COLLABORATIVE THREE-STREAM TRANSFORMERS FOR VIDEO CAPTIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

As the most critical components in a sentence, subject, predicate, and object require special attention in the video captioning task. In this paper, we design the collaborative three-stream transformers to model the interactions of objects, and the actions/relations of objects between different modalities. Specifically, it is formed by three branches of transformers used to exploit the visual-linguistic interactions of different granularities in spatio-temporal domain between videos and text, detected objects and text, and actions and text. Meanwhile, we design a cross-modality attention module to align the interactions modeled by the three branches of transformers. That is, an affinity matrix is computed to help align visual modalities by injecting the information from other interactions. In this way, the three branches of transformers can support each other to exploit the most discriminative semantic information in different modalities for accurate predictions of captions, especially for the subject, predicate, and object parts in a sentence. The whole model is trained in an end-to-end fashion. Extensive experiments conducted on two large-scale challenging datasets, *i.e.*, YouCookII and ActivityNet Captions, demonstrate that the proposed method performs favorably against the state-of-the-art methods.

1 INTRODUCTION

Video captioning aims to generate natural language descriptions of video content, which attracts much attention in recent years along with the rapidly increasing amount of videos recorded in daily life. It helps blind or D/deaf people are able to enjoy the videos. However, as noted in (Xiong et al., 2018; Park et al., 2019; Lei et al., 2020), it is very challenging to generate natural paragraph descriptions due to the difficulties of having relevant, less redundant, and semantic coherent sentences.

Recently, researchers attempt to use the transformer model to solve the video captioning task (Vaswani et al., 2017; Dai et al., 2019; Iashin & Rahtu, 2020; Zhu & Yang, 2020; Tang et al., 2021), which relies on the self-attention mechanism to describe the interactions between different modalities of the input data, such as video, audio, and text. In practice, the aforementioned methods generally concatenate the features extracted from individual modalities, or use self-attention to model the interactions between extracted features. Although they advance the state-of-the-art of video captioning, it is still far from satisfactory in real applications due to the domain gap between different modalities. Then, a question arises, "*How do we fill in the domain gap and capture the interactions among visual and linguistic modalities for video captioning?*"

Before answering this question, let us see the basic grammar rules at first. Generally, a sentence (Krishna et al., 2017a; Zhou et al., 2018a) is presented as the following form, *i.e.*,

Women wear Arabian skirts on a stage.

Notably, Subject, Object, and Predicate are the three most critical elements in a sentence. They indicates the objects, the actions of objects, and the interactions among different objects. To improve the accuracy of the Subject, Object, and Predicate predictions in a sentence, we propose a new COllaborative three-Stream Transformers (COST) to model the visual-linguistic interactions of different granularities in spatio-temporal domain between different modalities. Specifically, the COST model is formed by three branches of transformers, including the Video-Text, Detection-Text, and

Action-Text transformers. The Video-Text transformer is used to model the interactions between the global video appearances and linguistic texts. The Detection-Text is used to model the interactions between the objects in individual video frames, which enforces the model to focus on the objects being aligned in the visual and linguistic modalities, *i.e.*, indicating the Subjects and Objects in caption sentences. The Action-Text transformer is designed to model the actions/relations of objects between the visual and linguistic modalities, *i.e.*, indicating the Predicate in caption sentences. Meanwhile, to align the interactions modeled by the three branches of transformers, we introduce a cross-modality attention model. In particular, an affinity matrix is computed to represent the relevance among visual modalities and help inject the information from other interactions. In this way, different branches of transformers support each other to exploit the most discriminative semantic information in different modalities, and enforce the model to pay more attention on generating the accurate Subject, Object and Predicate predictions. The whole model is trained in an end-to-end fashion using the Adam algorithm (Kingma & Ba, 2015).

Several experiments are conducted on two publicly challenging datasets, *i.e.*, YouCookII (Zhou et al., 2018a) and ActivityNet Captions (Krishna et al., 2017a), to demonstrate the superior performance of the proposed method compared to the state-of-the-art methods (Zhou et al., 2018b; Dai et al., 2019; Park et al., 2019; Lei et al., 2020). Specifically, our COST method achieves the top CIDEr scores, *i.e.*, 60.78% and 29.64%, on the YouCookII *val* set and the ActivityNet *ae-test* set, improving 3.54% and 1.45% compared to the state-of-the-arts.

The main contributions of this paper are summarized as follows. (1) We develop the new collaborative three-stream transformers to learn the interactions between the visual-linguistic modalities of different granularities in spatio-temporal domain, which enforces the model to generate the accurate Subject, Object, and Predicate predictions. (2) To align the interactions described in the three branches of transformers, we design a cross-modality attention model. (3) Extensive experiments conducted on two challenging datasets show that our method performs favorably against the state-of-the-art methods.

2 Related Works

Network architecture. Video captioning has received much attention in recent years. Most of previous video captioning methods (Yu et al., 2016; Pan et al., 2017; Zhang et al., 2018) attempt to perform sequence-to-sequence learning using the encoder-decoder paradigm (Chen et al., 2019). The convolutional neural networks (CNNs) (Zheng et al., 2020) and long short term memory networks (LSTM) (Pei et al., 2019; Park et al., 2019) are adopted to learn discriminative feature embeddings for accurate predictions. Recently, the transformer model dominates this field due to its superior performance compared with CNNs and LSTM. Zhou et al. (2018b) propose an end-to-end trained transformer model, where the encoder is designed to encode the video into semantic representations, and the proposal decoder is used to decode from the encoding with different anchors to form video event proposals. Sun et al. (2019) design the VideoBERT model to learn bidirectional joint distributions over sequences of visual and linguistic tokens. Similarly, Lu et al. (2019) extend the BERT architecture (Devlin et al., 2019) to a multi-modal two-stream model, which processes both visual and textual inputs separately but interactions are conducted between streams. Lei et al. (2020) develop the memory-augmented recurrent transformer, which uses a highly summarized memory state from the video clips and the sentence history to facilitate better prediction of the next sentence. Different from the aforementioned methods, our method attempts to exploit the visual-linguistic interactions of different granularities in spatio-temporal domain across different modalities, with a focus on the most critical components in the sentence, *i.e.*, subject, predicate and object.

Multi-modal cross-attention mechanism. The interactions between different modalities are critical for the video captioning task. Recent transformer based methods (Iashin & Rahtu, 2020; Zhu & Yang, 2020; Tang et al., 2021) use the cross-attention module to learn correlations across different modalities. For example, Iashin & Rahtu (2020) concatenate the learned embeddings from multiple modalities, *e.g.*, video, audio and speech, for event description. Zhu & Yang (2020) propose the tangled transformer block to encode three sources of information, *i.e.*, global actions, local regional objects, and linguistic descriptions. Global-local correspondences are discovered by exploiting the contextual information. Tang et al. (2021) use frame-level dense captions as an auxiliary text input for better video and language associations, where the constrained attention loss is used to forces the model to automatically focus on the best matched caption from a pool of misalignment caption candidates. In contrast, we design the cross-modality attention module integrated in the collaborative three-stream transformers to align three types of visual-linguistic interactions, leading to more discriminative semantic cues for better caption generation.

Multi-modal pre-training models. Large-scale pre-training is another effective way to improve the accuracy of captioning models. Specifically, the jointly trained video and language models (Huang et al., 2020; Luo et al., 2020; Ging et al., 2020) on the large-scale datasets, such as YouTube-8M (Abu-El-Haija et al., 2016) and HowTo100M (Miech et al., 2019) with automatic speech recognition[1] transcripts, provide discriminative features for downstream tasks. Huang et al. (2020) construct a dense video captioning dataset, *i.e.*, Video Timeline Tags (ViTT), and explore several multi-modal sequence-to-sequence pre-raining strategies using transformers (Vaswani et al., 2017). Luo et al. (2020) also use transformers (Vaswani et al., 2017) with two single-modal encoders, a cross encoder, and a decoder. Recently, Ging et al. (2020) develop the Cooperative hierarchical Transformer (COOT) to model the interactions between different levels of granularity and modalities, which achieves superior results for video captioning.

3 Our Approach

As discussed above, we design the collaborative three-stream transformers to model the interactions of objects, and actions/relations of objects between different modalities, which is formed by three branches of transformers, *i.e.*, Video-Text, Detection-Text, and Action-Text transformers. Specifically, the video and text inputs are firstly encoded to extract the multi-modal feature embeddings. After that, the embeddings are fed into the three-stream transformers to exploit the visual-linguistic interactions between videos and text of different granularities in spatio-temporal domain, *i.e.*, global videos, detections, and actions. Meanwhile, the cross-modality attention module is designed to align the interactions modeled by the three branches of transformers. The overall architecture of the proposed method is shown in Figure 1.

3.1 Multi-Modality Tokens

Three kinds of tokens, *i.e.*, visual tokens, linguistic tokens, and special tokens, are used to encode the video and text inputs, which are described as follows. **Visual tokens.** For the visual tokens, we use three kinds of tokens with different granularities in spatio-temporal domain, that is the video tokens, the detection tokens, and the action tokens.

- *Video tokens* provide the global semantic information in the video sequence. In contrast to (Lei et al., 2020), we only use the appearance features extracted by Temporal Segment Networks (TSN) (Wang et al., 2016) (denoted as TSN-APP is Figure 1) as the video tokens, *i.e.*, $\{f_1^v, f_2^v, \cdots, f_{N_v}^v\}$, where f_i^v is the extracted feature of the i-th video clip, and N_v is the number of video clips. Notably, we can also leverage more powerful multi-modal feature extraction method COOT (Ging et al., 2020) to improve the performance, which is pre-trained on the large-scale HowTo100M dataset (Miech et al., 2019).
- *Detection tokens* are used to enforce the model to focus on the Subjects or Objects in caption sentences. Similar to (Park et al., 2019; Lu et al., 2019; Zhu & Yang, 2020), we use the Faster R-CNN method to detect the objects in each frame, which is pre-trained on the Visual Genome dataset (Krishna et al., 2017b). After that, the detection features in Faster R-CNN corresponding to the objects with the highest confidence scores in K categories[2] are used to generate the detection tokens for each frame. We use $\{f_1^d, f_2^d, \cdots, f_{N_d}^d\}$ to denote the set of detection tokens, where f_i^d is the i-th detection feature, and N_d is the total number of detections in the video sequence.
- *Action tokens* are designed to enforce the model to concentrate on the Predicates in caption sentences. Following (Lei et al., 2020), the optical flow features of video sequences are extracted by TSN (Wang et al., 2016) (denoted as TSN-MOT is Figure 1) to generate the

[1] `https://developers.google.com/youtube/v3/docs/captions`

[2] If the category number of the detected objects is less than K in a frame, we select the K detected objects with the highest confidence scores regardless the object categories to generate the detection tokens.

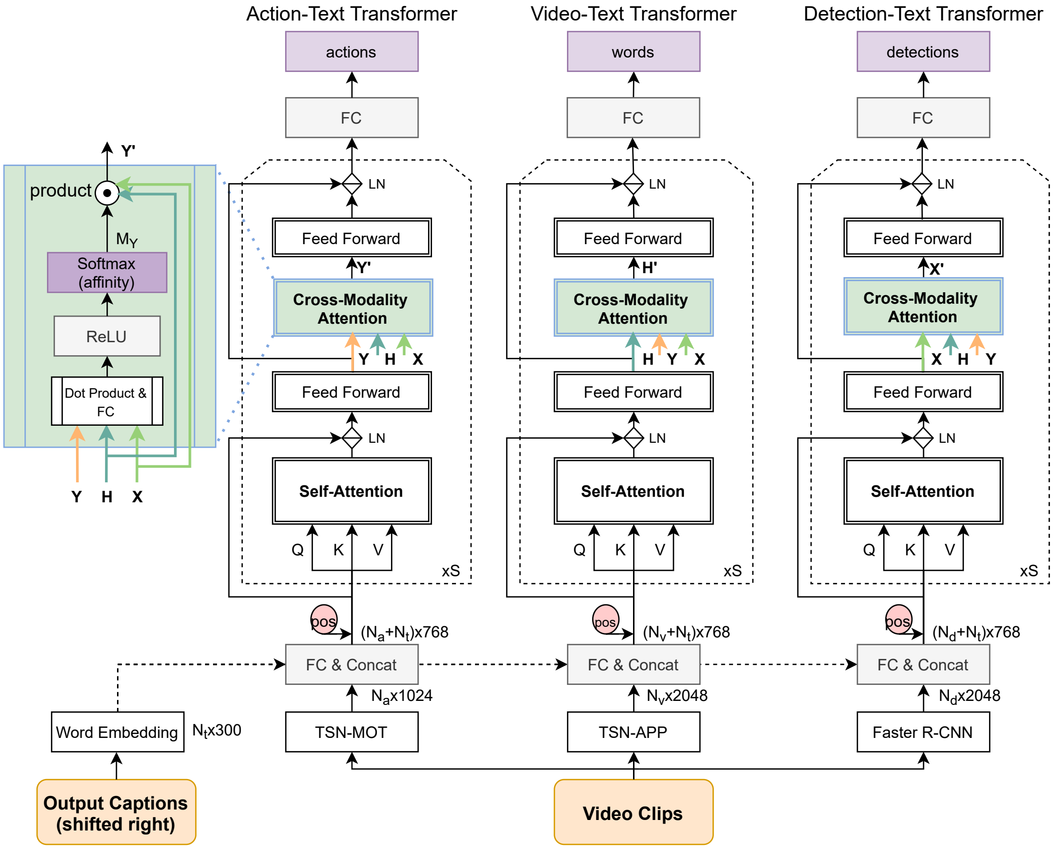


Figure 1: The network architecture of the proposed COST method, which is formed by three branches of transformers, *i.e.*, the Video-Text, Detection-Text and Action-Text transformers. LN denotes the layer normalization. The cross-modality attention module is designed to align the interactions described by different branches of transformers.

action tokens, which are used to describe the actions/relations of objects. The action tokens are denoted as $\{f_1^a, f_2^a, \cdots, f_{N_a}^a\}$, where f_i^a is the motion feature of the i-th video clip, and N_a is the total number of action tokens.

Linguistic tokens. We break down the captions of video sequences into individual words and compute the corresponding linguistic tokens using the GloVe model (Pennington et al., 2014). The linguistic tokens are denoted as $\{f_1^t, f_2^t, \cdots, f_{N_t}^t\}$, where f_i^t is the extracted features of the i-th word using the GloVe model, and N_t is the total number of words.

Special tokens. Besides the aforementioned tokens, we also introduce two kinds of special tokens in transformer, similar to BERT (Devlin et al., 2019). The first one is the modality type token [CLS], which is added at the beginning of visual features to denote which modality the following tokens come from. The second one is the three kinds of separation token, *i.e.*, [SEP], [BOS], and [EOS]. [SEP] is used at the end of the visual tokens to separate them from the linguistic tokens, [BOS] is used to denote the beginning of linguistic tokens, and [EOS] is used to denote the ending of the linguistic tokens, respectively. In addition, we use a fully-connected layer to encode the aforementioned tokens in the same dimension. Thus, the inputs for the three-stream transformer are computed as

$$\big\{[\text{CLS}^{(\cdot)}], f_1^{(\cdot)}, f_2^{(\cdot)}, \cdots, f_{N_{(\cdot)}}^{(\cdot)}, [\text{SEP}], [\text{BOS}], f_1^t, \cdots, f_{N_t}^t, [\text{EOS}]\big\}, \tag{1}$$

where $(\cdot) \in \{v, d, a\}$ indicates the video, detection, and action tokens, respectively. We also use the positional encoding strategy (Vaswani et al., 2017) in the Video-Text, Detection-Text, and Action-Text transformers to describe the order information of caption sentences.

3.2 Three-Stream Transformers

As shown in Figure 1, we feed the aforementioned tokens into the three-stream transformers. The Video-Text, Detection-Text, and Action-Text branches are formed by S basic blocks, and each block consists of a self-attention module and a cross-modality attention module. Both the self-attention and cross-modality modules are followed by a feed forward layer.

Self-attention module. The self-attention module is designed to model the visual-linguistic alignments in different branches of transformers, *i.e.*, Video-Text, Detection-Text, and Action-Text. Following (Vaswani et al., 2017), we compute the attention function between different tokens as follows.

$$\mathcal{A}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}})\boldsymbol{V}, \tag{2}$$

where $\boldsymbol{Q} \in \mathbb{R}^{N \times d}$ is the query matrix, $\boldsymbol{K} \in \mathbb{R}^{N \times d}$ is the key matrix, $\boldsymbol{V} \in \mathbb{R}^{N \times d}$ is the value matrix, and N and d are the number of tokens and the dimension of embeddings, respectively. We advocate h paralleled heads of scaled dot-product attentions to increase the diversity.

Cross-modality attention module. Besides the self-attention module, we use the cross-modality attention module to align the interactions modeled by the three branches of transformers. Specifically, we compute the affinity matrix to guide the alignments between different interactions by injecting the information from other branches of transformers. Given the feature embeddings $\mathcal{H}$, $\mathcal{X}$ and $\mathcal{Y}$ from the Video-Text, Detection-Text, and Action-Text transformers, we first calculate the corresponding affinity matrices $\mathcal{M}_{\mathcal{H}}$, $\mathcal{M}_{\mathcal{X}}$, and $\mathcal{M}_{\mathcal{Y}}$ using the dot-product operation followed by one FC layer and ReLU activation. Here, as shown in Figure 1, we take the affinity matrix $\mathcal{M}_{\mathcal{Y}} \in \mathbb{R}^{N_a \times (N_v + N_d)}$ as an example, which is computed as

$$\mathcal{M}_{\mathcal{Y}} = \mathrm{softmax}\big(\mathrm{ReLU}\big(\mathrm{FC}_{\mathcal{Y}}\big(\mathcal{Y} \odot \big(\oplus(\mathcal{H}, \mathcal{X})^{\mathrm{T}}\big)\big)\big)\big), \tag{3}$$

where $\odot$ indicates the dot product, and $\oplus(\cdot, \cdot)$ denotes the concatenation operation. We estimate the modality-wise normalization scores of the interactions using a softmax layer. $\mathcal{M}_{\mathcal{Y}}(i, j)$ denotes the normalized interaction score between the i-th entity in the Action-Text embeddings and the j-th entity in the Video-Text and Detection-Text embeddings. Based on the matrix, the feature embeddings of the Action-Text transformer can inject information from other branches of transformers, *i.e.*,

$$\mathcal{Y}' = \mathrm{FFN}\big(\mathcal{M}_{\mathcal{Y}} \odot \big(\oplus(\mathcal{H}, \mathcal{X})\big)\big), \tag{4}$$

where $\odot$ indicates the dot product, and FFN denotes the feed forward layer. Notably, we apply the cross-modality attention in all blocks for each branch of transformers. In this way, we can align the captions with the video, detection and action entities to enhance the discriminative representation for video captioning. It is noteworthy that we leverage the video-text features in history to obtain the long-term sentence-level recurrence to generate the next sentences (Lei et al., 2020).

3.3 Optimization

We use the multi-task loss to guide the training of our COST method, which is formed by three terms, *i.e.*, $\mathcal{L}_v(\cdot, \cdot)$ for the Video-Text transformer, $\mathcal{L}_d(\cdot, \cdot)$ for the Detection-Text transformer, and $\mathcal{L}_a(\cdot, \cdot)$ for the Action-Text transformer, *i.e.*,

$$\mathcal{L} = \mathcal{L}_v(\ell_v, [\mathrm{CLS}^v]) + \lambda_d \cdot \mathcal{L}_d(\ell_d, [\mathrm{CLS}^d]) + \lambda_a \cdot \mathcal{L}_a(\ell_a, [\mathrm{CLS}^a]), \tag{5}$$

where λ_d and λ_a are the preset parameters used to balance those three terms. $\mathcal{L}_v(\cdot, \cdot)$ is the cross-entropy loss used to penalize the errors of the predicted captions comparing to the ground-truth descriptions. The ground-truth of video tokens ℓ_v are the indexes of words in caption sentences. $\mathcal{L}_d(\cdot, \cdot)$ is also the cross-entropy loss used to penalize the errors of the predicted categories of objects comparing to the pseudo category labels generated by the Faster R-CNN detector[3]. $\mathcal{L}_a(\cdot, \cdot)$ is the multi-label classification loss used to handle multiple actions appearing in one video sequence. Specifically, we first aggregate all action tokens as the modality type token $[\mathrm{CLS}^a]$ and then compute

[3] Notably, we do not use the annotated objects for model training, but use the pseudo category label ℓ_d generated by the Faster R-CNN detector as the ground-truth label to enforce the network to maintain the original encoded semantic information of detector.

Table 1: Experimental results on the YouCookII *val* subset and ActivityNet Captions *ae-test* subset in the **paragraph-level** evaluation mode. *COOT* indicates that the evaluated methods use the multi-modal feature extraction model (Ging et al., 2020) pre-trained on HowTo100M (Miech et al., 2019).

Method	*COOT*	YouCookII (*val*)				ActivityNet Captions (*ae-test*)			
		B@4	M	C	R@4 ↓	B@4	M	C	R@4 ↓
Vanilla Transformer (Zhou et al., 2018b)	✗	7.62	15.65	32.26	7.83	9.31	15.54	21.33	7.45
Transformer-XL (Dai et al., 2019)	✗	6.56	14.76	26.35	6.30	10.25	14.91	21.71	8.79
Transformer-XLRG (Lei et al., 2020)	✗	6.63	14.74	25.93	6.03	10.07	14.58	20.34	9.37
MART (Lei et al., 2020)	✗	8.00	15.90	35.74	4.39	9.78	15.57	22.16	**5.44**
COST	✗	**9.47**	**17.67**	**45.54**	**4.04**	**11.14**	**15.91**	**24.77**	5.86
Vanilla Transformer (Zhou et al., 2018b)	✓	11.05	19.79	55.57	**5.69**	10.47	15.76	25.90	19.14
Transformer-XL (Dai et al., 2019)	✓	-	-	-	-	10.57	14.76	22.04	15.85
MART (Lei et al., 2020)	✓	11.30	**19.85**	57.24	6.69	10.85	**15.99**	28.19	6.64
COST	✓	**11.56**	19.67	**60.78**	6.63	**11.88**	15.70	**29.64**	**6.11**

the confidence score using a fully-connected layer, *i.e.*, $\text{FC}([\text{CLS}^a])$. Following (Zhang et al., 2021; Sun et al., 2020), the loss function is computed as

$$\mathcal{L}_a(\ell_a, \text{CLS}^a) = \log\big(1 + \sum_{i \in \Omega_{\text{pos}}(\ell_a)} e^{-s_i}\big) + \log\big(1 + \sum_{j \in \Omega_{\text{neg}}(\ell_a)} e^{s_j}\big), \tag{6}$$

which penalizes the confidence scores s_i of the detected actions $\Omega_{\text{pos}}(\ell_a)$ in video sequences are less than the threshold 0 while the confidence scores s_j of undetected actions $\Omega_{\text{neg}}(\ell_a)$ are larger than 0. The ground-truth action labels ℓ_a are the most common verbs in caption sentences, which are retrieved by the off-the-shelf part-of-speech tagger method (Sun et al., 2019).

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. We conducted several experiments on two challenging datasets, *i.e.*, YouCookII (Zhou et al., 2018a) and ActivityNet Captions (Krishna et al., 2017a). YouCookII includes $2,000$ long untrimmed videos describing 89 cooking recipes, where each video contains one reference paragraph and is further split into several event segments with annotated sentences. $1,333$ and 457 video sequences are used for training and validation, respectively. Meanwhile, ActivityNet Captions is a large-scale dataset formed by $10,009$ videos for training and $4,917$ videos for validation and testing. Notably, following (Zhou et al., 2019), the original validation set is split into the *ae-val* subset with $2,460$ videos for validation and the *ae-test* subset with $2,457$ videos for testing.

Evaluation metrics. Similar to (Park et al., 2019; Lei et al., 2020; Zhu & Yang, 2020), we use several standard metrics to evaluate our method, including BLEU@n (B@n) (Papineni et al., 2002) for n-gram precision, METEOR (M) (Denkowski & Lavie, 2014) for n-gram with synonym matching, CIDEr (C) (Vedantam et al., 2015) for tf-idf weighted n-gram similarity, and ROUGE (R@4) (Xiong et al., 2018; Park et al., 2019) for n-gram recall. Notably, two evaluation modes are considered, *i.e.*, micro-level and paragraph-level. The micro-level evaluation reports the average score on all video sequences; while the paragraph-level evaluation first concatenates the caption sentences of all video sequences and then computes the scores averaged across all videos based on the ground-truth paragraph caption sentences. In both modes, CIDEr is used as the primary metric for ranking.

4.2 Implementation Details

Our COST algorithm is implemented using PyTorch. The source code will be released after acceptance. All the experiments are conducted on a machine with 2 NVIDIA RTX-3090 GPUs. We train the model using the strategies similar to BERT (Devlin et al., 2019). Specifically, we use Adam (Kingma & Ba, 2015) with an initial learning rate of $1e-4$, β_1=0.9, β_2=0.999, L^2 weight decay of 0.01, and the learning rate warmup over the first 2 epochs. We use the early-stop strategy in the model training phase. That is, we train the model at most 20 epochs with the batch size 64. For each branch of transformers, we set the dimension of the feature embeddings $d = 768$, the number

Table 2: Comparison with the state-of-the-art methods on ActivityNet Captions *ae-val* subset in the **paragraph-level** evaluation mode. *Det.* and *Re.* indicate whether the model uses detection features and the sentence-level recurrence.

	Det.	*Re.*	B@4	M	C	R@4 ↓
LSTM based methods:						
MFT (Xiong et al., 2018)	✗	✓	10.29	14.73	19.12	17.71
HSE (Zhang et al., 2018)	✗	✓	9.84	13.78	18.78	13.22
LSTM based methods with detection feature:						
GVD (Zhou et al., 2019)	✓	✗	11.04	15.71	21.95	8.76
GVDsup (Zhou et al., 2019)	✓	✗	**11.30**	16.41	22.94	7.04
AdvInf (Park et al., 2019)	✓	✓	10.04	**16.60**	20.97	5.76
Transformer based methods:						
Vanilla Transformer (Zhou et al., 2018b)	✗	✗	9.75	15.64	22.16	7.79
Transformer-XL (Dai et al., 2019)	✗	✓	10.39	15.09	21.67	8.54
Transformer-XLRG (Lei et al., 2020)	✗	✓	10.17	14.77	20.40	8.85
MART (Lei et al., 2020)	✗	✓	10.33	15.68	23.42	**5.18**
COST	✓	✓	11.22	16.58	**25.70**	7.09

Table 3: Comparison with the state-of-the-arts on YouCookII *val* subset in the **micro-level** evaluation mode.

Method	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Masked Trans. (Zhou et al., 2018b)	7.53	3.84	11.55	27.44	0.38
S3D (Xie et al., 2017)	6.12	3.24	9.52	26.09	0.31
VideoBERT (Sun et al., 2019)	6.80	4.04	11.01	27.50	0.49
VideoBERT+S3D (Sun et al., 2019)	7.59	4.33	11.94	28.80	0.50
ActBERT (Zhu & Yang, 2020)	8.66	5.41	14.30	30.56	0.65
COST	**10.69**	**6.63**	**12.61**	**31.09**	**0.71**

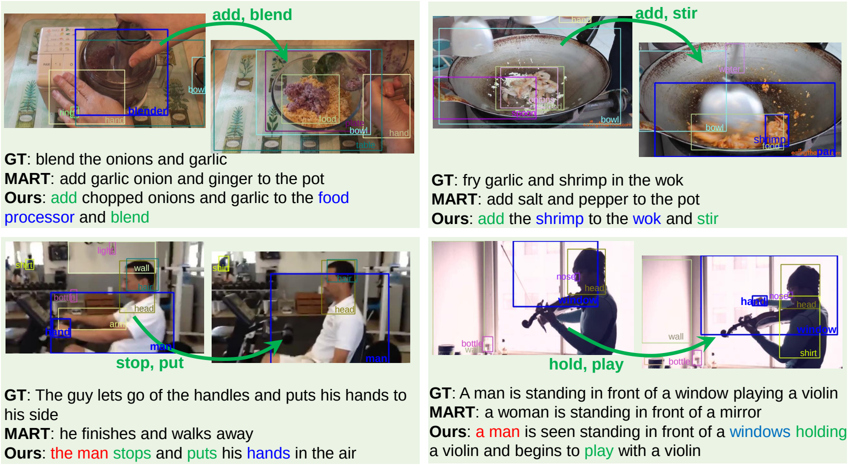
of transformer blocks $S = 2$, and the number of attention heads $h = 12$. The loss weights λ_a and λ_d in equation 5 are set to 2.0 and 0.02 empirically.

Considering the trade-off between accuracy and complexity, we extract 2048-dim video features, top $K = 5$ 2048-dim detection features and 1024-dim action features from at most 100 frames uniformly sampled from the video sequences, *i.e.*, $N_v = N_a = 102, N_d = 502$ with the two special tokens [CLS] and [SEP]. Meanwhile, we exploit the first 20 words in the caption sentences and compute the 300-dim GloVe features, *i.e.*, $N_t = 22$ with the two special tokens [BOS] and [EOS]. For the COOT features, we concatenate the local clip-level (384-dim) and the global video-level (768-dim) features to describe the videos. After the fully-connected layer, all the tokens are converted into the 768-dim features.

4.3 Evaluation Results

We compare the proposed COST method with the state-of-the-art methods on the two challenging datasets, *i.e.*, YouCookII and ActivityNet Captions, in Table 1. As shown in Table 1, our method achieves the best results on the YouCookII *val* subset and the ActivityNet Captions *ae-test* subset. Notably, without using the COOT features, our method improves near 10% CIDEr score compared to the second best method, *i.e.*, MART (Lei et al., 2020) on the YouCookII *val* subset. This is attributed to the proposed cross-modality attention module across the collaborative three-stream transformers. Besides, our method exploits the local appearance information from the Detection-Text transformer for better accuracy. Using the COOT features, the overall video captioning results are significantly improved, and our COST method also performs favorably against other algorithms by improving over 3% CIDEr score. We observe that the similar trend appears in the ActivityNet Captions *ae-test* subset. Although the ActivityNet Captions dataset is more challenging than the YouCookII dataset, our method improves the performance considerably compared to the state-of-the-art methods.

As presented in Table 2, we also compare the proposed method to the LSTM based methods with input detections on the ActivityNet Captions *ae-val* subset. Compared to AdvInf (Park et al., 2019), our method produces higher scores for both B@4 and CIDEr, demonstrating the superiority of the collaborative transformers to learn multi-model representations over LSTM. Meanwhile, MART (Lei et al., 2020) without input detections performs inferior than our method in terms of B@4, M



GT: blend the onions and garlic
MART: add garlic onion and ginger to the pot
Ours: add chopped onions and garlic to the food processor and blend

GT: fry garlic and shrimp in the wok
MART: add salt and pepper to the pot
Ours: add the shrimp to the wok and stir

GT: The guy lets go of the handles and puts his hands to his side
MART: he finishes and walks away
Ours: the man stops and puts his hands in the air

GT: A man is standing in front of a window playing a violin
MART: a woman is standing in front of a mirror
Ours: a man is seen standing in front of a windows holding a violin and begins to play with a violin

Figure 2: Qualitative results of the state-of-the-art MART (Lei et al., 2020) method and our COST method. The Subject, Predicate, and Object in a sentence are highlighted in the red, green and blue fonts, respectively.

Table 4: Ablation study on YouCookII *val* subset.

COST Variants	YouCookII (*val*)			
	B@4	M	C	R@4 ↓
COST-1 (v)	6.59	14.29	29.20	6.49
COST-1 (v+a)	7.72	15.45	33.98	4.60
COST-1 (v+a+d)	8.73	16.90	38.63	4.66
COST-2 (v+a)	7.73	15.89	34.73	5.74
COST-2 (v+d)	9.04	17.31	43.09	4.59
COST-3 (v+a+d)	**9.47**	**17.67**	**45.54**	**4.04**

and C metrics. It indicates that the Detection-Text transformer in our network can describe the Subjects and Objects in caption sentences more accurately.

According to Table 3, our method outperforms several BERT based methods in terms of the micro-level evaluation. In particular, the second best ActBERT (Zhu & Yang, 2020) relies on the same visual modalities as our method. However, they directly focus on the learning the alignment between text and other visual modalities, which is difficult to exploit the discriminative semantic information. In contrast, our cross-modality attention module learns from three visual-linguistic interactions using the three-stream transformers, producing better CIDEr score. It is worth mentioning that they use BERT (Devlin et al., 2019) and S3D (Xie et al., 2017) to generate more powerful text and video features than that generated by GloVe (Pennington et al., 2014) and TSN (Wang et al., 2016) in our method.

Furthermore, the qualitative results of our COST method and MART (Lei et al., 2020) are shown in Figure 2. It can be seen that our COST generates more accurate captions than MART (Lei et al., 2020). This is attributed to two reasons. First, using the Detection-Text transformer, the Objects in caption sentences can be learned explicitly (*e.g.*, *shrimp*, *hand*, and *windows*) or implicitly (*e.g.*, from *bowl* to *food processor*, and from *pan* to *wok*). Second, the Action-Text transformer in our method can extract the key verbs in caption sentences such as *add*, *blend*, *put* and *hold*. MART (Lei et al., 2020) fails to recognize these key Subjects, Objects and Verbs. The results indicate that these two branches of transformers can enforce the network learn the key elements in the caption sentences.

4.4 Ablation Study

To study the influence of different components in the proposed method, we conduct the ablation study on the YouCookII *val* subset, shown in Table 4. Notably, we use the appearance features extracted by TSN (Wang et al., 2016) in all COST-k variants, where k denotes the number of transformer branches retained in our COST method.

Effectiveness of multi-modal features. To verify the effectiveness of the multi-modal features, we construct three COST-1 variants, *i.e.*, COST-1 (v), COST-1 (v+a) and COST-1 (v+a+d). In particular,

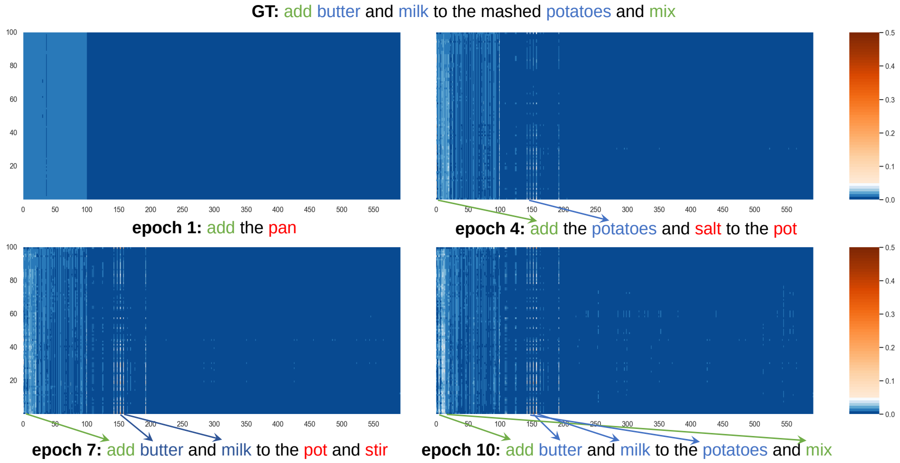


Figure 3: Heatmap used to indicate the affinity matrix $\mathcal{M}_{\mathcal{H}}$ in the Video-Text transformer, where the row denotes the video tokens and the column denotes the action and detection tokens. The false predictions of nouns and verbs are denoted in red font. For clarity, we only show a few epochs in the training phase.

we only use the Video-Text transformer, but change the input features as the combinations of the GloVe text features and the concatenated features from video (v), action (a) and detection (d). As shown in Table 4, the scores under all metrics are improved considerably by integrating the action or detection features. Moreover, the CIDEr score is boosted from 29.20% to 38.63%, if we include all the three-modal features. It indicates that multi-modal features definitely facilitate to generate more accurate video captions.

Effectiveness of three-stream transformers. To demonstrate the effectiveness of the three-stream transformers compared to the simple feature concatenation used in COST-1, we construct three COST-k ($k = 2, 3$) variants, shown in Figure 1. It can be seen that the accuracy can be improved by using the three-stream transformers, *i.e.*, the CIDEr score improved from 38.63% to 45.54%. Meanwhile, the accuracy is considerably improved by using more branches of transformers. This is because our cross-modality attention module is able to capture the most relevant semantic information from different visual-linguistic interactions. Compared to the COST-1 variants, our full model can obtain a more discriminative feature representations for video captioning.

Visualization of affinity matrix. To better understand the cross-modality attention module, we use the heapmap to visualize the affinity matrix $\mathcal{M}_{\mathcal{H}} \in \mathbb{R}^{N_v \times (N_d + N_a)}$ of the Video-Text transformer in Figure 3. At epoch 1, all values in the affinity matrix $\mathcal{M}_{\mathcal{H}}$ are similar and the maximal value in $\mathcal{M}_{\mathcal{H}}$ is only 0.001. The corresponding caption results are noisy with the false predictions of nouns and verbs, *e.g.*, *pan* instead of *butter*. It indicates that the affinity matrix is randomly initialized and interactions between different modalities are not aligned. After training for several epochs, a few entities dominate the affinity matrix with the maximal value of 0.3 (see the bright vertical lines in the heatmap). It indicates that our cross-modality attention module can successfully exploit the most relevant entities from other modalities to inject the information from other branches of transformers. In this way, the verb *stir* and the noun *pan* can be corrected to *mix* and *butter* for more accurate caption results.

5 Conclusion

In this paper, we propose the collaborative three-stream transformers to exploit the interactions of objects, and the actions/relations of objects between different modalities of different granularities in spatio-temporal domain. Meanwhile, the cross-modality attention module is designed to align the interactions modeled by the three branches of transformers, which focuses on improving the prediction accuracies of the Subject, Object, and Predicate in the caption sentences. Several experiments conducted on the YouCookII and ActivityNet Captions datasets demonstrate the effectiveness of the proposed method.

REFERENCES

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016.

Shaoxiang Chen, Ting Yao, and Yu-Gang Jiang. Deep learning for video captioning: A review. In *IJCAI*, pp. 6283–6290, 2019.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*, pp. 2978–2988, 2019.

Michael J. Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT@ACL*, pp. 376–380, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186, 2019.

Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. COOT: cooperative hierarchical transformer for video-text representation learning. In *NeurIPS*, 2020.

Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. Multimodal pretraining for dense video captioning. In *AACL/IJCNLP*, pp. 470–490, 2020.

Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *CVPRW*, pp. 4117–4126, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pp. 706–715, 2017a.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017b.

Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L. Berg, and Mohit Bansal. MART: memory-augmented recurrent transformer for coherent video paragraph captioning. In *ACL*, pp. 2603–2614, 2020.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pp. 13–23, 2019.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *CoRR*, abs/2002.06353, 2020.

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 *ICCV*, pp. 2630–2640, 2019.

Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, pp. 984–992, 2017.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318, 2002.

Jae Sung Park, Marcus Rohrbach, Trevor Darrell, and Anna Rohrbach. Adversarial inference for multi-sentence video description. In *CVPR*, pp. 6598–6608, 2019.

Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *CVPR*, pp. 8347–8356, 2019.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pp. 1532–1543, 2014.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pp. 7463–7472, 2019.

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pp. 6397–6406, 2020.

Zineng Tang, Jie Lei, and Mohit Bansal. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *NAACL-HLT*, pp. 2415–2426, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pp. 5998–6008, 2017.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pp. 4566–4575, 2015.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, volume 9912, pp. 20–36, 2016.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. *CoRR*, abs/1712.04851, 2017.

Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *ECCV*, volume 11215, pp. 489–505, 2018.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, pp. 4584–4593, 2016.

Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *ECCV*, volume 11217, pp. 385–401, 2018.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. Document-level relation extraction as semantic segmentation. In *IJCAI*, pp. 3999–4006, 2021.

Qi Zheng, Chaoyue Wang, and Dacheng Tao. Syntax-aware action targeting for video captioning. In *CVPR*, pp. 13093–13102, 2020.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, pp. 7590–7598, 2018a.

Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, pp. 8739–8748, 2018b.

Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, pp. 6578–6587, 2019.

Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, pp. 8743–8752, 2020.