807 ($p = 0.011$)	0.0978 ($p < 0.001$)
METEOR	0.0162 ($p = 0.3978$)	0.0057 ($p = 0.7650$)
BLEU-4	0.0044 ($p = 0.8157$)	0.0026 ($p = 0.8884$)
ROUGE	0.0110 ($p = 0.5631$)	0.0381 ($p = 0.1845$)
CIDEr	0.0037 ($p = 0.8445$)	0.0261 ($p = 0.1725$)

ence of distractors (Durga, 1980). While this is the case, we still find that proposed measures correlate with human judgement significantly more than existing measures, which we show in Table 3. To demonstrate the correlation of distributional measures with human judgement of distributional distance, humans were presented with two candidate caption sets (two image captioning models, OFA (Wang et al., 2022) and BLIP (Li et al., 2022) using different temperatures), and asked which candidate caption set correlated better with a reference caption set on two measures: how much they overlapped factually (correctness), and how much information they provided about the references (coverage). Additional experimental details are available in A.9.

Clearly, distributional measures correlate more, and with significantly less information than existing measures aggregated using the max function. Notably, despite evidence that existing decoding methods optimize for fooling humans over correctness (Ippolito et al., 2020), our method *is the only approach* which correlates *at all* with human judgement, suggesting that we have accomplished our goals of being distribution aware, improving the sensitivity of the base measures to human preferences.

5. Discussion and Limitations

Kernel-Based Metrics (KBMs) vs. Triangle-Rank Metrics (TRMs) A natural question to ask is: "which metric should practitioners choose when evaluating conditional language models?" KBMs have one major, distinct, advantage over the TRMs in that they are naturally differentiable, yet KBMs also have downsides. The first is that, unlike the TRMs, they require both a pre-trained BERT model and a kernel-density estimator which both have complex behavior affecting the performance of the model. The TRMs, however, can be specified on top of existing natural language distance functions, improving the ability of the user to intuit the model performance. Additionally, TRMs are bounded and have p-values that can be computed analytically. Finally, because the TRMs do not need a density estimate, they can be more sensitive with small sample sizes (see Figure 3), which is essential for conditional language generation where we have only a few gold-standard samples. Table 2 demonstrates another key benefit of TRMs: efficiency. The time per sample to compute TRMs, while higher than single metric standards, is lower than KBMs on average.

Perplexity We acknowledge that perplexity (likelihood of the test distribution) is another alternative metric to proposed methods. While methods **should** report the perplexity of their models, it is not standard practice, and it has been shown by Theis et al. (2016) that perplexity suffers from several

major issues when evaluating generative models. For example, a lookup table storing sufficiently many training examples will produce convincing results but have poor perplexity on the test data. On the other hand, van den Oord and Dambre (2015) demonstrate that even when perplexity is low, models may not generate high-quality test samples.

Reference-Free Metrics Some metrics, such as CLIP-score (Hessel et al., 2021) for visual description, are immune to ground truth aggregation effects as they are computed in a reference-free way, and focus on pre-trained models' ability to ground vision and language information. Unfortunately, such large, black-box, models represent a liability as a metric as their capabilities are largely unknown, and untested (Floridi and Chiriatti, 2020; Caglayan et al., 2020). Further, the metric is only as good as the model, and CLIP has been known to suffer from numerous issues including counting, attribute-association, and spatial reasoning (Blattmann et al., 2022; Ramesh et al., 2022).

Multi-Candidate Data Availability/Efficiency While multi-candidate evaluation of conditional language generation models represents a significantly more robust paradigm, it still has several drawbacks. One of the core drawbacks is the availability of multi-reference data. Outside the field of visual description, it is often not a standard practice to collect more than one gold-standard reference (even in fields such as summarization, where it makes sense to do so). While the availability of multi-reference data may be a bottleneck for the approach, fortunately, many canonical datasets in the image/video captioning domain (MS-COCO, Flickr-30K, MSR-VTT, VATEX, YouCook II) do contain more than one gold-standard reference, so the methods proposed in this work are immediately applicable to many popular datasets (and domains). Additionally, multi-candidate evaluation is less efficient than existing evaluation techniques, which may encourage an unintended reduction in evaluation. Further, such multi-candidate evaluation methods are somewhat less interpretable than single caption metrics, as they incorporate several axes at once, whereas existing pairwise metrics describe only a single axes of semantic similarity at any time. Still, existing metrics are often used as a "single number" for determining the quality of a model, a task that is better ascribed to multi-candidate metrics.

English-Language Experiments While in theory the TRMs and KBMs introduced in this work are transferable to other languages asides from English, it it important to acknowledge that our experiments were conducted on only English-language data. Transitioning to other languages my require additional work, for instance, languages with rich morphological structures, such as Finnish or Turkish, may require adjustments in kernel density estimations or the tuning of natural language distance (beyond METEOR or BLEU) functions within TRMs to accurately reflect the intricacies of these languages. Additionally, the availability of pre-trained models like BERT, which serve as the backbone for KBMs, is predominantly focused on English, with limited coverage and performance on low-resource languages. This gap necessitates the development or enhancement of multilingual or language-specific models to ensure the applicability and effectiveness of these metrics across diverse linguistic datasets.

6. Conclusion

In this work, we introduce a robust framework for multi-candidate evaluation of conditional language generation models, show that existing metrics for semantic similarity can be seamlessly extended to this framework, and demonstrate that multi-candidate evaluation paired with more sensitive distribution-aware metrics can provide novel insights into existing models and methods. This work is only the beginning. It is necessary for future work to explore how a wider range of existing generation techniques and models perform under this new paradigm, and to understand the implications of distribution-aware evaluation in fields beyond visual description.

Acknowledgements

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

7. Bibliographical References

Nayyer Aafaq, Ajmal Mian, et al. 2019. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, 52(6):1–37.

Martín Abadi, Paul Barham, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

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

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. VATT: transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 24206–24221.

Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Peter Anderson, Basura Fernando, et al. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society.

Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6816–6826. IEEE.

Song Bai, Philip Torr, et al. 2021. Visual parser: Representing part-whole hierarchies with transformers. *ArXiv preprint*, abs/2107.05790.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Natã Miccael Barbosa and Monchu Chen. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 543. ACM.

Michéle Basseville. 2013. Divergence measures for statistical data processing—an annotated bibliography. *Signal Processing*, 93(4):621–633.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR.

Shruti Bhargava and David Forsyth. 2019. Exposing and correcting the gender bias in image captioning datasets and models. *ArXiv preprint*, abs/1912.00578.

Andreas Blattmann, Robin Rombach, et al. 2022. Retrieval-augmented diffusion models. *ArXiv preprint*, abs/2204.11824.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ali Borji. 2019. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65.

Klaus Brinker. 2003. Incorporating diversity in active learning with support vector machines. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 59–66. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry,

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

Rémi Cadène, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 839–850.

Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Joao Carreira, Eric Noland, et al. 2018. A short note about kinetics-600. *ArXiv preprint*, abs/1808.01340.

David M. Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A. Ross, Bryan Seybold, and John F. Canny. 2022. What's in a caption? dataset-specific linguistic diversity and its effect on visual description models and metrics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 4739–4748. IEEE.

William Chan, Navdeep Jaitly, et al. 2015. Listen, attend and spell. *ArXiv preprint*, abs/1508.01211.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.

Ming Chen, Yingming Li, et al. 2018. Tvt: Two-view transformer network for video captioning. In *Asian Conference on Machine Learning*, pages 847–862. PMLR.

Shaoxiang Chen and Yu-Gang Jiang. 2021. Motion guided region message passing for video captioning. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1523–1532. IEEE.

Xinlei Chen, Hao Fang, et al. 2015. Microsoft coco captions: Data collection and evaluation server. *ArXiv preprint*, abs/1504.00325.

Yen-Chun Chen, Linjie Li, et al. 2019. Uniter: Learning universal image-text representations.

Yen-Chun Chen, Linjie Li, et al. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Po-Han Chi, Pei-Hung Chung, et al. 2021. Audio albert: A lite bert for self-supervised learning of audio representation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 344–350. IEEE.

Chung-Cheng Chiu, Arun Narayanan, et al. 2021. Rnn-t models fail to generalize to out-of-domain audio: Causes and solutions. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 873–880. IEEE.

Jinwoo Choi, Chen Gao, Joseph C. E. Messou, and Jia-Bin Huang. 2019. Why can't I dance in the mall? learning to mitigate scene bias in action recognition. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 851–863.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2020. Learning to model and ignore dataset bias with mixed capacity ensembles. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3031–3045, Online. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Brendan Collins, Jia Deng, et al. 2008. Towards scalable dataset construction: An active learning approach. In *Computer Vision – ECCV 2008*, pages 86–98, Berlin, Heidelberg. Springer Berlin Heidelberg.

Martin Cooke, Maria Luisa Garcia Lecumberri, et al. 2010. Language-independent processing in speech perception: Identification of english intervocalic consonants by speakers of eight european languages. *Speech Communication*, 52(11-12):954–967.

Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*, AAAI'05, page 746–751. AAAI Press.

Ido Dagan and Sean P. Engelson. 1995. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, page 150–157, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Stanislas Dehaene, Felipe Pegado, et al. 2010. How learning to read changes the cortical networks for vision and language. *science*, 330(6009):1359–1364.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Yue Deng, KaWai Chen, Yilin Shen, and Hongxia Jin. 2018. Adversarial active learning for sequences labeling and generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4012–4018. ijcai.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ramanand Durga. 1980. Semantic distance and semantic judgment. *Outstanding Dissertations in Bilingual Education Recognized by the National Advisory Council on Bilingual Education, 1979*, page 15.

Le Fang, Tao Zeng, et al. 2021. Transformer-based conditional variational autoencoder for controllable story generation. *ArXiv preprint*, abs/2101.00828.

Joshua Feinglass and Yezhou Yang. 2021a. 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, Online. Association for Computational Linguistics.

Joshua Feinglass and Yezhou Yang. 2021b. 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, Online. Association for Computational Linguistics.

Ronald Aylmer Fisher. 1992. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 776–780. IEEE.

Megan Gilliver, Linda Cupples, et al. 2016. Developing sound skills for reading: Teaching phonological awareness to preschoolers with hearing

loss. *Journal of Deaf Studies and Deaf Education*, 21(3):268–279.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Alex Graves, Abdel-rahman Mohamed, et al. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 5036–5040. ISCA.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391, Stockholm, Sweden. Association for Computational Linguistics.

LF Halliday and DVM Bishop. 2006. Is poor frequency modulation detection linked to literacy problems? a comparison of specific reading disability and mild to moderate sensorineural hearing loss. *Brain and language*, 97(2):200–213.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Lisa Anne Hendricks, Kaylee Burns, et al. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787.

Katherine L. Hermann, Ting Chen, and Simon Kornblith. 2020. The origins and prevalence of texture bias in convolutional neural networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Steven C. H. Hoi, Rong Jin, et al. 2009. Semisupervised svm batch mode active learning with applications to image retrieval. *ACM Trans. Inf. Syst.*, 27(3).

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, et al. 2021. Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6533–6537. IEEE.

Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1419–1429. IEEE.

Jiaji Huang, Rewon Child, et al. 2016. Active learning for speech recognition: the power of gradients. *ArXiv preprint*, abs/1612.03226.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. 2021. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR.

Ziwei Ji, Nayeon Lee, et al. 2022. Survey of hallucination in natural language generation. *ArXiv preprint*, abs/2202.03629.

Ye Jia, Michelle Tadmor Ramanovich, et al. 2021. Translatotron 2: Robust direct speech-to-speech translation. *ArXiv preprint*, abs/2107.08661.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019a. TIGEr: Text-to-image grounding for image caption evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2141–2152, Hong Kong, China. Association for Computational Linguistics.

Ming Jiang, Qiuyuan Huang, et al. 2019b. Tiger: Text-to-image grounding for image caption evaluation, human judgments and metric scores. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 2141–2152.

Jungseock Joo and Kimmo Kärkkäinen. 2020. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, pages 1–5.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. 2019. Learning not to learn: Training deep neural networks with biased data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9012–9020. Computer Vision Foundation / IEEE.

J Peter Kincaid, Robert P Fishburne Jr, et al. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 706–715. IEEE Computer Society.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online. Association for Computational Linguistics.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. 2017. MMD GAN: towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2203–2213.

Jinyu Li, Rui Zhao, Zhong Meng, Yanqing Liu, Wenning Wei, Sarangarajan Parthasarathy, Vadim Mazalov, Zhenghao Wang, Lei He, Sheng Zhao, and Yifan Gong. 2020a. Developing RNN-T models surpassing high-performance hybrid models with customization capability. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 3590–3594. ISCA.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Juncheng Li, Siliang Tang, Linchao Zhu, Haochen Shi, Xuanwen Huang, Fei Wu, Yi Yang, and Yueting Zhuang. 2021. Adaptive hierarchical graph reasoning with semantic coherence for video-and-language inference. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1847–1857. IEEE.

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

of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Xiujun Li, Xi Yin, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Yi Li and Nuno Vasconcelos. 2019. REPAIR: removing representation bias by dataset resampling. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9572–9581. Computer Vision Foundation / IEEE.

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

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51, Phildadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, et al. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Fenglin Liu, Xuancheng Ren, Xian Wu, Bang Yang, Shen Ge, and Xu Sun. 2021. O2NA: An object-oriented non-autoregressive approach for controllable video captioning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 281–292, Online. Association for Computational Linguistics.

Zhenyu Liu and Reza Modarres. 2011. A triangle test for equality of distribution functions in high dimensions. *Journal of Nonparametric Statistics*, 23(3):605–615.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10434–10443. IEEE.

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, Denver, CO, USA, May 06-11, 2017*, pages 5988–5999. ACM.

Vimal Manohar, Pegah Ghahremani, et al. 2018. A teacher-student learning approach for unsupervised domain adaptation of sequence-trained asr models. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 250–257. IEEE.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2020. Sparse text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4252–4273, Online. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2630–2640. IEEE.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ron Mokady, Amir Hertz, et al. 2021. Clipcap: Clip prefix for image captioning. *ArXiv preprint*, abs/2111.09734.

Mathew Monfort, SouYoung Jin, Alexander H. Liu, David Harwath, Rogério Feris, James R. Glass, and Aude Oliva. 2021. Spoken moments: Learning joint audio-visual representations from video descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14871–14881. Computer Vision Foundation / IEEE.

Ladislav Mosner, Minhua Wu, Anirudh Raju, Sree Hari Krishnan Parthasarathi, Ken'ichi Kumatani,

Shiva Sundaram, Roland Maas, and Björn Hoffmeister. 2019. Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 6475–6479. IEEE.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 14200–14213.

Aaron van den Oord, Sander Dieleman, et al. 2016. Wavenet: A generative model for raw audio. *ArXiv preprint*, abs/1609.03499.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, et al. 2017. Automatic differentiation in pytorch.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jesus Perez-Martin, Benjamin Bustos, et al. 2021. Improving video captioning with temporal composition of a visual-syntactic embedding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3039–3049.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 4816–4828.

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. 2021a. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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. 2021b. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, et al. 2022. Hierarchical text-conditional image generation with clip latents. *ArXiv preprint*, abs/2204.06125.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3202–3212. IEEE Computer Society.

Anna Rohrbach, Atousa Torabi, et al. 2017. Movie description. *International Journal of Computer Vision*, 123(1):94–120.

Yossi Rubner, Carlo Tomasi, et al. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121.

Aaqib Saeed, David Grangier, et al. 2021. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and*

Signal Processing (ICASSP), pages 3875–3879. IEEE.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234.

Tobias Scheffer, Christian Decomain, et al. 2001. Active hidden markov models for information extraction. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, IDA '01, page 309–318, Berlin, Heidelberg. Springer-Verlag.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Stanislau Semeniuta, Aliaksei Severyn, et al. 2018. On accurate evaluation of gans for language generation. *ArXiv preprint*, abs/1806.04936.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, Honolulu, Hawaii. Association for Computational Linguistics.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. 2020. Don't judge an object by its context: Learning to overcome contextual bias. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11067–11075. IEEE.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. 2019. Variational adversarial active learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5971–5980. IEEE.

P Slade and Tamás D Gedeon. 1993. Bimodal distribution removal. In *International Workshop on Artificial Neural Networks*, pages 249–254. Springer.

Alan F Smeaton, Yvette Graham, et al. 2019. Exploring the impact of training data bias on automatic generation of video captions. In *International Conference on Multimedia Modeling*, pages 178–190. Springer.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Khurram Soomro, Amir Roshan Zamir, et al. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tejas Srinivasan and Yonatan Bisk. 2022. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 77–85, Seattle, Washington. Association for Computational Linguistics.

Matteo Stefanini, Marcella Cornia, et al. 2021. From show to tell: A survey on image captioning. *ArXiv preprint*, abs/2107.06912.

Jonathan Stroud, David Ross, et al. 2020. D3d: Distilled 3d networks for video action recognition. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 625–634.

Jonathan C Stroud, David A Ross, et al. 2018. D3d: Distilled 3d networks for video action recognition. *ArXiv preprint*, abs/1812.08249.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7463–7472. IEEE.

Gabriel Synnaeve, Qiantong Xu, et al. 2019. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *ArXiv preprint*, abs/1911.08460.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ruixiang Tang, Mengnan Du, et al. 2021. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, pages 633–645.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. 2016. A note on the evaluation of generative models. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Romal Thoppilan, Daniel De Freitas, et al. 2022. Lamda: Language models for dialog applications. *ArXiv preprint*, abs/2201.08239.

Atousa Torabi, Christopher Pal, et al. 2015. Using descriptive video services to create a large data source for video annotation research. *ArXiv preprint*, abs/1503.01070.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Grigorios Tsoumakas and Min-Ling Zhang. 2009. Learning from multi-label data.

Aäron van den Oord and Joni Dambre. 2015. Locally-connected transformations for deep gmms. In *International Conference on Machine Learning (ICML): Deep learning Workshop*, pages 1–8.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Vijay V Vazirani. 2001. *Approximation algorithms*, volume 1. Springer.

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

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence - video to text. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4534–4542. IEEE Computer Society.

Ashwin K Vijayakumar, Michael Cogswell, et al. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *ArXiv preprint*, abs/1610.02424.

Sudheendra Vijayanarasimhan and Kristen Grauman. 2009. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 2262–2269. IEEE Computer Society.

Sudheendra Vijayanarasimhan and Kristen Grauman. 2011. Large-scale live active learning: Training object detectors with crawled data and crowds. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 1449–1456. IEEE Computer Society.

Sudheendra Vijayanarasimhan and Kristen Grauman. 2012. Active frame selection for label propagation in videos. In *Computer Vision – ECCV 2012*, pages 496–509, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sudheendra Vijayanarasimhan, Prateek Jain, and Kristen Grauman. 2010. Far-sighted active learning on a budget for image and video recognition. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3035–3042. IEEE Computer Society.

Elizabeth A Walker, Caitlin Sapp, et al. 2020. Language and reading outcomes in fourth-grade children with mild hearing loss compared to age-matched hearing peers. *Language, speech, and hearing services in schools*, 51(1):17–28.

Luyu Wang, Pauline Luc, et al. 2021a. Multimodal self-supervised learning of general audio representations. *ArXiv preprint*, abs/2104.12807.

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

Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. 2021b. Faier: Fidelity and adequacy ensured image caption evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14050–14059. Computer Vision Foundation / IEEE.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4580–4590. IEEE.

Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R. Hershey. 2017. Student-teacher network learning with enhanced features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 5275–5279. IEEE.

Daniel J Wilson. 2019. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200.

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

Rong Yan, Jie Yang, et al. 2003. Automatically labeling video data using multi-class active learning. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, page 516, USA. IEEE Computer Society.

Bang Yang, Yuexian Zou, Fenglin Liu, and Can Zhang. 2021a. Non-autoregressive coarse-to-fine video captioning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3119–3127. AAAI Press.

Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. 2019. Diversity-sensitive conditional generative adversarial networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kaiyu Yang, Klint Qinami, et al. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558.

Xu Yang, Hanwang Zhang, et al. 2021b. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Ilmi Yoon, Umang Mathur, et al. 2019. Video accessibility for the visually impaired. In *International Conference on Machine Learning AI for Social Good Workshop*.

Christine Yoshinaga-Itano and Allison Sedey. 1998. Early speech development in children who are deaf or hard of hearing: Interrelationships with language and hearing. *The Volta Review*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9127–9134. AAAI Press.

Arnold Zellner. 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021a. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

Ke Zhang, Wei-Lun Chao, et al. 2016. Video summarization with long short-term memory. In *ECCV*, pages 766–782.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5579–5588. Computer Vision Foundation / IEEE.

Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. 2020a. Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 7829–7833. IEEE.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020c. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zi-qiang Zhang, Yan Song, Jian-Shu Zhang, Ian McLoughlin, and Li-Rong Dai. 2020d. Semi-supervised end-to-end ASR via teacher-student learning with conditional posterior distribution. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 3580–3584. ISCA.

Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. 2020e. Object relational graph with teacher-recommended learning for video captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13275–13285. IEEE.

Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14810–14820. IEEE.

Bolei Zhou, Agata Lapedriza, et al. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13041–13049. AAAI Press.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018a. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances*

in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 7590–7598. AAAI Press.

Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. 2018b. End-to-end dense video captioning with masked transformer. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8739–8748. IEEE Computer Society.

Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. 2018c. End-to-end dense video captioning with masked transformer. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8739–8748. IEEE Computer Society.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM.

Appendix

A. Additional Experimental Details

In this section, we discuss additional experimental details for interested readers.

A.1. Code

We make all code/data publicly available for use at https://github.com/CannyLab/vdtk. We hope that releasing our code will help inspire further research and examination into the evaluation of models for visual description.

A.2. Datasets

MSR-VTT Dataset: The MSR-VTT dataset (Xu et al., 2016) is a dataset for video description consisting of 10,000 videos, with 20 reference ground truth descriptions for each video. It was collected by downloading 118 videos for each of 257 queries from a popular video sharing website. MSR-VTT contains 41.2 hours of video, with an average clip length lying between 10 to 30 seconds. It has a vocabulary size of 21,913. For more details about the diversity of the language present in the dataset, we refer readers to Chan et al. (2022).

MS-COCO Dataset: The MS-COCO dataset (Lin et al., 2014) is a large-scale dataset for image description, object detection and segmentation. MS-COCO contains 328K images, each with 5 ground truth descriptions generated by human AMT workers. For more details about the diversity of the language present in the dataset, we refer readers to Chan et al. (2022). MS-COCO is licensed under a Creative Commons Attribution 4.0 license.

A.3. Models

This paper explores the performance of our metrics over several models: two video captioning models, and two image captioning models.

TVT The Two-View Transformer (Chen et al., 2018) is a baseline method for video description, which consists of a transformer encoder/decoder structure. While we did not have access to the original code, we trained our own version of the model on the MSR-VTT dataset (standard splits), leveraging features from Perez-Martin et al. (2021). The model was trained for 300 epochs, with a batch size of 64, model hidden dimension of 512, 4 transformer encoder and decoder layers with 8 heads each, and dropout of 0.5. For optimization, we leveraged the Adam optimizer with a learning rate of $3e^{-4}$ and weight decay of $1e^{-5}$ with exponential learning rate decay with gamma 0.99. This model achieves a $CIDEr$ score of 56.39 on the test dataset. The model was trained using a Titan RTX-8000 GPU over the course of several hours.

O2NA O2NA (Liu et al., 2021) is a recent approach for non-auto-regressive generation of video captions. While the method had available code and checkpoints which we used for this experiment, the method is not designed to sample more than one candidate caption at any given time. To adjust the model to sample multiple candidate captions, we made several adjustments. First, the model was modified to sample a length according to a softmax distribution over the length likelihoods (instead of using a greedy choice of length, or beam search over lengths, as proposed in the paper). Second, the model was modified to sample tokens at each non-autoregressive step from a temperature-adjusted softmax distribution instead of greedily sampling tokens. We make our modified code available as a patch to the original repository, in the hopes that other users will continue to build on these alterations.

CLIPCap CLIPCap (Mokady et al., 2021) is a recent model for image description based on using the CLIP (Radford et al., 2021a) model for large vision and language pre-training as a feature encoder, and GPT (Brown et al., 2020) as a natural language decoder. CLIPCap code and MS-COCO trained model checkpoints are publicly available from the authors, however we made some alterations to support temperature-based and

nucleus sampling. We make our modified code available as a patch to the original repository, in the hopes that other users will continue to build on these alterations. CLIPCap is licensed under the MIT license.

VLP VLP (Zhou et al., 2020) is a unified vision and language pre-training model, designed to perform both image captioning and visual question answering. The model is pre-trained on the Conceptual Captions (Sharma et al., 2018) dataset, and fine-tuned on the MS-COCO captions dataset for image description. The authors make code and pre-trained models publicly available, however we modified the code somewhat to support additional sampling methods. We make our modified code available as a patch to the original repository, in the hopes that other users will continue to build on these alterations. VLP is licensed under the Apache License 2.0.

A.4. Distance Metrics

In this paper, we explore three base semantic metrics as distance underlying our TRM methods, CIDEr-D (Vedantam et al., 2015), METEOR (Agarwal and Lavie, 2008), and BERT Distance (Zhang et al., 2020b).

CIDEr-D CIDEr-D (Vedantam et al., 2015) is a n-gram-based metric designed for visual description, and based on the idea that common words are less useful in practice than uncommon words. In practice, this takes the form of a cosine similarity between TF-IDF weighted vectors representing the sentences. Because CIDEr-D is a score, and not a distance, we create a distance function: $d(c,r) = 10 - C(c,r)$, which works as CIDEr-D is bounded by 10. Note that because CIDEr-D is 10 if and only if and only if the two sentences are equal, this fulfills the TRM requirements.

METEOR METEOR (Agarwal and Lavie, 2008) is a score which evaluates the semantic distance between two text utterances based on one-to-one matches between tokens in the candidate and reference text. The score first computes an alignment between the reference and candidate, and computes a score based on the quality of the alignment. Because METEOR is a score, and not a distance function, we use the distance $d(c,r) = 1 - M(c,r)$, where M is the METEOR score of the reference. Because METEOR is bounded at 1 if and only if the two utterances are identical, this simple transformation satisfies the requirements of the TRM adjustment. While we could explore other ways of deriving a distance from METEOR, we found that this simple approach was sufficient to demonstrate the performance of our methods.

BERT Distance A recent method for determining the semantic distance between two samples is to leverage a pre-trained BERT embedding model to create a semantic embedding of the text, and computing the cosine distance between the test samples. In our work, we leverage the `MiniLM-L6-v2` model from the sentence-transformers package by Reimers and Gurevych (2019) to embed our descriptions. Because cosine distance is already a distance function, no additional transformation is necessary.

A.5. P-value Computations

For our experiments, our null hypothesis is that *the candidate samples and the ground truth samples are drawn from the same distribution*. Because most of the methods do not have an analytical way to compute the p-values (in fact, the TRMs are the only method which has an analytic p-value computation given in Liu and Modarres (2011)), we instead must compute the p-values though sampling. We thus enumerate the value of the statistic across all of the possible candidate/reference partitions given the joint set of candidates and references, and determine the probability of observing the sampled value, or some value more extreme.

The values in Table 1 represent the p-value obtained with a single candidate sentence, and 4 ground truth candidates for MS-COCO, or 19 ground truth candidates for MSR-VTT. We reserve one gorund truth description in both datasets to serve as the “Human" performance description. For TVT, CLIPCap and VLP, we sample the descriptions using beam search with 16 beams. For O2NA, which is a non-autoregressive model, we sample according to the method suggested in the original work (see Liu et al. (2021)). Because there are several thousand videos per dataset, computing all possible combinations across the dataset would be far from tractable. Thus, the p-values were computed on a per-visual-input basis, and then aggregated across videos using the harmonic mean, as suggested by Wilson (2019). Such an aggregation method is valid when the experiments are not independent (which they are not), unlike Fischer's method (Fisher, 1992).

Figure 3 demonstrates the log p-values for the proposed methods across several candidate samples. For MS-COCO, we use all five reference captions, and between one and ten candidate captions sampled from

CLIPCap using Nucleus Sampling (Holtzman et al., 2020) with a temperature of 1.0, top-p of 0.9 and top-k of 20. The caption set is generated once, meaning that the two-candidate set consists of the one-candidate set and one more additional caption. For MSR-VTT, we use 10 reference captions, and between one and seven candidate captions sampled from O2NA as described in appendix A.3 with a temperature of 1.0 for both the length and token samples. We do not go to the full 10 candidate captions for MSR-VTT due to tractability concerns, since adding an additional caption forces twice the number of partitions to be evaluated when computing p-values.

The above experiments were performed on several n2d-standard-32 cloud GCP instances, containing 32vCPUs and 128GB of RAM.

A.6. Frechet BERT Distance

The Frechet Inception Distance, originally proposed in Salimans et al. (2016), has often been used for the evaluation of the distance between samples of images generated by GANs. Images are first embedded in a latent space using a pre-trained inception network, and then the Frechet distance between the generated samples and the reference samples is computed. In our work, we replace the images with text, and the inception network with a pre-trained BERT embedding network (Devlin et al., 2019). For a set of candidate samples $(c_1,...,c_n) = C$, a set of reference samples $(r_1,...,r_m) \in R$, and a BERT embedding function $\phi_{\text{BERT}} : C \cup R \rightarrow \mathbb{R}^k$, we compute the Frechet BERT Distance as:

$$\begin{aligned} d^2 = &\left|\left| \frac{1}{n}\sum_{i=1}^{n}\phi_{BERT}(c_i) - \frac{1}{m}\sum_{i=1}^{n}\phi_{BERT}(r_i) \right|\right| \\ &+ \text{Tr}\left(C_C + C_R - 2\sqrt{C_C C_R}\right) \end{aligned} \tag{5}$$

where C_C andC_R are the covariance matrices of the C and R sets embedded with ϕ_{BERT} respectively.

To get the BERT embedding, we leverage the CLS token of a large pre-trained model, in this case, the `MiniLM-L6-v2` model from the sentence-transformers package by Reimers and Gurevych (2019).

The computation of p-values for the Frechet-BERT distance is largely bottle-necked by the slow performance of the `sqrtm` function, which, because the matrices are not symmetric, has no efficient algorithm for computation. Additionally, unlike the feature computation, this operation must occur for every partition, leading to significantly reduced efficiency compared to the other measures presented in this paper.

A.7. MMD-BERT

Another common metric in the GAN literature is the computation of a maximum-mean discrepancy between kernel-estimates of the samples introduced by Li et al. (2017). For a set of candidate samples $(c_1,...,c_n) = C$, a set of reference samples $(r_1,...,r_m) \in R$, and a BERT embedding function $\phi_{\text{BERT}} : C \cup R \rightarrow \mathbb{R}^k$, we compute the MMD-BERT distance as:

$$\begin{aligned} M\hat{M}D = &\sum_{i=1}^{N}\sum_{j=1}^{N} K(\phi_{\text{BERT}}(c_i), \phi_{\text{BERT}}(c_j)) \\ &+ \sum_{i=1}^{M}\sum_{j=1}^{M} K(\phi_{\text{BERT}}(r_i), \phi_{\text{BERT}}(r_j)) \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{M} K(\phi_{\text{BERT}}(c_i), \phi_{\text{BERT}}(r_j)) \end{aligned} \tag{6}$$

where K is a kernel function. In our experiments, we use an RBF kernel function with σ equal to the median distance pairwise distance divided by two.

A.8. Search Techniques

In section 3, Figure 6, we explore the performance of several different search techniques for our two-view transformer model on the MSR-VTT dataset. In this figure, we explore four decoding search techniques: Greedy Search, Beam Search, Temperature-Based Sampling, and Nucleus Sampling. For each method, and for each video in the test set, we sample 10 descriptions. For Greedy Search, we sample 10 repeated sentences. For beam search we sample the top beam search candidate, and repeat this ten times. While

we did explore using the top 10 results from a larger beam search, we found that a smaller beam search and repeated values produced better METEOR scores, so we chose to compare against this. Wider beam searches did produce higher $\text{TRM}_{\text{METEOR}}$ scores, but because optimizing for METEOR would be the current paradigm, we decided to include that in the referenced figure. For standard temperature based sampling, we sampled 10 results at each temperature. For Nucleus sampling, we sample 10 results at each temperature, however we freeze they hyper-paramters of top-p at 0.9 and top-k at 20, as we found these values to generate the best scores under the standard pairwise metrics. It remains relevant future work to perform a deep-dive into the different generative methods with respect to TRMs, as there are likely many interesting lessons that can be learned.

A.9. Correlation with Human Judgement

In our work, we run a human correlation experiment to determine how well human ratings correlate with our metric's judgements. The following study was granted exception by the University of California IRB, Protocol Number 2022-11-15846. A screenshot of our evaluation tool for mean opinion scores is given in Figure 7. In each HIT, raters from Mechanical Turk were presented with the reference captions, along with two sets of candidate captions. These candidate captions were sampled from two models: OFA (Wang et al., 2022) and BLIP (Li et al., 2022), at 11 different temperate settings: 0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0. We then query the subjects with two questions, both of which can be evaluated on a scale of $\{-2,2\}$, with 0 indicating a tie:

- Which group of candidate captions (as a whole) provides more useful information about the reference group for a person who cannot see the reference group?
- Which group of candidate captions (as a whole) matches best to the reference group factually?

Subjects are linked to the data collection interface on our server developed by us in a frame directly from an Amazon Mechanical Turk internal HIT using the ExternalQuestion API which allows external web content to be displayed within the internal HIT. No third-party software is used with the HITs and no reviewing data is collected by Amazon or any third-parties with the use of this API. The subjects are shown a consent form on the Amazon Mechanical Turk HIT prior to entering our data collection interface. Subjects are then required to click the "I Accept" button to confirm their agreement with the consent information of the study. They are then redirected to the data collection interface. For each image, users are presented with an image, and an associated image description. Images are drawn from the MSCOCO dataset (Lin et al., 2014). Human generated captions are drawn from the references collected by the authors of (Lin et al., 2014).

After completing all of the tasks in the session, users are given a randomly generated code, which is entered in the Amazon MTurk HIT page, and links the user's survey results to the Amazon worker ID. We collect these linkings to perform analysis on inter-rater agreement, as while the session itself is anonymous, users may complete multiple sessions, and some method is required to maintain identity between the sessions.

After each of these sessions, subjects will be given a brief survey regarding the task difficulty (Select from the options: "Very Easy", "Easy", "Normal", "Hard", "Very Hard") and prompted for any additional comments on the session in general for each session in an (optional) open-response format. Users are also encouraged to protect their privacy with the prompt: "After submitting your responses, you can protect your privacy by clearing your browser's history, cache, cookies, and other browsing data. (Warning: This will log you out of online services.)" Subjects were compensated with $0.18 USD per session (based on the recommended Amazon wage (federal minimum wage, $7.25/Hr), with an expected completion time of 1.5 minutes per session), and should be able to complete the session in under one and half minutes (based on several pilot examples). Subjects can participate in the task a maximum of 100 times. The maximum time commitment for each subject over two months of our study is 2 hours.

We analyze the experiments by first collecting all human ratings, and taking the mean of each score per image. We collect 5 ratings each for 794 images in the dataset, using 397 unique Mechanical Turk workers. We then compute the Pearson correlation for the standard max-aggregate scores, and for each of our methods against the mean of the human ratings. To compute the human-human correlation, we compute first the leave-one-out mean for each human rating, and compute the correlation of the leave-one-out mean with the existing images.

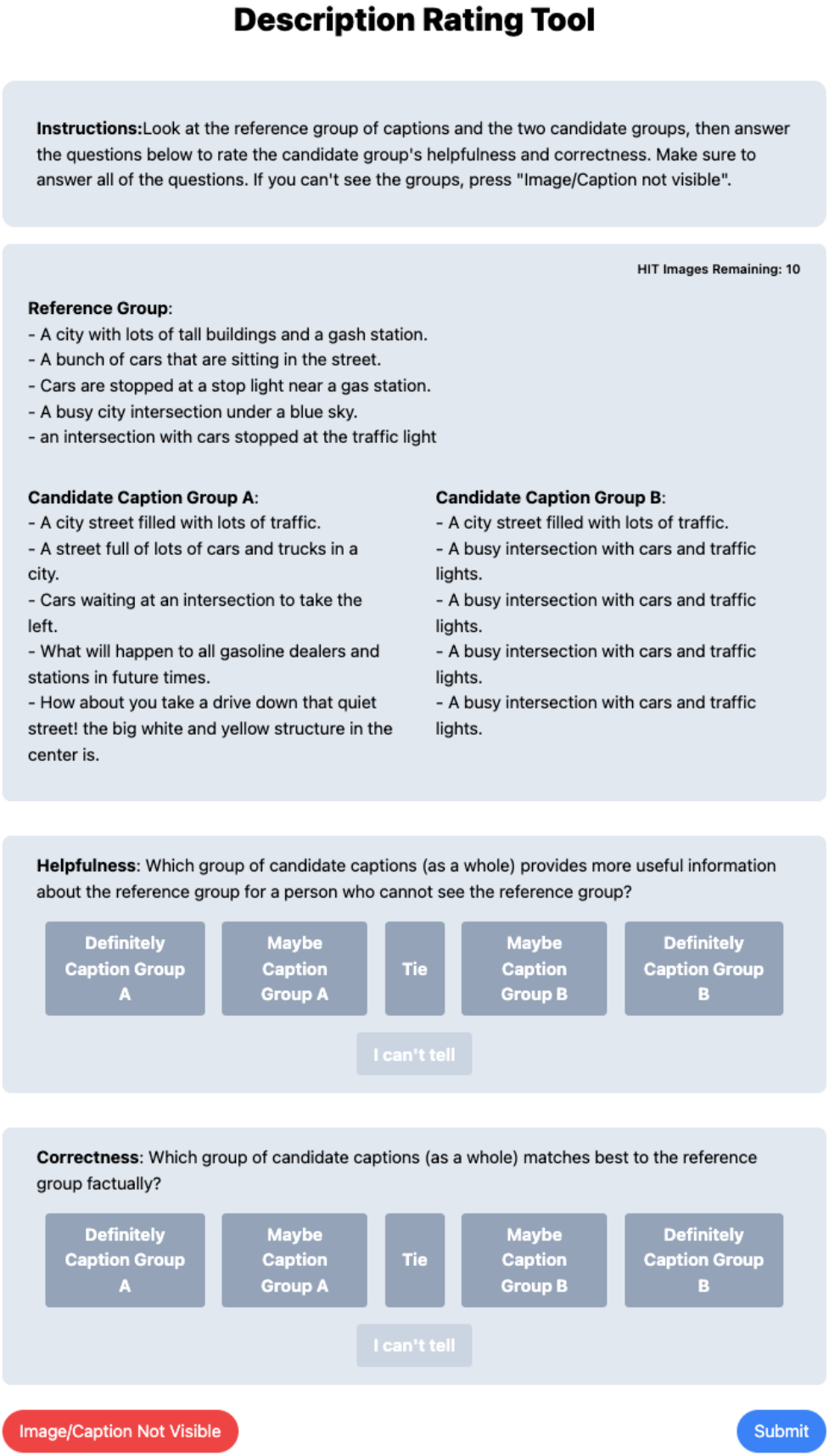


Figure 7: A screenshot of our human rating interface.

B. Additional Results

In this section we present several additional interesting results to augment those in the main discussion.

B.1. Embedding Methods for KBMs

In the main work, we primarily explore a BERT-based embedding method for the kernel-based methods. Such an exploration does not preclude the use of other embedding methods, each of which has different trade-offs, when looking at the quality of the resulting metric, what the resulting metric measures, the time required to compute the embedding, and the performance when the reference distribution is limited to small numbers of human samples (such as happens in practice). Figure Figure 8 shows a quick look at several possible choices for embedding methods in the MMD-* family, including Bag of words (with a 5K vocab), GLoVe (Pennington et al., 2014), FastText (Bojanowski et al., 2017), and CLIP (Radford et al., 2021b).

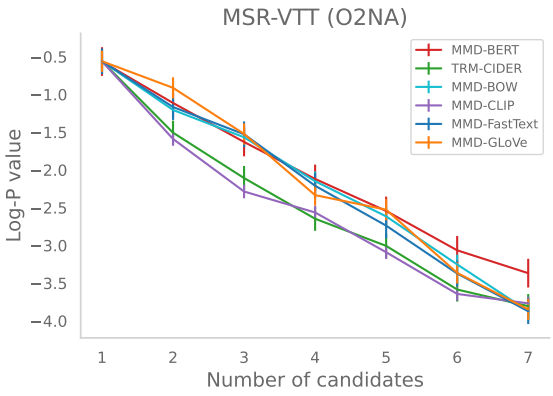
While we can see that some of the methods are more sensitive to deviations in the image distributions, such methods come with additional trade-offs. CLIP-style embeddings are the most sensitive to human versus generated captions with fewer captions created, but are significantly slower to evaluate at test time (almost 4x slower) than MMD-BERT, and also produce a higher p-value when computing the leave-one scores on the human captions (which is less desirable, as the human captions are drawn from the same distribution).

B.2. Unique vs. Correct Descriptions

In Figure 9, we explicitly demonstrate how TRMs enable evaluation of both caption diversity and quality. We artificially generate candidates for the MSR-VTT dataset by mixing human-generated exact descriptions with human-generated descriptions from other videos. On one axis we have the number of unique descriptions and on the other axis we have the number of correct (exactly-matching) descriptions. Clearly, unlike METEOR alone, $\text{TRM}_{\text{METEOR}}$ scores are affected by both correctness and diversity.

Each experiment consisted of 10 candidate captions from the MSR-VTT dataset, and 10 reference captions from the MSR-VTT dataset. We first split the 20 MSR-VTT reference captions into two sets of 10. One set of 10 captions formed the references. To select the candidate captions, we first sampled k unique captions from the remaining reference set (which formed the “correct pool"), and k unique captions from other videos in the dataset at random (forming the “incorrect pool"). We then selected m correct captions, from the correct pool (at random) and $10-m$ captions from the incorrect pool (at random). This was then plotted with m on the x-axis, and k on the y-axis, as a heat-map, where lighter colors represent better scores (higher METEOR, or lower TRM-METEOR), and darker colors represent poor scores.

We also explored the performance of the CIDEr metric across the same axes, the results of which are shown in Figure 10. We can see that they are largely similar to those from the METEOR metric, suggesting that regardless of the underlying metric, we are still making similar trade-offs between diversity and correctness.



Sensitivity and performance on Human-Generated Captions

Method	Log-P	Samples/Sec
TRM-CIDEr	-1.596	88.93
MMD-BERT	-1.786	56.68
MMD-CLIP	-1.887	14.41
MMD-GLoVe	-1.952	54.8
MMD-FastText	-1.954	57.45
MMD-BOW	-2.022	49.41

Figure 8: Performance of several different embedding functions for the MMD-* family of metrics. Left: Sensitivity when evaluated on the MSR-VTT dataset with ten reference captions and between one and seven candidate captions generated by O2NA. Right: Sensitivity and speed when evaluated on human reference samples with 5 references and 5 candidates.

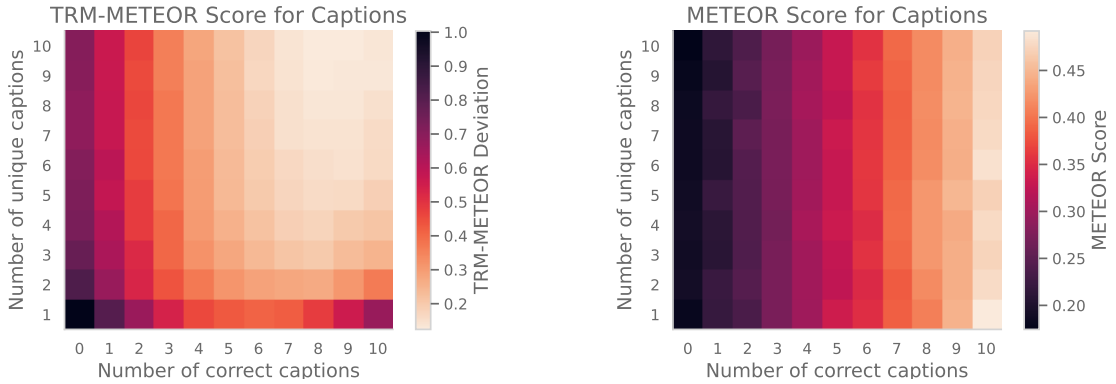


Figure 9: Plots showing how TRMs evaluate both diversity and quality. Left: $\text{TRM}_{\text{METEOR}}$, Right: METEOR. Lighter colors represent better scores. While $\text{TRM}_{\text{METEOR}}$ trades off between diversity and quality, METEOR focuses only on quality not diversity.

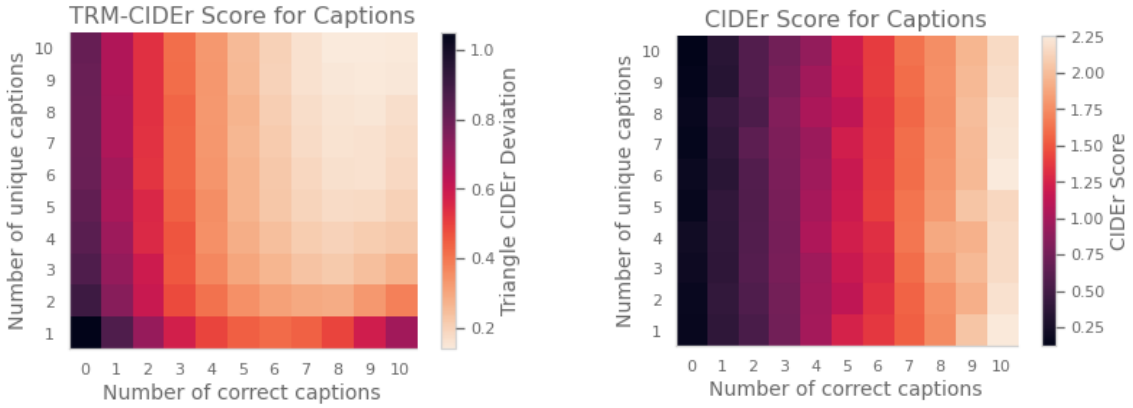


Figure 10: Plots showing diversity vs. quality tradeoffs. Left: $\text{TRM}_{\text{CIDEr}}$, Right: CIDEr. Lighter colors represent better scores. While $\text{TRM}_{\text{CIDEr}}$ trades off between diversity and quality, CIDEr focuses only on quality not diversity.

	METEOR	$\text{TRM}_{\text{METEOR}}$	CIDEr	$\text{TRM}_{\text{CIDEr}}$	BERT	TRM_{BERT}	MMD-BERT
MSCOCO	-0.6303	-0.5941	-0.5957	-0.4742	-0.6230	-0.5633	-0.6550
MSR-VTT	-1.0046	-0.9613	-1.0224	-0.9777	-1.0172	-1.040	-1.0374

Table 4: Log P-Values on human leave-one our samples. We can see that, surprisingly, none of the methods (even the standard aggregations) produce statistically signficant differences. That being said, TRMs often produce higher p-values, indicating that they may be more robust to noise in human caption sets. We do not compute the Frechet-BERT values for humans here, as it was prohibitively expensive.

B.3. Human p-values

Strong metrics for distributional comparison will have high sensitivity to samples coming from distinct distributions, and will produce high p-values for samples which come from the same distribution. To check that such a relationship holds, we also perform leave-one-out experiments using human-generated captions from the reference set for both MSR-VTT and MS-COCO. For MSR-VTT, we split the reference data into sets of 10 candidate samples and 10 reference samples, and compute the deviations using this partitioning. For MS-COCO, we leverage the c40 split which has 40 reference descriptions for 5000 samples of the ground truth. We partition the references for each video into groups of ten descriptions, and compute the p-values from pairs of these partitions. Table 4 gives the performance of the metrics on this human data.

Dataset	MAUVE Log p-value	METEOR Log p-value
MSR-VTT (O2NA)	-0.4414	-1.7881
MSR-VTT (Human Captions)	-0.1441	-0.6037
MS-COCO (CLIPCap)	-0.3980	-2.5585
MS-COCO (VLP)	-0.3234	-2.8609
MS-COCO (Human Captions)	-0.2189	-0.7233

Table 5: Log p-value estimates for MAUVE using five candidates, five references, and 100 samples (at nucleus sampling temperature 1.0 for O2NA, CLIPCap and VLP models). We can see that Log p-values for MSR-VTT and MS-COCO are signficantly worse than METEOR even with aggregation, likely due to the method using k-means to approximate the text distributions with only 5 samples.

B.4. MAUVE performance

In the main work, we found that MAUVE was prohibitively slow to use to compute p-values for the training data. Because our p-values were computed with 10 reference sentences, and up to 10 candidate sentences, at the existing rate, it could take several years to compute the MAUVE p-values for the 50,000 sample MS-COCO dataset. In Table 5, we present several high-variance estimates of the MAUVE p-values (computed using only 100 samples).

Two hot dogs sitting side by side with condiments.
Two hot dogs are laden with relish, ketchup, and mustard.
>>> two hot dogs on a plate loaded with condiments
Two hot dogs covered with ketchup and relish on a plate.
Two hot dogs in buns are smothered with condiments.

The meal is ready on the tray to be eaten.
A breakfast was delivered to a hotel room on a tray.
a bunch of food and stuff is laying on a tray
>>> Bananas, cereal, juice and other breakfast foods on a tray.
This tray includes several different items for a full breakfast.

Figure 11: Qualitative example of “central" captions. The caption marked with arrows is the ground truth caption which minimizes the expected METEOR distance to the other reference captions.

B.5. Visualizing Central Descriptions

We have found that descriptions which minimize the expected distance to the ground truth distribution are relatively sparse in detail compared to other descriptions. Figures 11, 12, 13 and 14 show qualitative examples of such descriptions for the MS-COCO dataset. Each plot shows qualitative examples of “central" captions. The caption marked with arrows is the ground truth caption which minimizes the expected METEOR distance to the other reference captions, and the other captions are the additional references in the MS-COCO dataset. Images are selected at random, and do not represent cherry-picked samples from MS-COCO.

A photo taken from a boat with a long bridge in the background.
A view of the coast from within a boat
The side of a boat and a bridge going over the ocean.
>>> A view of the lake, taken from a boat.
A boat flies its flag while sailing just off a pier.

a microwave on a kitchen counter above a dishwasher
this micro wave is black and silver and is on the counter
>>> A microwave oven sitting on top of a counter.
A microwave sitting on a counter, its stainless steel.
a silver microwave oven on a tan counter and a window

Figure 12: Qualitative example of “central" captions. The caption marked with arrows is the ground truth caption which minimizes the expected METEOR distance to the other reference captions.

A narrow city street has a leaning one way sign.
>>> a street with a line of cars parked on the side
Cars are parked alongside the road and a man is standing next to a sign.
A man is standing next to a road sign with a line of parked cars across the street in an urban area
A crooked one way sign pointing into the ground

A person pressing a button on a Wii controller.
A hand holds a remote that operates a video game.
There are no image to describe on this page..
>>> A person is holding a white Wii control
someone that is holding a wii remote in their hand

Figure 13: Qualitative example of “central" captions. The caption marked with arrows is the ground truth caption which minimizes the expected METEOR distance to the other reference captions.

Mom gives her daughter a lesson in using her baseball glove.
>>> Mother and her son playing in a few
two girls in red shirts grass and a baseball glove
A woman playing catch with her young child.
The mom is teaching her daughter to play baseball

a blue truck and a male in a purple shirt and a tree
Blue pickup truck filled with scrap pieces of household items.
A man has filled his truck with wheelchairs.
>>> A blue truck parked next to a tree and a man.
a man standing next to a truck full of bikes and a wheel chair

Figure 14: Qualitative example of “central" captions. The caption marked with arrows is the ground truth caption which minimizes the expected METEOR distance to the other reference captions.

B.6. Additional Qualitative Samples

MSCOCO Image 421060

Candidate Set 1
METEOR (↑): 0.236
TRM-METEOR (↓): 0.912

A person on a snowboard does a trick in the air.
A person on a snowboard does a trick in the air.
A person on a snowboard does a trick in the air.
A person on a snowboard does a trick in the air.
...

Candidate Set 2
METEOR (↑): 0.264
TRM-METEOR (↓): 0.362

A person in mid air on a snowboard in front of a TV.
A snowboarder getting some air after a jump.
A man is performing a ski jump on a green slope.
A person on skis going down a ramp.
...

References

Competitive spirit during a competition in mid air
A skier races down the track at a competition.
A person is skiing on a slope covered in snow.
The skier is jumping into the air in a half pipe.
Skier performing aerial jump during outdoor competition.

Figure 15: A qualitative sample from CLIPcap. Candidate set one uses beam search (8 beams), while candidate set two uses nucleus sampling (with temperature one, top-k of 20 and top-p of 0.9).