Temporal Language Grounding with Referring Attention and Weak Supervision

Zhiyuan Fang[1], Shu Kong[2], Zhe Wang[1], Fengyu Yan[1], Charless Fowlkes[2], Yezhou Yang[1]
[1]{zy.fang, fyan14, yz.yang}@asu.edu Arizona State University, Tempe, USA
[2]{skong2, fowlkes}@ics.uci.edu University of California, Irvine, USA

Abstract

*Videos contain rich information for humans to interpret the world while training a machine-aided system for video analysis often requires large-scale videos with manual annotations, which are prohibitively expensive to collect and largely limited in annotation diversity due to combinatorial nature. To alleviate this dilemma, we propose a simple weakly-supervised temporal grounding method with a top-down signal (e.g., language description, short phrase, and action category label). This makes our system different from current weakly-supervised video learning methods which generate attentional weights to localize video segments in a bottom-up manner. We conduct thorough experiments to validate our method on several benchmarks. Specifically, by training our model in a weakly-supervised way with only video-level signals varying in the format from the simplistic action category labels to complex free-form language descriptions, we show our method achieves state-of-the-art performance on action localization on THUMOS14 and phrase/sentence localization on Charades(-STA) datasets, respectively. As our method localizes the free-form language sentences in a retrieval fashion (*i.e. *the top-down referring signal), we show it is capable of verifying their temporal order given the video as visual evidence. To this end, we introduce the new problem called Vision-Language Causal Verification (VALVE), and augment the Charades-STA dataset with necessary annotations to study this problem. With promising results, we note that this is essentially a step toward causality understanding through learning on the abundant videos along with the captions.*

1. Introduction

Videos, abundant in number and usually coming with other modalities (e.g., language captions and audios), contain much richer information for human to interpret the world and learn highly abstract knowledge. Building an intelligent system to automate video analysis is beneficial to human life, from self-driving cars to senior-assistive robots and forensic evidence analysis [35, 76, 5, 45, 9].

However, developing machine-aided system for video analysis should require large-scale videos with manual annotations especially on temporal boundaries for localization or grounding. Apparently, such annotations are not only prohibitively expensive to collect, but also largely limited in terms of diversity, especially by the combinatorial nature of language-based phrase and sentences. As a result, many successful video analysis methods have particularly focused on supervised learning over small/medium-scale videos with manual annotations [53, 70, 13, 63, 7, 52]. Recently, some weakly-supervised methods have been proposed for action localization on untrimmed videos by modeling it as a classification problem over a set of fixed, predefined action categories [68, 62, 25, 61, 43, 4, 47, 65, 56, 54, 71, 6, 73, 42, 44]. However, they cannot be easily deployed as real-world applications where much more diverse actions exist in videos.

Even nowadays, very few work goes beyond weakly-supervised action localization to more complex scenarios, e.g., localizing phrases or free-form language sentences in videos. Gao *et al.* augment the Charades dataset [59] by generating complex language queries with temporal boundary annotations for language moment retrieval with a fully supervised learning regression-based method [19]. Anne Hendricks *et al.* also collect a new dataset containing videos and language descriptions, and train a fully-supervise model to localize video segments [1]. As these methods rely on training videos with well-annotated language sentences, they do not generalize well due to the combinatorial nature of languages (e.g., similar keywords and sentence structures) [24, 19, 30].

To alleviate this dilemma, we propose a simple, weakly-supervised framework allowing for training to ground/retrieve along the arrow of time the user-provided (top-down) signal [40], like a simplistic action category, describable moment, a short textual phrase or even a free-form language-based sentence. With this framework, we train our model over videos only with their video-level annotations, which are the input as top-down attention during training. Therefore, we term our system *Weakly-Supervised video analysis with Referring Attention* (WSRA) to distinguish from other weakly-supervised methods for action localization that they produce bottom-up attentive weights

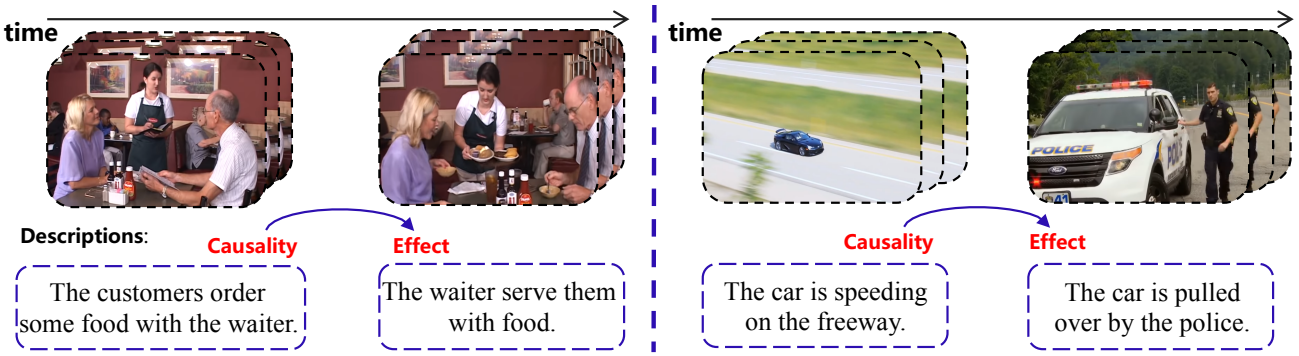

Figure 1: Demonstration of using sentence localization for vision-language causal verification. With our weakly supervised model with referring attention (WSRA), we can verify the temporal order of two sentences corresponding two temporal moments in a video to help verify their causal-effect relationship, though the model is not explicitly trained with any causal-effect signals.

to ground actions on untrimmed video with a classification loss. We validate our framework thorough experiments on three video analysis tasks in the weakly-supervised learning scenario, including action localization on THUMOS14 dataset[27], short phrase grounding on Charades dataset [58, 57] and free-form describable sentence grounding on the Charades-STA dataset [19], respectively.

Moreover, we note that the top-down grounding nature of our WSRA framework provides a unique opportunity to study the problem of Vision-Language Causal Verification (VALVE), which requires verifying causal-effect (temporal order in this paper) relationship between two phrases/sentences in front of the video as visual evidence, especially along the causal arrow of time [23]. We illustrate this in Fig. 1 through a scenario consisting two complex describable scenes, "*the customers order some food with the waiter*" and "*the waiter serve them with food*". Purely based on the textual description, a machine may find it difficult to parse the two sentences and connect them with a reasonable cause-effect relationship. Instead, our model, even though trained without any causal-effect signals, can draw a reasonable causal-effect relationship which tightly correlates with their temporal order by grounding the sentences temporally. To study this interesting VALVE problem, we augment the Charades-STA dataset with necessary annotations and conduct a series of experiments on it.

To summarize our contributions: (1) technically, we propose a simple, unified framework for weakly supervised method for video analysis; (2) conceptually, we show framework allows for learning over not only action category labels but also free-form language descriptions for video analysis, largely extending current weakly-supervised methods which mainly focus on action localization; (3) practically, we show how our method helps solve vision-language causal verification problem with a new dataset which will be released to public.

2. Related Work

Weakly-Supervised Action Localization gains attention in recent years as a specific form of learning a model over videos and their weak labels (i.e., action categories at video level) for video analysis [65, 55, 75, 44, 80, 47]; whereas the fully-supervised counterpart methods use full annotations at frame level for localizing the actions [6, 31, 71, 54, 66, 56], demanding expensive video dataset for training. Recent weakly-supervised methods largely model the problem as action recognition/classification [64, 12, 32, 82, 60, 60, 69, 8] and exploits the attentional mechanism that generates hidden weights for video segments to indicate positive and negative ones given an action label.

Temporal Grounding and Moments Retrieval with complex textual descriptions has also attracted a lot attention in recent years. This task is to ground a describable moment through a sentence or phrase along the arrow of time. Recent methods unanimously rely on well-annotated dataset for training with fully-supervised learning. For example, Gao *et al.* augment the Charades dataset [59] by generating complex language queries with temporal boundary annotations for language moment retrieval [19]; Anne Hendricks *et al.* also collect a new dataset containing videos and language descriptions to localize video segments [1]. Other followup methods [24, 40, 10, 67, 74, 21, 20] also depend on these or similar datasets and exploit the fully-supervised learning methods for training. This inevitably limits their generalizability due to the combinatorial nature of complex natural language sentences (e.g., synonymous words, grammatical tense and sentence structures) [24, 19].

Causal-Effect Detection on Videos is another separate line of work that investigates high-level causal induction from low-level visual signal [18, 17]. The methods exploit some heuristics in cognitive science and prove the concept and possibility to do this task. Different from these methods, we aim for learning such knowledge without prior heuristics from videos with weak labels (e.g., video-level descriptions). Specifically in this work, we study the problem of

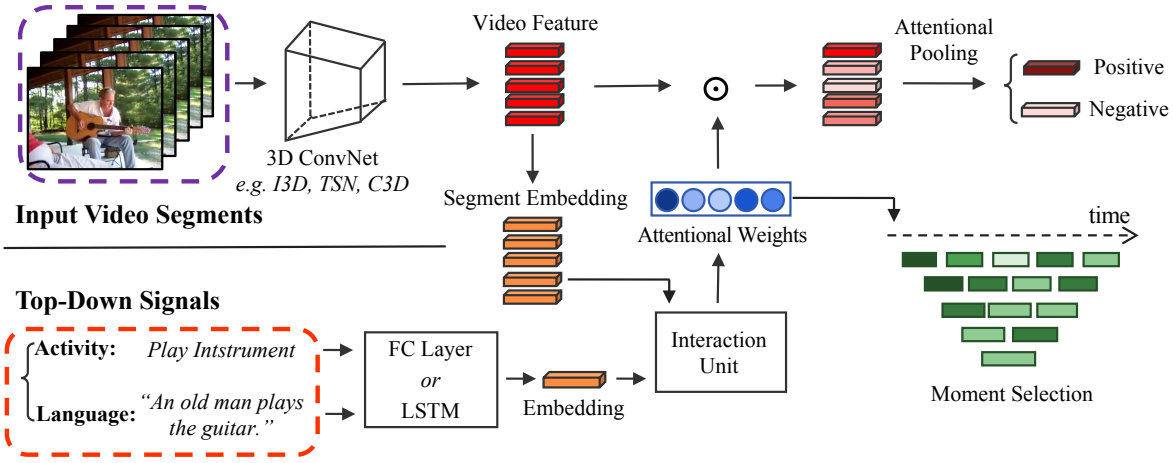

Figure 2: Illustrative diagram for the proposed weakly-supervised framework with referring attention (WSRA). Our model consists of a video stream encoding visual features, and another stream encoding the referring attentional signal spanning from the simplistic action category name to free-form language-based descriptions. We highlight that our framework takes as input the top-down signal provided by user for producing attentional weight for temporal grounding.

vision-language causal verification in the form of temporal order verification between sentences in front of the video as visual evidence.

3. Grounding with Referring Attention and Weak Supervision

Given a long and untrimmed video consisting of M snippets $\mathcal{V} = \{\boldsymbol{v}_i, ...\boldsymbol{v}_M\}$ and a user-provided signal T, our framework of weakly-supervised grounding with referring attention (***WSRA***) is to localize a video segment $\boldsymbol{v}$ from $\mathcal{V}$, which corresponds to the given signal T. The top-down signal is provided by user with different formats, such as the simplistic action category names, short phrases or even free-form language sentences. Based on the form of top-down signals, we can solve different tasks like action localization, temporal grounding, and describable moment retrieval [56, 19, 1]. It is worth mentioning that, WSRA model is trained in a weakly supervised way along with their weak labels at video level; and such weak label can be from either cheap annotations or free-form language captions, like in movie, or even other modalities (e.g., audio) though we did not cover this within the scope of this paper. We elaborate details of our model and training in this section.

3.1. Two-Stream Architecture

As illustrates in Fig. 2, we exploit a two-stream architecture that encodes the videos and top-down referring signals, respectively. Inside the video stream, we also use another mini two-stream model (which combines RGB and optical flow) to extract the global feature and embedding for a video segment. This pre-trained model for video feature extraction is widely used for action classification in literature [15, 8, 69, 68]. But in our work, the key idea of using two-stream architecture is to project the features from video and the top-down referring signals into a common embedding space, in which we build another interaction unit to learn their associations.

The video stream takes as input a stack of M video segments and outputs their feature presentations, denoted by $\{\boldsymbol{f}_1, ..., \boldsymbol{f}_M\}$ corresponding to the M video segments in $\mathcal{V}$, respectively. This is mainly carried out by the so-called 3D ConvNet (shown in Fig. 2, which jointly transforms the RGB and optical flow of each video segment). Moreover, over the feature presentations, it projects these segment features individually into an embedding space $\{\hat{\boldsymbol{t}}_1, ...\hat{\boldsymbol{t}}_M\}$ with two more fully-connected layers and a dropout layer (dropout rate=0.7). The video segment feature $\boldsymbol{f}_i$ is used later for classification or regression, while the embedding $\hat{\boldsymbol{t}}_i$ alone is used for association learning to output the similarity score with the embedding from second stream for the top-down referring signal. We build this stream upon a pre-trained 3D convolutional network [8, 28, 69]. We also tried to use a common feature/embedding on the video stream for both classification and associative learning, but it worked worse, probably due to the fact that video feature and language feature encoded from LSTM contain very different information for classification and associative learning, respectively.

The referring attention stream takes as input the top-down signal and project it into the same embedding space, as shown in the bottom part of Fig. 2. The top-down signal can come with various forms, like one-hot vector to specify the action category, phrase representations by GloVe [48] and the feature representation from an LSTM [22] over a natural language sentence. Even though the top-down sig-

nals may come with different format, this stream can have the same architecture consisting of multiple fully-connected layers.

This top-down signal provides the referring attention to tell the model where to look at temporally thus grounding is fulfilled. We denote the output embedding for the input top-down signal by $\boldsymbol{t}^j$, where subscript j is the video index.

3.2. The Interaction Unit

Over the embeddings from the two streams for video and top-down signal respectively, we build this interaction unit to learn their association measure, which in this work is the similarity score. Mathematically, it outputs the cosine similarity score $\mathcal{S}_{(i,j)}$ between the i^{th} segment in video j ($\hat{\boldsymbol{t}}_i^j$), and the top-down signal for the same video ($\boldsymbol{t}^j$), i.e., $\mathcal{S}_{(i,j)} = \frac{<\boldsymbol{t}^j \cdot \hat{\boldsymbol{t}}_i^j>}{||\boldsymbol{t}^j|| \cdot ||\hat{\boldsymbol{t}}_i^j||}$ The similarity score naturally acts as a soft weight used for gating the visual feature $\mathbf{f}_i$ in the first stream, such that we are able to pool toward a positive hypothesis which is assumed to be associated well with the top-down signal. On the other hand, we can also use $(1-\mathcal{S}_{(i,j)})$ to pool video segment features $\mathbf{f}_i$'s to get a negative hypothesis, corresponding to segments that do not contain the specific action category specified by the top-down signal.

In practice for pooling the segment features with the gating mechanism, instead of using the raw cosine similarity score, we further normalize across all video segments belong to the video in question using a softmax transform [73]. Mathematically, we have the final weight $\mathcal{A}_i^j$ for weighted pooling video segments

$$\mathcal{A}_i^j = \frac{\exp(\mathcal{S}_{(i,j)})}{\sum_{i=1}^{K} \exp(\mathcal{S}_{(i,j)})}. \tag{1}$$

Finally, we have the pooled feature $\mathbf{f}^j$ representing all the positive segments w.r.t the top-down referring signal in video j as below

$$\boldsymbol{f}^j = \frac{\sum_{i=1}^{K} \mathcal{A}_i \cdot \boldsymbol{f}_i}{\sum_{i=1}^{K} \mathcal{A}_i}. \tag{2}$$

Over the pooled features, as well as the top-down referring attention, we are able to exploit sampling tricks and multiple losses to train our model, as detailed in the next section.

3.3. Sampling and Loss for Training

We note that a successful training requires careful sampling and loss functions. In this section, we introduce the sampling trick and losses used in our work.

Sampling for mini-Batch To train a weakly-supervised model for action classification over a fixed number of categories, besides the multi-way classification loss, it is straightforward to additionally exploit a margin-based loss over positive&negative hypothesis pooled with bottom-up attentional weights for better learning [47]. However, when training with regression loss instead of classification over more complex signals like short phrases and natural language sentences which have a combinatorial nature (e.g., complex structure, grammatical tenses and many synonymous words), it poses a challenge of how to define the negative hypothesis and how to measure its similarity to the positive one. For this reason, we propose a sampling strategy that constructs the mini batch during training, within which we can easily exploit the margin-based loss. Specifically, given a video and the top-down referring signal of interest, we sample a few relative and irrelative videos based on the phrase/sentence to form a mini-batch. Within the batch, we are able to not only use the margin-based loss with ease, but also exploit the simple multi-way classification loss just for this batch dynamically during training. We find this simple sampling trick makes training much easier. Our sampling is actually a reflection of curriculum learning [3].

Multi-Instance Loss is widely used in the Multiple-Instance Learning (MIL) [41]. It can be formulated as a classification task that predicts labels of bags, each of which is a collection of unlabeled instances. MIL labels a bag as positive class if at least one instance in it is positive, while negative bag contains no positive instances. In our weakly supervised video grounding, a video can be thought as a bag of segments as individual instances represented by features $\{\mathbf{f}_1, \ldots, \mathbf{f}_M\}$. For example, an action or a natural language sentence can span multiple segments which are not necessarily consecutive in time.

Depending on whether using classification (e.g., action classification) or regression (natural language sentence grounding) to train our weakly-supervised grounding model, we can formulate this MIL loss as below[1]:

$$\mathcal{L}_{mil} = \begin{cases} \sum_{c=1}^{C} -y_c \cdot \log(\bar{p}_c), & C\text{-way classification} \\ \|\mathbf{t} - \bar{\mathbf{p}}\|, & \text{regression} \end{cases} \tag{3}$$

where $\bar{p}_c$ and $\bar{\mathbf{p}}$ are the c^{th}-class confidence score or pooled feature over top-k best segments in the positive hypothesis:

$$\begin{cases} p_c = \frac{1}{k} \max_{\mathbf{p}_i \in \mathbf{S}} \sum_{i=1}^{k} p_c^i, & C\text{-way classification} \\ \bar{\mathbf{p}} = \frac{1}{k} \min_{\mathbf{p}_i \in \mathbf{S}} \sum_{i=1}^{k} \mathbf{p}_i, & \text{regression} \end{cases} \tag{4}$$

wherein $\mathbf{S} = [\mathbf{p}_1, \ldots, \mathbf{p}_M]$ lists the M segments in ascending/descending order for classification/regression based on confidence score or Euclidean distance. Basically, it means the MIL loss only focuses on the best k segments which are supposed to be the positive hypothesis [38, 47]. In our work, we set k proportional to the length of the video in question, *i.e.* 1/8 of its segment number.

Multi-Instance Multi-Label Loss In practice, merely using MIL loss does not guarantee a successful training when

[1] For brevity, we show only the loss on a single video.

a whole video contains many annotations, e.g., phrases and sentences. This situation is exact the Multi-Instance Multi-Label learning (MIML) [86]. For example, a single video not only contains multiple action categories, but also each action spans over multiple video segments. As MIL loss focuses on each class probability independently (as seen in Eq. 3), it cannot handle well in this MIML situation, which usually contain multiple instance and labels with complex correlations among each other. Therefore, we additionally adopt the MIML loss [86, 26, 16], which is based on C-way classification. Owing to our sampling strategy for training, we can also treat the regression problem with as a multi-way classification problem within the sampled videos in the mini batch. Specifically, we use the MIML loss as detailed in [26], and the the loss error for video $\mathcal{V}$ in question w.r.t a specific class c of interest is defined as:

$$\mathcal{L}_{miml} = \sum_{y \in \bar{Y}} \Big(\epsilon(\mathcal{V}, c) \cdot |1 + P_y - P_c|_+ / R(\mathcal{V}, c) \Big) \quad (5)$$

where P_i indicates the classification probability of video $\mathcal{V}$ to the i^{th} class; $|q|_+$ is the function returning 0 if $q < 0$ otherwise q; $\bar{Y}$ records the irrelevant label set; $R(\mathcal{V}, c) = \sum_{y \in \bar{Y}} I(P_y > P_c)$ that sums up indicator function I when condition over the confidence score satisfies $P_y > P_c$; and $\epsilon(\mathcal{V}, c) = \sum_{i=1}^{R(\mathcal{V},c)} \frac{1}{i}$ is the ranking error w.r.t the video $\mathcal{V}$ and label c of interest. However, in practical implementation, we note that by removing the ranking term ϵ and scaling term R, we are able to expedite training without harming the performance in our experiments.

Cross Video Margin-based Loss works over multiple videos that encourages to find precise temporal location while at the same time forces the positive and negative hypothesis to have larger distance. Owing to our sampling strategy that forms the training batch from multiple videos, we are able to exploit this loss easily. Taking a pair of videos ($\mathcal{V}_i$ and $\mathcal{V}_j$) for example, we have positive hypothesis y standing for the common action or phrase or language description, then with the negative hypothesis y', we have our loss as below:

$$\mathcal{L}_{margin} = \big| \|\boldsymbol{f}_y^i - \boldsymbol{f}_y^j\|_2 - \|\boldsymbol{f}_y^i - \boldsymbol{f}_{y'}^j\|_2 \ + \alpha \big|_+, \quad (6)$$

where α is the margin set as 0.5 in our work.

Regression Loss is used for regressing the video segment embedding to the natural language embedding during natural language grounding only:

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_{j=1}^{N} \|\hat{\boldsymbol{t}}^j - \boldsymbol{t}^j\|_2. \quad (7)$$

In practice, we find that using this regression loss alone achieves better performance in sentence grounding than additionally using margin based loss. This suggests grounding a natural language sentence is quite challenging in pooling toward the negative hypothesis, even though our sampling strategy alleviates this issue to some extent.

To summarize, we use the overall loss by combining all the above terms to train our system in an end-to-end manner. We list details on the use of these loss terms in the supplementary material for specific tasks involved in our experiments. During training, we use the Adam optimizer [36] with constant learning 10^{-4} and coefficients 0.9 and 0.999 for computing running averages of gradient and its square. We implement our algorithm using PyTorch toolbox [46] on a single GTX1080 Ti GPU[2].

4. Experiments

The goal of our experiments is to validate the effectiveness of the proposed weakly-supervised framework with the top-down referring attention (*WSRA*) for temporal grounding. We use the mean Average Precision (mAP) under various Intersection over Union (IoU) threshold to measure the performance. We first compare our WSRA with other weakly/fully-supervised approaches three different tasks, action localization, short phrase grounding and natural sentence grounding, on three public benchmarks, THUMOS14 [27], Charades [58, 57] and Charades-STA [19]. To show our WSRA model also allows for verifying the causality between two sentences, we also carry out another experiment with our re-assembled dataset from Charades-STA.

4.1. Action Localization on THUMOS14

The simplistic top-down signal is the one-hot vector specifying one of the fixed classes of interest. Therefore we train our model for the action localization with weakly-supervised setup. We use the popular THUMOS14 dataset, which is a subset of UCF101 [64] and provides 20-class 200/213 videos for training/testing, as commonly done in literature [44, 47]. This dataset is challenging since the videos have varied duration (ranging from seconds to minutes); and there exist multiple actions in a single video segment. To encode the video segment, as done by other compared weakly-supervised methods [44, 47], we turn to the I3D feature from a pre-trained model over the Kinetics dataset [33, 8]. During inference, we input each of the 20 classes separately, and the model predicts frame-wise probability for each class. Then we generate the final localization results from the probabilities matrix of all frames and all classes. We note such a practice is a technique used in energy-based model [37, 50] which predicts the label into the class which produces the minimum entropy. Moreover, to reduce false positive predictions due to the background class, we turn to a thresholding mechanism that removes all the predictions of a segment with its class-specific entropy

[2] Our code and models can be found in https://XXX.XXX/XXX

Table 1: Action localization on THUMOS14. We compare our WSRA model with other weakly supervised approaches, as well as some state-of-the-art fully-supervised methods.

	Method	Fea.	*mAP*@IoU					
			0.1	0.2	0.3	0.4	0.5	0.7
Fully-Supervised	Richard et al. [51]	*Fisher*	39.7	35.7	30.0	23.2	15.2	–
	Shou et al. [56]	*C3D*	47.7	43.5	36.3	28.7	19.0	–
	Yeung et al. [77]	*VGG*	48.9	44.0	36.0	26.4	17.1	–
	PSDF [81]	*TSN*	51.4	42.6	33.6	26.1	18.8	–
	DAPS [14]	*C3D*	–	–	–	–	13.9	–
	Shou et al. [54]	*C3D*	–	–	40.1	29.4	23.3	07.9
	SMS [79]	*PSDF*	51.0	45.2	36.5	27.8	17.8	–
	R-C3D [73]	*C3D*	54.5	51.5	44.8	35.6	28.9	–
	SSN [85]	*TSN*	66.0	59.4	51.9	41.0	29.8	–
Weakly-Supervised	UntrimNet [68]	*TSN*	44.4	37.7	28.2	21.1	13.7	–
	Singh [61]	*C3D*	36.4	27.8	19.5	12.7	6.8	–
	STPN [44]	*UNTF*	45.3	38.8	31.1	23.5	16.2	05.1
	STPN [44]	*I3D*	52.0	44.7	35.5	25.8	16.9	04.3
	W-TALC [47]	*I3D*	55.2	49.6	40.1	31.1	22.8	07.3
	W-TALC^{miml}	*I3D*	**56.3**	**52.1**	**41.9**	**32.5**	**24.5**	**08.5**
	WSRA^{mil}	*I3D*	55.7	51.2	40.1	31.0	22.7	07.4
	WSRA	*I3D*	**56.1**	**51.2**	**41.0**	**31.9**	**24.0**	**08.2**

Table 2: Phrase Grounding on the Charades dataset. Besides reporting our result, we also post-process the results by averaging class-wise predicting probabilities across 15 neighboring segments as used in [57] (marked by *).

	Approach	Features	mAP (%)
Fully Supervised	Random [57]	–	2.42
	RGB [57]	*TSN*	7.89
	PCN [11]	*VGG*	8.90
	Two-Stream [57]	*TSN*	8.94
	Two-stream+LSTM [57]	*TSN*	9.60
	R-C3D [73]	*C3D*	12.07
	I3D [8]	*I3D*	17.22
	Two-stream+LSTM [49]	*I3D*	18.20
	Super Events [49]	*I3D*	**19.41**
Weak	W-TALC	*TSN*	5.79
	W-TALC	*I3D*	14.46
	WSRA	*TSN*	6.27
	WSRA	*I3D*	14.45
	WSRA*	*I3D*	**16.79**

E_c satisfying: $E_c > 0.5 * (E_{max} - E_{min})$, where E_{max} and E_{min} are the max/min entropy over all classes.

We list detailed comparisons in Table 1, including both fully and weakly supervised approaches. We note that the proposed WSRA makes use of the top-down signal thus generating the top-down referring attention, whereas other compared method generates bottom up attentional weights during training. Among them, W-TALC also uses the multi-instance loss during training and achieves the state-of-the-art performance [47]. Then we are interested in how MIML loss works in W-TALC. We re-implement W-TALC using the released code by replacing their MIL loss with our MIML loss. As seen from the table, our re-implemented W-TALC^{miml} achieves even better results, demonstrating that our MIML loss brings performance gain over the original W-TALC trained with MIL loss. With our WSRA model trained with the top-down referring attention and MIML loss, we are able to achieve better performance than W-TALC and on-par to our implemented W-TALC^{miml}. This series of experiments demonstrate the effectiveness of our WSRA framework in weakly supervised action localization task, although we did not tailor our WSRA to this specific action localization task. Next we show through experiments how our WSRA performs for grounding more complex top-down signals, *i.e.* short phrases and natural language sentences.

4.2. Phrase Grounding on Charades

We evaluate our method for phrase grounding on the Charades dataset [58, 57], which is a recent large-scale dataset consisting of 157 phrase descriptions and 11,848 untrimmed videos. With the provided split, we train/test on its 7,985/2,000 videos. On average, each video in Charades contains ∼7 activities, all of which are with structured phrase names from a fixed vocabulary. We train with margin-based loss on regression terms within the mini batch sampled using our strategy. We use the released benchmarking code from the dataset [58] to report the performance. In Table 2, we compare our WSRA method with other fully-supervised approaches. To the best of our knowledge, there is no weakly-supervised methods for phrase grounding on this dataset. So to gain a clear insight how our WSRA performs, we also train the W-TALC model using the release code [47] with different visual features, TSN [69] and I3D [8], respectively. We can see from Table 2 that our WSRA, as well as our implemented W-TALC, achieves very competitive performance compared to the fully-supervised methods. Moreover, by using a post-processing method that averages the neighboring 15 frames for a more temporally consistent predictions as described in [57], we improve the performance further, which is close to the best fully-supervised method. Similar to action localization on THUMOS14, we observe that our WSRA and W-TALC perform on-par to each other without exhibiting the advantage of using top-down referring attention brought by our method. We believe the reason is that grounding phrases on this dataset is similar to action classification on THUMOS14, and both datasets have fixed number of actions/phrases. In Figure. 3, we further conduct an ablation study on effects of different features. We see from the figure that, our model enjoys a performance boosting than W-TALC when using only TSN features in all modalities. And I3D features clearly show an obvious improvement. In next experiment, we highlight the absolute advantage of our WSRA framework on grounding the most complex signals, *i.e.* the natural language sentence-based descriptions.

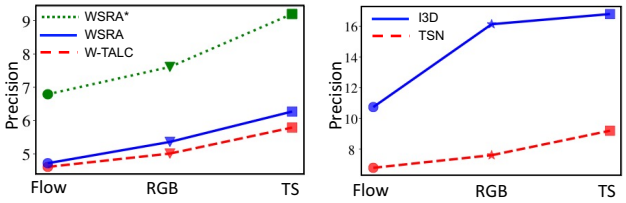

Figure 3: Ablation study on the choice of video features in our WSRA system on the Charades dataset, where Flow, RGB and "TS" denote using the features on optical flow only, RGB only or both. Left figure demonstrates the best performance is achieved when using both flow and RGB (ref. "TS") for both our WSRA and W-TALC method. Right figure shows the performance of our WSRA with different (stronger) video features (I3D vs. TSN).

Table 3: Natural language grounding results on Charades-STA dataset. "WSRA-Verb", "WSRA-Lang" and "WSRA-Fuse" denote the models trained with languages, verb and the fusion of them.

Approach	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
	Fully-Supervised			
Random [19]	08.51	03.03	37.12	14.06
CTRL [19]	21.42	07.15	59.11	26.91
Xu *et al.* [72]	35.60	15.80	79.40	45.40
MAN [84]	46.53	22.72	86.23	53.72
	Weakly-Supervised			
W-TALC	14.65	05.00	44.46	22.66
WSRA-Verb	**29.35**	**10.05**	**56.26**	**33.73**
WSRA-Lang	27.02	08.87	55.62	32.37
WSRA-Fuse	**31.20**	**11.01**	**60.21**	**37.52**

4.3. Language Grounding on Charades-STA

We highlight the advantage of our WSRA through a more challenging task, *i.e.* natural language grounding on the Charades-STA dataset [19], which augments the Charades dataset [58, 57] with manual annotations in the form of natural language descriptions at precise start-end timestamp of each video. Charades-STA contains 12,408/3,720 video-query pairs for training/testing, respectively.

The free-form style of natural language descriptions enables us to explore three ways to use them: 1) encoding the verbs (the phrase from standard Charades dataset) using word2vec embedding from GloVe [48] (denoted by "WSRA-Verb"), 2) encoding the whole sentence with a follow-up LSTM encoder (denoted by "WSRA-Lang"), 3) fusing the above two features (denoted by "WSRA-Fuse"). As for the encoder, we find that adapting from a pre-trained LSTM on other cross-modal tasks, *e.g.* Image Captioning or VQA [78, 2], gives a better performance. In this paper, we use the pre-trained LSTM from [34] to encode sentences, and fix it without updating during training. We use dynamic unrolling setting in our implementation [83, 46] that flexibly encodes sentences in any length.

Event 1: The man opens the refrigerator and takes out a water.

Event 2: He closes the refrigerator and starts drinking the water.

Figure 4: The task of VALVE requires one to verify if two describable moments convey the correct temporal order. We note that such a temporal order usually correlates with causal-effect relationship.

To return the final grounding results, we consider the moment selection methods [19, 84], which generate multi-scale sampled moment candidates for retrieval with fixed length of frames $\{64, 28, 256, 512\}$. After simple statistical study on the training set, we found video length varies significantly across the dataset. Thus, we modify above strategies by sampling the moment candidates with the length to be proportioned to the video duration. Specifically, sampled candidate clips are with $\{20\%, 30\%, 40\%, 50\%\}$ of the whole videos and in 80% overlap. We report accuracy of our system under different mean Intersection-over Union (mIoU = $\{0.5, 0.7\}$) and Recall@$\{1, 5\}$ to measure the performance. In Table 3, we show results of the baseline method and our models. We re-implementing W-TALC, which is a bottom-up attentional method achieving state-of-the-art on weakly supervised action localization, and take sentences as learning target for comparison. With the comparison between our WSRA and W-TALC, we can see the advantage of top-down referring attention becomes obvious: WSRA significantly outperforms W-TALC. When taking as input the complex sentence, WSRA actually performs a little worse than taking the verb as input. This suggests the challenge of encoding free-form natural languages. However, when combining the two encoding methods (verb+language) by multiplying their probabilistic predictions for each class as in [57], we are able to achieve even better performance than using verb alone. We believe this is owing to the fact that, though encoding a complex language is difficult, such language features still provide unique information for grounding itself temporally.

4.4. Vision-Language Causal Verification (VALVE)

The successful training of WSRA on free-form language grounding provides us a unique opportunity to study the new exciting problem, which we call Vision-Language

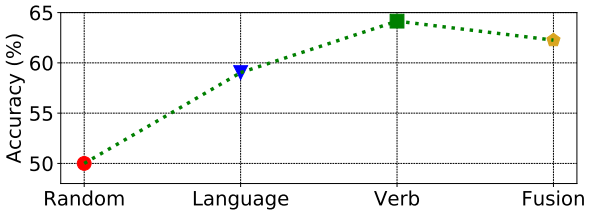

Figure 5: VALVE results on our dataset with different methods/inputs, "Random" is the native baseline; while "Language", "Verb" and "Fusion" indicate our model is trained with different format of top-down referring signals.

Causal Verification (VALVE). VALVE requires one to verify the temporal order of two natural language descriptions with the visual evidence (the video here). While it is essentially a binary classification problem, it is a step toward understanding causal relationship of two describable moments with the help of the visual evidence, *i.e.* the video.

To conduct this study, we collect a new dataset (VALVE dataset) from the Charades [58, 57] and Charades-STA [19] testing sets, with our manual efforts: three different human annotators verify the temporal order of sentence descriptions by watching the video, and answer if the pair of descriptions satisfy the following two conditions, 1) multiple events exist within the single video, 2) events demonstrate high relevance or causal-effect relationships. We consider the event pair as a correct one if at least two annotators agree on the same annotations. In total, we obtain 400 verification pairs, and their corresponding textual descriptions obtained from Charades-STA. Fig. 4 shows an example in our VALVE dataset; and Fig. 6 shows the histogram of video number vs. moment interval. As we do not train any NLP model to understand causal-effect relationship over the two describable sentences, the random guess serves a naive baseline which should be 50% accuracy.

With the models trained previously on the Charades-STA dataset, we overlay in Figure. 6 the VALVE performance vs. the describable moments interval (in seconds). We manually set three groups of testing samples in terms of moment interval, in each of which there are almost the same number of samples. Without surprise, VALVE performance on short-interval moments is worse than those with long-interval moments. This reflects a clear correlation with the performance of sentence grounding[3].

Moreover, we plot in Fig. 5 the VALVE performance vs. feature input taken by our WSRA model. Interestingly, we observe that WSRA trained with verb input only

[3]We note that our model achieves worse than random guess in group-1 within 5sec interval. The reason is that the samples we collect is not balanced. At the same time, it also demonstrates our WSRA is actually learning without random guess even though it is struggling in such short-interval situation.

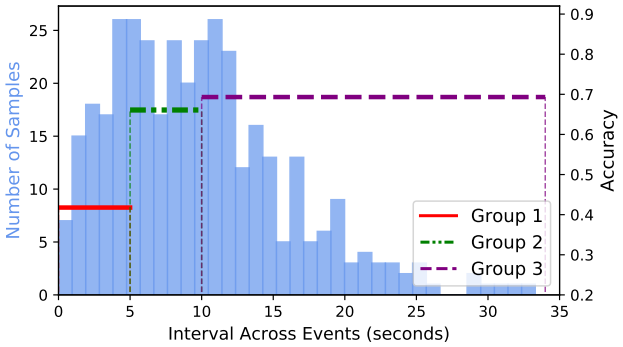

Figure 6: We show statistics of time interval across events (in second). We divide event pairs into three groups according to their time intervals and their corresponding precision. "Group 1, 2, 3" denote the event pairs under different time intervals (0-5s, 5-10s and lager than 10s).

gives a higher accuracy than using language input only or using both. This is different from the observation of sentence grounding in the previous experiment on Charades-STA where the fusion achieves the best grounding performance. We believe a good sentence grounding performance does not guarantee a better VALVE performance; and this suggests VALVE is actually a more challenging task.

5. Conclusion

We introduce a simple, unified weakly-supervised temporal grounding framework learned with top-down referring attention (WSRA). Owing to the top-down grounding nature, we are able to train models over video dataset only with video-level labels. We show our WSRA performs quite well over a range of grounding tasks, spanning from action localization, phrase grounding and natural language grounding. Owing to the top-down referring nature of our model, we are able to explore the possibility of causality understanding with both natural language sentences and the video as visual evidence, in the form of temporal verification. We show promising results on this new problem which we call Vision-Language Causal Verification (VALVE).

To the best of our knowledge, we provide the first benchmark evaluation for grounding short phrases and natural languages by training a model in a weakly supervised learning setup. Moreover, our model provides a unique opportunity to study the new problem of VALVE. This is actually a step toward causality understanding through learning over the abundant videos along with the captions. By demonstrating the effectiveness of our system, as future work, we expect to train our model over more free-form videos with either weakly-supervised learning or even self-supervised learning where the signals are from the free captions or audios, for more specific tasks or more abstract knowledge learning.

References

[1] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5803–5812, 2017. 1, 2, 3, 12

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 7

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009. 4

[4] P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014. 1

[5] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016. 1

[6] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Carlos Niebles. Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2911–2920, 2017. 1, 2

[7] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 1

[8] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 3, 5, 6

[9] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013. 1

[10] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 162–171, 2018. 2, 12

[11] A. Dave, O. Russakovsky, and D. Ramanan. Predictive-corrective networks for action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 981–990, 2017. 6

[12] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[13] A. Ekin, A. M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image processing*, 12(7):796–807, 2003. 1

[14] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem. Daps: Deep action proposals for action understanding. In *European Conference on Computer Vision*, pages 768–784. Springer, 2016. 6

[15] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016. 3

[16] J. Feng and Z.-H. Zhou. Deep miml network. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 5

[17] A. Fire and S.-C. Zhu. Using causal induction in humans to learn and infer causality from video. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35, 2013. 2

[18] A. Fire and S.-C. Zhu. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):23, 2016. 2

[19] J. Gao, C. Sun, Z. Yang, and R. Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5267–5275, 2017. 1, 2, 3, 5, 7, 8, 12

[20] J. Gao, Z. Yang, and R. Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017. 2

[21] K. Gavrilyuk, A. Ghodrati, Z. Li, and C. G. Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966, 2018. 2

[22] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. 1999. 3

[23] T. Gold. The arrow of time. *American Journal of Physics*, 30(6):403–410, 1962. 2

[24] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. Localizing moments in video with temporal language. *arXiv preprint arXiv:1809.01337*, 2018. 1, 2

[25] D.-A. Huang, L. Fei-Fei, and J. C. Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *European Conference on Computer Vision*, pages 137–153. Springer, 2016. 1

[26] S.-J. Huang, W. Gao, and Z.-H. Zhou. Fast multi-instance multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 5

[27] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The thumos challenge on action recognition for videos in the wild. *Computer Vision and Image Understanding*, 155:1–23, 2017. 2, 5

[28] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013. 3

[29] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014. 12

[30] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[31] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4405–4413, 2017. 2

[32] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2

[33] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[34] V. Kazemi and A. Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017. 7

[35] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–578, 2018. 1

[36] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[37] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 5

[38] W. Li and N. Vasconcelos. Multiple instance learning for soft bags via top instances. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4277–4285, 2015. 4

[39] B. Liu, S. Yeung, E. Chou, D.-A. Huang, L. Fei-Fei, and J. Carlos Niebles. Temporal modular networks for retrieving complex compositional activities in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–568, 2018. 12

[40] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T.-S. Chua. Attentive moment retrieval in videos. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 15–24. ACM, 2018. 1, 2

[41] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Advances in neural information processing systems*, pages 570–576, 1998. 4

[42] A. Mishra, V. K. Verma, M. S. K. Reddy, S. Arulkumar, P. Rai, and A. Mittal. A generative approach to zero-shot and few-shot action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 372–380. IEEE, 2018. 1

[43] M. H. Nguyen, L. Torresani, F. De La Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1925–1932. IEEE, 2009. 1

[44] P. Nguyen, T. Liu, G. Prasad, and B. Han. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6752–6761, 2018. 1, 2, 5, 6

[45] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160. IEEE, 2011. 1

[46] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 5, 7

[47] S. Paul, S. Roy, and A. K. Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018. 1, 2, 4, 5, 6

[48] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3, 7

[49] A. Piergiovanni and M. S. Ryoo. Learning latent super-events to detect multiple activities in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5304–5313, 2018. 6

[50] C. Poultney, S. Chopra, Y. L. Cun, et al. Efficient learning of sparse representations with an energy-based model. In *Advances in neural information processing systems*, pages 1137–1144, 2007. 5

[51] A. Richard and J. Gall. Temporal action detection using a statistical language model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3131–3140, 2016. 6

[52] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, pages 1593–1600. IEEE, 2009. 1

[53] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1234–1241. IEEE, 2012. 1

[54] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5734–5743, 2017. 1, 2, 6

[55] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang. Autoloc: weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–171, 2018. 2

[56] Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058, 2016. 1, 2, 3, 6

[57] G. A. Sigurdsson, S. Divvala, A. Farhadi, and A. Gupta. Asynchronous temporal fields for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2017. 2, 5, 6, 7, 8

[58] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, 2018. 2, 5, 6, 7, 8

[59] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data

collection for activity understanding. In *European Conference on Computer Vision*, 2016. 1, 2

[60] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 2

[61] K. K. Singh and Y. J. Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3544–3553. IEEE, 2017. 1, 6

[62] P. Siva and T. Xiang. Weakly supervised action detection. In *BMVC*, volume 2, page 6, 2011. 1

[63] C. G. Snoek, M. Worring, and A. W. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM, 2005. 1

[64] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 5

[65] C. Sun, S. Shetty, R. Sukthankar, and R. Nevatia. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 371–380. ACM, 2015. 1, 2

[66] D. Tran and J. Yuan. Max-margin structured output regression for spatio-temporal action localization. In *Advances in neural information processing systems*, pages 350–358, 2012. 2

[67] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7190–7198, 2018. 2

[68] L. Wang, Y. Xiong, D. Lin, and L. Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4325–4334, 2017. 1, 3, 6

[69] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 2, 3, 6

[70] X. Wang, A. Farhadi, and A. Gupta. Actions˜ transformations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2658–2667, 2016. 1

[71] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE international conference on computer vision*, pages 3164–3172, 2015. 1, 2

[72] H. Xu, K. He, L. Sigal, S. Sclaroff, and K. Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, volume 2, page 7, 2019. 7

[73] H. Xu, B. Kang, X. Sun, J. Feng, K. Saenko, and T. Darrell. Similarity r-c3d for few-shot temporal activity detection. *arXiv preprint arXiv:1812.10000*, 2018. 1, 4, 6

[74] H. Xu, B. Li, V. Ramanishka, L. Sigal, and K. Saenko. Joint event detection and description in continuous video streams. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 396–405. IEEE, 2019. 2

[75] Y. Xu, C. Zhang, Z. Cheng, J. Xie, Y. Niu, S. Pu, and F. Wu. Segregated temporal assembly recurrent networks for weakly supervised multiple action detection. *arXiv preprint arXiv:1811.07460*, 2018. 2

[76] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos. Robot learning manipulation action plans by" watching" unconstrained videos from the world wide web. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 1

[77] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2678–2687, 2016. 6

[78] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 7

[79] J. Yuan, B. Ni, X. Yang, and A. A. Kassim. Temporal action localization with pyramid of score distribution features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2016. 6

[80] Y. Yuan, Y. Lyu, X. Shen, I. W. Tsang, and D.-Y. Yeung. Marginalized average attentional network for weakly-supervised learning. 2018. 2

[81] Z. Yuan, J. C. Stroud, T. Lu, and J. Deng. Temporal action localization by structured maximal sums. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3684–3692, 2017. 6

[82] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 2

[83] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014. 7

[84] D. Zhang, X. Dai, X. Wang, Y.-F. Wang, and L. S. Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. *arXiv preprint arXiv:1812.00087*, 2018. 7, 12

[85] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017. 6

[86] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In *Advances in neural information processing systems*, pages 1609–1616, 2007. 5

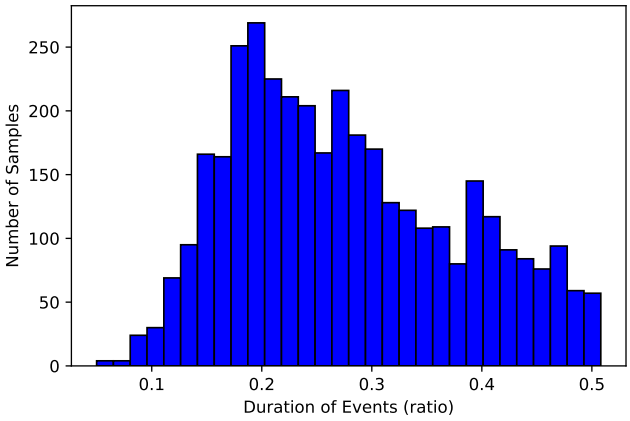

Figure 7: Histogram distribution between video number and even duration in Charades-STA dataset.

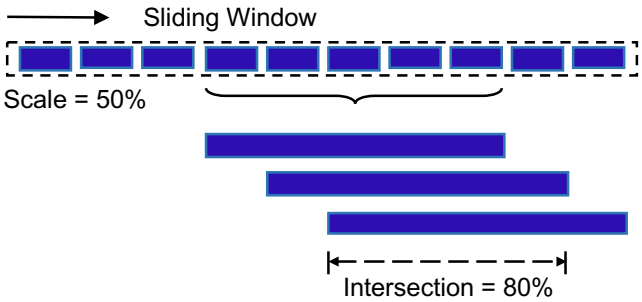

Figure 8: We generate the moment candidates using the sliding window method under different scales. The above example shows the generation process when scale is 50%.

Appendix

In this appendix, we provide more details of our language grounding for moment retrieval, and show more visualizations on weakly-supervised action localization and examples of the Vision-language Causal Verification (VALVE) dataset. We also analyze several failure examples on the action localization task.

1. Language Grounding for Moment Retrieval

In Figure 7, we show the histogram distribution of video numbers vs. event length in the training set of Charades-STA dataset [19]. We observe that the majority length of events distributed over the range from 20% to 50%. We generate our proposals in a sliding window method (as shown in Figure 8) under different scales and with intersection 80% across consecutive moments. We average-pool the segment features within each moment as the moment level features. Note that when trained on Charades-STA, the attentional weights are directly applied on each moment.

2. Visualization of Action Localization

We demonstrate several action localization results on THUMOS14 [29] in Figure 10, 11 and 9. These results are generated by our model trained in a weakly supervised fashion. The attentional weights demonstrate good discrimination across the temporal axis in various activities. In Figure 11, the predicted range of *LongJump* shows a large difference with the ground-truth, and the predicted boundaries for most of the events are not perfectly matching with the groundtruth. Twh reaons explain this. First, we do not exploit any strong supervision (precise time boundaries of events) during training, so it is difficult to capture precise temporal boundaries. Second, the temporal boundary for an action is hard to define by nature, and even human annotators may be confused by the action temporal boundaries.

Approach	Rank@1	Rank@5	mIoU
	Fully-Supervised		
TMN [39]	18.71	72.97	30.14
TGN [10]	24.28	71.43	38.62
MCN [1]	24.42	75.40	37.69
MAN [84]	**27.02**	**81.70**	**41.16**
	Weakly-Supervised		
WSRA-Lang	07.65	33.55	22.10
WSRA-Lang*	12.14	53.24	26.07

Table 4: Natural language moment retrieval results on DiDeMo dataset.

3. More examples of VALVE

We list several more examples of the Vision-Language Causal Verification (VALVE) in Figure 12. Samples in the first and second row are cases when event pairs are of short time intervals, most of which consist of successive human actions. Last sample in Figure 12 shows the events of a long time interval, however with a higher correlation or stronger causal effect relationship.

4. Additional Results on DiDeMo Datasets

We report additional results on DiDeMo dataset in Table. 4. We adopt official provided frame-wise VGG features in [1]. However, considering videos in DiDeMo dataset are adapted from YFCC100M [**?**], which are mostly about outdoors and diverse activities, we only train it with languages on regression loss. According to [1], annotated moments in DiDeMo are mostly shorter than 10 seconds, we post-process our prediction on only short moments and report results as "WSRA-Lang*".

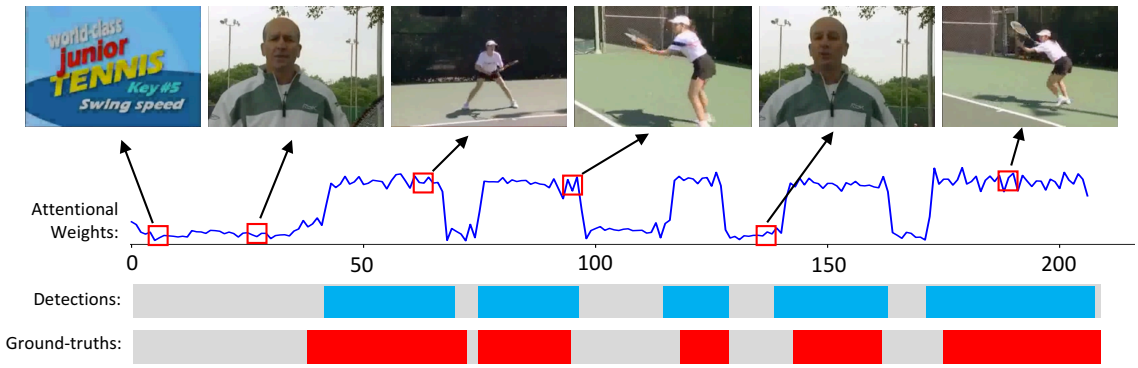

Figure 9: Example of *TennisSwing* Action from THUMOS14 dataset.

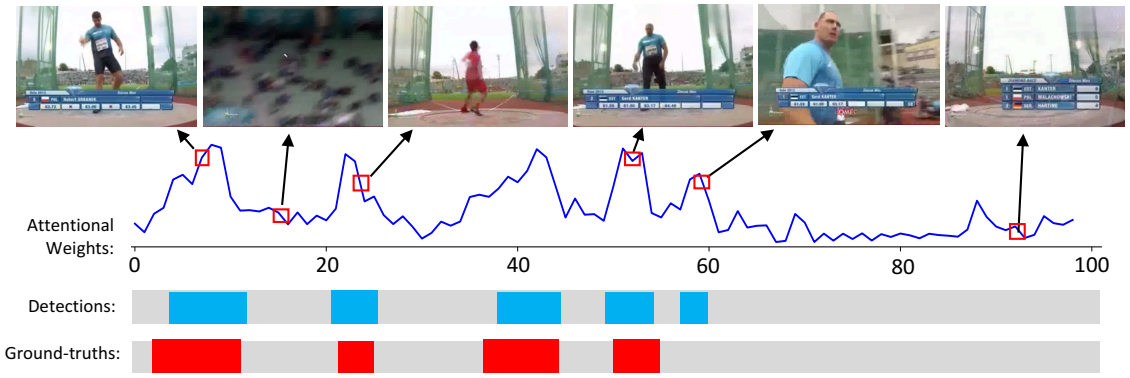

Figure 10: Example of *ThrowDiscus* Action from THUMOS14 dataset.

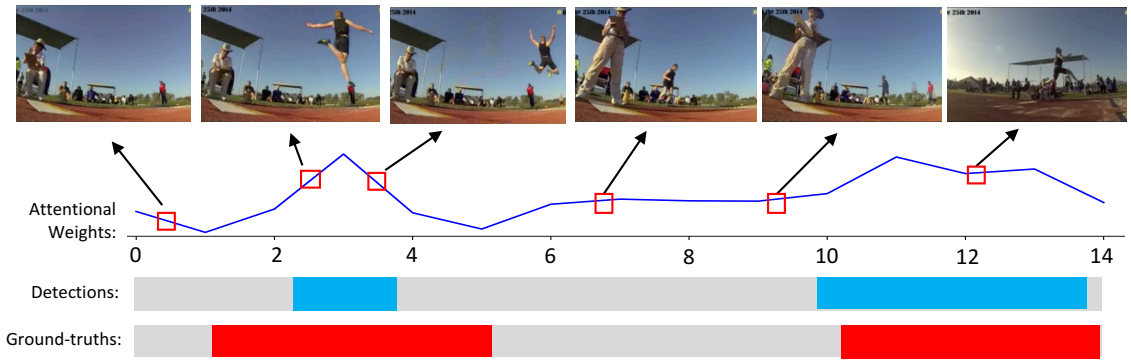

Figure 11: Example of *LongJump* Action from THUMOS14 dataset.

Figure 12: Example of Vision-Language Causal Verification tasks.