Multi-Modal Relational Graph for Cross-Modal Video Moment Retrieval

Yawen Zeng[1], Da Cao[1*], Xiaochi Wei[2], Meng Liu[3], Zhou Zhao[4], Zheng Qin[1*]
[1]Hunan University, [2]Baidu Inc., [3]Shandong Jianzhu University, [4]Zhejiang University
{yawenzeng11, caoda0721, mengliu.sdu}@gmail.com
weixiaochi@baidu.com, zhaozhou@zju.edu.cn, zqin@hnu.edu.cn

Abstract

Given an untrimmed video and a query sentence, cross-modal video moment retrieval aims to rank a video moment from pre-segmented video moment candidates that best matches the query sentence. Pioneering work typically learns the representations of the textual and visual content separately and then obtains the interactions or alignments between different modalities. However, the task of cross-modal video moment retrieval is not yet thoroughly addressed as it needs to further identify the fine-grained differences of video moment candidates with high repeatability and similarity. Moveover, the relation among objects in both video and sentence is intuitive and efficient for understanding semantics but is rarely considered.

Toward this end, we contribute a multi-modal relational graph to capture the interactions among objects from the visual and textual content to identify the differences among similar video moment candidates. Specifically, we first introduce a visual relational graph and a textual relational graph to form relation-aware representations via message propagation. Thereafter, a multi-task pre-training is designed to capture domain-specific knowledge about objects and relations, enhancing the structured visual representation after explicitly defined relation. Finally, the graph matching and boundary regression are employed to perform the cross-modal retrieval. We conduct extensive experiments on two datasets about daily activities and cooking activities, demonstrating significant improvements over state-of-the-art solutions.

1. Introduction

Entering the era of information explosion, individuals spend more time in seeking their desired information and the video is not an exception. However, traditional video retrieval methods are specifically designed for whole video retrieval and are not suitable for more fine-grained video moment retrieval scenario. To alleviate people's expectation of quickly retrieving a desired video moment, the task of cross-modal video moment retrieval [1, 8] is proposed. In particular, given an untrimmed video and a query sentence, the task of cross-modal video moment retrieval aims to extract a video moment from the untrimmed video that best matches the query.

(a) Video moment candidates with high similarity

Query sentence: *The person puts a book in a bag.*

(b) Interactions of limited objects

Figure 1: Challenges in cross-modal video moment retrieval. Fig.1a reveals the difficulty of retrieving desired video moment from candidates with high similarity, while Fig.1b exhibits the difficulty of modeling the spatial-temporal interactions of objects.

In fact, a great effort has been made to address the cross-modal video moment retrieval issue. Existing work mostly relies on multi-scale pre-segmented video moment candidates via the sliding window strategy, and then retrieves a suitable video moment from them [36]. Similar to the cross-modal retrieval task [2], the cross-modal video moment retrieval needs to understand and stitch text-video semantics. The typical method is to extract the global [5] and local [3, 17] information of the sentence and video first, then leverage attention mechanism [21, 22, 24] and semantic matching [34] to fuse modalities, and finally rank video moment candidates based on the learned representation. As

*Corresponding authors.

compared to cross-modal video retrieval, the task of cross-modal video moment retrieval is more complicated since it needs to further identify the slight differences of video moment candidates generated from a same video. As shown in Fig.1a, video moment candidates are of high similarity due to the segmentation via the sliding window strategy, which requires more sophisticated intra-modal recognition capabilities. Although recent work has emerged to find the relationship among video moment candidates [38] or generate some more reasonable candidates instead of pre-segmented clips [4], they are not specifically designed for understanding semantics on video frames.

Further observations have found that the background of video moment candidates changes slightly, while the sematic differences of generated candidates are determined by limited objects. As revealed in Fig.1b, for the query sentence, the essential difference between the expected and the deviated candidates is whether the moment of "book enters bag" is covered, which brings the dawn of distinguishing video moment candidates with high similarity. In other words, exploring the interaction pattern among limited objects (i.e., person, book, and bag) is helpful to reduce redundant information and highlight key clues to distinguish video moment candidates. Especially, in the pattern where an object disappears or two objects no longer interact, modeling the interaction of objects can be regarded as a significant signal. Therefore, how to understand the relation among objects in the video and its query sentence, and sensitively capture the differences of video moment candidates with high similarity is of great importance.

To address aforementioned issues, we propose a multi-modal relational graph (MMRG) framework to investigate the cross-modal video moment retrieval task comprehensively. The general framework of MMRG is illustrated in Fig.2. To be specific, we first construct graphs for visual and textual objects separately, where the visual objects are constrained by textual objects instead of modeling all visible objects. Meanwhile, the relations among objects is explicitly treat as nodes to solve the problem of ambiguous relation definition. Moreover, we innovatively propose a customized multi-task pre-training strategy in the visual relation understanding, which can highlight objects and relations, and enhance the performance of visual representation with explicitly defined relation. Finally, both graph matching and boundary regression are introduced to regularize the cross-modal retrieval.

The main contributions of this work are three-fold:

- To the best of our knowledge, this is the first work that attempts to perform the cross-modal video moment retrieval by investigating the interactions among visual and textual objects, which is able to distinguish video moment candidates with high intra-modal similarity.
- We propose a graph-based solution, MMRG, to improve the performance of cross-modal video moment retrieval, which is well suited for modeling the cross-modal semantic consistency and interactions among objects.
- Extensive experiments are conducted on two well-known datasets, which demonstrate the effectiveness of our method. Meanwhile, we have released the dataset and implementation to facilitate the research community[1].

2. Related Work

2.1. Video Moment Retrieval

As an application of artificial intelligence in the multimedia field, cross-modal video moment retrieval has drawn great attention in the research community [6]. Technically, the majority of prior work devotes to handle the cross-modal semantic matching via generating video moment candidates with multi-scale sliding windows. Furthermore, [11] utilizes reinforcement learning to locate the boundary. [4] employs adversarial learning to optimize the candidate generation. Some other work employs an interactive graph [38] or 2D adjacent temporal relation [40] to extract the relation among candidate moments. In terms of enhancing semantic understanding, the attention mechanism is utilized to promote cross-modal semantic fusion [21, 22, 24]. Further, researchers refine query sentences to word level [17, 36, 42] and explore visual temporal relations [41]. Among the work mentioned above, few efforts have been made to explore the interaction among visual and textual objects, which is more intuitive and crucial to capture the differences of video moment candidates with high similarity.

2.2. Visual Relational Reasoning

As computer vision technology continues to be explored, visual relational reasoning performs outstandingly in image/video understanding, such as image captioning [14, 37], visual question answering [20], and action recognition [35]. [23] propose a language-guided graph representation to capture entities and their relations, and develop a cross-modal graph matching strategy for the multiple-phrase visual grounding task. [43] design an object relation graph and a teacher-recommended learning to integrate the abundant linguistic knowledge into the caption model. [28] abstract videos as fully-connected spatial-temporal 3D graphs with object trajectory for video relation detection. However, modeling object relations via the spatio-temporal graph [27] is still not thoroughly investigated. In our work, we construct the connections among objects by explicitly expressing relation nodes to ease the ambiguity

[1] https://cvpr-2021.wixsite.com/mmrg

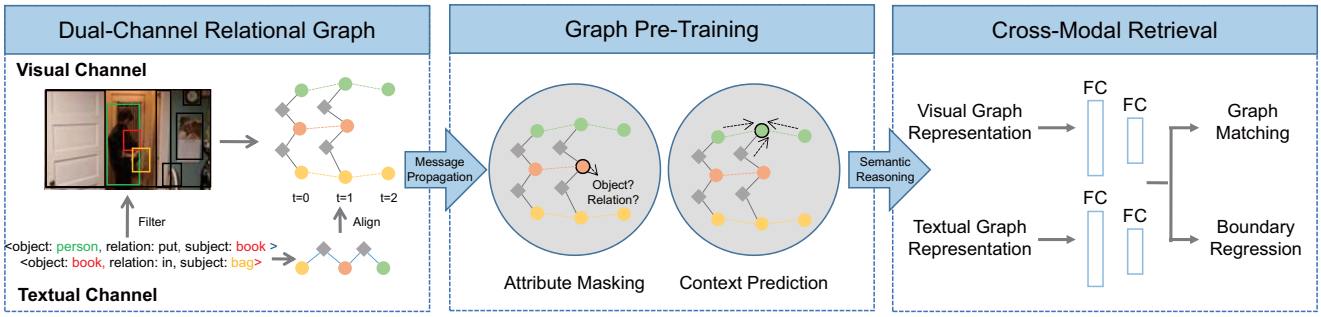

Figure 2: The graphical representation of our proposed MMRG framework. The input is an untrimmed video and its query sentence, while the output is the alignment score and location offset.

of visual relation, further digging into object interactions, and performing the cross-modal fusion of object graphs.

2.3. Graph Pre-Training

Despite the pre-training model is of great significance in computer vision [19] and natural language processing [7], few studies have applied it to the graph. In fact, applying pre-training to the graph can capture domain-specific knowledge at the node/edge level or even graph level since the graph has a common and transferable structural pattern. [16] suggest that pre-training at the node level and graph level can excellently enhance downstream tasks. [39] fuse pre-training and graph to learn graph representation when only attention is needed. [29] present graph contrastive coding to measure the structural similarity. In this paper, we pioneered to explore an appropriate pre-training scheme in the cross-modal video moment retrieval scenario to enhance the heterogeneous representations of objects and relations after explicitly defined relation.

3. Method

As illustrated in Fig.2, our proposed MMRG framework consists of three modules: dual-channel relational graph, graph pre-training, and cross-modal retrieval. Specifically, the dual-channel relational graph module constructs textual relational graph and visual relational graph, respectively. Thereinto, the textual relational graph is utilized to filter irrelevant objects in the visual relational graph and further apply multi-head attention to align object semantics. Thereafter, the pre-training module customizes two pre-training tasks, i.e., attribute masking and context prediction, to enhance the visual relation reasoning after explicitly defined relation. Finally, the graph matching and boundary regression are utilized to perform the cross-modal retrieval.

3.1. Problem Formulation

Let v and q denote a long untrimmed video and a query sentence, respectively. The query sentence q is affiliated with a temporal annotation $loc_q = [l_s, l_e]$ on the video v, where l_s and l_e are the start and end points of the target video moment. Given the video v and its query sentence q, the goal of cross-modal video moment retrieval is to identify the desired video moment with the boundary of $loc_o = [o_s, o_e]$ to be close to the ground truth loc_q.

3.2. Duel-Channel Relational Graph

To capture the interaction pattern among objects from both visual and textual content, we design a duel-channel relational graph, which define the explicit relations and apply message propagation.

3.2.1 Textual Relational Graph.

The textual relational graph is constructed by extracting phrase relations. Specifically, the sentence parser[2] is employed to identify the phrase nouns H^p and the relations $r^p_{ij} \in R^p$ between the nouns $h^p_i, h^p_j \in H^p$ from the query sentence q. Due to the objects are not born with relational structures, simply treating nouns as nodes and ignoring relational phrase [14, 44] will lose and cannot understand the semantic information of the relation explicitly.

To this end, we explicitly define relations R^p as nodes to construct graphs $\mathcal{G}_p$. It means that noun objects will be associated by relational phrases in graph $\mathcal{G}_p$. Formally, the textual relational graph is denoted as $\mathcal{G}_p = \{H^p \cup R^p, E^p, \mathbf{X}^p\}$, where E^p are the edges connecting nodes and $\mathbf{X}^p = \{\mathbf{X}^p_H \cup \mathbf{X}^p_R\}$ are the phrase (noun and relation) embedding vectors extracted by word2vec[3]. Meanwhile, the noun similarity is performed as post-processing strategy on the Flickr30K Entities dataset[4] , which ensures isolated nodes are not existed in the textual relational graph $\mathcal{G}_p$.

Two message propagation operators are then leveraged to get higher-level representation of node features $\mathbf{X}^p_H$ and $\mathbf{X}^p_R$. Primarily, we optimize the expression of the relation node $\mathbf{X}_{r^p_{ij}}$, which is determined by two nouns (subject $\mathbf{X}_{h^p_i}$ and object $\mathbf{X}_{h^p_j}$) and its own feature. Then we update the

[2] https://github.com/vacancy/SceneGraphParser
[3] https://code.google.com/archive/p/word2vec
[4] http://shannon.cs.illinois.edu/DenotationGraph/

relation node by aggregation, as,

$$\mathbf{X}'_{r^p_{ij}} = \mathbf{X}_{r^p_{ij}} + f^r_p([\mathbf{X}_{h^p_i}, \mathbf{X}_{h^p_j}, \mathbf{X}_{r^p_{ij}}]), \tag{1}$$

where f^r_p is feature mapping function with the implementation of fully connected layers (FC).

Different from the relation node which is the connection of two nouns, the noun node has indefinite neighbors. Therefore, we aggregate neighbor object nodes and their relations via the attention mechanism [33]:

$$\mathbf{X}'_{h^p_i} = \mathbf{X}_{h^p_i} + \sum_{h^p_j \in \mathcal{N}(h^p_i)} w_{r^p_{ij}} f^h_p([\mathbf{X}_{h^p_j}, \mathbf{X}'_{r^p_{ij}}]), \tag{2}$$

$$w_{r^p_{ij}} = \underset{h^p_j \in \mathcal{N}(h^p_i)}{softmax}(f^h_p([\mathbf{X}_{h^p_i}, \mathbf{X}'_{r^p_{ij}}])^T f^h_p([\mathbf{X}_{h^p_j}, \mathbf{X}'_{r^p_{ij}}])), \tag{3}$$

where f^h_p is FC, the attention weight $w_{r^p_{ij}}$ of the neighborhood $\mathcal{N}(h^p_i)$ and updated relation node $\mathbf{X}'_{r^p_{ij}}$ is determined by the distance of their mapping features.

3.2.2 Visual Relational Graph.

Like the textual relational graph, we also strive to capture the relations among visual objects by constructing a visual relational graph $\mathcal{G}_v$. To maintain the consistency of the phrase object and the visual object in frame, we resort to the phrase object H^p to filter the proposals from Faster R-CNN[5] [31]. Specifically, only top-1 proposals with the highest similarity of phrases is applied as the nodes H^v of the visual relational graph $\mathcal{G}_v$, and the node vectors $\mathbf{X}^v_H$ are extracted through RoI-Align [12]. It is worth noting that if the similarity between the word2vec feature of the proposal label and the phrase is less than 0.5, we assume that there is no corresponding visual object in the moment. Meanwhile, if the similarity of multiple proposal regions is larger than 0.9, these regions will be merged, and their joint region is regarded as the visual object.

Interaction relation modeling. To capture the relations R^v among visual objects, especially the interaction pattern in spatial (frame), we combine some regional features as the explicit representation of relations. To simplify the definition of the formula, the default frame time is t. Therefore, the initial relations expression $\mathbf{X}_{r^v_{ij}} \in \mathbf{X}^v_R$ of the visual object relation $r^v_{ij} \in R^v$ is represented as:

$$\mathbf{X}_{r^v_{ij}} = [\mathbf{X}_{un_{ij}}, \mathbf{X}_{mu_{ij}}, \mathbf{X}_{pos_i}, \mathbf{X}_{pos_j}], \tag{4}$$

where $\mathbf{X}_{un_{ij}}$ is the visual feature extracted from the union box region of two objects, which is the minimum box region covering both objects. The mutual feature $\mathbf{X}_{mu_{ij}}$ indicates whether two objects are overlapped, and are expressed by the mutual position between the two objects. Let $[x_i, y_i, w_i, h_i]$ and $[x_j, y_j, w_j, h_j]$ denote the coordinates of the object $h^v_i, h^v_j \in H^v$ in time t respectively, where (x, y) indicates the position of top left corner, and w, h are the width and height of the box. $\mathbf{X}_{mu_{ij}}$ is formally defined as:

$$\mathbf{X}_{mu_{ij}} = [\frac{x_i - x_j}{w_j}, \frac{y_i - y_j}{h_j}, log\frac{w_i}{w_j}, log\frac{h_i}{h_j}]. \tag{5}$$

Moreover, the location representation based on union region $\mathbf{X}_{pos_i} = [\frac{x_i}{W_{u_{ij}}}, \frac{y_i}{H_{u_{ij}}}, \frac{x_i+w_i}{W_{u_{ij}}}, \frac{y_i+h_i}{H_{u_{ij}}}, \frac{S_i}{S_{u_{ij}}}]$ of the object itself is also important, where u_{ij} is the union region between the objects $h^v_i, h^v_j \in H^v$.

Spatial Propagation. Since visual nodes have both spatial interactivity and temporal continuity, we perform message propagation in spatial and temporal respectively. Similar to the textual relational graph $\mathcal{G}_p$, the visual relational graph $\mathcal{G}_v$ is defined as $\mathcal{G}_v = \{H^v \cup R^v, E^v, \mathbf{X}^v\}$, where E^v is the edge connecting the nodes and $\mathbf{X}^v = \{\mathbf{X}^v_H \cup \mathbf{X}^v_R\}$. Then the explicit relation features are optimized via message propagation. There are only two object nodes connecting to a relation node r^v_{ij}, so the representation of r^v_{ij} is updated as:

$$\mathbf{X}'_{r^v_{ij}} = \mathbf{X}^v_{r_{ij}} + f^r_v([\mathbf{X}_{h^v_i}, \mathbf{X}_{h^v_j}, \mathbf{X}_{r^v_{ij}}]), \tag{6}$$

where f^r_v is FC. We then aggregate neighbor object nodes and their relations via the attention mechanism similar to the textual relational graph:

$$\mathbf{X}'_{h^v_i} = \mathbf{X}_{h^v_i} + \sum_{h^v_j \in \mathcal{N}(h^v_i)} w_{r^v_{ij}} f^h_v([\mathbf{X}_{h^v_j}, \mathbf{X}'_{r^v_{ij}}]), \tag{7}$$

$$w_{r^v_{ij}} = \underset{h^v_j \in \mathcal{N}(h^v_i)}{softmax}(f^h_v([\mathbf{X}_{h^v_i}, \mathbf{X}'_{r^v_{ij}}])^T f^h_v([\mathbf{X}_{h^v_j}, \mathbf{X}'_{r^v_{ij}}])), \tag{8}$$

where f^h_v is FC, the attention weight $w_{r^v_{ij}}$ of the neighborhood $\mathcal{N}(h^v_i)$ and the updated relation node $\mathbf{X}'_{r^v_{ij}}$ are determined by the distance of their mapping features.

Cross-modal temporal propagation. Under the cross-modal paradigm, the expression of the visual object node $\mathbf{X}^v_H$ should be consistent with the phrase object node $\mathbf{X}^p_H$. Therefore, $\mathbf{X}^v_H$ should perform information aggregation under the constraint of the phrase. In other words, the visual object node's neighbors $\mathcal{N}(h^v_i)$ also include corresponding cross-modal phrase object. To this end, we present cross-modal graph attention over neighbors to learn which neighbors are more relevant and weigh their contributions accordingly.

$$\mathbf{X}'_{h^v_i} = \sigma(\sum_{j \in \mathcal{N}(h^v_i)} softmax(e^v_{ij}) \mathbf{W}_{ij^h} \mathbf{X}_{h^v_j})), \tag{9}$$

$$e^v_{ij} = \mathbf{a}^T [\mathbf{W}_{ij^h} \mathbf{X}_{h^v_i}, \mathbf{W}_{ij^h} \mathbf{X}_{h_j}], j \in \mathcal{N}(h^v_i), \tag{10}$$

[5] https://github.com/rbgirshick/py-faster-rcnn

where $\mathbf{W}_{ij^h}$ is a learnable linear transformation matrix, $\sigma(\cdot)$ is the LeakyReLU, and $\mathbf{a}^T$ is a learnable shared vector. We then extend attention to multi-head strategy [32] by repeating K times so that the training process is more stable, which is formally formulated as:

$$\mathbf{X}'_{h_i^v} = \Big\|_{k=1}^{K} \sigma\Big(\sum_{j \in \mathcal{N}(h_i^v)} softmax(e_{ij}^v)\mathbf{W}_{ij^h}\mathbf{X}_{h_j^v}\Big), \quad (11)$$

where $||$ represents the concatenation and $\sigma(\cdot)$ is the sigmoid function. The learned embeddings are connected as the semantic-specific graph embedding.

3.3. Pre-Training in Graph

Recently, there are rarely pre-training discussions in the graph, at least not in the video domain. The main difficulty is that the ambiguous connection definition of visual relation among two pixel areas (objects) easily lead to failures in semantic understanding. In the previous section, we have presented multi-modal relational graph with explicit relation to enhance the semantic expression, but its understanding of nodes is still incomplete. Especially, after explicitly defined relation, the solution is required to deal with two issues, the heterogeneity of nodes in the graph, and the semantic gap across modalities. Therefore, we pioneered the introduction of task-adaptive pre-training (TAPT) strategy [9, 16] in visual relation understanding. Consequently, two types of practical self-supervised pre-training tasks at the node level and graph level are considered.

Attribute Masking. This pre-training task is designed to optimize the heterogeneous feature of relation/object nodes under the explicit expression. The premise of capturing domain–specific knowledge is that our model can distinguish these two types of nodes with different meanings. Specifically, we label 20% of the objects and relation nodes of visual relational graph $\mathcal{G}_v$, 80% of which are replaced with [MASK] labels, and the remaining parts are kept the original attributes. Consequently, this attribute masking task can force our model to predict these attributes based on neighboring nodes. Further, the objects and relations can be better distinguished at the node level, and more neighborhood knowledge and clearer relation can be captured, which significantly helps the learning of explicit relation features.

Context Prediction. To ensure the vector representations of nodes can capture the global information of graph structure under the cross-modal paradigm, the cross-modal context prediction is also used as a pre-training task. Hence, the visual relational graph structure is reconstructed from the neighbor subgraphs of nodes under textual semantic constraints, so that nodes appearing in the context of similar structures can be mapped to similar embeddings. Specifically, the subgraph structure of negative sampling is employed to randomly sample the context that is not adjacent to the current object. Finally, the reconstruction is optimized by the pairwise loss. Through this task, the model can learn structural information at the graph level to alleviate semantic gap across modalities.

3.4. Cross-Modal Retrieval

Based on previous efforts, each node feature captures interactions about objects. In this section, these relation-aware representations are merged with global information for cross-modal video moment retrieval. Following the popular methods [8, 21], we retrieve the most suitable video moment from pre-segmented candidates with high similarity. This involves two related downstream tasks, namely, graph matching and boundary regression. Graph matching determines whether the semantics of the visual relational graph $\mathcal{G}_v$ and textual relational graph $\mathcal{G}_p$ are related, while boundary regression to further adjust the boundary.

Graph Matching. Since graph embedding and graph pooling either employ subgraph sampling [26] or aggregate node information [10], which still lost the structure information of the graph, we introduce cross-graph matching to calculate the similarity $s_z(q_n, v_m)$ between graphs as follows:

$$f_h^s([\mathbf{X}'_{h_i^p}, O(\sum_{t \in T}\mathbf{X}'_{h_{it}^v})]) + f_r^s([\mathbf{X}'_{r_{ij}^p}, O(\sum_{t \in T}\mathbf{X}'_{r_{ijt}^v})]), \quad (12)$$

where O refers to the max pooling, f_h^s and f_r^s are two-layer FC utilized. We divide the moment-query pairs into two groups, $\mathcal{P}$ are treated as the positive matched pairs, while $\mathcal{N}$ are considered as the negative mismatched pairs. Finally, the matching loss of our model is constructed as follows:

$$\begin{aligned} L_{mah} = &\sum_{(q_n, v_m) \in \mathcal{P}} log(1 + exp(-s_z(q_n, v_m))) \\ &+ \sum_{(q_n, v_m) \in \mathcal{N}} \lambda_1 log(1 + exp(s_z(q_n, v_m))), \end{aligned} \quad (13)$$

where λ_1 is a hyper parameter balancing the weights between the positive and negative pairs.

Boundary Regression. As the multi-scale sliding window is adopted to segment videos, fixed duration of the moment candidate needs more flexible boundary offset supplement. Here we adopt the moment boundary adjustment strategy and denote the offset values $loc_z = [l_{s'}, l_{e'}]$, where $l_{s'}, l_{e'}$ are offset of the start and end points. The boundary offset regression is obtained through a FC f^l:

$$loc_z = f^l([\widetilde{\mathbf{X}}_{q_n}, \widetilde{\mathbf{X}}_{v_m}]), \quad (14)$$

where $\widetilde{\mathbf{X}}_{q_n}$ and $\widetilde{\mathbf{X}}_{v_m}$ concatenate the node features, relation features and global features (extracted from complete video

or sentence). After predicting the boundary, the final output $loc_o = [o_s, o_e]$ of the model is:

$$o_s = \tau_s + l_{s'}, o_e = \tau_e + l_{e'}. \tag{15}$$

The boundary regression loss $\mathcal{L}_{reg}$ is defined as the IoU (Intersection over Union) value of ground truth loc_{q_n} and loc_o:

$$\mathcal{L}_{reg} = \sum_{(q_n, v_m) \in \mathcal{P}} |IoU(loc_{q_n}, loc_o)|, \tag{16}$$

$$IoU(loc_{q_n}, loc_o) = \frac{min(l_e, o_e) - max(l_s, o_s)}{max(l_e, o_e) - min(l_s, o_s)}. \tag{17}$$

In the training phase, the boundary regression loss is only performed on positive samples, while in the testing scenario, the candidate with the highest matching score is added an offset value to relocate the boundary. Eventually, the overall loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{mah} + \lambda_2 \mathcal{L}_{reg}, \tag{18}$$

where λ_2 is a hyper parameter balancing the weights of graph matching and boundary regression.

4. Experiments

In this section, we conduct extensive experiments on two datasets to answer the following three research questions:

RQ1 How does our proposed MMRG framework perform as compared to other state-of-the-art competitors?

RQ2 How do different components in MMRG framework contribute to its performance?

RQ3 Can we visualize the retrieval performance of various methods and interaction pattern among objects?

4.1. Datasets and Evaluation Metric

We experimented with two publicly accessible datasets: Charades-STA [8] and TACoS [30], one is related to daily activities at home[6] and the other one is cooking activities in lab kitchen[7]. We downloaded original datasets and further constructed the moment candidates with different unit sizes of $[64, 128, 256, 512]$ via 80% overlap. In summary, we ultimately obtained $12,541$ and $7,463$ video moment-query sentence pairs for Charades-STA and TACoS, respectively.

The experimental datasets are divided into training, verification and testing according to 70%, 10%, and 20%. To evaluate the performance of MMRG and other baselines, we adopted "R@n, IoU=m" proposed by [15] as the evaluation metric, which calculates the IoU between the top-n retrieved video moments and the ground truth. In the rest of this article, we use R(n,m) to mean "R@n, IoU=m", which is the percentage of IoU greater than m.

[6] https://allenai.org/plato/charades

[7] http://www.coli.uni-saarland.de/projects/smile/tacos

4.2. Implementation Details

In the feature representation, visual object feature is a $1,024$-dimensional vector extracted by RoI-Align [12], while textual object/relation feature is $1,024$-dimensional vector obtained by employing word2vec [25]. In addition, global features of query and video are $4,800$-dimension and $4,096$-dimension extracted by Sentence2vec [18] and C3D [13], respectively. For the dimensions of the FC layer, the input of f_p^r, f_p^h, f_v^r and f_v^h is the concatenated dimension and the output is $1,024$. To initialize the hidden layers in our method, we randomly set their parameters with a gaussian distribution (a mean of 0 and a stand deviation of 0.1). The number of multi-head K is set as 6, and balance parameters λ_1 and λ_2 are set as 0.8 and 0.7, respectively.

4.3. Overall Performance Comparison (RQ1)

To demonstrate the effectiveness of our proposed method, we compared it with several state-of-the-art approaches: 1) CTRL [8]; 2) MCN [1]; 3)ROLE [22]; 4) ACRN [21]; 5) READ [11]; 6) MAN [38]; 7) CMIN [42]; 8) ORG [43]; and 9) STVC [27]. CTRL and MCN employ sliding window strategy to generate video moment candidates, ROLE and ACRN are algorithms that apply attention mechanism to align local semantics, MAN and CMIN are models based on the graph technique, and ORG and STVC are frameworks for applying graph to understand relation. It worth to mention that READ is a reinforcement learning-based method, which is designed for boundary localization and only returns a boundary value. Therefore, its performance is only compared on R(1,m).

Experimental results are shown in Table 1, we have the following observations: 1) Both attention-based methods, ROLE and ACRN, and reinforcement learning-based algorithm READ have better performance than that of CTRL and MCN, which indicates that it is necessary to understand and integrate local features. 2) The two graph-based methods, MAN and CMIN beats other baselines on a great margin, which verifies that it is critical to capture the structural information. Meanwhile, the performance of the two relation-based frameworks, ORG and STVC are not as excellent as expected, which may be due to the visual indirect relation is too vague to get a higher level of understanding. 3) Our approach MMRG outperforms prior methods on both Charades-STA and TACoS. This manifests that the multi-modal relational graph with pre-training strategy is effective, which helps to improve the performance of feature representation to further identify the differences of video moment candidates with high repeatability and similarity.

4.4. Ablation Study (RQ2)

To better understand the contribution of different components in our framework, we conduct ablation studies on the

Table 1: Overall performance comparison among various methods on Charades-STA and TACoS datasets.

Method	Charades-STA						TACoS					
	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5
CTRL	80.67%	64.88%	37.54%	88.53%	73.81%	53.50%	76.01%	50.52%	34.02%	81.57%	70.55%	49.39%
MCN	79.11%	66.45%	39.72%	88.75%	75.26%	55.11%	77.85%	49.45%	35.21%	80.72%	71.25%	50.08%
ROLE	84.75%	67.56%	41.08%	90.01%	76.38%	57.38%	79.22%	55.20%	37.42%	81.80%	74.30%	53.27%
ACRN	84.57%	68.25%	40.85%	90.41%	76.92%	57.44%	81.34%	54.24%	36.25%	80.73%	74.76%	53.02%
READ	87.00%	70.24%	42.06%	-	-	-	83.03%	57.53%	38.04%	-	-	-
MAN	86.82%	70.11%	42.35%	91.06%	76.77%	58.01%	83.11%	55.86%	37.82%	82.39%	75.96%	53.96%
CMIN	87.36%	70.55%	42.67%	91.58%	77.39%	58.37%	83.80%	56.26%	37.42%	82.64%	76.29%	54.35%
ORG	87.02%	70.43%	42.73%	91.67%	77.55%	58.58%	83.40%	56.31%	37.37%	82.77%	76.41%	54.58%
STVC	86.25%	69.72%	41.35%	90.68%	76.44%	57.46%	82.66%	55.30%	36.58%	81.54%	75.51%	53.70%
MMRG	**88.27%**	**71.60%**	**44.25%**	**92.35%**	**78.67%**	**60.22%**	**85.34%**	**57.83%**	**39.28%**	**84.37%**	**78.38%**	**56.34%**

Table 2: Performance comparison of MMRG and its variants on Charades-STA and TACoS datasets.

Method	Charades-STA						TACoS					
	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5
Backbone	80.80%	64.38%	37.27%	88.72%	73.27%	53.28%	76.17%	50.28%	34.82%	81.81%	70.33%	49.55%
+GCN	80.34%	64.37%	37.13%	88.35%	73.42%	53.11%	75.42%	50.01%	34.98%	81.55%	70.49%	49.29%
+GAT	82.27%	66.72%	40.22%	90.45%	76.68%	57.24%	78.38%	53.36%	36.72%	82.36%	72.36%	51.30%
+Cross	81.88%	66.35%	39.72%	89.72%	75.39%	56.72%	78.05%	53.11%	36.22%	81.97%	72.32%	51.08%
+STG	83.21%	67.73%	41.72%	89.27%	76.20%	57.09%	80.23%	54.09%	37.44%	81.88%	72.07%	52.11%
+RSTG	84.77%	68.27%	42.07%	90.86%	77.23%	58.37%	82.48%	55.68%	38.57%	82.78%	73.99%	53.56%
+PreAM	87.04%	70.08%	43.73%	91.05%	77.28%	59.38%	83.82%	56.75%	38.27%	83.45%	74.83%	55.18%
+PreCP	86.72%	69.73%	43.58%	91.33%	77.65%	59.86%	84.53%	57.01%	38.88%	83.05%	75.28%	55.86%
MMRG	**88.27%**	**71.60%**	**44.25%**	**92.35%**	**78.67%**	**60.22%**	**85.34%**	**57.83%**	**39.28%**	**84.37%**	**78.38%**	**56.34%**

graph-based visual representation learning and pre-training tasks. Specifically, we compared our model to its variants: 1) Backbone is a model similar to CTRL, which combines global and local objects information. The following variants are all extended based on this method. 2) +GCN and +GAT are variants of utilizing GCN or GAT to optimize features for visual graph. 3) +Cross implements cross-graph attention on visual graph and textual graph. 4) +STG is the spatio-temporal graph introduced in this paper and +RSTG designs interaction relation attributes. 5) +PreAM and +PreCP are two pre-training tasks of attribute masking and context prediction respectively, and their training is based on +RSTG.

The performance of MMRG and its variants is shown in Table 2. We have the following observations: 1) The variant +GCN has little improvement as compared to Backbone, which shows that using GCN to capture a complete visual graph is less effective. 2) The performance of +GAT is better than that of +GCN, which implies that the attention mechanism is effective for connecting temporal and spatial nodes. However, +Cross, which introduces cross-attention to visual graph and textual graph, performs worse as compared to +GAT. We suspect that the two graphs interfere with each other due to the indistinguishability among spatio-temporal nodes. 3) The performance of +STG is outstanding, indicating that carefully handling spatio-temporal nodes is helpful for visual understanding. Meanwhile, the +RSTG has been further improved due to being strengthened by the explicit relation, which shows the richness of the relation information. 4) The two pre-training tasks +PreAM and +PreCP have better performance than that of +RSTG, which proves that the customized pre-training methods are effective for the learning of node features. Meanwhile, MMRG employs two pre-training methods and achieves the best performance, which verifies that two pre-training tasks designed at the node level and graph level are able to highlight objects and relations and promote the representation learning performance.

4.5. Qualitative Analysis (RQ3)

The understanding of video and sentence only involves limited objects, which are closely connected. To gain deep insight into the impact of objects on retrieval performance, we exploited micro-level case studies on video moment candidates with high similarity. Specifically, we randomly selected two video-language pairs accompanied with the ground truth of video moment from two datasets. In the above instance of Fig.3, the given query sentence is "the person puts a book in a bag". It is easy to observe that when the book (i.e., the red bounding box) is completely putted in the bag, the object "book" is disappeared. In the bottom instance of Fig.3, the query sentence is "the person puts down a bag". When the girl puts down the bag, there is no interaction between "person" and "bag".

As revealed in Fig.3, we visualized the performance of Backbone, +RSTG, and MMRG to analyze the impact

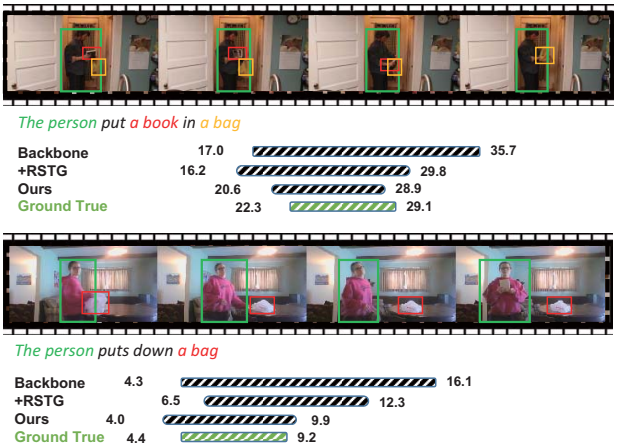

Figure 3: The performance of Backbone, +RSTG and MMRG on two instances. Black bars represent retrieved video moments, while green bars represent the ground truth.

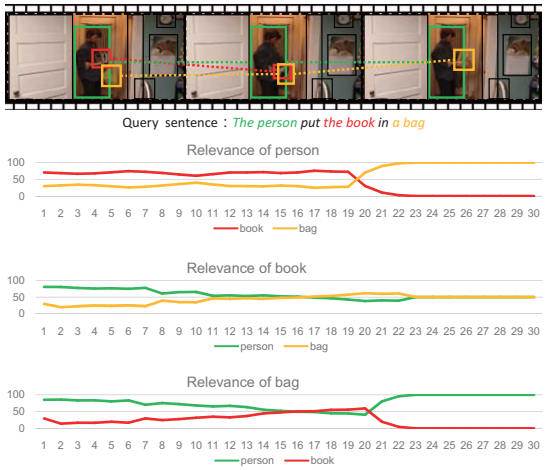

Figure 4: The visualization of relevance among objects. The x-axis represents the timestamp of frame and the y-axis is the normalized probability of relevance.

of object relations and pre-training strategies. Accordingly, we have the following observations: 1)+RSTG and MMRG have obtained better matching video moments than Backbone, mainly because +RSTG and MMRG can sensitively perceive dynamic changes in objects and interaction patterns. This advantage allows these two models to optimize the boundary more precisely, which manifests the significance of capturing objects and their interactions. 2) The performance of +RSTG is better than that of Backbone. This is largely because +RSTG narrows down the boundary to be relevant to the limited objects. 3) MMRG integrates pre-training tasks into +RSTG to improves the retrieval performance. This proves the superiority of MMRG in objects and semantics understanding by employing pre-training tasks.

To further analyze the interaction pattern among objects, we zoom in on the moment when the action "put" occurs, and then calculate the mutual similarity among objects via learned representation. The visualization is shown in Fig.4, the whole moment is roughly divided into three stages: before (1-7), during (8-20) and after (21-30). For each object before the action is performed, MMRG considers "people" is closely related to "bag" and "book", but there is no strong relation between "book" and "bag". When the action occurs, "book" and "bag" have the violent interaction, and the relevance of them also increases significantly. Eventually, when the action is completed, "book" disappears, leaving "person" and "bag" in the scene. It is worth noting that the disappearance of "book" is appeared around the 20th step, and the extension of the relevance curve is smooth. It is probably caused by the introduction of multi-modal relational graph spreading to multiple hops. In summary, MMRG has demonstrated its great ability to capture the interaction pattern among objects, which improves the discrimination of video moment candidates with high intra-modal similarity.

5. Conclusions and Future Work

In this paper, we address the cross-modal video moment retrieval issue by employing a multi-modal relational graph to identify the differences of video moment candidates generated from a same video with high intra-modal similarity. Specifically, we first introduce dual-channel relational graph to form relation-aware representations via message propagation. Thereafter, customized pre-training tasks are designed to enhance the visual representation. Finally, graph matching and boundary regression are employed to perform the cross-modal retrieval. Extensive experiments have verified the effectiveness of our proposed solution.

In the future, we are interested in realizing the video moment retrieval in a personalized manner. As such, the retrieved results are relevant to the personal interests of users. Along this line, the personal query history can be treated as the user-item interactions to better capture a user's preference towards video moments.

Acknowledgements

We are highly grateful to the anonymous reviewers for their careful reading and insightful comments. This work is supported by the the National Natural Science Foundation of China (No. 61802121, No. 61772191, and No. U20A20174), the Natural Science Foundation of Hunan Province (No. 2019JJ50057), the Science and Technology Projects of Hunan Province (No. 2020SK2066, No. 2019WK2072, No. 2018TP2023, and No. 2015TP1004), the Changsha Science and Technology Project (No. kq2006029), and the Fundamental Research Funds for the Central Universities.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *CVPR*, pages 5803–5812, 2017. 1, 6

[2] Da Cao, Zhiwang Yu, Hanling Zhang, Jiansheng Fang, Liqiang Nie, and Qi Tian. Video-based cross-modal recipe retrieval. In *Proceedings of the ACM International Conference on Multimedia*, pages 1685–1693. ACM, 2019. 1

[3] Da Cao, Yawen Zeng, Meng Liu, Xiangnan He, Meng Wang, and Zhen Qin. STRONG: Spatio-temporal reinforcement learning for cross-modal video moment localization. In *MM*, pages 4162–4170, 2020. 1

[4] Da Cao, Yawen Zeng, Xiaochi Wei, Liqiang Nie, Richang Hong, and Zeng Qin. Adversarial video moment retrieval by jointly modeling ranking and localization. In *MM*, pages 898–906, 2020. 2

[5] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, pages 162–171, 2018. 1

[6] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, pages 10635–10644, 2020. 2

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

[8] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *CVPR*, pages 5267–5275, 2017. 1, 5, 6

[9] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*, pages 8342–8360, 2020. 5

[10] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, pages 1024–1034, 2017. 5

[11] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*, pages 8393–8400, 2019. 2, 6

[12] Kaiming He, Georgia Gkioxari, Piotr Dollr, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988, 2017. 4, 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[14] Jingyi Hou, Xinxiao Wu, Xiaoxun Zhang, Yayun Qi, Yunde Jia, and Jiebo Luo. Joint commonsense and relation reasoning for image and video captioning. In *AAAI*, pages 10973–10980, 2020. 2, 3

[15] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, pages 4555–4564, 2016. 6

[16] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *ICLR*, 2020. 3, 5

[17] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. Cross-modal video moment retrieval with spatial and language-temporal attention. In *ICMR*, pages 217–225, 2019. 1, 2

[18] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *NIPS*, pages 3294–3302, 2015. 6

[19] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 2019. 3

[20] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. *arXiv preprint arXiv:1903.12314*, 2019. 2

[21] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *SIGIR*, pages 15–24, 2018. 1, 2, 5, 6

[22] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *MM*, pages 843–851, 2018. 1, 2, 6

[23] Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. Learning cross-modal context graph for visual grounding. In *AAAI*, pages 11645–11652, 2020. 2

[24] Hahn Meera, Kadav Asim, M. Rehg James, and Peter Graf Hans. Tripping through time: Efficient localization of activities in videos. *arXiv preprint arXiv:1904-09936*, pages 1–13, 2019. 1, 2

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 6

[26] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, and Yang Liu. graph2vec: Learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005*, 2017. 5

[27] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Carlos Juan Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *CVPR*, pages 10867–10876, 2020. 2, 6

[28] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. Video relation detection with spatio-temporal graph. In *MM*, pages 84–93, 2019. 2

[29] Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. Gcc: Graph contrastive coding for graph neural network pre-training. *arXiv preprint arXiv:2006.09963*, 2020. 3

[30] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *TACL*, 1:25–36, 2013. 6

[31] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *ACL*, pages 91–99, 2015. 4

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the NeurIPS*, pages 5998–6008, 2017. 5

[33] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018. 4

[34] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, pages 334–343, 2019. 1

[35] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 2

[36] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, pages 9062–9069, 2019. 1, 2

[37] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, pages 10685–10694, 2019. 2

[38] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. MAN: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, pages 1247–1257, 2019. 2, 6

[39] Jiawei Zhang, Haopeng Zhang, Congying Xia, and Li Sun. Graph-bert: Only attention is needed for learning graph representations. *arXiv preprint arXiv:2001.05140*, 2020. 3

[40] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, pages 12870–12877, 2020. 2

[41] Songyang Zhang, Jinsong Su, and Jiebo Luo. Exploiting temporal relationships in video moment localization with natural language. In *MM*, pages 1230–1238, 2019. 2

[42] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*, pages 655–664, 2019. 2, 6

[43] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zhengjun Zha. Object relational graph with teacher-recommended learning for video captioning. In *CVPR*, pages 13278–13288, 2020. 2, 6

[44] Hao Zhou, Chongyang Zhang, and Chuanping Hu. Visual relationship detection with relative location mining. In *MM*, pages 30–38, 2019. 3