Jointly Cross- and Self-Modal Graph Attention Network for Query-Based Moment Localization

Daizong Liu[†]
School of Electronic Information and Communication, Huazhong University of Science and Technology
dzliu@hust.edu.cn

Xiaoye Qu[†]
Huawei Cloud, Hangzhou, China
quxiaoye@huawei.com

Xiao-Yang Liu
Department of Electrical Engineering, Columbia University
xl2427@columbia.edu

Jianfeng Dong
School of Computer and Information Engineering, Zhejiang Gongshang University
dongjf24@gmail.com

Pan Zhou[*]
The Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology
panzhou@hust.edu.cn

Zichuan Xu
School of Software, Dalian University of Technology
z.xu@dlut.edu.cn

ABSTRACT

Query-based moment localization is a new task that localizes the best matched segment in an untrimmed video according to a given sentence query. In this localization task, one should pay more attention to thoroughly mine visual and linguistic information. To this end, we propose a novel **C**ross- and **S**elf-**M**odal **G**raph **A**ttention **N**etwork (CSMGAN) that recasts this task as a process of iterative messages passing over a joint graph. Specifically, the joint graph consists of **C**ross-**M**odal relation **G**raph (CMG) and **S**elf-**M**odal relation **G**raph (SMG), where frames and words are represented as nodes, and the relations between cross- and self-modal node pairs are described by an attention mechanism. Through parametric message passing, CMG highlights relevant instances across video and sentence, and then SMG models the pairwise relation inside each modality for frame (word) correlating. With multiple layers of such a joint graph, our CSMGAN is able to effectively capture high-order interactions between two modalities, thus enabling a further precise localization. Besides, to better comprehend the contextual details in the query, we develop a hierarchical sentence encoder to enhance the query understanding. Extensive experiments on two public datasets demonstrate the effectiveness of our proposed model, and GCSMAN significantly outperforms the state-of-the-arts. The code is available at https://github.com/liudaizong/CSMGAN.

CCS CONCEPTS

• **Information systems** → **Video search**; **Novelty in information retrieval**.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3414026

KEYWORDS

Query-based moment localization; cross-modal interaction graph; self-modal relation graph; hierarchical query embedding

ACM Reference Format:
Daizong Liu[†], Xiaoye Qu[†], Xiao-Yang Liu, Jianfeng Dong, Pan Zhou[*], and Zichuan Xu. 2020. Jointly Cross- and Self-Modal Graph Attention Network for Query-Based Moment Localization. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3394171.3414026

1 INTRODUCTION

Localizing activities in videos [16, 17, 19, 31, 44] is an important topic in multimedia information retrieval. However, in realistic scenario, YouTube videos normally contain complicated background contents, and cannot be directly indicated by a pre-defined list of action classes. To address this problem, query-based moment localization is proposed recently [1, 18] and attracts increasing interests from the multimedia community [27, 50]. It aims to ground the most relevant video segment according to a given sentence query. This task is challenging because most part of video contents are irrelevant to the query while only a short segment matches the sentence. Therefore, video and sentence information need to be deeply incorporated to distinguish the fine-grained details of different video segments and perform accurate segment localization.

Most existing methods [5, 26, 45, 46, 51] for this task focus on learning the cross-modal relations between video and sentence. Specifically, they develop attention based interaction mechanisms to enhance the video representation with sentence information. Meanwhile, few algorithms [6, 51] attempt to learn the self-modal relations. For example, Zhang *et al.* [51] leverage self-attention to capture long-range semantic dependencies just in video encoding. However, the cross- and self-modal relations are never jointly

[†]Equal Contribution.
[*]This work was supported in part by the National Natural Science Foundation of China (No. 61972448, No. 61902347), and Zhejiang Provincial Natural Science Foundation (No. LQ19F020002). Corresponding author: Pan Zhou.

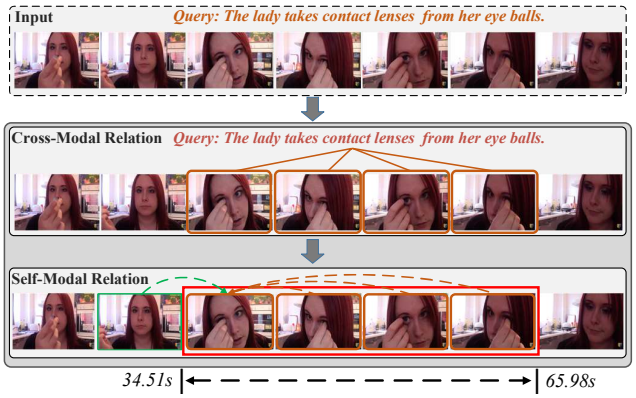

Figure 1: Given a query and an untrimmed video, CSMGAN considers cross-modal relation for highlighting relevant instances (brown rectangles), and self-modal relation for correlating sequential elements (red rectangle) and distinguishing components near the boundary (green rectangle).

investigated in a joint framework for this task. As shown in Figure 1, for the video modality, each frame should not only obtain information from its associated words in the query to highlight relevant frames (brown rectangles), but also need to correlate these highlighted frames to infer the sequential activity (red rectangle). At the same time, as the adjacent frame (green rectangle) near the boundary shows different visual appearance, such self-modal relation also contributes to distinguishing the segment boundaries for more precise localization. Similarly, for the query modality, a better understanding of sentence can be acquired in conjunction with both frames and other words. Such cases motivate us to propose a joint framework for modelling both cross- and self-modal relations.

In this paper, we develop a novel cross- and self-modal graph attention network (CSMGAN) for query-based moment localization, which recasts this task as an end-to-end, message passing based joint graph information fusion procedure. The joint graph consists of a cross-modal relation graph (CMG) and a self-modal relation graph (SMG), and represents both video frames and sentence words as nodes. Specially, in each joint graph layer, CMG first establishes the edges between each word-frame pair for cross-modal information passing, where the directed pair-wise relations are efficiently captured by a heterogeneous attention mechanism. Subsequently, SMG is designed to capture the complex self-modal relations by establishing the edges within each modality. The combination of CMG and SMG makes it possible to obtain more contextual representations by correlating highlighted cross-modal instances with sequential elements. Moreover, by stacking multiple layers to recursive propagate messages over the joint graph, our CSMGAN can capture higher-level relationships among multi-modal representations, and comprehensively integrates the localization information for precise moment retrieval.

Besides, traditional methods [9, 29, 39, 45–47] adopt RNN for sentence query embedding. However, they fail to explicitly consider the multi-granular textual information, such as specific phrases which are crucial to understanding the sentence. To capture the fine-grained query representations, we build a hierarchical structure to understand the query at three levels: word-, phrase- and sentence-level. These hierarchical representations are then merged to stand for a more informative understanding of the sentence query.

In summary, the main contributions of our work are:

- We present a cross- and self-modal graph attention network (CSMGAN), which is made up of cross- and self-modal graph for localizing desired moments. To the best of our knowledge, it is the first time that a joint framework is proposed to consider both cross- and self-modal relations for query-based moment localization.
- We design a hierarchical structure to capture the fine-grained sentence representation at three different levels: word-level, phrase-level and sentence-level.
- We conduct experiments on Activity Caption and TACoS datasets and our CSMGAN outperforms the state-of-the-arts with clear margins.

2 RELATED WORKS

Query-based localization in images. Early works of localization task mainly focus on localizing the image region corresponding to a language query. They first generate candidate image regions using image proposal method [32], and then find the matched one with respect to the given query. Some works [22, 28, 33] try to extract target image regions based on description reconstruction error or probabilities. There are also several studies [6, 7, 43, 48] considering incorporating contextual information of region-phrase relationship into the localization model. [40] further models region-region and phrase-phrase structures. Some other methods exploit attention modeling in queries, images, or object proposals [12, 15, 42].

Query-based moment localization in videos. Different from traditional video retrieval [13, 14], it is a new task introduced recently [1, 18], which aims to localize the most relevant video segment from a video with text descriptions. Traditional methods [18, 26] sample candidate segments from a video first, and subsequently integrate query with segment representations via a matrix operation. To further mine the cross-modal interaction more effectively, some works [9, 20, 41, 49] integrate the sentence representation with those video segments individually, and then evaluated their matching relationships. For instance, Xu *et al.* [41] introduce a multi-level model to integrate visual and textual features earlier and further re-generate queries as an auxiliary task. Chen *et al.* [5] capture the evolving fine-grained frame-by-word interactions between video and query to enhance the video representation understanding. Recently, other works [5, 29, 39, 45, 47, 51] propose to directly integrate sentence information with fine-grained video clip, and predict the temporal boundary of the target segment by gradually merging the fusion feature sequence over time. Zhang *et al.* [47] model relations among candidate segments produced from a convolutional neural network with the guidance of the query information. To modulate temporal convolution operations, Yuan *et al.*[45] and Mithun *et al.* [29] introduce the sentence information as a critical prior to compose and correlate video contents. Although these methods achieve relatively superior performances by capturing cross-modal information, they ignore to utilize the self-modal relation which is complementary to the cross-modal relation. Different from them, we propose a cross- and self-modal graph attention network to jointly consider both cross- and self-modal relations.

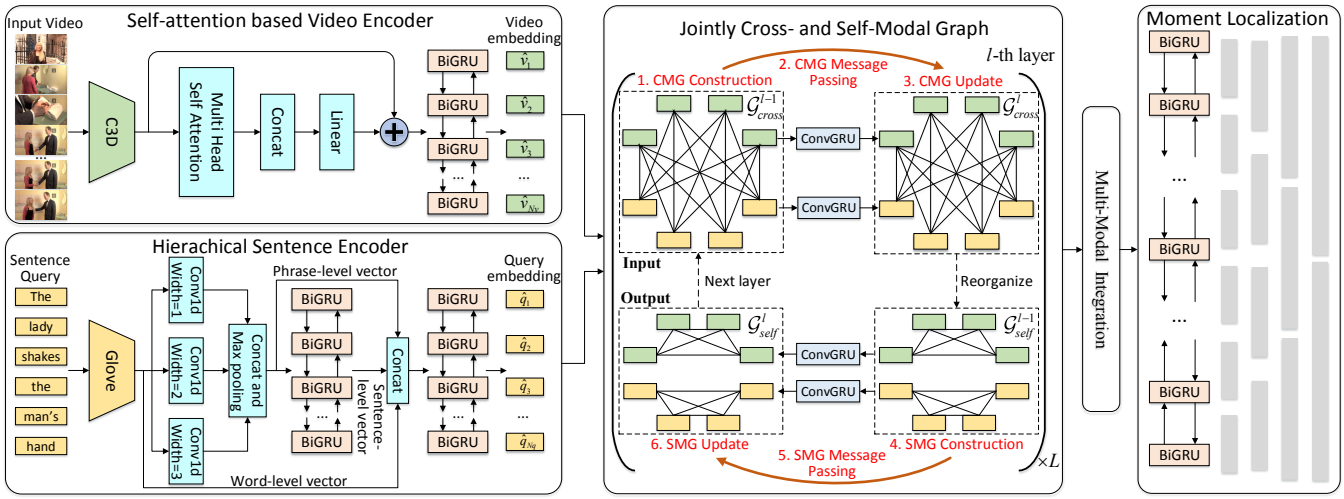

Figure 2: Illustration of our proposed CSMGAN. We first utilize a self-attention based video encoder and a hierarchical sentence encoder to extract corresponding features. Then, a jointly cross- and self-modal graph is devised for multi-modal interaction. In the joint graph, words and frames both represented as nodes first construct CMG to mine cross-modal relations and update their states through ConvGRU. Then the nodes are reorganized as SMG to model the self-modal relationships and updated for the next layer graph input. At last, we conduct multi-modal integration and perform moment localization.

The successive cross- and self-modal graphs enable our model to capture much higher-level interactions.

Graph neural networks. Graph neural network (GNN) [34] is an extension for recursive neural networks and random walk based models for graph structured data. As a follow-up work, Gilme *et al.* [21] further adapt GNN to sequential outputs with a learnable message passing module. As GNN is wildly used in sequential information processing, in this paper, we design a novel GNN module for cross- and self-modal relations mining. Different from original GNN, we represent edge weights by an attention mechanism and aggregate messages with a gate function. Moreover, we utilize a ConvGRU [2] layer for node state updating.

3 THE PROPOSED CSMGAN FRAMEWORK

3.1 Overview

Given an untrimmed video $\boldsymbol{V}$ and a sentence query $\boldsymbol{Q}$, the task aims to determine the start and end timestamps (s, e) of specific video segment referring to the sentence query. Formally, we represent the video as $\boldsymbol{V} = \{\boldsymbol{v}_t\}_{t=1}^{N_v}$ frame-by-frame, where $\boldsymbol{v}_i$ is the i-th frame in the video and N_v is the total frame number. We also denote the given sentence query as $\boldsymbol{Q} = \{\boldsymbol{q}_n\}_{n=1}^{N_q}$ word-by-word, where $\boldsymbol{q}_n$ is the n-th word. With the training set $\{V, Q, (s, e)\}$, we aim to learn to predict the most relevant video segment boundary $(\hat{s}, \hat{e})$ which conforms to the sentence query information.

We present our method CSMGAN in Figure 2. First of all, a self-attention based video encoder and a hierarchical sentence encoder are utilized to extract contextual sentence and video embeddings. Then, in order to better interact multi-modal features, we capture both cross- and self-modal relations by developing a jointly cross- and self-modal graph network. Specially, in the joint graph, a cross-modal relation graph (CMG) establishes weighted edges between frame-word pairs to pass the message flows across the modalities. Following it, a self-modal relation graph (SMG) reorganizes the previous nodes and edges to model the relationships within each modality. With the self-relation complemented to the cross-relation, the joint graph can perform richer interaction of multi-modalities. Moreover, with multiple layers of such joint graph, the CSMGAN can capture higher-order relationships. At last, the enhanced two modal representations are integrated to score different candidate video segments by a moment localization module.

3.2 Video and Sentence Encoder

Video encoder. Following [51], we first extract the frame-wise features by a pre-trained C3D network [36], and then employ a self-attention [37] module to capture the long-range dependencies among video frames. Considering the sequential characteristic in video, a bi-directional GRU [11] is further utilized to capture the contextual information in time series. We denote the encoded representation as $\hat{\boldsymbol{V}} = \{\hat{\boldsymbol{v}}_t\}_{t=1}^{N_v} \in \mathbb{R}^{N_v \times d}$, where the $\hat{\boldsymbol{v}}_t \in \mathbb{R}^{1 \times d}$ is the feature of the t-th frame.

Sentence encoder. Most previous works generally adopt recurrent neural networks to model the contextual information for each word during the sentence encoding process. However, considering the query "He continues playing the instrument", it is reasonable to focus on the phrase "continues playing" instead of each single word to obtain more detailed temporal clues for precise localization. Therefore, to fully mine the guiding information, we develop a hierarchical structure with word-, phrase-, and sentence-level feature extracting for sentence query encoding.

We first generate the word-level features for the query by using the Glove word2vec embedding [30], and denote them as $\boldsymbol{Q}^w = \{\boldsymbol{q}_n^w\}_{n=1}^{N_q} \in \mathbb{R}^{N_q \times d_g}$, where N_q is the number of words in the sentence and d_g is the Glove embedding dimension. To discover the potential phrase-level features, we apply 1D convolutions on the word-level features with different window sizes. Specially, at each

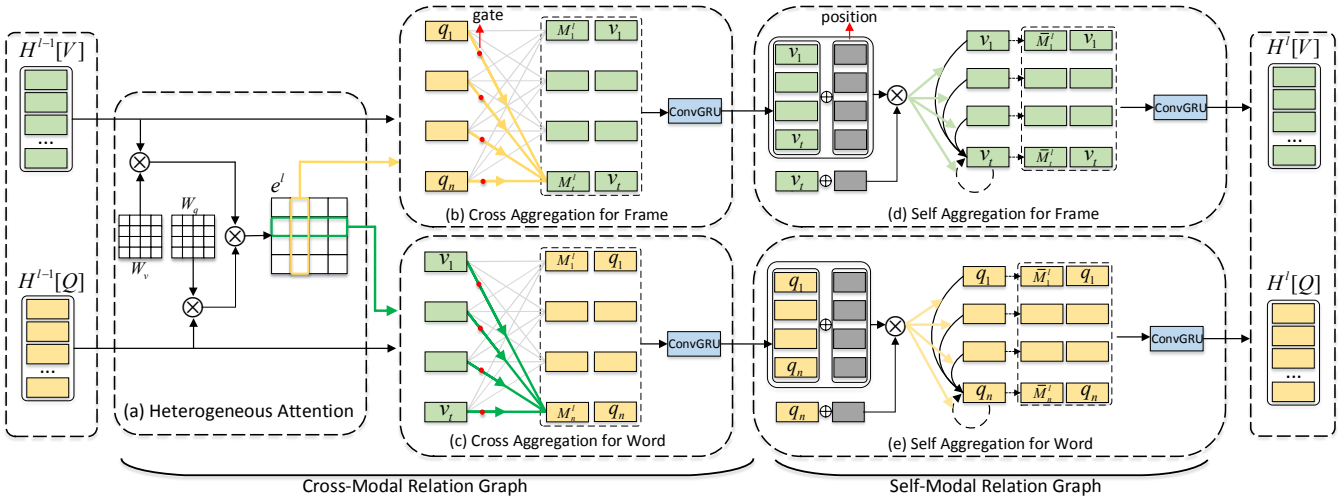

Figure 3: Illustration of our cross-modal relation graph and self-modal relation graph. Both two graphs compute attention matrices to stand for the attention weights on the corresponding edges. Each node aggregates the messages from its neighbor nodes in an edge-weighted manner and updates its state with both aggregated message and current state through ConvGRU. We apply a gate function in cross-modal graph and consider the temporal position for all nodes in self-modal graph.

word location, we compute the inner product of the word feature vectors with convolution filters of three kinds of window sizes, which captures unigram, bigram and trigram features. To maintain the sequence length after convolution process, we zero-pad the sequence vectors when convolution window size is larger than one. The output of the n-th word location with window size $k \in \{1, 2, 3\}$ is formulated as follows:

$$\boldsymbol{q}^p_{n,k} = \tanh(\text{Conv1d}(\boldsymbol{q}^w_{n:n+k-1})) \in \mathbb{R}^{1\times d_g}, \tag{1}$$

where Conv1d$(\cdot)$ operates on the windowed features with d_g kernels. $\boldsymbol{q}^p_{n,k}$ is the phrase-level feature corresponding to n-th word location with window size k. To find the most contributed phrase at each word location, we then apply max-pooling to obtain the final phrase-level feature $\boldsymbol{Q}^p = \{\boldsymbol{q}^p_n\}^{N_q}_{n=1} \in \mathbb{R}^{N_q\times d_g}$ by:

$$\boldsymbol{q}^p_n = \max(\boldsymbol{q}^p_{n,1}, \boldsymbol{q}^p_{n,2}, \boldsymbol{q}^p_{n,3}), n \in \{1, 2, ..., N_q\}. \tag{2}$$

After obtaining the phrase-level feature vector $\boldsymbol{Q}^p$, we encode them with a bi-directional GRU network to produce the sentence-level feature $\boldsymbol{Q}^s = \{\boldsymbol{q}^s_n\}^{N_q}_{n=1} \in \mathbb{R}^{N_q\times d_g}$. At last, we concat these three-level features and leverage another bi-directional GRU network to integrate them by:

$$\hat{\boldsymbol{Q}} = \text{Bi-GRU}(\text{Concat}[\boldsymbol{Q}^w, \boldsymbol{Q}^p, \boldsymbol{Q}^s]) \in \mathbb{R}^{N_q\times d}. \tag{3}$$

Here the contextual query representation $\hat{\boldsymbol{Q}} = \{\hat{\boldsymbol{q}}_n\}^{N_q}_{n=1}$ is projected from the length $3d_g$ to d to keep same dimension as video representation. After the hierarchical embedding structure, the given query can obtain comprehensive understanding and provide robust representation for later localization.

3.3 Jointly Cross- and Self-modal Graph

As shown in Figure 3, we develop a jointly cross- and self-modal graph to capture both cross- and self-modal information for multi-modal representation interaction. Specially, the joint graph consists of two subgraphs: cross-modal relation graph (CMG) and self-modal relation graph (SMG). In the CMG, each frame (word) integrates information from the other modality according to the cross-modal attentive relations. Subsequently, in the SMG, the self-attentive contexts within modality are further captured. By stacking L layers of such joint graphs, we can comprehensively perform the interaction between two modalities. Next, we will describe the detailed process of each subgraph in the l-th joint graph layer.

3.3.1 *Cross-Modal Relation Graph.*

Graph construction. In the CMG, we build a directed graph as $\mathcal{G}_{cross} = (\mathcal{V}_{cross}, \mathcal{E}_{cross})$, where $\mathcal{V}_{cross} = \boldsymbol{V} \cup \boldsymbol{Q} = \{\boldsymbol{v}_t\}^{N_v}_{t=1} \cup \{\boldsymbol{q}_n\}^{N_q}_{n=1}$, containing all frames and words as nodes, and $\mathcal{E}_{cross}$ is the edge set between all word-frame node pairs, namely edge $\boldsymbol{e}_{(n,t)} = (\boldsymbol{q}_n, \boldsymbol{v}_t)$ represents the cross-modal interaction from word node $\boldsymbol{q}_n$ to frame node $\boldsymbol{v}_t$ and $\boldsymbol{e}_{(t,n)} = (\boldsymbol{v}_t, \boldsymbol{q}_n)$ denotes the reverse interaction. To initialize the input features for each node, we set the encoded video representation of nodes $\boldsymbol{V}$ and query representation of nodes $\boldsymbol{Q}$ as initial hidden states: $\text{H}^0(\boldsymbol{V}) = \hat{\boldsymbol{V}}$ and $\text{H}^0(\boldsymbol{Q}) = \hat{\boldsymbol{Q}}$ in the CMG, respectively.

Cross-modal attention. To update CMG, the first step is to compute the attention weights between frame and word nodes $\boldsymbol{V}$, $\boldsymbol{Q}$ which represent their pair-wise relations. As in Figure 3 (a), the attention weight on each pair-wise edge can be computed as below:

$$\boldsymbol{e}^l = (\text{H}^{l-1}(\boldsymbol{Q}) \cdot \boldsymbol{W}_{\boldsymbol{q}})(\text{H}^{l-1}(\boldsymbol{V}) \cdot \boldsymbol{W}_{\boldsymbol{v}})^T \in \mathbb{R}^{N_q\times N_v}. \tag{4}$$

$\text{H}^{l-1}(\boldsymbol{Q})$ and $\text{H}^{l-1}(\boldsymbol{V})$ are the feature vectors for word and frame nodes in $(l-1)$-th layer. As $\boldsymbol{Q}$ and $\boldsymbol{V}$ come from different feature distributions, $\boldsymbol{W}_{\boldsymbol{q}}, \boldsymbol{W}_{\boldsymbol{v}} \in \mathbb{R}^{d\times d}$ are linear projection used to embed the heterogeneous nodes [23] into a joint latent space instead of direct computing in the node embedding space. Each row of $\boldsymbol{e}^l$ denotes the similarity of all frame nodes $\boldsymbol{V}$ to the specific word node $\boldsymbol{q}_n$, and each column represents the similarity of all word nodes $\boldsymbol{Q}$ to the specific frame node $\boldsymbol{v}_t$.

Node message aggregation. For message aggregation, we aggregate the assigned features for each node from its neighbors in an edge-weighted manner [38]. Here we introduce Figure 3 (b) which aggregates all neighboring word nodes $\{\boldsymbol{q}_n\}_{n=1}^{N_q}$ for frame node $\boldsymbol{v}_t$, and Figure 3 (c) performs the reverse aggregation process. For word node $\boldsymbol{q}_n$ in $\boldsymbol{Q}$, the assigned feature to $\boldsymbol{v}_t$ is:

$$\boldsymbol{M}^l_{(n,t)} = \text{Softmax}(\boldsymbol{e}^l_{(n,t)})\text{H}^{l-1}(\boldsymbol{q}_n) \in \mathbb{R}^{1\times d}. \tag{5}$$

The softmax procedure makes the sum of all word nodes' attention vectors to one. However, not all neighborhood nodes share same semantic importance, several neighborhood nodes contribute less to target node. For example, word "the" in the query is not informative enough to the frame, and the frames only containing one stationary basketball should have less significance to highlight "playing basketball". To emphasize informative neighborhood nodes and weaken inessential ones, we apply a learnable gate function $G(\cdot)$ to measure the confidence of each neighbor message by:

$$\boldsymbol{g}^l_{(n,t)} = G(\boldsymbol{M}^l_{(n,t)}) = \sigma(\boldsymbol{M}^l_{(n,t)}\boldsymbol{W}_g + \boldsymbol{b}_g) \in (0,1), \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function, $\boldsymbol{W}_g \in \mathbb{R}^{d\times 1}$ and $\boldsymbol{b}_g \in \mathbb{R}^1$ are the trainable weight parameter and bias. Then, we can aggregate the gated messages for node $\boldsymbol{v}_t$ by:

$$\boldsymbol{M}^l_t = \sum_{n=1}^{N_q} \boldsymbol{g}^l_{(n,t)}\boldsymbol{M}^l_{(n,t)} \in \mathbb{R}^{1\times d}. \tag{7}$$

With the help of such gate mechanism, the irrelevant aggregated messages are filtered and messages from relevant node pairs are further enhanced.

Node representation update. After aggregating the information from all neighbors, node $\boldsymbol{v}_t$ gets a new state by taking into account its current state and its received messages $\boldsymbol{M}^l_t$. To preserve the sequential information conveyed in the prior state and the messages, we do not utilize a simple element-wise matrix addition on $\text{H}^{l-1}(\boldsymbol{v}_t)$ and $\boldsymbol{M}^l_t$. Instead, we leverage a ConvGRU [2] layer to update the node state with two inputs by:

$$\bar{\text{H}}^l(\boldsymbol{v}_t) = \text{ConvGRU}(\text{H}^{l-1}(\boldsymbol{v}_t), \boldsymbol{M}^l_t) \in \mathbb{R}^{1\times d}. \tag{8}$$

This ConvGRU is proposed as a convolutional counterpart to original fully connected GRU [10]. In the same way, the representations for nodes of two modalities can be updated.

3.3.2 *Self-Modal Relation Graph.*

Graph construction. Following CMG, our SMG aims to capture the complex self-modal relations within each modality. It only connects edges between word-word or frame-frame node pairs. Like CMG, we denote this graph as $\mathcal{G}_{self} = (\mathcal{V}_{self}, \mathcal{E}_{self})$, where $\mathcal{V}_{self}$ is the node set containing all frames and words, and each edge in edge set $\mathcal{E}_{inter}$ indicates the self-modal relation.

Self-modal attention. Figure 3 (d) and (e) depict the process of self-modal information passing. Given a node $\boldsymbol{v}_t$ in Figure 3 (d) for example, we first compute a self-attention matrix to stand for the relations from its neighbor nodes to itself. To better correlate the relevant nodes, in this stage, we consider both the semantic information as well as the temporal position in the sequence of each node. We argue that the temporal index of the node is critical to our localization task as less attention should be given to distant frame (word) nodes, even if they are semantically similar to the current node. Inspired by TransformerâĂŹs positional encoding [37], we denote the position encoding for each node as:

$$\text{PE}(\boldsymbol{v}_t) = \begin{cases} \sin(t/10000^{j/d}), & \text{if } j \text{ is even} \\ \cos(t/10000^{j/d}), & \text{if } j \text{ is odd} \end{cases}, \tag{9}$$

where $\text{PE}(\boldsymbol{v}_t) \in \mathbb{R}^{1\times d}$, and j varies from 1 to d dimension. With the positional and semantic information combined, the self-attention matrix can be calculated by:

$$\bar{\boldsymbol{e}}^l_t = (\bar{\text{H}}^l(\boldsymbol{v}_t) + \alpha\text{PE}(\boldsymbol{v}_t))(\bar{\text{H}}^l(\boldsymbol{V}) + \alpha\text{PE}(\boldsymbol{V}))^T, \tag{10}$$

where $\bar{\boldsymbol{e}}^l_t \in \mathbb{R}^{1\times N_v}$ calculates the similarity for all frame nodes $\boldsymbol{V}$ to current nodes $\boldsymbol{v}_t$, α is to balance the two types of information.

Node message aggregation. The aggregation process can be formulated as following:

$$\bar{\boldsymbol{M}}^l_t = \text{Softmax}(\bar{\boldsymbol{e}}^l_t)\bar{\text{H}}^l(\boldsymbol{V}) \in \mathbb{R}^{1\times d}. \tag{11}$$

Here we only aggregate the information from $\bar{\text{H}}^l(\boldsymbol{v}_t)$ as positional information is designed for auxiliary similarity computing.

Node representation update. Similar to CMG, we also exploit a ConvGRU layer to update the node state and get the output as:

$$\text{H}^l(\boldsymbol{v}_t) = \text{ConvGRU}(\bar{\text{H}}^l(\boldsymbol{v}_t), \bar{\boldsymbol{M}}^l_t) \in \mathbb{R}^{1\times d}. \tag{12}$$

At last, following the same procedure, we can get the final representations for all nodes of each modality in the l-th jointly cross- and self-modal graph layer as $\text{H}^l(\boldsymbol{V}) \in \mathbb{R}^{N_v\times d}$ and $\text{H}^l(\boldsymbol{Q}) \in \mathbb{R}^{N_q\times d}$. Subsequently, these two modal features will be feed to the $(l+1)$-th joint graph layer as input.

3.4 Multi-Modal Integration Module

After L joint graph layer, we can get mutual sentence-aware video representation $\tilde{\boldsymbol{V}} = \{\tilde{\boldsymbol{v}}_t\}_{t=1}^{N_v} = \text{H}^L(\boldsymbol{V})$ and video-aware sentence representation $\tilde{\boldsymbol{Q}} = \{\tilde{\boldsymbol{q}}_n\}_{n=1}^{N_q} = \text{H}^L(\boldsymbol{Q})$. To integrate both two representations, we first compute the cosine similarity between each pair of word feature $\tilde{\boldsymbol{q}}_n$ and frame feature $\tilde{\boldsymbol{v}}_t$ as:

$$c_{n,t} = \frac{(\tilde{\boldsymbol{v}}_t)(\tilde{\boldsymbol{q}}_n\boldsymbol{W}_c)^T}{\parallel \tilde{\boldsymbol{v}}_t \parallel_2 \parallel \tilde{\boldsymbol{q}}_n\boldsymbol{W}_c \parallel_2}, \tag{13}$$

where $\boldsymbol{W}_c \in \mathbb{R}^{d\times d}$ is the linear parameter. $c_{n,t}$ is used to further extract the implicit query clues for each frame. Once we get the similarity scores between entire words $\{\tilde{\boldsymbol{q}}_n\}_{n=1}^{N_q}$ and a specific frame $\tilde{\boldsymbol{v}}_t$, we integrate the query information for frame $\tilde{\boldsymbol{v}}_t$ as:

$$\boldsymbol{h}_t = \sum_{n=1}^{N_q} \text{Softmax}(c_{n,t})\tilde{\boldsymbol{q}}_n \in \mathbb{R}^{1\times d}, \tag{14}$$

where $\boldsymbol{h}_t$ is aggregated with the query representation relevant to the t-th frame. We concat such aggregated query feature vectors with frame features to get the final multi-modal semantic representations $\boldsymbol{f} = \{\boldsymbol{f}_t\}_{t=1}^{N_v}$, where $\boldsymbol{f}_t = \text{Concat}[\tilde{\boldsymbol{v}}_t, \boldsymbol{h}_t] \in \mathbb{R}^{1\times 2d}$.

3.5 Moment Localization

We first apply a bi-directional GRU network on $\boldsymbol{f}$ to further absorb the contextual evidences in temporal domain. To predict the target video segment, we pre-define a set of candidate moments $\Phi_t = \{(\hat{s}_{t,i}, \hat{e}_{t,i})\}_{i=1}^{N_\Phi}$ with multi-scale windows [45] at each time t, where

N_Φ is the number of moments at current time-step. Then, we need to score these candidate moments and predict the offsets $\hat{\delta}_t = \{(\hat{\delta}^s_{t,i}, \hat{\delta}^e_{t,i})\}_{i=1}^{N_\Phi}$ of them relative to the ground-truth. In details, we produce the confidence scores $cs_t = \{cs_{t,i}\}_{i=1}^{N_\Phi}$ for these moments at time t by a Conv1d layer:

$$cs_{t,i} = \sigma(\text{Conv1d}(\boldsymbol{f}_t)) \in (0,1), \tag{15}$$

where $\sigma(\cdot)$ is the sigmoid function. The temporal offsets are predicted by another Conv1d layer:

$$(\hat{\delta}^s_{t,i}, \hat{\delta}^e_{t,i}) = \text{Conv1d}(\boldsymbol{f}_t). \tag{16}$$

Therefore, the final predicted moment i of time t can be presented as $(\hat{s}_{t,i} + \hat{\delta}^s_{t,i}, \hat{e}_{t,i} + \hat{\delta}^e_{t,i})$.

Training. We first compute the IoU (Intersection over Union) score $IoU_{t,i}$ between each candidate moment $(\hat{s}_{t,i}, \hat{e}_{t,i})$ with the ground truth (s_t, e_t). If $IoU_{t,i}$ is larger than an IoU threshold τ, we treat this candidate moment as a positive sample. We adopt an alignment loss to learn the confidence scoring rule for candidate moments, where the moments with higher IoUs will get higher confidence scores. The alignment loss function can be formulated as follows:

$$\mathcal{L}_{align} = -\frac{1}{N_v N_\Phi} \sum_{t=1}^{N_v} \sum_{i=1}^{N_\Phi} IoU_{t,i} log(cs_{t,i}) + (1 - IoU_{t,i}) log(1 - cs_{t,i}). \tag{17}$$

Since parts of the pre-defined candidates are coarse in boundaries, we only fine-tune the localization offsets of positive moment samples by a boundary loss:

$$\mathcal{L}_b = \frac{1}{N_{pos}} \sum_{j}^{N_{pos}} \mathcal{R}_1(\hat{\delta}^s_j - \delta^s_j) + \mathcal{R}_1(\hat{\delta}^e_j - \delta^e_j), \tag{18}$$

where N_{pos} denotes the number of positive moments, and $\mathcal{R}_1$ is the smooth L1 loss. Therefore, the joint loss can be represented as:

$$\mathcal{L} = \mathcal{L}_{align} + \beta \mathcal{L}_b, \tag{19}$$

where β is utilized to control the balance.

Inference. We first rank all candidate moments according to their predicted confidence scores, and then adopt a non-maximum suppression (NMS) to select "Top n" moments as the prediction.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

Activity Caption. Activity Caption [24] contains 20k untrimmed videos with 100k descriptions from YouTube. The videos are 2 minutes on average, and the annotated video clips have much larger variation, ranging from several seconds to over 3 minutes. Since the test split is withheld for competition, following public split, we adopt "val 1âĂİ as validation subset, "val 2âĂİ as our test subset.

TACoS. TACoS [31] is widely used on this task and contain 127 videos. The videos from TACoS are collected from cooking scenarios, thus lacking the diversity. They are around 7 minutes on average. We use the same split as [18], which includes 10146, 4589, 4083 query-segment pairs for training, validation and testing.

Evaluation Metrics. Following previous works [18, 45], we adopt "R@n, IoU=mâĂİ as our evaluation metrics. The âĂIJR@n, IoU=mâĂİ is defined as the percentage of at least one of top-n selected moments having IoU larger than m. Following [26, 39, 45], we choose the evaluation criteria âĂIJR@n, IoU=mâĂİ with n $\in \{1, 5\}$, m $\in \{0.3, 0.5, 0.7\}$ and âĂIJR@n, IoU=mâĂİ with n $\in \{1, 5\}$, m $\in \{0.1, 0.3, 0.5\}$ for Activity Caption and TACoS datasets, respectively.

Table 1: Performance compared with previous methods on the Activity Caption dataset.

Method	R@1 IoU=0.3	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.3	R@5 IoU=0.5	R@5 IoU=0.7
MCN [1]	39.35	21.36	6.43	68.12	53.23	29.70
TGN [5]	45.51	28.47	-	57.32	43.33	-
CTRL [18]	47.43	29.01	10.34	75.32	59.17	37.54
ACRN [26]	49.70	31.67	11.25	76.50	60.34	38.57
QSPN [41]	52.13	33.26	13.43	77.72	62.39	40.78
CBP [39]	54.30	35.76	17.80	77.63	65.89	46.20
SCDM [45]	54.80	36.75	19.86	77.29	64.99	41.53
ABLR [46]	55.67	36.79	-	-	-	-
GDP [8]	56.17	39.27	-	-	-	-
CMIN [51]	63.61	43.40	23.88	80.54	67.95	50.73
CSMGAN	**68.52**	**49.11**	**29.15**	**87.68**	**77.43**	**59.63**

Table 2: Performance compared with previous methods on the TACoS dataset.

Method	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
MCN [1]	3.11	1.64	1.25	3.11	2.03	1.25
CTRL [18]	24.32	18.32	13.30	48.73	36.69	25.42
ABLR [46]	34.70	19.50	9.40	-	-	-
ACRN [26]	24.22	19.52	14.62	47.42	34.97	24.88
QSPN [41]	25.31	20.15	15.23	53.21	36.72	25.30
TGN [5]	41.87	21.77	18.90	53.40	39.06	31.02
GDP [8]	39.68	24.14	13.50	-	-	-
CMIN [51]	32.48	24.64	18.05	62.13	38.46	27.02
SCDM [45]	-	26.11	21.17	-	40.16	32.18
CBP [39]	-	27.31	24.79	-	43.64	37.40
CSMGAN	**42.74**	**33.90**	**27.09**	**68.97**	**53.98**	**41.22**

4.2 Implementation Details

For training our CSMGAN, we first resize every frame of videos to 112×112 pixels as input, and then apply a pre-trained C3D [36] to obtain 4096 dimension features. After that we apply PCA to reduce the feature dimension from 4096 to 500 for decreasing the model parameters. These 500-d features are used as the frame features in our model. Since some videos are overlong, we set the length of video feature sequences to 200 for both Activity Caption and TACoS datasets. As for sentence encoding, we utilize Glove word2vec [30] to embed each word to 300 dimension features. The hidden state dimension of BiGRU networks is set to 512. We set α to 1 for positional encoding. During moment localization, we adopt convolution kernel size of [16, 32, 64, 96, 128, 160, 192] for Activity Caption, and [8, 16, 32, 64] for TACoS. We set the stride of them as 0.5, 0.125, respectively. We then set the high-score threshold τ to 0.45, and the balance hyper-parameter β to 0.001 for Activity Caption, 0.005 for TACoS. The number of our joint graph layer is set to 2. We train our model with an Adam optimizer with leaning rate 8×10^{-4}, 3×10^{-4} for Activity Caption and TACoS, respectively. The batch size is set to 128 and 64 for two datasets, respectively.

Table 3: Ablation study on the Activity Caption and TACoS datasets, where the reference is our full model.

Components	Module	Activity Caption						TACoS					
		R@1 IoU=0.3	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.3	R@5 IoU=0.5	R@5 IoU=0.7	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
Reference	**full**	**68.52**	**49.11**	**29.15**	**87.68**	**77.43**	**59.63**	**42.74**	**33.90**	**27.09**	**68.97**	**53.98**	**41.22**
Encoder	**w/o HS**	66.32	46.80	26.54	85.97	74.43	56.49	39.56	30.42	24.40	65.50	51.39	39.38
Joint Graph	**w/o CSG**	64.13	44.47	25.49	84.35	72.97	54.47	36.91	28.45	22.27	63.18	49.19	37.11
Cross-Modal Graph	**w/o EM**	67.48	47.94	28.09	86.02	75.27	56.32	41.11	32.24	25.93	66.67	53.05	40.17
	w/o MG	67.28	47.39	28.13	86.46	74.91	56.47	40.48	31.66	25.73	66.39	52.56	40.21
Self-Modal Graph	**w/o SMG**	66.53	46.62	27.57	85.68	73.97	55.68	39.64	30.86	24.61	65.10	50.90	39.11
	w/o PE	67.41	48.45	28.56	86.51	75.20	57.20	40.23	31.32	25.30	66.17	51.41	39.43
Node Update	**w/o CG**	67.37	47.51	28.07	86.06	75.66	56.96	40.97	31.96	25.66	66.42	52.00	40.04

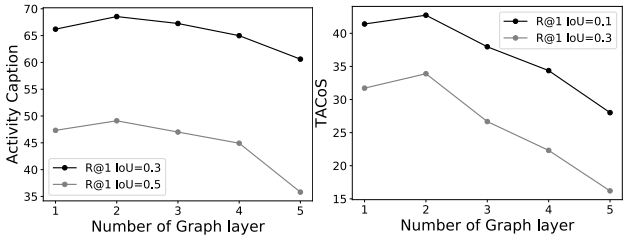

Figure 4: Effect of the number of graph layers on the Activity Caption and TACoS Datasets.

4.3 Performance Comparison and Analysis

Activity Caption. Table 1 shows the performance evaluation results of our method and all comparing methods on Activity Caption dataset. Compared to the state-of-the-arts methods, our model surpasses them with clear margin on all R@1 and R@5 metrics. Specially, our method brings 5.27% and 8.90% absolute improvements in the strict metrics "R@1, IoU=0.7âĂİ and "R@5, IoU=0.7âĂİ.
TACoS. Table 2 shows the performance results of our method and all baselines on TACoS dataset. On this challenging dataset, we find that our method still achieves significant improvements. In details, our method brings 2.30% and 3.82% improvements in the strict metrics"R@1, IoU=0.5âĂİ and "R@5, IoU=0.5âĂİ, respectively.
Analysis. Specifically, the compared methods can be divided into two classes: 1) Sliding window based methods: MCN [1], CTRL [18], and ACRN [26] first sample candidate video segments using sliding windows, and directly integrate query representations with window-based segment representations via a matrix operation. They do not employ a comprehensively structure for effective cross-modal interaction, leading to relatively lower performances than other methods. 2) Cross-modal interaction based methods: TGN [5], QSPN [41], CBP [39], SCDM [45], ABLR [46], GDP [8], and CMIN [51] integrate query representations with the whole video representations in an attention-guided manner, and can generate contextual query-guided video representation for precisely boundary localization. However, they ignore to capture the self-modal relation which helps to correlate relevant instances within each modality. Compared to them, our method emphasizes the importance of capturing both cross- and self-modal relation during the effective integration of multi-modal feature. Our jointly cross- and self-modal graph can mine much richer and higher-level interactions, thus achieving better results than both two kinds of methods.

4.4 Ablation Study

In this section, we perform ablation studies to examine the effectiveness of our proposed CSMGAN. Specifically, we re-train our model with the following settings:

- **w/o HS**: We first remove the hierarchical structure from the sentence encoder, and only take a bi-directional GRU to encode the sentence query.
- **w/o CSG**: We then discard the jointly cross- and self-modal graph to validate the importance of cross- and self-modal relations capturing in multi-modal interaction.
- **w/o EM**: To explore the effect of heterogeneous attention in CMG, we remove the embedding matrices in Eq. 4, and compute the attentive matrix in the node embedding space.
- **w/o MG**: To further analyze the gate mechanism in CMG, we remove the gate function during the message passing in the cross-modal graph layer.
- **w/o SMG**: To evaluate the effect of SMG, we remove the self-modal graph, and only apply cross-modal graph for multi-modal interaction.
- **w/o PE**: To assess the component of SMG, we remove the positional encoding from the SMG.
- **w/o CG**: Finally, we replace the ConvGRU with a simple matrix element-wise addition during the node updating.
- **full**: The full model.

The ablation study conducted on Activity Caption and TACoS datasets are shown in Table 3. By analyzing the ablation results, we have the conclusions as follows:

- First of all, our full model outperforms all the ablation models on both two datasets, which demonstrates each component is definitely helpful for this task.
- Compared to other ablation models, the w/o CSG model performs worst on two datasets. This means that our jointly cross- and self-modal graph takes an important role in effective multi-modal features interaction. Besides, the hierarchical structure (HS) for sentence embedding also has significant contribution to the full model.
- At last, almost all ablation models still yield better results than all state-of-the-arts methods. This fact demonstrates that the excellent performance of our graph based framework does not only rely on one specific key component, and our full model is robust to address this task.

Figure 5: Visualization on the edge weights of both cross-modal graph (CMG) and self-modal graph (SMG). Left: column-wise weights of the attention matrix of different CMG layer, where the weights stands for the relations from all words to a specific frame. Right: self-attention weights of different SMG layer, where the relevant frames have higher context weights.

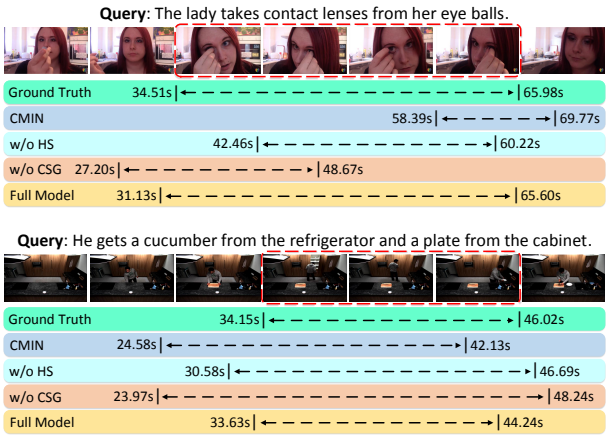

Figure 6: Qualitative visualization on both two datasets (top: Activity Caption, bottom: TACoS).

To further investigate the influence on the various number of our joint graph layers, we show the impact of different layer numbers on two datasets in Figure 4. We can observe that our model achieves best result when the number of layer is set to 2. Then the performance will drop if the number of layers increases. The propagated messages between the instances in cross-modality and self-modality will be accumulated if we use more graph layers, resulting in over-smoothing [25] problem, namely the representations of both video and sentence converge to the same value.

4.5 Qualitative Results

To qualitatively validate the effectiveness of our method, we show examples from two datasets in Figure 6. Although the sentences are very diverse, our full model can still localize more accurate boundaries than CMIN. In two variant models, the w/o CSG has the most coarse boundaries because it lacks the detailed interaction of multi-modal features. The w/o HS fails to capture more contextual sentence guiding clues for localization, leading to relatively coarse boundaries. As a comparison, our full model achieves the most precise localization.

We further give a deep visualization on the cross- and self-modal relations in each joint graph layer. Specially, we first visualize the relations from all words to a specific frame in the cross-modal graph. As shown in Figure 5 (left), sentence "she begins brushing the horse while still speaking" has two activities. For the non-relevant frame, the attention weights on these eight words are more inclined to be an even distribution. But for the relevant frame, the contributed words like "begins", "brushing", "still", and "speaking" obtain higher attention weights since the described action indeed happens there. Moreover, with GNN layer increasing, the distribution of these words weights are sharper and more distinguishable. However, too many GNN layers will result in over-smoothing problem, where each frame-word pair has almost the same activation. We also plot the context weights in the self-modal graph as shown in Figure 5 (right). The weights are calculated by a softmax function, and they represent the relations from surrounding frames to one specific frame. We find that frame containing "brushing while speaking" is more relevant to the frame "begin brushing". Although the frames near the segment boundaries are visually similar to the frames in the segment, the self-modal relation can effectively distinguish them and produce lower attention weights for such noisy frames.

5 CONCLUSION

In this paper, we propose a jointly cross- and self-modal graph attention network (CSMGAN) for query-based moment localization in video. We consider both cross- and self-modal relations in a joint framework to capture much higher-level interactions. Specially, cross-modal relation highlights relevant components across video and sentence, and then self-modal relation models the pairwise correlation inside each modality for frames/words association. Besides, we also develop a hierarchical structure for more contextual sentence understanding in a word-phrase-sentence process. The experimental results on various datasets demonstrate the effectiveness of our proposed method.

REFERENCES

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

[2] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. 2015. Delving deeper into convolutional networks for learning video representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 961–970.

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

[5] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 162–171.

[6] Kan Chen, Rama Kovvuri, Jiyang Gao, and Ram Nevatia. 2017. MSRC: Multimodal Spatial Regression with Semantic Context for Phrase Grounding. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. 23–31.

[7] Kan Chen, Rama Kovvuri, and Ram Nevatia. 2017. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 824–832.

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

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

[10] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[11] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Advances in Neural Information Processing Systems (NIPS)*.

[12] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. 2018. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7746–7755.

[13] Jianfeng Dong, Xirong Li, and Cees GM Snoek. 2018. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia* 20, 12 (2018), 3377–3388.

[14] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2019. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9346–9355.

[15] Ko Endo, Masaki Aono, Eric Nichols, and Kotaro Funakoshi. 2017. An Attention-based Regression Model for Grounding Textual Phrases in Images.. In *IJCAI*. 3995–4001.

[16] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Spatio-temporal video re-localization by warp lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1288–1297.

[17] Yang Feng, Lin Ma, Wei Liu, Tong Zhang, and Jiebo Luo. 2018. Video re-localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 51–66.

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

[19] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. 2018. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5958–5966.

[20] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. 2019. Mac: Mining activity concepts for language-based temporal localization. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 245–253.

[21] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 1263–1272.

[22] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4555–4564.

[23] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous Graph Transformer. *arXiv preprint arXiv:2003.01332* (2020).

[24] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 706–715.

[25] Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[26] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *Proceedings of the 41nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 15–24.

[27] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*. 843–851.

[28] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 11–20.

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

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[31] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics* 1 (2013), 25–36.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*. 91–99.

[33] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 817–834.

[34] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE Transactions on Neural Networks* 20, 1 (2008), 61–80.

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

[36] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 4489–4497.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*. 5998–6008.

[38] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

[40] Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, and Jia Deng. 2016. Structured matching for phrase localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 696–711.

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

[42] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1307–1315.

[43] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 69–85.

[44] Jun Yuan, Bingbing Ni, Xiaokang Yang, and Ashraf A Kassim. 2016. Temporal action localization with pyramid of score distribution features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3093–3102.

[45] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In *Advances in Neural Information Processing Systems (NIPS)*. 534–544.

[46] Yitian Yuan, Tao Mei, and Wenwu Zhu. 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*, Vol. 33. 9159–9166.

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

[48] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. 2018. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR). 4158–4166.

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

[50] Songyang Zhang, Jinsong Su, and Jiebo Luo. 2019. Exploiting Temporal Relationships in Video Moment Localization with Natural Language. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1230–1238.

[51] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 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.

A ADDITIONAL DATASETS

In this section, we provide detailed performance analysis on additional datasets, including Charades-STA [18] and DiDeMo [1].

A.1 Charades-STA

Dataset. Charades-STA is built on the Charades dataset [35], which focuses on indoor activities. As Charades dataset only provides video-level paragraph description, the temporal annotations of Charades-STA are generated in a semi-automatic way, which involves sentence decomposition, keyword matching, and human check. In total, the video length on the Charades-STA dataset is 30 seconds on average, and there are 12408 and 3720 moment-query pairs in the training and testing sets respectively.

Settings. Following previous settings [39, 45], we extract 1024 dimension features by I3D [4], and then apply PCA to reduce the feature dimension to 500 for decreasing the model parameters. During moment localization, we adopt convolution kernel size of $[16, 24, 32, 40]$, and set the stride as 0.25. We set the high-score threshold τ to 0.5, and the balance hyper-parameter β to 0.005.

Analysis. Table 4 shows the performance evaluation results of our CSMGAN and all comparing methods on Charades-STA dataset. Compared to the state-of-the-arts methods, our model surpasses them with clear margin both on R@1 and R@5 metrics. Specially, compared with the previous state-of-the-art method in absolute values, our method brings 3.91% and 3.77% improvements in the strict âĂIJR@1, IoU=0.7âĂİ and âĂIJR@5, IoU=0.7âĂİ metrics, respectively.

A.2 DiDeMo

Dataset. DiDeMo is recently proposed in [1], specially for natural language moment retrieval in open-world videos. DiDeMo contains 10464 videos with 33005, 4180 and 4021 annotated moment-query pairs in the training, validation and testing sets respectively. To annotate moment-query pairs, videos in DiDeMo are trimmed to a maximum of 30 seconds, divided into 6 segments of 5 seconds long each, and each moment contains one or more consecutive segments. Therefore, there are 21 candidate moments in each video and the task is to select the moment that best matches the query.

Settings. Following previous settings [26], we extract 4096 dimension features by C3D [36], and then apply PCA to reduce the feature dimension to 500 for decreasing the model parameters. During moment localization, we adopt convolution kernel size of $[16, 32, 64, 96]$, and set the stride as 0.25. We set the high-score threshold τ to 0.5, and the balance hyper-parameter β to 0.005.

Analysis. Table 5 shows the performance comparisons on the DiDeMo dataset. Compared to the state-of-the-arts methods, our model surpasses them with clear margin both on R@1 and R@5 metrics. Specially, compared to the previous state-of-the-art method in absolute values, our method brings 2.51% improvements in the strict âĂIJR@1, IoU=0.7âĂİ. At the same time, it is worth noticing that our CSMGAN achieves significant improvements (12.16%) in the âĂIJR@5, IoU=0.7âĂİ metrics.

B MORE QUALITATIVE RESULTS

In this section, we provide more qualitative visualization results of our method on widely used datasets:

Table 4: Performance compared with previous methods on the Charades-STA dataset.

Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
TGA [29]	17.04	6.93	58.17	26.80
MCN [1]	17.46	8.01	48.22	26.73
ACRN [26]	20.26	7.64	71.99	27.79
CTRL [18]	23.63	8.89	58.92	29.57
SAP [9]	27.42	13.36	66.37	38.15
QSPN [41]	35.60	15.80	79.40	45.40
CBP [39]	36.80	18.87	70.94	50.19
GDP [8]	39.47	18.49	-	-
SCDM [45]	54.44	33.43	74.43	58.08
CSMGAN	**60.04**	**37.34**	**89.01**	**61.85**

Table 5: Performance compared with previous methods on the DiDeMo dataset.

Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
MCN [1]	23.33	15.37	41.03	20.37
VSA-STV [18]	25.38	14.49	68.56	26.92
VSA-RNN [18]	24.94	14.52	68.39	26.10
CTRL [18]	26.45	15.36	68.78	28.43
ACRN [26]	27.44	16.65	69.43	29.45
CSMGAN	**29.44**	**19.16**	**70.77**	**41.61**

- Figure 7 shows results on Activity Caption [3] dataset.
- Figure 8 shows results on TACoS [31] dataset.
- Figure 9 shows results on Charades-STA [18] dataset.
- Figure 10 shows results on DiDeMo [1] dataset.

For each dataset, we present four kinds of localization results on two sample videos, including three variants of our method and ground truth. For precise localization, our CSG (Cross- and Self-model Graph) mines the deep interaction between two modalities, and relates the instances within each modality. Meanwhile, HS (hierarchical structure) also contributes sentence understanding for better grounding. It is obvious that our full model achieves most precise localization result with the help of these two modules, which demonstrates the effectiveness of our proposed CSMGAN.

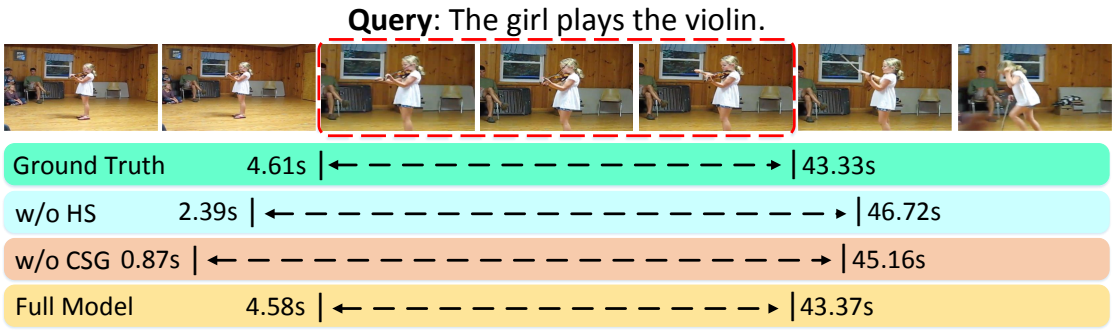

Figure 7: Qualitative visualization on Activity Caption dataset.

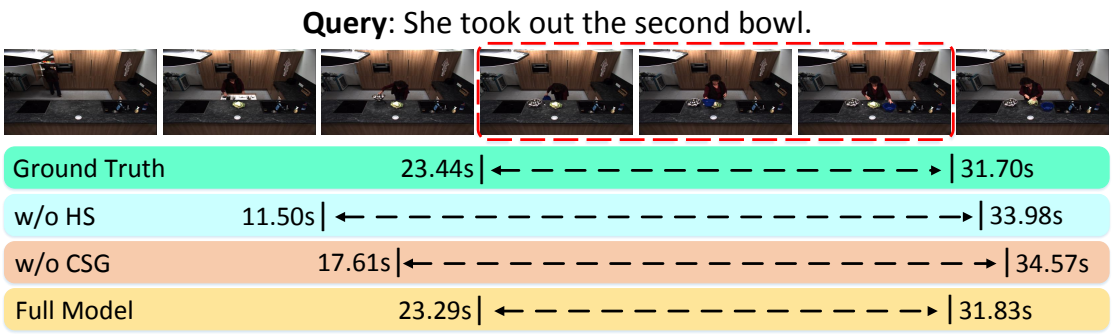

Figure 8: Qualitative visualization on TACoS dataset.

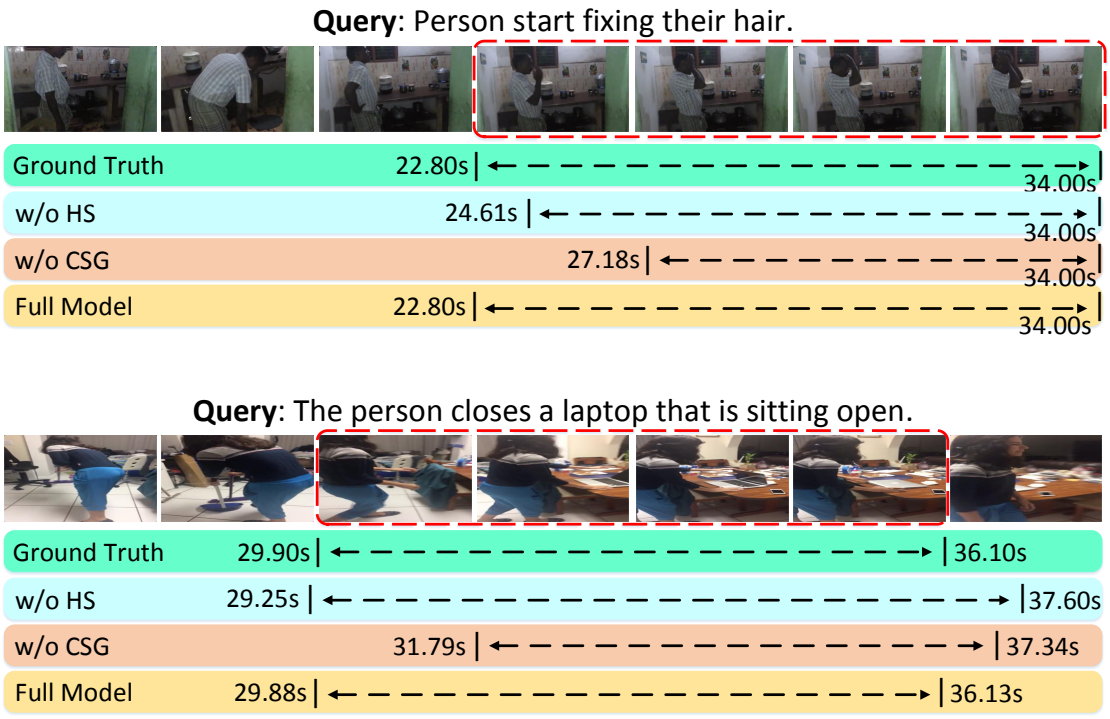

Figure 9: Qualitative visualization on Charades-STA dataset.

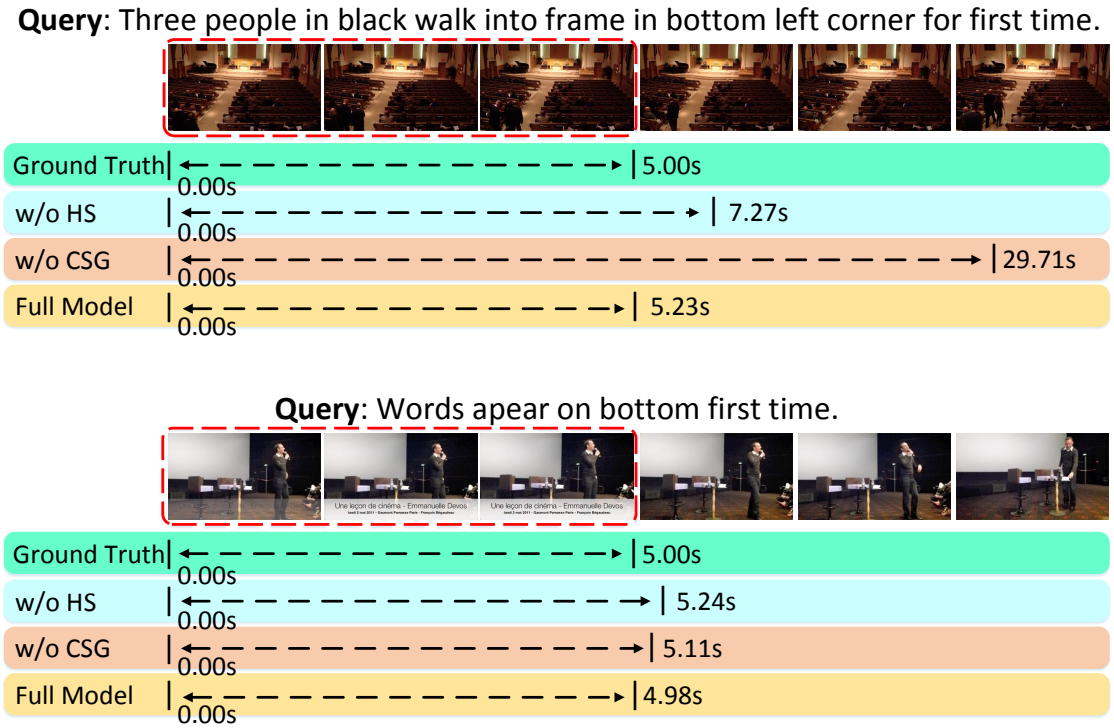

Figure 10: Qualitative visualization on DiDeMo dataset.