Text-based Temporal Localization of Novel Events

Sudipta Paul[1], Niluthpol Chowdhury Mithun[2], and Amit K Roy-Chowdhury[1]

[1] University of California, Riverside CA USA
[2] SRI International, Princeton NJ USA
{spaul, amitrc}@ece.ucr.edu, niluthpol.mithun@sri.com

Abstract. Recent works on text-based localization of moments have shown high accuracy on several benchmark datasets. However, these approaches are trained and evaluated relying on the assumption that the localization system, during testing, will only encounter events that are available in the training set (i.e., *seen* events). As a result, these models are optimized for a fixed set of seen events and they are unlikely to generalize to the practical requirement of localizing a wider range of events, some of which may be *unseen*. Moreover, acquiring videos and text comprising all possible scenarios for training is not practical. In this regard, this paper introduces and tackles the problem of text-based temporal localization of novel/unseen events. Our goal is to temporally localize video moments based on text queries, where *both the video moments and text queries are not observed/available during training*. Towards solving this problem, we formulate the inference task of text-based localization of moments as a relational prediction problem, hypothesizing a conceptual relation between semantically relevant moments, e.g., a temporally relevant moment corresponding to an unseen text query and a moment corresponding to a seen text query may contain shared concepts. The likelihood of a candidate moment to be the correct one based on an unseen text query will depend on its relevance to the moment corresponding to the semantically most relevant seen query. Empirical results on two text-based moment localization datasets show that our proposed approach can reach up to 15% absolute improvement in performance compared to existing localization approaches.

Keywords: Temporal Localization, Moment Retrieval, Novel/Unseen Events.

1 Introduction

Event localization in a long and untrimmed video is an important video analysis problem. Recently, there has been a surge of works that address the task of temporal grounding of text/sentence in untrimmed videos [3, 12, 29, 36, 71, 74]. Most of these works utilize a set of fully supervised training data containing videos, text descriptions, and temporal boundary annotations. These works try to optimize over a fixed set of events and queries (which we call seen events and seen queries) that are available during training. However, in a real-world dynamic environment, a system is expected to encounter *previously unseen events and queries*, as shown in Figure 1, and is required to *localize corresponding moments based on unseen text queries* in the videos. As a result, a system optimized over a fixed set of events is unlikely to generalize and perform well for unseen events. Moreover, as textual annotations are expensive and time consuming [35],

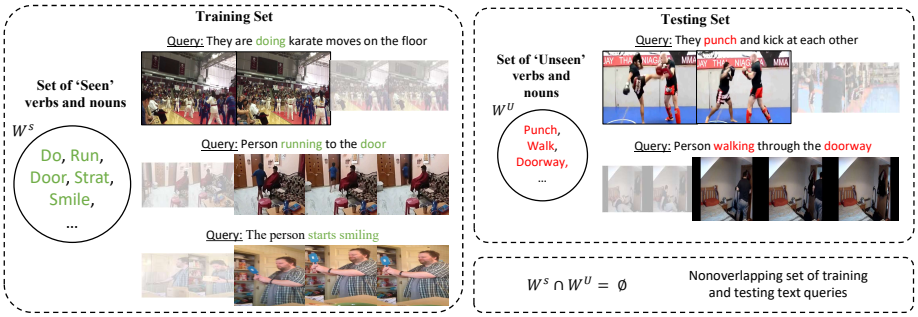

Fig. 1. Example illustration of our proposed task. We consider the task of localizing novel moments for unseen queries. The set of verbs and nouns present in the testing set is absent in the training set, e.g., training data does not have any text with verb 'walk' or noun 'doorway'. Hence, the system is required to learn transferable knowledge from the training data to perform localization for novel events based on unseen queries.

it is impossible to collect videos of all possible events and textual descriptions and learn models with the collected data. Hence, the applicability of current text-based temporal localization systems are severely limited to a small set of events and the problem of localizing novel/unseen events based on unseen text queries remains unaddressed in the current literature.

In this work, our goal is to temporally localize video moments based on text queries, where both the video moments and text queries *are not observed/available during training*. Towards this goal, we learn transferable knowledge from seen events and queries and utilize it to localize novel/unseen events. We hypothesize that temporally relevant moments corresponding to unseen text queries and those corresponding to seen text queries are likely to contain shared concepts, if the unseen query and the seen query are semantically relevant. For instance, in Figure 1, moment corresponding to the unseen text query *'They punch and kick at each other'* from the testing set has similarities to the moment corresponding to seen text query *'They are doing karate moves on the floor'* from the training set. Therefore, instead of localizing moments only based on its encoded representation, we formulate the inference task of localization as a relational prediction problem. The likelihood of a candidate moment to be the correct one based on an unseen text query depends on its relevance to the moment corresponding to the semantically most relevant seen query. We term this moment corresponding to the semantically most relevant seen query as the *support moment*. To learn a proper relational system that can localize novel events, we simulate the support moment based relational inference on the available training data during training. As a result, the system learns to localize moments based on relational reasoning, instead of directly localizing based on observed moment representations. Our motivation behind the approach is that a relational system learned on seen events/queries is transferable to the unseen events/queries [54]. We term our approach as **T**emporal **L**ocalization using **R**elational **R**easoning (**TLRR**).

Our problem is related to the zero-shot paradigm (where the objective is to adapt models to perform different tasks on the unseen or unobserved classes) as we utilize seen moment-text pairs to infer on the unseen events [26, 39, 64, 78, 87]. However, those zero-shot approaches are not directly applicable to our problem setup. For example, [79] assumes unseen classes are known in advance and uses the information to mine common semantics for seen classes and unseen classes for zero-shot temporal activity detection. However, text-based annotations of events are not limited to a fixed set of classes and the unseen queries are not known beforehand. Again, [8, 27, 68] perform retrieval across multiple modality data in the zero-shot setting. These works consider images with specific classes, and utilize the word embedding space to transfer knowledge between seen classes and unseen classes. However, in a video, textual descriptions refer to multiple entities, interactions of multiple entities, and different activities in a combined manner that is not expressible by a single class. As a result, directly utilizing label embeddings is not enough to transfer knowledge from seen events/queries to unseen events/queries. We will demonstrate the advantage of our proposed TLRR approach over zero-shot approaches and other recent temporal localization approaches on two benchmark datasets. The following are the main **contributions** of our work.

- We address a novel and practical problem of temporal localization of video moments based on unseen text queries.
- We hypothesize a conceptual relation between semantically relevant moments and propose a relational reasoning based temporal localization approach, TLRR, which can learn transferable knowledge from seen events and localize novel events based on unseen text queries.
- We reorganize two existing text-based temporal localization datasets (Charades-STA [12] and ActivityNet Captions [24]) for our proposed novel problem setting. Empirical results on these two text-based video moment localization datasets show that our proposed approach can reach up to 15% absolute improvement in performance compared to existing localization approaches.

2 Related Works

Temporal Localization of Moments. Temporal localization of moments in a video based on text query was introduced by [3, 12]. Recently, there are many works that address the problem both in presence of strong supervision (temporal endpoints are known for each query) [5, 6, 9, 13, 14, 15, 16, 17, 18, 21, 22, 30, 31, 32, 33, 34, 36, 38, 43, 47, 53, 57, 59, 60, 62, 65, 66, 67, 71, 72, 74, 75, 76, 77, 80, 81, 82, 83, 84, 86] and weak supervision (only video-text correspondence is known) [7, 28, 35, 55, 56, 61, 70]. Among the recent works on temporal localization of moments in the fully supervised setting, [71] performs semantic conditioned dynamic modulation, [74] relies on dense regression based approach, [36] utilizes both local and global interaction for video grounding. Recently, [37] proposed text-based temporal localization without query annotation. Unlike our setting, they have access to videos of all types of events and can optimize their model for such events in a weakly supervised manner. Hence, none of these works address the problem of localizing novel events based on unseen text queries.

Zero-shot Learning (ZSL). ZSL aims to do inference task on classes whose instances may not have been seen during training [26, 39, 64, 78, 87]. Initial works on ZSL were attribute-based [25, 41]. However, attribute-based ZSL has poor scalability and semantic embedding of labels are a good alternative for attributes [69]. Most of the works that utilize semantic embedding based learning focus on the association of visual and semantic information by linear compatibility [1, 2, 11, 48], non-linear compatibility [52, 63] or in a hybrid way [40]. To the best of our knowledge, only [79] works on activity detection in ZSL setup. However, [79] is limited to work on activity labels and can not be adapted directly for moment localization of unseen text queries.

Zero-shot Cross Modal Retrieval (ZS-CMR). Conventional cross modal retrieval work [10] considers similar type of events are present in both training set and testing set. However, ZS-CMR aims to perform retrieval across multiple modality data in the zero-shot setting. They train the retrieval model with limited categories to support cross-modal retrieval on new categories [27]. There are few works that consider retrieval between visual and textual modality with ZS-CMR setting [8, 27, 68]. However, these works are limited by the use of specific class information of the images to transfer knowledge between seen classes to unseen classes.

Relational Reasoning. Relational reasoning concept has been applied to different vision applications, i.e., visual question answering [46, 49], deep reinforcement learning [73], few-shot learning [54], self supervised learning [42], activity recognition [44, 85]. [54] is the closest to the proposed TLRR and uses relational reasoning for zero-shot learning. However, our work differs in several ways: (i) we do not work with a fixed set of labels, (ii) our relational module learns to identify relations between visual information rather than learning to identify relations between visual and semantic information, and (iii) our proposed problem setup requires the model to identify intra-video subtle differences between moments, whereas [54] learns to differentiate classes.

3 Methodology

3.1 Problem Statement

Let $\mathcal{S}^{tr} = \{(v, q, (\tau_s, \tau_e))|v \in \mathcal{V}^{tr}, q \in \mathcal{Q}^{tr}, \tau_s, \tau_e \in [0, T]\}$ be the training set of video-sentence pairs for seen queries where $\mathcal{V}^{tr}$ is the set of all training videos with maximum duration T, $\mathcal{Q}^{tr}$ is the set of seen queries, (τ_s, τ_e) are the ground truth temporal endpoints for a query. For a given test-set $\mathcal{S}^{te} = \{(v, q)|v \in \mathcal{V}^{te}, q \in \mathcal{Q}^{te}\}$ with video-sentence pairs, our task is to predict the set of temporal endpoints $\{(\tau_s, \tau_e)\}$. We consider that $\mathcal{Q}^{tr} \cap \mathcal{Q}^{te} = \emptyset$, i.e., queries in test-set are not seen during training. As a result, $\mathcal{V}^{te}$ contains events that are not present in $\mathcal{V}^{tr}$. Additionally, we consider that S^{tr} is available during inference.

3.2 Localization Inference Schema

Existing temporal localization approaches [36, 71, 81] learn to encode fused moment-text representations. They either follow candidate moment sampling and encoding process to predict overlap scores (Figure 2 (a)) [71, 81] or summarize the whole video

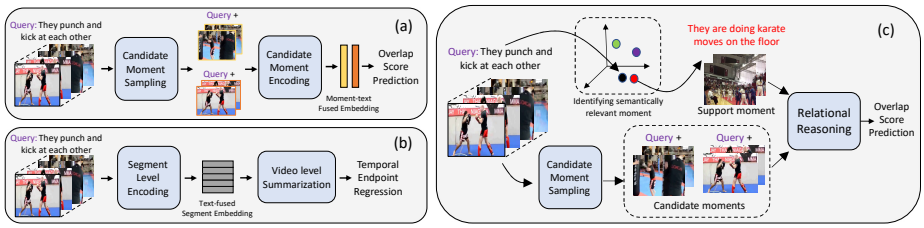

Fig. 2. A brief illustration of our novel text-based temporal localization approach. While existing works learn to encode video segments to identify the correct moment ((a) and (b)), we consider relational reasoning between two semantically relevant moment for localization purpose (c).

based on query encoding and segment level encoding of video to regress temporal endpoints (Figure 2 (b)) [36]. In both cases, moment representations are directly optimized for available seen events. As a result, the models get tuned to the available events in the training set and do not necessarily learn to generalize for unseen events. Since, our objective is to localize events which are not available during training, we deviate from the conventional approaches and propose a novel approach on how to address the text-based temporal localization task. For our proposed TLRR, we hypothesize that the correct moment corresponding to the unseen text query and the moments corresponding to the semantically relevant seen queries will contain shared concepts or similarities. Therefore, to identify the correct moment in a video based on an unseen text query, instead of directly predicting based on the moment-text representation, we utilize semantically relevant seen events. In that regard, we formulate the localization inference as a relational reasoning problem between two semantically relevant moments.

For a given video and an unseen text query, semantically relevant moments can be identified based on the semantics of the text query. Recent advances in Natural Language Processing (NLP) unfold many sentence encoder models which are trained on large corpus of text data in self-supervised or unsupervised manner. These models are able to capture wide range of sentence semantics and can be transferred to other NLP tasks. Our idea is to use these sentence encoders to find semantically relevant moments. In our work, we utilize universal sentence encoder [4], which is also able to capture sentence semantics, to find semantically relevant moments. Figure 2 (c) clearly illustrates our localization inference scheme. Given the unseen query, instead of directly inferring overlap scores from moment-text fused representation, we first identify semantically relevant query and its corresponding moment using universal sentence encoder. We utilize this semantically relevant moment as the support moment and consider relational reasoning between the support moment and the candidate moments to identify the correct moment. Our motivation behind this approach is that this relational inference system can be learned using available training data and the learned relational model is transferable to unseen cases [54]. Our framework consists of candidate moment encoder, fusion network, support moment encoder and relational reasoning module. In the following sections, we discuss the framework and how we utilize available training data to learn a proper relational inference system.

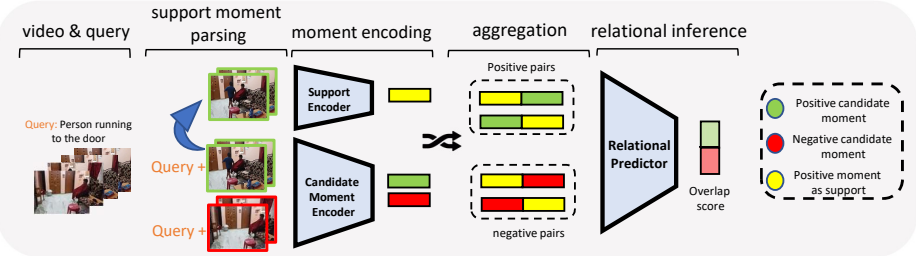

Fig. 3. Overview of the framework and the training of the relational reasoning based temporal localization approach. Candidate moment and support moment representations are aggregated to form positive pairs (positive candidate, positive support) and negative pairs (negative candidate, positive support)/(positive candidate, negative support). The relational module is trained to estimate the relational scores based on the pairs.

3.3 Framework

As illustrated in Figure 3, our framework consists of a candidate moment encoder that generates a text-fused representation of candidate moments, a support moment encoder that encodes the support moment, and a relational prediction module to infer based on the relational reasoning between candidate moment and support moment. To learn the relational reasoning system utilizing available training samples, we mimic the relational inference task during training. At train-time, for seen queries in training set, we infer the overlap scores based on the relation between candidate moment and support moment, where the ground truth moment is used as the positive support moment. All the modules and the learning procedure are described in the following sections.

Visual Feature Extraction. We perform fixed interval sampling over the frames of the videos and sample l non-overlapping clips per video. For each clip, we extract 2D/3D convolutional feature, resulting in a set of l clip features $\{\boldsymbol{c}_i\}_{i=1}^{l}$. Here, $\boldsymbol{c}_i$ is the feature representation of the i^{th} clip.

Text Feature Extraction. We use GloVe word embedding [45] and Bi-directional LSTM network [19] for representing text queries. For each word s of the query sentence q, we use Glove word embeddings to obtain its initial embedding vectors, which are fed sequentially into a three-layer bidirectional LSTM network. The last hidden state $\hat{\boldsymbol{q}}$ is used as the feature representation of the input sentence.

Candidate Moment Encoding and Modality Fusion. Clip representations $\{\boldsymbol{c}_i\}_{i=1}^{l}$, sampled from each video is used to construct candidate moment representations. For each candidate moment, we max-pool the corresponding clip features across the specific time span. For example, moment corresponding to i^{th} to $(i+n)^{th}$ clips will be represented by $\boldsymbol{f}_{i:i+n} = MaxPool(\boldsymbol{c}_i, \ldots, \boldsymbol{c}_{i+n})$, where $\boldsymbol{f} \in \mathcal{R}^{d_f}$ (d_f is the feature dimension). Moment encodings and text encodings are projected in the same subspace and their dot product is taken as the fused moment-text representation by $\boldsymbol{e} = (\boldsymbol{W}^q\hat{\boldsymbol{q}}).(\boldsymbol{W}^f\boldsymbol{f})$. Here, $\boldsymbol{W}^q$ and $\boldsymbol{W}^f$ are the learnable parameters. We stack all moment-text representations of a video as a 2D feature map, similar to [81], and use L convolutional layers to further encode the representations. As a result, we obtain a set

of candidate moment representations $\{\boldsymbol{m}_i\}_{i=0}^{N}$, where N is the total number of candidate moments from a video and $\boldsymbol{m}_i \in R^{d_m}$, where d_m is the feature dimension of the candidate moment representations.

Support Moment Encoder. We use a feed-forward network as the support moment encoder. For a support moment consisting of n consecutive clips $\{\boldsymbol{c}_i\}_{i=1}^{n}$, where $\boldsymbol{c}_i \in \mathcal{R}^{d_m}$, we first average pool the n clip representations to a single representation $\boldsymbol{s'} \in \mathcal{R}^{d_m}$. If we have multiple support moments, then we average pool all the support moment representations into a single representation. Then we use a feed-forward network to obtain the final support representation $\boldsymbol{s}$ by

$$\boldsymbol{s} = ReLU(\boldsymbol{W}^s \boldsymbol{s'} + \boldsymbol{b}^s). \tag{1}$$

Here, $\boldsymbol{W}^s$ and $\boldsymbol{b}^s$ are the learnable parameters and $\boldsymbol{s} \in \mathcal{R}^{d_m}$. We keep the feature dimension of support moment same as the candidate moment feature dimension d_m. The input to the support moment encoder varies in the training stage and inference stage. In the training stage, the correct candidate moment is used as the support moment. In the inference/testing stage, based on the unseen test query, most semantically relevant moments from the training set are used as the support moments. These moments work as the helper to find the correct moment from the video.

3.4 Relational Prediction

The relational module is a function $\mathcal{Z}_\theta(\cdot)$ parameterized by learnable weights θ and modeled by a feed forward neural network. Input to the relational module is a pair of two representations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where one element represents the selected support moment $\boldsymbol{s}$ and the other element represents a candidate moment $\boldsymbol{m}_i$ from the set of candidate moment representations $\{\boldsymbol{m}_i\}_{i=1}^{N}$. We use concatenation as the aggregation function to get aggregated representation of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as $a_{cat}(\boldsymbol{x}_i, \boldsymbol{x}_j)$. For a pair of support moment representation $\boldsymbol{s}$ and i^{th} candidate moment representation $\boldsymbol{m}_i$, the relational module outputs a overlap score ϕ_i by

$$\phi_i = \mathcal{Z}_\theta(a_{cat}(\boldsymbol{s}, \boldsymbol{m}_i)). \tag{2}$$

To confirm that the relational reasoning module $\mathcal{Z}_\theta$ predicts based on the relation between pair of representations and not based on a single representation, $\mathcal{Z}_\theta$ requires to maintain the commutative property, i.e., $\mathcal{Z}_\theta(a_{cat}(\boldsymbol{s}, \boldsymbol{m}_i))$ = $\mathcal{Z}_\theta(a_{cat}(\boldsymbol{m}_i, \boldsymbol{s}))$. However, the concatenation operation $a_{cat}(\cdot, \cdot)$ is not commutative. Therefore, to enforce the commutative property of the relational module, we compute the overlap score for the pair of elements $\boldsymbol{s}$ and $\boldsymbol{m}_i$ by

$$\phi_i = \frac{1}{2}\big[\mathcal{Z}_\theta(a_{cat}(\boldsymbol{m}_i, \boldsymbol{s})) + \mathcal{Z}_\theta(a_{cat}(\boldsymbol{s}, \boldsymbol{m}_i))\big]. \tag{3}$$

3.5 Learning Relational Inference

In our learning setup, a training sample consists of a video v, a text query q, and temporal ground truth information for the query (τ_s, τ_e). Instead of learning to directly predict

the overlap score for each candidate moment, we learn to infer the overlap scores based on the relation with most relevant support moments. To train this relational inference system, we sample two types of support moment: i) positive support moment and ii) negative support moment. For each query in a video, we extract the ground truth segment of the video and use it as the positive support moment $\boldsymbol{s}^+$. Again, for each query in a video, we select semantically unrelated query in the trainset and use its corresponding moment as the negative support moment $\boldsymbol{s}^-$. Our objective is to distinguish intra-video candidate moments based on the support moment. To do so, we compute overlap prediction loss $\mathcal{L}^{intra}$ for a set of pairs $\mathcal{X}^1$=$\{(\boldsymbol{m}_i, \boldsymbol{s}^+)\}$, which consists of pairs of all candidate moments and positive support moment in a video. To guide the learning of distinguishing intra-video candidate moments through relational inference system, we use scaled $tIoU$ (temporal Intersection-over-Union) value with ground-truth segment as the supervision signal. We compute the scaled $tIoU$ by

$$y_i = \begin{cases} 0 & g_i \leq t_{min}, \\ \frac{g_i - t_{min}}{t_{max} - t_{min}} & t_{min} < g_i < t_{max}, \\ 1 & g_i > t_{max}. \end{cases} \tag{4}$$

Here, g_i is the ground truth $tIoU$ for the i^{th} candidate moment and t_{min}, t_{max} are two thresholds to compute y_i. For a video with N candidate moments, $\mathcal{L}^{intra}$ is realized by binary cross entropy loss as

$$\mathcal{L}^{intra} = -\frac{1}{N} \sum_{\mathcal{X}^1} \big[y_i \log(\phi_i) + (1 - y_i) \log(1 - \phi_i) \big]. \tag{5}$$

Here, ϕ_i is the overlap score computed using Eqn. 3. To ensure that the model predicts the overlap score based on the relationship between the candidate moment and the support moment, we use the sampled negative support moments s^- to train the model. In each video, candidate moments with $tIoU > t_{min}$ are considered as positive candidate moment m^+. For each video with P positive candidate moments, we formulate a set of pairs $\mathcal{X}^2$=$\{(\boldsymbol{m}_i^+, \boldsymbol{s}^-)\}$ and compute negative relational loss $\mathcal{L}^{neg}$ by

$$\mathcal{L}^{neg} = -\frac{1}{P} \sum_{\mathcal{X}^2} \log(1 - \phi_i). \tag{6}$$

The two losses are jointly considered for training our relational inference model, with λ balancing contributions as in

$$\mathcal{L}^{total} = \mathcal{L}^{intra} + \lambda \mathcal{L}^{neg}. \tag{7}$$

We compute $\mathcal{L}_{total}$ for all seen video-text query pairs in the training set and optimize the relational inference model by minimizing the total loss.

3.6 Inference for Unseen Queries

During inference, given a video and an unseen text query, we are required to localize the correct moment. We use the universal sentence encoder [4] to find semantically

relevant queries from the training set. Then the corresponding moment to the relevant query is used as a support moment. Based on the video, support moments, and the unseen query, the learned relational model predicts overlap score ϕ for different temporal granularities in one forward pass. All the predicted segments are ranked and refined with non-maximum suppression (NMS) according to the predicted ϕ. Afterwards, the final temporal grounding result is obtained.

4 Experiments

4.1 Reorganized Datasets

Existing benchmark temporal moment localization dataset splits are not designed for the task of temporal localization of novel events based on unseen text queries. Instead, training set (trainset for short) and testing set (testset for short) data are sampled from the same distribution, and text queries in the testset overlap with text queries in the trainset. We reorganize two of the benchmark datasets namely Charades-STA [12] and ActivityNet Captions [24] to create splits according to our problem setting. For both datasets, we create splits based on the verbs and nouns present in the text queries. First, we combine all the annotations of the trainset and testset videos of the dataset. To create the splits, we consider a set of n_V verbs and n_N nouns present in the combined annotation. We consider it the set of seen verbs and seen nouns. Then, we identify videos that contain at least a single query that has a verb or noun not present in the mentioned set. In the selected videos, queries which do not have verbs or nouns from the mentioned set are collected as unseen testset split and, queries which have verbs or nouns from the mentioned set are collected as seen testset split. The training set is created from the rest of the videos, with queries that contain either verb or noun present in the mentioned set. We exclude queries which contains verb or noun from both seen set and unseen set. We use spaCy [20] to parse verbs and nouns from text queries. These reorganized datasets reflect a realistic setting as datasets are usually composed of recurring events of limited concepts. However, a localization system may encounter varied types of events in real-world applications. Details of the nouns and verbs selected to create the split are provided in the supplementary material. Excluding queries which contains verb or noun from both seen set and unseen set results in reduced number of moment-sentence pairs in the reorganized dataset. However, the size of the dataset doesn't have impact on the significance of our proposed problem setup, which is experimentally evaluated in the supplementary material.

Charades-STA Unseen. Charades-STA dataset contains a total of $6{,}670$ videos where $5{,}336$ and $1{,}334$ are the number of training and testing videos. Textual annotations in Charades-STA has direct temporal correspondence with activity annotation of the Charades dataset [50]. We combine training and testing set annotations and consider $n_V = 20$ and $n_N = 40$ (excluding 'person' noun) for creating Charades-STA Unseen dataset. In this way, we have Charades-STA Unseen dataset with 5525, 1665, and 867 training, unseen testing, and seen testing moment-sentence pairs respectively.

ActivityNet Captions Unseen. ActivityNet Captions [24] dataset is proposed for dense video captioning task. Each video contains at least two ground truth segments and each segment is paired with one ground truth caption [66]. This dataset contains around 20k

Table 1. This table reports ***unseen*** text query based temporal moment localization performance of TLRR, compared against several approaches, on Charades-STA Unseen dataset.

Method	R@1, IoU@0.5	R@1, IoU@0.7	R@5, IoU@0.5	R@5, IoU@0.7	mIoU
DeViSE [11]	29.98	11.29	71.42	39.81	-
ESZSL [48]	23.90	10.13	60.50	34.53	-
SCDM [71]	28.22	11.89	54.25	32.95	28.63
LGI [36]	29.01	12.85	-	-	29.62
2D-TAN [81]	31.05	13.33	70.75	36.94	29.88
TLRR	**33.15**	**16.22**	**77.66**	**42.40**	**31.29**

videos which are split into training, validation, and testing set with 50%, 25%, and 25% ratio respectively. Textual description for only the training and validation set is given. We combine training and validation set and consider $n_V = 70$ and $n_N = 250$ for creating ActivityNet Captions Unseen dataset. In this way, we have ActivityNet Captions Unseen dataset with 5669, 2553, and 710 training, unseen testing, and seen testing moment-sentence pairs respectively.

4.2 Evaluation Metric

We use "$R@k, IoU@m$", which reports the percentage of at least one of the top-k results having Intersection-over-Union (IoU) larger than m [12]. For a text query, "$R@k, IoU@m$" reflects if one of the top-k retrieved moments has IoU with the ground truth moment larger than the specified threshold m. So, "$R@k, IoU@m$" is either 1 or 0 for each text query. We compute it for all the text queries in the testing sets and report the average results for $k \in \{1, 5\}$ and $m \in \{0.50, 0.70\}$. We also compute mIoU where mIoU is the average IoU over all testing samples.

4.3 Implementation Details

We use VGG feature [51] for Charades-STA Unseen dataset. For ActivityNet Captions Unseen dataset, we use extracted C3D features [58]. The number of frames in a clip is set to 4 for Charades-STA Unseen, and 16 for ActivityNet Captions Unseen and we use non-overlapping clips for both datasets. The number of sampled clips N is set to 16 for Charades-STA Unseen, 64 for ActivityNet Captions Unseen. For the candidate moment encoder, we adopt a 4-layer convolution network with a kernel size of 5 for Charades-STA Unseen and a 4-layer convolution network with a kernel size of 9 for ActivityNet Captions Unseen. For both datasets, the support moment encoder is a single-layer feed-forward network and the relational prediction network is a two-layer feed-forward network. The proposed network is implemented in TensorFlow and trained using a single RTX 2080 GPU. We use mini-batches containing 32 video-sentence pairs and use Adam [23] optimizer with a learning rate of 0.0001. The dimension of both candidate moment representation d_m and support moment representation d_s is set to 512 for both datasets. We set λ=3 empirically in Eqn 7 for both datasets. The scaling thresholds t_{min} and

Table 2. This table reports ***unseen*** text query based temporal moment localization performance of TLRR, compared against several approaches, on ActivityNet Captions Unseen dataset.

Method	R@1, IoU@0.5	R@1, IoU@0.7	R@5, IoU@0.5	R@5 IoU@0.7	mIoU
DeViSE [11]	5.07	2.00	10.46	4.05	-
ESZSL [48]	4.72	1.85	11.83	4.48	-
SCDM [71]	19.22	8.22	46.38	23.58	23.97
2D-TAN [81]	19.15	10.26	38.78	24.01	21.70
VSLNet [77]	19.23	9.99	-	-	25.32
TLRR	**23.19**	**13.24**	**53.31**	**36.66**	**26.35**

t_{max} of Eqn. 4 are set to 0.5 and 1.0 respectively for both datasets. Non-maximum suppression (NMS) with a threshold of 0.5 is applied during the inference. We train TLRR for 50 epochs. We select the checkpoint which has the best average performance across metrics for seen queries.

4.4 Result Analysis

Temporal Localization Performance of Novel/Unseen Events. Since ours is the first work on temporal localization of novel events, there are no existing approaches to directly compare with. As our problem setup is closely related to zero-shot settings, we adapt two zero-shot learning approaches namely **DeViSE** [11] and **ESZSL** [48] for this problem setup. We also compare with some of the state-of-the-art temporal localization approaches with publicly available codes, e.g., **2D-TAN** [81], **SCDM** [71], **LGI** [36], and **VSLNet** [77], by training those models using our reorganized training splits.

Table 1 and Table 2 illustrate the TLRRs' performance for temporal localization of novel event based on unseen text query and compare it with other approaches for Charades-STA Unseen and ActivityNet Captions Unseen dataset respectively. For the Charades-STA Unseen dataset, the performance of different baseline approaches are comparable among them. However, TLRR provides $2\% - 7\%$ absolute improvement over the best scores of compared approaches over all the reported metrics. In Table 2, baseline zero-shot approaches (DeViSE, ESZSL) are performing poorly for ActivityNet Captions Unseen dataset. This is because the text queries are complex compared to Charades-STA Unseen and it requires fine-grained analysis of longer videos in ActivityNet Caption Unseen. We observe $3\% - 15\%$ absolute improvement over best scores of compared approaches in the ActivityNet Captions Unseen dataset.

Relational Reasoning Performance Analysis. Since TLRR's performance is dependent on its ability to reason on the relationship of two different moments, in Table 3, we analyze the competence of our relational reasoning module $\mathcal{Z}_\theta$ for Charades-STA Unseen dataset. We consider three scenarios: i) **Irrelevant**: based on the unseen text query, retrieve the seen query from the semantic embedding space that are furthest away or most irrelevant and use the corresponding moment as the support information, ii) **Random**: retrieve random seen query from the training set and use the corresponding moment as the support information, and iii) **Relevant**: retrieve the nearest/most relevant

Table 3. This table reports *unseen* text query based novel event localization performance using different types of support moments to analyze TLRR for Charades-STA Unseen dataset.

Support Moment	R@1, IoU@0.5	R@1, IoU@0.7	R@5, IoU@0.5	R@5, IoU@0.7	mIoU
Irrelevant	20.30	11.05	62.58	33.93	22.48
Random	28.71	14.47	73.57	40.24	28.40
Relevant	**33.15**	**16.22**	**77.66**	**42.40**	**31.29**

seen query from the semantic embedding space and use the corresponding moment as the support information (i.e., our proposed TLRR). We observe that when irrelevant queries are retrieved and their corresponding moment is used as the support, the performance goes down. Since the moment corresponding to a irrelevant query does not contain shared concept/ similarities with the correct moment, the relational module expectedly fails to identify the correct moment. When random seen queries are selected, the performance is better compared to the irrelevant case. We obtain the best performance when the closest seen query is selected from the semantic embedding space.

Temporal Localization Performance of Seen Events. We further report the performance of different approaches when evaluated on the testing split of seen queries in both the datasets on Table 4 and Table 5. Although the main focus of this paper is temporal localization of unseen events, this experiment is presented to evaluate how the performance of different methods changes for seen events compared to localization of unseen events (Table 1 and Table 2). We expect any method to work slightly better on localizing the seen events compared to the unseen ones; however, a drastic/large change would indicate poor generalization ability of the model.

For the compared methods and baselines, we observe that there is a significant difference in performance when the same model is evaluated in the testing split of seen queries and testing split of unseen queries for both datasets comparing Table 1 and Table 2 with Table 4 and Table 5 respectively. Not surprisingly, both the conventional temporal localization approaches (i.e., SCDM and 2D-TAN) show a drastic change in performance across metrics in both datasets. The average difference in performance is reported by Δ_{avg} in Table 4 and Table 5. SCDM shows 19.80% average difference in Charades-STA and 13.24% average difference across metrics in ActivityNet in localization performance of seen queries compared to localization performance of unseen queries. Similarly, 2D-TAN shows average difference (across metrics) of 5.89% in Charades-STA and 16.18% in ActivityNet in localizing seen queries compared to unseen. Though the zero-shot based approaches (DeViSE and ESZSL) show small gap in performance between seen and unseen events, which is expected due to the approaches generalization ability, they are unable to maintain a proper level of localization performance compared to other methods. However, the proposed TLRR approach shows a significantly lower change in performance, e.g., 3.37% average in Charades-STA and 7.13% average in ActivityNet Captions.

This indicates the significance of the problem setup and generalization ability of TLRR. Unlike the conventional temporal localization approaches, TLRR is not designed to specifically focus on the seen events. In Table 4 and Table 5, we observe

Table 4. This table reports ***seen*** text query based temporal moment localization performance of TLRR on Charades-STA Unseen dataset. Here, Δ_{avg} refers to average performance difference for seen events and unseen events (Table 1) for a specific method. From the lower value of Δ_{avg}, it is evident that TLRR generalizes significantly better than other temporal localization approaches.

Method	**R@1, IoU@0.5**	**R@1, IoU@0.7**	**R@5, IoU@0.5**	**R@5, IoU@0.7**	$\boldsymbol{\Delta_{avg}} \downarrow$
DeViSE [11]	36.34	15.86	77.66	44.10	5.36
ESZSL [48]	37.50	18.40	72.34	42.13	10.34
SCDM [71]	50.46	28.00	73.49	54.86	19.80
2D-TAN [81]	37.95	18.45	76.70	42.56	5.89
TLRR	34.83	20.76	78.78	48.56	**3.37**

Table 5. This table reports ***seen*** text query based temporal moment localization performance of TLRR on ActivityNet Captions Unseen dataset. Δ_{avg} refers to average performance difference for seen events and unseen events (Table 2) for a specific method. From the lower value of Δ_{avg}, it is evident that TLRR generalizes significantly better than other temporal localization approaches.

Method	**R@1, IoU@0.5**	**R@1, IoU@0.7**	**R@5, IoU@0.5**	**R@5, IoU@0.7**	$\boldsymbol{\Delta_{avg}} \downarrow$
DeViSE [11]	12.07	5.40	18.18	8.52	**5.64**
ESZSL [48]	12.64	5.40	19.74	8.66	5.89
SCDM [71]	34.66	20.74	59.51	35.37	13.24
2D-TAN [81]	34.65	22.39	57.18	42.68	16.18
TLRR	27.46	17.61	60.42	49.44	7.13

that model optimized to do localization inference directly based on the candidate moment representation overall performs better compared to TLRR for types of events that are already seen in training. However, direct localization limits these models' capacity to a small set of events which is evident by the significant gap between performances for seen and unseen events. Instead, our proposed TLRR approach is able to retain a competitive performance for the seen queries and boost the performance for unseen queries resulting in reducing the performance gap between seen and unseen events. Also, our proposed TLRR is able to show comparable performance (please refer to supplementary material) on the original temporal localization dataset, even though TLRR is not optimized for seen events and have a relatively simple base architecture.

Effect of $\mathcal{L}^{neg}$ in learning TLRR. TLRR uses $\mathcal{L}^{intra}$ and $\mathcal{L}^{neg}$ to learn relational localization system. Effectiveness of these two loss components for distinguishing intra-video moments by relational prediction is evident from Table 1, Table 2, and Table 3. We consider two setups, i) TLRR trained with $\mathcal{L}^{intra}$ and ii) TLRR trained with $\mathcal{L}^{intra}$ + $\lambda\mathcal{L}^{neg}$. We observe that when only $\mathcal{L}^{intra}$ is used to train TLRR, there is almost no difference in performance (difference within 1%) for using relevant or irrelevant moments as input to the support encoder. However, there is $5\% - 15\%$ difference in Charades-STA Unseen dataset for using relevant or irrelevant moments as input to the support encoder when $\mathcal{L}^{intra}$ + $\lambda\mathcal{L}^{neg}$ is used to train TLRR. So, $\mathcal{L}^{neg}$ enforces the model to predict based on the relation.

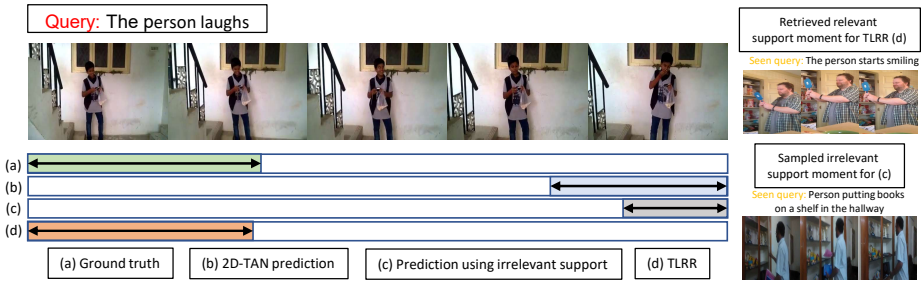

Fig. 4. Given the query 'The person laughs' and the corresponding video, this figure shows: (a) ground truth segment of the video which corresponds to the text query, (b) predicted moment by 2D-TAN, (c) predicted moment when irrelevant moment is used as support, and (d) predicted moment using retrieved relevant support moment (TLRR). While (b) and (c) result in failure, TLRR is able to detect the correct moment using relational reasoning.

Qualitative Result. In Figure 4, we illustrate an example case of our system's success. Given the query 'The person laughs' and the corresponding video, Figure 4 shows: (a) ground truth segment of the video which corresponds to the text query, (b) predicted moment by 2D-TAN, (c) predicted moment when irrelevant moment is used as support, and (d) predicted moment using retrieved relevant support moment. Person laughing is a difficult event to detect as it encompasses a small region of the frame and results in small temporal variation in the feature. Without any notion/previous knowledge of how the activity/event is, it becomes even harder, which is reflected by the failure case of (b) and (c). However, TLRR is able to detect the correct moment using relational reasoning.

5 Conclusion

In this paper, we address the novel problem of temporal localization of unseen/novel events based on unseen text queries. The problem of identifying novel events in video is important and practical because not every kind of event can be expected to be within the training set. This allows for generalization of temporal localization methods to novel scenarios. We propose a relational reasoning based framework hypothesizing a conceptual relation between moments corresponding to semantically relevant queries. Extensive experiments on reorganized Charades-STA and ActivityNet Captions datasets demonstrate the effectiveness of the proposed framework compared to several baselines in localizing video moments from text queries. Our code and dataset splits will be publicly available. Though support moment based relational prediction can reduce the performance gap between seen and unseen events, it is burdened with the extra computation of relevant moments, which is computationally expensive. Future work can focus on this issue.

Acknowledgments. This work was partially supported by ONR grant N00014-19-1-2264 and NSF grant 1901379.

Bibliography

[1] Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. IEEE transactions on pattern analysis and machine intelligence **38**(7), 1425–1438 (2015)

[2] Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2927–2936 (2015)

[3] Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE international conference on computer vision. pp. 5803–5812 (2017)

[4] Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., et al.: Universal sentence encoder. arXiv preprint arXiv:1803.11175 (2018)

[5] Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.S.: Temporally grounding natural sentence in video. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 162–171 (2018)

[6] Chen, S., Jiang, W., Liu, W., Jiang, Y.G.: Learning modality interaction for temporal sentence localization and event captioning in videos. In: European Conference on Computer Vision. pp. 333–351. Springer (2020)

[7] Chen, S., Jiang, Y.G.: Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8425–8435 (2021)

[8] Chi, J., Peng, Y.: Dual adversarial networks for zero-shot cross-media retrieval. In: IJCAI. pp. 663–669 (2018)

[9] Ding, X., Wang, N., Zhang, S., Cheng, D., Li, X., Huang, Z., Tang, M., Gao, X.: Support-set based cross-supervision for video grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11573–11582 (2021)

[10] Dong, J., Li, X., Xu, C., Ji, S., He, Y., Yang, G., Wang, X.: Dual encoding for zero-example video retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9346–9355 (2019)

[11] Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al.: Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems (NIPS). pp. 2121–2129 (2013)

[12] Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)

[13] Gao, J., Xu, C.: Fast video moment retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1523–1532 (2021)

[14] Ge, R., Gao, J., Chen, K., Nevatia, R.: Mac: Mining activity concepts for language-based temporal localization. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 245–253. IEEE (2019)

[15] Ghosh, S., Agarwal, A., Parekh, Z., Hauptmann, A.: Excl: Extractive clip localization using natural language descriptions. arXiv preprint arXiv:1904.02755 (2019)
[16] Hahn, M., Kadav, A., Rehg, J.M., Graf, H.P.: Tripping through time: Efficient localization of activities in videos. arXiv preprint arXiv:1904.09936 (2019)
[17] He, D., Zhao, X., Huang, J., Li, F., Liu, X., Wen, S.: Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8393–8400 (2019)
[18] Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with temporal language. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1380–1390 (2018)
[19] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
[20] Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear
[21] Huang, J., Liu, Y., Gong, S., Jin, H.: Cross-sentence temporal and semantic relations in video activity localisation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7199–7208 (2021)
[22] Jiang, B., Huang, X., Yang, C., Yuan, J.: Cross-modal video moment retrieval with spatial and language-temporal attention. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. pp. 217–225 (2019)
[23] Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
[24] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: International Conference on Computer Vision (ICCV) (2017)
[25] Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE transactions on pattern analysis and machine intelligence **36**(3), 453–465 (2013)
[26] Li, J., Jing, M., Lu, K., Ding, Z., Zhu, L., Huang, Z.: Leveraging the invariant side of generative zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7402–7411 (2019)
[27] Lin, K., Xu, X., Gao, L., Wang, Z., Shen, H.T.: Learning cross-aligned latent embeddings for zero-shot cross-modal retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11515–11522 (2020)
[28] Lin, Z., Zhao, Z., Zhang, Z., Wang, Q., Liu, H.: Weakly-supervised video moment retrieval via semantic completion network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11539–11546 (2020)
[29] Lin, Z., Zhao, Z., Zhang, Z., Zhang, Z., Cai, D.: Moment retrieval via cross-modal interaction networks with query reconstruction. IEEE Transactions on Image Processing **29**, 3750–3762 (2020)
[30] Liu, B., Yeung, S., Chou, E., Huang, D.A., Fei-Fei, L., Carlos Niebles, J.: Temporal modular networks for retrieving complex compositional activities in videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 552–568 (2018)

[31] Liu, D., Qu, X., Dong, J., Zhou, P., Cheng, Y., Wei, W., Xu, Z., Xie, Y.: Context-aware biaffine localizing network for temporal sentence grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11235–11244 (2021)

[32] Liu, D., Qu, X., Liu, X.Y., Dong, J., Zhou, P., Xu, Z.: Jointly cross-and self-modal graph attention network for query-based moment localization. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4070–4078 (2020)

[33] Liu, M., Wang, X., Nie, L., He, X., Chen, B., Chua, T.S.: Attentive moment retrieval in videos. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 15–24 (2018)

[34] Liu, M., Wang, X., Nie, L., Tian, Q., Chen, B., Chua, T.S.: Cross-modal moment localization in videos. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 843–851 (2018)

[35] Mithun, N.C., Paul, S., Roy-Chowdhury, A.K.: Weakly supervised video moment retrieval from text queries. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

[36] Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10810–10819 (2020)

[37] Nam, J., Ahn, D., Kang, D., Ha, S.J., Choi, J.: Zero-shot natural language video localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1470–1479 (October 2021)

[38] Nan, G., Qiao, R., Xiao, Y., Liu, J., Leng, S., Zhang, H., Lu, W.: Interventional video grounding with dual contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2765–2775 (2021)

[39] Niu, L., Cai, J., Veeraraghavan, A., Zhang, L.: Zero-shot learning via category-specific visual-semantic mapping and label refinement. IEEE Transactions on Image Processing **28**(2), 965–979 (2018)

[40] Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G.S., Dean, J.: Zero-shot learning by convex combination of semantic embeddings. arXiv preprint arXiv:1312.5650 (2013)

[41] Parikh, D., Grauman, K.: Relative attributes. In: 2011 International Conference on Computer Vision. pp. 503–510. IEEE (2011)

[42] Patacchiola, M., Storkey, A.: Self-supervised relational reasoning for representation learning. arXiv preprint arXiv:2006.05849 (2020)

[43] Paul, S., Mithun, N.C., Roy-Chowdhury, A.K.: Text-based localization of moments in a video corpus. IEEE Transactions on Image Processing **30**, 8886–8899 (2021)

[44] Paul, S., Torres, C., Chandrasekaran, S., Roy-Chowdhury, A.K.: Complex pairwise activity analysis via instance level evolution reasoning. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2378–2382. IEEE (2020)

[45] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)

[46] Raposo, D., Santoro, A., Barrett, D., Pascanu, R., Lillicrap, T., Battaglia, P.: Discovering objects and their relations from entangled scene representations. arXiv preprint arXiv:1702.05068 (2017)
[47] Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics **1**, 25–36 (2013)
[48] Romera-Paredes, B., Torr, P.: An embarrassingly simple approach to zero-shot learning. In: International conference on machine learning. pp. 2152–2161. PMLR (2015)
[49] Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. arXiv preprint arXiv:1706.01427 (2017)
[50] Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision. pp. 510–526. Springer (2016)
[51] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
[52] Socher, R., Ganjoo, M., Sridhar, H., Bastani, O., Manning, C.D., Ng, A.Y.: Zero-shot learning through cross-modal transfer. arXiv preprint arXiv:1301.3666 (2013)
[53] Soldan, M., Xu, M., Qu, S., Tegner, J., Ghanem, B.: Vlg-net: Video-language graph matching network for video grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3224–3234 (2021)
[54] Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1199–1208 (2018)
[55] Tan, R., Xu, H., Saenko, K., Plummer, B.A.: Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2083–2092 (2021)
[56] Tang, H., Zhu, J., Gao, Z., Zhuo, T., Cheng, Z.: Attention feature matching for weakly-supervised video relocalization. In: Proceedings of the 2nd ACM International Conference on Multimedia in Asia. pp. 1–7 (2021)
[57] Tang, H., Zhu, J., Wang, L., Zheng, Q., Zhang, T.: Multi-level query interaction for temporal language grounding. IEEE Transactions on Intelligent Transportation Systems (2021)
[58] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: International Conference on Computer Vision (ICCV). pp. 4489–4497. IEEE (2015)
[59] Wang, H., Zha, Z.J., Chen, X., Xiong, Z., Luo, J.: Dual path interaction network for video moment localization. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4116–4124 (2020)
[60] Wang, H., Zha, Z.J., Li, L., Liu, D., Luo, J.: Structured multi-level interaction network for video moment localization via language query. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7026–7035 (2021)

[61] Wang, Y., Deng, J., Zhou, W., Li, H.: Weakly supervised temporal adjacent network for language grounding. IEEE Transactions on Multimedia (2021)
[62] Wu, A., Han, Y.: Multi-modal circulant fusion for video-to-language and backward. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 1029–1035 (2018)
[63] Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 69–77 (2016)
[64] Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4582–4591 (2017)
[65] Xiao, S., Chen, L., Zhang, S., Ji, W., Shao, J., Ye, L., Xiao, J.: Boundary proposal network for two-stage natural language video localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2986–2994 (2021)
[66] Xu, H., He, K., Plummer, B.A., Sigal, L., Sclaroff, S., Saenko, K.: Multilevel language and vision integration for text-to-clip retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9062–9069 (2019)
[67] Xu, M., Pérez-Rúa, J.M., Escorcia, V., Martinez, B., Zhu, X., Zhang, L., Ghanem, B., Xiang, T.: Boundary-sensitive pre-training for temporal localization in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7220–7230 (2021)
[68] Xu, X., Song, J., Lu, H., Yang, Y., Shen, F., Huang, Z.: Modal-adversarial semantic learning network for extendable cross-modal retrieval. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. pp. 46–54 (2018)
[69] Xu, X., Hospedales, T., Gong, S.: Semantic embedding space for zero-shot action recognition. In: 2015 IEEE International Conference on Image Processing (ICIP). pp. 63–67. IEEE (2015)
[70] Yang, W., Zhang, T., Zhang, Y., Wu, F.: Local correspondence network for weakly supervised temporal sentence grounding. IEEE Transactions on Image Processing **30**, 3252–3262 (2021)
[71] Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In: Advances in Neural Information Processing Systems. pp. 534–544 (2019)
[72] Yuan, Y., Mei, T., Zhu, W.: 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, pp. 9159–9166 (2019)
[73] Zambaldi, V., Raposo, D., Santoro, A., Bapst, V., Li, Y., Babuschkin, I., Tuyls, K., Reichert, D., Lillicrap, T., Lockhart, E., et al.: Deep reinforcement learning with relational inductive biases. In: International Conference on Learning Representations (2018)
[74] Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., Gan, C.: Dense regression network for video grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10287–10296 (2020)
[75] Zhang, D., Dai, X., Wang, X., Wang, Y.F., Davis, L.S.: 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. pp. 1247–1257 (2019)

[76] Zhang, H., Sun, A., Jing, W., Zhen, L., Zhou, J.T., Goh, R.S.M.: Natural language video localization: A revisit in span-based question answering framework. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
[77] Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. arXiv preprint arXiv:2004.13931 (2020)
[78] Zhang, H., Long, Y., Guan, Y., Shao, L.: Triple verification network for generalized zero-shot learning. IEEE Transactions on Image Processing **28**(1), 506–517 (2018)
[79] Zhang, L., Chang, X., Liu, J., Luo, M., Wang, S., Ge, Z., Hauptmann, A.: Zstad: Zero-shot temporal activity detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
[80] Zhang, M., Yang, Y., Chen, X., Ji, Y., Xu, X., Li, J., Shen, H.T.: Multi-stage aggregated transformer network for temporal language localization in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12669–12678 (2021)
[81] Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. arXiv preprint arXiv:1912.03590 (2019)
[82] Zhang, S., Su, J., Luo, J.: Exploiting temporal relationships in video moment localization with natural language. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 1230–1238 (2019)
[83] Zhang, Z., Lin, Z., Zhao, Z., Xiao, Z.: 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. pp. 655–664 (2019)
[84] Zhao, Y., Zhao, Z., Zhang, Z., Lin, Z.: Cascaded prediction network via segment tree for temporal video grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4197–4206 (2021)
[85] Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: Proceedings of the European conference on computer vision (ECCV). pp. 803–818 (2018)
[86] Zhou, H., Zhang, C., Luo, Y., Chen, Y., Hu, C.: Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8445–8454 (2021)
[87] Zhu, Y., Long, Y., Guan, Y., Newsam, S., Shao, L.: Towards universal representation for unseen action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9436–9445 (2018)