DeCo : Decomposition and Reconstruction for Compositional Temporal Grounding via Coarse-to-Fine Contrastive Ranking

Lijin Yang[1*],Quan Kong[2], Hsuan-Kung Yang[2], Wadim Kehl[2], Yoichi Sato[1], Norimasa Kobori[2]
[1]The University of Tokyo, [2]Woven by Toyota
{yang-lj, ysato}@iis.u-tokyo.ac.jp
{quan.kong, hsuan-kung.yang, wadim.kehl, norimasa.kobori}@woven-planet.global

Abstract

Understanding dense action in videos is a fundamental challenge towards the generalization of vision models. Several works show that compositionality is key to achieving generalization by combining known primitive elements, especially for handling novel composited structures. Compositional temporal grounding is the task of localizing dense action by using known words combined in novel ways in the form of novel query sentences for the actual grounding. In recent works, composition is assumed to be learned from pairs of whole videos and language embeddings through large scale self-supervised pre-training. Alternatively, one can process the video and language into word-level primitive elements, and then only learn fine-grained semantic correspondences. Both approaches do not consider the granularity of the compositions, where different query granularity corresponds to different video segments. Therefore, a good compositional representation should be sensitive to different video and query granularity. We propose a method to learn a coarse-to-fine compositional representation by decomposing the original query sentence into different granular levels, and then learning the correct correspondences between the video and recombined queries through a contrastive ranking constraint. Additionally, we run temporal boundary prediction in a coarse-to-fine manner for precise grounding boundary detection. Experiments are performed on two datasets, Charades-CG and ActivityNet-CG, showing the superior compositional generalizability of our approach.

1. Introduction

Over the last years, robust results have been shown for the detection of predefined, simpler action classes in video [11, 38, 41, 48]. On the other hand, the detection of dense action, e.g. action contents described by a rich description, still poses a significant challenge due to the large diversity of possible language descriptions and semantic associations [16]. Recent work in this area treats complex actions as monolithic events via end-to-end predictions [27, 44]. They produce a single label to densely describe a long video sequence, and they are difficult to scale up to more varied patterns. Fundamentally, most complex actions consist of a series of simpler events, and thus a dense action can be treated as a composition of known event primitives [13, 25, 26]. Such a compositional representation can allow for a higher degree of model generalization. By learning from a finite number of action composites, and recombining their constituent event primitives in novel ways for unseen action descriptions, the representation can expand to large numbers of novel scenarios that have not been observed in the original action space. In this case, the action description is not treated as a single label but as a language modality that allows to learn finer-grained video and language correspondence.

The task of temporal video grounding concentrates on the described setting. Given a video and an action-related query sentence, the model has to output the start and end times of the specific moment that semantically corresponds to the given query sentence.

On top of temporal grounding, Compositional Temporal Grounding is a new task for testing whether the model can generalize to sentences that contain novel compositions of seen words. Several recent methods [12, 47, 50] encode both sentence and video segments into unstructured global representations and then devise specific cross-modal interaction modules to fuse them for final prediction. Unfortunately, such global representations fail to explicitly model video structure and language compositions, and fail for cases where higher granularity is required. Similar to the above approach of correspondence learning via global information, composite representations learn from paired monolithic video and language respectively, through large-scale self-supervised pre-training [32]. In either case, the main challenge of utilizing datasets of limited action size to achieve generalization towards novel actions remains challenging due to its combinatorial complexity.

Differently to the global representation approaches, VISA [17] introduces a compositional method toward the

* Work done while Lijin Yang was an Intern at Woven by Toyota.

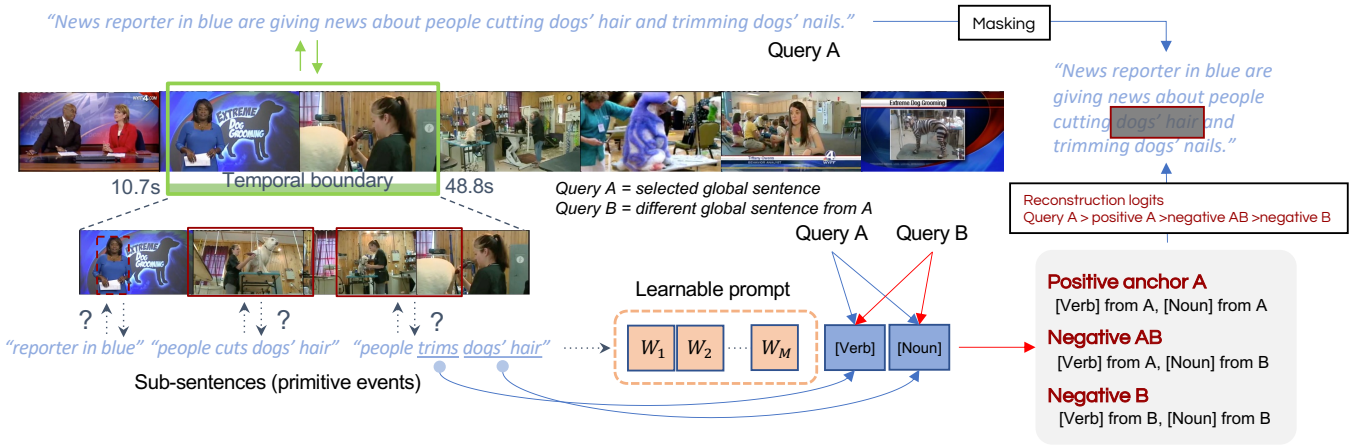

Figure 1. Our proposed method of decomposition and reconstruction of masked words with a coarse-to-fine scheme.

temporal grounding task. They parse video and language into several primitive elements, and then learn fine-grained semantic correspondences. Specifically, the composite representation is extracted from a graph that is structured with primitive elements, and these elements serve as a common representation for vision-language correspondence learning. However, both the global and compositional approaches disregard the granularity of the action composition during the learning phase, and can have issues when generalizing to more varied action spaces.

Based on these considerations, we argue that a good composite representation should be sensitive to different levels of granularity of both the action and the query language. We provide an overview of our idea in Fig. 1. Concretely, we propose to learn a composite representation in a coarse-to-fine manner, where we first decompose the whole query sentence into several simple phrases as subsentences, and then learn correspondences not only globally across query and video, but also between the subsentences and their related primitive sequences from the video. Since there is no ground truth to relate subsentences and their corresponding sequences, we propose to learn correspondence by decomposing and reconstructing them under weak supervision. For each subsentence as an anchor sample, we generate a temporal proposal and learn the positive representation through the mining of negative samples.

To structure the overall weak supervision, we mask the words in a given anchor query, and use the embedding of the positive and negative query with the related pseudo temporal segments from the video respectively to do the reconstruction for the masked words. The negative sample is another subsentence but includes the words from the other query action description, to allow the model sensitive to the word level variation for composition. According to the fact that the negative subsentence contains the novel word compared to the positive anchor sub-sentence, we could rank the reconstruction quality according to the given query sub-sentence as a natural prior constraint during the training.

Additionally, the ground truth that links the temporal boundary to the global sentence provides supervision for the coarse compositional representation. The generated subsentences from the same global query are recombined into a new sentence that should be as informative as the original global query, and then this recombined sentence is used with the original query sentence to estimate the temporal boundary via supervised learning. The whole process is trained end-to-end.

Our contributions are summarized below:

(i). We argue that a good compositional representation should be sensitive to action granularity of video and query language, and propose a coarse-to-fine decomposition approach to this end.

(ii). To learn a compositional representation without composite-level labels, we decompose events from global queries and learn primitive-level correspondences between video and language with a weakly-supervised contrastive ranking in form of word-masked reconstruction.

(iii). The decomposed events are recombined to a novel query along with the original query for temporal boundary estimation supervised, jointly learned end-to-end way with a coarse-to-fine structure.

(iv). Experiments on Charades-CG and ActivityNet-CG [17] demonstrate our method significantly outperforms existing methods about compositional temporal grounding tasks.

2. Related Work

Compositional Video Understanding. Some recent works explore compositional generalization for certain applications, including image captioning [28, 45, 46], image recognition [24, 39] visual question answering [5, 8], image synthesis [19], and zero-shot learning [14, 18, 34]. Also,

such an approach gained traction for video understanding in recent years due to the complexity of spatial-temporal information, and the difficulty of generalizing supervised models from monolithic event labels to novel concepts in unseen video. Hou et al. [10] propose a novel self-compositional learning framework to demonstrate the effectiveness of the proposed method for novel Human Object Interaction concept discovery. Action genome [13] provides a representation that decomposes actions into spatial-temporal scene graphs and learns the temporal changes in visual relationships that result in an action. They demonstrated increased performance over monolithic single-label-based supervised learning with a significant $+18$ mAP points difference. A similar prominent gain also has been shown in [25, 26, 31], demonstrating the high potential for generalizability by explicitly leveraging compositional structures for video representations. Motivated by the promising performance of compositional generalization on video understanding, we propose a novel compositional temporal grounding approach towards natural language sentences in videos to solve dense action understanding.

Temporal Grounding. Temporal grounding is a task first proposed by [6]. It focuses on localizing the target video for a given natural language sentence query. Existing methods first generate candidate video segments via sliding windows [6, 22], a proposal network [36, 37] or predefined anchors [21, 42, 42], and then semantically match each candidate with the sentence query. However, both proposal generation and semantic matching for all regarded proposals is computationally costly. To discard proposals and increase efficiency, proposal-free methods encode the video modality only once and directly model the interaction between each video frame and the sentence query [2, 3]. As a proposal-free variant, Hao et al. [9] introduce two auxiliary tasks, i.e. cross-modal matching and temporal order discrimination, to steer the training of the grounding model. However, none of the above methods pay attention to novel query sentences for grounding. To close the gap between learning from known sentences and dealing with unknown sentences as novel queries, Compositional Temporal Grounding is proposed by VISA [17] with two new benchmarks for novel compositional query tests named Charades-CG and ActivityNet-CG. In addition to these datasets, VISA proposed a variational cross-graph reasoning framework that explicitly decomposes video and language into multiple structured hierarchies and learns fine-grained semantic correspondence among them. However, their approach does not consider the granularity of the composition during the learning process, and focuses only on the fine-grained level of compositional correspondence learning. Moreover, during testing, the decomposed information of VISA comes from an external object detector that is run for the video. This provides model-external ontological knowledge as word-level information that might be novel to the given query. Differently, we propose a coarse-to-fine decomposition for varying granularity of primitive events generated from the given video and language without using any external knowledge to build compositional factors.

3. Method

Problem formulation and overview. We first present the problem formulation before going into details of our proposed method. Given a set of N videos $\{v_1, \cdots, v_N\}$ and their corresponding query sentences $\{q_1, \cdots, q_N\}$ that describe each video, our goal is to ground each sentence to a specific temporal segment in the video with start timestamps $\{st_1, \cdots, st_N\}$ and end timestamps $\{en_1, \cdots, en_N\}$.

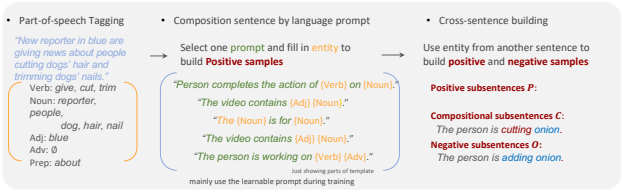

Figure 2. Decomposition process for a query sentence.

The overview of our method is shown in Figure 3. Given a video v and its corresponding sentence query q, we first decompose each sentence into k subsentences $P = \{p^1, \cdots, p^k\}$ and each subsentences is generated with 2 negative samples as $P_N = \{p^{k1}, \cdots, p^{kn}\}$, where $n \in \{1, 2\}$. To ensure that our decomposed sentences contain the same information with the original query, we further recompose the subsentences P as r as another input to the following modules. The given video v, the original sentence q and the decomposed sentences set $P \in \mathbb{R}^k$ and $P_N \in \mathbb{R}^{2k}$ are fed together into the temporal boundary grounding network, acquiring temporal proposals respectively as $T \in \mathbb{R}^{3k+2}$, that contains each query sentence-related proposal segment clip from video v (Sec. 3.1). The temporal proposal duration is represented as Gaussian weights [49] where frame features within each proposal will be aggregated based on the weight in the Gaussian curve. We then apply a mask-conditioned transformer [49] to process the proposed video segments T, sentences q, subsentences P and P_N, and Gaussian weights of each temporal proposals T to reconstruct the masked words in the original sentence q (Sec. 3.4). Finally, we use contrastive ranking as a constraint for encouraging the model to learn the coarse-to-fine correspondence between the global sentence q, subsentences P with its negative sampling set P_N, and given video v with its temporal segments T (Sec. 3.5).

3.1. Sentence decomposition

The key to achieving compositional generalization is to encourage the model to learn the correspondence between

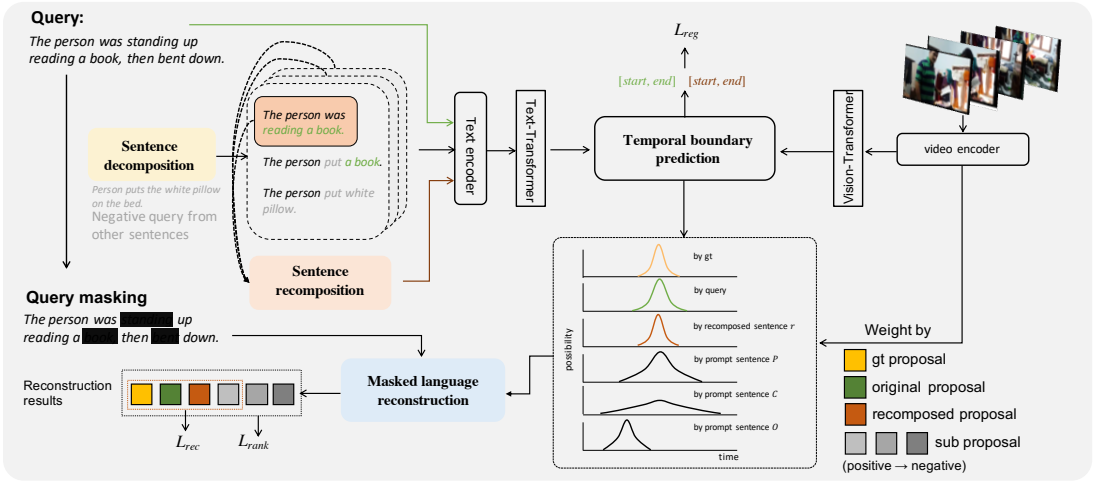

Figure 3. Illustration of our coarse-to-fine contrastive ranking. We decompose each query sentence into k subsentences P and construct k compositional subsentences C and k negative subsentences O. To ensure the decomposed subsentences are highly related to the original query, we recompose P into r. We predict the temporal boundaries in the video of all the $3k+1$ sentences. We then use the predicted boundary for masked language reconstruction and encourage the recomposed sentence r to be as informative as the original query sentence, while the compositional and negative sentences are less informative.

subsentences and primitive events. Since no such correspondence is given as ground truth, we need to learn them in a weakly-supervised manner. Figure 2 shows one example of this process.

The first step is to construct the subsentences by sentence decomposition. For a given query sentence q^i, we first conduct a part-of-speech tagging [23] on the sentence q^i, getting the tags (e.g, verb, noun, etc.) of each word. Then we choose a template from one of the five composition templates (Verb, Noun), (Adjective, Noun), (Preposition, Noun), (Noun, Noun), (Verb, Adverb) and fill the template with randomly selected words of the corresponding tag TAG, forming compositional word tuples $S = \{s^1, \cdots, s^n\}$. These composition word tuples are then used to form the subsentences $P_i = \{p_i^1, \cdots, p_i^k\}$. Inspired by [51], we use a learnable prompt token $[w]_m$ ($m \in \{1, \cdots, M\}$) and append it in front of each compositional word tuple to form the subsentences. Specifically, the prompt t given to the text transformer is designed in the following form:

$$t = [w]_1[w]_2 \cdots [w]_M[TAG]_1[TAG]_2, \qquad (1)$$

where each $[w]_m$ is a vector with the same dimension as word embeddings for each tagged word. To allow fair comparison with existing work, we also use GloVe [30] as our word embedding. M is a hyperparameter specifying the number of prompt tokens, $[w]_1[w]_2 \cdots [w]_M$ are shared among all word tuples.

To allow the model to learn the temporal grounding from word-level variance such that novel composition can be better performed, we further generate subsentence compositions of word tuples that do not exist in the training set by mixing words in s_i with s_j, where s_i is a word tuple decomposed from q^i, and s_j is decomposed from another query sentence q^j. We choose a word from s_j, and swap this word with a word in s_i with the same tag , to generate novel subsentences $C = \{c^1, \cdots, c^k\}$. C provides each subsentence with composition word tuples, such that the words are partially from P_i and partially from P_j as one kind of negative sample.

We also generate another subsentence set $O = \{o^1, \cdots, o^k\}$ where composition word tuples only come from another query sentence q^j and are completely unrelated to S. This set will be used to get our second negative sample. All of these subsentences are used in the latter steps for learning correspondences between subsentences and primitive events.

3.2. Recomposition

To make sure our decomposed subsentences are highly related to the original query, the sentence recombined from the subsentences should be as informative as the original query for the temporal grounding. Following this thinking, we generate a sentence r by recomposing the subsentences $\{p^1, \cdots, p^k\} \in P$ from query q. This recomposition is done by a transformer decoder, which takes a set of N learnable tokens as its query vector and the concatenation of $\{p^1, \cdots, p^k\}$ as the key and value vectors. By this means, we get a recomposed sentence r with length N.

3.3. Temporal Proposal/Boundary Prediction

We use a transformer-based model to generate a temporal proposal in T for each input sentence. To be specific, for sentence query q, we use GloVe [30] following a single-layer MLP to extract word embeddings and to build a prompt embedding $t \in \mathbb{R}^{(m+2)\times D}$, where D is the dimension of

the embedding. The prompt embedding t is then forwarded to a text transformer as a feature projector for acquiring self-attention sentence features for temporal boundary estimation. For video v, we use a pre-trained C3D [33]/I3D [1] model for extracting the D dimension frame feature f_i of i-th frame within a segment size L, to construct a video feature $F = \{f_i\} \in \mathbb{R}^{L \times D}_{i \in L}$. Then, we append a prediction token $< pred >$ with dimension D by random initialization at the end of the video feature F following [50], forming $F_T = [F, < pred >]$. F_T is used as the video features fed to a vision transformer as a video feature projector. Finally, the projected text and video features are fed into a cross-modal attention module to estimate the temporal boundary of the given query sentence. In the cross-modal attention module, $Query$ and $Value$ is $t \in \mathbb{R}^{(m+2) \times D}$, Key is the transpose of $F_T \in \mathbb{R}^{L \times D}$, and the resulting feature F_v is used for temporal proposals prediction. We predict the Gaussian proposal duration center $u^i \in \mathbb{R}^{3k+2}$ and width $d^i \in \mathbb{R}^{3k+2}$ for each query sentence/subsentence through a fully-connected layer activated by a $Sigmoid$ function. Based on the estimated center and width, we can further acquire the start st and end en timestamps as below:

$$\begin{aligned} st &= \max((u^i - d^i/2), 0) * Duration \\ en &= \min((u^i - d^i/2), 1) * Duration \end{aligned} \tag{2}$$

where $Duration$ is the video length. The estimated temporal boundary timestamp corresponding with the original query q and the recomposed sentence r are minimized with a $L2$ regression loss $\mathcal{L}_{reg}$.

3.4. Mask Reconstruction

The reconstruction of masked words is helpful for analyzing whether the compositional understanding at the fine-grained level is good or not. Thus, we randomly replace 30% of the words in the original query q with a $< mask >$ token and let the mask-conditioned transformer [49] predict the next word given a prefix of the query and visual features inside the proposal. The mask-conditioned transformer will only use the video features inside each proposal by multiplying the proposal mask with the attention map before aggregating contextual information. We then use a cross-entropy loss to measure the reconstruction quality of each proposal using the mask-conditioned reconstruction completion module to reconstruct the original sentence q.

The mask-conditioned transformer uses cross attention, and so *Key* and *Value* are the frame-wise features F_v within each estimated temporal proposal to be aggregated based on the weight in the Gaussian curve. *Query* is the text embedding acquired from the given query sentence/subsentence as t. The cross-attention feature is then fed to a FC to project the feature to the word number space for calculating the cross entropy for the masked word. We denote the cross-entropy loss of the ground-truth boundary, original query proposals, compositional subsentence proposals, cross-sentence subsentence proposals, and the irrelevant subsentence proposals as $\mathcal{L}_{gt}, \mathcal{L}_q, \mathcal{L}_p, \mathcal{L}_c$, and $\mathcal{L}_o$, respectively, and only use $\mathcal{L}_{gt}$, $\mathcal{L}_q$ and $\mathcal{L}_p$ to learn the reconstruction.

In this way we can measure the semantic relevance between the proposal and the query, as we assume that the most-relevant proposal can best reconstruct q using only the visual features within the proposal.

3.5. Contrastive Ranking

Since we have no ground truth labels for the subsentences, we design a weakly-supervised method based on contrastive ranking to learn the similarity between subsentences and videos. Intuitively, the original query q and the recomposed sentence r should contain all information that describes a part of the video v, and the compositional subsentence contains relevant parts of that information as well, thus the similarity between q, v and between r, v should be larger than that of p^k, v. In the same fashion, the similarity of p^k, v should be larger than the similarity of c^k, v, and all the previous similarities should be larger than o, v since o is irrelevant to the video v. Based on this idea, we design a contrastive ranking loss to model this similarity difference. We use cross entropy logits as an estimate of semantic similarity between a sentence and its corresponding temporal proposal, so that this ranking loss can be expressed by:

$$\begin{aligned} \mathcal{L}_{rank} = &\max(\mathcal{L}_{gt} - \mathcal{L}_q + h_0, 0) \\ &+ \max(\mathcal{L}_q - \mathcal{L}_p + h_1, 0) \\ &+ \max(\mathcal{L}_p - \mathcal{L}_c + h_2, 0) \\ &+ \max(\mathcal{L}_c - \mathcal{L}_o + h_3, 0) \end{aligned} \tag{3}$$

where L is the cross entropy logits respectively and h is the hyperparameter to control the threshold of constrastive loss.

3.6. Model training and inference

Our model is jointly optimized by three loss functions: the regression loss $\mathcal{L}_{reg}$ is a strong supervision signal that directly supervises the proposal output. The reconstruction loss $\mathcal{L}_{rec}$ is used to help the model learn how to reconstruct the query sentence using video features within each proposal, serving as an estimation of the semantic similarity between a proposal and a sentence. Lastly, the ranking loss $\mathcal{L}_{rank}$ uses the relative reconstruction quality to train the model to generate the most semantically relevant positive proposals for each subsentence. The full loss function can be represented as:

$$\mathcal{L} = \lambda_{reg}\mathcal{L}_{reg} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{rank}\mathcal{L}_{rank}, \tag{4}$$

where $\lambda_{reg}, \lambda_{rec}, \lambda_{rank}$ are hyperparameters for loss balancing. As for model inference, we take the prediction of the recomposed sentence r as the final output.

Method	Setting	Test-Trivial			Novel composition		
		R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
WSSL [4]	WS	15.33	5.46	18.31	3.61	1.21	8.26
CPL [50]	WS	47.28	21.85	42.27	39.11	15.60	35.53
TMN [20]	FL	18.75	8.16	19.82	8.68	4.07	10.14
CTRL [7]	FL	18.53	8.59	22.03	4.62	0.17	11.21
TSP-PRL [35]	FL	39.86	21.07	38.41	16.30	2.04	13.52
VSLNet [43]	FL	45.91	19.80	41.63	24.25	11.54	31.43
SCDM [40]	FL	46.63	24.17	42.08	27.73	12.25	30.84
2D-TAN [44]	FL	48.58	26.49	44.27	30.91	12.23	29.75
LGI [27]	FL	49.45	23.80	45.01	29.42	12.73	30.09
DeCo (*Ours*)	FL	**58.75**	**28.71**	**49.06**	**47.39**	**21.06**	**40.70**
VISA [17]	FL	53.20	26.52	47.11	45.41	22.71	42.03

Table 1. IoU@{0.5, 0.7} and mIoU results on the Charades-CG *Test-trivial* and *Novel-composition* set. The bold numbers represent the top-1 result. The dark row indicates external detector knowledge. WS: Weakly-supervised learning, FL: Fully-supervised learning.

Method	Setting	Test-Trivial			Novel composition		
		R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
WSSL [4]	WS	11.03	4.14	15.07	2.89	0.76	7.65
CPL [50]	WS	26.53	11.31	32.06	19.49	7.00	26.95
TMN [20]	FL	16.82	7.01	17.13	8.74	4.39	10.08
CTRL [7]	FL	13.25	4.49	17.51	5.22	1.55	11.21
TSP-PRL [35]	FL	34.27	18.80	37.05	14.74	1.43	12.61
SCDM [40]	FL	37.86	22.41	40.09	21.32	9.34	28.52
LGI [27]	FL	43.56	23.29	41.37	23.21	9.02	27.86
2D-TAN [44]	FL	44.50	26.03	42.12	22.80	9.95	28.49
VSLNet [43]	FL	39.27	23.12	42.51	20.21	9.18	29.07
DeCo (*Ours*)	FL	43.98	24.25	43.47	27.35	11.66	31.27
LGI [27] + DeCo (*Ours*)	FL	**47.38**	**28.43**	**46.03**	**28.69**	**12.98**	**32.67**
VISA [17]	FL	47.13	29.64	44.02	31.51	16.73	35.85

Table 2. IoU@{0.5, 0.7} and mIoU results on the ActivityNet-CG *Test-trivial* and *Novel-composition* set. The bold numbers represent the top-1 result. The dark row indicates external detector knowledge. WS: Weakly-supervised learning, FL: Fully-supervised learning.

4. Experiments

4.1. Implementation Detail

We use PyTorch [29] for our implementation. For a fair comparison, we follow [27] to use C3D [33] features for the ActivityNet-CG dataset and the I3D [1] feature for the Charades-CG dataset. In more detail, the features are extracted by first downsampling each video by a factor of 8, and we follow [50] by setting the maximum video feature length as 200. In all experiments, we use the Adam optimizer [15] with an initial learning rate of 4e-4 for 30 epochs. For loss balancing we empirically set the factors to $\lambda_{reg} = 10, \lambda_{rec} = 1$, and $\lambda_{rank} = 1$. Please refer to the supplementary material for more implementation details.

4.2. Datasets

To evaluate our method, we conduct experiments on the newly proposed Charades-CG and ActivityNet-CG datasets [17]. Each of these two datasets contains a new testing split of *Novel-Composition* of the original Charades-STA [7] and ActivityNet Captions [16] datasets. In the novel composition test split, each sentence contains one of the five types of novel compositions, *i.e.*, the constituents are both observed during training but have never appeared in the same sentence. We also test on the *Test-trivial* split proposed in [17] for a complete comparison. In the Charades-CG dataset, the number of video-sentence pairs in the *Training* / *Novel-Composition* / *Test-Trivial* sets is 8281 / 3442 / 3096. The ActivityNet-CG dataset is much larger, with 36724 / 12028 / 15712 video-sentence pairs in the *Training* / *Novel-Composition* / *Test-Trivial* sets.

Configurations	Test-Trivial			Novel-Composition		
	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
(a) q	50.32	22.67	43.92	40.41	15.66	36.48
(b) $gt > q$	53.62	25.29	45.46	41.40	17.75	36.62
(c) $gt > q, p$	52.03	25.16	45.08	40.94	17.49	36.92
(d) $gt > q > p$	54.20	27.39	47.05	43.64	18.54	38.29
(e) $gt > q > p > c$	56.59	26.32	47.65	44.25	19.52	39.20
(f) $gt > q > p > o$	55.94	27.62	47.82	43.99	19.20	38.65
(g) $gt > q > p > c, o$	57.17	28.36	48.21	45.26	19.58	39.46
(h) $gt > q > p > c > o$	57.66	28.68	48.47	46.05	20.40	40.14
(h) + *recomposition*	58.75	28.71	49.06	47.39	21.06	40.70

Table 3. Ablation study for contrastive ranking on Charades-CG dataset with both *Test-Trivial* and *Novel-Composition* setup. q: original query, p: positive sub-sentence, c: negative sub-sentence (part of the words comes from another sentence), o: negative sub-sentence (words all come from another sentence)).

Setting	R1@0.5	R1@0.7	mIoU
fixed template	45.61	20.16	39.68
$M = 0$	43.35	18.30	38.57
$M = 2$	46.05	20.40	40.14
$M = 4$	45.99	20.69	39.95
$M = 8$	45.64	19.55	39.59
$M = 2$, $N = 1$	45.26	18.74	39.52
$M = 2$, $N = 2$	46.72	18.65	40.32
$M = 2$, $N = 4$	47.39	21.06	40.70
$M = 2$, $N = 8$	44.33	19.23	39.34

Table 4. Ablation study on the prompt engineering.

4.3. Quantitative Results

The quantitative results are shown in Tables 1 and 2, respectively. We compare our proposed approach with other temporal grounding approaches. On the Charades-CG dataset, it can be observed that the proposed approach outperforms all the baselines in terms of R1@0.5 and mIoU scores in both *Test-Trivial* and *Novel-Composition* setups. We believe these performance improvements come from the usage of contrastive ranking constraints along with the composition sentence-building scheme, which helps to understand the correspondence of the video and query sentence from the perspective of granularity.

On the other hand, when testing on the ActivityNet-CG dataset, purely adopting the proposed approach outperforms all the baselines without the compositional design, which again proves the importance of compositional logic for the visual grounding tasks.

However, the proposed approach fails to surpass VISA [17] in the ActivityNet-CG dataset (Table 2). Our hypothesis is two-fold: (a) VISA adopts external object detector and action detector to obtain prior instance-level knowledge, which provides post-decomposition information thus helps it to understand individual components of seen compositions during training. (b) The proposed approach highly relies on the quality of the encoded visual features, thus it is prone to fail when the visual input is complicated and novel and unseen samples exist. Please refer to the supplementary material for further analysis.

Based on the above observation, we further propose a variant version of our approach, in which we utilize the LGI-based backbone [27] for the visual feature extraction. The result is shown in Table 2 and denoted as *LGI + DeCo (Ours)*. By introducing a better feature extractor, the performance significantly improves and achieves comparable results compared to VISA [17].

4.4. Ablation Study

4.4.1 Contrastive Ranking

In this subsection, we examine the effectiveness of the proposed contrastive ranking approach by considering different ranking configurations, and the results are shown in Table. 3. The observations and the insights from it are discussed in the following paragraphs:

The performance boost from contrastive ranking. In this ablation study, we incrementally include more contrastive ranking constraints and compare the performance accordingly. The results can be referred from the rows $(a), (b), (d), (e), (h)$ in Table 3. It can be observed that the accuracy increases in both *Test-Trivial* and *Novel-Composition* setups, as we add more ranking constraints in the training phase, thus proving the effectiveness of the proposed concept.

Contrastive ranking v.s. sample and augmentation. To further dive into the rationale behind the performance boost coming from contrastive ranking, we conducted another set of experiments. Here, we consider the number of components (i.e., the recomposed sentences) to be used for training with and without the contrastive ranking. The results are shown in the set of rows $(b,\ c,\ d)$ in Table 3. It can be seen that by simply introducing the composition subsentences without contrastive ranking we cannot guarantee better performance (i.e., row (c) has worse performance than row (b)). On the other hand, introducing the composition subsentences along with the concept of contrastive ranking increase the performance (i.e., row (d) outperforms row (c)). This hints towards our proposed concept of having advantages for visual temporal grounding.

4.4.2 Prompt Engineering

Hand-craft prompt template v.s. learnable context prompt. In our method, we use learnable prompt tokens to

formulate the subsentences from compositional word tuples. Here we test their effectiveness by comparing the performance between using learnable prompt tokens and using hand-crafted prompt templates. The result is shown in Table 4, where the first row indicate the performance using a fixed prompt template, while the other rows show the results of different number of learnable prompt tokens. In the upper block, the tokens are used for both sentence decomposition (Sec. 3.1) and recomposition (Sec. 3.2), while in the lower block different learnable tokens are used for decomposition and recomposition. As can be seen from the Table, fixing a hand-crafted prompt template performs comparably well against using one type of template only (upper block of Table 4), but does not perform as well as using two types of learnable prompts (lower block of Table 4).

Ablation on the number of M and N. Following [51], we find that learnable prompt context can work better than hand-craft prompt template while eliminating the need of detailed template design. We further test the performance of varying-length learnable context prompts by changing M. As can be seen in Table 4, the model can achieve best performance when $M = 2$, and the performance decreases when the number of M increases. We argue that $M = 2$ is enough for learning the context of the two-words composition. After fixing M, we also test the length of N in learnable token used during recomposition and set $N = 4$ for both ActivityNet-CG and Charades-CG datasets.

4.5. Qualitative Results

Effectiveness of contrastive ranking. We visualize two qualitative examples of grounding results by our model without and with proposed contrastive ranking loss in Figure 4. Since the composition "turns on a light" is novel and has never been seen in the training set, the model without contrastive ranking loss lacks composition generalizability and the predicted temporal boundary is far from the ground-truth results (example (a)). For example (b) we can see that "puts some cloth" is manageable by the model without contrastive ranking can also give general grounding results, as it is closer to the composition available in the training set (such as "put cup"), but fails to generate accurate start/end predictions. However, by taking advantage of multi-granularity decomposition as well as contrastive ranking, the model with contrastive loss can better understand the primitive knowledge of words "put" and "clothes", and thus predict the temporal boundary accurately.

Effectiveness of sentence recomposition. To qualitatively show the effectiveness of the subsentence recomposition, in Figure 5 we show the temporal boundary prediction results generated by the original query sentence (light blue) and our recomposed sentence r (orange). As shown in Figure 5 (a) and (b), the temporal boundary predicted by the recomposed sentence can achieve better performance than that with the original query sentence as input. From these

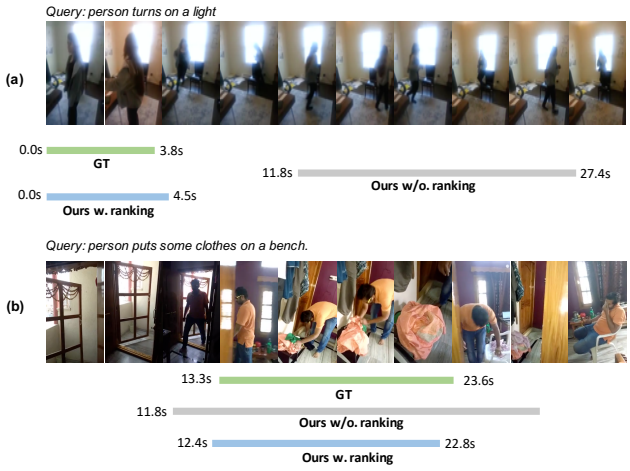

Figure 4. Qualitative examples of the ground truth (GT), Ours (without contrastive ranking), and Ours (with contrastive ranking). Examples are from the Charades-CG dataset.

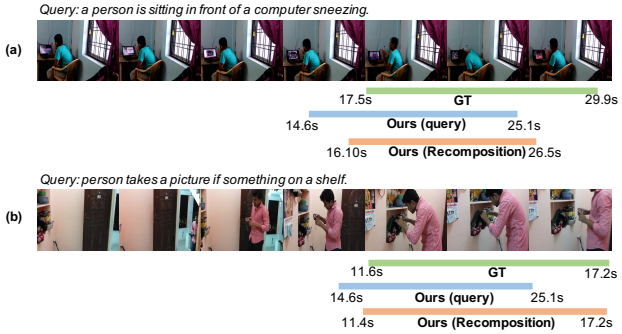

Figure 5. Qualitative examples of the ground truth (GT), Ours (with the query as input), and Ours (with the recomposed sentence r as input). Examples are from the Charades-CG dataset.

cases, we can clearly see that the recomposed sentence can summarize the knowledge in all of the subsentences, which is as informative as the original query sentence to accurately ground the boundary in the video. Also, the decomposition-recomposition operation can help the model to better understand the contribution of each subsentence, and thus generate even more accurate results than using the original query sentence.

5. Conclusions

We propose DeCo, a novel method for compositional temporal grounding. Our method tries to learn a good composite representation sensitive to different granularity levels of both video and language, by decomposing and recomposing query sentences while learning the vision-language correspondence via coarse-to-fine contrastive ranking. Experiments show the outstanding performance of our method.

Acknowledgement This work was supported by JST SPRING, Grant Number JPMJSP2108.

References

[1] Joao Carreira and Andrew 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] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. Learning modality interaction for temporal sentence localization and event captioning in videos. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 333–351. Springer, 2020.

[3] Shaoxiang Chen and Yu-Gang Jiang. Hierarchical visual-textual graph for temporal activity localization via language. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XX*, volume 12365 of *Lecture Notes in Computer Science*, pages 601–618. Springer, 2020.

[4] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. *Advances in Neural Information Processing Systems*, 31, 2018.

[5] Mona Gandhi, Mustafa Omer Gul, Eva Prakash, Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Measuring compositional consistency for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5046–5055, June 2022.

[6] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: temporal activity localization via language query. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5277–5285. IEEE Computer Society, 2017.

[7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[8] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA: A benchmark for compositional spatio-temporal reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11287–11297. Computer Vision Foundation / IEEE, 2021.

[9] Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pages 130–147. Springer, 2022.

[10] Zhi Hou, Baosheng Yu, and Dacheng Tao. Discovering human-object interaction concepts via self-compositional learning. In *ECCV*, 2022.

[11] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14024–14034, 2020.

[12] Yifei Huang, Lijin Yang, and Yoichi Sato. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[13] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10233–10244. Computer Vision Foundation / IEEE, 2020.

[14] Michael C. Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P. Xing. Rethinking knowledge graph propagation for zero-shot learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11479–11488, 2019.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[17] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3032–3041, 2022.

[18] Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, and Muli Yang. Siamese contrastive embedding network for compositional zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9316–9325. IEEE, 2022.

[19] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18197–18207, June 2022.

[20] Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan 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.

[21] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11235–11244. Computer Vision Foundation / IEEE, 2021.

[22] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei, editors, *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 843–851. ACM, 2018.

[23] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[24] M. Mancini, M. F. Naeem, Y. Xian, and Z. Akata. Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5222–5230. Computer Vision Foundation / IEEE, June 2021.

[25] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1046–1056. Computer Vision Foundation / IEEE, 2020.

[26] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[27] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020.

[28] Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. Compositional generalization in image captioning. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 87–98, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[30] Jeffrey Pennington, Richard Socher, and Christopher D 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.

[31] Gorjan Radevski, Marie-Francine Moens, and Tinne Tuytelaars. Revisiting spatio-temporal layouts for compositional action recognition. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 110. BMVA Press, 2021.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[33] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[34] X. Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6857–6866, 2018.

[35] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12386–12393, 2020.

[36] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2986–2994. AAAI Press, 2021.

[37] Huijuan Xu, Kun He, Bryan A. Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9062–9069. AAAI Press, 2019.

[38] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14722–14732, 2022.

[39] Muli Yang, Yuehua Zhu, Jiaping Yu, Aming Wu, and Cheng Deng. Divide and conquer: Compositional experts for generalized novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14268–14277, June 2022.

[40] Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Dataset and metric. In *Proceedings of the 2nd International Workshop on Human-centric Multimedia Analysis*, pages 13–21, 2021.

[41] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7094–7103, 2019.

[42] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S. Davis. MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1247–1257. Computer Vision Foundation / IEEE, 2019.

[43] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020.

[44] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877, 2020.

[45] Wenqiao Zhang, Haochen Shi, Jiannan Guo, Shengyu Zhang, Qingpeng Cai, Juncheng Li, Sihui Luo, and Yueting Zhuang. MAGIC: multimodal relational graph adversarial inference for diverse and unpaired text-based image captioning. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 3335–3343. AAAI Press, 2022.

[46] Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. Consensus graph representation learning for better grounded image captioning. In *Proceedings of the Thirty-Fifth Conference on Association for the Advancement of Artificial Intelligence (AAAI)*, pages 3394–3402, 2021.

[47] Yanyi Zhang, Xinyu Li, and Ivan Marsic. Multi-label activity recognition using activity-specific features and activity correlations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14625–14635. Computer Vision Foundation / IEEE, 2021.

[48] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017.

[49] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, page 3, 2022.

[50] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15555–15564, 2022.

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021.