Cross-Sentence Temporal and Semantic Relations in Video Activity Localisation

Jiabo Huang[1*] jiabo.huang@qmul.ac.uk
Yang Liu[2*] yangliu@pku.edu.cn
Shaogang Gong[1] s.gong@qmul.ac.uk
Hailin Jin[3] hljin@adobe.com

[1]Queen Mary University of London [2]WICT, Peking University [3]Adobe Research

Abstract

Video activity localisation has recently attained increasing attention due to its practical values in automatically localising the most salient visual segments corresponding to their language descriptions (sentences) from untrimmed and unstructured videos. For supervised model training, a temporal annotation of both the start and end time index of each video segment for a sentence (a video moment) must be given. This is not only very expensive but also sensitive to ambiguity and subjective annotation bias, a much harder task than image labelling. In this work, we develop a more accurate weakly-supervised solution by introducing Cross-Sentence Relations Mining (CRM) in video moment proposal generation and matching when only a paragraph description of activities without per-sentence temporal annotation is available. Specifically, we explore two cross-sentence relational constraints: (1) Temporal ordering and (2) semantic consistency among sentences in a paragraph description of video activities. Existing weakly-supervised techniques only consider within-sentence video segment correlations in training without considering cross-sentence paragraph context. This can mislead due to ambiguous expressions of individual sentences with visually indiscriminate video moment proposals in isolation. Experiments on two publicly available activity localisation datasets show the advantages of our approach over the state-of-the-art weakly supervised methods, especially so when the video activity descriptions become more complex.

1. Introduction

Video activity localisation by natural language is an important yet challenging task, which aims to localise temporally a video segment (moment[1]) that best corresponds to a query sentence in an untrimmed (and often unstructured) video [21, 8]. Most of the existing methods address this task in a fully supervised manner [22, 6], *i.e.* the untrimmed video data are annotated by both a paragraph description, in which each sentence is describing a video *moment-of-interest* (MoI), and per-sentence temporal boundaries on the precise start and end time indices of every MoI. Given such fine-grained labelling, models can generate MoIs from the original videos to learn the best alignment of MoIs with their descriptions (Fig. 1 (a)). To avoid the high annotation cost and subjective annotation bias[2], recent works focus on

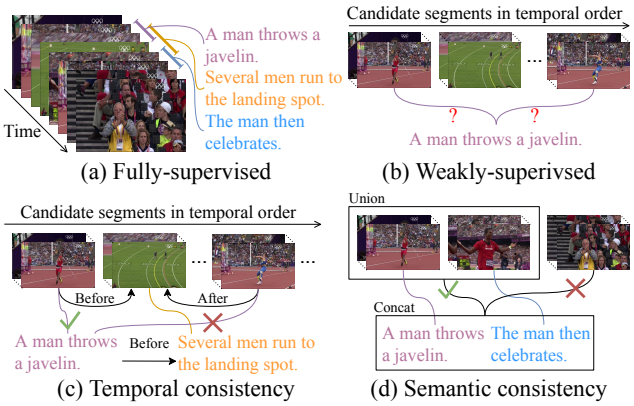


Figure 1: Different video activity localisation methods: **(a)** Given a paragraph description and the per-sentence temporal annotation (start and end time index), fully-supervised methods learn to align sentences with ground-truth semantically matching video moments [6, 22]. **(b)** Without fine-grained temporal annotations, weakly-supervised models often generate proposals of video segments corresponding to sentences in a paragraph before learning the best visual-text alignment [20, 18]. **(c)** The CRM model explores the temporal order of different sentences in a paragraph to minimise the ambiguities in matching the best video moments to specific sentences in the context of a paragraph. **(d)** To deal with ambiguous expressions in descriptions, CRM further explore plausible sentence expansion, *e.g.* pairing two sentences (concatenation) as a more complex query to constrain the localisation of pairwise video moment proposals. This explores cross-sentencing semantic consistency.

*Corresponding authors.
[1]Video segment and moment are used interchangeably in this paper.
[2]Different temporal boundaries are marked for the same sentences [1].

weakly-supervised learning without per-sentence temporal boundary annotations in training [8, 11, 21].

Existing weakly-supervised solutions [37, 25, 18] localise different MoIs individually (Fig. 1 (b)), which is not optimal as it neglects the fact that the cross-sentence relations in a paragraph play an important role in temporally localising multiple MoIs. Critically, an individual sentence is sometimes ambiguous out of its paragraph context [30, 24, 39]. For example in Fig. 1 (c), without the consideration of the *temporal relations* with the second sentence, the first query sentence (purple) can be easily mismatched with incorrect video segment, which is visually indiscriminate from the ground-truth moment. Our analysis on the ActivityNet-Captions [15] shows that the temporal relations of over 65% moment pairs predicted by a latest model [18] are contradictory with the true order of their descriptions. Yet, MoIs described by a paragraph are often semantically related to each other in their corresponding sentences. For example in Fig. 1 (d), "the man" in the blue query exhibits ambiguity if its *semantic relations* with previous sentences are ignored. We also observed that more than 38% descriptions in ActivityNet-Captions [15] contain ambiguous ways of referring to expressions, *e.g*. pronouns. To conclude, there are large error-margins in mis-localising individual sentences to video segments in isolation.

In this work, we introduce a weakly-supervised method for video activity localisation by natural language called *Cross-sentence Relations Mining* (CRM). The key idea is to explore the cross-sentence relations in a paragraph as constraints to better interpret and match complex moment-wise temporal and semantic relations in videos. Given the one-to-one moment-sentence mappings, the inherent cross-moment relations are unknown and not straightforward to be modelled in videos but intrinsically available in the paragraph descriptions. Hence, we impose the same cross-sentencing relations to their potentially matching video moments for more reliable proposal selections. The proposed CRM method differs significantly from the existing weakly-supervised models [37, 20, 25] which localise per-sentence queries *individually*. They lack fundamentally any ability to make use of the cross-sentence relations for moment proposal selection in model training. Even though such relational information is less complete than per-sentence fine-grained temporal annotation, it requires no annotation and avoids subjective bias from inherent ambiguity in temporal labelling [1]. Specifically, by assuming different activities in videos are described sequentially, we formulate a *temporal consistency* constraint to encourage the selected moments to be temporally ordered according to their descriptions in a paragraph (Fig. 1 (c)). This is different from the temporal pretext tasks in self-supervised video learning where the temporal constraint is adopted within a *single modality*. We exploit it in a *cross modality* setup, *i.e*., constraining the temporal order of event in visual modality by the sentences order in text modality. Moreover, we encourage moment proposal selections to satisfy cross-sentence broader semantics in context to minimise video-text matching ambiguities. To that end, we introduce a *semantic consistency* constraint to ensure that a moment selected for any pairing of two sentences (concatenation) in a paragraph is consistent (overlapping) with the union of the selected segments per sentence (Fig. 1 (d)).

Our **contributions** are: **(1)** To our best knowledge, this is the first idea to develop a model using *cross-sentence relations* in a paragraph to explicitly represent and compute *cross-moment relations* in videos, so as to alleviate the ambiguity of each individual sentence in video activity localisation. **(2)** We formulate a new weakly-supervised method for activity localisation by natural language called *Cross-sentence Relations Mining* (CRM), that trains a model with both temporal and semantic cross-sentence relations to improve per-sentence temporal boundary prediction in testing. **(3)** Our approach achieves the state-of-the-art performance on two available activity localisation benchmarks, especially so given more complex query descriptions.

2. Related Works

Early studies of video activity localisation by natural language mostly concentrate on making use of temporal annotations to learn visual-text alignment with *strong supervision* [9, 2, 12, 34, 33, 6]. However, due to the unaffordable annotation cost of the fine-grained temporal boundary, a growing number of works in recent years have turned to tackle this task with only the video-level moment's description, *i.e*. *weak supervision* [8, 11, 21, 18, 28, 37].

Strong Supervision. With the help of temporal annotation, fully-supervised methods localise activity in untrimmed videos either in frame or segment-level. SAP [5] proposed to compute the visual-linguistic correlation scores of the sentences and every frame in videos and group the highly correlated frames as the predicted moments. MCN [13] instead pre-divided videos into candidate segments (proposals) with variant lengths in different positions so to conduct segment-level semantic alignment. The latest methods either follow SAP to predict the probabilities of boundary across frames [3, 33, 6, 4, 22] or in the same spirit as MCN to select from a set of pre-defined proposals constructed by explicit sliding windows [9, 19] or implicit multi-granularity anchors [34, 29, 32]. Recently, DPIN [27] proposed to combine the two localisation strategies by a dual path interaction network so to take the advantage of both. Regardless of their remarkable success, fully-supervised methods rely heavily on the fine-grained temporal annotation, which is not only expensive but also prone to subjective bias [1]. In this work, we propose to

further exploit the video-level descriptions of MoIs as well as their relations so to reduce the gaps between the weakly and fully-supervised models without extra annotation cost.

Weak Supervision. In the absence of temporal boundary annotations, most of the existing weakly-supervised approaches are either based on multi-instance learning [14] (MIL) or jointly learn with reconstruction task. The MIL-based methods [11, 25, 20, 37] learn the visual-text alignment in the video-level by maximising the matching scores of the videos and their corresponding descriptions manually annotated on the datasets while suppressing that of the videos and the descriptions of others. Such learned visual-text alignment is then applied to localise the moments which are best matched with the given queries in inference. Another commonly adopted strategy [18, 8] aims at selecting the video segments which can help accomplish the reconstruction task to the largest extent, *e.g.* WS-DEC [8] jointly optimises the sentence localisation and video captioning tasks so to identify the video segments which yield consistent captions with the queries. Even though remarkable progress has been made in the past few years, none of these methods fully exploit the video-level descriptions but treat different sentences in the paragraph independently. In this work, we propose to explore the relations of sentences in paragraphs to constrain the selections of moments in training so that only the reliable video segments with consistent relations will be aligned with the query sentences.

Temporal action localisation [10, 17, 38] is a similar task which localises the pre-defined action classes in untrimmed videos. However, the language query is usually composed of multiple actions with intricate correlations, which make it more practical but challenging to be localised.

3. Weakly-Supervised Activity Localisation

Suppose we have N untrimmed video $\mathcal{V} = \{V_i\}_{i=1}^{N}$ with each composed of L_c disjoint clips $V_i = \{c_i^j\}_{j=1}^{L_c}$ in fixed duration. Corresponding to each video, we have a description paragraph consisting of L_q text query sentences $Q_i = \{Q_i^j\}_{j=1}^{L_q}$ one-to-one describing the MoIs in V_i. Given a video-query pair (V_i, Q_i^j), by dividing the untrimmed video V_i into L_s candidate segments $\{S_i^k\}_{k=1}^{L_s}$ using sliding windows [18, 20] as the *proposals*, our objective is to select the S_i^k from all the proposals which is most aligned with Q_i^j *in semantic*. For simplicity, we take a single video V and its description paragraph $Q = \{Q^j\}_{j=1}^{L_q}$ as example in the following discussion and deprecate the subscript i. Although the video-query (multi-sentences) relations are available in training, there is no access to the ground-truth per-sentence temporal boundary. This is a weakly-supervised learning problem where video proposals S^k interact with the text queries Q^j to discover the most plausible matches between video segments and text sentences.

Here we formulate a *Cross-sentence Relations Mining* (CRM) method for this task. Fig. 2 shows an overview. We first learn the visual-text alignment in video-level with the same spirit of MIL to feed a video-query pair into a *modalities matching network* (MMN), which predicts the matching score of the query and every proposal and supervise the max-pooling of scores by binary cross-entropy loss. We then explore the order of two descriptions in the paragraph and optimise their joint matching scores to a proposals pair with consistent temporal relations. Furthermore, we synthesise a longer query by forming pairs of sentences in a paragraph (concatenation) and encourage its pairwise localisation to be semantically consistent with the union of proposals individually selected for each sentence. This is to minimise the ambiguities in sentences so to improve the model's interpretation of multiple video moments in a more complex sentencing context.

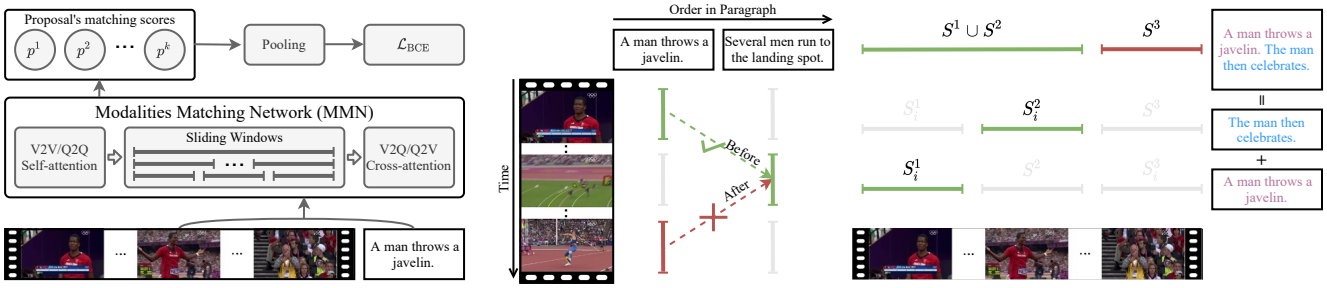
3.1. Video-Sentence Alignment

We start with the alignment of representations from two different modalities, *i.e.* an untrimmed video $V = \{c_i^1, c_i^2, \cdots, c_i^{L_c}\} \in \mathbb{R}^{L_c \times D_v}$ composed of L_c clips and a query sentence $Q^j = \{w^{j,1}, w^{j,2}, \cdots, w^{j,L_w}\} \in \mathbb{R}^{L_w \times D_t}$ with L_w words. To explore the relation of V and Q^j and enable visual-text interaction, both the representations are first projected into D-dimensional spaces by two independent fully-connected layers, respectively. For clarity concern, we reuse the symbols $V \in \mathbb{R}^{L_v \times D}$ and $Q^j \in \mathbb{R}^{L_w \times D}$ after projections. Both the video V and the query Q^j will then be fed into a *Modalities Matching Network* (MMN), which will generate a set of candidate moments (proposals) $\{S^1, S^2, \cdots, S^{L_s}\}$ by sliding windows [18, 20] and predicts their individual matching scores with the input query $\{p(S^k|Q^j)\}_{k=1}^{L_s}$ (Fig. 2 (a)). Motivated by the remarkable success of Transformer [26, 7] on sequence analysis, the MMN is composed of a stack of attention units to explore both the within and cross-modal correlation.

Attention Unit. As the building block of our MMN, the attention unit plays a significant role to learn the representation of a target sequence in terms of its correlations with every element in a reference sequence. Given a target sequence $X^t \in \mathbb{R}^{L_t \times D}$ and a reference $X^r \in \mathbb{R}^{L_r \times D}$, an attention unit $\mathcal{F}(X^t, X^r)$ attends X^t using X^r as follows:

$$\begin{aligned} \mathcal{A} &= \text{Softmax}(X^t W^{q\top} W^k X^{r\top} / \sqrt{D}) \in \mathbb{R}^{L_t \times L_r} \\ \mathcal{F}(X^t, X^r) &= \text{FC}(X^t + \mathcal{A} X^r W^{v\top}) \in \mathbb{R}^{L_t \times D}. \end{aligned} \tag{1}$$

The notions $\{W^q; W^k; W^v\} \in \mathbb{R}^{3 \times D \times D}$ in Eq. (1) are three learnable matrices and the coefficient $1/\sqrt{D}$ is to counteract the effect of small gradients caused by large D [26]. The Softmax$(\cdot)$ is the row-wise softmax normalisation and $\mathcal{A}$ is the correlation scores of target-reference element pairs. The FC$(\cdot)$ is a linear projection with consis-



(a) Multi-instance Learning (b) Temporal consistency (c) Semantic consistency

Figure 2: Overview of the proposed *Cross-sentence Relations Mining* (CRM) method. **(a)** The modalities matching network (MMN) is composed of self and cross-attention units and trained by MIL objective. **(b)** The joint matching scores of two queries to a pair of proposals are optimised to encourage the consistency of cross-sentence and cross-moment temporal relations. **(c)** A longer query is synthesised by pairs of sentences in a paragraph (concatenation), whose their pairwise localisation is constrained to be consistent with the union of the two proposals selected for each sentence.

tent input-output dimensions. The attended result serves as the updated representation of the target sequence.

To investigate the visual-text matching relations, it is essential to explore not only the within-modal context but also the cross-modal interaction [20]. Hence, the MMN is constructed by both self-attention and cross-attention blocks. The video V and the query Q^j are first fed into two independent self-attention blocks respectively, in which the target and reference inputs are from the same modalities:

$$V \leftarrow \mathcal{F}^{\text{V2V}}(V, V), \quad Q^j \leftarrow \mathcal{F}^{\text{Q2Q}}(Q^j, Q^j). \tag{2}$$

By doing so, the salient clips/words in the input video/query are highlighted by considering the context of the video or sentence. Conventional sliding window strategy [18, 20] is then adopted to divide the video into L_s proposals $V = \{S^k\}_{k=1}^{L_s} \in \mathbb{R}^{L_s \times D}$. Each proposal is composed of arbitrary continual clips in V and represented by max-pooling the features of its included clips. After that, the two representations are interacted by cross-attention blocks:

$$V \leftarrow \mathcal{F}^{\text{Q2V}}(V, Q^j), \quad Q^j \leftarrow \mathcal{F}^{\text{V2Q}}(Q^j, V), \tag{3}$$

which attends one modality by another so to suppress the redundant text and irrelevant visual information.

Matching Score. Given the visual features $V = \{S^k\}_{k=1}^{L_s}$ and the text representation $Q^j = \{w^{j,k}\}_{k=1}^{L_w}$, the matching score $p(S^k|Q^j)$ of a proposal-query pair is predicted according to both the modalities. The sentence representation is first computed by aggregating all the words: $Q^j \leftarrow \text{cmax}(\{w^{j,k}\}_{k=1}^{L_w}) \in \mathbb{R}^{1\times D}$ where $\text{cmax}(\cdot)$ denotes the column-wise max-pooling function, which is then fused with every proposal's representation [9, 13]:

$$E^{k,j} = (S^k + Q^j) \| (S^k \otimes Q^j) \| \text{FC}(S^k \| Q^j). \tag{4}$$

The notion $(\cdot \otimes \cdot)$ indicates the element-wise multiplication and $(\cdot \| \cdot)$ is the concatenation of two vectors while $\text{FC}(\cdot)$ standing for a linear projection. After that, the joint representations $\{E^{k,j}\}_{k=1}^{L_s}$ are fed into a linear classifier:

$$p(S^k|Q^j) = \sigma(E^{k,j}W^\top + B). \tag{5}$$

The variable $\{W, B\}^\top \in \mathbb{R}^{D+1}$ is the weights of classifier and $\sigma(\cdot)$ is the sigmoid function. The yielded probabilities $\{p(S^k|Q^j)\}_{k=1}^{L_s} \in (0, 1)$ serve as the matching scores between proposals and query, which is abbreviated to $p^{k,j}$.

Multi-Instance Learning. In the absence of temporal boundary, the ground-truth moment is agnostic. Therefore, we optimise the matching scores in video-level to facilitate visual-text alignment. To that end, the matching score between the video V and the query Q^j is obtained by the max-pooling of all the proposals' score $p(V|Q^j) \leftarrow \max(\{p^{k,j}\}_{k=1}^{L_s})$. For each positive pair (V, Q^j) given manually on the dataset, we construct two negative counterparts by replacing either V or Q^j by a randomly sampled video V^- or sentence Q^- from the mini-batch and compute their matching scores in the same way as $p(V|Q^j)$. The binary cross-entropy (BCE) loss function is then adopted as the video-query alignment supervision signal:

$$\begin{aligned} \mathcal{L}_{\text{BCE}}(V, Q^j) = 2* &- \log p(V|Q^j) \\ &- \log(1 - p(V|Q^-)) - \log(1 - p(V^-|Q^j)), \end{aligned} \tag{6}$$

where the coefficient 2 is applied to the positive term considering the balance of positive and negative pairs. The rationale behind Eq. (6) is assuming that the MoIs in one video doesn't exist in any other videos so (V, Q^-) and (V^-, Q^j) should be semantically unmatched. By minimising $p(V|Q^-)$ and $p(V^-|Q)$, the predictions of the incorrect proposals in V with different semantics from Q^j will also be minimised implicitly so that the learned matching scores can reveal the inherent visual-text relations. This takes the spirit of MIL [14] by treating the proposals as the instances in a bag (video) and learning with the bag-level annotations.

3.2. Cross-Sentence Relations Mining

The $\mathcal{L}_{\text{BCE}}$ in Eq. (6) aligns queries with the proposals yielding the largest matching scores among all the candidates. However, the predicted scores can be unreliable due to the visually indiscriminate moment proposals existed in videos and text ambiguities in individual sentences which will lead to visual-text misalignment in training. Therefore, we explore the cross-sentence relations to select reliable proposals with consistent cross-moment relations.

Temporal Consistency. As the video frames are naturally exhibited to the viewers in time order, the temporal relations of different MoIs should intrinsically be encoded in the order of their descriptions in the paragraph. With such an assumption, we can identify the pairs of proposals both yielding high predicted matching score with the corresponding queries but inconsistent in temporal relations, which are likely to be incorrect. Given arbitrary query sentences pair $(Q^j, Q^{j'})$ from the description paragraph of video V, their respective selected segments $(S^k, S^{k'})$ should satisfy similar temporal structure with them, *i.e.* S^k should occur before $S^{k'}$ in the video if Q^j is in front of $Q^{j'}$ in the paragraph and vice versa. The temporal order of two proposals $\mathcal{R}(S^k, S^{k'}) = 0$ if S^k starts before $S^{k'}$ in the video, otherwise $\mathcal{R}(S^k, S^{k'}) = 1$. Similarly, $\mathcal{R}(Q^j, Q^{j'}) = \mathbb{1}[j >= j']$ where j and j' are the position of sentences in the paragraph. The temporal constraint is then formulated to ensure $\mathcal{R}(S^k, S^{k'}) = \mathcal{R}(Q^j, Q^{j'})$.

By assuming the matching scores of different queries to any proposals are independent, the joint probability of Q^j and $Q^{j'}$ are respectively matching with S^k and $S^{k'}$ is:

$$p(S^k, S^{k'}|Q^j, Q^{j'}) = p(S^k|Q^j) \cdot p(S^{k'}|Q^{j'}). \quad (7)$$

As shown in Fig. 2 (b), we take the queries' order as the ground-truth for the temporal relation of the proposal pair. Given Q^j and $Q^{j'}$, the joint probabilities set $\{p(S^k, S^{k'}|Q^j, Q^{j'})\}_{k,k'=1}^{L_s}$ is then divided into two subsets: for all the proposal pairs $(S^k, S^{k'})$, the joint probability $p(S^k, S^{k'}|Q^j, Q^{j'}) \in P_t^+$ if $\mathcal{R}(S^k, S^{k'}) = \mathcal{R}(Q^j, Q^{j'})$, otherwise belonging to P_t^-. The MIL loss is re-formulated with the temporal constraint:

$$\begin{aligned}\mathcal{L}_{\text{TMP}}(V, Q^j, Q^{j'}) = &- \log(\max(P_t^+)) \\ &- \log(1 - \max(P_t^-)).\end{aligned} \quad (8)$$

By training with $\mathcal{L}_{\text{TMP}}$, the model learns to align the proposals with queries only if they are temporally consistent. This refrains the model from visual-text misalignment in the absence of ground-truth temporal annotations.

Semantic Consistency. To minimise the negative impact from ambiguous per-sentence expressions in isolation and to explore the context of a paragraph, it is beneficial for a model to consider broader semantics beyond individual sentences by relating other expressed objects/actions in a wider context [22]. However, it is nontrivial to explicitly do so since the object/action's information is missing without fine-grained annotation. In this case, we propose to form pairs of MoIs by concatenation in the same videos: $Q^{j,j'} = Q^j \| Q^{j'}$ and train the model to localise the concatenated longer query with the consideration of both sentences in each pair. Given the proposals S^k and $S^{k'}$ with the largest $p(S^k, S^{k'}|Q^j, Q^{j'})$ in Eq. (8), the matching scores of $Q^{j,j'}$ and the video segments S^l is optimised to encourage the consistency of S^l and $S^k \cup S^{k'}$ (Fig. 2 (c)). As in the temporal constraint, we divide the predicted scores $p(S^l|Q^{j,j'})$ into two subsets: for all the proposals S^l in the video V, $p(S^l|Q^{j,j'}) \in P_s^-$ if $\text{IoU}(S^l, S^k \cup S^{k'}) < \tau$, and P_s^+ is composed of the S^l which is most consistent with $S^k \cup S^{k'}$. The τ decides how two proposals are deemed inconsistent regarding their intersection over union score (IoU) which is set to 0.5 in practice. The constraint on the semantic consistency of S^l and $S^k \cup S^{k'}$ is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{SMT}}(V, Q^j, Q^{j'}) = &- \log(\max(P_s^+)) \\ &- \log(1 - \max(P_s^-)).\end{aligned} \quad (9)$$

To minimise $\mathcal{L}_{\text{SMT}}$, the model is explicitly trained to consider the semantics of both Q^j and $Q^{j'}$ when localising $Q^{j,j'}$ so to ensure the overlap of S^l and $S^k \cup S^{k'}$. By introducing additional longer queries synthesised from pairwise sentences in model training, it enhances the model's capacity to interpret and match more complex descriptions to video moments, critical in practice due to that untrimmed raw videos are often unstructured.

3.3. Model Training

In each training iteration, we randomly sample n videos with a pair of queries for each from its paragraph description as a mini-batch and the overall loss is computed by:

$$\begin{aligned}\mathcal{L} = &\frac{1}{2*n}\sum_{i=1}^{n}\sum_{j=1}^{2}\mathcal{L}_{\text{BCE}}(V_i, Q_i^j) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{\text{TMP}}(V_i, Q_i^1, Q_i^2) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{\text{SMT}}(V_i, Q_i^1, Q_i^2).\end{aligned} \quad (10)$$

Since the objective function $\mathcal{L}$ in Eq. (10) is differentiable, conventional stochastic gradient descent algorithm is adopted for end-to-end model training. The overall process of a training iteration is summarised in Alg. 1.

Algorithm 1 Video activity localisation by CRM

Input: Untrimmed videos $\mathcal{V}$, Paragraph descriptions $\mathcal{Q}$.
Output: An updated video activity localisation model.
Sampling a random mini-batch of videos;
Sampling two queries for each video from its paragraph;
foreach video-query pair **do**
  Mapping video and query to D-dimensional spaces;
  Conducting V2V and Q2Q self-attention (Eq. (2));
  Generating proposals by sliding windows;
  Conducting V2Q and Q2V cross-attention (Eq. (3));
  Fusing each proposal's feature with the query (Eq. (4));
  Computing the proposal-query matching scores (Eq. (5));
end foreach
Computing the objective loss (Eq. (10));
Updating model weights by back-propagation.

4. Experiment

Datasets. Experiments were conducted on two video activity localisation datasets: (1) Charades-STA [9] contains 12,408/3720 video-query pairs from 5338/1334 videos for training and testing, respectively. The query sentences are composed of 7.2 words on average and the average duration of the target video moments and untrimmed videos are 8.1 and 30.6 seconds; (2) ActivityNet-Captions [15] is a much larger-scale dataset composed of 19,290 videos with 37,417/17,505/17,031 MoIs in the train/val_1/val_2 split. The average length of queries is 14 words while that of the MoIs and untrimmed videos are 36.2 and 117.6 seconds.

The activities captured in those two datasets are of various complexity: Only 6% of the descriptions involve more than one actions in Charades whilst 44% in ActivityNet with 12% *vs*. 44% regarding the number of people [16].

Performance Metric. We followed previous works [8, 28, 6] to evaluate the activity localisation results by the "IoU@m" metric where m is the pre-defined temporal Intersection over Union (IoU) thresholds. Given the temporal boundary (s, e) of a target moment and the selected segment proposal $(\tilde{s}, \tilde{e})$ with the largest predicted matching score, the IoU between the two video segments is computed by $\frac{\max(0,\min(e,\tilde{e})-\max(s,\tilde{s}))}{\max(e,\tilde{e})-\min(s,\tilde{s})}$. A prediction is considered correct if its IoU with the ground-truth is greater than the pre-defined IoU thresholds set to $\{0.1, 0.3, 0.5\}$ on ActivityNet and $\{0.3, 0.5, 0.7\}$ on Charades [8, 28].

Implementation. We used VGG (4096-D) and ResNet152 (2048-D) feature representations officially released with the datasets for per-frame representations in Charades and ActivityNet, respectively. The videos were truncated evenly (and zero-padded) into 128 clips in Charades and 256 in ActivityNet, with each clip represented by the max-pooling of 5 continual frame's features. The pre-trained GloVe embedding [23] was adopted as the word feature representation (300-D) and the maximal sentence length was set to 20 words. Both the clip and word representations were linearly mapped to 256-D spaces before being fed into MMN. The sliding windows stride was 8 and the window sizes were $\{8, 12, 20, 32, 64\}$ in Charades and $\{8, 16, 32, 64, 128\}$ in ActivityNet. The temporal dependencies of video segments in terms of the same query sentences were explored by an additional self-attention unit before predicting their matching scores. As the paragraph descriptions were pre-divided into individual sentences on both datasets, we restored the order of sentences in the paragraph by the ground-truth start time of MoIs. Note that timestamps were unavailable in proposal selections, neither in training nor testing. The proposed CRM was trained 50 epochs by Adam optimiser with a batch size of 64 and learning rate of $1e-4$. Cross-sentence relations were only used in training with no extra computational cost in testing.

4.1. Comparisons to the State-Of-The-Art

Table 1 compares the performance of CRM against the state-of-the-art video activity localisation models including both fully- and weakly-supervised methods. We observe: **(1)** Not surprisingly, fully-supervised models outperform weakly-supervised models clearly. However, CRM reduces that performance gap by over 41% on the ActivityNet at IoU $= 0.3$. **(2)** Discovering different video moments correlating to the *same-sentence* for proposal selection has been exploited to a good effect by existing methods in the form of attention [18, 20] or 2D temporal convolution [36, 35]. However, the notably better performance of CRM compared to those methods further demonstrates the additional advantage of using *cross-sentence* temporal and semantic relations within a paragraph for learning better visual-text alignment and benefiting per-sentence localisation in testing. **(3)** CRM surpasses the state-of-the-art weakly-supervised methods across the board except for IoU@0.3 on Charades. This demonstrates compellingly the effectiveness of CRM from modelling explicitly cross-sentence relations. Our advantages on the OOD split of ActivityNet-Captions [31] further indicate CRM's better multi-modal understanding rather than driven by annotation biases.

4.2. Components Analysis

We investigated the effects of different components in CRM model design to study their individual contributions. The "val_1" split of ActivityNet was adopted.

Effects of Cross-sentence Relations. We evaluated the effectiveness of imposing cross-sentence relational consistency by training the baseline model (BCE) with either the temporal (BCE+TMP) or semantic (BCE+SMT) constraint as well as with both (BCE+TMP+SMT). Fig. 3 shows that both constraints are beneficial individually and the benefits become more clear when they are jointly adopted. More-

Split	Method	Moment	Query	IoU@0.1	IoU@0.3	IoU@0.5
val_2	DPIN [27]	✓	✗	-	62.40	47.27
	2D-TAN [35]	✓	✗	-	59.45	44.51
	DRN [33]	✗	✗	-	-	42.49
	LGI [22]	✓	✗	-	58.52	41.51
	HVTG [6]	✗	✗	-	57.60	40.15
val_1	WS-DEC [8]	✗	✗	62.71	41.98	23.34
	WSLLN [11]	✗	✗	75.4	42.8	22.7
	BAR [28]	✓	✗	-	49.03	30.73
	CRM (Ours)	✓	✓	**76.66**	**51.17**	**31.67**
val_2	SCN [18]	✓	✗	71.48	47.23	29.22
	RTBPN [36]	✓	✗	73.73	49.77	29.63
	CCL [37]	✓	✗	-	50.12	31.07
	CRM (Ours)	✓	✓	**81.61**	**55.26**	**32.19**
OOD	WS-DEC [8]	✓	✗	30.71	17.00	7.17
	CRM (Ours)	✓	✓	**38.35**	**22.77**	**10.31**

(a) ActivityNet-Captions

Method	Moment	Query	IoU@0.3	IoU@0.5	IoU@0.7
DPIN [27]	✓	✗	-	47.98	26.96
2D-TAN [35]	✓	✗	-	39.81	23.25
DRN [33]	✗	✗	-	53.09	31.75
LGI [22]	✓	✗	72.96	59.46	35.48
HVTG [6]	✗	✗	61.37	47.27	23.30
TGA [21]	✗	✗	29.68	17.04	6.93
SCN [18]	✓	✗	42.96	23.58	9.97
LoGAN [25]	✓	✗	51.67	34.68	14.54
BAR [28]	✓	✗	44.97	27.04	12.23
RTBPN [36]	✓	✗	**60.04**	32.36	13.24
VLANet [20]	✓	✗	45.24	31.83	14.17
CCL [37]	✓	✗	-	33.21	15.68
CRM (Ours)	✓	✓	53.66	**34.76**	**16.37**

(b) Charades-STA

Table 1: Performance comparisons on video activity localisation methods. Fully and weakly-supervised methods are shown in the upper and lower part of each table, respectively. The 'Moment' column refers to methods trained by exploiting multiple video moments corresponding to the same-sentence, whilst the 'Query' column refers to training by cross-sentence temporal ordering and sentence pairing in the context of a paragraph. The 'Split' column denotes the different data splits in the ActivityNet-Captions used in the evaluations. The discounted recall rates [31] are reported for the 'OOD' split of ActivityNet-Captions.

over, the performance improvement is more significant on ActivityNet than Charades. Given the generally more complex activities in ActivityNet, this shows that training CRM on combinations of pairwise sentencing as semantic consistency constraint (Eq. (9)) has its unique advantages in activity localisation against more complex query descriptions.

Temporal Consistency. To verify our assumption on temporal order, we compared how many correct predictions learned with and without $\mathcal{L}_{\text{TMP}}$ (Eq. (8)) against the ground-truth. Specifically, for each video consists of n MoIs, we constructed C_n^2 MoI pairs and measured the ratio of consistent pairs by comparing the order of the two ground-truth moments and that of the selected proposals. Table 2 shows that by explicitly training CRM with cross-sentence temporal order constraint, the video segments selected by CRM is much more consistent in temporal relations on ActivityNet than the base models without it. Although different moments in the test set are localised independently, such advantages are still clear. Besides, it is surprising to see that the cross-moment temporal relations yielded by the base model on Charades are reasonably consistent with the true order but the temporal constraint still benefited the localisation results. This implies the potential advantages of optimising joint matching scores of moment pairs with their descriptions in learning effective visual-text alignment.

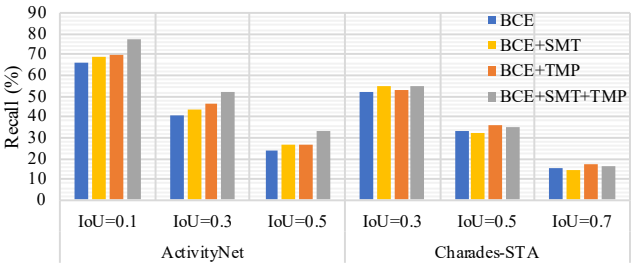


Figure 3: Effects of cross-sentence relations mining. BCE is the base model trains with only the MIL objective (Eq. (6)). TMP and SMT are the proposed constraints on temporal (Eq. (8)) and semantic (Eq. (9)) relational consistency.

	Train		Test	
Temporal	ActivityNet	Charades	ActivityNet	Charades
✗	64.28	73.88	45.02	73.91
✓	82.43	74.88	70.82	74.65

Table 2: Temporal consistency between the descriptions of MoI pairs and their selected proposals. Metric: accuracy.

	Train		Test	
Semantic	ActivityNet	Charades	ActivityNet	Charades
✗	55.76	35.34	57.84	31.01
✓	68.14	55.46	71.30	51.33

Table 3: Semantic consistency between the union of two MoIs' segments and the one selected for the concatenation of their descriptions. Metric: prediction recall at IoU $= 0.5$.

Semantic Consistency. As in the analysis of temporal consistency, we enumerated all the possible MoI pairs in the same videos and quantify the semantic consistency by taking the union of MoI pairs as the ground-truth moment corresponding to the concatenation of their descriptions. More specifically, given the sentence description of two MoIs and their temporal boundary S^i and S^j, we concatenated the

two per-sentence queries and identified the video segment S^k yielding the largest matching scores with the concatenation. We then computed the temporal IoU between $S^i \cup S^j$ and S^k, where S^k is deemed semantically consistent with $S^i \cup S^j$ if $\text{IoU}(S^i \cup S^j, S^l) > 0.5$. Note that it is not necessary for the two moments to be consecutive in time so that our semantic assumption can hold, as the boundary defined by the concatenated description always matches their temporal union. Table 3 shows that the baseline model trained without semantic constraint in Eq. (9) yields sensible performances in localising the paired queries. This demonstrates that CRM implicitly learns to consider the semantic context of queries by the attention units. The superior results of CRM trained with explicit semantic constraint shows that it encourages broader consensus in semantics across sentences. This explains why the performance advantages of CRM is more significant when localising more complex activities in ActivityNet.

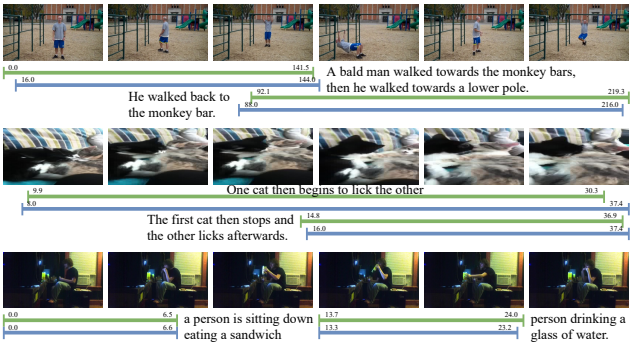
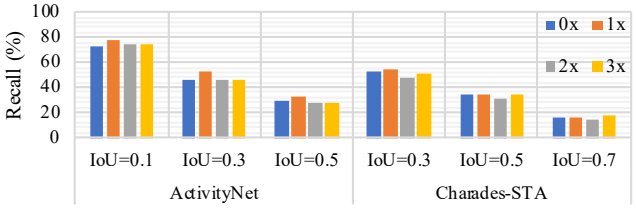


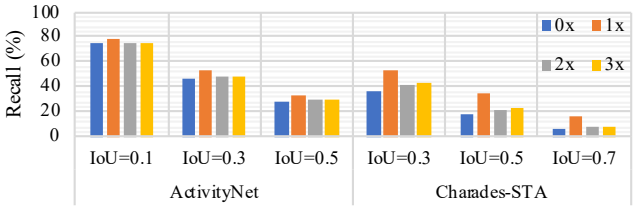
Figure 4: Qualitative examples show the interaction between MoIs in the same videos. The green bars indicate the ground-truth MoI's boundaries whilst the blue bars show the model predictions by CRM. The query sentences are simplified for illustrations only given the space limit.

Qualitative Examples. Fig. 4 shows some qualitative examples from both ActivityNet and Charades. They show how different MoIs in the same videos may interact with each other so that their relations can be used to optimise per-sentence activity localisation in the context of a paragraph. It is evident that localising video moments by per-sentence independently is unreliable, *e.g.* in the first example (top-row), the man reaches the monkey bars both before and after he walks toward the lower pole. "The first cat" example in the middle-row is ambiguous without context. By explicitly exploring the cross-sentence relations, CRM avoids such ambiguities and minimises video-text misalignment.

Effects of Attention Units. As the building block of our MMN backbone in section 3.1, the attention units play a significant role in exploring the videos and sentences data as well as their correlations. We investigated its effect by comparing the prediction recall of CRM constructed with



(a) Self-attention



(b) Cross-attention

Figure 5: Effects of attention units. Models are constructed and trained with different numbers of self-attention and cross-attention units to investigate their effects.

different numbers of attention units, showing its benefits in sequence analysis and visual-text interactions (Fig. 5). On the other hand, due to the limited video data available for training (10K/5K on ActivityNet/Charades), stacking up attention layers fails to further benefit CRM, leading to model performance degradation possibly due to overfitting.

5. Conclusion

In this work, we presented a novel *Cross-sentence Relations Mining* (CRM) method for learning video activity localisation in the absence of per-sentence temporal annotation. CRM explores cross-sentence relations within each paragraph description of a long video to optimise video moment proposal selections in training so to improve per-sentence localisation in testing. CRM minimises mismatching individual sentences to video moment proposals during training by constraining their selections according to the temporal ordering and pairwise sentencing as expanded queries in the context of a paragraph description of video. This improves notably CRM's capacity to localise more accurately video activities against more complex language descriptions. Experiments on two available activity localisation benchmark datasets show the performance advantages of the proposed CRM method over a wide range of state-of-the-art weakly-supervised models. Extensive ablation studies further provided in-depth analysis of the effectiveness of the individual components in CRM.

Acknowledgements

This work was supported by the China Scholarship Council, Vision Semantics Limited, the Alan Turing Institute Turing Fellowship, and Adobe Research.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5803–5812, 2017. 1, 2

[2] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the Empirical Methods in Natural Language Processing*, pages 162–171, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 2

[3] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8175–8182, Jul. 2019. 2

[4] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10551–10558, 2020. 2

[5] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8199–8206, 2019. 2

[6] Shaoxiang Chen and Yu-Gang Jiang. Hierarchical visual-textual graph for temporal activity localization via language. In *Proceedings of the European Conference on Computer Vision*, pages 601–618. Springer, 2020. 1, 2, 6, 7

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

[8] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 3059–3069, 2018. 1, 2, 3, 6, 7

[9] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5267–5275, 2017. 2, 4, 6

[10] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Cascaded boundary regression for temporal action detection. 2017. 3

[11] Mingfei Gao, Larry S Davis, Richard Socher, and Caiming Xiong. Wslln: Weakly supervised natural language localization networks. In *Proceedings of the Empirical Methods in Natural Language Processing*, 2019. 2, 3, 7

[12] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1984–1990, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2

[13] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. *arXiv preprint arXiv:1809.01337*, 2018. 2, 4

[14] James D Keeler, David E Rumelhart, and Wee Kheng Leow. Integrated segmentation and recognition of hand-printed numerals. In *Advances in neural information processing systems*, pages 557–563, 1991. 3, 4

[15] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2, 6

[16] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Proceedings of the European Conference on Computer Vision*, 2020. 6

[17] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018. 3

[18] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11539–11546, 2020. 1, 2, 3, 4, 6, 7

[19] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *Proceedings of the ACM International Conference on Multimedia*, page 843–851, New York, NY, USA, 2018. Association for Computing Machinery. 2

[20] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 156–171. Springer, 2020. 1, 2, 3, 4, 6, 7

[21] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2019. 1, 2, 7

[22] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020. 1, 2, 5, 7

[23] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. 6

[24] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000. 2

[25] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. 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, pages 2083–2092, 2020. 2, 3, 7

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 5998–6008, 2017. 3

[27] Hao Wang, Zheng-Jun Zha, Xuejin Chen, Zhiwei Xiong, and Jiebo Luo. Dual path interaction network for video moment localization. In *Proceedings of the ACM International Conference on Multimedia*, pages 4116–4124, 2020. 2, 7

[28] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In *Proceedings of the ACM International Conference on Multimedia*, pages 1283–1291, 2020. 2, 6, 7

[29] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069, 2019. 2

[30] DanFeng Yan and Shiyao Guo. Leveraging contextual sentences for text classification by using a neural attention model. *Computational intelligence and neuroscience*, 2019, 2019. 2

[31] Yitian Yuan, Xiaohan Lan, Long Chen, Wei Liu, Xin Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Datasets and metrics. *arXiv preprint arXiv:2101.09028*, 2021. 6, 7

[32] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *Proceedings of the Conference on Neural Information Processing Systems*, 2019. 2

[33] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020. 2, 7

[34] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 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*, pages 1247–1257, 2019. 2

[35] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877, 2020. 6, 7

[36] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the ACM International Conference on Multimedia*, pages 4098–4106, 2020. 6, 7

[37] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Proceedings of the Conference on Neural Information Processing Systems*, 33, 2020. 2, 3, 7

[38] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017. 3

[39] Matthias Zimmermann. Joint segmentation and classification of dialog acts using conditional random fields. In *Tenth Annual Conference of the International Speech Communication Association*, 2009. 2