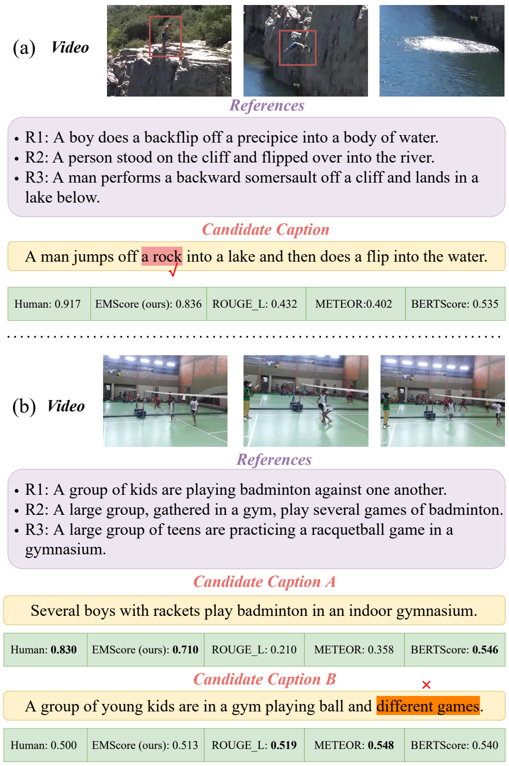
EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embedding Matching

Yaya Shi[1], Xu Yang[2], Haiyang Xu[3], Chunfeng Yuan[4]*, Bing Li[4], Weiming Hu[4,5,6], Zheng-Jun Zha[1]
[1]University of Science and Technology of China [2]Southeast University [3]Alibaba Group
[4]NLPR, Institute of Automation, Chinese Academy of Sciences
[5]School of Artificial Intelligence, University of Chinese Academy of Sciences
[6]CAS Center for Excellence in Brain Science and Intelligence Technology
shiyaya@mail.ustc.edu.cn 101013120@seu.edu.cn shuofeng.xhy@alibaba-inc.com
{cfyuan, bli, wmhu}@nlpr.ia.ac.cn zhazj@ustc.edu.cn

Abstract

Current metrics for video captioning are mostly based on the text-level comparison between reference and candidate captions. However, they have some insuperable drawbacks, e.g., *they cannot handle videos without references, and they may result in biased evaluation due to the one-to-many nature of video-to-text and the neglect of visual relevance. From the human evaluator's viewpoint, a high-quality caption should be consistent with the provided video, but not necessarily be similar to the reference in literal or semantics. Inspired by human evaluation, we propose* ***EMScore*** *(Embedding Matching-based score), a novel reference-free metric for video captioning, which directly measures similarity between video and candidate captions. Benefiting from the recent development of large-scale pre-training models, we exploit a well pre-trained vision-language model to extract visual and linguistic embeddings for computing EMScore. Specifically, EMScore combines matching scores of both coarse-grained (video and caption) and fine-grained (frames and words) levels, which takes the overall understanding and detailed characteristics of the video into account. Furthermore, considering the potential information gain, EMScore can be flexibly extended to the conditions where human-labeled references are available. Last but not least, we collect VATEX-EVAL and ActivityNet-FOIl datasets to systematically evaluate the existing metrics. VATEX-EVAL experiments demonstrate that EMScore has higher human correlation and lower reference dependency. ActivityNet-FOIL experiment verifies that EMScore can effectively identify "hallucinating" captions. Code and datasets are available at https://github.com/shiyaya/emscore.*



- R1: A boy does a backflip off a precipice into a body of water.
- R2: A person stood on the cliff and flipped over into the river.
- R3: A man performs a backward somersault off a cliff and lands in a lake below.

A man jumps off a rock ✓ into a lake and then does a flip into the water.

Human: 0.917	EMScore (ours): 0.836	ROUGE_L: 0.432	METEOR:0.402	BERTScore: 0.535

- R1: A group of kids are playing badminton against one another.
- R2: A large group, gathered in a gym, play several games of badminton.
- R3: A large group of teens are practicing a racquetball game in a gymnasium.

Several boys with rackets play badminton in an indoor gymnasium.

Human: **0.830**	EMScore (ours): **0.710**	ROUGE_L: 0.210	METEOR: 0.358	BERTScore: **0.546**

A group of young kids are in a gym playing ball and different games ×.

Human: 0.500	EMScore (ours): 0.513	ROUGE_L: **0.519**	METEOR: **0.548**	BERTScore: 0.540

Figure 1. Two examples of caption evaluation. All the metric scores are scaled to [0, 1], including human scores. For example (a), reference-based metrics over-penalize for this correct candidate caption due to "a rock" is not contained in the references. Our reference-free metric EMScore gives a reasonable high score with the help of using video as ground truth. For example (b), some reference-based metrics (e.g., ROUGE_L and METEOR) under-penalize the hallucination (e.g., "different games") which is not related to the video, and give an unreasonable higher score for "hallucinating" caption B than correct caption A.

*Corresponding author

1. Introduction

Video Captioning [4] aims to generate a text describing the visual content of a given video. Driven by the neural encoder-decoder paradigm, research in video captioning has made significant progress [29, 35]. To make further advances in video captioning, it is essential to accurately evaluate generated captions. The most ideal metric is human evaluation while carrying human judgments is time-consuming and labor-intensive. Thus, various automatic metrics are applied for video caption evaluation.

However, most of the widely applied video caption metrics like BLEU [19], ROUGE [12], CIDEr [28], and BERTScore [34] come from the other tasks, such as machine translation, text summarization and image captioning, which may neglect the special characteristic of video captioning and then limit the development of video captioning. Furthermore, these automatic metrics require human-labeled references — and thus they are called reference-based metrics — and such requirements cause three intrinsic drawbacks: (1) They can not be used when provided videos have no human-labeled references, which is not uncommon in this age that millions of reference-free videos are produced online every day. (2) They may over-penalize the correct captions since references hardly describe all details of videos due to the one-to-many nature [32] of captioning task, especially when the number of references is limited. Fig.1 (a) shows one such example where a candidate caption correctly describes the "a rock" while reference-based metrics punish this word since references do not contain it. (3) As pointed by [23], these reference-based metrics may under-penalize the captions with "hallucinating" descriptions since these metrics only measure similarity to references, and the visual relevance cannot be fully captured. For example, as shown in Fig.1 (b), due to the word "games" appearing in the references, some reference-metrics return higher scores for caption B than caption A, even though "different games" is a "hallucinating" phrase which is not related to the video.

These drawbacks inspire us to develop a reference-free metric. From the human evaluator's viewpoint, if a caption is *consistent* with the source video, *i.e.*, the visual contents in the video are comprehensively and accurately described by the caption, this caption is a high-quality one, and not necessarily be similar to the reference in literal or semantics. A promising evaluation metric should imitate the human evaluation process, and introduce video content into the evaluation. Nowadays, due to the boom of the large-scale vision-language pre-training models [11, 17, 21], the gaps between the visual and linguistic embeddings have been further narrowed, enabling us to judge whether a caption is consistent with a video.

Motivated by these research progresses, we propose a *reference-free* metric **EMScore** (Embedding Matching-based score) for evaluating video captions, which exploits a pre-trained large-scale vision-language model to extract visual and linguistic embeddings. Specifically, to obtain a comprehensive comparison between the video and caption, EMScore averages the matching scores of both coarse-grained (video and caption) and fine-grained (frames and words) levels. For the coarse-grained one, we compute the similarity between the global embeddings of the video and the candidate caption, which take the overall understanding of the video into account and evaluate candidates from a global perspective. For the fine-grained embedding matching, we compute the sum of cosine similarities between the frame and word embeddings, which takes the detailed characteristic of the video (visual elements change over time) into account. Also, it provides more interpretability for EMScore. Furthermore, considering the potential information gain, such as syntactic structure in references, and doing embedding matching in the same language domain is easier than cross-modal domains, we extend EMScore to the conditions where human-labeled references are available and name the extended metric EMScore_ref.

Currently, there is no available video caption quality dataset that can be used to evaluate metrics. To facilitate the development of video captioning evaluation metrics, we are the first to collect a video caption quality dataset VATEX-EVAL which contains 54,000 human ratings for video-caption pairs. Experiments on VATEX-EVAL show the following advantages of our EMScore by introducing the video in evaluating. First, EMScore has a higher human correlation compared with some popular automatic metrics like BLEU, ROUGE, or CIDEr. Second, EMScore has low reference dependency, *e.g.*, EMScore's 0-reference Kendall's correlation with humans is similar to BLEU_1's 4-reference correlation or EMScore_ref's 1-reference is similar to CIDEr's 9-reference correlations. Therefore, EMScore can significantly reduce the cost of manually annotating references. Third, EMScore is more robust to quality drift that it achieves higher correlations compared with the other automatic metrics when evaluating captions of different qualities. Furthermore, we collect another dataset ActivityNet-FOIL which contains "hallucinating" captions to verify the sensitivity of EMScore. Experiment results show that EMScore is more effective to identify "hallucinating" captions than the other metrics.

Our contributions are summarized as follows:

- We propose a reference-free video captioning metric EMScore that directly measures consistency with video contents in both coarse-grained and fine-grained levels, and extend it to reference-available condition.
- We collect two datasets VATEX-EVAL and ActivityNet-FOIL for researchers to study the metrics' correlation with human judgments and sensitivity in the "hallucinating" case, respectively.

- Exhaustive experimental results verify that EMScore has a higher human correlation and is able to effectively identify the “hallucinating” captions.

2. Related work

2.1. Caption Evaluation

Rule-Based Evaluation The most widely used caption metrics are based on n-gram matching — BLEU [19], ROUGE [12] and METEOR [3]. Especially, CIDEr [28] weights each n-gram by tf-idf. However, they are sensitive to lexical variation and hard to capture semantics of a caption, so they correlate poorly with human judgments [34].

Embedding-Based Evaluation Embedding-based metrics which use pre-trained models to extract embeddings and perform semantic matching in the embedding space, have been proven to correlate better with human judgments. BERTScore [34] uses contextual word embeddings generated by BERT, and measures the semantic similarity of two texts by computing token-level cosine similarity. BERTScore can be regarded as a special case of ours, it only uses references for evaluation and performs single fine-grained embedding matching. Among these embedding metrics, some works try to take into account the vision information. Tiger [8] uses a trained image-text matching SCAN model [10] to compare the ground outputs between candidate caption and reference. ViLBERTScore [9] uses a pre-trained ViLBERT model [16] to compare the visually-grounded text representation between candidate caption and reference. In these two evaluation metrics, the image is used as a visual ground in the evaluation rather than as ground truth, and they are still reference-based metrics. CLIP-Score [7] and FAIEr [30] are recently proposed reference-free evaluation metrics. CLIPScore [7] uses the pre-trained image-language model CLIP [21] to obtain image and text embeddings, and compute the cosine similarity. But they only consider coarse-grained matching and ignore fine-grained ones, so that CLIPScore lacks interpretability and ignores that a more precise score comes from fine-grained matching. FAIEr [30] introduces the scene graph to evaluate the fidelity and adequacy of the image captions. The above metrics are all proposed for image captioning. In this paper, we propose an evaluation metric specifically for video captioning by introducing video content. We consider not only coarse-grained embedding matching between video and text but also the fine-grained embedding matching between frames and words to take into account the characteristic of the visual elements of the video over time.

2.2. Pre-trained Vision-Language Models

Inspired by the success of the large-scale pre-training in NLP [5, 22], large-scale pre-training models [11, 16, 17, 27] also become the research hotspot in the vision-language community. Generally, these models are pre-trained by pre-text tasks on large-scale datasets, such as Conceptual captions [24] and HowTo100M [18] . During the pre-training, the models learn to narrow the gaps between the vision and language embeddings, which enables them to generalize well to various down-stream tasks like VQA [2], Visual Grounding [14], Image/Video-Text retrieval and Image/Video Captioning [13, 26, 33]. Motivated by the narrowed embedding gaps, we exploit one large-scale pre-trained model: CLIP [21], which is pre-trained via contrastive learning on 400 million image-text pairs, to design a video caption metric. CLIP-straight [20] shows that straight forward applying CLIP to video-text retrieval can achieve excellent zero-shot performance, which proves that the gaps between the extracted video and text embeddings are reduced. Therefore, by CLIP, measuring the consistency between the video content and the candidate caption is transformed to computing the cosine similarity between the extracted video and caption embeddings.

3. EMScore

Fig.2 shows the pipeline of EMScore, which computes the embedding similarity of the generated captions and the source video to achieve reference-free caption evaluation.

3.1. Embedding Extraction

We use CLIP [21] to extract video and text embeddings at both fine-grained and coarse-grained levels. Specifically, the visual encoder E_v (ViT-B/32) [6] extracts the embeddings of individual frame and total video. The language encoder E_t (Transformer) [22] extracts the embeddings of each token and whole sentence.

Frame and Video Representation Given a video $V = \{v_i\}_{i=1}^{|V|}$ ($|V|$ is the number of frames), each fine-grained frame embedding f_{v_i} is obtained as follows:

$$\mathbf{f}_{v_i} = \text{Norm}\left(E_v(v_i)\right), \mathbf{f}_v \in \mathbb{R}^d, \tag{1}$$

where Norm$(\cdot)$ is a L2 normalization function.

The coarse-grained video embedding $\mathbf{f}_V$ is the normalization of the mean-pooling of all the frame embeddings:

$$\mathbf{f}_V = \text{Norm}\left(\frac{1}{|V|} \cdot \sum_{i=1}^{|V|} \mathbf{f}_{v_i}\right), \mathbf{f}_V \in \mathbb{R}^d. \tag{2}$$

Word and Text Representation Given a caption, we first use the default tokenizer of CLIP to obtain word tokens and then add two special tokens [SOS] and [EOS] to construct a new token sequence $X = \{x_j\}_{j=1}^{|X|}$ ($|X|$ is the number of tokens). The contextual token embeddings are:

$$\begin{gathered}\{\mathbf{f}_{sos}, \mathbf{f}_{x_1}, \cdots, \mathbf{f}_{x_{|X|-2}}, \mathbf{f}_{eos}\} = \text{Norm}\left(W \cdot LN(E_t(X))\right), \\ \mathbf{f}_x \in \mathbb{R}^d,\end{gathered} \tag{3}$$

where LN is Layer Normalization, $W \in \mathbb{R}^{h \times d}$ are fixed parameters from CLIP, and h is the hidden size of text encoder. All these $|X|$ token embeddings are used for fine-grained embedding matching and the last $\mathbf{f}_{eos}$ is treated as the global embedding $\mathbf{f}_X$ for coarse-grained embedding matching.

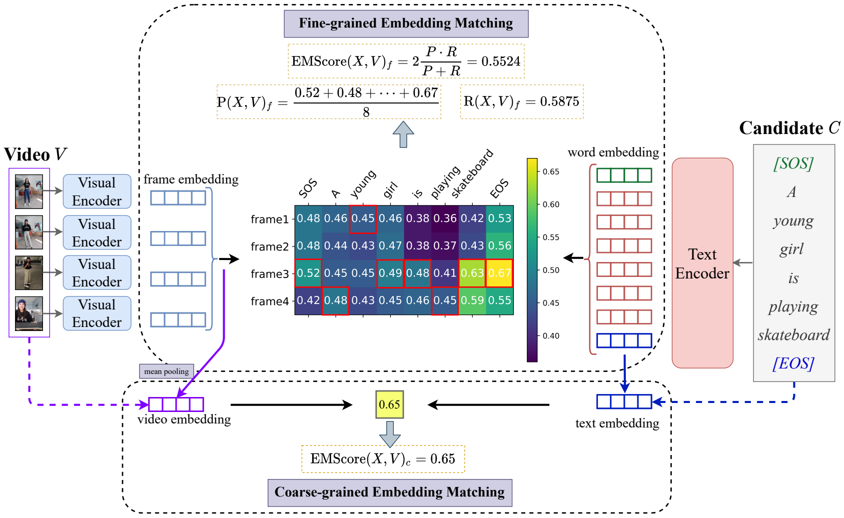

Fine-grained Embedding Matching
$\mathrm{EMScore}(X,V)_f = 2\frac{P \cdot R}{P+R} = 0.5524$
$\mathrm{P}(X,V)_f = \frac{0.52+0.48+\cdots+0.67}{8}$
$\mathrm{R}(X,V)_f = 0.5875$
Video V
Visual Encoder
Visual Encoder
Visual Encoder
Visual Encoder
frame embedding
SOS A young girl is playing skateboard EOS
frame1 0.48 0.46 0.45 0.46 0.38 0.36 0.42 0.53
frame2 0.48 0.44 0.43 0.47 0.38 0.37 0.43 0.56
frame3 0.52 0.45 0.45 0.49 0.48 0.41 0.63 0.67
frame4 0.42 0.48 0.43 0.45 0.46 0.45 0.59 0.55
0.65
0.60
0.55
0.50
0.45
0.40
word embedding
Text Encoder
Candidate C
[SOS]
A
young
girl
is
playing
skateboard
[EOS]
mean pooling
video embedding
0.65
text embedding
$\mathrm{EMScore}(X,V)_c = 0.65$
Coarse-grained Embedding Matching

Figure 2. Illustration of the computation of the EMScore which uses video as ground truth. Given the video V and candidate caption C, we extract global representations of video and caption for coarse-grained vector matching EMScore(X,V)$_c$, and local representations of frames and words for fine-grained greedy matching EMScore(X,V)$_f$. We highlight the precision greedy matching in red, and for simplicity, we give the calculation without idf weighting. The overall EMScore is the average score of EMScore(X,V)$_c$ and EMScore(X,V)$_f$.

3.2. Embedding Matching

Coarse-grained Embedding Matching Given the source video V and the generated caption X, the coarse-grained embedding matching $\mathrm{EMScore}(X,V)_c$ is:

$$\mathrm{EMScore}(X,V)_c = \mathbf{f}_X^\top \mathbf{f}_V, \text{[1]} \tag{4}$$

where $\mathbf{f}_V$ and $\mathbf{f}_X$ are embeddings of the video and caption, respectively. Such process is shown in lower part of Fig. 2.

Fine-grained Embedding Matching For videos, since visual elements in the frame change over time, only performing the coarse-grained embedding matching may lose detailed information, which inspires us to design a fine-grained embedding matching to achieve frame-token alignment. The upper part of Fig.2 shows the applied fine-grained matching. Given the video frame embedding $\mathbf{f}_v$ and the sentence token embedding $\mathbf{f}_x$, we first compute the precision (P) and recall (R) and then combine them to get the F1 score (F) as our fine-grained embedding matching score $\mathrm{EMScore}(X,V)_f$:

$$P(X,V)_f = \frac{1}{|X|}\sum_{x_i \in X} \max_{v_j \in V} \mathbf{f}_{x_i}^\top \mathbf{f}_{v_j}, \tag{5}$$

$$R(X,V)_f = \frac{1}{|V|}\sum_{v_j \in V} \max_{x_i \in X} \mathbf{f}_{x_i}^\top \mathbf{f}_{v_j}, \tag{6}$$

$$\mathrm{EMScore}(X,V)_f = 2\frac{P \cdot R}{P+R}. \tag{7}$$

By such token-frame matching in the calculation of precision, it is easy to figure out which visual frame is aligned with a specific word. The precision evaluates the correctness of the caption, such as whether descriptions are related to the video content without incorrect details. Similarly, it is easy to figure out which word is aligned with a specific visual frame in the calculation of recall. The recall evaluates the comprehensiveness of the caption, such as whether the content in the video is described without omission. The F1 measure combines the evaluation of these two aspects.

IDF Weighting A caption usually consists of two kinds of words: visual content words like nouns and function words like "the", "and", *etc*. For these function words, it is hard to align them with the video frames and thus we should lower their importance weight during token-frame matching. Since the more visual-irrelevant words will appear more times in the whole caption corpus, *e.g*., the word "a" may appear in every sentence, we calculate the inverse document frequency (idf) to weigh the importance of each word and integrate it into EMScore. Given a corpus $\{X^{(i)}\}_{i=1}^N$, the idf value of a token x is:

$$\mathrm{idf}(x) = -\log \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[x \in X^{(i)}\right], \tag{8}$$

where $\mathbb{I}[\cdot]$ is an indicator function. The special token [EOS] appears in each caption and Eq. (8) will assign its weight as 0, while this token contains comprehensive contextual information of the whole sentence since it is used as the discriminative signal for classification during the pre-training in CLIP. To remedy this, we empirically set the idf value of the [EOS] token to the average value of the entire idf set.

After calculating the idf values, the Precision in Eq. (5) is changed to:

$$P(X,V)_f = \frac{\sum_{x_i \in X} \mathrm{idf}\left(x_i\right) \max_{v_j \in V} \mathbf{f}_{x_i}^\top \mathbf{f}_{v_j}}{\sum_{x_i \in X} \mathrm{idf}\left(x_i\right)}. \tag{9}$$

When calculating Precision and Recall, IDF is applied for X and V, respectively. Note that idf weighting will not affect the calculation of Recall in Eq. (6) since each frame is equally important.

[1] Since all the embeddings are L2 normalized, the cosine similarity is reduced to the inner product.

System	GT	Top-Down	ORG-TRL	AM_1	AM_2	AM_3
Average Score	4.750	3.920	4.003	3.916	3.854	3.793

Table 1. Average scores for the six different caption source.

3.3. EMScore & EMScore_ref

When calculating EMScore, we do not need any reference and only use the video V. Specifically, EMScore is defined as the average of EMScore$_c$ and EMScore$_f$:

$$\text{EMScore}(X,V)=\frac{\text{EMScore}(X,V)_c+\text{EMScore}(X,V)_f}{2}. \tag{10}$$

The score is in the range [-1, 1]. A higher EMScore indicates a better caption, as it is more consistent with the video.

When the reference caption X^* is available, we can incorporate it to get EMScore_ref. First, EMScore(X, X^*) are calculated as in Eq. (10) by replacing V with X^*, and the ground truth embeddings are changed from the frame and video representations to word and text representations. Second, we define enhanced EMScore_ref as the average of EMScore(X, V) and EMScore(X, X^*).

$$\text{EMScore\_ref}(X,V,X^*)=\frac{\text{EMScore}(X,V)+\text{EMScore}(X,X^*)}{2}. \tag{11}$$

If there are multiple reference sentences $\{X_i^*\}_{i=1}^{M}$, $\text{EMScore}\,(X, X^*) = \max_i \text{EMScore}\,(X, X_i^*)$. In the following, unless otherwise specified, EMScore refers to EMScore(X, V), and EMScore_ref refers to EMScore(X, V, X*).

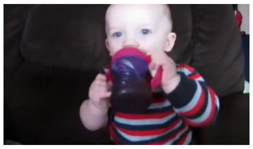
4. The Collected Datasets

4.1. The VATEX-EVAL Dataset

The VATEX-EVAL dataset is collected to evaluate the correlation of automatic metrics with human judgment.

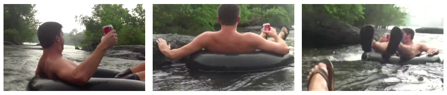
Candidate Caption Collection We use all 3000 validation videos from VATEX [31] and collect a total of 18,000 candidate captions with 6 captions per video. To span the full range of caption quality, for each video, we collect three kinds of captions: one high-quality, two medium-quality, and three low-quality captions. Specifically, for high-quality captions (GT), they are randomly selected from original ground-truth reference captions; for medium-quality captions (Top-Down and ORG-TRL), they are generated from Top-Down [1] and ORG-TRL [35] captioning models; for low-quality captions (AM_1, AM_2, AM_3), they are selected from other videos in the VATEX validation dataset by adversarial matching. More details of the caption collection are in the Appendix.

Human Evaluation Setup To ensure high quality of the human evaluation, each candidate caption is scored by 3 English-speaking annotators, amounting to 54,000 human ratings. For each video, we ask 3 annotators to rate the consistency degrees between the captions between the video. The rate scales from 1 to 5 where 1 denotes inconsistent and 5 denotes consistent. Fig.3 shows one example where incorrect details are red highlighted. Annotators are provided with detailed instruction (refer to Appendix), which is written to minimize subjectivity in annotations.



sentences	score
A baby is sitting on a couch and drinking from a sippy cup.	5
A baby is drinking from a sippy cup while **a woman** talks to him.	4
A baby is drinking from a sippy cup **while a boy is playing a ball.**	3
A baby is **playing with toys and a puppy is next to it.**	2
A gril is playing with toys and a puppy is next to it.	1

Figure 3. An annotation example for the VATEX-EVAL dataset. Incorrect details in the captions are red highlighted.



original correct paragraph	A man is floating on an intertube in a ***river***. He is drinking a soda and laughing as he floats. Another man runs into him several times.
foil paragraph	A man is floating on an intertube in a ***pool***. He is drinking a soda and laughing as he floats. Another man runs into him several times.

Figure 4. A correct-foil pair example for ActivityNet-FOIL.

Dataset Analysis We demonstrate the reliability of our collected VATEX-EVAL dataset from two aspects. Firstly, to check the agreements among different annotators, we compute the Kendall and Spearman correlation coefficients, which are 0.568 and 0.628 respectively. These inter-annotator correlations indicate strong inter-annotator agreements. Secondly, Tab.1 presents the average annotation scores for the six candidate caption collection sources. The average score of the original ground-truth captions strongly outperforms those of all other caption types, which is in line with the fact that GT captions have the highest quality intuitively. The ORG-TRL model gets a higher annotation score than the Top-Down model, which is also positively correlated with model complexity. The three captions of adversarial matching also give reasonable and reliable scores in the order of the adversarial matching score. The above analysis proves that our annotations are reliable.

4.2. The ActivityNet-FOIL Dataset

Prior work demonstrates that current captioning models generally generate "hallucinating" descriptions [23] that are not actually in the source visual scene. To test how sensitive EMScore is to identify foil captions that contain inaccurate visual concepts, we follow FOIL-COCO dataset [25] to change ActivityNet-Entities test dataset [36] for constructing an ActivityNet-FOIL dataset. In ActivityNet-Entities, each video has two corresponding paragraphs. We use one of two paragraphs to construct correct-foil pair, and use the other as a reference for reference-based metrics. Each paragraph has about 3 sentences in different time stamps, and a visual concept in each sentence is grounded to an annotation bounding box. A foil caption is created by replacing the original visual concept with a similar but false one.

Our data generation process has three main steps: First, we collect all visual concepts and filter out the ones with low frequency. Then we pair together words belonging to the same supercategory (such as river-pool, shirt-shoe, cat-dog). At last, we obtain 2,191 correct-foil pairs, in which each visual concept has approximately 13 foil ones. Second, we replace a visual concept in the original correct caption with paired foil candidate to construct a candidate foil caption. Each correct caption has multiple candidate foil captions. Third, for each correct caption, we mine the hardest foil caption by selecting the lowest perplexity candidate. Finally, we create 1900 correct-foil paragraph pairs, and at least one caption in the foil paragraph contains a foil visual concept. As shown in Fig.4, it contains a correct-foil paragraph pair per video. We compute the accuracy of each metric in its capacity to assign a higher score to the correct candidate paragraph versus the foil. More details about the ActivityNet-FOIL collection can be seen in the Appendix.

5. Experiments

We conduct experiments to evaluate our EMScore and EMScore_ref on VATEX-EVAL (cf. Section 5.1) and ActivityNet-FOIL (cf. Section 5.2) datasets. To measure caption-level human correlation, we compute Kendall's correlation τ and Spearman's rank correlation ρ.

We compare EMScore with four rule-based metrics, *e.g.*, BLEU [19], ROUGE_L [12], METEOR [3] and CIDEr [28][2], and two embedding-based metrics, *e.g.*, BERTScore [34][3] and Improved BERTScore [32][4]. For these two embedding-based metrics, we utilize RoBERTa-base [15] as the backbone, and use F1-measure with idf and Recall with idf, respectively, which are the best setting in their paper. For our EMScore and EMScore_ref, they can also optionally combine with idf. Specifically, the training caption corpus from the source dataset (VATEX and ActivityNet) is used to calculate idf. For the value of $|V|$, we use all frames in video. The value of h and d are both 512.

5.1. Results on VATEX-EVAL

5.1.1 Ablation Study

Effect of P, R, F, and idf weighting From Tab.2, we can see that F1-measure has achieved relatively stable performance regardless of whether idf weighting is used or not. After adding idf weighting, the performance on precision and F1-measure of EMScore is improved. The result proves that idf weighting is effective. The best performance is obtained under the combination of calculating F1-measure and using idf weighting. Therefore, in the following, we use F1-measure combined with idf weighting as the default setting.

Metric	τ	ρ
EMScore$_f$ (P)	0.1843	0.2404
EMScore$_f$ (R)	0.2263	0.2946
EMScore$_f$ (F)	0.2228	0.2900
EMScore$_f$ (P-idf)	0.2052	0.2674
EMScore$_f$ (R-idf)	0.2263	0.2946
EMScore$_f$ (F-idf)	**0.2296**	**0.2989**

Table 2. The performance difference of different calculation methods and the effect of idf weighting on fine-grained EMScore$_f$. τ/ρ indicates the Kendall/Spearman correlation, respectively.

#	Metric	GT	τ	ρ
1	EMScore$_c$	V	0.2269	0.2955
2	EMScore$_f$ (F-idf)	V	0.2296	0.2989
3	EMScore (F-idf)	V	**0.2324**	**0.3026**
4	EMScore$_c$	X*	0.2390	0.3104
5	EMScore$_f$ (F-idf)	X*	0.2495	0.3240
6	EMScore (F-idf)	X*	**0.2550**	**0.3307**
7	EMScore_ref$_c$	V+X*	0.2738	0.3548
8	EMScore_ref$_f$ (F-idf)	V+X*	0.2779	0.3599
9	EMScore_ref (F-idf)	V+X*	**0.2863**	**0.3705**

Table 3. The effect of different granularities embedding matching and the effect of different ground truths. GT, V, X* are denoted as ground truth, video, and reference. For X*, there is one reference. τ/ρ indicates the Kendall/Spearman correlation, respectively.

Metric	No Ref		1 Ref		9 Refs	
	τ	ρ	τ	ρ	τ	ρ
BLEU_1	-	-	0.1219	0.1591	0.289	0.3697
BLEU_4	-	-	0.0806	0.0881	0.216	0.256
ROUGE_L	-	-	0.1249	0.1631	0.2378	0.3085
METEOR	-	-	0.1644	0.2149	0.2763	0.3574
CIDEr	-	-	0.1732	0.2263	0.2781	0.3606
BERTScore (F-idf)	-	-	0.1824	0.2373	0.293	0.3775
Improved_BERTScore (R-idf)	-	-	0.1516	0.198	0.2442	0.3167
EMScore (F-idf)	**0.2324**	**0.3026**	-	-	-	-
EMScore_ref (F-idf)	-	-	**0.2863**	**0.3705**	**0.3681**	**0.4719**

Table 4. Human correlation on the VATEX-EVAL dataset. τ/ρ indicates the Kendall/Spearman correlation respectively.

Effect of Different Granularities and Ground Truths In Tab.3, we first observe the impact of different granularities. For EMScore which uses video as ground truth (GT), the fine-grained EMScore in line 2 achieves better results than the coarse-grained one in line 1. The result verifies our motivation that it is correct to consider the characteristics of the visual elements of the video over time in the video caption evaluation process. Moreover, the performance of the combination of two granularities in line 3 is further improved. The result verifies that multi-granularity combination is beneficial. Next, we observe the impact of using different ground truths. When using both video and reference as GT, a better correlation result is achieved than using them alone. The result proves our conjecture that the information in the video and references are complementary, and additional use of references can bring information gain. Therefore we recommend using EMScore_ref when references are available.

5.1.2 Comparsion with the other metrics

In the following experiment, we prove that our EMScore achieves higher human correlation and lower reference dependency, which benefits from the introduction of the video content. We also show that our metric is robust to quality

[2] These metrics are implemented in MS COCO evaluation tool https://github.com/tylin/coco-caption.

[3] https://github.com/Tiiiger/bert_score

[4] https://github.com/ck0123/improved-bertscore-for-image-captioning-evaluation

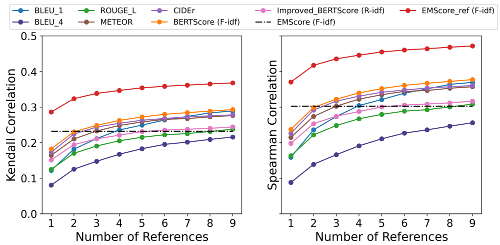


Figure 5. The Kendall and Spearman correlations between automatic metrics and human judgments with the different numbers of references on the VATEX-EVAL dataset. The dashed line indicates EMScore, which does not rely on any reference.

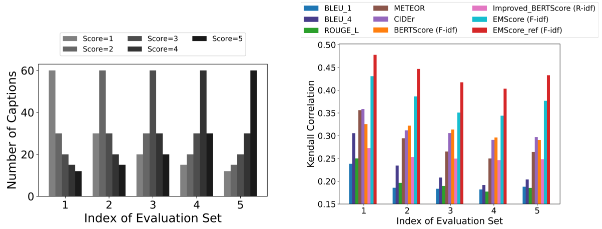
drift and has a consistent system-level ranking with humans.

High Human Correlation Tab.4 shows the correlation results of metrics with 0, 1, 9 references per candidate caption. We have several observations as follows: (1) In situations where no references are available, our EMScore still works well, and achieves surprisingly competitive results. The result demonstrates the advantage of taking video content into account, while other reference-based metrics cannot handle this situation; (2) When using the same number of references (e.g., 1 or 9), our EMScore_ref outperforms other prior metrics by a large margin. The comparison results prove that our metric achieves higher human correlation and we propose a more reliable metric.

Low Reference Dependency The Kendall and Spearman correlations between automatic metrics and human judgments with the different numbers of references are shown in Fig.5. Our EMScore without any references can achieve competitive results, compared with reference-based metrics which need at least 4 or 5 references, such as BLEU_1 and Improved_BERTScore. Besides, our EMScore_ref with only one reference can achieve comparable results with reference-based metrics, which need at least 8 or 9 references, such as CIDEr and BERTScore. The results show that our metric has lower reference dependency, which benefits from the introduction of video content in evaluating.

Robust to Quality Drift It is extremely important for metrics to deal with quality drift, since the quality of generated captions can vary significantly across different video captioning models. To assess the robustness of metrics to quality drift, we create biased sets from our annotated VATEX-EVAL dataset by sampling candidate captions of different quality levels with different probabilities. Specifically, the annotation score of each caption ranges from 1 to 5. We then create 5 biased sets, indexed by the variable $I \in \{1, 2, \cdots, 5\}$. For the I^{th} set, we sample the candidate captions whose annotation score is k with a probability of $\frac{1}{|I-k|+1}$, where $k \in \{1, 2, \cdots, 5\}$.

In this way, the 5 sets have different distributions of candidate captions with different qualities, as shown in Fig.6 (a). We compute the Kendall correlation between different metrics and human judgments on the 5 sets. One reference is used in the reference-based metrics and our EMScore_ref. Fig.6 (b) shows that: (1) Our metrics EMScore and EMScore_ref have a higher correlation than other metrics on all biased sets, which proves that our metrics are robust to the quality drift; (2) We find that rule-based metrics, e.g., BLEU_4, perform much better on low-quality captions (set 1) than on high-quality (set 5). With the development of video captioning, they will become increasingly unreliable because they struggle to judge high-quality captions.



(a) Distribution of captions of different annotation scores in different test sets.

(b) The Kendall correlation of different metrics with human judgments on different test sets.

Figure 6. Robustness of metrics over different caption quality biased sets. One reference is used in reference-based metrics and our EMScore_ref.

System-level Ranking on VATEX-EVAL Video captioning researchers generally report system-level scores to verify the effectiveness of their methods, so, it is essential to measure the metric's system-level human correlation. A reliable metric is expected to have the same system ranking as humans. In Tab.5, we compare the ranking of six system average scores rated by metrics and humans on the VATEX-EVAL datasets. All the metric scores are scaled to [0, 1], including human scores. For the human scores, the GT system gets the highest score, and following by ORG-TRL, Top-Down, AM_1, AM_2, AM_3. We use red fonts to highlight that the metric's ranking is inconsistent with human ranking. We can see that CIDEr and BERTScore cannot correctly rank GT, ORG-TRL, and Top-Down systems, *e.g.*, they give the highest score to the ORG-TRL system instead of GT. Our EMScore and EMScore_ref are consistent with human ranking. The result shows that our EMScore and EMScore_ref are more reliable in system-level evaluation than other metrics, and it will be beneficial for the video captioning development.

5.1.3 EMScore Visualization

Fig.7 visualizes how fine-grained EMScore matches the most similar visual elements to the tokens (as the calculation of precision). For the first example, "bubbles" is occurred in the 106th frame, "another boy" is occurred in the 160th and 187th frames, and compared with other frames, "face paint" appears in a larger proportion in the 4th and 6th frames. For the second example, the visual concept "boy" appears as the main visual element in the 53th frame, so the token 'boy' matches this frame instead of 84th∼298th frames where multiple visual elements appear. Compared with coarse-grained embedding matching, our fine-grained

System	Human		EMScore (F-idf)		EMScore_ref (F-idf)		CIDEr		BERTScore (F-idf)	
GT	0.937	(1)	0.581	(1)	0.639	(1)	0.178	(2)	0.498	(3)
ORG-TRL	0.751	(2)	0.539	(2)	0.606	(2)	0.185	(1)	0.527	(1)
Top-Down	0.730	(3)	0.530	(3)	0.591	(3)	0.173	(3)	0.515	(2)
AM_1	0.729	(4)	0.522	(4)	0.584	(4)	0.146	(4)	0.464	(4)
AM_2	0.714	(5)	0.515	(5)	0.571	(5)	0.140	(5)	0.451	(5)
AM_3	0.698	(6)	0.512	(6)	0.566	(6)	0.134	(6)	0.447	(6)

Table 5. System-level ranking on the VATEX-EVAL dataset. Nine references are used in the reference-based metrics and our EMScore_ref. Each column for the metrics in the table gives the score for each system and the ranking of the six systems. The red fonts highlight denote that the metric's ranking is inconsistent with humans.

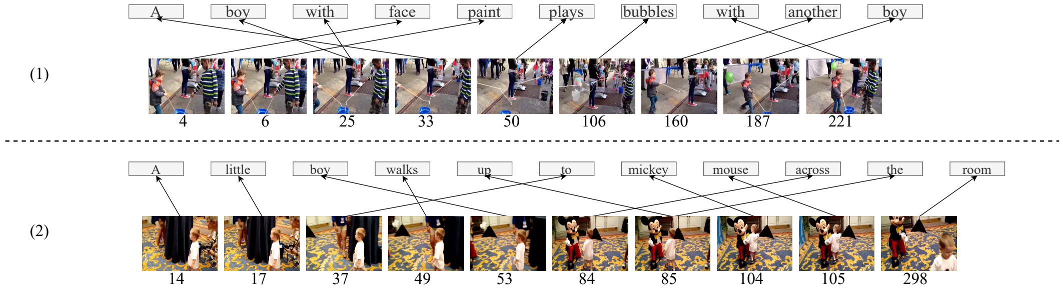


Figure 7. EMScore precision visualization. Each token is matched to the most similar frame. The temporal index is shown under the frame.

Metric	Accuracy(%)	Metric	Accuracy(%)
BLEU_1	60.11	EMScore$_c$	87.95
BLEU_4	66.11	EMScore$_f$ (F-idf)	90.32
ROUGE_L	56.74	EMScore (F-idf)	89.47
METEOR	72.89	EMScore_ref$_c$	90.21
CIDEr	77.89	EMScore_ref$_f$ (F-idf)	93.00
BERTScore (F-idf)	86.68	EMScore_ref (F-idf)	92.42

Table 6. Pairwise ranking accuracy on ActivityNet-FOIL dataset.

one can take into account the characteristic of the video, and provide more interpretability for EMScore.

5.2. Experiments on ActivityNet-FOIL

To test the capability of EMScore to identify "hallucinating" captions, we compute the accuracy of pairwise ranking for each evaluation metric in their capacity to assign a higher score to the correct candidate paragraph versus the foil on the ActivityNet-FOIL dataset. Each candidate paragraph has multiple captions, so we first compute the caption score, then calculate the overall score of the paragraph as the average score of multiple captions. The ground truth for each caption is obtained by cutting the video and the reference paragraph into multiple segments and reference captions, respectively, according to the timestamp of the candidate captions. The accuracy results are shown in the Tab.6, we have the following findings: (1) Even without any reference, our EMScore outperforms all reference-based metrics. Moreover, our EMScore reaches a noteworthy improvement in terms of accuracy by about 3% compared to the best prior metric (BERTScore 86.68%). The excellent result proves that it is effective to take the video content as ground truth in the hallucination caption identification; (2) When enhanced by the reference, our EMScore_ref$_f$ achieves the highest accuracy (93.00%); (3) Due to the large changes in the visual scene of the video in the ActivityNet dataset, it will be more effective to consider fine-grained embedding matching than coarse-grained one. At the same time, the multi-granularity combination does not bring performance improvement, and the results suggest that it is sufficient to use fine-grained matching alone for videos with large visual scene changes.

6. Conclusion

In this paper, we have conducted a systematic study on the video captioning evaluation metrics. First, to solve the drawbacks of reference-based metrics, we have proposed a novel video captioning evaluation metric EMScore by measuring the consistency between the video and caption. Second, we have collected two datasets (VATEX-EVAL and ActivityNet-FOIL) to systematically analyze the reliability of the existing metrics. The VATEX-EVAL experiments have demonstrated that our EMScore has a higher human correlation and lower reference dependency. Moreover, it is robust to quality drift, and consistent with humans on the system-level ranking. The ActivityNet-FOIL experiments have shown that our EMScore is sensitive to identifying "hallucinating" captions.

Limitations. EMScore is an embedding-based metric, and relies on the performance of used vision-language pretrained (VLP) model. More reliable evaluation scores can be obtained by leveraging better VLP models to extract better representations. More discussion about the effect of VLP models is in the Appendix.

Acknowledgements. This work is supported by Beijing Natural Science Foundation (Grant No. JQ21017), the national key R&D program of china (No. 2020AAA0105702), the Natural Science Foundation of China (Grant No. 61972397, 62036011, 62192782, 61721004, U19B2038, 61906192), the Key Research Program of Frontier Sciences, CAS, Grant No. QYZDJ-SSW-JSC040, the University Synergy Innovation Program of Anhui Province under Grants GXXT-2019-025, the Science and Technology Service Network Initiative, CAS (Grant No. KFJ-STS-SCYD-317).

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086, 2018. 5

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015*, pages 2425–2433, 2015. 3

[3] Satanjeev Banerjee and Alon Lavie. 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, 2005. 3, 6

[4] Jincan Deng, Liang Li, Beichen Zhang, Shuhui Wang, Zhengjun Zha, and Qingming Huang. Syntax-guided hierarchical attention network for video captioning. *IEEE Trans. Circuits Syst. Video Technol.*, 32(2):880–892, 2022. 2

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186, 2019. 3

[6] 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. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021*. 3

[7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 7514–7528. 3

[8] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 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 2019*, pages 2141–2152, 2019. 3

[9] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. ViLBERTScore: Evaluating image caption using vision-and-language BERT. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 34–39, 2020. 3

[10] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Computer Vision - ECCV 2018 - 15th European Conference, 2018*, Lecture Notes in Computer Science, pages 212–228, 2018. 3

[11] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: hierarchical encoder for video+language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 2046–2065. 2, 3

[12] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004. 2, 3, 6

[13] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for sequence-level image captioning. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018*, pages 1416–1424. 3

[14] Daqing Liu, Hanwang Zhang, Zheng-Jun Zha, and Feng Wu. Learning to assemble neural module tree networks for visual grounding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pages 4672–4681. 3

[15] Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019. 6

[16] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pages 13–23. 3

[17] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 9876–9886. 2, 3

[18] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pages 2630–2640. 3

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. 2, 3, 6

[20] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using CLIP. *CoRR*, abs/2102.12443, 2021. 3

[21] 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. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, pages 8748–8763. 2, 3

[22] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 3

[23] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018. 2, 5

[24] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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, ACL 2018*, pages 2556–2565. 3

[25] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 255–265, 2017. 5

[26] Ganchao Tan, Daqing Liu, Meng Wang, and Zheng-Jun Zha. Learning to discretely compose reasoning module networks for video captioning. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 745–752. 3

[27] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019*, pages 5099–5110. 3

[28] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 2, 3, 6

[29] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence - video to text. In *2015 IEEE International Conference on Computer Vision, ICCV 2015*, pages 4534–4542. 2

[30] Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. Faier: Fidelity and adequacy ensured image caption evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, pages 14050–14059. 3

[31] 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 *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pages 4580–4590. 5

[32] Yanzhi Yi, Hangyu Deng, and Jinglu Hu. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 985–994. 2, 6

[33] Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for fine-grained image captioning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(2):710–722, 2022. 3

[34] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*. 2, 3, 6

[35] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 13275–13285. 2, 5

[36] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, and Marcus Rohrbach. Grounded video description. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6571–6580, 2019. 5