VLG-Net: Video-Language Graph Matching Network for Video Grounding

Sisi Qu*, Mattia Soldan*, Mengmeng Xu*, Jesper Tegner, Bernard Ghanem
King Abdullah University of Science and Technology (KAUST), Saudi Arabia
{sisi.qu, mattia.soldan, mengmeng.xu, jesper.tegner, bernard.ghanem}@kaust.edu.sa

Abstract

Grounding language queries in videos aims at identifying the time interval (or moment) semantically relevant to a language query. The solution to this challenging task demands the understanding of videos' and queries' semantic content and the fine-grained reasoning about their multi-modal interactions. Our key idea is to recast this challenge into an algorithmic graph matching problem. Fueled by recent advances in Graph Neural Networks, we propose to leverage Graph Convolutional Networks to model video and textual information as well as their semantic alignment. To enable the mutual exchange of information across the domains, we design a novel Video-Language Graph Matching Network (VLG-Net) to match video and query graphs. Core ingredients include representation graphs, built on top of video snippets and query tokens separately, which are used for modeling the intra-modality relationships. A Graph Matching layer is adopted for cross-modal context modeling and multi-modal fusion. Finally, moment candidates are created using masked moment attention pooling by fusing the moment's enriched snippet features. We demonstrate superior performance over state-of-the-art grounding methods on three widely used datasets for temporal localization of moments in videos with natural language queries: ActivityNet-Captions, TACoS, and DiDeMo.

1. Introduction

Temporal action understanding is at the forefront of computer vision research. The task of temporally grounding language queries in videos was recently introduced by Hendricks *et al.* [1] and Gao *et al.* [14], as a generalization of the temporal action localization task and so as to overcome the constraint of a predefined set of actions. This novel interdisciplinary task has gained momentum within the vision and language community for its relevance and possible applications in video retrieval [11, 52, 77], video question answering [20, 26], human-computer interaction [86], and video storytelling [17]. Recently, two new tasks (video corpus moment retrieval [60] and video-subtitle moment retrieval [27]) were introduced to extend the localization task to a large corpus of videos. However, we maintain that the complexity of the single video setup requires further investigation to devise better methods.

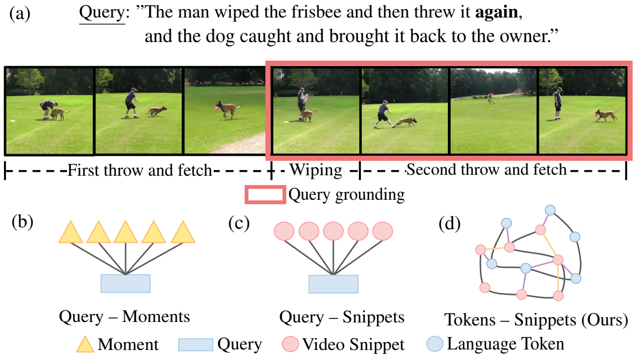


Figure 1. **Temporal video grounding task and multi-modality interaction schemes.** (a) A video grounding example showcasing the importance of fine-grained semantic understanding and necessity of proper context modelling. (b, c) Approaches for multi-modal interactions in previous works. (d) Our multi-modal interaction scheme gives finer alignments by snippet-token matching.

Natural language grounding in videos faces multiple challenges, some of which are inherited from temporal action localization, such as context awareness modelling and candidate moment generation. Semantic context is a fundamental cue necessary to boost the performance of localization methods [10, 13, 30, 62, 72]. To enrich video representation, Hendricks *et al.* [1] adopted *global-context*, which is moment independent, leading to sub-optimal performance. Conversely, a moment specific *local-context*, defined as a moment's temporal neighbourhood, was adopted in [14, 15, 23, 34, 55, 67]. In our view, *non-local context* merits deeper analysis, since it has the potential to identify relevant information not restricted to the temporal neighbourhood within one data modality. For example, in Fig. 1(a), although "First throw and fetch" is not in the temporal vicinity of "Second throw and fetch", it is still semantically related with our target moment, which is a good tes-

*denotes equal contributions. First author names are ordered with alphabetical ordering of surnames.

timony of the importance of non-local context modeling for video grounding. Moreover, as showcased by the example, the free-form nature of the language modality introduces additional challenges, as a model must operate not only with the semantic content of both videos and language queries, but also reason about their multi-modal interactions. Previous studies [14, 16, 55] employ a cross-modal processing unit designed to jointly model text and visual features by calculating element-wise addition, multiplication, and direct concatenation between a moment's representation and the query embedding, as depicted in Fig 1 (b). Alternatively, Mun *et al*. [44] only adopt Hadamard product to fuse this information at query-snippet level in Fig. 1 (c).

Motivated by the work in [72], we propose to leverage the representational capability of graphs to encode snippet-snippet, token-snippet, and token-token relations as edge connections. As such, we adopt Graph Convolutional Networks (GCN) [57] to gather intra- and inter-modality context. This sets the stage for resolving finer alignments by recasting inter-modality interactions as an algorithmic graph matching problem working at the level of snippets and tokens, as shown in Fig. 1(d), different from previous methods that focus on relatively coarse query-moment or query-snippet interactions. Inspired by [29, 71], we adopt a cross-graph attention-based matching mechanism to enable the mutual exchange of information across the modalities.

In this paper, we design a new architecture referred to as Video-Language Graph Matching Network (VLG-Net). First, representation graphs for both video and language are constructed. The video graph models each snippet as a node and takes advantage of two sets of edges to represent both local temporal relations and non-local semantic relations between video snippets. Similarly, we construct a language graph, where each node is a token and each edge reflects token-to-token relations, *e.g.* syntactic dependencies [39, 40]. These modality-specific graphs are used to model local and non-local context through graph convolutions. Then, we resort to the attention-based graph matching layer to fuse the modalities together and enable inter-modality context aggregation, allowing for fine-grained alignment through the adoption of a specialized set of learnable edges. With this design, we avoid the need for heuristics of context modelling and simultaneously learn the best way for multi-modal fusion.

Contributions. Our contributions are two-fold. **(1)** We provide a new representation for modeling contextual information and multi-modal fusion for the video grounding task. To the best of our knowledge, we are the first to utilize attention-based graph matching of language and video graphs to learn video-language interactions for grounding. **(2)** Our VLG-Net improves localization performance by enabling finer alignment between the two modalities. Extensive experiments demonstrate best results with significant improvement over state-of-the-art grounding methods in three commonly used datasets, which verify the effectiveness of our approach.

2. Related work

2.1. Video Grounding

Moment candidates. Previous works can be organized into two major categories: (i) proposal-free and (ii) proposal-based methods. Proposal-free methods, such as [9, 18, 32, 36, 38, 44, 47, 53, 75, 79, 84], aim at directly regressing the temporal boundaries of the queried moment from the multi-modal fused feature. In contrast, proposal-based approaches [1, 4, 6, 7, 15, 22, 34, 61, 67, 83, 85] adopt a propose-and-rank pipeline: by firstly generating moment proposals and then ranking them according to their similarity with the textual query. In 2D-TAN, Songyang *et al*. [85] show the removal of heavily overlapped moments can significantly reduce the amount of computation while retaining good performance. Accordingly, we design the candidates and score the moments without boundary refinements [14, 15, 33, 34, 61, 67].

Moments' context. For moment context modelling, some works [6, 18] attempt to use the memory property of LSTM cells [51] to contextualize the video features. However, attention-based mechanisms [59] adopted in [31, 33, 61] improve the aggregation of long-range semantic dependencies. Similar to [61], we suggest that visual context modelling should be dynamic and query-dependent. Songyang *et al*. [85] claim that neighbouring proposals hold valuable context and apply 2D convolutions with large kernel size on top of the moment representations to gather context information in the latter stages of their pipeline. In contrast to [85], we only use a Multi-Layer Perceptron (MLP) network for moment score computation, which reduces the overall computation, while delegating the context gathering to earlier stages of the pipeline.

Multi-modal fusion. Moving beyond the cross-modal processing unit presented in Section 1, the works of [67] devise a new cross-modality interaction scheme based on circular matrices. In [6, 34, 61], frame features are concatenated with frame-guided attention-pooled features of the language. Lu *et al*. [38] take advantage of the QANet[76] architecture, which is based on cross-attention and convolution operations, for multi-modal fusion. Dynamic filters generated from language features are adopted in [47, 83] in order to modulate, through convolutions, the visual information given the query. Recently, using the Hadamard product to fuse/gate information [44, 79, 85] has become popular. In contrast to these methods, our graph matching layer specifically models local, non-local, and query-guided context, thereby exploiting the semantic neighbourhood of snippets and tokens to fuse the modalities through graph convolutions.

2.2. Graphs and Graph Neural Networks

Graphs in Videos. In various video understanding tasks, such as action recognition [8, 35, 65], and action localization [72, 78], graphs equipped with extensive representational power have been used for representation of data/features. Each video is represented as a space-time region graph in [65] and as a 3D point cloud in the spatial-temporal space in [35]. Alternatively, Zeng *et al.* [78] define temporal action proposals as nodes to form a graph, while Xu *et al.* [72] consider video snippets as nodes in a graph. Inspired by [72], video snippets in our work are represented as nodes in a graph and different edges are constructed between them to model various relationships.

Graphs in Language. In natural language processing (NLP), both sequential and non-local relations are crucial. The former is usually captured by recurrent neural networks [49], while the latter can be represented using graph neural networks [2, 3, 28, 41, 43, 55, 80]. Syntactic information has been shown successful for language modelling when combined with GCN in [21, 31, 40, 81]. Driven by these arguments, we stack together LSTM and a Syntactic Graph Convolution Network (SyntacGCN) [21, 31, 40, 81] for modelling and enriching the language features.

Graph Neural Networks in Graph Matching. Graph matching is one of the core problems in graph pattern recognition, aiming to find node correspondences between different graphs [5]. Given the powerful capability of Graph Neural Networks (GNNs) for encoding graph structure information, approaches leveraging GNNs/GCNs have recently surfaced in the graph matching field [63, 71]. For example, Li *et al.* [29] propose a GNN-based graph matching network to represent graphs as feature vectors to measure their similarity. Following [29], a neighborhood matching network is introduced by [69] to match graph nodes by estimating similarities of their neighborhoods. Due to their superiority in finding consistent correspondences between sets of features, graph matching methods have been widely applied in various tasks, such as cross-lingual knowledge graph alignment [69, 71], object shape matching [73], video re-identification [74], 3D action recognition [19], *etc.* Motivated by these works, we resort to applying graph matching in the video grounding task by employing a cross-graph attention matching mechanism.

3. Methodology

3.1. Problem Formulation

Given an untrimmed video and a language query, the video grounding task aims to localize a temporal moment in the video that matches the query. Each video-language pair has one associated ground-truth moment, defined as a temporal interval with boundary (τ_s, τ_e), which best matches the query. Our method scores m candidate moments, where the k-th moment consists of start time $t_{s,k}$, end time $t_{e,k}$, and confidence score p_k.

The video stream is represented as a sequence of **snippets** $V = \{v_i\}_{i=1}^{n_v}$, where each snippet has ϵ consecutive frames and n_v is the number of snippets. Similarly, a language query is represented by n_l **tokens** $L = \{l_i\}_{i=1}^{n_l}$. The inputs to our VLG architecture are snippet features $X_v \in \mathbb{R}^{c_v \times n_v}$ and token features $X_l \in \mathbb{R}^{c_l \times n_l}$ extracted using pre-trained models, where c_v and c_l are the snippet and token feature dimensions. We describe the details of feature extraction in Sec. 4.2.

3.2. VLG-Net Architecture

Our video grounding architecture is illustrated in Fig. 2. First, we feed both the video features X_v and the language embeddings X_l into a stack of computation blocks, respectively. On the video path, we use 1D convolutions and GCNeXt [72] blocks to enrich the visual representation with local and non-local intra-modality context. On the language path, we apply LSTM and SyntacGCN to aggregate temporal and syntactic context, which models the grammatical structure of the language query. The two paths converge in the graph matching layer for cross-modal context modelling and multi-modal fusion. After the graph matching layer, we apply masked moment attention pooling to produce the representations of possible moment candidates. Finally, we use Multi-Layer Perceptron (MLP) to score the query-moment pair based on their representation and post-process the score by non-maximum suppression (NMS). We report top-κ ranked moments as the final predictions.

3.3. Video and Language Representations

Here, we detail the set of operations performed on each modality. The stack of computation blocks of each branch is specifically designed to model intra-modality context to enrich each snippet and token features.

Video Representation. We first apply 1D convolutions to map each input visual feature to a desired dimension. The video is then cast as a graph, where each node represents a snippet and each edge represents a dependency between a snippet pair. We design two types of edges: (i) *Ordering Edges* and (ii) *Semantic Edges*. Static *Ordering Edges* connect consecutive snippets and model the temporal order. Conversely, *Semantic Edges* are dynamically constructed to connect semantically similar snippets based on their feature representations. Specifically, an ordering or semantic snippet neighborhood is determined, and its aggregated representation is computed through edge convolutions $\mathcal{F}$, similar to [66]. Each edge convolution employs a split-transform-merge strategy [70] to increase the diversity of transformations. These graph operations (called GCNeXt) were proposed in [72] to enrich video snippet representations for the purpose of temporal action localization. In our architecture,

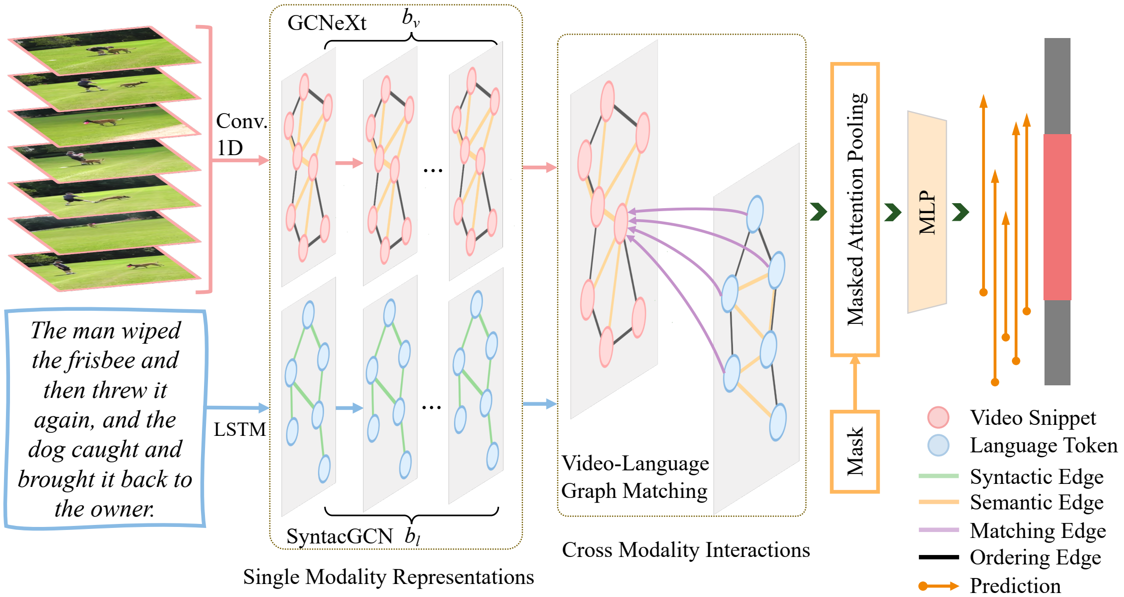

Conv. 1D
GCNeXt
b_v
...
LSTM
The man wiped the frisbee and then threw it again, and the dog caught and brought it back to the owner.
SyntacGCN b_l
Single Modality Representations
Video-Language Graph Matching
Cross Modality Interactions
Masked Attention Pooling
Mask
MLP
Video Snippet
Language Token
Syntactic Edge
Semantic Edge
Matching Edge
Ordering Edge
Prediction

Figure 2. **VLG-Net Architecture**. The inputs are snippet features and token embeddings. 1D convolutions and b_v GCNeXt operations are applied to the video snippets to enrich their representations. Correspondingly, tokens are fed to a stack of LSTM layers and b_l SyntacGCN layers. A graph matching layer is adopted for cross-modal context modelling and multi-modal fusion. Masked attention pooling lists all possible moment candidates, and a Multi-Layer Perceptron (MLP) computes each moment' score to rank them as final predictions.

we stack b_v GCNeXt blocks together and refer to the input of each block as $X_v^{(i)}$, where $X_v^{(0)}$ is the output of the convolutional layer. The GCNeXt block is formulated as:

$$X_v^{(i+1)} = \text{GCNeXt}(X_v^{(i)}) = \qquad (1)$$

$$\sigma\left(\mathcal{F}(X_v^{(i)}, \mathcal{A}_o^{(i)}, W_o^{(i)}) + \mathcal{F}(X_v^{(i)}, \mathcal{A}_s^{(i)}, W_s^{(i)}) + X_v^{(i)}\right),$$

where $\mathcal{A}_o^{(i)}$, and $\mathcal{A}_s^{(i)}$ are respectively the adjacency matrices of *Ordering Edges* and *Semantic Edges*, and $W_\cdot^{(i)}$ are the trainable weights for the i-th GCNeXt block. We use Rectified Linear Unit (ReLU) as the activation function σ. More details about this block can be found in [72]. The output of the last block is referred to as $X_v^{(b_v)}$, which is the input to the graph matching layer.

Language Representation. The query token features X_l are fed to a set of b_s LSTM layers to capture semantic information and the temporal context. Moreover, given that language follows a predefined set of grammatical rules, we set out to leverage syntactic information [40, 82] to model grammatical inter-word relations. For this purpose, we adopt SyntacGCN, as shown in Fig. 1b. Syntactic graphs are preferred over fully connected graphs, since the former's sparsity property makes them more robust against noise in language [21]. Our SyntacGCN represents a query as a sparse directed graph, in which each output of the last LSTM layer, referred to as $X_l^{(0)}$, is viewed as a node, and each syntactic relation as an edge. The adjacency matrix $\mathcal{A}_l$ is directly constructed from the query's syntactic dependencies [39] and the graph convolution can be formulated as in Eq. 2.

$$X_{l,j}^{(i+1)} = \sigma\left(X_{l,j}^{(i)} + \sum_{k\in\mathcal{N}(j)} \mathcal{A}_{l,jk} W_l^{(i)} \alpha_{jk}^{(i)} X_{l,k}^{(i)}\right) \qquad (2)$$

Here, $X_{l,j}^{(i)}$ is the j-th token feature from the i-th graph input, $\mathcal{N}(j)$ is the syntactic neighbourhood of node j, $W_l^{(i)}$ is the learnable weight in the i-th block, and σ is ReLU. Moreover, $\alpha_{jk}^{(i)}$ is the edge weight learned from the feature of paired nodes $X_{l,j}^{(i)}$ and $X_{l,k}^{(i)}$, shown in Eq. 3, where $\mathbf{w}_\alpha$ and W_α are learnable parameters and $||$ denotes vector concatenation.

$$\alpha_{jk}^{(i)} = \text{SoftMax}(\mathbf{w}_\alpha^{(i)\top} \sigma(W_\alpha^{(i)}(X_{l,j}^{(i)} || X_{l,k}^{(i)}))) \qquad (3)$$

We refer to the last output of the SyntacGCN as $X_l^{(b_l)}$, which will be used to match the video representation $X_v^{(b_v)}$.

3.4. Video-Language Graph Matching

The enriched video representation $X_v^{(b_v)}$ and language representation $X_l^{(b_l)}$ meet and interact in the graph matching layer. This layer is adopted for cross-modal context modelling and multi-modal fusion. It achieves this goal by evaluating the correlation between each snippet in the video and each token in the language by exploiting the semantic neighbourhood of snippets/tokens. The process is shown in

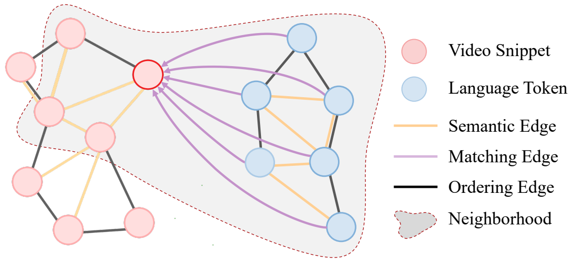


Figure 3. **The video-language matching graph.** The nodes represent video snippets and language tokens. Ordering Edge (black) models the sequential nature of both modalities. Semantic Edge (yellow) connects graph nodes in the same modality according to their feature similarity. Matching Edge (purple) captures the cross-modality relationships. We apply relation graph convolution on the constructed video-language graph for cross-modal context modelling and multi-modal fusion. The neighborhood is related to the node highlighted in red.

Fig. 3. First, we create a video-language matching graph, where each node can be either a video snippet or a language token. We include three types of edges: (i) *Ordering Edge* *($\mathcal{O}$)*, (ii) *Semantic Edge ($\mathcal{S}$)*, and (iii) *Matching Edge ($\mathcal{M}$)*.

As depicted in Fig 3, we use *Ordering Edge* and *Semantic Edge* on both video and language sub-graphs (as defined in Sec. 3.3), while the *Matching Edge* reflects the inter-modality relation. *Ordering Edge* models the sequential nature of both modalities. For example, if an *Ordering Edge* links two tokens, the words corresponding to the two tokens are consecutive in the input query. *Semantic Edge* is used to connect graph nodes in the same modality according to their feature similarity, providing non-local dependencies over the entire graph. Importantly, *Matching Edge* is employed to explicitly model and learn the cross-modality interaction, so as to extract meaningful alignment information and learn an aggregation policy. The *Matching Edge* weights are referenced as β. We use *Matching Edge* to densely connect all possible snippet-token pairs, and set the edge weight proportional to the correlation between the matched node features. Similar to *Semantic Edges*, *Matching Edges* are dynamic and evolve in the training process.

To combine all three types of edges, we employ relation graph convolution [50] on the constructed video-language matching graph. Eq. 4 explicitly shows the convolutions in this layer.

$$
\begin{aligned}
\mathbf{x}_i^{(GM)} = & \sum_{j\in\mathcal{N}_i^{\mathcal{O}}} (W_{\mathcal{O}}\mathbf{x}_j + W_0\mathbf{x}_i) \\
& + \sum_{j\in\mathcal{N}_i^{\mathcal{S}}} (W_{\mathcal{S}}\mathbf{x}_j + W_0\mathbf{x}_i) \\
& + \sum_{j\in\mathcal{N}_i^{\mathcal{M}}} (\beta_{i,j} W_{\mathcal{M}}\mathbf{x}_j + W_0\mathbf{x}_i)
\end{aligned}
\qquad (4)
$$

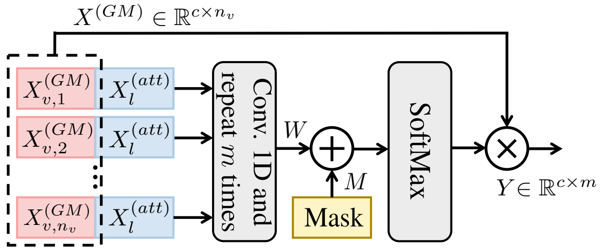


Figure 4. **Masked attention pooling.** Sequence of operations for the learnable cross-attention configuration. Inputs are video nodes $X_v^{(GM)}$ from the graph matching layer and the query embedding $X_l^{(att)}$ computed through self-attention pooling on top of the last SyntacGCN layer's output. The output Y represents all moment candidates.

Here, $\mathbf{x}_i \in \{X_{v,1}^{(b_v)}, \ldots, X_{v,n_v}^{(b_v)}, X_{l,1}^{(b_l)}, \ldots, X_{l,n_l}^{(b_l)}\}$ is the feature representation of the i-th node in the video-language matching graph. $\mathcal{N}_i^r$ collects all nodes connected to the i-th node by edge type $r \in \{\mathcal{O}, \mathcal{S}, \mathcal{M}\}$. $\beta_{i,j}$ is proportional to $\mathbf{x}_i^T\mathbf{x}_j$, and $W_.$ are the learnable weights.

Since $\mathbf{x}_i^{(GM)}$ is enriched with intra-modality and inter-modality context, we use it to represent the video-language relation for each snippet and token. As output, we stack all the $\mathbf{x}_i^{(GM)}$ to form $X^{(GM)} \in \mathbb{R}^{c\times n_v}$ and pass it to the masked moment pooling layer.

3.5. Masked Attention Pooling

The graph matching layer returns a new video graph and language graph fused with the other modality. Then, a masked attention pooling operation is applied to the new video graph to list the relevant sub-graph representations as candidate moments. The output of this module is denoted as $Y = [\mathbf{y}_k]_{k=1}^m, \mathbf{y}_k \in \mathbb{R}^c$, where m is the number of candidate moments, and c is the feature dimension of each moment. For efficiency purposes, the operation is implemented as a masked attention, allowing us to process each snippet feature only once while computing each moment's representation.

Specifically, we implement three different schemes, namely: (i) learnable self-attention, (ii) cross-attention, and (iii) learnable cross-attention. In (i), we obtain the unnormalized attention weights by applying a 1D convolutional layer that maps each snippet feature to a single score. In (ii) and (iii), we compute the query representation $X_l^{(att)}$ by applying self-attention pooling on top of the last SyntacGCN layer $X_l^{(b_l)}$. Cross-attention obtains the unnormalized weights by computing the inner product between the snippet and query feature, while learnable cross-attention concatenates each snippet feature with the query feature and uses a 1D convolutional layer to obtain the weights. Configuration (iii) is depicted in Fig. 4. In all cases, the unnormalized weight vector has shape $w \in \mathbb{R}^{n_v\times 1}$ for each video. The vector is repeated m times to obtain the matrix

$W \in \mathbb{R}^{n_v \times m}$, and a fixed mask $M \in \mathbb{R}^{n_v \times m}$ is applied to it. Similar to Songyang *et al.* [85], We generate moment candidates and apply a sparse sampling strategy to discard redundant moments. Therefore, the mask is generated according to the sampled moments, highlighting for each of them, which are the snippets that must be taken into account when computing the moment's pooled feature. The attention scores are then obtained by a softmax operation. Thanks to the masking operation, snippets not related to the n-th moment will not be considered. Finally, the moments' features are obtained simply as a matrix multiplication: $Y = X^{(GM)}\text{SoftMax}(W + M)$. Ablation results are reported in Supplementary Material.

3.6. Moment Localization

The output of the previous module is then fed to a Multi-Layer Perceptron (MLP) network to compute the score p_k for each moment candidate. This score predicts the Intersection-over-Union (IoU) of each moment with the ground truth one. For training, we supervise this process using a cross-entropy loss, shown in Eq. 5. We assign the label $t_k = 1$ if the IoU is greater than a threshold θ and $t_k = 0$ otherwise.

$$\mathcal{L} = \frac{1}{m}\sum_{k=1}^{m} t_k \log p_k + (1 - t_k)\log(1 - p_k), \qquad (5)$$

At inference time, moment candidates are ranked according to their predicted scores and non-maximum suppression is adopted to discard highly overlapping moments. The remaining top-κ moments are involved in the recall computation. The temporal boundaries $(t_{s,k}, t_{e,k})$ associated with the top-κ moments are used to calculate the Intersection-over-Union (IoU) with the ground-truth video moments (τ_s, τ_e) to determine the alignment performance. A formal definition of the metric used and details about the training strategy are presented in Sec. 4.2

4. Experiments

4.1. Datasets

ActivityNet-Captions [25] is a popular benchmark dataset initially collected for the task of dense captioning, and recently adopted for the task of moment localization with natural language [6, 31]. The dataset is subdivided into four splits: train, val_1, val_2, and test. The test set is withheld for competition purposes while leaving the rest publicly available. See Tab. 1 for details about the publicly available splits. Following the setting in [31], we use val_1 as the validation set and val_2 as the testing set.

TACoS [46] consists of videos selected from the MPII Cooking Composite Activities video corpus [48]. It is comprised of 18818 video-language pairs of different cooking activities in the kitchen. On average, every video in TACoS contains 148 queries, some of which are annotations of short video segments.

Dataset	Num. Videos	Video-Sentence pairs train	val	test	Vocab. Size
Activitynet-Captions [25]	14926	37421	17505	17031	15406
TACoS [46]	127	10146	4589	4083	2255
DiDeMo [1]	10642	33005	4180	4021	7523
Charades-STA [14]	6670	12404	0	3720	1289

Table 1. **Video-language grounding dataset statistics.**

DiDeMo [1] contains unedited video footage from Flickr with sentences aligned to unique moments in its 10642 videos. It contains more than 40k video-language pairs with coarse temporal annotations. Moment start and end points are aligned to five-second intervals and the maximum annotated moment length is 30 seconds.

Concerns related to Charades-STA [14] dataset led us to the decision of not evaluating our method on it. Limitations include the absence of a validation set, reduced vocabulary size, and a smaller number of video-language pairs compared to the other datasets (Tab. 1). See Supplementary Material for more discussion.

4.2. Implementation

Evaluation Metrics. We follow the commonly used setting in [12], where the Rank@κ for IoU=θ serves as our evaluation metric. For example, given a video-query pair, the result is positive if any of the top-κ predictions has IoU with the ground-truth larger or equal to θ; otherwise the result is negative. We average the results across all testing samples. Specifically, we set $\kappa \in \{1, 5\}$ with $\theta \in \{0.3, 0.5, 0.7\}$ for ActivityNet Captions, $\kappa \in \{1, 5\}$ with $\theta \in \{0.1, 0.3, 0.5\}$ for TACoS, and $\kappa \in \{1, 5\}$ with $\theta \in \{0.5, 0.7, 1.0\}$ for DiDeMo.

Language and Video Features. After lower-case conversion and tokenization, we use the pretrained GloVe model [45] to obtain the initial language embedding for every token and extract the syntactic dependencies using the Stanford CoreNLP 4.0.0 parser [39]. The b_s layers of LSTM with 512 hidden units are used as the query encoder. Then, the syntactic GCN encodes syntactic information of the queries and returns a new embedding with 512 dimensions. For visual features, we use pretrained C3D [58] for ActivityNet Captions and TACoS, and VGG16 [54] for DiDeMo, while holding their parameters fixed during training, as they are readily available and commonly used by state-of-the-art methods. We use 1D convolutions to project the input visual features to a fixed dimension (512), and the hyper-parameters of GCNeXt blocks are the same as in [72].

Training and Inference. We use Adam [24] with a CosineAnnealingLR scheduler [37]. We adopt learning rates ranging from 10^{-3} to 10^{-4} for different datasets. The number of sampled snippets n_v is set to 64 for ActivityNet

Captions, and 256 for TACoS, and 18 for DiDeMo. In post-processing, we apply NMS to the m predictions to filter out highly overlapping moments.

4.3. Comparison with State-of-the-Art

ActivityNet Captions. In Table 2, we report the performance of our method in comparison with several recently proposed methods. VLG-Net reaches the highest scores for all evaluation metrics except for R@5 IoU=0.5. For R@1 IoU=0.7, VLG-Net achieves the improvement of 2.04 over the previous best method.

	R@1		R@5	
	IoU0.5	IoU0.7	IoU0.5	IoU0.7
MCN [1]	21.36	6.43	53.23	29.70
CTRL [14]	29.01	10.34	59.17	37.54
TGN [6]	27.93	–	44.20	–
ACRN [32]	31.67	11.25	60.34	38.57
CMIN [31]	44.62	24.48	69.66	52.96
QSPN [22]	33.26	13.43	62.39	40.78
ABLR [75]	36.79	–	–	–
TripNet [42]	32.19	13.93	–	–
PMI [53]	38.28	17.83	–	–
2D-TAN [85]	44.05	27.38	76.65	62.26
2D-TAN [85]	44.51	26.54	77.13	61.96
DRN [79]	45.45	24.39	**77.97**	50.30
VLG-Net	**46.74**	**29.42**	76.42	**63.17**

Table 2. **State-of-the-art comparison on ActivityNet Captions.** We report the results at different Recall@k and different IoU thresholds. VLG-Net reaches the highest scores for IoU0.7 for both R@1 and R@5. The top-2 performance values are highlighted by bold and underline, respectively.

	R@1			R@5		
	IoU0.1	IoU0.3	IoU0.5	IoU0.1	IoU0.3	IoU0.5
MCN [1]	14.42	–	5.58	37.35	–	10.33
CTRL [14]	24.32	18.32	13.30	48.73	36.69	25.42
MCF [68]	25.84	18.64	12.53	52.96	37.13	24.73
TGN [6]	41.87	21.77	18.9	53.40	39.06	31.02
ACRN [33]	24.22	19.52	14.62	47.42	34.97	24.88
ROLE [34]	20.37	15.38	9.94	45.45	31.17	20.13
VAL [56]	25.74	19.76	14.74	51.87	38.55	26.52
ACL-K [16]	31.64	24.17	20.01	57.85	42.15	30.66
CMIN [31]	36.68	27.33	19.57	64.93	43.35	28.53
QSPN [22]	25.31	20.15	15.23	53.21	36.72	25.30
SM-RL [64]	26.51	20.25	15.95	50.01	38.47	27.84
SLTA [23]	23.13	17.07	11.92	46.52	32.90	20.86
ABLR [75]	34.70	19.50	9.40	–	–	–
SAP [7]	31.15	–	18.24	53.51	–	28.11
TripNet [42]	–	23.95	19.17	–	–	–
2D-TAN [85]	47.59	37.29	25.32	70.31	57.81	45.04
2D-TAN [85]	46.44	35.22	25.19	74.43	56.94	44.21
DRN [79]	–	–	23.17	–	–	33.36
VLG-Net	**55.26**	**42.94**	**31.39**	**77.41**	**63.66**	**51.71**

Table 3. **State-of-the-art comparison on TACoS.** Our model outperforms all previous methods achieving significantly higher performance with great margins on all metrics.

TACoS. Evaluation on TACoS dataset using the same input features is illustrated in Table 3. Our model outperforms all the competing methods and achieves the highest scores for all IoU thresholds with significant improvement. Specifically, VLG-Net exceeds the previously best method 2D-TAN [85] by a margin ranging from 2.98 to 7.67 across all evaluation settings.

DiDeMo. Table 4 shows the video grounding results of our VLG-Net as compared to state-of-the-art methods using the same VGG features. Our proposed technique outperforms the top ranked methods ROLE and ACRN with respect to R@1 and R@5 for IoU=0.5 and 0.7 with evident increases. It also reaches the highest performance in regards to R@1 IoU1.0. For completeness, we report the performances of TGN [32] and TMN [6]; with the caveat that their performance could not be verified as the code is not publicly available.

	R@1			R@5		
	IoU0.5	IoU0.7	IoU1.0	IoU0.5	IoU0.7	IoU1.0
MCN [1]	–	–	13.10	–	–	44.82
TMN [32]	–	–	18.71	–	–	**72.97**
TGN [6]	–	–	24.28	–	–	71.43
ACRN [33]	27.44	16.65	–	69.43	29.45	–
ROLE [34]	29.40	15.68	–	70.72	33.08	–
VLG-Net	**32.28**	**24.82**	**24.75**	**84.80**	**68.09**	67.79

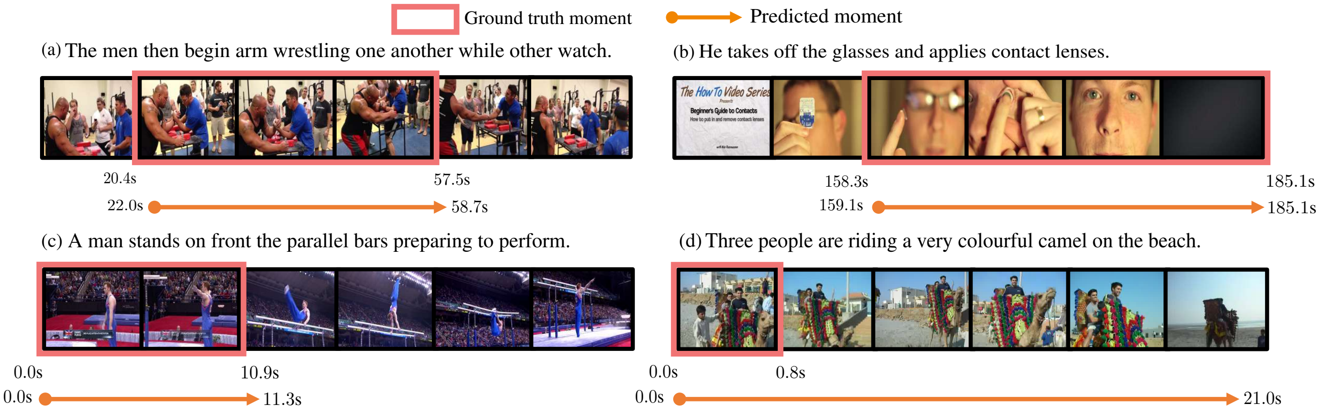
Table 4. **State-of-the-art comparison on DiDeMo.** Our proposed model outperforms the top ranked method ROLE and ACRN with respect to IoU0.5 and 0.7 for R@1 and R@5 with clear margins. It also reaches the highest performance in regards to R@1 IoU1.0.

4.4. Ablation Study

To evaluate the effectiveness of every module/component in VLG-Net and to justify our design choices, we perform several ablation experiments.

First, we investigate the impact of the architecture's main modules, including GCNeXt for video modeling, SyntacGCN for language modeling, and graph matching for cross-modality interactions (Tab. 5). In this experiment, we train four models on two datasets, by ablating modules one-by-one and focus on R@1 IoU=0.7 as being the most challenging metric. By removing the GCNeXt module, the video intra-modality context is not modelled as a graph, leading to the drop of performance, thus motivating our choice for intra-modality context gathering for videos via graphs. Similarly, ablating SyntacGCN results in inferior performance. Notably, when the graph matching module is eliminated, the performance is severely impaired, indicating its crucial importance. Combining all these modules together registers the best performance, which attests to the superiority of our proposed VLG-Net architecture in modeling intra- and inter- interactions between video snippets and language queries.

We also analyze the contributions of different types of



(a) The men then begin arm wrestling one another while other watch.

(b) He takes off the glasses and applies contact lenses.

(c) A man stands on front the parallel bars preparing to perform.

(d) Three people are riding a very colourful camel on the beach.

Figure 5. **Qualitative Results.** Examples of grounding results, we compare ground truth annotations (box) and predicted temporal endpoints (arrow). See Section 4.5 for more details.

Dataset	Modules			R@1 IoU0.7
	GCNeXt.	SyntacGCN.	Graph Matching	
ActivityNet-Captions	✓	✓	✓	**29.42**
	✗	✓	✓	28.16
	✓	✗	✓	28.41
	✓	✓	✗	14.34
TACoS	✓	✓	✓	**31.39**
	✗	✓	✓	29.87
	✓	✗	✓	30.49
	✓	✓	✗	7.42

Table 5. **Ablation of main modules.** We report the performance of our architecture when specific modules are removed, as well as our best performance for ActivityNet-Captions and TACoS datasets.

Dataset	Edge Types			R@1 IoU0.7
	Ordering	Semantic	Matching	
ActivityNet-Captions	✓	✓	✓	**29.42**
	✗	✓	✓	28.89
	✓	✗	✓	29.18
	✓	✓	✗	13.87
TACoS	✓	✓	✓	**31.39**
	✗	✓	✓	29.12
	✓	✗	✓	28.44
	✓	✓	✗	6.12

Table 6. **Ablation of different edges.** We investigate the impact of edges within the graph matching layer. We report the performance of our VLG-Net when specific edges are removed, as well as our best performance for ActivityNet-Captions and TACoS datasets.

edges considered in the graph matching module for two datasets, as shown in Tab. 6. For each dataset, removing Ordering Edges or Semantic Edges leads to inferior results. This supports the notion that temporal local and non-local relationships modelling is beneficial for snippets/tokens features enrichment. Importantly, excluding Matching Edges impairs the performance significantly. This is because this setup prevents the exchange of information between video and language in the graph matching layer, while the two modalities only interact in the moment attention pooling module. Again, the best performance is obtained when all three types of edges are considered, thus justifying our design for the multi-modal fusion operation.

4.5. Visualization

We show several qualitative grounding results from ActivityNet-Captions in Fig. 5. Our VLG-Net can generate precise moment boundaries that match the query well in different scenarios.

Worth mentioning, our method can sometimes give predictions that are more meaningful than the ground truth annotation. As shown in Fig. 5(d), although the ground truth aligns to the very beginning of the video only, the query “Three people are riding a very colourful camel on the beach” can semantically match the whole video. Obviously, our VLG-Net, in this case, gives a more reasonable grounding result.

5. Conclusion

In this paper, we address the problem of text-to-video temporal grounding and we recast it into an algorithmic graph matching problem. We formulate the untrimmed video and language query as video and language graphs respectively, where video snippets and language tokens are graph nodes, and snippet/token correlations are edges. We propose Video-Language Graph Matching Network (VLG-Net) to match the video graph and language graph through GCN, and explore four type of edges, Syntactic Edge, Ordering Edge, Semantic Edge, and Matching Edge, that encode local, non-local, and cross modality relationships for finer alignment of the video-query pair. Extensive experiments show that our VLG-Net can efficiently model inter- and intra-modality context, learn the best way for multi-modal fusion, and surpass the current state-of-the-art performance on three widely adopted datasets: ActivityNet-Captions, TACoS, and DiDeMo.

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 (ICCV)*, 2017. 1, 2, 6, 7

[2] Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. Graph convolutional encoders for syntax-aware neural machine translation. *arXiv preprint arXiv:1704.04675*, 2017. 3

[3] Daniel Beck, Gholamreza Haffari, and Trevor Cohn. Graph-to-sequence learning using gated graph neural networks. *arXiv preprint arXiv:1806.09835*, 2018. 3

[4] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. SST: Single-Stream Temporal Action Proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[5] Tibério S Caetano, Julian J McAuley, Li Cheng, Quoc V Le, and Alex J Smola. Learning graph matching. *IEEE transactions on pattern analysis and machine intelligence*, 31(6):1048–1058, 2009. 3

[6] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally Grounding Natural Sentence in Video. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 2, 6, 7

[7] 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*, 2019. 2, 7

[8] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[9] Chen Shaoxiang, Jiang Yu-Gang. Hierarchical Visual-Textual Graph for Temporal Activity Localization via Language. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[10] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S. Davis, and Yan Qiu Chen. Temporal Context Network for Activity Localization in Videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[11] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[12] 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 (ICCV)*, 2017. 6

[13] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[14] Gao Jiyang, Sun Chen, Yang Zhenheng, Nevatia, Ram. TALL: Temporal Activity Localization via Language Query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 6, 7

[15] R. Ge, J. Gao, K. Chen, and R. Nevatia. MAC: Mining Activity Concepts for Language-Based Temporal Localization. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. 1, 2

[16] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. MAC: Mining Activity Concepts for Language-based Temporal Localization. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. 2, 7

[17] Spandana Gella, Mike Lewis, and Marcus Rohrbach. A dataset for telling the stories of social media videos. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 1

[18] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. 2

[19] Michelle Guo, Edward Chou, De-An Huang, Shuran Song, Serena Yeung, and Li Fei-Fei. Neural graph matching networks for fewshot 3d action recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 653–669, 2018. 3

[20] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-Aware Graph Convolutional Networks for Video Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 1

[21] Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. Aligned Dual Channel Graph Convolutional Network for Visual Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. 3, 4

[22] Xu Huijuan, He Kun, Sigal Leonid, Sclaroff Stan, and Saenko Kate. Multilevel Language and Vision Integration for Text-to-Clip Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 2, 7

[23] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. Cross-Modal Video Moment Retrieval with Spatial and Language-Temporal Attention. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval (ICMR)*, 2019. 1, 7

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[25] 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 (ICCV)*, 2017. 6

[26] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. TVQA: Localized, Compositional Video Question Answering. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 1

[27] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. TVR: A Large-Scale Dataset for Video-Subtitle Moment Re-

trieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[28] Lei Li, Li Jin, Zequn Zhang, Qing Liu, Xian Sun, and Hongqi Wang. Graph Convolution Over Multiple Latent Context-Aware Graph Structures for Event Detection. *IEEE Access*, 2020. 3

[29] Yujia Li, Chenjie Gu, Thomas Dullien, Oriol Vinyals, and Pushmeet Kohli. Graph matching networks for learning the similarity of graph structured objects. *arXiv preprint arXiv:1904.12787*, 2019. 2, 3

[30] 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 (ECCV)*, 2018. 1

[31] Zhijie Lin, Zhou Zhao, Zhu Zhang, Zijian Zhang, and Deng Cai. Moment Retrieval via Cross-Modal Interaction Networks With Query Reconstruction. *IEEE Transactions on Image Processing*, 2020. 2, 3, 6, 7

[32] Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Temporal Modular Networks for Retrieving Complex Compositional Activities in Videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 7

[33] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive Moment Retrieval in Videos. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018. 2, 7

[34] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-Modal Moment Localization in Videos. In *Proceedings of the 26th ACM International Conference on Multimedia*, 2018. 1, 2, 7

[35] Xingyu Liu, Joon-Young Lee, and Hailin Jin. Learning Video Representations from Correspondence Proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[36] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, Xiaolin Li. Rethinking the Bottom-Up Framework for Query-based Video Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 2

[37] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with warm Restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[38] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. DEBUG: A Dense Bottom-Up Grounding Approach for Natural Language Video Localization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 2

[39] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014. 2, 4, 6

[40] Diego Marcheggiani and Ivan Titov. Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017. 2, 3, 4

[41] Diego Marcheggiani and Ivan Titov. Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826*, 2017. 3

[42] Meera Hahn, Asim Kadav, James M. Rehg, and Hans Peter Graf. Tripping through time: Efficient Localization of Activities in Videos. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020. 7

[43] Gaku Morio and Katsuhide Fujita. Syntactic graph convolution in multi-task learning for identifying and classifying the argument component. In *Proceeding of the International Conference on Semantic Computing (ICSC)*. IEEE, 2019. 3

[44] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-Global Video-Text Interactions for Temporal Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[45] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 6

[46] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding Action Descriptions in Videos. *Transactions of the Association for Computational Linguistics (ACL)*, 2013. 6

[47] Rodriguez Cristian, Marrese-Taylor Edison, Saleh Fatemeh Sadat, Li Hongdong, Gould Stephen. Proposal-free Temporal Moment Localization of a Natural-Language Query in Video using Guided Attention. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2

[48] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012. 6

[49] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985. 3

[50] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling Relational Data with Graph Convolutional Networks, 2017. 5

[51] Hochreiter Sepp and Schmidhuber Jürgen. Long Short-Term Memory. *Neural Computation*, 1997. 2

[52] Dian Shao, Yu Xiong, Yue Zhao, Qingqiu Huang, Yu Qiao, and Dahua Lin. Find and Focus: Retrieve and Localize Video Events with Natural Language Queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[53] Shaoxiang Chen, Wenhao Jiang, Wei Liu, Yu-Gang Jiang. Learning Modality Interaction for Temporal Sentence Localization and Event Captioning in Videos, 2020. 2, 7

[54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[55] Xiaomeng Song and Yahong Han. VAL: Visual-Attention Action Localizer. In Richang Hong, Wen-Huang Cheng, Toshihiko Yamasaki, Meng Wang, and Chong-Wah Ngo, editors, *Proceedings of the Advances in Multimedia Information Processing (PCM)*, 2018. 1, 2, 3

[56] Xiaomeng Song and Yahong Han. VAL: Visual-Attention Action Localizer. In *Proceedings of the Advances in Multimedia Information Processing (PCM)*, 2018. 7

[57] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks, 2017. 2

[58] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 2015. 6

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 2

[60] Escorcia Victor, Soldan Mattia, Sivic Josef, Ghanem Bernard, and Russell Bryan. Temporal Localization of Moments in Video Collections with Natural Language, 2019. 1

[61] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally Grounding Language Queries in Videos by Contextual Boundary-aware Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 2

[62] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1

[63] Runzhong Wang, Junchi Yan, and Xiaokang Yang. Learning combinatorial embedding networks for deep graph matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3056–3065, 2019. 3

[64] Weining Wang, Yan Huang, and Liang Wang. Language-driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[65] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[66] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay Sarma, Michael Bronstein, and Justin Solomon. Dynamic Graph CNN for Learning on Point Clouds. *ACM Transactions on Graphics*, 2018. 3

[67] Aming Wu and Yahong Han. Multi-modal Circulant Fusion for Video-to-Language and Backward. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, 2018. 1, 2

[68] Aming Wu and Yahong Han. Multi-modal Circulant Fusion for Video-to-Language and Backward. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, 2018. 7

[69] Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao. Neighborhood matching network for entity alignment. *arXiv preprint arXiv:2005.05607*, 2020. 3

[70] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[71] Kun Xu, Liwei Wang, Mo Yu, Yansong Feng, Yan Song, Zhiguo Wang, and Dong Yu. Cross-lingual knowledge graph alignment via graph matching neural network. *arXiv preprint arXiv:1905.11605*, 2019. 2, 3

[72] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, and Bernard Ghanem. G-TAD: Sub-Graph Localization for Temporal Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 4, 6

[73] Junchi Yan, Minsu Cho, Hongyuan Zha, Xiaokang Yang, and Stephen M Chu. Multi-graph matching via affinity optimization with graduated consistency regularization. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1228–1242, 2015. 3

[74] Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C Yuen. Dynamic label graph matching for unsupervised video re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5142–5150, 2017. 3

[75] Yuan Yitian, Mei Tao, and Zhu Wenwu. To Find Where You Talk: Temporal Sentence Localization in Video with Attention Based Location Regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 2, 7

[76] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 2

[77] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[78] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph Convolutional Networks for Temporal Action Localization. *arXiv preprint arXiv:1909.03252*, 2019. 3

[79] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense Regression Network for Video Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 7

[80] Junchi Zhang, Qi He, and Yue Zhang. Syntax grounded graph convolutional network for joint entity and event extraction. *Neurocomputing*, 2020. 3

[81] Yuhao Zhang, Peng Qi, and Christopher D Manning. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 3

[82] Yuhao Zhang, Peng Qi, and Christopher D. Manning. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. 4

[83] Zhang Da, Dai Xiyang, Wang Xin, Wang Yuan-Fang, and Davis Larry S. MAN: Moment Alignment Network for Natural Language Moment Retrieval via Iterative Graph Adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[84] Zhang Hao, Sun Aixin, Jing Wei, Zhou Joey Tianyi. Span-based Localizing Network for Natural Language Video Localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 2

[85] Zhang Songyang, Peng Houwen, Fu Jianlong, Luo, Jiebo. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 2, 6, 7

[86] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-Language Navigation With Self-Supervised Auxiliary Reasoning Tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1