What's in a Caption? Visual Description Linguistic Patterns and Their Effects on Models and Metrics

Supplementary Material

A. Datasets

We investigate four primary datasets in this work. An overview of the datasets is given in Table S1.

MSR-VTT: The MSR-VTT (MSR Video to Text) [XMYR16] dataset is a medium-scale open domain benchmark for visual description. It was originally collected using 257 YouTube search queries across 20 categories, with 118 videos collected for each query (41.2 Hours). The dataset is annotated with 20 captions per video by 1,327 Amazon Mechanical Turk workers. Each video has a duration between 10 and 30 seconds, with an average of two shots per clip.

VATEX: The VATEX dataset [WWC+19] is a medium-scale open domain video description benchmark, based on a subset of the Kinetics-600 dataset for action recognition. VATEX consists of 41,269 video clips, and each clip is annotated with 10 unique descriptive captions by 2,159 Amazon Mechanical Turk workers.

MSVD: The MSVD (Microsoft Video Description) dataset [CD11] is a small-scale open domain benchmark for video description comprised of 1,970 YouTube clips of 4-10 seconds each, collected by asking Amazon Mechanical Turk workers to link a video, start time, and end time from YouTube that depicts a specific, short action. Each video is then annotated with an average of 41 ground truth descriptions by 835 Amazon Mechanical Turk workers.

MSCOCO: The Microsoft Common Objects in Context (MS-COCO) [LMB+14] dataset is a large-scale open-domain benchmark for image description. MS-COCO consists of more than 120,000 images of complex scenes including people, animals, and common objects. Each image is annotated with five ground truth descriptions.

B. Experimental Details

In this section, we present detailed experimental details corresponding to our experiments. Along with these experimental details, we make the code for our work available at https://github.com/CannyLab/vdtk. Note that numbers may differ slightly between the released code, and our presented experiments due to the tokenization scheme. For our released code, we use the Spacy[3] tokenizer to compute all metrics, as it is significantly more efficient in practice than the Stanford tokenizer[4], however for academic purposes, we compute the metrics with the Stanford tokenizer to avoid tokenization shift. In most cases, the difference in the metrics between tokenization methods is negligible (or very small).

B.1. Motivation: Leave One Out Ground Truth Performance

To generate an estimate of human performance on the selected datasets, we use a procedure called "leave one out" performance. Let a dataset $\mathcal{D}$ be composed of N samples $S_0 \dots S_N$. For each sample S_i, there may be K_i possible reference captions, $C_0^i \dots C_{K_i}^i$. In order to compute the leave one out performance of human samples for the dataset, we first select a hypothesis caption $H_i \in \{C_0^i \dots C_{K_i}^i\}$. We then compute the updated reference set $R_i = \{C_0^i \dots C_{K_i}^i\}/\{H_i\}$. In the case that H_i is duplicated within R_i, *we allow the duplicate to remain* to maximize the possible human score. In the case that there is only one (or fewer) captions for a video, we drop those captions from the computation. We then use the reference sets $R_0 \dots R_N$ and hypotheses $H_0 \dots H_N$ to compute the "leave-one-out" score for the dataset. Clearly, this is an estimate of the ground truth performance, as it is a random sample of the possible "leave-one-out" hypotheses sets.

Because some of the metrics (particularly CIDEr) are dataset dependent, it would be intractable to compute all possible hypotheses sets. Instead of computing all possible hypotheses sets, we perform 750 iterations of this sampling procedure and use the mean of the iterations to achieve our final "leave-one-out" estimates presented in the paper. We found empirically that 750 iterations were sufficient across all of the datasets to achieve a stable mean. The raw values of the "leave-one-out" estimates are presented in Table S2, alongside the state of the art results.

B.2. Motivation: Semantically Masked Leave One Out performance

To test the performance of ground truths without semantic information, we devised an experiment based on the leave-one-out experiments above, however, focused on removing semantic information. To compute this value, we select hypotheses as in subsection B.1, however for both the captions in the reference and the captions in the ground truth, we replace any token identified by the Spacy part of speech

[3] https://spacy.io/

[4] https://nlp.stanford.edu/software/tokenizer.html

Dataset	Domain	Categories	Videos	Avg. Length	Length (hrs)	Annotations / Video	Annotation Method
MSR-VTT	open	20	10K	20s	41.2	20	AMT
VATEX	open	600	42K	-	-	10	AMT
MSVD	open	218	1970	10s	41	35.5	AMT
MS-COCO	open	-	120K	-	-	5	AMT

Table S1. An overview of the datasets that we analyze in this paper. All of the datasets are open-domain, with a focus on video description. Additionally, each of the datasets include more than one ground truth description per video, which we use to validate the performance of ground truth data, without collecting additional human results. Notably, all of these methods use AMT as their annotation method.

Dataset	BLEU@4	METEOR	ROUGE	CIDEr
MSVD	0.453 (0.644)	0.370 (0.419)	0.689 (0.795)	1.038 (1.115)
MSR-VTT	0.209 (0.472)	0.247 (0.312)	0.487 (0.648)	0.426 (0.600)
VATEX	0.234 (0.342)	0.249 (0.235)	0.478 (0.503)	0.611 (0.576)
MS-COCO	0.152 (0.410)	0.228 (0.311)	0.438 (0.609)	0.788 (1.409)

Table S2. Raw leave-one-out score estimates for each of the datasets (SOTA in parentheses).

Dataset	BLEU@4	METEOR	ROUGE	CIDEr
MSVD	0.289 (0.453)	0.097 (0.370)	0.442 (0.689)	0.502 (1.038)
MSR-VTT	0.123 (0.209)	0.085 (0.247)	0.387 (0.487)	0.327 (0.426)
VATEX	0.132 (0.234)	0.201 (0.249)	0.391 (0.478)	0.511 (0.611)
MS-COCO	0.079 (0.152)	0.198 (0.228)	0.396 (0.438)	0.684 (0.788)

Table S3. Raw leave-one-out score estimates under semantic masking for each of the datasets (Non-masked in parentheses).

analysis as a noun, proper noun, or verb with a *unique* mask token. This means that this unique mask token will achieve a 0 in any associated token-based metric, as it will not match any semantic token in the ground truth. Table Table S3 gives the full performance on each of the datasets in the masked setup.

B.3. Caption Diversity: Token Metrics

In this work, we compute several metrics based on token-level diversity, demonstrated in Table 1 from the main paper. The number of unique tokens is equal to the number of tokens in the dataset as computed by the Stanford PTB tokenizer. This number does not do any lemmatizing or stemming, thus, is an upper bound for the vocabulary complexity. We then compute three additional metrics, the within-sample uniqueness, the between-sample uniqueness, and the 90% head of the vocabulary. The within-sample uniqueness corresponds to the percentage of tokens that are unique within a sample - i.e. the percentage of tokens that appear exactly once among the references for any particular image or video. We then average this number over all of the samples to get the number presented in Table 1. The between-sample uniqueness is a measure of the percentage of tokens in each sample that are unique at the *dataset* level, i.e. the percentage of tokens among the tokens in the reference set of a single sample that do not appear in any other caption in the dataset. These per-sample numbers are then averaged across the dataset to get the number presented in Table S4. Finally, the 90% head corresponds to the number of tokens that make up 90% of the mass of the total number of tokens in the dataset. This is an approximate measure of how long-tailed the distribution is. The 90% number is selected empirically (further analysis could look at the full cumulative distribution of the token counts). Table S4 replicates Table 1 from the main paper, however includes between-sample token uniqueness.

Dataset	Unique	BS-Unique	WS-Unique	Head
MSVD	9455	1.21%	11.8%	944
MSR-VTT	22780	0.76%	21.55%	1636
VATEX	31364	0.33 %	24.87%	1363
MS-COCO	35341	0.22%	33.76%	824

Table S4. Vocabulary metrics for each of the datasets. Unique: The number of unique tokens. BS-Unique: Average percent of tokens per description that are unique. WS-Unique: Average percent of of tokens that are unique within a sample. Head: The number of unique tokens comprising 90% of the total tokens.

We also compute many of the same metrics restricted to counting nouns and verbs (as identified by the Spacy POS tagger). Each of the above metrics is computed the same way, however instead of considering all tokens, we consider only tokens that are tagged as either nouns or verbs during the computation of the metrics. Table S5 demonstrates the full results of this experiment, plus an additional metric: the average number of tokens per caption which also appears in Table 2 in the main paper.

B.4. Caption Diversity: N-Gram Metrics

To explore the diversity of samples at an n-gram level, we introduce two novel metrics, the Expected Vocab Size @ N (EVS@N), and the Expected Number of Decisions @ N (ED@N). Both of these metrics measure the diversity of the language at an n-gram level by exploring the properties of an n-gram language model trained on the dataset. In this section, we discuss the explicit definition of these metrics. For all n-grams, we use an n-gram language model based on tokens extracted with the Stanford PTB tokenizer. In all

Dataset	WSNU	BSNU	WSVU	BSVU	NC	VC	NH	VH	NPC	VPC	TPC
MSVD	12.6%	1.9%	14.8%	1.5%	4985	1773	755	229	2.39	1.10	7.03
MSR-VTT	23.1%	1.2%	29.4%	0.8%	12697	3639	1512	293	3.28	1.32	9.32
VATEX	26.9%	0.67%	35.7%	0.3%	16670	4975	1161	338	4.37	2.10	15.29
MS-COCO	34.9%	0.41%	55.8%	0.2%	20155	4200	723	184	3.71	1.02	11.33

Table S5. Part of speech distributions for each of the datasets. DS: Dataset. WSNU: Within sample noun uniqueness. BSNU: Between sample noun uniqueness. WSVU: Within sample verb uniqueness. BSVU: Between sample verb uniqueness. NC: Unique noun count. VC: Unique verb count. NH: Noun head (90% of mass). V: Verb Head (90% of mass). VPC: Average number of verbs per caption. NPC: Average number of nouns per caption. TPC: Average number of tokens per caption.

cases, we pad the references with $[BOS]$ and $[EOS]$ tokens to allow the model to handle the beginning and end of the sequences. For WikiText-103, we create individual reference sentences by splitting on '.' tokens, and pad each of these references individually with $[BOS]$ and $[EOS]$ tokens.

B.4.1 Expected Vocab Size @ N

The EVS@N metric is a measure of how many n-grams *do not* act as 1-grams in practice in the dataset. This measure is computed by looking at the entropy of the next-token distribution of an n-gram language model. For a sequence of words $w_0, \ldots w_{n-1}$, we first compute the distribution $P(w_n|w_0, \ldots, w_{n-1})$. If this distribution has 0 entropy (i.e. it assigns all of the probability mass to a single next token), then we consider this n-gram a "static n-gram". If the entropy is non-zero, then we consider it a "dynamic n-gram". The EVS@N can then be computed as the proportion of dynamic n-grams

$$\text{EVS@N} = \frac{|\text{dynamic n-grams}|}{|\text{static n-grams}| + |\text{dynamic n-grams}|}$$

This measures a set of effective n-grams in the data (i.e. the size of the n-gram vocab), as it coalesces n-grams where no decisions are made into a single logical unit.

B.4.2 Expected Decisions @ N

The ED@N metric is a measure of how many decisions an n-gram language model has to make for a sequence of N tokens. ED@N is a counting measure of the EVS@N - i.e. how many dynamic n-grams are expected in a sequence of length n. For a $K-gram$ language model, this measure is explicitly computed as:

$$\text{ED@N} = 1 + \sum_{i=1}^{N-1} (1 - EVS@K)(0) + (EVS@K)(1)$$

In this work, for the first token we use a 2-gram language model ($K = 2$), for the second token we use a 3-gram language model ($K = 3$), and for any additional tokens, we use a 4-gram language model ($K = 4$).

B.5. Sample Diversity: Within Sample Diversity

We use several techniques to measure the within-sample semantic diversity of the data. In all of these cases, the notion of semantics is somewhat subjective. In this work, we use a BERT-style embedding trained for sentence similarity, called MP-Net [STQ+20] to embed each reference description as a 384-dimensional vector. We leverage the implementation in Sentence Transformers[5], which is pre-trained on over 1 billion sentence pairs.

Figure 2 measures the minimum within-sample distances, i.e. it looks for the closest pair of references in each sample, and plots the distance between them. Thus, for a dataset of length N with a set of samples $S_0 \ldots S_N$ and captions $S_i^0 \ldots S_i^{K_i}$, this histogram plots the distribution over all descriptions of

$$H_{ij} = \min_{k \neq n} ||S_i^k - S_i^j||$$

In order to avoid obvious issues with repetition in the semantics, we use only the unique set of captions in a sample, as opposed to allowing for duplicates, which would force H_i to zero for any sample with repeated captions (actually exaggerating the effect in Figure 2. We don't allow this in order to avoid biasing our experiments to datasets such as VATEX, which explicitly remove exact duplicates. Close duplicates are not affected, as can clearly be seen by MSVD, which contains a lot of semantic redundancy. Note that this is a distribution over all references (as opposed to samples).

Another method of measuring semantic diversity is by looking at the spread of the semantics in the sample. While we use the literal variance of the within-sample pairwise distance distribution in Figure 3, we can also look at other measures of spread. Figure S1 demonstrates the difference (as a percent of the mean) between the mean of the inter-sample distances and the closest inter-sample distance. When this percentage is high, the descriptions are relatively spread out for a sample, with clusters of descriptions that are close together in semantic space. If the percentage is low, the descriptions for a sample are well-distributed (mostly equidistant) in the semantic space.

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

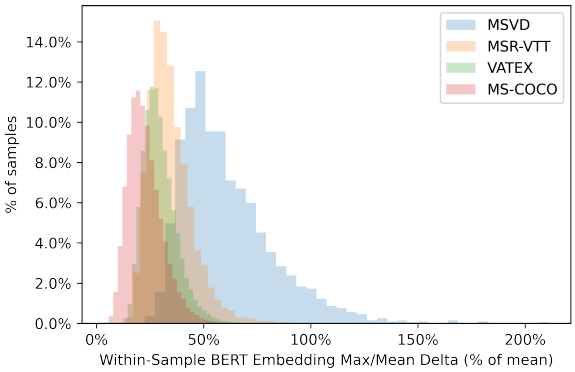


Figure S1. Plot demonstrates the difference between the closest semantic vector, and the mean of the semantic vectors. In all cases, the mean will always be further than the closest sample, however, a low delta suggests a more equal spread of references, while a high delta represents highly redundant samples.

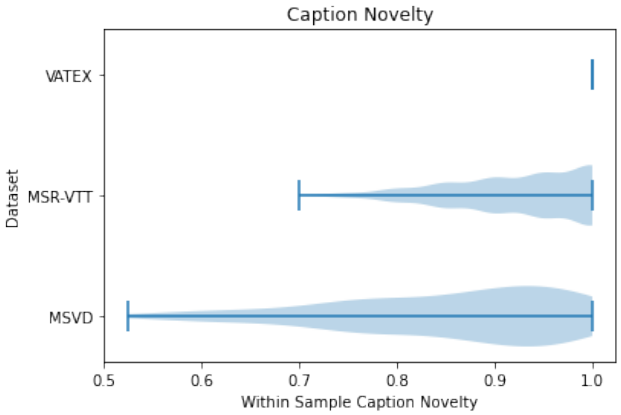


Figure S2. Violin plot demonstrating the distribution of caption novelty - i.e. how many captions in each sample are not exact matches in the text space. As we can see, while the vast majority of captions are novel in some datasets, in datasets like MSVD, there some samples which have high *exact* redundancy.

Figure S2 gives a general overview for the video description datasets of the exact-duplicate distribution of the descriptions. While most of the samples have high within-sample uniqueness, there are some samples that are highly redundant (and in the case of MSVD, have exact-redundancy of as much as $\sim 50\%$.

B.6. Dataset Diversity: Number of Ground Truths

To investigate how the number of ground truth metrics impacts the computation of the metrics, we performed several leave one out experiments as in subsection B.1 where we restricted the size of R_i for each sample to a certain number of references r by randomly sampling r elements without replacement from the original reference set. This allows us to measure the approximate performance of the methods if the number of ground truths was reduced. The results of this experiment are given in Figure S3. We can see from Figure S3 that except for CIDEr, increasing the number of ground truths increases the leave one out performance of the metrics. In fact, we can see that in most cases, the performance is nowhere near saturated, and collecting more ground truths will allow metrics to better capture the semantic variance of a scene. The standout among the group is CIDEr, in which the score does not increase as we increase the number of ground truths. This is primarily due to the IDF component of the CIDEr score, which penalizes increasing the number of tokens harshly. We can see that here, as we increase the number of ground truths, the CIDEr score *decreases*! This suggests that CIDEr is relatively robust to adding more ground truth, however cannot capture as much semantic variance as the other metrics, as the CIDEr score does not materially account for new information from the ground truth samples.

B.7. Concept-Diversity: Captions Required for BLEU Score

One of the key experiments we perform is designed to measure the minimum number of captions from the training set that are required to "solve" the test set of the dataset for a particular BLEU score. We first compute a set of all hypothesis descriptions from the training set. Then, for each sample in the test set, we compute the BLEU@4 score using that hypothesis for every sample in the test set. In the case of large datasets such as MS-COCO, which contains $591,435$ unique hypothesis captions, this can be time-consuming, even for the (relatively quick) BLEU@4 metrics. Each hypothesis thus has a score for each sample in the test dataset. Finding the minimal core-set of captions that covers this test dataset to a specified BLEU threshold is a weighted set-cover problem, which can be solved to an $O(\log N)$ approximation with a randomized rounding algorithm [Vaz01], however, we found that it was sufficient to use the greedy approximation algorithm for set cover, which selects the caption which covers the largest number of new samples at each iteration. Thus, the results in Figure 5 provide an upper bound on the possible number of captions required.

Figure 5 plots the required number of captions to achieve a BLEU@4 score of X (the value on the X-axis) on every sample. Note that this requirement is *more restrictive* than the plotted SOTA scores, which achieve a mean of X. Thus, the effect of this figure may be even more dramatic than is pictured. The reason for this discrepancy is we compute the core-set using a greedy set cover, and due to our implementation details, it is difficult to terminate the cover efficiently when a mean score is reached.

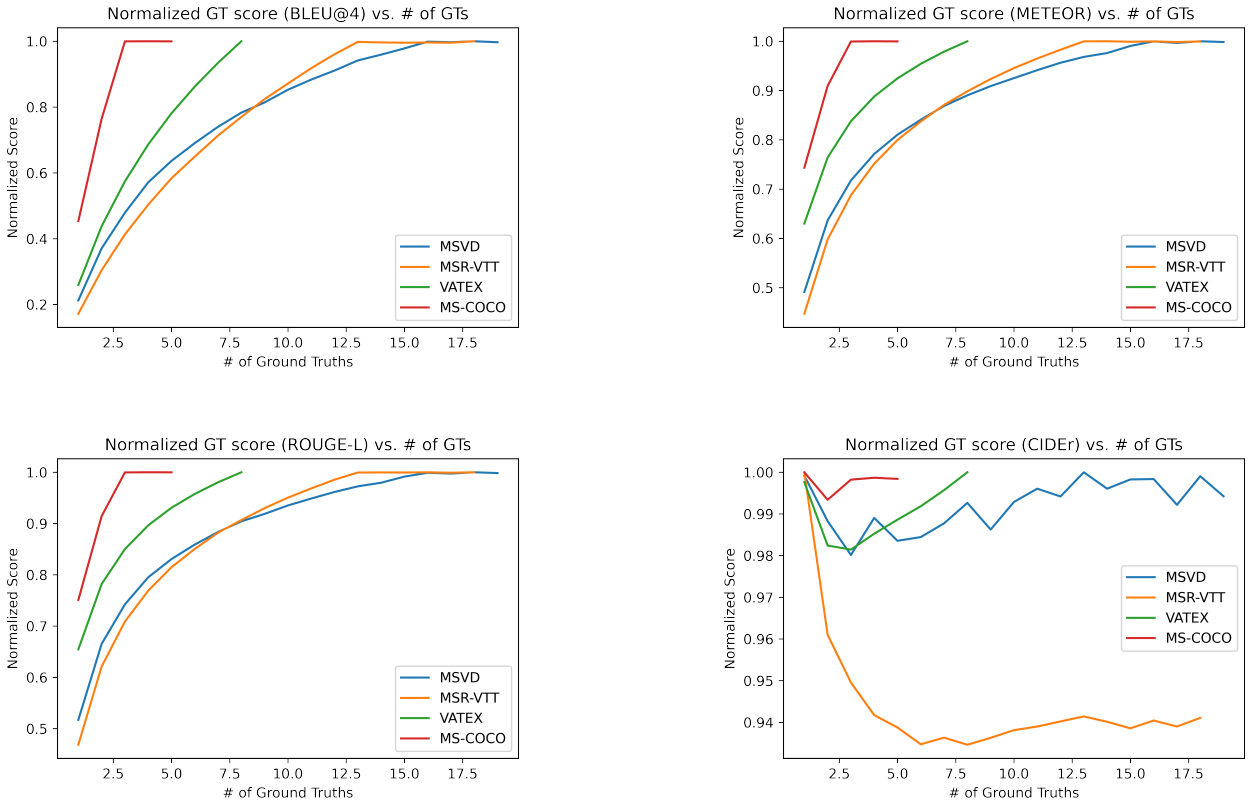


Figure S3. Performance of different metrics with respect to the number of ground truths considered in leave-one-out experiments. Raw scores are normalized to a maximum of 1, so we can compare the different datasets on the same plot.

While our work only computes the core-set for BLEU@4, we believe it would be interesting to see the numbers for other metrics, however, with current implementations, it may be intractable, as the computations require a full pairwise computation of the metrics between the hypotheses and the test-set samples. Additionally, metrics such as CIDEr which have dataset-wide effects would have to be estimated, requiring several hundred iterations of this experiment to achieve high-quality estimates of the performance. It thus remains interesting (and important) future work to explore how many captions are required to perform well on any given dataset for other metrics.

B.8. Concept-Diversity: Feature Sets

To measure the diversity of the datasets at a concept level, we look at how the ground truth captions overlap with the label sets from common feature extractors. If we find that this overlap is high, it suggests that features may have the ability to bias the model along the classification lines of the feature-extractor label set (since a lot of the time, the information extracted by the features is useful primarily for segmenting data along feature class boundaries).

B.8.1 Computing Label-Set overlap

We discuss two methods for computing label-set overlap in the main paper: exact match and fuzzy matching. Exact match is implemented as a string substring: i.e. does the label string appear as a direct substring of the caption. This method provides a lower bound on the true conceptual overlap, as it does not account for misspellings (which are surprisingly common in datasets such as MSR-VTT, and others collected using AMT without additional review steps), and other close matches. While this is a lower bound, it has the benefit of not introducing false-positive matches (as any match is guaranteed to be label overlap). We also discuss the use of fuzzy matching, which we implement using the fuzzywuzzy[6] library for approximate string matching with a threshold of 90. This library uses Levenshtein distance to compute approximate matching, however introduces false-positives which makes it difficult to analyze the overlap. In all cases, the numbers in Table 3 represent the percentage of samples that have at least one reference description that has exact overlap with a label from the dataset.

[6] https://github.com/seatgeek/thefuzz

We explore overlap on four common datasets for feature extraction:

ImageNet-1K: [DDS+09] is a popular image classification dataset consisting of 1K labels for object classification ranging across a very wide variety of objects. We can see this from the overlap scores in Table 3, which are relatively high on almost all of the datasets. MSR-VTT is relatively low, suggesting that it is one of the most open-domain datasets among the datasets we explore.

Kinetics-600: [CNBH+18] is a popular dataset for action recognition, which contains 600 activities. We can see that the video datasets have a much higher overlap with kinetics, but even though MS-COCO is an image dataset only, there is still some overlap, suggesting that captions of static data still contain human inferences about motion and activity.

MS-COCO: [LMB+14] is a dataset for object detection (and also for visual description) containing object-detection labels over 80 object classes from everyday life. Even though COCO has a relatively restricted object set, we can see that it consists of a set of very popular objects, as the overlap is more than 50% for all captions. Additionally, it's interesting that the object labels for MS-COCO don't always appear in the captions themselves (as the self-overlap is only 92%).

Places-365: [ZLK+17] is a dataset for scene recognition, consisting of 365 labels of scenes or settings for an image. We find empirically that the overlap for places is likely low, not due to a lack of descriptions of setting, but rather a lack of wide coverage of the variance of settings in Places.

B.8.2 Feature-Set Core-Sets and BLEU@4 performance

To directly measure how transferable descriptions are along feature-extractor label axes, we explore the leave-one-out performance of captions sharing the same feature label, but from different samples in the dataset. The results of this experiment using BLEU@4 scores are given in Table 4. In order to compute the leave-one-out performance, we begin by computing a set of reference captions R_c for each label in each feature-extractor label set, drawing from the training dataset. These concept-level reference sets consist of all captions containing that label as an exact sub-string. Then, for each sample S_i with references R_i, we compute the set of all concepts overlapping that sample's references C_i. We then compute the hypothesis set for sample S_i as

$$H_i = \left[\bigcup_{c \in C_i} R_c \right] / \{R_i\}$$

Next, for each hypothesis in H_i, we compute the BLEU@4 score for that hypothesis using ground truths R_i. The table Table 4 reports the mean over all samples of the maximum across H_i for each sample in the test set. The results of this metric are clear - when you use the best caption from another sample along feature boundaries, then these captions are relatively transferable (and almost always outperform samples from even the same sample).

B.9. Tools & Hardware

The experiments in this paper are computed using the metric implementations provided by the MSCOCO evaluation toolkit in order to compute numeric metric values that are comparable with state of the art methods. In the experiments in the paper, we use the Stanford PTB[7] tokenizer provided as part of the toolkit for tokenization and standardization. Unfortunately, because the MSCOCO toolkit does not explicitly specify a tokenization scheme and most works in video description do not subscribe to a standard tokenization tool, we are unable to be certain that the metric is consistent between our work, and the work presented in the state of the art papers.

The experiments are run in parallel on a machine with 96 AMD EPYC 7B12 cores and 378 GB of RAM running on Google Cloud Platform. Notably, the caption concept-overlap experiments require a very large amount of compute, with this machine requiring almost 10 hours to compute the BLEU score for the core-set concept overlap. We found scores such as METEOR [AL08] and SPICE [AFJG16] to be computationally prohibitive (requiring several months of sustained compute) for some of these experiments, thus, we do not include those scores in this work. We also do not report several modern metrics for this reason - as a major downside to many of the automated metrics that have recently been developed is their forward inference speed (up to 1000s of times slower than the computation of the BLEU score). A key area of future work is improving the computational performance of metrics, as this will allow such metrics to not only be used for more detailed analysis but will allow such metrics to be optimized directly using techniques such as self-critical sequence training [RMM+17].

[7] https://nlp.stanford.edu/software/tokenizer.html

C. Additional Qualitative Examples

Additional qualitative examples are selected at random from the datasets using a random number generator over the length of each dataset. Some randomly selected samples are omitted due to explicit content in the visual data or descriptions (which is an additional cause for concern, but out of scope of the current research).

- a person mixing potatoe salad in a bowl (Dist: 0.10) (BLEU@4: 0.0000)
- someone mixes a potato salad (Dist: 0.13) (BLEU@4: 0.0001)
- this is someone stirring a bowl of potato salad (Dist: 0.13) (BLEU@4: 0.3156)
- a person in a kitchen is mixing a bowl of potato salad with a spatula in a white bowl on a cutting board (Dist: 0.14) (BLEU@4: 0.1613)
- the bowl of potato salad is stirred (Dist: 0.18) (BLEU@4: 0.3267)
- someone is mashing potato salad together (Dist: 0.19) (BLEU@4: 0.0000)
- someone is making food (Dist: 0.22) (BLEU@4: 0.0000)
- a women stirring and mixing some egg salad (Dist: 0.23) (BLEU@4: 0.0000)
- a person is cooking (Dist: 0.24) (BLEU@4: 0.0000)
- a bowl is being stirred (Dist: 0.26) (BLEU@4: 0.0000)
- a person stirring food (Dist: 0.26) (BLEU@4: 0.0000)
- someone mixing pieces of boiled eggs and yogurt in a bowl (Dist: 0.27) (BLEU@4: 0.0000)
- mixing up food in a bowl in the kitchen (Dist: 0.30) (BLEU@4: 0.0000)
- making some potato salad (Dist: 0.30) (BLEU@4: 0.0000)
- a great way to make potato salad (Dist: 0.36) (BLEU@4: 0.0000)
- how to prepare a potato salad (Dist: 0.36) (BLEU@4: 0.0001)
- preparation of some egg dish (Dist: 0.44) (BLEU@4: 0.0000)

msrvtt+train+video481
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.10,0.24,0.44]
Mean Leave One Out BLEU@4 score: 0.0473

Figure S4. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).

- baby muppets are talking to a judge kermit (Dist: 0.18) (BLEU@4: 0.3457)
- cartoon muppets are having a court proceeding (Dist: 0.18) (BLEU@4: 0.0000)
- kermit is a judge and is talking to 2 other muppet babies (Dist: 0.20) (BLEU@4: 0.1499)
- some muppet characters are interacting with each other (Dist: 0.22) (BLEU@4: 0.0000)
- a scene from the muppet babies where baby kermit is a judge in a court room is playing in a movie theater (Dist: 0.24) (BLEU@4: 0.1336)
- a cartoon clip of baby muffets in a court room (Dist: 0.24) (BLEU@4: 0.0876)
- muppets are talking to each other (Dist: 0.25) (BLEU@4: 0.3641)
- the scene is from baby muppets the green frog character kermit is a judge and two other characters are standing before him (Dist: 0.26) (BLEU@4: 0.1759)
- muppet babies playing in a theater (Dist: 0.29) (BLEU@4: 0.0000)
- kermit judges gonzo and another muppet baby (Dist: 0.31) (BLEU@4: 0.0000)
- some cartoon characters speak to a judge (Dist: 0.32) (BLEU@4: 0.0000)
- cartoon babies standing in front of a judge (Dist: 0.34) (BLEU@4: 0.0000)
- the muppets are being shown in a theater (Dist: 0.35) (BLEU@4: 0.0000)
- the muppet babies are on a theater screen (Dist: 0.36) (BLEU@4: 0.0000)
- cartoon characters are talking in a tv show (Dist: 0.38) (BLEU@4: 0.0000)
- a cartoon frog is talking to other cartoon characters (Dist: 0.41) (BLEU@4: 0.0000)
- a clip taken from the cartoon the muphets (Dist: 0.42) (BLEU@4: 0.0000)
- cartoons are being displayed (Dist: 0.43) (BLEU@4: 0.0000)

msrvtt+train+video1902
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.18,0.30,0.43]
Mean Leave One Out BLEU@4 score: 0.0698

Figure S5. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).

- a man is playing guitar and singing (Dist: 0.07) (BLEU@4: 0.4468)
- a man playing the guitar and singing (Dist: 0.07) (BLEU@4: 0.5314)
- a man is playing the guitar and singing outside (Dist: 0.09) (BLEU@4: 0.5874)
- a man is singing while playing his guitar (Dist: 0.09) (BLEU@4: 0.7071)
- a man is singing while playing the guitar (Dist: 0.09) (BLEU@4: 0.7071)
- a man playing a guitar in various locations singing a song (Dist: 0.13) (BLEU@4: 0.0000)
- a man plays a guitar and sings a song outside (Dist: 0.13) (BLEU@4: 0.4111)
- a man sings while playing the acoustic guitar (Dist: 0.16) (BLEU@4: 0.4214)
- a man is singing on a doorstep and playing his guitar (Dist: 0.18) (BLEU@4: 0.3393)
- a man singing a song (Dist: 0.19) (BLEU@4: 0.0000)
- man is playing a guitar (Dist: 0.22) (BLEU@4: 0.0001)
- a man sings while playing a guitar throughout a town (Dist: 0.22) (BLEU@4: 0.4111)
- a man is playing the guitar (Dist: 0.23) (BLEU@4: 0.6065)
- a man plays a guitar outside (Dist: 0.25) (BLEU@4: 0.4083)
- this is a video of a man playing guitar and singing (Dist: 0.25) (BLEU@4: 0.2900)
- a man is singing a melodious song (Dist: 0.29) (BLEU@4: 0.3564)
- guy performing music on building (Dist: 0.40) (BLEU@4: 0.0000)

msrvtt+train+video5073
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.07, 0.18, 0.40]
Mean Leave One Out BLEU@4 score: 0.3661

Figure S6. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).

- a football coach talking to his players on the sideline (Dist: 0.16) (BLEU@4: 0.0001)
- a person talking to football players (Dist: 0.18) (BLEU@4: 0.0000)
- a coach is talking to his team of football players (Dist: 0.19) (BLEU@4: 0.0000)
- a football coach talks on the sideline (Dist: 0.23) (BLEU@4: 0.0001)
- a football coach watches the players (Dist: 0.26) (BLEU@4: 0.0001)
- football players are on the field (Dist: 0.27) (BLEU@4: 0.0001)
- football players on the field (Dist: 0.28) (BLEU@4: 0.0001)
- a coach is speaking with his team (Dist: 0.29) (BLEU@4: 0.0000)
- a man on the sidelines at a football game is talking continuously (Dist: 0.29) (BLEU@4: 0.0000)
- a football coach is shown losing his patience with his players (Dist: 0.36) (BLEU@4: 0.0000)
- a football player is on the sidelines (Dist: 0.37) (BLEU@4: 0.0000)
- people are playing sports (Dist: 0.38) (BLEU@4: 0.0000)
- footage of a football game (Dist: 0.39) (BLEU@4: 0.0000)
- there are men at a sports match (Dist: 0.44) (BLEU@4: 0.0000)
- players playing soccer game (Dist: 0.49) (BLEU@4: 0.0000)
- a man is chewing and watching football players (Dist: 0.50) (BLEU@4: 0.0000)
- a trainer guide a player (Dist: 0.52) (BLEU@4: 0.0000)
- a man is talking (Dist: 0.54) (BLEU@4: 0.0000)
- some guys are geting ready to play rugby in a ground (Dist: 0.57) (BLEU@4: 0.0000)
- a man talking or yelling with his lid backward (Dist: 0.71) (BLEU@4: 0.0000)

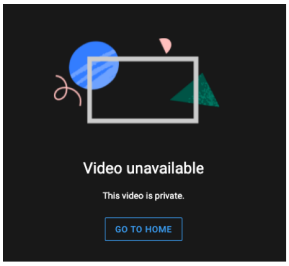
msrvtt+train+video63
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.16, 0.37, 0.71]
Mean Leave One Out BLEU@4 score: 0.0000

Figure S7. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).

- While outside, a man is using a chainsaw to cut through a wood log. (Dist: 0.07) (BLEU@4: 0.2804)
- A man cuts into a large log with a chainsaw. (Dist: 0.10) (BLEU@4: 0.0000)
- A man using a chain saw to cut a board out of a log. (Dist: 0.11) (BLEU@4: 0.4547)
- A man uses a chainsaw to cut a log into smaller segments. (Dist: 0.12) (BLEU@4: 0.3787)
- A man holds a chainsaw and carves a log with the saw. (Dist: 0.14) (BLEU@4: 0.0000)
- A man is outside using a chainsaw to shape a large tree trunk. (Dist: 0.15) (BLEU@4: 0.4002)
- A man is using a chain saw to cut a large tree trunk. (Dist: 0.17) (BLEU@4: 0.4575)
- A man uses a chainsaw to carve a fallen tree in the woods. (Dist: 0.19) (BLEU@4: 0.3760)
- A man is using an electric saw to cut a block of wood. (Dist: 0.22) (BLEU@4: 0.3662)
- A man is taking a video showing us how to cut a good lumber. (Dist: 0.37) (BLEU@4: 0.0000)

vatex+train+c7R_RDH6s_A_000122_000132
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.07, 0.16, 0.37]
Mean Leave One Out BLEU@4 score: 0.2714

Figure S8. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).



- A man demonstrates a lunge with a barbell with weights on the end. (Dist: 0.14) (BLEU@4: 0.0000)
- A man is using barbells to teach how to do a lunge with them. (Dist: 0.16) (BLEU@4: 0.0000)
- A man is in the gym performing a barbell single side lunge. (Dist: 0.18) (BLEU@4: 0.1623)
- a man in a gym is demonstrating how to properly do an exercise (Dist: 0.18) (BLEU@4: 0.0000)
- A person demonstrates the proper technique for a barbell exercise while text explains the technique. (Dist: 0.19) (BLEU@4: 0.0000)
- A man place a set of barbells on his shoulder and steps forward. (Dist: 0.21) (BLEU@4: 0.0000)
- The guy is in the gym lifting some barbel on the back of his shoulders. (Dist: 0.25) (BLEU@4: 0.1956)
- A man is holding a barbell on his shoulders as he steps forward. (Dist: 0.26) (BLEU@4: 0.0000)
- A person doing exercises and telling how to do them the right way. (Dist: 0.29) (BLEU@4: 0.0000)
- A man is standing in the gym lifting weights from the ground to his shoulders (Dist: 0.36) (BLEU@4: 0.1967)

vatex+train+j53UDMQ8mxo_000000_000010
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.14, 0.22, 0.36]
Mean Leave One Out BLEU@4 score: 0.0555

Figure S9. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5). The visual content of this video is missing (as the video has become private since the collection of the dataset), however we include the video as it is one of the randomly sampled instances.

- A man is pouring tea from a tea kettle into two mugs while keeping his other hand in his pocket. (Dist: 0.17) (BLEU@4: 0.1573)
- Man places hand in pocket before pouring water from kettle into two mugs. (Dist: 0.18) (BLEU@4: 0.1492)
- A man in a kitchen puts his hand in his pocket while pouring a beverage and explains why. (Dist: 0.23) (BLEU@4: 0.1579)
- The man is pouring something from the kettle into the two coffee mugs on the counter. (Dist: 0.26) (BLEU@4: 0.0000)
- He places his hand in his pant pocket then pours coffee into the cups. (Dist: 0.29) (BLEU@4: 0.1395)
- A man wearing a gray tank top is pouring tea into a cup. (Dist: 0.29) (BLEU@4: 0.1880)
- A man wearing a tee shirt pours coffee in two cups. (Dist: 0.29) (BLEU@4: 0.1898)
- A man in his kitchen has a kettle pouring himself hot water (Dist: 0.30) (BLEU@4: 0.0000)
- There is a man making tea and he is explaining where to put your arms so you don't burn them while pouring the water. (Dist: 0.31) (BLEU@4: 0.0000)
- A guy is talking about a cooking video that he saw a couple weeks ago. (Dist: 0.53) (BLEU@4: 0.0000)

vatex+train+DbJWd2K2Hw0_000229_000239
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.17, 0.29, 0.53]
Mean Leave One Out BLEU@4 score: 0.0982

Figure S10. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).

- A person peels and slices an apple using a knife onto a plate. (Dist: 0.07) (BLEU@4: 0.3161)
- A person peels an apple and chops it to tiny pieces. (Dist: 0.09) (BLEU@4: 0.0000)
- A person is cutting up an apple with a knife onto a green plate. (Dist: 0.11) (BLEU@4: 0.2460)
- A person is peeling apples and then cutting out the cores. (Dist: 0.11) (BLEU@4: 0.0000)
- A person is cutting a peeled and cored red apple into slices. (Dist: 0.12) (BLEU@4: 0.1907)
- A person is taking the skin off of some apples using a small knife. (Dist: 0.12) (BLEU@4: 0.0000)
- A person peeling an apple using a knife while sitting at a table (Dist: 0.13) (BLEU@4: 0.3184)
- A man uses a knife to peel of the side of apple. (Dist: 0.13) (BLEU@4: 0.0000)
- A person inside sitting at a table peels an apple on a green plate. (Dist: 0.18) (BLEU@4: 0.2289)
- A person cuts an onion into pieces using a knife. (Dist: 0.30) (BLEU@4: 0.0000)

vatex+train+W3orrcZAb2w_000200_000210
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.07, 0.14, 0.30]
Mean Leave One Out BLEU@4 score: 0.1300

Figure S11. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).

- On a beach three men are doing cartwheels after another (Dist: 0.07) (BLEU@4: 0.1949)
- Three young men are doing cartwheels across the beach, towards the surf. (Dist: 0.09) (BLEU@4: 0.1967)
- Three males are doing cartwheels down the beach towards the ocean. (Dist: 0.09) (BLEU@4: 0.4125)
- Three guys do cartwheels on a beach toward the water until the stop and fall. (Dist: 0.11) (BLEU@4: 0.0000)
- Three men wearing shorts are doing cartwheels on the sand on the beach. (Dist: 0.12) (BLEU@4: 0.1967)
- Group of men doing numerous cartwheels down the beach towards the water. (Dist: 0.13) (BLEU@4: 0.4090)
- Three kids do cartwheels down the beach together. (Dist: 0.17) (BLEU@4: 0.2367)
- Three men doing cart wheels across a beach, stop then laughing. (Dist: 0.20) (BLEU@4: 0.0000)
- A group of people are successively doing cart wheels on the beach. (Dist: 0.23) (BLEU@4: 0.1982)
- Three boys are doing cart wheeling on the beach all by themselves. (Dist: 0.33) (BLEU@4: 0.0000)

vatex+train+qyHurkZFOp0_000002_000012
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.07, 0.15, 0.33]
Mean Leave One Out BLEU@4 score: 0.1845

Figure S12. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).

- A man is using specialized ice climbing gear and metal cleats on his shoes to climb an icy cliff. (Dist: 0.10) (BLEU@4: 0.1532)
- A man using proper equipment and shoe's to ice climb. (Dist: 0.10) (BLEU@4: 0.0000)
- A man is climbing up the side of an icy mountains using climbing shoes and tools. (Dist: 0.11) (BLEU@4: 0.4925)
- A man is wearing shoes with spikes and using a ice pick to climb a wall of ice. (Dist: 0.12) (BLEU@4: 0.3709)
- A man is attached to a harness as he is climbing a wall of snow and ice. (Dist: 0.13) (BLEU@4: 0.0000)
- A man is climbing up the side of an ice wall. (Dist: 0.14) (BLEU@4: 0.4808)
- A man wearing a harness and spiked shoes climbs up a snow covered wall. (Dist: 0.14) (BLEU@4: 0.0000)
- A man uses a pick and a harness to scale an icy cliff. (Dist: 0.18) (BLEU@4: 0.1480)
- a person tries their best to climb up a snowy mountain (Dist: 0.30) (BLEU@4: 0.0000)
- A man cheers as another man is using crampons to climb a wall of ice. (Dist: 0.30) (BLEU@4: 0.3714)

vatex+train+keBAoE5tC44_000011_000021
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.10, 0.16, 0.30]
Mean Leave One Out BLEU@4 score: 0.2017

Figure S13. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).

msvd+train+_WRC7HXBJpU_360_370
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.09, 0.30, 0.74]
Mean Leave One Out BLEU@4 score: 0.2693

- someone is melting butter in a pan (Dist: 0.09) (BLEU@4: 0.7014)
- butter is being melted in a pan (Dist: 0.09) (BLEU@4: 0.4889)
- the butter is melting in the pan (Dist: 0.10) (BLEU@4: 0.5329)
- a man melts a piece of butter in a pan (Dist: 0.10) (BLEU@4: 0.5170)
- a person melts butter on a pan (Dist: 0.10) (BLEU@4: 0.0000)
- butter is melted in a pan (Dist: 0.11) (BLEU@4: 0.5115)
- a man is heating butter in a pan (Dist: 0.12) (BLEU@4: 0.5000)
- a man is melting butter in a pan (Dist: 0.12) (BLEU@4: 0.8409)
- someone is melting a piece of butter in a skillet (Dist: 0.13) (BLEU@4: 0.5170)
- butter is melting in a pan (Dist: 0.14) (BLEU@4: 0.6732)
- a man is stirring melting butter in a pan (Dist: 0.14) (BLEU@4: 0.5969)
- butter melts in a pan (Dist: 0.14) (BLEU@4: 0.0001)
- butter is melting in a frying pan (Dist: 0.15) (BLEU@4: 0.6435)
- melting the butter on the frying pan (Dist: 0.15) (BLEU@4: 0.0001)
- some butter is melting in a hot skillet (Dist: 0.16) (BLEU@4: 0.5170)
- butter melted in the pan (Dist: 0.16) (BLEU@4: 0.8187)
- the butter melted in the pan (Dist: 0.16) (BLEU@4: 0.7598)
- a man is melting butter in a skillet (Dist: 0.17) (BLEU@4: 0.8409)
- butter is melting on a skillet (Dist: 0.17) (BLEU@4: 0.0001)
- a cook spreads melted butter around a pan (Dist: 0.18) (BLEU@4: 0.0000)
- a man is spreading some butter in a pan (Dist: 0.19) (BLEU@4: 0.4317)
- the man mmelted butter in the frying pan (Dist: 0.19) (BLEU@4: 0.0001)
- a man demonstrates how to prepare a pan with butter (Dist: 0.21) (BLEU@4: 0.0000)
- the man is meliting butter in the pan (Dist: 0.22) (BLEU@4: 0.0001)
- a man is pressing a butter with mixing utensil in a black pan (Dist: 0.26) (BLEU@4: 0.0000)
- somebody is cooking (Dist: 0.36) (BLEU@4: 0.0000)
- a man is cooking (Dist: 0.39) (BLEU@4: 0.0001)
- a cheese cub is melting in the pan (Dist: 0.39) (BLEU@4: 0.5170)
- he is cooking on the plate (Dist: 0.48) (BLEU@4: 0.0000)
- a man is melting the cheese (Dist: 0.50) (BLEU@4: 0.3641)
- a man coking pork chops (Dist: 0.52) (BLEU@4: 0.0000)
- a man makes food for him self (Dist: 0.54) (BLEU@4: 0.0000)
- cheese in the oil (Dist: 0.55) (BLEU@4: 0.0000)
- a man cooking his kichen (Dist: 0.58) (BLEU@4: 0.0000)
- a person is making pork chops (Dist: 0.59) (BLEU@4: 0.0000)
- cooking the pork chops (Dist: 0.60) (BLEU@4: 0.0000)
- anyone is frying (Dist: 0.61) (BLEU@4: 0.0000)
- a man in a blue shirt and yellow cap preparing pork chops (Dist: 0.67) (BLEU@4: 0.0000)
- clear cooking details for pork chops (Dist: 0.70) (BLEU@4: 0.0001)
- racipe for pork chops (Dist: 0.74) (BLEU@4: 0.0001)

Figure S14. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).

msvd+test+pzq5fPfsPZg_51_57
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.06, 0.26, 0.75]
Mean Leave One Out BLEU@4 score: 0.4139

- a woman is hugging a little boy (Dist: 0.06) (BLEU@4: 0.6435)
- a woman is hugging a boy (Dist: 0.08) (BLEU@4: 0.7598)
- a woman hugs a little boy (Dist: 0.08) (BLEU@4: 0.6732)
- a woman is hugging a child (Dist: 0.10) (BLEU@4: 0.7598)
- a young woman hugs a little boy (Dist: 0.11) (BLEU@4: 0.6435)
- the woman is hugging the boy (Dist: 0.11) (BLEU@4: 0.0001)
- a woman is hugging her child (Dist: 0.12) (BLEU@4: 0.7598)
- the lady hugged her little boy close to her (Dist: 0.16) (BLEU@4: 0.0000)
- a woman is hugging her son (Dist: 0.18) (BLEU@4: 0.7598)
- a woman is hugging a toddler (Dist: 0.21) (BLEU@4: 0.7598)
- a girl hugs a little boy (Dist: 0.23) (BLEU@4: 0.5373)
- a girls is hugging a little boy (Dist: 0.23) (BLEU@4: 0.6435)
- a woman hugs a small boy then sets him down on the floor (Dist: 0.25) (BLEU@4: 0.2045)
- the mother is hugging the baby (Dist: 0.28) (BLEU@4: 0.5373)
- a teenage girl is hugging a little boy (Dist: 0.29) (BLEU@4: 0.5411)
- two people are hugging (Dist: 0.33) (BLEU@4: 0.0000)
- the girl hugged the little boy tightly (Dist: 0.34) (BLEU@4: 0.0000)
- the girl is hugging the baby (Dist: 0.44) (BLEU@4: 0.5373)
- whitney is hugging her son in her house (Dist: 0.44) (BLEU@4: 0.3457)
- a lady is playing with child (Dist: 0.47) (BLEU@4: 0.0000)
- one lady and child are standing together (Dist: 0.50) (BLEU@4: 0.0000)
- the baby is doing the something (Dist: 0.75) (BLEU@4: 0.0000)

Figure S15. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).

- Young boy leaping into air with tennis racket with dog below him. (Dist: 0.05) (BLEU@4: 0.0000)
- A boy jumping into the air with a tennis racket. (Dist: 0.10) (BLEU@4: 0.3352)
- A boy with a tennis racket jumping in the air. (Dist: 0.10) (BLEU@4: 0.3352)
- A young man with a tennis racket jumping high into the air near a dog on a street side. (Dist: 0.19) (BLEU@4: 0.2195)
- A kid is in the air with a dog looking on (Dist: 0.32) (BLEU@4: 0.2481)

mscoco+train+174229
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.05, 0.15, 0.32]
Mean Leave One Out BLEU@4 score: 0.2276

Figure S16. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).

- Grey cat laying on a green floor near a sandel. (Dist: 0.10) (BLEU@4: 0.0000)
- A small gray and white cat laying on the floor (Dist: 0.11) (BLEU@4: 0.0000)
- White and grey kitten lying on a messy green carpet. (Dist: 0.15) (BLEU@4: 0.0000)
- A young cat on a mat with a flip flop shoe. (Dist: 0.20) (BLEU@4: 0.0000)
- The grey and white cat is beside a rubber show. (Dist: 0.22) (BLEU@4: 0.0000)

mscoco+validate+186296
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.10, 0.16, 0.22]
Mean Leave One Out BLEU@4 score: 0.0000

Figure S17. Qualitative example of metrics presented in the paper. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).

References [Supplementary]

[AFJG16] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016. S6

[AL08] Abhaya Agarwal and Alon Lavie. 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, 2008. S6

[CD11] David Chen and William B Dolan. 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, 2011. S1

[CNBH+18] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. S6

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. S6

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. S1, S6

[RMM+17] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. S6

[STQ+20] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*, 2020. S3

[Vaz01] Vijay V Vazirani. *Approximation algorithms*, volume 1. Springer, 2001. S4

[WWC+19] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. S1

[XMYR16] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. S1

[ZLK+17] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. S6