

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

A. The VATEX-EVAL Dataset

A.1. Candidate Caption Collection

System-Generated Caption Collection To generate diverse captions, we train two video captioning models on the VATEX dataset which consists of 25,991 training and 3,000 validation videos. Each video is paired with 10 English and 10 Chinese captions, and we only use English captions in this paper. The first caption model is a traditional Top-Down model which was first proposed for image captioning [1]. The Top-Down model has two LSTM layers. The first LSTM layer is a top-down visual attention model that utilizes soft attention to dynamically attend to video features. The second LSTM layer is used as a language model for caption generation. The video features used in this model contain InceptionResNetV2 [10] features and C3D [3] features. The second caption model is ORG-TRL [12] which achieves excellent performance in the video captioning task. Compared with the Top-Down model, ORG-TRL adds the object-relational graph to combine object features in the encoder and adds the external language model for teacher-recommended learning in the decoder. We use the default setting in the original paper to train the ORG-TRL model, use the baseline setting of ORG-TRL to train the Top-Down model.

Adversarial Matching Selection By observing system-generated captions, we find above video captioning models usually generate medium-quality captions. And human-written captions are usually in high-quality apparently. Considering the diversity of candidate caption quality, we add some low-quality captions by adversarial matching. Similar to VCR [11] which proposes adversarial matching to construct wrong answers for multiple-choice QA, by selecting a correct answer from another question, we use the same strategy to rich the diversity of caption quality. We denote the reference set for video V_i as $\{r_i^k\}_{k=1}^{n_{refs}}$, where n_{refs} is the number of references. We select a reference caption in other videos that have a top similar score with this video V_i . Specifically, we measure the sum sentence similarity score between other video references r_j and the

reference set of this video V_i : $\arg \max_j \sum_k \text{sim} \{r_j, r_i^k\}$,

and we use Sentence-BERT [7] to measure sentence similarity. In practice, we don't use the top-1 sentence but select top-3 to top-5 sentences to avoiding too many high-quality captions.

A.2. Annotation Instructions

Tab.1 shows annotation instructions used to guide human quality annotation, and an annotation example is shown in Fig.3 in main paper.

B. The ActivityNet-FOIL Dataset

Generation of replacement word pairs To construct a foil paragraph, we replace a visual concept in the original correct caption with an incorrect but similar word (the foil). To achieve this, we construct the replaceable correct-foil pairs. Firstly, we collect all visual concepts that are pre-annotated and filter out which less than 100 times, obtaining 235 unique visual concepts. Then, we use spacy¹ to extract semantic embedding and use spectral cluster [9] to obtain 15 supercategory. We pair together words belonging to the same supercategory (e.g., shirt-shoe, mountain-park, cat-dog). Wrong pairs such as man-male are filtered out manually. Finally, we obtain 2,191 correct-foil pairs, and each visual concept has approximately 13 foil ones.

Mining the hardest foil caption Above correct-foil pairs are used to generate the candidate foil captions for each correct caption. But not all generated foil captions are natural, such as "A man in a blue *toy* is kneeling down". These captions are simple samples that can be easily detected by learnable language models. To eliminate the possible language bias of foil captions, we only select the hardest ones. For this purpose, we use GPT2 [6]² to compute the Perplexity (PPL) for each candidate foil caption and the one with the lowest PPL will be selected as the hardest foil. Finally, we create 1900 correct-foil paragraph pairs, and at least one caption in the foil paragraph contains a foil visual concept.

¹We use "en_core_web_lg" model

²We use "gpt2-medium" model provided in the <https://>

Annotation Instructions	
You will be given a video and several sentences.	
Your task is to rate each of the sentences on correctness, that is, whether a sentence is consistent with the content of the video. Specifically, A sentence should be rewarded for describing the content in the video and penalized for describing content unrelated to the video.	
We wish you according to the rules introduced below, rate the correctness of the sentence with respect to the video.	
Score	Rule
5	The description in the sentence is completely correct, without any error
4	The description in the sentence is almost completely correct, but there is a little error
3	The proportion of wrong and right is nearly equal
2	The description in the sentence is almost completely incorrect, but there is something right
1	The description in the sentence is completely incorrect, nothing is correct

Table 1. Annotation instructions of VATEX-EVAL dataset.



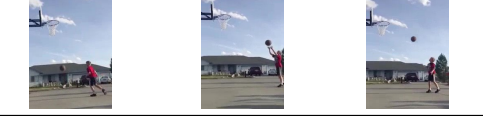
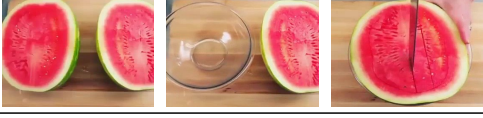

No.	Video	References and Candidate	Human	EMScore	EMScore_ref	BERTScore
(a)		<ul style="list-style-type: none"> R1: The man is doing some hand and arm exercises as he lifts weights. R2: A man holding dumbbell weights is moving his arms up and down first to his sides, and then in front of him , before repeating the exercise. R3: A man uses dumbbells to exercise his arms and shoulders. C: A man in a gym is demonstrating how to lift weights. 	1.0	0.682	0.705	0.590
(b)		<ul style="list-style-type: none"> R1: Two young girls are standing on a trampoline practicing synchronized hand jive. R2: Two young girls are standing on a trampoline and playing a patty-cake game and talking to each other. R3: Two young girls stand on a trampoline playing a hand clapping game. C: A group of children are jumping on a trampoline and playing basketball. 	0.333	0.499	0.45	0.652
(c)		<ul style="list-style-type: none"> R1: The boy tried three times to make a shot with the basketball. R2: A person is counting as a boy is trying to throw a basketball through the hoop. R3: A young boy is practising how to shoot a basketball on the street. C: A boy throws a basketball at a hoop where it bounces back, he then throws it again this time through the hoop. 	0.917	0.766	0.798	0.415
(d)		<ul style="list-style-type: none"> R1: A half of a watermelon is placed in a glass bowl and cut into smaller pieces. R2: Two halves of a watermelon are sitting on a counter and one half is placed in a glass bowl then using a knife the interior of the melon is sliced in a grid pattern. R3: A person places a watermelon inside of a bowl and slices it. C: A watermelon in halves are seen on the table and one of them is placed on a bowl for slicing. 	1.0	0.767	0.705	0.494
(e)		<ul style="list-style-type: none"> R1: Several couples are on the dance floor dancing and twirling. R2: Two people dance and spin around on the dance floor. R3: A couple having a good time with others dancing the night away. C: A man and a woman dance with each other among a group of other dancers while music plays in the background. 	1.0	0.778	0.705	0.346

Figure 1. Examples of evaluation scores assigned by Human, EMScore (F-idf), EMScore_ref (F-idf), and BERTScore (F-idf). R and C denote reference and candidate caption, respectively. Red highlights indicate descriptions that do not appear in the reference text. Green highlights indicate descriptions that are inconsistent with the video content. Purple highlights indicate that there is some common sense that can be inferred from the appearance of the video.

C. Qualitative Analysis

C.1. Qualitative Analysis on VATEX-EVAL

Fig. 1 shows some evaluation example of EMScore (F-idf), EMScore_ref (F-idf), and BERTScore (F-idf) on the VATEX-EVAL dataset. Ideally, the score assigned by these automatic metrics should be similar to the score assigned by the human score. All the metric scores are scaled to [0, 1], including human scores. We have the following observations: (1) From case (a), we

can see that even if there is a difference between candidate and reference, such as “in the gym” is not occur in references, our two metrics EMScore and EMScore_ref give the candidate caption a reasonably high score similar to the human judgment, but BERTScore gives it a much lower score. The result demonstrates that using video as ground truth for evaluation is effective, especially when the description in references cannot comprehensively express the content of the video. And the result proves that our metric can solve the problem of over-punishing the correct captions in the reference-based metrics; (2) Case (b) shows that when there is an “hallucinating” description





No.	Video	References and Candidate	EMScore	EMScore_ref	BERTScore
(a)		<p>Correct/Foil: People play holding a pole/bucket and hitting heavy balls to a target. Two poles are on front the ball with handle.</p> <p>Reference: The stones heat each other causing them to move forward. Three people are watching after the stones. Two people are brushing the front of the stone.</p>	✓	✓	✗
(b)		<p>Correct/Foil: A woman is seen speaking to the camera while holding up various ingredients. She begins pouring the ingredients together into a bowl/bottle while still speaking to the camera. She then pours dressing into the bowl/dish and presents a salad she had made.</p> <p>Reference: Woman is holding a clear plastic container. Woman is emptying salad onto the plate. Woman is opening a clear plastic container holding tomatoes.</p>	✓	✓	✗
(c)		<p>Correct/Foil: Several metal pieces are lying on a black cloth/towel. A man shows the pieces up close He uses them as darts outside , throwing them at a wooden board/box.</p> <p>Reference: Various tools are laid out in fabric with a person unraveling them. The person shows off the tools to the camera. The person is then seen out back throwing the object against wood and showing it to the camera.</p>	✗	✓	✓
(d)		<p>Correct/Foil: A man is lifting weights in a weight room , pulling a large barbell up to his chest. He stands shaking before lifting it above his head/neck, trying to hold it in place He drops the barbell hard onto the ground/mat.</p> <p>Reference: A man bends on front a wight. Then, the man raises the weight until the shoulders. After, the man holds the weight above the head, and then he falls the weight to the floor.</p>	✗	✓	✓

Figure 2. Pairwise ranking examples for EMScore (F-idf), EMScore_ref (F-idf), and BERTScore (F-idf) on the ActivityNet-FOIL dataset. ✓ denotes that the metrics give a right pairwise ranking, and vice versa.

(such as “a group of children”) of the video content in the caption, our EMScore and EMScore_ref give a low score similar to humans. But BERTScore failed to handle this case and give an unreasonable high score because they only measure text-level comparison and ignore the visual relevance. However, our EMScore and EMScore_ref which use video as ground truth can effectively avoid the problem of under-punishing the incorrect captions. (3) Case (c) and (d) show that the BRETScore cannot handle long candidate captions. The reason is that BERTScore only calculates fine-grained embedding matching, and cannot handle global semantics in long texts. But our metric EMScore not only use fine-grained embedding matching but also coarse-grained one, and gives a reasonably high score; (4) From case (e), we can see that our metric EMScore can correctly evaluate common-sense descriptions (such as, “while music plays in the background”) which can be inferred from the video content. We attribute success to the pre-trained vision-language model which is not only used to extract embeddings but also provides common-sense knowledge related to vision. But BERTScore lacks visual common sense because it uses a pure-language pre-trained model, and when the information is lost in the references, it cannot handle this case.

C.2. Qualitative Analysis on ActivityNet-FOIL

Fig. 2 shows some pairwise ranking examples of EMScore (F-idf), EMScore_ref (F-idf), and BERTScore (F-idf)

Metric	Time (s)	Speed (cands/s)
BERTScore	13.54	1,329
EMScore(X,V)	18.84	955
EMScore(X,X*)	15.40	1,169
EMScore_ref(X,V,X*)	24.28	741

Table 2. Speed analysis on the VATEX-EVAL dataset. The reported values are an average of 3 runs.

on ActivityNet-FOIL dataset. Ideally, these automatic metrics should be assigned a higher score to the original correct paragraph than the corresponding foil one. We find that EMScore and EMScore_ref can effectively use video information. For case (a), the visual concept “a pole” is included in the video content but not described in the reference. The reference-based metric BERTScore failed to handle this case. But our metrics EMScore and EMScore_ref take advantage of its use of video as ground truth and give the right pairwise ranking. Case (b) also supports this finding. Besides, we find a challenge for EMScore. Similar visual concepts are difficult to detect by EMScore, such as, “cloth” and “towel” in case (c), “head” and “neck” in case (d). We look forward to a better vision-language model in the future that can be able to distinguish these similar visual concepts.

Models		R@1	R@10	EMScore _c		EMScore _f		EMScore	
				τ	ρ	τ	ρ	τ	ρ
other VLP models	MIL-NCE	9.6	33.7	0.1505	0.1972	-	-	-	-
	Frozen-in-Time	21.6	62.7	0.2096	0.2734	-	-	-	-
various CLIP versions	RN50	27.5	66.9	0.2111	0.2754	0.2149	0.2803	0.2163	0.2820
	ViT-B/32	31.9	73.4	0.2269	0.2955	0.2296	0.2989	0.2324	0.3026
	ViT-B/16	35.0	77.3	0.2352	0.3059	0.2369	0.3081	0.2405	0.3127

Table 3. The effect of vision-language pre-trained (VLP) Models. Text-to-video retrieval performance R@k on the VATEX validation dataset is used to measure VLP models. When the VLP model is used as an embedding extractor, the retrieval performance of the VLP model is positively correlated with the human correlation of EMScore. τ/ρ indicates the Kendall/Spearman correlation, respectively.

D. Speed Analysis

Despite the use of a large pre-trained model (CLIP), our metric is relatively fast. BERTScore uses a pre-trained model (BERT) similar to ours, so it’s fair to compare the speed with it. BERTScore is the most similar to our EMScore(X, X^*), except that they don’t conduct coarse-grained embedding matching. As shown in Tab.2, to complete the VATEX-EVAL dataset which includes 18,000 candidates, using a GeForce RTX-2080Ti GPU, BERTScore and EMScore(X, X^*) take nearly 13.54 and 15.40 seconds to process, respectively. The additional coarse-grained embedding only needs to spend 15.40-13.54=1.86 seconds. For the EMScore(X, V) and EMScore_{ref}(X, V, X^*), which take video as ground truth, we pre-extract the video feature at first. Each video takes nearly 1.89 seconds, in which video frames extraction costs about 1.61 seconds and embedding extraction costs about 0.28 seconds. EMScore(X, V) and EMScore_{ref}(X, V, X^*) are able to process 1,169 and 741 candidates/second, respectively. Compared to the time costs of video captioning development stages, the time cost of EMScore is relatively less. Therefore, EMScore is very suitable for use during validation and testing.

E. The Effect of Pre-training Models

To analyze the impact of the vision-language pre-trained (VLP) models on our EMScore, we use various CLIP versions and two other VLP models as our metric embedding extractors. The different CLIP versions are mainly determined by the various visual encoder network, such as RN50 (ResNet-50), ViT-B/16 (the base Vision Transformer variant with 16×16 input patch size), ViT-B/32 (the base Vision Transformer variant with 32×32 input patch size). We additionally use two other VLP models, such as MIL-NCE [4] which propose a multiple instance learning approach and pre-trained on the HowTo100M [5] dataset, and Frozen-in-Time [2] which propose an end-to-end trainable model that is designed to take advantage of both large-scale image dataset Google Conceptual Captions [8] and video captioning dataset WebVid-2M. To quantitatively analyze the performance of the VLP models, we check their text-to-

video retrieval performance on the validation set of the VATEX captioning dataset. As shown in Tab.3, we use Recall at K (R@K and K=1, 10) to represent retrieval performance. Because MIL-NCE and Frozen-in-Time are pre-trained on video-level contrastive learning, we only use these two models for coarse-grained EMScore. To measure human correlation, we compute Kendall correlation τ and Spearman rank correlation ρ .

We find that when the VLP model is used as an embedding extractor, the retrieval performance of the VLP model is positively correlated with the human correlation of EMScore. The reason is that a better VLP model which has better retrieval performance means it can project vision and language input into a shared space and get better generic cross-modal representations. Furthermore, our EMScore based on embedding matching can effectively leverage obtained representations.

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. 1
- [2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1708–1718. IEEE, 2021. 4
- [3] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pages 6546–6555. IEEE Computer Society, 2018. 1
- [4] 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. 4
- [5] 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. 4
- [6] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 1
- [7] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 1
- [8] 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. 4
- [9] X Yu Stella and Jianbo Shi. Multiclass spectral clustering. In *ICCV*, pages 313–319, 2003. 1
- [10] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In Srinivas Aravamudan, Ilya Sutskever, and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017*, pages 4278–4284. AAAI Press, 2017. 1
- [11] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE, 2019. 1
- [12] 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. 1