Exploring Motion and Appearance Information for Temporal Sentence Grounding

Daizong Liu[1,2], **Xiaoye Qu**[3], **Pan Zhou**[1*], **Yang Liu**[4]
[1]The Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering,
Huazhong University of Science and Technology
[2]School of Electronic Information and Communication, Huazhong University of Science and Technology
[3]Huawei Cloud [4]Harbin Institute of Technology
{dzliu, panzhou}@hust.edu.cn, quxiaoye@huawei.com, liu.yang@hit.edu.cn

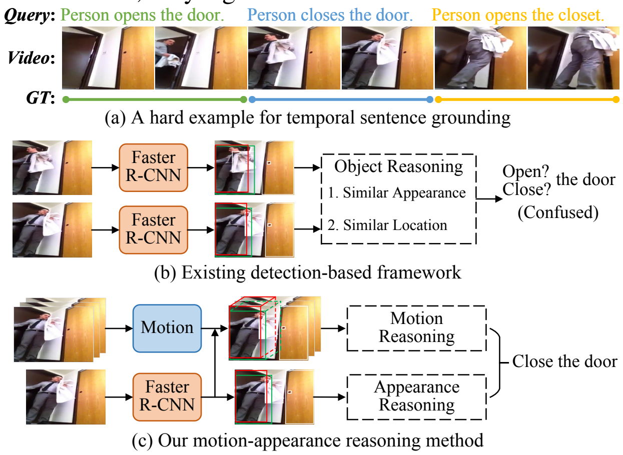
Abstract

This paper addresses temporal sentence grounding. Previous works typically solve this task by learning frame-level video features and align them with the textual information. A major limitation of these works is that they fail to distinguish ambiguous video frames with subtle appearance differences due to frame-level feature extraction. Recently, a few methods adopt Faster R-CNN to extract detailed object features in each frame to differentiate the fine-grained appearance similarities. However, the object-level features extracted by Faster R-CNN suffer from missing motion analysis since the object detection model lacks temporal modeling. To solve this issue, we propose a novel **M**otion-**A**ppearance **R**easoning **N**etwork (MARN), which incorporates both motion-aware and appearance-aware object features to better reason object relations for modeling the activity among successive frames. Specifically, we first introduce two individual video encoders to embed the video into corresponding motion-oriented and appearance-aspect object representations. Then, we develop separate motion and appearance branches to learn motion-guided and appearance-guided object relations, respectively. At last, both motion and appearance information from two branches are associated to generate more representative features for final grounding. Extensive experiments on two challenging datasets (Charades-STA and TACoS) show that our proposed MARN significantly outperforms previous state-of-the-art methods by a large margin.

Introduction

Temporal sentence grounding is an important topic of cross-modal understanding in computer vision. Given an untrimmed video, it aims to locate a segment that contains the interested activity corresponding to the sentence description. There are several related tasks proposed involving both video and language, such as video summarization (Song et al. 2015; Chu, Song, and Jaimes 2015), video question answering (Gao et al. 2019; Le et al. 2020), and temporal sentence grounding (Gao et al. 2017; Anne Hendricks et al. 2017). Among them, temporal sentence grounding is the most challenging task due to its detailed multi-modal interaction and complicated context reasoning.

To localize the target segment, most previous works (Anne Hendricks et al. 2017; Gao et al. 2017; Chen et al. 2018; Zhang et al. 2019; Liu et al. 2020a,b; Zhang et al. 2020b; Liu et al. 2021a; Liu, Qu, and Zhou 2021) first pre-define abundant video segments as proposals, and then match them with the sentence query for ranking. The best segment proposal with the highest matching degree is finally selected as the target segment. Instead of proposal-based paradigm, some proposal-free works (Rodriguez et al. 2020; Chen et al. 2020; Yuan, Mei, and Zhu 2019; Mun, Cho, and Han 2020; Zhang et al. 2020a) propose to directly regress the start and end timestamps of the target segment for each frame. Although above two kinds of works bring significant improvements in recent years, all of them extract frame-level video features to model the semantic of the target activity, which captures the redundant background information and fails to perceive the fine-grained differences among video frames with high similarity. As shown in Figure 1 (a), for the semantically similar queries "Person opens the door" and "Person opens the closet", modeling the temporal relations by frame-wise features can capture the same action "open", but it is not enough to distinguish the local details of different objects ("door" and "closet") in these frames.



(a) A hard example for temporal sentence grounding

(b) Existing detection-based framework

(c) Our motion-appearance reasoning method

Figure 1: (a) A hard example where the video contains several semantically similar segments. (b) Existing detection-based framework only extracts appearance-aware object information, and fails to distinguish similar motions "open" and "close". (c) Our method develops additional branch to learn motion-aware object contexts, and associates the information of two branches to reason the query.

*Corresponding author.

Recently, detection-based approaches (Zeng et al. 2021; Zhang et al. 2020c,d) have been proposed to capture fine-grained object appearance features in each frame and achieved promising results. Among them, (Zeng et al. 2021) focus on temporal sentence grounding and learn spatio-temporal object relations to reason the semantic of target activity, while other works consider spatio-temporal object grounding task where an object rather than a video segment is retrieved. These works can well alleviate the issue of indistinguishable local appearances, such as "door" and "closet", by learning the representations of objects in the frame. However, methods like (Zeng et al. 2021) generally extract object features by the object detection model like Faster R-CNN (Ren et al. 2015), which lacks the object-level motion context to model the temporal action of a specific object, thus degenerating the performance on similar events. As shown in Figure 1 (b), it is hard for detection-based methods to distinguish the similar motions "open" and "close" by learning the object relations in the successive frames, since the objects extracted by Faster R-CNN have similar appearance and spatial positions in these frames. Therefore, the motion context plays an important role in modeling the consecutive states or actions for objects. **How to effectively integrate the action knowledge from motion contexts and the appearance knowledge from detection model to compose the complicated activity is an emerging issue.**

To this end, in this paper, we propose a novel **M**otion-**A**ppearance **R**easoning **N**etwork (MARN), which incorporates motion contexts into appearance-based object features for better reasoning the semantic relations among objects. Specifically, we detect and obtain appearance-aware object representations by a Faster R-CNN model, and simultaneously apply RoIAlign (He et al. 2017) on the 3D feature maps from C3D network (Tran et al. 2015) for motion-aware object features extraction. Then, we develop separate branches to reason the motion-guided and appearance-guided object relations, respectively. In each branch, we interact object features with query information for query-related object semantic learning and adopt a fully-connected object graph for spatio-temporal semantic reasoning. At last, we represent frame-level features by aggregating object features inside the frame, and introduce a motion-appearance associating module to integrate representative information from two branches for final grounding.

The main contributions of this work are three-fold:

- To the best of our knowledge, we are the first work that explores both motion-aware and appearance-aware object information, and proposes a novel Motion-Appearance reasoning network for temporal sentence grounding.
- We devise motion and appearance branches to capture action-oriented and appearance-guided object relations. A motion-appearance associating module is further proposed to integrate the most representative features from two branches for final grounding.
- We conduct extensive experiments on two challenging datasets Charades-STA and TACoS. The experimental results show that our proposed MARN outperforms other state-of-the-art approaches with a large margin.

Related Work

Temporal Sentence Grounding. The task of temporal sentence grounding is introduced by (Gao et al. 2017) and (Anne Hendricks et al. 2017), which aims to identify the start and end timestamps of one specific video segment semantically corresponding to the given sentence query. Most previous works (Anne Hendricks et al. 2017; Chen et al. 2018; Zhang et al. 2019; Yuan et al. 2019; Zhang et al. 2020b; Qu et al. 2020; Liu et al. 2022a) localize the target segment via generating video segment proposals. They utilize sliding windows or pre-defined segment proposals to generate segment candidates, and then match them with the query. Instead of using segment candidates, some works (Chen et al. 2020; Yuan, Mei, and Zhu 2019; Mun, Cho, and Han 2020; Zhang et al. 2020a; Liu et al. 2022b) directly regress the start and end timestamps after interacting the whole video with query. However, these two types of methods are all based on frame-level features to capture the semantic of video activity, which fails to capture the fine-grained difference among video frames with high similarity, especially the adjacent frames near the segment boundary. Recently, detection-based approaches (Zeng et al. 2021; Zhang et al. 2020c,d) have been proposed to capture object appearances in the frame, which leads to more precise localization. However, they only adopt object features extracted by detection models, thus cannot obtain the motion information of each object. As for (Zeng et al. 2021), it builds dual textual and visual object graphs for cross-modal graph matching and directly utilize object features for grounding. In this paper, we only construct visual graph and interact object features with textual information for semantic enhancement, then integrate object features inside each frame to represent frame-level features for grounding. Moreover, considering motion contexts play a key role (Seo et al. 2021) in modeling the consecutive states or actions, we extract both motion-aware and appearance-aware object features to capture action-oriented and appearance-aspect contexts for more accurate spatio-temporal object reasoning.

Model

Given an untrimmed video V and a sentence query Q, we present the video as $V = \{v_t\}_{t=1}^{T}$ where v_t denotes the t-th frame and T denotes the frame number. Analogously, the sentence query is represented as $Q = \{q_n\}_{n=1}^{N}$ where q_n denotes the n-th word and N denotes the total number of words. The temporal sentence grounding (TSG) task aims to retrieve the video segment including both start and end timestamps that is most relevant to the sentence query.

In this section, we propose the Motion-Appearance Reasoning Network (MARN), which incorporates both motion and appearance information to reason object relations for modeling the activity among successive frames. As shown in Figure 2, we first extract motion- and appearance-aware object features, and then develop separate motion and appearance branches to learn fine-grained action-oriented and appearance-aspect object relations. Specifically, in each branch, after interacting object-query features to filter out irrelevant object features, we reason the relations between

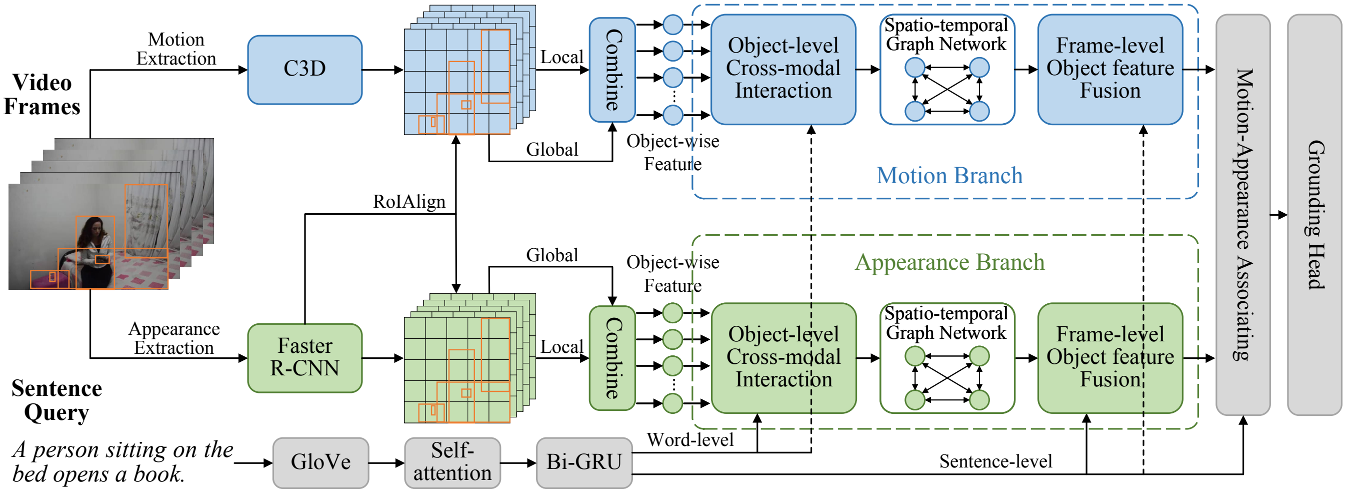

Video Frames
Motion Extraction
C3D
Local
Global
Combine
Object-wise Feature
Object-level Cross-modal Interaction
Spatio-temporal Graph Network
Frame-level Object feature Fusion
Motion Branch
RoIAlign
Appearance Branch
Appearance Extraction
Faster R-CNN
Local
Global
Combine
Object-wise Feature
Object-level Cross-modal Interaction
Spatio-temporal Graph Network
Frame-level Object feature Fusion
Motion-Appearance Associating
Grounding Head
Sentence Query
A person sitting on the bed opens a book.
GloVe
Self-attention
Bi-GRU
Word-level
Sentence-level

Figure 2: Overall pipeline of the proposed network MARN. We first utilize video and query encoders to extract motion- and appearance-aware object features, word- and sentence-level query features. Then, we develop separate motion and appearance branches for specific cross-modal object reasoning. At last, we associate motion-appearance information for final grounding.

foremost objects with a spatio-temporal graph and represent frame-level features by fusing its contained objects features. At last, we associate motion and appearance frame-level features for final grounding.

Video and Query Encoders

Video encoder. Unlike previous detection-based methods only using Faster R-CNN (Ren et al. 2015) pre-trained on image detection dataset to extract appearance-aware object features, we consider additionally extracting motion-aware object information to obtain action-oriented features for temporal modeling. Specifically, **for appearance features**, we first sample fixed T frames from the original video averagely, and then obtain K objects from each frame using Faster R-CNN that is built on a ResNet (He et al. 2016) backbone. Therefore, there are total $T \times K$ objects in a single video, and we can represent their appearance features as $\boldsymbol{V}^a_{local} = \{\boldsymbol{o}^a_{t,k}, \boldsymbol{b}_{t,k}\}^{t=T,k=K}_{t=1,k=1}$, where $\boldsymbol{o}^a_{t,k} \in \mathbb{R}^D, \boldsymbol{b}_{t,k} \in \mathbb{R}^4$ denotes the local appearance feature and bounding-box position of the k-th object in t-th frame. Since the global feature of the whole frame also contains the non-local information of its internal objects, we utilize another ResNet model with a linear layer to generate frame-wise appearance representation $\boldsymbol{V}^a_{global} \in \mathbb{R}^{T \times D}$. **For motion features**, we first extract the feature maps of each video clip from the last convolutional layer in C3D (Tran et al. 2015) network, and then apply RoIAlign (He et al. 2017) on such feature maps and use object bounding-box locations $\boldsymbol{b}_{t,k}$ to generate motion-aware object features $\boldsymbol{V}^m_{local} = \{\boldsymbol{o}^m_{t,k}, \boldsymbol{b}_{t,k}\}^{t=T,k=K}_{t=1,k=1}$. To extract the clip-wise global features $\boldsymbol{V}^m_{global} \in \mathbb{R}^{T \times D}$, we directly apply average pooling and linear projection to the extracted feature maps of C3D.

Since it is necessary to consider each object's both spatial and temporal locations for reasoning object-wise relations, we add a position encoding to object-level local features in both appearance and motion representations as:

$$\boldsymbol{v}^a_{t,k} = \text{FC}([\boldsymbol{o}^a_{t,k}; \boldsymbol{e}^b; \boldsymbol{e}^t]), \boldsymbol{v}^m_{t,k} = \text{FC}([\boldsymbol{o}^m_{t,k}; \boldsymbol{e}^b; \boldsymbol{e}^t]), \quad (1)$$

where $\boldsymbol{e}^b = \text{FC}(\boldsymbol{b}_{t,k})$, $\text{FC}(\cdot)$ is the fully connected layer, $\boldsymbol{e}^t$ is obtained by position encoding (Mun, Cho, and Han 2020) according to each frame's index. Thus, $\widehat{\boldsymbol{V}}^a_{local} = \{\boldsymbol{v}^a_{t,k}\}^{t=T,k=K}_{t=1,k=1}$, $\widehat{\boldsymbol{V}}^m_{local} = \{\boldsymbol{v}^m_{t,k}\}^{t=T,k=K}_{t=1,k=1}$. Similarly, we add position encoding into two global representations as:

$$\widehat{\boldsymbol{V}}^a_{global} = \text{FC}([\boldsymbol{V}^a_{global}; \mathbf{e}^T]), \widehat{\boldsymbol{V}}^m_{global} = \text{FC}([\boldsymbol{V}^m_{global}; \mathbf{e}^T]). \quad (2)$$

At last, we expand the above two global features from size of $T \times D$ to size of $(T \times K) \times D$, and concatenate the local object features with corresponding global features to reflect the context in objects as:

$$\boldsymbol{F}^a = \text{FC}([\widehat{\boldsymbol{V}}^a_{local}; \widehat{\boldsymbol{V}}^a_{global}]), \boldsymbol{F}^m = \text{FC}([\widehat{\boldsymbol{V}}^m_{local}; \widehat{\boldsymbol{V}}^m_{global}]), \quad (3)$$

where $\boldsymbol{F}^a = \{\boldsymbol{f}^a_{t,k}\}^{t=T,k=K}_{t=1,k=1}, \boldsymbol{F}^m = \{\boldsymbol{f}^m_{t,k}\}^{t=T,k=K}_{t=1,k=1} \in \mathbb{R}^{(T \times K) \times D}$ denote the final encoded object-level features.

Query encoder. We first utilize the Glove (Pennington, Socher, and Manning 2014) to embed each word into dense vector, and then employ multi-head self-attention (Vaswani et al. 2017) and Bi-GRU (Chung et al. 2014) to encode its sequential information. The final word-level features can be denoted as $\boldsymbol{Q} = \{\boldsymbol{q}_n\}^N_{n=1} \in \mathbb{R}^{N \times D}$, and the sentence-level feature $\boldsymbol{q}_{global} \in \mathbb{R}^D$ can be obtained by concatenating the last hidden unit outputs in Bi-GRU.

Cross-Modal Object Reasoning

After extracting appearance- and motion-aware object representations, we develop two separate branches to reason both motion-guided and appearance-guided object relations with cross-modal interaction. In each branch, we first interact object features with the query to enhance their semantic, and then reason object relations in a spatio-temporal graph. A query-guided attention module is further developed to fuse the object information within each frame to represent frame-level features.

Cross-modal interaction. Learning correlations between visual features and query information is important for query-based video grounding, which helps to highlight the relevant object features corresponding to the query while weakening the irrelevant ones. Specifically, for the k-th object in the t-th frame in the motion branch, we interact its feature $\boldsymbol{f}^m_{t,k}$ with word-level query features $\{\boldsymbol{q}_n\}^N_{n=1}$ by:

$$\boldsymbol{M}^m_{t,k,n} = \mathrm{w}^\top \tanh(\boldsymbol{W}^m_1 \boldsymbol{f}^m_{t,k} + \boldsymbol{W}^m_2 \boldsymbol{q}_n + \boldsymbol{b}^m_1), \tag{4}$$

where $\boldsymbol{W}^m_1, \boldsymbol{W}^m_2$ are learnable matrices, $\boldsymbol{b}^m_1$ is the bias vector and the $\mathrm{w}^\top$ is the row vector as in (Zhang et al. 2019). The query-enhanced object features $\widehat{\boldsymbol{f}}^m_{t,k}$ can be obtained by:

$$(\boldsymbol{f}^m_{t,k})' = \sum_{n=1}^{N} \mathrm{softmax}(\boldsymbol{M}^m_{t,k,n})\boldsymbol{q}_n, \tag{5}$$

$$\widehat{\boldsymbol{f}}^m_{t,k} = \sigma(\boldsymbol{W}^m_3(\boldsymbol{f}^m_{t,k})' + \boldsymbol{b}^m_2) \odot \boldsymbol{f}^m_{t,k}, \tag{6}$$

where σ is the sigmoid function, $\odot$ represents element-wise product, $\boldsymbol{W}^m_3, \boldsymbol{b}^m_2$ are parameters. $\widehat{\boldsymbol{F}}^m = \{\widehat{\boldsymbol{f}}^m_{t,k}\}^{t=T,k=K}_{t=1,k=1} \in \mathbb{R}^{(T\times K)\times D}$, and the enhanced object features $\widehat{\boldsymbol{F}}^a$ of appearance branch can be obtained in the same way.

Spatio-temporal graph. Since the detected objects have both spatial interactivity and temporal continuity, as shown in Figure 2, we construct object graph to capture spatio-temporal relations in each branch, respectively. For motion branch, we define object-wise features $\widehat{\boldsymbol{F}}^m = \{\widehat{\boldsymbol{f}}^m_{t,k}\}^{t=T,k=K}_{t=1,k=1}$ including all objects in all frames as nodes and build a fully-connected motion graph. We adopt graph convolution network (GCN) (Kipf and Welling 2016) to learn the object-relation features via message propagation. In details, we first measure the pairwise affinity between object features by:

$$\boldsymbol{A}^m = \mathrm{softmax}((\widehat{\boldsymbol{F}}^m \boldsymbol{W}^m_4)(\widehat{\boldsymbol{F}}^m \boldsymbol{W}^m_5)^\top), \tag{7}$$

where $\boldsymbol{W}^m_4, \boldsymbol{W}^m_5$ are learnable parameters. $\boldsymbol{A}^m \in \mathbb{R}^{(T\times K)\times(T\times K)}$ is obtained by calculating the affinity edge of each pair of objects. Two objects with strong semantic relationships will be highly correlated and have an edge with high affinity score in $\boldsymbol{A}^m$. Then, we apply single-layer GCN with residual connections to perform semantic reasoning:

$$\widetilde{\boldsymbol{F}}^m = (\boldsymbol{A}^m \widehat{\boldsymbol{F}}^m \boldsymbol{W}^m_6)\boldsymbol{W}^m_7 + \widehat{\boldsymbol{F}}^m, \tag{8}$$

where $\boldsymbol{W}^m_6$ is the weight matrix of the GCN layer, $\boldsymbol{W}^m_7$ is the weight matrix of residual structure. The output $\widetilde{\boldsymbol{F}}^m = \{\widetilde{\boldsymbol{f}}^m_{t,k}\}^{t=T,k=K}_{t=1,k=1} \in \mathbb{R}^{(T\times K)\times D}$ is the updated features for motion-aware objects. Feature $\widetilde{\boldsymbol{F}}^a$ for appearance-aware objects can be obtained in the same way.

Object feature fusion. After obtaining the object features, we aim to integrate object features within each frame to represent fine-grained frame-level information under the guidance of query information. To this end, we compute the cosine similarity between object feature $\widetilde{\boldsymbol{f}}^m_{t,k}$ and the sentence-level query feature $\boldsymbol{q}_{global}$ at frame t as :

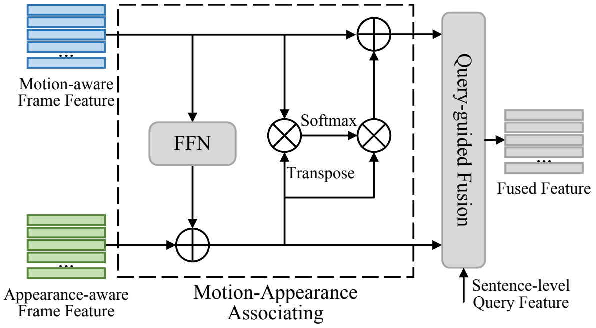


Figure 3: Illustration of the Motion-Appearance Associating (MAA) module, where FFN consists of three linear layers.

$$c^m_{t,k} = \frac{(\widetilde{\boldsymbol{f}}^m_{t,k})(\boldsymbol{q}_{global}\boldsymbol{W}_q)^\top}{||\widetilde{\boldsymbol{f}}^m_{t,k}||_2||\boldsymbol{q}_{global}\boldsymbol{W}_q||_2}, \tag{9}$$

where $\boldsymbol{W}_q$ is the linear parameter, $c^m_{t,k}$ indicates the relational score between the visual object and the given query. Then, we integrate the object features within each frame t to represent query-specific frame-level feature as:

$$\boldsymbol{h}^m_t = \sum_{k=1}^{K} \mathrm{softmax}(c^m_{t,k})\widetilde{\boldsymbol{f}}^m_{t,k}. \tag{10}$$

The final query-specific frame-level features in both appearance and motion branches can be denoted as $\boldsymbol{H}^a = \{\boldsymbol{h}^a_t\}^T_{t=1}, \boldsymbol{H}^m = \{\boldsymbol{h}^m_t\}^T_{t=1} \in \mathbb{R}^{T\times D}$.

Associating Motion and Appearance

After generating appearance- and motion-aware frame-level features $\boldsymbol{H}^a$ and $\boldsymbol{H}^m$, we develop a motion-appearance associating (MAA) module to associate their features and decide which features are most discriminative for final grounding. Module details are shown in Figure 3.

Motion-guided appearance enhancement. Compared to appearance-aware feature $\boldsymbol{H}^a$, the motion-aware feature $\boldsymbol{H}^m$ captures the temporal contexts of objects. Considering the motion context contains implicit appearance information, we abstract the motion context and integrate it into the appearance features for generating motion-enhanced appearance features. Firstly, the motion feature $\boldsymbol{H}^m$ is adapted to appearance feature space through a frame-independent feed-forward network (FFN) which consists of three linear layers. Then, We add the adapted motion information to the appearance feature $\boldsymbol{H}^a$ in an element-wise way, leading to motion-enhanced appearance feature $\widehat{\boldsymbol{H}^a}$. This procedure can be formulated as:

$$\widehat{\boldsymbol{H}^a} = \boldsymbol{H}^a + \mathrm{FFN}(\boldsymbol{H}^m). \tag{11}$$

In this soft and learnable way, contexts appeared in the motion space are aggregated into the appearance feature space.

Appearance-fused motion enhancement. Compared to motion-aware feature $\boldsymbol{H}^m$, the feature $\boldsymbol{H}^a$ represents more on the appearance details and spatial location clues of a certain object. We utilize the dot-product attention to attend appearance-aware object features into motion space for inferring motion contexts as:

$$\widehat{\boldsymbol{H}}^m = \boldsymbol{H}^m + \text{softmax}(\boldsymbol{H}^m(\widehat{\boldsymbol{H}}^a)^\top)\widehat{\boldsymbol{H}}^a, \tag{12}$$

where $\widehat{\boldsymbol{H}}^m$ is the appearance-enhanced motion feature.

Query-guided motion-appearance fusion. To decide which information are most discriminative among appearance and motion features $\widehat{\boldsymbol{H}}^a, \widehat{\boldsymbol{H}}^m$ corresponding to the query, we integrate them under the guidance of sentence-level query features through an attention-based weighted summation as:

$$\begin{aligned}\widetilde{\boldsymbol{H}} &= \text{softmax}(\widehat{\boldsymbol{H}}^a(\boldsymbol{q}_{global})^\top) \odot \widehat{\boldsymbol{H}}^a \\ &+ \text{softmax}(\widehat{\boldsymbol{H}}^m(\boldsymbol{q}_{global})^\top) \odot \widehat{\boldsymbol{H}}^m,\end{aligned} \tag{13}$$

where $\widetilde{\boldsymbol{H}} = \{\widetilde{\boldsymbol{h}}_t\}_{t=1}^T \in \mathbb{R}^{T\times D}$ is the integrated frame-level features to be fed into the last grounding head.

Grounding Head

With the fine-grained query-specific video features $\widetilde{\boldsymbol{H}}$, many grounding heads are plug and play. In this paper, we follow previous works (Zhang et al. 2019; Yuan et al. 2019) to retrieve the target video segment with pre-defined segment proposals. In details, we first define multi-size candidate segments on each frame t, and adopt multiple fully-connected layers to process frame-wise features $\widetilde{\boldsymbol{h}}_t$ to produce the confidence scores and temporal offsets of the segment proposals. Suppose there are R proposals within the video, for each proposal whose start and end timestamp is (τ_s, τ_e), we denote its confidence score and offsets as o and (δ_s, δ_e), where s, e means the start and end. Therefore, the predicted segments of each proposal can be presented as $(\tau_s + \delta_s, \tau_e + \delta_e)$. During training, we compute the Intersection over Union (IoU) score o^{gt} between each pre-defined segment proposal and the ground truth, and utilize it to supervise the confidence score as:

$$\mathcal{L}_{iou} = -\frac{1}{R}\sum o^{gt}\text{log}(o) + (1-o^{gt})\text{log}(1-o). \tag{14}$$

As the boundaries of pre-defined segment proposals are relatively coarse, we further utilize a boundary loss for positive samples (if o is larger than a threshold value λ, the sample is viewed as positive sample) to promote localizing precise start and end points as follows:

$$\mathcal{L}_{boundary} = \frac{1}{R_{pos}}\sum \mathcal{L}_1(\delta_s - \delta_s^{gt}) + \mathcal{L}_1(\delta_e - \delta_e^{gt}), \tag{15}$$

where R_{pos} is the number of positive samples, $\mathcal{L}_1$ denotes the smooth L1 function. The overall loss function can be formulated with a balanced parameter α as:

$$\mathcal{L} = \mathcal{L}_{iou} + \alpha\mathcal{L}_{boundary}. \tag{16}$$

Experiments

Datasets and Evaluation Metric

Charades-STA. Charades-STA is a benchmark dataset for the video grounding task, which is built upon the Charades (Sigurdsson et al. 2016) dataset. It is collected for video action recognition and video captioning, and contains 6672 videos and involves 16128 video-query pairs. Following previous work (Gao et al. 2017), we utilize 12408 video-query pairs for training and 3720 pairs for testing.

TACoS. TACoS is collected by (Regneri et al. 2013) for video grounding and dense video captioning tasks. It consists of 127 videos on cooking activities with an average length of 4.79 minutes. In video grounding task, it contains 18818 video-query pairs. We follow the same split of the dataset as (Gao et al. 2017) for fair comparisons, which has 10146, 4589, and 4083 video-query pairs for training, validation, and testing respectively.

Evaluation metric. We adopt "R@n, IoU=m" proposed by (Hu et al. 2016) as the evaluation metric, which calculates the IoU between the top-n retrieved video segments and the ground truth. It means the percentage of IoU greater than m.

Implementation Details

For appearance-aware object features, We utilize ResNet50 (He et al. 2016) based Faster R-CNN (Ren et al. 2015) model pretrained on Visual Genome dataset (Krishna et al. 2016) to obtain appearance-aware object features, and extract its global feature from another ResNet50 pretrained on ImageNet (Deng et al. 2009). The number K of extracted objects is set to 20. For motion-aware object features, we define continuous 16 frames as a clip and each clip overlaps 8 frames with adjacent clips. We first extract clip-wise features from a pretrained C3D (Tran et al. 2015) or I3D (Carreira and Zisserman 2017) model, and then apply RoIAlign (He et al. 2017) on them to generate object features. Since some videos are overlong, we uniformly downsample frame- and clip-feature sequences to $T = 256$. As for sentence encoding, we utilize Glove (Pennington, Socher, and Manning 2014) to embed each word to 300 dimension features. The head size of multi-head self-attention is 8, and the hidden dimension of Bi-GRU is 512. The dimension D is set to 1024, and the balance hyper-parameter α is set to 0.005. For segment proposals in grounding head, we have 800 samples for each video on both Charades-STA and TACoS datasets, and set $\lambda = 0.55$. We train the whole model for 100 epochs with batch size of 16 and early stopping strategy. Parameter optimization is performed by Adam optimizer with leaning rate 4×10^{-4} for Charades-STA and 3×10^{-4} for TACoS, and linear decay of learning rate and gradient clipping of 1.0.

Comparison with State-of-the-Arts

Compared methods. To demonstrate the effectiveness of our MARN, we compared it with several state-of-the-art methods: (1) Proposal-based: CTRL (Gao et al. 2017), QSPN (Xu et al. 2019), BPNet (Xiao et al. 2021), DRN (Zeng et al. 2020), CBLN (Liu et al. 2021b); (2) Proposal-free: CBP (Wang, Ma, and Jiang 2020), GDP (Chen et al. 2020), VSLNet (Zhang et al. 2020a), IVG-DCL (Nan et al. 2021); (3) Detection-based: MMRG (Zeng et al. 2021).

Comparison on Charades-STA. We compare our MARN with the state-of-the-art proposal-based and proposal-free methods on the Charades-STA dataset in Table 1, where

Method	Charades-STA					TACoS				
	Feature	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7	Feature	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5
CTRL	C3D	23.63	8.89	58.92	29.57	C3D	18.32	13.30	36.69	25.42
QSPN	C3D	35.60	15.80	79.40	45.50	C3D	20.15	15.32	36.72	25.30
CBP	C3D	36.80	18.87	70.94	50.19	C3D	27.31	24.79	43.64	37.40
GDP	C3D	39.47	18.49	-	-	C3D	24.14	-	-	-
VSLNet	I3D	47.31	30.19	-	-	C3D	29.61	24.27	-	-
BPNet	I3D	50.75	31.64	-	-	C3D	25.96	20.96	-	-
IVG-DCL	I3D	50.24	32.88	-	-	C3D	38.84	29.07	-	-
DRN	I3D	53.09	31.75	89.06	60.05	C3D	-	23.17	-	33.36
CBLN	I3D	61.13	38.22	90.33	61.69	C3D	38.98	27.65	59.96	46.24
Ours	C3D+Object	64.47	43.09	93.61	71.55	C3D+Object	46.33	35.74	63.97	53.18
	I3D+Object	**66.43**	**44.80**	**95.57**	**73.26**	I3D+Object	**48.47**	**37.25**	**66.39**	**54.61**

Table 1: Overall performance comparison among our method with proposal-based and proposal-free methods on the Charades-STA and TACoS datasets under the official train/test splits.

Method	Charades-STA					TACoS				
	Feature	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7	Feature	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5
MMRG	Object	44.25	-	60.22	-	Object	57.83	39.28	78.38	56.34
Ours	C3D+Object	48.63	34.25	65.79	41.10	C3D+Object	61.07	42.48	81.91	60.33
	I3D+Object	**50.07**	**35.82**	**68.46**	**42.97**	I3D+Object	**63.69**	**44.37**	**85.28**	**62.14**

Table 2: Comparison with detection-based method MMRG on Charades-STA and TACoS datasets under MMRG's train/test splits. We do not compare with (Zhang et al. 2020c,d) since they address different tasks and datasets and are close source.

we reach the highest results over all evaluation metrics. Particularly, our C3D+Object variant outperforms the best proposal-based method CBLN by 4.87% and 9.86% absolute improvement in terms of R@1, IoU=0.7 and R@5, IoU=0.7, respectively. Compared to the proposal-free method IVG-DCL, the C3D+Object model outperforms it by 14.23% and 10.21% in terms of R@1, IoU=0.5 and R@1, IoU=0.7, respectively. We also compare our model with the detection-based method MMRG in Table 2. To make a fair comparison, we follow the same data splits for training/testing. It shows that our C3D+Object model brings a further improvement of 4.38% and 5.57% in terms of R@1, IoU=0.5 and R@5, IoU=0.5. We further utilize the I3D to present a new I3D+Object variant, which performs better than C3D+Object since I3D can obtain stronger features.

Comparison on TACoS. Table 1 and 2 also report the grounding results on TACoS. Compared to CBLN, our C3D+Object model outperforms it by 7.35%, 8.09%, 4.01%, and 6.94% in terms of all metrics, respectively. Our model also outperforms IVG-DCL by a large margin. Compared to the detection-based method MMRG, our C3D+Object model brings the improvements of 3.20% and 3.99% in strict metrics of R@1, IoU=0.5 and R@5, IoU=0.5, respectively. The I3D+Object variant further achieves better results.

Ablation Study

In this section, we will perform in-depth ablation studies to evaluate the effectiveness of each component in our MARN on Charades-STA dataset. Since most previous works utilize C3D to extract features in this task and our C3D+object variant already achieves the state-of-the-art performance, we utilize the C3D+Object variant as our backbone here.

Main ablation. We first perform main ablation studies to demonstrate the effectiveness of each component. To construct the baseline model, we utilize a general ResNet50 based Faster-RCNN for appearance-aware object extraction and another ResNet50 for frame-level global feature extraction, and do not encode any motion contexts. Instead of building the appearance branch, we employ a co-attention (Lu et al. 2016) module to interact object-level cross-modal information and simply concatenate query-object features for semantic enhancement. We also utilize another co-attention module to capture object-relations, and a mean-pooling layer to fuse object features to represent frame-level features. We utilize the same grounding head in all ablation variants. The performances of each variants are shown in Table 3, and we can observe the following conclusions: (1) It is worth noticing that the baseline model achieves better performance than all existing methods in Table 1, demonstrating that the detection-based method is more effective in distinguishing the frames with high similarity. Our object-level features filter out redundant background information in frame-level features of previous works, thus leading to fine-grained activity understanding and more precise localization. (2) The proposed appearance branch (AB) can cap-

Method	Charades-STA	
	R@1, IoU=0.7	R@5, IoU=0.7
baseline	34.76	63.09
+ AB	37.20	65.97
+ AB&ME	38.91	67.64
+ AB&ME&MB	40.85	69.19
+ AB&ME&MB&MAA	**43.09**	**71.55**

Table 3: Main ablation study on MARN under the official train/test splits. It investigates the appearance branch (AB), the motion encoder (ME), the motion branch (MB), and the motion-appearance associating module (MAA).

Method	Charades-STA	
	R@1, IoU=0.7	R@5, IoU=0.7
MARN	**43.09**	**71.55**
w/o global feature	41.46	70.03
w/o position encoding	40.73	68.69

Table 4: Ablation study on the video encoding.

Module	Changes	Charades-STA	
		R@1, IoU=0.7	R@5, IoU=0.7
Cross-modal Interaction	w/ attention	**43.09**	**71.55**
	w/ concatenation	41.62	70.17
Graph Network	w/ graph	**43.09**	**71.55**
	w/o graph	41.38	69.93
	layer=1	**43.09**	**71.55**
	layer=2	42.74	71.26
Object-feature Fusion	w/ attention	**43.09**	**71.55**
	w/ pooling	41.81	70.44

Table 5: Ablation study on the reasoning branches.

Motion-guided	Appearance-fused	Charades-STA	
		R@1, IoU=0.7	R@5, IoU=0.7
×	×	40.85	69.19
✓	×	41.97	70.23
×	✓	42.14	70.60
✓	✓	**43.09**	**71.55**

Table 6: Ablation study on the MAA module.

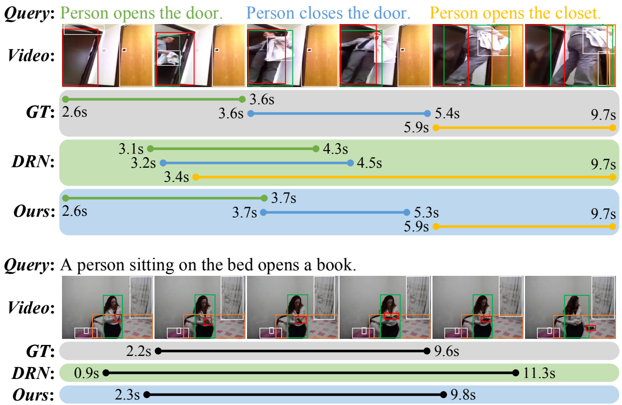

Query: Person opens the door. Person closes the door. Person opens the closet.
Video:
GT: 2.6s 3.6s 3.6s 5.4s 5.9s 9.7s
DRN: 3.1s 4.3s 3.2s 4.5s 3.4s 9.7s
Ours: 2.6s 3.7s 3.7s 5.3s 5.9s 9.7s
Query: A person sitting on the bed opens a book.
Video:
GT: 2.2s 9.6s
DRN: 0.9s 11.3s
Ours: 2.3s 9.8s

Figure 4: The qualitative results on Charades-STA dataset.

ture more fine-grained object relations, thus bringing a significant improvement. (3) The motion encoder (ME) and the corresponding motion branch (MB) can incorporate action-oriented contexts into appearance-based features for better understanding the activity. (4) Motion-appearance association module (MAA) further brings improvement, which proves the effectiveness of incorporating appearance and motion features for bi-directional enhancement.

Analysis on the video encoder. As shown in Table 4, we conduct the investigation on different video encoding. We find that the full model performances worse if we remove the global feature learning. It demonstrates that the frame-wise features help to better explore the non-local object information in the frame. Besides, it also presents the effectiveness of the position encoding in identifying temporal semantic and improving the accuracy of temporal grounding.

Analysis on the reasoning branches. We also conduct ablation study within the reasoning branches as shown in Table 5. For object-level cross-modal interaction, our attention based mechanism outperforms the mechanism of concatenating query-object features by 1.47% and 1.38%. For spatio-temporal graph network, replacing the graph model with simple co-attention model will reduce the performance, since the latter lacks temporal modeling. We can observe that our model achieves best result when the number of graph layer is set to 1, and more graph layers will result in over-smoothing problem (Li, Han, and Wu 2018). For frame-level object-feature fusion, attention based mechanism performs better than the mean-pooling.

Analysis on the associating module. As shown in Table 6, equipped with only motion-guided appearance enhancement, there is a 1.12% and 1.04% point gain compared to the model w/o MAA. This is because the motion context contains implicit appearance information. Beside, when adopting appearance-fused motion enhancement alone, we achieve an improvement of 1.29% and 1.41%, since the appearance-aware objects can also contribute to motion modeling. Applying both of them together, the performance boost is larger than using only one of them.

Visualization

We provide qualitative results in Figure 4, where we choose DRN for comparison due to its open-source. Here, we only show a fixed number of bounding boxes, and color the best matching ones according to the attentive weights in Eq. (5) and (9). For the first video, DRN relies on the frame-level video features, thus failing to distinguish the similar object “door”, “closet”, leading to worse grounding results. By contrast, our model utilizes a detection-based framework that easily captures the appearance differences between “door” and “closet”. Besides, we also incorporate the object-level motion contexts into appearance features, providing more fine-grained details for action understanding. For the second video, our method captures the “open” relations between “person” and “book”, and performs more accurate grounding than DRN.

Conclusion

In this paper, we proposed a novel Motion-Appearance Reasoning Networks (MARN) for temporal sentence grounding, which incorporates both motion contexts and appearance features for better reasoning spatio-temporal semantic relations between objects. Through the developed motion and appearance branches, our MARN manages to mine both motion and appearance clues which matches the semantic of query, and then we devise an associating module to integrate the motion-appearance information for final grounding. Experimental results on two challenging datasets show the effectiveness of our proposed MARN. In the future, we will explore more robust query and video encoders (e.g. bert (Devlin et al. 2018) and X3D (Feichtenhofer 2020)) to further improve the grounding performance.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No.61972448, and in part by the Shenzhen Basic Research (General Project) under Grant No.JCYJ20190806142601687.

References

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 5803–5812.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6299–6308.

Chen, J.; Chen, X.; Ma, L.; Jie, Z.; and Chua, T.-S. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 162–171.

Chen, L.; Lu, C.; Tang, S.; Xiao, J.; Zhang, D.; Tan, C.; and Li, X. 2020. Rethinking the Bottom-Up Framework for Query-based Video Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Chu, W.-S.; Song, Y.; and Jaimes, A. 2015. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3584–3592.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Advances in Neural Information Processing Systems (NIPS)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Feichtenhofer, C. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 203–213.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 5267–5275.

Gao, L.; Zeng, P.; Song, J.; Li, Y.-F.; Liu, W.; Mei, T.; and Shen, H. T. 2019. Structured two-stream attention network for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6391–6398.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hu, R.; Xu, H.; Rohrbach, M.; Feng, J.; Saenko, K.; and Darrell, T. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4555–4564.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* .

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332* .

Le, T. M.; Le, V.; Venkatesh, S.; and Tran, T. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9972–9981.

Li, Q.; Han, Z.; and Wu, X.-M. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Liu, D.; Qu, X.; Di, X.; Cheng, Y.; Xu, Z. X.; and Zhou, P. 2022a. Memory-Guided Semantic Learning Network for Temporal Sentence Grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liu, D.; Qu, X.; Dong, J.; and Zhou, P. 2020a. Reasoning Step-by-Step: Temporal Sentence Localization in Videos via Deep Rectification-Modulation Network. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1841–1851.

Liu, D.; Qu, X.; Dong, J.; and Zhou, P. 2021a. Adaptive Proposal Generation Network for Temporal Sentence Localization in Videos. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9292–9301.

Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Cheng, Y.; Wei, W.; Xu, Z.; and Xie, Y. 2021b. Context-aware Biaffine Localizing Network for Temporal Sentence Grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11235–11244.

Liu, D.; Qu, X.; Liu, X.-Y.; Dong, J.; Zhou, P.; and Xu, Z. 2020b. Jointly Cross-and Self-Modal Graph Attention Network for Query-Based Moment Localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4070–4078.

Liu, D.; Qu, X.; Wang, Y.; Di, X.; Zou, K.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022b. Unsupervised Temporal Video Grounding with Deep Semantic Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liu, D.; Qu, X.; and Zhou, P. 2021. Progressively Guide to Attend: An Iterative Alignment Framework for Temporal

Sentence Grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9302–9311.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems (NIPS)*, 289–297.

Mun, J.; Cho, M.; and Han, B. 2020. Local-Global Video-Text Interactions for Temporal Grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10810–10819.

Nan, G.; Qiao, R.; Xiao, Y.; Liu, J.; Leng, S.; Zhang, H.; and Lu, W. 2021. Interventional Video Grounding with Dual Contrastive Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2765–2775.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Qu, X.; Tang, P.; Zou, Z.; Cheng, Y.; Dong, J.; Zhou, P.; and Xu, Z. 2020. Fine-grained iterative attention network for temporal language localization in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4280–4288.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics* 1: 25–36.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 91–99.

Rodriguez, C.; Marrese-Taylor, E.; Saleh, F. S.; Li, H.; and Gould, S. 2020. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2464–2473.

Seo, A.; Kang, G.-C.; Park, J.; and Zhang, B.-T. 2021. Attend What You Need: Motion-Appearance Synergistic Networks for Video Question Answering. *arXiv preprint arXiv:2106.10446* .

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, 510–526.

Song, Y.; Vallmitjana, J.; Stent, A.; and Jaimes, A. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5179–5187.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4489–4497.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, 5998–6008.

Wang, J.; Ma, L.; and Jiang, W. 2020. Temporally Grounding Language Queries in Videos by Contextual Boundary-aware Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xiao, S.; Chen, L.; Zhang, S.; Ji, W.; Shao, J.; Ye, L.; and Xiao, J. 2021. Boundary Proposal Network for Two-Stage Natural Language Video Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xu, H.; He, K.; Plummer, B. A.; Sigal, L.; Sclaroff, S.; and Saenko, K. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9062–9069.

Yuan, Y.; Ma, L.; Wang, J.; Liu, W.; and Zhu, W. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In *Advances in Neural Information Processing Systems (NIPS)*, 534–544.

Yuan, Y.; Mei, T.; and Zhu, W. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9159–9166.

Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense regression network for video grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10287–10296.

Zeng, Y.; Cao, D.; Wei, X.; Liu, M.; Zhao, Z.; and Qin, Z. 2021. Multi-Modal Relational Graph for Cross-Modal Video Moment Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2215–2224.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2020a. Span-based Localizing Network for Natural Language Video Localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6543–6554.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020b. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, Z.; Lin, Z.; Zhao, Z.; and Xiao, Z. 2019. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 655–664.

Zhang, Z.; Zhao, Z.; Lin, Z.; Huai, B.; and Yuan, N. J. 2020c. Object-Aware Multi-Branch Relation Networks for Spatio-Temporal Video Grounding. In *Proceedings of the International Joint Conferences on Artificial Intelligence*.

Zhang, Z.; Zhao, Z.; Zhao, Y.; Wang, Q.; Liu, H.; and Gao, L. 2020d. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10668–10677.