Explore Inter-Contrast Between Videos via Composition for Weakly Supervised Temporal Sentence Grounding

Jiaming Chen,[*1] **Weixin Luo,**[*2] **Wei Zhang,**[†1] **Lin Ma**[†2]

[1]School of Control Science and Engineering, Shandong University
[2]Meituan
{ppjmchen, forest.linma}@gmail.com, luowx@shanghaitech.edu.cn, davidzhang@sdu.edu.cn

Abstract

Weakly supervised temporal sentence grounding aims to temporally localize the target segment corresponding to a given natural language query, where it provides video-query pairs without temporal annotations during training. Most existing methods use the fused visual-linguistic feature to reconstruct the query, where the least reconstruction error determines the target segment. This work introduces a novel approach that explores the inter-contrast between videos in a composed video by selecting components from two different videos and fusing them into a single video. Such a straightforward yet effective composition strategy provides the temporal annotations at multiple composed positions, resulting in numerous videos with temporal ground-truths for training the temporal sentence grounding task. A transformer framework is introduced with multi-tasks training to learn a compact but efficient visual-linguistic space. The experimental results on the public Charades-STA and ActivityNet-Caption dataset demonstrate the effectiveness of the proposed method, where our approach achieves comparable performance over the state-of-the-art weakly-supervised baselines. The code is available at https://github.com/PPjmchen/Composition_WSTG.

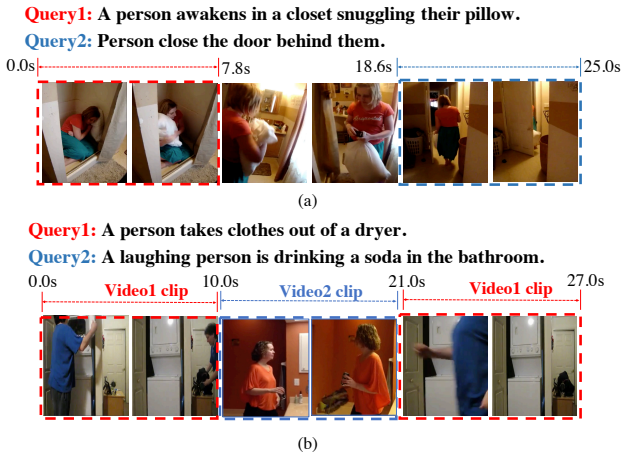

Figure 1: (a) Temporal sentence grounding in videos. Camera movements and scene transitions cause intra-contrast inside the same video. (b) Composing clips from different videos produces numerous artificial video samples and pseudo temporal labels.

Introduction

Temporal sentence grounding (Gao et al. 2017) is regarded as a crucial vision-language task that aims at localizing the temporal boundary of the target segment in an untrimmed video when given a natural language query. As a fundamental problem in video analysis and understanding, it has attracted increasing research interest over the last few years.

Many fully supervised methods (Anne Hendricks et al. 2017; Mithun, Paul, and Roy-Chowdhury 2019; Yuan, Mei, and Zhu 2019; Yuan et al. 2020; Wang, Ma, and Jiang 2020; Zhang et al. 2020a, 2021) has been developed, which directly learn from the videos, the text queries, and the temporal boundaries. Such methods rely on a variety of multi-modal information fusion and supervision from accuracy timestamp labels to model the boundary and content information of the target segments. However, the fully supervised setting needs precise temporal boundary annotations that are expensive and time-consuming. This inevitable limitation prevents the temporal sentence grounding from being applied at a large scale.

The limitations come from the precise but labor-intensive boundary annotations, while the text query annotations are much easier to collect (Lin et al. 2020). To this end, the weakly supervised temporal sentence grounding setting is introduced, where coarse text-video pair annotations are provided only during training. Without boundary annotations of the target segments, recent methods (Lin et al. 2020; Song et al. 2020) fuse the visual and linguistic features and reconstruct either themselves or the masked query. The target segment accompanied by the text query is expected to yield the least reconstruction error during testing. These methods, however, cannot guarantee that the target segments always correspond to the best reconstruction due to the powerful representation of deep neural networks that can even well reconstruct the irrelevant clips.

Videos describe a sequence of actions accompanied by various participants and complex backgrounds, leading to a

[*]These authors contributed equally.
[†]Corresponding authors

contrast of these elements among different videos. We name it inter-contrast in this paper. Due to camera movements and scene transitions, one video can be treated as a combination of a series of clips, which undoubtedly brings temporal correlations and intra-contrast between these clips, as shown in Figure 1 (a). Inspired by this observation, we intend to use the composition of clips from different videos to simulate a collection of clips inside a natural video. As shown in Figure 1 (b), the built-up video-video pairs can be regarded as the simulation of clip-clip pairs inside a single video. Under the weakly supervised setting, this approach can generate a series of re-organized samples with pseudo-labels of temporal boundaries for training. That means we can easily tackle the weakly supervised setting by the fully supervised view. Suppose the model can learn the inter-contrast between clips sampled from different videos and successfully locates all the clips based on their natural language descriptions. In that case, such a mechanism can be seamlessly applied to a regular testing video to locate the target segment given the corresponding text query. We propose a novel method for weakly supervised temporal sentence grounding via the proposed video composition strategy based on this assumption.

The key idea of this paper is that the training videos are composed of different videos, yielding numerous and precise composition timestamps. With the crafted video, we further carefully investigate how to separate it into different semantic components for different text queries. Given one of the text queries from the two sampled videos, the composed video can be treated as a collection of text-related and text-unrelated segments. The former explicitly indicates several temporal annotations corresponding to the given query. We then reverse the temporal annotations for the other query text. Thus, it provides supervision signals to train the weakly supervised temporal sentence grounding task. As shown in Figure 1 (b), we can create a reliable synthetic training sample with precise temporal boundaries. The network takes it as input and performs temporal grounding by learning to separate the composed video to different semantic components conditioned on the respective texts. Besides, a negative query sample is introduced, which is critical to prevent the model from learning to locate the target segments using the obvious scene transitions between different videos. To effectively capture rich semantics in the video and text query as well as encode videos and texts into a unified sequence representation, we utilize a transformer-based model adopted in (Zhang et al. 2021), which is trained by the above three tasks together.

To summarize, our main contributions are three-fold: i) we tackle the weakly supervised temporal sentence grounding task with a novel approach based on learning to ground segments composed from different videos in training, where such composition brings pseudo temporal annotations for training; ii) we propose a multi-queries grounding task in the fully supervised manner where it recognizes the composed video from different query sentences, predicts composition ratios in a global view and predicts masked word prediction task to better align visual-linguistic features. All these tasks are seamlessly integrated into an emerging transformer-based model; iii) our method achieves comparable performance over the state-of-the-art weakly-supervised approaches (Gao et al. 2019; Lin et al. 2020; Song et al. 2020; Wu et al. 2020; Zhang et al. 2020c; Ma et al. 2020; Tan et al. 2021; Zhang et al. 2020b) on the public Charades-STA and ActivityNet-Caption dataset.

Related Work

Fully Supervised Temporal Grounding

As one of multi-modalities tasks (Zhang et al. 2019b; Arbelle et al. 2021; Zhang et al. 2021; Wang et al. 2018, 2019), temporal sentence grounding is a task to temporally localize the corresponding segments in a video based on a given sentence query, initially proposed by (Gao et al. 2017; Anne Hendricks et al. 2017). Until now, it remains challenging since it requires understanding the semantics of both vision and language and realizing alignment between these two modalities. Early work (Gao et al. 2017) utilizes sliding windows to extract candidate video segments. The segment features are extracted and then fused with text representation by various multi-modal fusion operations (i.e., Add, Multiply, and Fully-connected) for estimating the correlation between queries and clips. (Anne Hendricks et al. 2017) proposes to minimize the squared distance between candidate clips and query sentences in the joint representation space. The two early methods follow a simple pipeline of extracting a set of candidate clips and then matching the clip and the given query. The following work adopts more sophisticated strategies such as Graph Convolutional Network used in (Zhang et al. 2019a), attention mechanisms used in (Yuan, Mei, and Zhu 2019), and reinforcement learning used in (He et al. 2019). In addition, (Wang, Ma, and Jiang 2020) introduces a boundary-aware method that keeps the paradigm of candidate matching and proposes an additional branch to predict the temporal boundaries of the target segments. 2D Temporal Adjacent Network (Zhang et al. 2020a) proposes to embed segment representation and language representation into a 2D feature map and utilizes 2D convolutions to align these two modalities and extract adjacent relationships among candidate segments. Inspired by recently proposed multi-modal BERT models, (Zhang et al. 2021) presents a visual-language transformer backbone and utilizes multi-stage feature aggregation to produce discriminative representation for temporal sentence grounding.

Weakly Supervised Temporal Grounding

Although temporal sentence grounding methods succeed under the fully supervised setting, they rely on labor-intensive labeling. In contrast, weakly supervised temporal sentence grounding learns from coarse video-query pairs without timestamp ground truths. In (Mithun, Paul, and Roy-Chowdhury 2019), a text guided attention module (TGA) is proposed to extract frame-wise attention scores, and sliding windows with different sizes are utilized to aggregate segment attention scores. The segments with the highest matching scores are considered as predictions in the testing stage. (Tan et al. 2019) proposes a multi-level co-attention mechanism for multi-modal alignment and utilizes the positional encoding to construct the temporally-aware multi-

modal representations. (Gao et al. 2019) designs a multi-instance learning framework consisting of a classification module and a ranking module. The former measures the matching scores between candidate segments and query sentences, while the latter intends to rank and select the candidates. Inspired by sentence reconstruction in visual grounding (Liu et al. 2019), (Song et al. 2020) proposes to reconstruct the whole sentence from weighted features of candidate video segments. (Lin et al. 2020) proposes to reconstruct the masked keywords of the query sentence from the candidate video segments, as the reconstruction loss will guide the network to discover high-quality proposal segments by assigning higher confidence scores to them. (Ma et al. 2020) proposes a Video-Language Alignment Network (VLANet) based on contrastive learning. Firstly, it utilizes the Surrogate Proposal Selection module to select candidate proposals in a multi-modal representation space. Then the Cascaded Cross-modal Attention module learns to align two modalities based on feature interactions and multi-directional attention flows. (Wu et al. 2020) proposes a Boundary Adaptive Refinement (BAR) network based on reinforcement learning, which designs a cross-modal evaluator for computing the matching scores between video segments and queries and providing boundary-flexible and content-aware rewards. (Zhang et al. 2020b) proposes a Regularized Two-Branch Proposal Network (RTBPN), which designs a language-aware filter to generate an enhanced video stream and a suppressed one and utilizes a two-branch framework to learn from inter-sample and intra-sample confrontments simultaneously. (Tan et al. 2021) leverages a multi-level co-attention mechanism and proposes a Latent Graph Co-Attention Network (LoGAN). The network extracts the context cues from all possible pairs of frames through frame-by-word interactions and learns contextualized visual-semantic representations.

Proposed Method

Problem Formulation

Mathematically, given a set of video-query pairs $P_{test} = \{(V_1, Q_1), (V_2, Q_2), ..., (V_N, Q_N)\}$, the weakly supervised temporal sentence grounding task intends to predict the temporal range $\{(s_1, e_1), (s_2, e_2), ..., (s_N, e_N)\}$ of the target segments. Here s_i and e_i denote the start timestamp and the end timestamp of the i-th target segment and N denotes the number of pairs in the testing set. During training, it provides another set of video-query pairs $P_{train} = \{(V_1, Q_1), (V_2, Q_2), ..., (V_M, Q_M)\}$ only without timestamp annotations. Here M represents the number of pairs in the training set and $P_{train} \cap P_{test} = \emptyset$.

Following (Zhang et al. 2020a, 2021), each video is evenly divided into N_C clips. The feature vectors in each clip are extracted from a pre-trained CNN model and averaged across time dimensions, resulting in a single clip feature vector with a dimension of D_V. A downsampling 1D average pooling layer is applied over all the clip feature vectors so that each video can be represented to $V = \{f_i\}_{i=1}^{N_V}$. Similarly, we use a pre-trained GloVe (Pennington, Socher, and Manning 2014) model to embed each word in the query sentence and pad them to a fixed length N_Q, which can be denoted as $Q = \{w_i\}_{i=1}^{N_Q}$.

The following sections are organized as follows. First, our used temporal sentence grounding framework is introduced. Afterward, we present the proposed composition strategy, including composing input videos and generating pseudo-labels for training the network in detail. Last but not least, the training losses and the inference details are illustrated.

Temporal Sentence Grounding Framework

Since our proposed composition strategy naturally provides pseudo labels for training, any advanced, fully supervised method can be seamlessly integrated into our framework. This paper constructs our approach based on the cutting-edge transformer introduced.

Recently, a multi-modal transformer (Su et al. 2019; Chen et al. 2020; Lin et al. 2020; Zhang et al. 2021) is proposed to address a series of vision-language tasks. Such networks have a similar framework as shown in Figure 2, where the visual data and text data are embedded into an input-sequence of feature tokens that interact inside the multi-head self-attention layers of the transformer network. Thus, the produced output-sequence aggregates cross-modalities information. We then introduce the technical details of such a robust network. More specifically, the visual features and the text features are mapped to the same dimension, where each of them is regarded as a feature token. We then put them together as the sequence inputs added with the standard position embedding. As discussed in (Zhang et al. 2021), to better model different modalities, the parameters of the self-attention layers are decoupled to two groups for visual and language, respectively. In the process of interaction inside the self-attention layers, two modalities are fused in joint representation space. Similar to other multi-modal transformers, the output sequence is divided back into two groups, which corresponds to the original visual sequence and text sequence as shown in Figure 2.

To our best knowledge, this is the first work to tackle the weakly supervised temporal sentence grounding in a fully supervised manner. We modify the multi-modal transformer (Zhang et al. 2021) by introducing video composition and pseudo-label generation, including query pseudo-label, localization pseudo-label, and video-query alignment pseudo-label. It turns out that our method not only intends to align the visual target segment and the linguistic query but also explores the inter-contrast between videos due to the composition.

Composition Strategy

Video Composition

In this work, we propose a video composition method to produce artificial video samples for training. There exist an inter-contrast between different videos naturally. We assume that the model can well learn the visual-linguistic corresponding within a regular video as long as it has successfully captured the counterpart within a composed video. That means we replace the investigation of intra-contrast within a single video with the exploration of inter-contrast

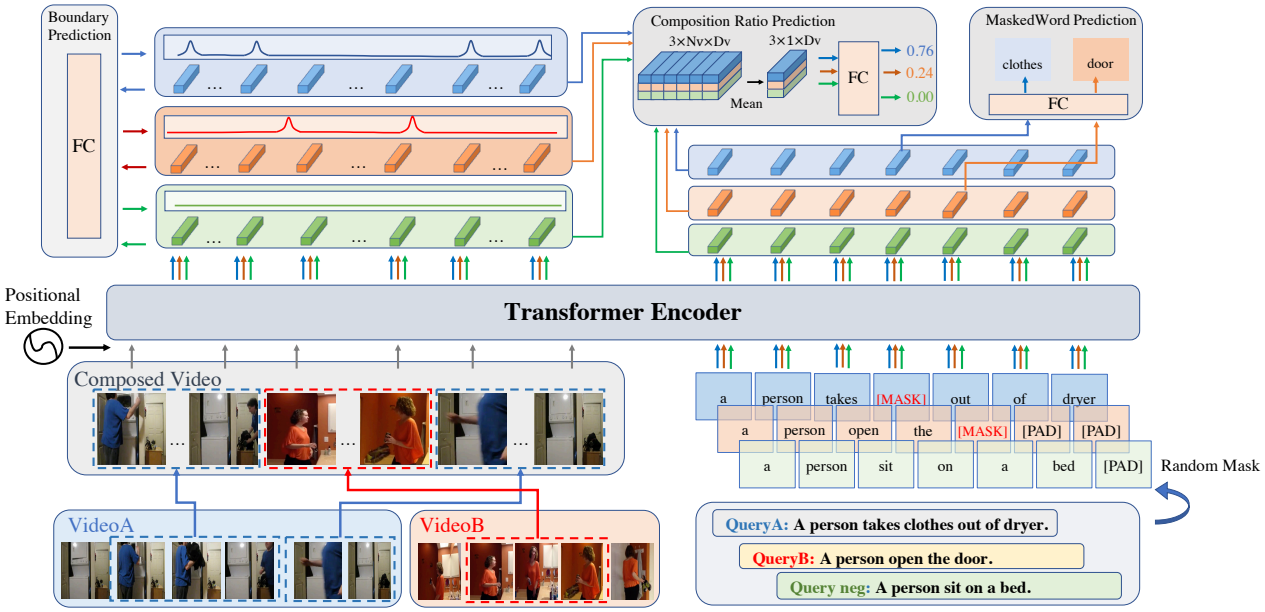

Figure 2: The main framework of the proposed method. Two video-query pairs are randomly sampled from the training set, and the videos are broken into clips and composed randomly. The visual and text sequences are mapped to the same dimension feature vectors and put together and fed into the transformer encoder. Conditioned by different relevant queries, the output sequence is aggregated in the prediction head, which performs multi-task learning: boundary prediction, composition ratio prediction and masked word prediction.

between different videos when temporal annotations related to queries are absent in the training set. More specifically, as shown in Figure 1 (b), two video-query pairs $(V_1, Q_1), (V_2, Q_2)$ are sampled randomly from the training set. Recall each video holds the same length N_V. Therefore we randomly select a clip in the second video and its length ratio l accounts for the whole video and satisfies $r_1 \leq l \leq r_2$, where $0 < r_1, r_2 < 1$ are hyperparameters empirically selected. The selected clip is inserted into another random position of the first sampled video. It should be noticed that the two queries remain the same.

Pseudo-Labels Generation

As shown in Figure 2, composing videos naturally produces artificial videos with clear boundary information, which is absent in most of existing weakly supervised temporal sentence grounding methods. Our framework provides three kinds of pseudo-labels for training a network:

Query Pseudo-Labels Aforementioned, contrast between clips always happens within a regular video due to camera movements and scene transitions. This property can be easily simulated via video composition because of the intercontrast between different videos. During the composition of two different videos from the training set, the two corresponding queries remain the same. Meanwhile, we randomly choose a query sentence corresponding to the third video not involved in composition. Conditioned by these three query sentences, the network should yield different responses. We train the network with the first two query sentences relevant to the composed video as the fully supervised setting always does.

It is worth noting that inserting V_2 clips will interrupt V_1 and split V_1 into two parts in the composed video where the Q_1 is used to recover the starting and ending timestamps twice. The network, however, may easily locate the target segments due to the obvious boundary artifacts in the composed videos. To this end, we also add the irrelevant query as a regularization used in (Arbelle et al. 2021), in which the network should refuse to respond when it takes the irrelevant query as an input.

Localization Pseudo-Labels Besides target classification and boundary localization commonly used in many grounding tasks, the previous work (Wang, Ma, and Jiang 2020; Zhang et al. 2021) proposes to carry out classification on moments such as starting time and ending time, where it can contribute to the grounding scores. Thus, we construct the localization pseudo-labels by each clip's temporal boundaries, i.e. starting timestamp t_s and ending timestamp t_e, as well as their middle timestamp $t_m = \frac{t_s+t_e}{2}$. Rather than producing a single positive label at each moment, we generate the moment classification ground-truths with multiple Gaussian distributions following (Zhang et al. 2021). Specifically, for a clip spanning over $[t_s, t_e]$, the moment classification ground-truths d^s, d^e, and d^m of the t-th output visual feature in the transformer encoder can be formulated as follows:

$$\begin{cases} d^s = e^{-\frac{(t-t_s)^2}{2\sigma^2}} \\ d^e = e^{-\frac{(t-t_e)^2}{2\sigma^2}} \\ d^m = e^{-\frac{\left(t-\frac{t_s+t_e}{2}\right)^2}{2\sigma^2}} \end{cases}$$

Here the range of t_s, t_e is normalized to $[0, N_C]$. The $\sigma = \alpha(t_s - t_e)$ is the deviation of the Gaussian distribution and the α denotes the controlling factor for sharpness. Such localization pseudo-labels from the composed videos will be verified in experiments where they can effectively guide the network to conduct accurate grounding in regular videos even under the weakly supervised setting.

Video-query Alignment Pseudo-Labels Perform localization for all the composed clips may ignore the global information of the whole composed video. To remedy this problem, we propose to predict the proportions of the clips corresponding to different queries, which requires the network to aggregate all the information to produce the predictions. Concretely, an average operation is carried out over all the visual features, and a fully-connected layer is used to produce proportion scalars, as shown in Figure 2.

Training

As introduced above, three pseudo-labels are constructed from the composed video, and we then design three tasks as follows: (1) Proposal classification and moment classification; (2) Composition ratio prediction; (3) Masked word prediction. For simplicity, we sample two different video-query pairs (V_a, Q_a) and (V_b, Q_b) to compose a video, and we also sample another video-query pair (V_n, Q_n) as a negative sample. Since we paste a clip C_b randomly sampled from V_b into V_a, the composed video consists of a clip from V_b and two clips from V_a. Here we take the C_b with the query Q_b as an example for simplicity.

Proposal Classification and Moment Classification The temporal sentence grounding intends to temporally locate the target segment related to the query, which can be regarded as a classification task and a localization task. Thus, we follow the proposed classification and the moment classification used in (Zhang et al. 2021). Initially, video segment proposals are proposed in a sparse way introduced in (Zhang et al. 2020a). Those proposals of which Intersection over Union(IoU) with C_b is higher than a predefined threshold t_h are treated as positive samples, and those of which IoU is below than another predefined threshold t_l are treated as negative samples during training. Other proposals that do not satisfy these requirements will not participate in training.

The objective function designed for the proposed classification is defined as follows.

$$L_P = \sum_{i \in \{a,b,n\}} L_{CE}(\hat{g}_i, g_i)$$

where L_{CE} is a cross-entropy loss and $\hat{g}$ is the proposal score. For Q_a and Q_b, the ground-truth g is set to 1 for positive proposals, otherwise 0 for negative ones. Conditioned by the negative query Q_n, the ground-truth g_n is set to all zeros.

As for the moment classification, the model is required to classify the moment for each one of N_V steps. The objective function of the moment classification is defined below.

$$L_B = \sum_{i \in \{a,b,n\}} L_{CE}(\hat{d}_i^s, d_i^s) + L_{CE}(\hat{d}_i^m, d_i^m) + L_{CE}(\hat{d}_i^e, d_i^e)$$

where $\hat{d}^s$, $\hat{d}^m$, and $\hat{d}^e$ are the boundary predictions for the starting moment, the middle moment, and the ending moment, respectively. Meanwhile, the ground-truths d^s, d^m, and d^e have been illustrated in the previous subsection 3.3.2. Conditioned by the negative query Q_n, we set the ground-truth d_n to all zeros.

Composition Ratio Prediction We also construct a task on a video level. After averaging all the output features, a fully-connected layer is used to predict the ratio $\hat{r}$ of C_b in the composed video given Q_b. Conditioned by the two relevant queries Q_a and Q_b, the model should predict their own proportion in the composed video. Conditioned by the negative query Q_n, we set the ground-truth r_n to all zeros. The loss is designed as follows:

$$L_R = \sum_{i \in \{a,b,n\}} L_{CE}(\hat{r}_i, r_i)$$

Masked Word Prediction To better align visual and linguistic features, we mask each word in Q_a, Q_b with a probability of 0.15 and use the "[MASK]" token to replace them. For the output feature of the masked words, the transformer should reconstruct the masked words w from other unmasked ones, and the visual information, which is widely used in other multi-modal transformer frameworks (Zhang et al. 2021) and the standard cross-entropy loss is used here:

$$L_M = \sum_{i \in \{a,b\}} L_{CE}(\hat{w}_i, w_i)$$

where $\hat{w}$ is the predicted masked words.

Total loss During the training phase, we compute the weighted sum for all the losses above as follows:

$$L = w_P * L_P + w_B * L_B + w_R * L_R + w_M * L_M$$

where the w_P, w_B, w_R, w_M are hyper-parameters to indicate the importance of each task.

Inference

In the inference stage, the composed video is replaced by the testing video, which is fed to the transformer encoder with the unmasked query sentence together. All of the candidate clips are sorted by the mean of proposal classification and moment classification scores as the prediction results.

Experiments

To verify the effectiveness of our composition strategy, we perform experiments on two public benchmark datasets ActivityNet-Captions (Caba Heilbron et al. 2015) and Charades-STA (Sigurdsson et al. 2016). Since our method is the first to tackle the weakly supervised temporal sentence grounding task in a fully supervised manner, we compare the proposed method with both the recent state-of-the-art fully supervised methods MCN (Anne Hendricks et al. 2017),ABLR (Yuan, Mei, and Zhu 2019), TGN (Mithun, Paul, and Roy-Chowdhury 2019), CBP (Wang, Ma, and Jiang 2020), SCDM (Yuan et al. 2020) 2D-TAN (Zhang et al. 2020a), MAT (Zhang et al. 2021) and weakly supervised approaches WSLLN(Gao et al. 2019), SCN (Lin et al.

Method	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5
Fully supervised Methods						
MCN	42.80	21.37	9.58	-	-	-
ABLR	73.30	55.66	36.79	-	-	-
TGN	70.06	45.51	28.47	79.10	57.32	44.20
CBP	-	54.30	35.76	-	77.63	65.89
SCDM	-	54.80	36.75	-	77.29	64.99
2D-TAN	-	59.45	44.51	-	85.53	77.13
MAT	-	-	48.02	-	-	78.02
Weakly supervised Methods						
WSLLN	**75.40**	42.80	22.70	-	-	-
SCN	71.48	47.23	29.22	90.88	71.45	55.69
MARN	-	47.01	29.95	-	72.02	57.49
BAR	-	49.03	30.73	-	-	-
VGN	-	**50.12**	**31.07**	-	77.36	61.29
RTBPN	73.73	49.77	29.63	**93.89**	79.89	60.56
Ours	71.86	46.62	29.52	93.75	**80.92**	**66.61**

Table 1: Comparison with state-of-the-arts on ActivityNet-Captions dataset ($n \in \{1, 5\}, m \in \{0.1, 0.3, 0.5\}$).

2020), MARN (Song et al. 2020), BAR (Wu et al. 2020) , VGN (Zhang et al. 2020c), VLANet (Ma et al. 2020), LoGAN (Tan et al. 2021), RTBPN (Zhang et al. 2020b). We also perform ablation study and qualitative analysis in this section.

Datasets

ActivityNet-Captions. The ActivityNet-Captions dataset is built on ActivityNet v1.3 dataset(Caba Heilbron et al. 2015), which is originally developed for human activity understanding. It consists of 19209 videos and is the largest dataset for temporal sentence grounding. Following the splitting in (Zhang et al. 2020a, 2021), we use val-1 as the validation set and val-2 as the testing set. In conclusion, we have 37417, 17505, and 17031 moment-sentence pairs for training, validation, and testing, respectively.

Charades-STA. The Charades-STA dataset is built upon Charades (Sigurdsson et al. 2016) by adding query sentence annotations, which are all about indoor activities. (Gao et al. 2017) proposes a semi-automatic method to generate moment-query annotations for fully supervised temporal sentence grounding. Following the splitting as (Lin et al. 2020), 12408 video-query pairs are used for training and 3720 pairs for testing in our experiments.

Evaluation Metric

Following the evaluation metric proposed in (Gao et al. 2017), we adopt the "R@n, IoU=m" metric to evaluate the performance for all methods. Specifically, it computes the percentage of testing samples that have at least one correct grounding prediction (i.e., the IoU between the prediction and the ground truth is larger than m) in the top-n predictions. Following previous methods(Lin et al. 2020), we set $n \in \{1, 5\}, m \in \{0.3, 0.5, 0.7\}$ for Charades-STA and $n \in \{1, 5\}, m \in \{0.1, 0.3, 0.5\}$ for ActivityNet-Captions.

Method	R@1 IoU=0.3	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.3	R@5 IoU=0.5	R@5 IoU=0.7
Fully supervised Methods						
ABLR	-	24.36	9.01	-	-	-
CBP	-	36.80	18.87	-	70.94	35.74
SCDM	-	54.44	33.43	-	74.43	58.08
2D-TAN	-	39.70	23.31	-	80.32	51.26
Weakly supervised Methods						
SCN	42.09	23.58	9.97	**95.56**	71.80	38.87
MARN	**48.55**	31.94	14.81	90.70	70.00	37.40
BAR	44.97	27.04	12.23	-	-	-
VGN	-	**33.21**	15.68	-	73.50	41.87
VLANet	45.24	31.83	14.17	95.70	**82.85**	33.09
LoGAN	51.76	**34.68**	14.54	92.74	74.30	39.11
RTBPN	**60.04**	32.36	13.24	**97.48**	71.85	41.18
Ours	43.31	31.02	**16.53**	95.54	77.53	**41.91**

Table 2: Comparison with state-of-the-arts on Charades-STA dataset ($n \in \{1, 5\}, m \in \{0.3, 0.5, 0.7\}$).

Implementation Details

Following (Lin et al. 2020), we utilize C3D (Tran et al. 2015) to extract visual features for the ActivityNet-Captions dataset and VGG (Simonyan and Zisserman 2014) for the Charades-STA dataset. N_C and N_V are set to 256 and 32, respectively, where the kernel size and stride of the 1D average pooling layer are both 8. The clip length r_1 and r_2 are 0.2 and 0.9 for the ActivityNet-Captions dataset while 0.2 and 0.5 for the Charades-STA dataset. The α factor is set as 0.25. The proposal threshold th_h and th_l are set as 0.7 and 0.5. Our model is trained on AdamW (Loshchilov and Hutter 2017) optimizer with the reduce-on-plateau learning rate decay strategy. The initial learning rate is set to 2e-4 and the training batch size is set to 64. The loss weights w_B, w_P, w_R, w_M are set as 0.4, 1.0, 1.0, 10.0.

Performance Comparisons

ActivityNet-Captions Dataset. We show the performance comparison on ActivityNet-Captions in Table 1. Our method almost outperforms all weakly supervised methods on "R@5", denoting a high recall for grounding. For "R@1", Our method obtained competitive performance compared to SCN and MARN.

Charades-STA Dataset. We further show the performance comparison on Charades-STA in Table 2. Our method outperforms other methods on "R@1, IoU=0.7" and "R@5, IoU=0.7", which means that our proposed method is able to predict more precise boundaries than other methods. For "R@5, IoU=0.1", there exists only a 0.02% gap between our method and the best result of SCN. We also achieve competitive performance compared to SCN and MARN for "R@1, IoU=0.3" and "R@1, IoU-0.5".

Ablation Study

In this section, we design the ablation experiments on ActivityNet-Captions, as shown in Table 3:

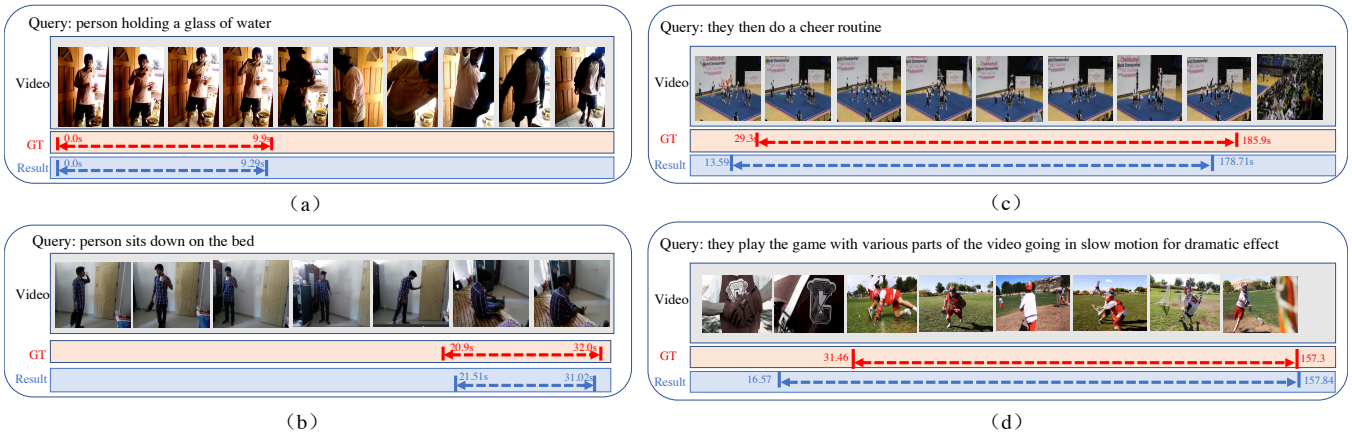

Figure 3: The visualization of examples. (a), (b) are from Charades-STA dataset and (c), (d) are from ActivityNet-Captions.

Methods	R@1, IoU=0.1	R@1, IoU=0.3	R@1, IoU=0.5
QueryA-Only	67.72	44.60	27.24
w/o Negative Query	68.42	44.62	28.76
w/o CRP	71.52	45.98	29.01
w/o MWP	69.12	44.31	27.89
Complete Model	71.86	46.62	29.52

Table 3: Ablation study on ActivityNet-Captions, where the first 'QueryA-Only' means the model is trained with a single query sentence rather than both. The following three denote the cases by removing the corresponding module only.

• **Multi-queries Evaluation** To evaluate the effect of the three queries (two relevant and one negative), we train the network on (1) QueryA only, (2) QueryA+QueryB, and (3) all the three queries. The results in Table 3 show that the model trained with the last configuration achieves better performance than the model trained with a single Query or without the negative query. We believe that training on two relevant queries can fully explore the composed video while the negative one can prevent the model from over-fitting the obvious boundary artifacts caused by composition.

• **Training w/o Ratio Prediction** To verify the effect of the composition ratio prediction, we remove the video-query alignment branch directly and keep other settings the same as the proposed method. The results in Table 3 show that it slightly hurts the performance when removing the ratio prediction module. However, learning holistic features by global reasoning tasks is meaningful for temporal grounding in those videos that do not hold obvious scene transitions. We leave this exploration in the future.

• **Training w/o Masked Word Prediction** To verify the effect of the masked word prediction, we feed the complete query sentences to the network and remove the masked word prediction module. Masked word prediction requires the linguistic features to consider their linguistic contexts and fully utilize the visual components to accurately predict the masked words, which help learn a compact and effective visual-linguistic representation. The results in Table 3 have supported this claim, where the performance drops by about 2% on all the metrics.

Qualitative Analysis

Further, we visualize several examples from Charades-STA and ActivityNet-Captions as shown in Figure 3. We can observe that there exists clear intra-contrast of (a) and (b) in Figure3 but also inter-contrast between (a) and (b). Since our method is designed for exploring such inter-contrast under a weakly-supervised setting, it can also localize the target segments well within the regular videos (a) and (b). In Figure3 (c), the intra-contrast is not so obvious, and our network also yields excellent predictions, which means that our method can also well tackle the videos with tiny intra-contrast and capture the fine-grained semantic information inside the videos. In Figure3 (d), the whole video presents large-scale scene transitions, and the text query contains complex descriptions. Our network, however, still performs well in most cases except for the beginning prediction.

Conclusion

This paper proposes a simple yet effective video composition strategy for weakly supervised temporal sentence grounding. Specifically, we compose a video by sampling different clips from videos where inter-contrast is always present. Then, we tackle the weakly supervised temporal sentence grounding in a fully supervised manner. By imposing multi-tasks including proposal classification, boundary refinement, clip ratio prediction, and masked word prediction, the network explores the inter-contrast between videos within a composed video when given different queries. The promising results on the Charades-STA and the ActivityNet-Caption demonstrate the effectiveness of our proposed method. We hope that this research work could provide insight for temporal sentence grounding and other research fields such as pre-training.

Acknowledgments

We gratefully acknowledge the support of the National Natural Science Foundation of China under Grants 61991411 and U1913204, the National Key Research and Development Plan of China under Grant 2018AAA0102504, the Natural Science Foundation of Shandong Province for Distinguished Young Scholars under Grant ZR2020JQ29, the Shandong Major Scientific and Technological Innovation Project 2019JZZY010428. This work is partially supported by Meituan.

Reference

Lin, Z.; Zhao, Z.; Zhang, Z.; Wang, Q.; and Liu, H. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11539–11546.

Mithun, N. C.; Paul, S.; and Roy-Chowdhury, A. K. 2019. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11592–11601.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Song, Y.; Wang, J.; Ma, L.; Yu, Z.; and Yu, J. 2020. Weakly-Supervised Multi-Level Attentional Reconstruction Network for Grounding Textual Queries in Videos. arXiv:2003.07048.

Gao, M.; Davis, L. S.; Socher, R.; and Xiong, C. 2019. Wslln: Weakly supervised natural language localization networks. *arXiv preprint arXiv:1909.00239*.

Tan, R.; Xu, H.; Saenko, K.; and Plummer, B. A. 2019. wman: Weakly-supervised moment alignment network for text-based video segment retrieval.

Tan, R.; Xu, H.; Saenko, K.; and Plummer, B. A. 2019. wman: Weakly-supervised moment alignment network for text-based video segment retrieval.

Liu, X.; Li, L.; Wang, S.; Zha, Z.-J.; Meng, D.; and Huang, Q. 2019. Adaptive reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2611–2620.

Ma, M.; Yoon, S.; Kim, J.; Lee, Y.; Kang, S.; and Yoo, C. D. 2020. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *European Conference on Computer Vision*, 156–171. Springer.

Wu, J.; Li, G.; Han, X.; and Lin, L. 2020. Reinforcement Learning for Weakly Supervised Temporal Grounding of Natural Language in Untrimmed Videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1283–1291.

Zhang, Z.; Zhao, Z.; Lin, Z.; He, X.; et al. 2020c. Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding. *Advances in Neural Information Processing Systems*, 33: 18123–18134.

Wang, B.; Ma, L.; Zhang, W.; Jiang, W.; Wang, J.; and Liu, W. 2019. Controllable video captioning with pos sequence guidance based on gated fusion network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2641–2650.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 961–970.

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

Yuan, Y.; Ma, L.; Wang, J.; Liu, W.; and Zhu, W. 2020. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020a. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12870–12877.

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

Arbelle, A.; Doveh, S.; Alfassy, A.; Shtok, J.; Lev, G.; Schwartz, E.; Kuehne, H.; Levi, H. B.; Sattigeri, P.; Panda, R.; et al. 2021. Detector-Free Weakly Supervised Grounding by Separation. *arXiv preprint arXiv:2104.09829*.

Zhang, M.; Yang, Y.; Chen, X.; Ji, Y.; Xu, X.; Li, J.; and Shen, H. T. 2021. Multi-Stage Aggregated Transformer Network for Temporal Language Localization in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12669–12678.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Zhang, W.; Wang, B.; Ma, L.; and Liu, W. 2019b. Reconstruct and represent video contents for captioning via reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 42(12): 3088–3101.

Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; and Schmid, C. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7464–7473.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yuan, Y.; Mei, T.; and Zhu, W. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Con-*

ference on Artificial Intelligence, volume 33, 9159–9166.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, 5267–5275.

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, 5803–5812.

Zhang, D.; Dai, X.; Wang, X.; Wang, Y.-F.; and Davis, L. S. 2019a. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1247–1257.

He, D.; Zhao, X.; Huang, J.; Li, F.; Liu, X.; and Wen, S. 2019. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8393–8400.

Wang, J.; Ma, L.; and Jiang, W. 2020. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12168–12175.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Zhang, Z.; Lin, Z.; Zhao, Z.; Zhu, J.; and He, X. 2020b. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4098–4106.

Tan, R.; Xu, H.; Saenko, K.; and Plummer, B. A. 2021. 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*, 2083–2092.

Wang, B.; Ma, L.; Zhang, W.; and Liu, W. 2018. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7622–7631.