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

Yaya Shi\textsuperscript{1}, Xu Yang\textsuperscript{2}, Haiyang Xu\textsuperscript{3}, Chunfeng Yuan\textsuperscript{4,*}, Bing Li\textsuperscript{4}, Weiming Hu\textsuperscript{4,5,6}, Zheng-Jun Zha\textsuperscript{1}
\textsuperscript{1}University of Science and Technology of China \textsuperscript{2}Southeast University \textsuperscript{3}Alibaba Group \textsuperscript{4}NLPR, Institute of Automation, Chinese Academy of Sciences \textsuperscript{5}School of Artificial Intelligence, University of Chinese Academy of Sciences \textsuperscript{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. \textsc{VATEX-EVAL} experiments demonstrate that EMScore has higher human correlation and lower reference dependency. \textsc{ActivityNet-FOIL} experiments verify that EMScore can effectively identify “hallucinating” captions. Code and datasets are available at https://github.com/shiyaya/emscore.

*Corresponding author

![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.](https://example.com/figure1.png)
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. CLIPScore [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 downstream 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 straightforward 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 $V = \{v_1\}_{i=1}^{|V|}$ ($|V|$ is the number of frames), each fine-grained frame embedding $f_{v_i}$ is obtained as follows:

$$f_{v_i} = \text{Norm}(E_v(v_i)), f_{v} \in \mathbb{R}^d,$$

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

The coarse-grained video embedding $f_v$ is the normalization of the mean-pooling of all the frame embeddings:

$$f_v = \text{Norm} \left( \frac{1}{|V|} \cdot \sum_{i=1}^{|V|} f_{v_i} \right), f_v \in \mathbb{R}^d. \quad (2)$$

**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:

$$\{f_{eos}, f_{s1}, \ldots, f_{s_{|X|-2}}, f_{eos} = \text{Norm}(W \cdot \text{LN}(E_t(X)))) \in \mathbb{R}^d,$$

where $\text{LN}$ 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 $f_{eos}$ is treated as the global embedding $f_X$ for coarse-grained embedding matching.
3.2. Embedding Matching

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

$$EMScore(X, V)_c = f_X^\top f_V,$$

where $f_V$ and $f_X$ 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 $f_i$ and the sentence token embedding $f_j$, 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 $EMScore(X, V)_f$:

$$P(X, V)_f = \frac{1}{|X|} \sum_{x_i \in X} \max_{v_j \in V} f_i^\top f_j,$$

$$R(X, V)_f = \frac{1}{|V|} \sum_{v_j \in V} \max_{x_i \in X} f_i^\top f_j,$$

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

By such token-frame matching in the calculation of precision, 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 $\{X^{(i)}\}^N_{i=1}$, the idf value of a token $x$ is:

$$idf(x) = -\log \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[x \in X^{(i)}],$$

where $\mathbb{1}[\cdot]$ 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:

$$P(X, V)_f = \frac{\sum_{x_i \in X} \text{idf}(x_i) \max_{v_j \in V} f_i^\top f_j}{\sum_{x_i \in X} \text{idf}(x_i)}.$$  

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.
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_e and EMScore_f:

\[
\text{EMScore}(X, V) = \frac{\text{EMScore}(X, V)_e + \text{EMScore}(X, V)_f}{2}. \tag{10}
\]

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\((X, X^*)\).

\[
\text{EMScore}_{\text{ref}}(X, V, X^*) = \frac{\text{EMScore}(X, V) + \text{EMScore}(X, X^*)}{2}. \tag{11}
\]

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 in the captions are red highlighted.

![An annotation example for the VATEX-EVAL dataset.](image)

**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 anno-

<table>
<thead>
<tr>
<th>System</th>
<th>GT</th>
<th>Top-Down</th>
<th>ORG-TRL</th>
<th>AM_1</th>
<th>AM_2</th>
<th>AM_3</th>
<th>Average Score</th>
</tr>
</thead>
</table>

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

![A correct-foil pair example for ActivityNet-FOIL.](image)
tation 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], and two embedding-based metrics, e.g., BERTScore [34] and Improved BERTScore [32]. 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 b 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.

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 Comparison 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

<table>
<thead>
<tr>
<th>Metric</th>
<th>τ</th>
<th>ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMScore (P)</td>
<td>0.1843</td>
<td>0.2504</td>
</tr>
<tr>
<td>EMScore (R)</td>
<td>0.2263</td>
<td>0.2946</td>
</tr>
<tr>
<td>EMScore (F)</td>
<td>0.2228</td>
<td>0.2900</td>
</tr>
<tr>
<td>EMScore (F-idf)</td>
<td>0.2052</td>
<td>0.2674</td>
</tr>
<tr>
<td>EMScore (R-idf)</td>
<td>0.2263</td>
<td>0.2946</td>
</tr>
<tr>
<td>EMScore (F-idf)</td>
<td>0.2296</td>
<td>0.2989</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>τ</th>
<th>ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERTScore (F-idf)</td>
<td>0.1824</td>
<td>0.2373</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.1644</td>
<td>0.2149</td>
</tr>
<tr>
<td>CIDEr</td>
<td>0.1732</td>
<td>0.2263</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Metric</th>
<th>τ</th>
<th>ρ</th>
<th>τ</th>
<th>ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>METEOR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIDEr</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Human correlation on the VATEX-EVAL dataset. τ/ρ indicates the Kendall/Spearman correlation respectively.
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 \( \frac{1}{|I-I'|+1} \), 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.

**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
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, 62192782, 61721004, U19B2038, 61906192), the Key Research Program of Frontier Sciences, CAS, Grant No. QYZDJ-SWW-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


