Cross Interaction Network for Natural Language Guided Video Moment Retrieval

Xinli Yu
Temple University
Philadelphia, US
yxinli92@gmail.com

Mohsen Malmir
Amazon.com
Boston, US
malmim@amazon.com

Xin He
Amazon.com
Boston, US
xih@amazon.com

Jiangning Chen
Amazon.com
Boston, US
cjiangni@amazon.com

Tong Wang
Amazon.com
Boston, US
tonwng@amazon.com

Yue Wu
Amazon.com
Boston, US
wuayue@amazon.com

Yue Liu
Amazon.com
Boston, US
liuyue.ly@gmail.com

Yang Liu
Amazon.com
Boston, US
yangliud@amazon.com

ABSTRACT

Natural language query grounding in videos is a challenging task that requires comprehensive understanding of the query, video and fusion of information across these modalities. Existing methods mostly emphasize on the query-to-video one-way interaction with a late fusion scheme, lacking effective ways to capture the relationship within and between query and video in a fine-grained manner. Moreover, current methods are often overly complicated resulting in long training time. We propose a self-attention together with cross interaction multi-head-attention mechanism in an early fusion scheme to capture video-query intra-dependencies as well as inter-relation from both directions (query-to-video and video-to-query). The cross-attention method can associate query words and video frames at any position and account for long-range dependencies in the video context. In addition, we propose a multi-task training objective that includes start/end prediction and moment segmentation. The moment segmentation task provides additional training signals that remedy the start/end prediction noise caused by annotator disagreement. Our simple yet effective architecture enables speedy training (within 1 hour on an AWS P3.2xlarge GPU instance) and instant inference. We showed that the proposed method achieves superior performance compared to complex state of the art methods, in particular surpassing the SOTA on high IoU metrics (R@1, IoU=0.7) by 3.52% absolute (11.09% relative) on the Charades-STA dataset.

KEYWORDS

Information retrieval; Natural language guided; Video moment retrieval; Cross attention; Self attention

ACM Reference Format:
Xinli Yu, Mohsen Malmir, Xin He, Jiangning Chen, Tong Wang, Yue Wu, Yue Liu, and Yang Liu. 2021. Cross Interaction Network for Natural Language Guided Video Moment Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3463021

1 INTRODUCTION

Natural language guided *video moment retrieval* (VMR) refers to the general process of retrieving relevant video contents associated with the text queries. Manual search is a very time consuming process, and with the surge of user-generated video clips, it calls for an automated solution to retrieve video segments given any text descriptions.

However, correctly identifying the most relevant time interval is a challenging cross-modality problem that requires comprehensive understanding of both the video and the text. For example, localizing the query "mixing the ingredients" in the video requires the understanding of what "ingredients" refer to, and being able to resolve the "mixing" action. Moreover, VMR is also a temporal learning problem that needs information aggregation across the temporal dimension. For example, resolving "the first time a cat jumps up" requires not only the understanding of what "cat" and "jump" refer to but also the chronological event order indicated by the word "first".

There are generally two fusing schemes to tackle this cross-modality problem - early and late[4, 7, 10, 11, 16, 18] to transform the video moment and the query to the same space. Late fusion is commonly used [4, 7, 10] where video moment is first transformed to the shared space without the knowledge of the query, and different moments are later ranked by a distance metric. We adapt the early fusion [11, 16, 18] scheme in this paper, where video features are mapped to the shared space guided by query information. This allows more fine-grained video-query integration before matching.

For the fusing mechanism, one idea in the current VMR research is to use iterative message passing [13, 18]. This iterative procedure enriches the representation of each frame by incorporating information across all frames and query words. However, it is computationally expensive and introduces many additional meta parameters for the number of iterative stages. Methods [11, 16, 17] without iterative message passing are still overly complicated with multiple stacked layers.

We propose to fuse with a *cross interaction multi-head attention* (CI-MHA) mechanism inspired by ViL-BERT [9], which is just a single-layer attention module, Moreover, by utilizing CI-MHA,

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463021

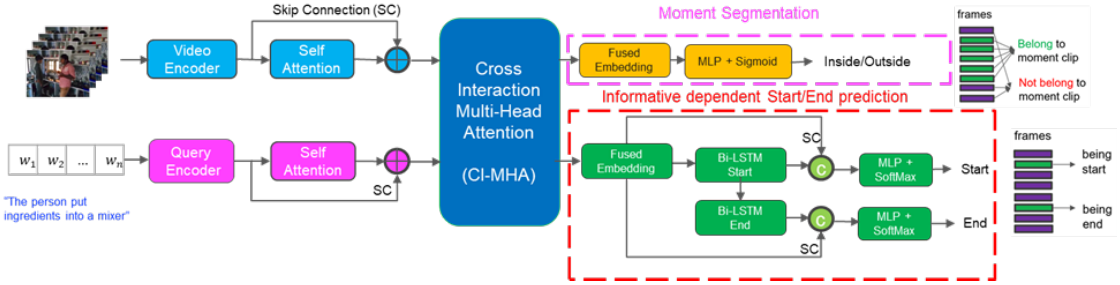

Figure 1: Overall architecture of our model for video moment retrieval.

we derive both the query-enriched video features and the video-enriched query features to capture interactions from both directions. Our approach is much simpler and verified to achieve higher accuracy compared to complex state of arts methods.

For moment localization, we adopt a discriminative approach and further propose a novel multi-task training method to jointly train the *start/end predictions* task with the *moment segmentation* task for better performance. With the discriminative approach, the start/end prediction task directly generates the probability distribution along the temporal dimension for where the moment starts/ends. The moment segmentation task performs binary classification of each frame into the "moment" or the "background". This is more efficient than an the popular ranking-based approach [10, 15, 20] where similarity between each visual and textual input pair needs to be calculated to minimize the pairwise ranking loss and help to save training and inference time.

The motivation for joint moment segmentation task is start/end prediction picks one specific point along the time axis as the positive label, and can be noisy due to annotators' disagreement. The moment segmentation task leverages the continuous intervals with every frame inside as a positive sample, which is more balanced in labels and intuitively provides more temporal dependency information to exploit because the model must learn to predict consecutive positive labels while capturing the boundaries.

In summary, the **key contributions** of this work are two-fold:

- We propose to use the cross interaction multi-head attention mechanism under an early fusion scheme to associate video features with natural language query from both directions. This is a much simpler and computationally-efficient way compared with complex methods including iterative message passing based ones, and it achieves superior results through experiments.
- We propose a multi-task training objective, including 1) start/end prediction task that predicts the target start/end positions, and 2) moment segmentation task that predicts if a frame belongs to the ground-truth time interval. Start/end prediction ground truth consists two specific time points and can be noisy due to annotators' disagreement. The moment segmentation task is aiming to identify a continuous time interval, which has a more balanced positive sample ratio and implies the intra-dependencies of frames within the time interval. Hence, it could provide more robust information for moment localization.

Our simple yet effective architecture enables speedy training completion within 1 hour on an AWS P3.2xlarge GPU instance, and other methods mentioned above are found to take much longer time, for example LGI [11] takes about 3 hours with the same setting. Meanwhile, our method is able to achieve superior SOTA performance on R@1 metrics, surpassing the SOTA on high IoU metrics (R@1, IoU=0.7) by 3.52% absolute (11.09% relative) on the Charades-STA dataset.

2 APPROACH

Given a video and a query sentence, our model maps the video and query features to the same space and then predicts the moment start/end position in the video. At inference time, the most-likely start and end video-query pair is returned as the candidate moment.

2.1 Model Architecture

As illustrated in Figure 1, query and video are each encoded into features by an encoding module. After the encoding step, self-attention is applied to the video feature embeddings and the query embeddings respectively to capture the intra-dependecies, and then the two embeddings are fused by CI-MHA to enrich the video representation with query context and the query representation with video context. The two enriched representations are concatenated and passed to the multi-task training module to predict the target moment clip.

2.2 Video Encoders

C3D network [14] pre-trained on sport1M is used as feature encoder. The untrimmed video is divided into a sequence of 16-frame segments. A 3D CNN module is applied to extract C3D features from each segment. The output of the 3D CNN video feature encoder module is a tensor of size $N \times D$ where $D = 500$ dimensional features and $N = M/16$ where M is the number of frames in the untrimmed video.

I3D network [2] pre-trained on Kinetics is used as feature encoder. The videos are first pre-processed to a frame rate of 24 frames per second. The I3D network takes 64 consecutive frames as input and outputs a snippet-level feature vector.

Positional Encoding A temporal positional embedding is added to the corresponding video segment feature, which provides information about the relative position of each frame and improves accuracy. The positional encoding is formulated as a mapping from

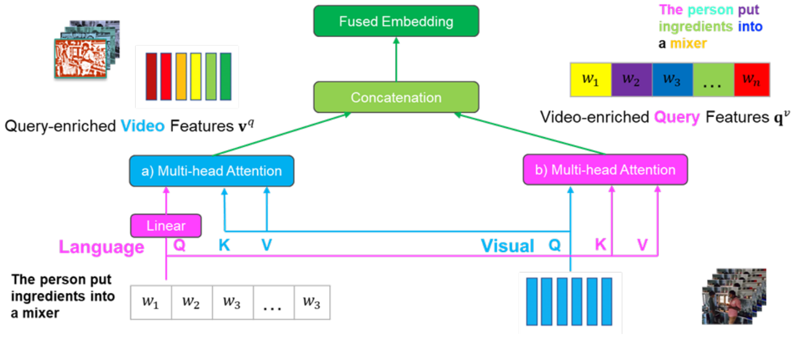
Fused Embedding
The person put ingredients into a mixer
w_1 w_2 w_3 ... w_n
Concatenation
Query-enriched Video Features $\mathbf{v}^q$
Video-enriched Query Features $\mathbf{q}^v$
a) Multi-head Attention
b) Multi-head Attention
Linear
Language
Q
K
V
Visual
Q
K
V
The person put ingredients into a mixer
w_1 w_2 w_3 ... w_3

Figure 2: Cross-attention module of the proposed approach

each position to the correspondingly learned embedding in a trainable embedding matrix as such in BERT [3].

2.3 Video/Query Self Interaction

Let V be the raw video input, and the video encoder embedding output

$$\mathbf{v}_0 = f_{\text{encoder}}(V), \tag{1}$$

which is of size (N, D_v), where N is the video length, D_v is the video embedding size.

Let W be the query input, and the query embedding output

$$\mathbf{q}_0 = f_{\text{encoder}}(W), \tag{2}$$

which is of size (L, D_q), where L is the query length, D_q is the query embedding size.

We use self attention to model intra-dependencies within each modality in a fine-grained manner before applying the CI-MHA module. The video self-attention captures temporal interactions between frames, and the query self-attention captures word-level dependencies. The self interaction allows association between any positions and accounts for both short term and long term temporal and word dependencies. A skip connection via addition is applied afterwards such that the original video/query embedding is also incorporated to avoid information loss. In more details, we apply

$$\mathbf{q} = \text{MHA}(Q = \mathbf{q}_0, K = V = \mathbf{q}_0) + \mathbf{q}_0, \tag{3}$$

$$\mathbf{v} = \text{MHA}(Q = \mathbf{v}_0, K = V = \mathbf{v}_0) + \mathbf{v}_0. \tag{4}$$

2.4 Visual-Language Fusion

We propose to use CI-MHA for Visual-Language fusion. A feed-forward layer is applied to the video encoder embedding $\mathbf{v}$ and the query encoder embedding $\mathbf{q}$, and then positional encoding is added to the video embedding. The resulting video embedding and query embedding are both of size (N, D).

$$\hat{\mathbf{v}} = \text{FeedForward}(\mathbf{v}) + \text{Positional Encoding}(\mathbf{v}), \tag{5}$$

$$\hat{\mathbf{q}} = \text{FeedForward}(\mathbf{q}). \tag{6}$$

For natural language guided VMR, it is intuitive to capture the query-to-video relationship via query enriched video feature embedding (i.e. attention weighted video feature embedding) such as previous work by [19]. Yet, it is also crucial to not overlook the video-to-query association (i.e. attention weighted query embedding). CI-MHA enables us to account for both and ensure more detailed interaction information is apprehended. An illustration is shown in Figure 2. Specifically, we applied

$$\mathbf{v}^q = \text{MHA}(Q = \hat{\mathbf{q}}, K = V = \hat{\mathbf{v}}), \tag{7}$$

$$\mathbf{q}^v = \text{MHA}(Q = \hat{\mathbf{v}}, K = V = \mathbf{q}). \tag{8}$$

These two embeddings are then concatenated along feature dimension as the input for the following prediction task. The fused embeddings of the two modalities $\mathbf{s}_{v,q}$ can be summarized as

$$\mathbf{s}^{v,q} = [\mathbf{v}^q, \mathbf{q}^v]. \tag{9}$$

2.5 Multi-task Training Objectives

We trained the model with two tasks that are complementary to each other and help avoid overfitting, as described below.

Informative Dependent Start/End To predict start/end time point, we adopt an approach similar to ExCL [5]. We use two bi-directional LSTM layers that condition the end prediction's hidden states on the start prediction's hidden states to better factor in temporal dependencies, i.e.,

$$\mathbf{h}_t^{\text{start}} = \text{Bi-LSTM}_{\text{start}}(\mathbf{s}_t^{v,q}, \mathbf{h}_{t-1}^{\text{start}}), \tag{10}$$

$$\mathbf{h}_t^{\text{end}} = \text{Bi-LSTM}_{\text{end}}(\mathbf{h}_{t-1}^{\text{start}}, \mathbf{h}_{t-1}^{\text{end}}), \tag{11}$$

where $\mathbf{h}_0^{\text{start}}, \mathbf{h}_0^{\text{end}}$ are the random initialization for LSTM's first hidden states.

Then Multi-Layer Perceptron (MLP) and softmax layers are applied over the temporal dimension to generate the start/end distributions as the following

$$c_t^{\text{start}} = \text{MLP}_{\text{start}}([\mathbf{h}_t^{\text{start}}, \mathbf{s}_t^{v,q}]), \tag{12}$$

$$c_t^{\text{end}} = \text{MLP}_{\text{end}}([\mathbf{h}_t^{\text{end}} \mathbf{s}_t^{v,q}]). \tag{13}$$

We further add skip connections via concatenation (i.e. $[\mathbf{h}_t^{\text{end}}, \mathbf{s}_t^{v,q}]$) to help prevent possible information loss compared to only using the hidden states $\mathbf{h}_t^{\text{start}}, \mathbf{h}_t^{\text{end}}$ in [5]. Next, we normalize the prediction over the temporal dimension.

$$P_{\text{start}} = \text{SoftMax}(\mathbf{C}^{\text{start}}), \quad P_{\text{end}} = \text{SoftMax}(\mathbf{C}^{\text{end}}), \tag{14}$$

where $\mathbf{C}^{\text{start}} = [c_0^{\text{start}}, c_1^{\text{start}}, ..., c_T^{\text{start}}]$, and same for $\mathbf{C}^{\text{end}}$.

The loss function is negative log-likelihood. The ground-truth labels are two sparse one-hot vectors labeling the ground-truth start/end time point.

Moment Segmentation It is easy for annotators to disagree over the exact point in time as the ground-truth boundaries. This renders the start/end ground truth labels being not only sparse but also noisy. In moment segmentation we let the ground truth cover a continuous time interval and every frame within the ground truth becomes a positive training sample which leads to more balanced prediction. Empirical results show that this task is beneficial to our model performance.

At each time stamp t, we apply a MLP and a sigmoid function to the fused embedding $\mathbf{s}_{v,q}$ and output the probability whether the time stamp t belongs to the target moment or not. The loss function is binary cross entropy loss.

$$P_{in}(t) = \text{Sigmoid}(\text{MLP}_{\text{segment}}(\mathbf{s}_t^{v,q})), \quad (15)$$

$$P_{out}(t) = 1 - P_{in}(t). \quad (16)$$

3 EXPERIMENTS

3.1 Datasets

ActivityNet Captions The ActivityNet Captions dataset [1] connects videos to a series of temporally annotated sentences. On average, each of the 20K videos in ActivityNet contains 3.65 temporally localized sentences, resulting in a total of 100K sentences. **Charades-STA** Charades-STA [12] contains 9848 videos across 157 activities. Each video contains temporal activity annotation and sentence descriptions with start and end time to make them suitable for language-based temporal localization task.

3.2 Implementation Details

We trained the model end-to-end, with raw video frames and natural language query as input. For query encoding, a bi-LSTM is adopted to encode the query GLOVE embedding to 512 dimension, and we use a max query length of 25 words. For video encoding, C3D features has a dimension of 500 and I3D features has a dimension of 1024, and the max video frame length we use is 128 frames. In training, we set the batch size at 100 using Adam optimization with 0.0001 learning rate.

3.3 Comparison with State-of-the-art

We compare our approach on both the ActivityNet Caption and Charades-STA datasets against several recent researches. For example, iterative message passing based methods, including LOGAN[13], MAN[18], as well as other SOTA such as DRN[17], ExCL[5], SCN[8] etc. Following previous work, we adopt "R@1, IoU=θ" as the evaluation metrics.

From **Table 1** and **Table 2**, our method outperforms the competing methods on R@1 with medium to high IOU metrics, which demonstrates the effectiveness of our approach. Note the "MHA" method refers to using uni-direction-fused embedding $\mathbf{v}^q$ produced by equation (7) , while the CI-MHA uses bi-direction-fused embedding $\mathbf{s}^{v,q}$ produced by equation (7), (8) and (9).

We provide example moments predicted by the proposed approach on the test videos in Figure 3. We found that the natural language queries are very diverse and often contain successive temporal actions. The first one shows a simple query where our approach localizes the moment when the person pours the ingredients.

Table 1: Evaluation on Charades-STA

Method	R@1, IoU=0.7	R@1, IoU=0.5	R@1, IoU=0.3
Random	3.03%	8.51%	-
LOGAN [13]	14.54%	34.68%	51.67%
MLVI [16]	15.80%	35.60%	-
ExCL [5]	22.40%	44.10%	-
MAN [18]	22.72%	46.23%	-
DRN [17]	31.75%	53.09%	-
MHA	30.49%	51.04%	60.85%
CI-MHA w/o MS	31.39%	52.41%	62.78%
CI-MHA	**35.27%**	**54.68%**	**69.87%**

Table 2: Evaluation on ActivityNet Captions

Method	R@1, IoU=0.7	R@1, IoU=0.5	R@1, IoU=0.3
MLVI [16]	13.60%	27.70%	45.30%
TripNet [6]	13.93%	32.19%	45.42%
ExCL [5]	23.9%	41.46%	**62.21%**
DRN [17]	23.24%	43.78%	-
MHA	21.76%	40.80%	59.73%
CI-MHA w/o MS	23.41%	41.73%	60.49%
CI-MHA	**25.13%**	**43.97%**	61.49%

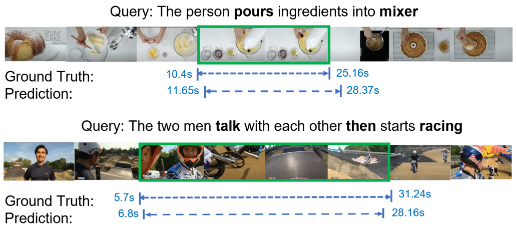
Query: The person pours ingredients into mixer
Ground Truth:
Prediction:
10.4s
25.16s
11.65s
28.37s
Query: The two men talk with each other then starts racing
Ground Truth:
Prediction:
5.7s
31.24s
6.8s
28.16s

Figure 3: Case studies of queries/videos at inference time.

In the second example, which shows a rather complex (temporal) sentence structure requesting two actions, the approach localizes the moment the two people talk with each other until starts racing. The inference examples demonstrate the moment retrieval ability of the proposed approach and its ability for real world applications.

4 CONCLUSIONS

We proposed a cross interaction multi-head attention based fusion architecture with multi-task training objectives to address the video moment retrieval task. The novel cross-attention fusion approach and moment segmentation joint training task allow our model to be simple and effective, as shown on two widely used VMR datasets. It is fast in training with instant inference, which demonstrates clear computationally efficiency strength over complicated iterative message passing based methods while still outperforming on the results. Our model also achieves superior or competitive performance compared with multiple recent works using other more complex structures.

REFERENCES

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 961–970.

[2] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*. 5267–5275.

[5] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. 2019. Excl: Extractive clip localization using natural language descriptions. *arXiv preprint arXiv:1904.02755* (2019).

[6] Meera Hahn, Asim Kadav, James M Rehg, and Hans Peter Graf. 2019. Tripping through time: Efficient localization of activities in videos. *arXiv preprint arXiv:1904.09936* (2019).

[7] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing moments in video with temporal language. *arXiv preprint arXiv:1809.01337* (2018).

[8] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-Supervised Video Moment Retrieval via Semantic Completion Network. *AAAI* (2020).

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

[10] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. 2019. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11592–11601.

[11] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10810–10819.

[12] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*. Springer, 510–526.

[13] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. 2021. LoGAN: Latent Graph Co-Attention Network for Weakly-Supervised Video Moment Retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2083–2092.

[14] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[15] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5005–5013.

[16] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9062–9069.

[17] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10287–10296.

[18] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1247–1257.

[19] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931* (2020).

[20] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. 2017. Dual-path convolutional image-text embedding with instance loss. *arXiv preprint arXiv:1711.05535* (2017).