Compositional Temporal Grounding with Structured Variational Cross-Graph Correspondence Learning

Juncheng Li [1] Junlin Xie [1] Long Qian [1] Linchao Zhu [2] Siliang Tang [1] Fei Wu [1]
Yi Yang [1] Yueting Zhuang [1*] Xin Eric Wang [3]

[1] Zhejiang University, [2] University of Technology Sydney, [3] University of California, Santa Cruz
{junchengli, 22051289, qianlong0926, siliang, wufei, yangyics, yzhuang}@zju.edu.cn
linchao.zhu@uts.edu.au, xwang366@ucsc.edu

Abstract

*Temporal grounding in videos aims to localize one target video segment that semantically corresponds to a given query sentence. Thanks to the semantic diversity of natural language descriptions, temporal grounding allows activity grounding beyond pre-defined classes and has received increasing attention in recent years. The semantic diversity is rooted in the principle of compositionality in linguistics, where novel semantics can be systematically described by combining known words in novel ways (**compositional generalization**). However, current temporal grounding datasets do not specifically test for the compositional generalizability. To systematically measure the compositional generalizability of temporal grounding models, we introduce a new Compositional Temporal Grounding task and construct two new dataset splits,* i.e.*, Charades-CG and ActivityNet-CG. Evaluating the state-of-the-art methods on our new dataset splits, we empirically find that they fail to generalize to queries with novel combinations of seen words. To tackle this challenge, we propose a variational cross-graph reasoning framework that explicitly decomposes video and language into multiple structured hierarchies and learns fine-grained semantic correspondence among them. Experiments illustrate the superior compositional generalizability of our approach. The repository of this work is at* https://github.com/YYJMJC/Compositional-Temporal-Grounding.

1. Introduction

Understanding rich and diverse activities in videos is a prominent and fundamental goal of video understanding. While there have been significant works in activity recognition [3, 8] and localization [28, 38], one major limitation of these works is that they are restricted to pre-defined action classes, thus suffering from scaling to various complex activities. A natural solution to this problem is to utilize the systematic compositionality [4, 9, 31] of human language, which allows us to form novel compositions by combining known words in novel ways to describe unseen activities (*i.e.* **compositional generalization**). Therefore, a new task, namely temporal grounding in videos [10, 16], has recently received increasing attention. Formally, give an untrimmed video and a query sentence, it aims to identify the start and end timestamps of one specific moment that semantically corresponds to the given query sentence.

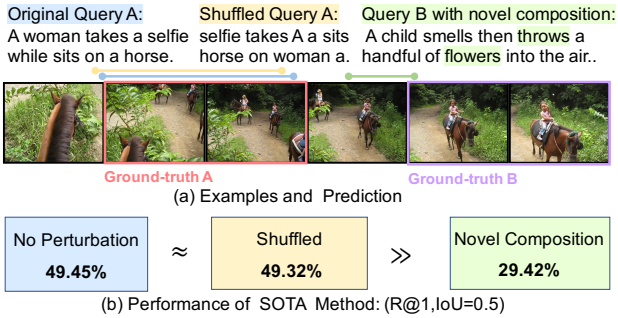

Figure 1. (a) On the top, we show three examples of two queries. (b) On the bottom, we report comparisons on Charades-CG with metric R@1, IoU@0.5. The left blue box represents the original model. The middle yellow box represents the model with shuffled queries as input. The right green box represents the performance on the queries that contain novel compositions.

Although the compositional generalization is a key property of human language that allows temporal grounding beyond pre-defined classes, current temporal grounding datasets do not specifically test for this ability. The training and testing splits of existing datasets contain almost the same compositions (*e.g. verb-noun pair, adjective-noun pair, etc*). Our statistical results show that only 1.37% and 5.19% of testing sentences contain novel compositions in the Charades-STA [10] and ActivityNet Captions [16]

*Yueting Zhuang is the corresponding author.

datasets, respectively. To systematically measure the compositional generalizability (CG) of existing methods, we introduce a new task, **Compositional Temporal Grounding**. Our compositional temporal grounding task aims to test whether the model can generalize to the sentences that contain novel compositions of seen words. We construct two re-organized datasets **Charades-CG** and **ActivityNet-CG**. Our dataset split protocols enable us to measure whether a model can generalize to novel compositions, of which the individual components have been observed during training but the combination is novel.

Using our newly constructed datasets, we evaluate modern state-of-the-art (SOTA) temporal grounding models, and empirically find that SOTA models fail to achieve compositional generalization, though they have achieved promising progress on the typical temporal grounding task. We observe that their performance drops dramatically (Figure 1.b, left vs. right). The results indicate that the SOTA models may not well generalize to novel compositions. Furthermore, as word order is a crucial factor for the compositionality of language, we analyze the word order sensitivity of SOTA models to gain more intuitive insight. Specifically, we randomly shuffled queries in advance and then use the shuffled sentences to train and evaluate the models. Surprisingly, we find that they are insensitive to the word order, even though permuting word order destroys the complete semantics of original sentences (Figure 1.b, left vs. middle). These observations confirm with recent studies [35,42] suggesting that current models are heavily driven by superficial correlations. This pushes us to rethink the solution of temporal grounding.

When we systematically analyze the SOTA models, we find that previous temporal grounding methods largely neglect the structured semantics in video and language, which is crucial for compositional reasoning. These methods [10, 32,46,48] mainly encode both sentence and video segments into unstructured global representations, respectively, and then devise specific cross-modal interaction modules to fuse them for final prediction. These global representations fail to explicitly model video structure and language compositions. Take the novel composition of "throws flowers" in Figure 1.a as example. If the model infers the individual semantics of the two words, as well as establish the correspondence of them to specific semantics in video (*i.e. the action "throw" and the object "flower" in video*), the model can easily localize the novel composition in video by composing the corresponding video semantics of the two words.

Motivated by this insight, we propose a novel **V**ar**I**ational cro**S**s-graph re**A**soning (**VISA**) framework for compositional temporal grounding. By explicitly modeling the semantic structures of video and language, and inferring the fine-grained correspondence between them, our VISA model can achieve joint compositional reasoning. Specifically, we first introduce a hierarchical semantic graph that explicitly decomposes both video and language into three semantic hierarchies (*i.e. global events, local actions, and atomic objects*). The hierarchical semantic graph serves as unified structured representations for both video and language, which tightly couple multi-granularity semantics between the two modalities. Second, we propose a variational cross-graph correspondence learning that establishes fine-grained semantic correspondence between the semantic hierarchical graphs of video and language.

Our contributions are summarized as follows:

- We introduce a new task, Compositional Temporal Grounding, as well as new splits of two prevailing temporal grounding datasets, which are able to measure the compositional generalizability of existing methods.
- We perform in-depth analyses on several SOTA models and empirically find that they fail to achieve compositional generalization
- We propose a **V**ar**I**ational cro**S**s-graph re**A**soning (**VISA**) framework that decomposes video and language into hierarchical graphs and learns fine-grained cross-graph correspondence between them.
- Experimental results validate the significant superiority of our approach on compositional generalizability.

2. Related Work

Temporal Grounding. Recently, the development of deep learning [20,25] promotes the prosperity of computer vision [**?**, 24, 54] and vision-and-language understanding [21, 23, 47,49,50,53]. Temporal grounding in videos via language is a recently proposed task [10,16]. Existing supervised methods can be categorized into two groups. 1) Proposal-based methods [10,43,45,48] first extract candidate proposals by temporal sliding windows and then match the query sentence with them by multi-modality fusion. 2) Proposal-free methods [26,32,44,46] directly predict the temporal boundaries of target segments without pre-defining proposals. In this paper, we evaluate the compositional generalizability of current methods.

Compositional Generalization. Recently, compositional generalization has received increasing attention as its advantages on robustness and sample efficiency. To evaluate the compositional generalization, Lake *et al.* [17] propose the SCAN benchmark, which requires translating instructions generated by a phrase-structure grammar to action sequences. The SCAN is split such that the testing set contains unseen compositions in the training set. The following works have proposed several techniques to improve SCAN, including data augmentation [1], meta-learning [6, 18, 34], and architectural design [5, 12]. Some recent works also explore compositional generalization on other applications,

including image captioning [33, 51, 52], visual question answering [13, 14], action recognition [30, 40, 55], and state-object recognition [29]. In this paper, we systematically study the compositional generalization on temporal grounding natural language sentences in videos.

3. Compositional Temporal Grounding

3.1. Problem Formulation

To systematically benchmark the progress of current methods on compositional generalization, we introduce a new task, **Compositional Temporal Grounding**. Our compositional temporal grounding task aims to evaluate how well a model can generalize to query sentences that contain novel compositions or novel words. We construct new splits of two prevailing datasets Charades-STA [10] and ActivityNet Captions [16], named **Charades-CG** and **ActivityNet-CG**, respectively. Specifically, we define two new testing splits: Novel-Composition and Novel-Word. Each sentence in the novel-composition split contains one type of novel composition. We define the composition as novel composition if its constituents are both observed during training but their combination way is novel. Each sentence in the novel-word splits contains a novel word, which aims to test whether a model can infer the potential semantics of the unseen word based on the other learned composition components appearing in the context.

3.2. Dataset Re-splitting

For each dataset, we first combine all instances in the original training set and testing set, and remove the instances that can be easily predicted solely based on videos. We then re-split each dataset into four sets: training, novel-composition, novel-word, and test-trivial. The test-trivial set is similar to the original testing set, where most of the compositions are seen during training. Concretely, We use AllenNLP [11] to lemmatize and label all nouns, adjectives, verbs, adverbs, prepositions in language queries. Based on dependency parsing results, we define 5 types of compositions: verb-noun, adjective-noun, noun-noun, verb-adverb, and preposition-noun. For each type of composition, we construct a statistical table, where the row indexes are all possible first components of the composition and the column indexes are all possible second components of the composition. Taking verb-noun as an example, the element in row i and column j corresponds to the composition that consists of the $i-th$ verb and the $j-th$ noun in the dataset. For each table, we first sample an element from each row and each column, and then add all queries that contain the sampled compositions to the training set, which ensures that all components of compositions can be observed in the training set. Next, for each type of composition, we sample compositions from tables and add the corresponding queries into the novel-composition split. Meanwhile, we sample some words as new words and add the queries that contain the new words into the novel-word split. Since each video is associated with multiple text queries, if one query is selected into the training set, we will add other queries of the same video into the training set. If one query is selected into the novel-composition or novel-word split, we will add the remaining queries of the same video into the test-trivial set. Thus, we make sure that there is no video overlap between training and testing sets. Table 1 summarizes detailed statistics. We provide more details in supplementary materials.

Dataset	Split	Videos	Queries
Charades-CG	Training	3555	8281
	Novel-Composition	2480	3442
	Novel-Word	588	703
	Test-Trivial	1689	3096
ActivityNet-CG	Training	9659	36724
	Novel-Composition	4202	12028
	Novel-Word	2011	3944
	Test-Trivial	4775	15712

Table 1. Statistics of Charades-CG and Activity-CG.

4. Method

As illustrated in Figure 2, our VISA framework mainly consists of two components: a **hierarchical semantic graph** and a **variational cross-graph correspondence learning**. Given an untrimmed video V and a query sentence Q, the hierarchical semantic graph first decomposes them into three semantic hierarchies (*i.e. global events, local actions, and atomic objects*), respectively. Then, the variational cross-graph correspondence learning establishes fine-grained semantic correspondence between two graphs. Finally, based on the fine-grained semantic correspondence between video and sentence, our VISA infers the target moment that semantically corresponds to the given query.

4.1. Hierarchical Semantic Graph

Language queries describe some semantic events [22], which can be further parsed to central predicates and their corresponding arguments. Similarly, videos naturally record some relevant events in our lives, which consist of a variety of actions and each action involves multiple objects. Therefore, language and videos are both inherently organized in hierarchical structures. Based on this observation, given a video V and a query Q, we decompose both of them into three semantic hierarchies, which correspond to global events, local actions, and atomic objects, respectively. Such a hierarchical semantic graph provides a unified structured representation for modeling fine-grained semantic correspondence between videos and language queries.

Graph Initialization. For an untrimmed video V, we first divide it into a sequence of segments with a fixed length and extract the features using pre-trained 3D CNN: $\{V_t\}_{t=1}^{T}$, where $V_t = \{f_i^t\}_{i=1}^{K}$ and f_i^t denotes the C3D features of

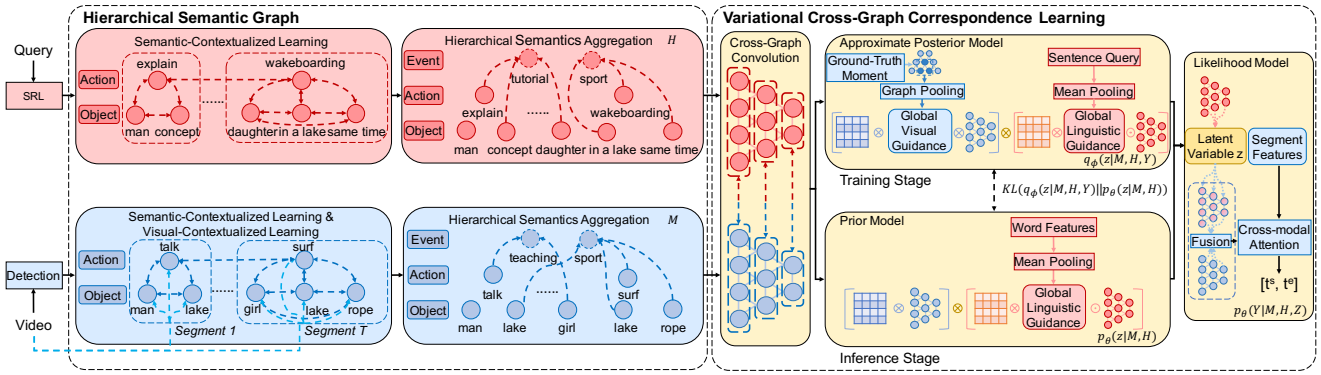

Figure 2. Overview of our VISA framework. We omit the details of the input video and sentence.

frame i in segment t. Then, we adopt off-the-shelf object detection and action recognition models to extract objects and actions for each segment, where each segment contains N_1 object nodes $\{\overline{s}^o_{t,i}\}_{i=1}^{N_1} \in \mathbb{R}^{d\times N_1}$ and N_2 action nodes $\{\overline{s}^a_{t,i}\}_{i=1}^{N_2} \in \mathbb{R}^{d\times N_2}$. We initialize both object and action nodes by the sum of the GloVe [36] vectors of each word in the object labels and action labels. Finally, all the object nodes $\overline{S}^o$ and action nodes $\overline{S}^a$ across segments constitute the first and second hierarchies of the video semantic graph.

For query Q, we use semantic role labeling (SRL) to decompose the query into multiple semantic structures. Each semantic structure contains a central predicate (verb) and some corresponding arguments (noun phrases including prep, adj, and adv). The predicates are considered as action nodes denoted by $\{\overline{c}^a_i\}_{i=1}^{L_2} \in \mathbb{R}^{d\times L_2}$, and the arguments are considered as object nodes denoted by $\{\overline{c}^o_{i,j}\}_{j=1}^{L_1} \in \mathbb{R}^{d\times L_1}$. If a word serves as multiple arguments for different predicates, we duplicate it for each action node. Similarly, we initialize them using GloVe word vectors. Finally, all the object nodes $\overline{C}^o$ and action nodes $\overline{C}^a$ constitute the first and second hierarchies of the language semantic graph.

Semantic-Contextualized Learning. Events are high-level semantic abstractions of video context and involve complicated interactions between different semantic concepts. For example, the query "the camel stands up and walks off with the family riding on its back" is composed of objects (*camel*, *the family*), actions (*stands up, walks off, and riding on*), and the underlying relations among them such as the spatial relation (*on its back*), the temporal relation (*stands up and walks off*), and the agentive relation (*riding on*). Therefore, to achieve comprehensive understanding of video events, we present semantic-contextualized learning to model the complicated interaction between the semantic nodes and learn fine-grained contextual information beyond the coarse semantic labels. Further, semantic contextual information is crucial for resolving semantic ambiguity of individual semantic nodes as the pre-trained detector might be noisy and the detected actions and objects might have dramatic variations in appearance.

Concretely, we define three types of undirect edges: *action-action*, *action-object*, and *object-object*. For the video semantic graph (/the language semantic graph), the object nodes in the same segment (/semantic structure) are connected by the *object-object* edges, the action and object nodes in the same segment (/semantic structure) are connected by the *action-object* edges, and all the action nodes are connected by the *action-action* edges. Afterward, we perform relation-aware graph convolution on video semantic graph. For a semantic node $\overline{s}_i \in \{\overline{S}^a, \overline{S}^o\}$, we calculate the adjacency correlation for each edge type r as:

$$\tilde{\alpha}^r_{ij} = (W^r\overline{s}_i)^T \cdot (W^r\overline{s}_j), \quad \alpha^r_{ij} = \frac{exp(\tilde{\alpha}^r_{ij})}{\sum_{j\in\mathcal{N}^r_i} exp(\tilde{\alpha}^r_{ij})} \quad (1)$$

where $\mathcal{N}^r_i$ is the neighborhood nodes of s_i on edge type r and W_r is the relation-specific projection matrix. Then, we refine s_i using the neighboring nodes of all edge types as:

$$\hat{s}_i = \sum_{r\in R}\sum_{j\in\mathcal{N}^r_i} \alpha^r_{ij} \cdot (U^r\overline{s}_j) \quad (2)$$

where R is the three types of edges and U_r is another transformation matrix. $\hat{s}_i$ is the result of the first relation-aware graph convolution layer. To model multi-order relations, we perform M layers of relation-aware graph convolution and learn final semantic-contextualized node features $S = \{s_i\}_{i=1}^{N_v} \in \mathbb{R}^{N_v\times d}$, where N_v is the total number of action and object nodes. In the same manner, we can obtain semantic-contextualized node features $C = \{c_i\}_{i=1}^{N_s} \in \mathbb{R}^{N_s\times d}$ of language semantic graph.

Visual-Contextualized Learning. We further propose visual-contextualized learning to collect relevant visual context from videos to the video semantic graph. Specifically, for a semantic node s_i, let $V_i = \{f^i_j\}_{j=1}^{K}$ denotes the corresponding segment, and f^i_j is the frame feature (following, we omit the superscript i for simplicity). We first compute the visual filter for each frame f_j in the segment and obtain the filtered visual feature as:

$$g^i_j = \sigma(W^g[s_i; \overline{f}; f_j] + b_g), \quad f^{'}_j = f_j \odot g^i_j \quad (3)$$

where $\odot$ denotes the Hadamard product, and $\overline{f}$ is obtained by performing average pooling on the V_i. Then, we perform max-pooling across the filtered frame features to get the semantic-relevant visual context as $F_i = MaxPool(f_1^{'}, ..., f_K^{'})$. Finally, we concatenate s_i with F_i and transform them to the original dimension by a transformation matrix W_v as $s_i = W^v[s_i, F_i]$. Here, we reuse s_i to represent the final visual-contextualized semantic node representation for simplicity.

Hierarchical Semantics Aggregation. Based on the observation that semantic events are composed of a series of interactional actions and objects, we propose the hierarchical semantic aggregation mechanism, which aggregates the semantics from the contextualized action nodes and object nodes to compose the global event nodes. Inspired by the success of positional query encoding [2] in object detection, we initialize the event nodes as a set of learnable query vectors $\{p_i\}_{i=1}^{N_p}$ and then aggregate relevant semantics from action nodes and object nodes to refine the event nodes. Here we take the video semantic graph as an illustration. For an event query p_i, we calculate the attention weights over semantic nodes $\{s_j\}_{j=1}^{N_v}$ and update the p_i, given by:

$$\tilde{p_i} = \sum_{j=1}^{N_v} \alpha_{ij}^e \cdot s_j, \alpha_{ij}^e = \frac{exp((W_1^e p_i)^T \cdot (W_2^e s_j))}{\sum_{j=1}^{N_v} exp((W_1^e p_i)^T \cdot (W_2^e s_j))} \tag{4}$$

where W_1^e, W_2^e are projection matrices, and $\tilde{p_i}$ is the semantics-aware event node. Subsequently, we stack multiple such graph self-attention layers and merge the final event nodes into $\{s_j\}_{j=1}^{N_v}$ to form the complete hierarchical semantic graph of video, denoted by $M = \{m_i\}_{i=1}^{N_m} \in \mathbb{R}^{d \times N_m}$. In the same manner, we can obtain the complete hierarchical semantic graph of language, denoted by $H = \{h_i\}_{i=1}^{N_h} \in \mathbb{R}^{d \times N_h}$. M and H are the unified structure of three semantic hierarchies, which tightly couple multi-granularity semantics between the two modalities.

4.2. Variational Cross-Graph Correspondence

After parsing both videos and language queries into individual hierarchical semantic graphs, we then model the cross-modality interactions between two graphs by cross-graph convolution, and induce the fine-grained semantic correspondence between them for final prediction. The objective function can be formulated as $P(Y|M, H)$, where Y is the target time interval. Since the ground-truth correspondence between two graphs is not available, we treat the cross-graph correspondence as a latent variable z. The problem can then be formulated into a variational inference framework [39] and the objective function can be rewritten as $P(Y|M, H, z)P(z|M, H)$. Instead of directly maximizing $P(Y|M, H)$, we propose to maximize its evidence lower bound (ELBO) [15] as follows:

$$\begin{aligned} \mathcal{L}^{ELBO}(\phi, \theta) &= E_{q_\phi(z|M,H,Y)} log p_\theta(Y|M, H, z) \\ &\quad - KL(q_\phi(z|M, H, Y)||p_\theta(z|M, H)) \\ &\leq log p(Y|M, H) \end{aligned} \tag{5}$$

Specifically, we characterize $P(Y|M, H)$ using three components: a prior model $p_\theta(z|M, H)$, a posterior model $q_\phi(z|M, H, Y)$, and a likelihood model $p_\theta(Y|M, H, z)$. In the following, we first introduce cross-graph convolution to capture the semantic correlation between two graphs and then describe these three models in detail.

Cross-Graph Convolution. Given the graphs M and H, we perform cross-graph convolution between the same hierarchical levels of two graphs. For a video semantic node m_i^k, the cross-convolution from H to M is formulated as:

$$\alpha_{ij}^{h2m} = \frac{exp((W_1^c m_i^k)^T \cdot (W_2^c h_j^k))}{\sum_{j \in \mathcal{N}_H^k} exp((W_1^c m_i^k)^T \cdot (W_2^c h_j^k))} \tag{6}$$

$$\tilde{m}_i^k = (1 - \beta_i^k) \odot m_i^k + \beta_i^k \odot \sum_{j \in \mathcal{N}_H^k} \alpha_{ij}^{h2m} \cdot h_j^k, k \in \{e, a, o\} \tag{7}$$

where $\beta_i^k = \sigma(U^g m_i^k + b)$ controls the information flow from H to M, k denotes three semantic levels (*i.e. event, action, object*), $\mathcal{N}_H^k$ denotes the nodes of H in level k. In a similar manner but reversed order, we can obtain $\tilde{H}$.

Prior Model. Given $\tilde{M}$ and $\tilde{H}$, the prior model $p_\theta(z|M, H)$ aims to infer the cross-graph correspondence captured by a latent variable $z \in \mathbb{R}^{N_m \times N_h}$, where z_{ij} corresponds to the semantic correspondence between $\tilde{m}_i$ and $\tilde{h}_j$. Specifically, the z_{ij} can be formulated as:

$$\tilde{z}_{ij} = (W_1^s \tilde{m}_i)^T \cdot (W_2^s q \odot \tilde{h_j}), z_{ij} = \frac{exp(\tilde{z}_{ij})}{\sum_{j=1}^{N_h} exp(\tilde{z}_{ij})} \tag{8}$$

where q is the global sentence feature that guides the semantic correspondence inference.

Approximate Posterior Model. The posterior model $q_\phi(z|M, H, Y)$ infers the cross-graph correspondence with additional information of ground-truth Y. According to the temporal boundary Y, we can determine the segments in Y and the action and object nodes that correspond to these segments. These nodes in the video graph contain the most relevant semantics to the language semantic graph, which can better guide cross-graph correspondence learning. Therefore, we obtain m^* through mean-pooling over these nodes

and use m^* to guide the correspondence learning:

$$\tilde{z}_{ij} = (W_3^s m^* \odot \tilde{m}_i)^T \cdot (W_4^s q \odot \tilde{h}_j),\ z_{ij} = \frac{exp(\tilde{z}_{ij})}{\sum_{j=1}^{N_h} exp(\tilde{z}_{ij})} \tag{9}$$

where m^* and q serve as global visual and linguistic guidance, respectively.

Likelihood Model. The likelihood model $p_\theta(Y|M, H, z)$ predicts the temporal boundary based on the latent correspondence z and hierarchical semantic graphs $\tilde{M}$ and $\tilde{H}$. Specifically, we first integrate two graphs based on the learned cross-graph correspondence to obtain joint multi-modality representations:

$$M^{'} = z\tilde{H} \in \mathbb{R}^{d \times N_m}, \quad M^J = W^J[\tilde{M}; M^{'}] \in \mathbb{R}^{d \times N_m} \tag{10}$$

where projection matrix $W^J \in \mathbb{R}^{d \times 2d}$ and M^J is the joint multi-modality representations of the hierarchical semantic graph. Next, we use M^J to refine segment features $X = \{x_t\}_{t=1}^T \in \mathbb{R}^{d \times T}$. We perform mean-pooling over frame features $\{f_i^t\}_{i=1}^K$ of segment V_t to obtain the segment features x_t. We adopt multi-head cross-modal attention to softly select relevant information from M^J to X. Concretely, we take X as queries and M^J as keys and values:

$$X^* = MultiAttn(X, M^J, M^J) \tag{11}$$

where X^* is the semantics-aware segment representations. Subsequently, we summarize the segment representations using attentive pooling based on the sentence feature q:

$$v^* = \sum_{i=1}^{T} \alpha_i^q \cdot x_i^*, \quad \alpha_i^q = \frac{exp((W_1^q q)^T \cdot (W_2^q x_i^*))}{\sum_{i=1}^{T} exp((W_1^q q)^T \cdot (W_2^q x_i))} \tag{12}$$

where v^* is the summarized video feature. Finally, we predict the time interval (t^s, t^e) as $t^s, t^e = MLP(v^*)$.

4.3. Optimization

As described in Equation 5, the ELBO objective function consists of two terms. The first term corresponds to the negative number of the regression loss. Specifically, following [32], we minimize the sum of smooth L_1 distances between the normalized ground-truth time interval $(\hat{t^s}, \hat{t^e}) \in [0, 1]$ and our prediction (t^s, t^e). This term not only teaches the likelihood model to predict the correct time interval but also encourages the approximate posterior model to learn more accurate cross-graph correspondence. The second term corresponds to the KL-divergence loss. Concretely, as the latent variable z is a correlation matrix, we compute the KL-divergence by rows. Intuitively, through minimizing this term, we can teach the prior model to capture the cross-graph semantic correspondence as well as the approximate posterior model. During testing without access to the ground-truth, we can use the learned prior model to replace the approximate posterior model to infer the cross-graph correspondence. Note that we use the approximate posterior model to generate z during training.

5. Experiments

5.1. Benchmarking the SOTA Methods

We evaluate the compositional generalizability of SOTA methods on the proposed Charades-CG and ActivityNet-CG datasets. Specifically, these methods can be categorized into four groups: 1) Proposal-based methods: **TMN** [27], **2D-TAN** [48]; 2) Proposal-free methods: **LGI** [32], **VL-SNet** [46]; 3) RL-based method: **TSP-PRL** [41]; 4) Weakly-supervised method: **WSSL** [7]. Due to the space limitation, we provide more experimental results and implementation details in supplemental materials.

Evaluation Metrics. Following previous works, we adopt "R@n, IoU=m" and mIoU (*i.e.* the average temporal IoU) as our evaluation metrics. Specifically, given a testing query, it first calculates the Intersection-over-Union (IoU) between the predicted moment and the ground truth, and "R@n, IoU=m" is defined as the percentage of at least one of top-n predictions with IoU larger than m.

5.2. Results on Compositional Temporal Grounding

Table 2 and Table 3 summarize the results of the above methods on compositional temporal grounding. Overall, our VISA achieves the highest performance on all dataset splits, demonstrating the superiority of our proposed model. Notably, we observe that the performance of all tested SOTA models drops significantly on the novel-composition and novel-word splits. The difference in performance between test-trivial and novel-composition (novel-word) ranges up to 20%. In contrast, our VISA surpasses them by a large margin on novel-composition and novel-word splits, demonstrating superior compositional generalizability. Particularly, for the novel-composition splits of Charades-CG and ActivityNet-CG datasets, our method significantly surpasses the SOTA methods by 30.86% and 23.32% relatively on mIoU, respectively.

5.3. In-Depth Analysis

Effect of Individual Components. We conduct an ablation study to illustrate the effect of each component in Table 4. Specifically, we train the following ablation models. 1) w/o SCL: we remove the Semantic-Contextualized Learning (SCL). 2) w/o VCL: we remove the Visual-Contextualized Learning (VCL). 3) w/o HSA: we remove the Hierarchical Semantics Aggregation (HSA). 4) w/o VCC: we replace the Variational Cross-graph Correspondence learning (VCC) by directly using cross-modal self-

	Method	*Test-Trivial*			*Novel-Composition*			*Novel-Word*		
		IoU=0.5	IoU=0.7	mIoU	IoU=0.5	IoU=0.7	mIoU	IoU=0.5	IoU=0.7	mIoU
Weakly-supervised	WSSL	15.33	5.46	18.31	3.61	1.21	8.26	2.79	0.73	7.92
RL-based	TSP-PRL	39.86	21.07	38.41	16.30	2.04	13.52	14.83	2.61	14.03
Proposal-based	TMN	18.75	8.16	19.82	8.68	4.07	10.14	9.43	4.96	11.23
	2D-TAN	48.58	26.49	44.27	30.91	12.23	29.75	29.36	13.21	28.47
Proposal-free	LGI	49.45	23.80	45.01	29.42	12.73	30.09	26.48	12.47	27.62
	VLSNet	45.91	19.80	41.63	24.25	11.54	31.43	25.60	10.07	30.21
	Ours-VISA	**53.20**	**26.52**	**47.11**	**45.41**	**22.71**	**42.03**	**42.35**	**20.88**	**40.18**

Table 2. Performances (%) of SOTA temporal grounding models and our VISA on the proposed Charades-CG datasets.

	Method	*Test-Trivial*			*Novel-Composition*			*Novel-Word*		
		IoU=0.5	IoU=0.7	mIoU	IoU=0.5	IoU=0.7	mIoU	IoU=0.5	IoU=0.7	mIoU
Weakly-supervised	WSSL	11.03	4.14	15.07	2.89	0.76	7.65	3.09	1.13	7.10
RL-based	TSP-PRL	34.27	18.80	37.05	14.74	1.43	12.61	18.05	3.15	14.34
Proposal-based	TMN	16.82	7.01	17.13	8.74	4.39	10.08	9.93	5.12	11.38
	2D-TAN	44.50	26.03	42.12	22.80	9.95	28.49	23.86	10.37	28.88
Proposal-free	LGI	43.56	23.29	41.37	23.21	9.02	27.86	23.10	9.03	26.95
	VLSNet	39.27	23.12	42.51	20.21	9.18	29.07	21.68	9.94	29.58
	Ours-VISA	**47.13**	**29.64**	**44.02**	**31.51**	**16.73**	**35.85**	**30.14**	**15.90**	**35.13**

Table 3. Performances (%) of SOTA temporal grounding models and our VISA on the proposed ActivityNet-CG datasets.

	Method	Charades-CG		ActivityNet-CG	
		Comp	Word	Comp	Word
1	w/o SCL	43.75	40.16	29.03	29.41
2	w/o VCL	42.26	38.62	29.34	28.09
3	w/o HSA	44.22	41.09	30.29	29.31
4	w/o VCC	41.08	37.54	27.32	26.37
5	Detection	12.97	11.70	10.92	10.07
6	**VISA**	**45.41**	**42.35**	**31.51**	**30.14**

Table 4. Ablation results with metric R@1, IoU=0.5 on novel-composition (Comp) and novel-word (Word) splits.

attention to fuse two graphs. 5) Detection-based: we directly use the detection results and SRL labels as features.

The results of Row 1 and Row 2 indicate that learning fine-grained contextualized information is crucial for compositional reasoning. Also, the results of Row 3 validate the importance of event-level hierarchy on global semantic understanding. Ours w/o VCC does not achieve satisfying results, because directly fusing the graphs of video and sentence could possibly disrupt the semantic correspondence between them, which causes a negative effect on temporal grounding performance. In contrast, the proposed VCC establishes fine-grained cross-graph correspondence by variational inference, which is meticulous and achieves the best results. Furthermore, Row 5 suggests that the main performance gain does not directly come from the pre-trained detection models. Instead, these detected semantic labels serve as unified symbols for joint compositional reasoning.

Results on Different Composition Types. To gain further insight, we examine the results (R@1, IoU=0.5) of our models on different types of compositions. Table 5 shows that generalizing to "Verb-Noun" compositions is the most difficult, as it requires the model to accurately identify the corresponding action and objects in video and jointly reason over them to infer the semantics of the novel composition.

Type	Charades-CG			ActivityNet-CG		
	w/o VCC	w/o SCL	VISA	w/o VCC	w/o SCL	VISA
Verb-Noun	36.56	38.82	41.37	24.41	26.32	28.89
Adj-Noun	42.17	44.04	45.06	26.76	28.31	30.67
Noun-Noun	40.38	42.56	43.41	29.51	30.20	33.93
Verb-Adv	43.81	46.37	47.83	31.08	33.46	35.60
Prep-Noun	44.12	47.86	48.61	34.78	36.03	37.35

Table 5. Performance of our models on each composition type.

Word Order Sensitivity. To gain more intuitive insight, we explore whether the models are sensitive to the word order, which is a crucial factor for the compositionality of language. Intuitively, if we change the word order of a sentence, its semantics might change greatly and thus the original ground-truth temporal boundaries might not be suitable for the shuffled sentence. Specifically, we randomly shuffle queries in advance and then use the shuffled queries to train and evaluate the models. We define the sensitivity metric as the relative performance degradation of the shuffled version on R@1, IoU=0.5. The higher value indicates a higher sensitivity. According to Table 6, we surprisingly find that SOTA models are insensitive to the word order. In contrast, our method are sensitive to the linguistic structure of sentences. Moreover, we observe the highest sensitivity of our VISA on novel-word splits, indicating that the linguistic structure is important for inferring the semantics of novel words. In the end, we observe that the proposed SCL and VCC promote our method to capture the linguistic structure of sentences in a mutually rewarding way.

5.4. Qualitative Analysis

Sensitivity on Specific Shuffling. We manually select some query sentences and change their word order in some specific ways, such that the changed query can still semantically correspond to other segments in their original videos.

Method	Charades-CG			ActivityNet-CG		
	Trivial	Comp	Word	Trivial	Comp	Word
2D-TAN	0.41	0.52	0.43	0.29	0.30	0.41
LGI	0.28	0.23	0.16	0.31	0.22	0.19
VLSNet	0.07	0.24	0.10	0.24	0.31	0.48
VISA	24.14	29.80	33.97	22.09	27.60	31.89
w/o SCL	19.64	24.31	29.72	18.07	24.64	28.73
w/o VCC	21.32	26.73	30.88	20.15	25.46	29.79

Table 6. The word order sensitivity of SOTA models and VISA.

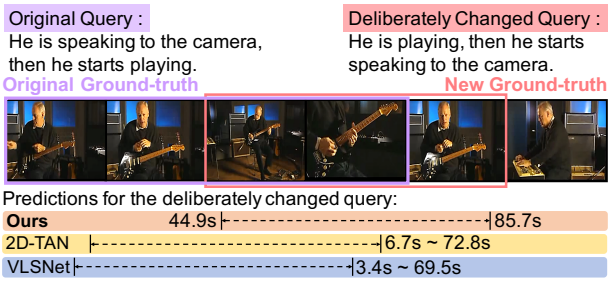

Figure 3. Qualitative examples on specifically shuffled queries.

As shown in Figure 3, we annotate the changed query with a new ground-truth (red box) and use the changed query to test models. Interestingly, the predictions of SOTA methods have higher IoU with the original temporal boundary, though the semantics of the sentence has been deliberately modified. In contrast, our VISA keenly captures the semantics change and locates to the new temporal boundary.

Qualitative Examples. Figure 4 visualizes three qualitative examples, which indicate the importance of compositionality. In the first case, the baseline fails to understand the composition meaning of "prepares to jump ", so it mistakenly localizes to the "jump" segment. In contrast, our VISA successfully captures the compositional meanings. The second case contains complex compositions, which describe two events. Without inferring their temporal relationship from the composition structure, the baseline localizes a wrong segment, even though it also contains the two individual events (*i.e.* *"a man talk"* and *"the reporter in the street talks"*). Conversely, our VISA understands the correct temporal order of these two events. The third case shows that our VISA successfully generalizes to novel composition. While *pulling* (*e.g.* *pulling rope*) and *horse* (*e.g.* *lead horse*) are both observed in the training split, the baseline suffers from generalizing to this novel composition.

Visualizing Learned Graph. In Figure 5, we present the learned hierarchical semantic graph. We visualize some key nodes and the edges with high weights. The yellow dotted lines represent the cross-graph semantic correspondence. If the semantic correspondence score between two nodes is greater than a specific threshold, we connect them with a yellow dotted line. We represent the event nodes by their most related semantics according to their attended nodes. Our VISA successfully aligns the visual semantics "punching person" and "boxing ring' to the linguistic words "perform kickboxing". Also, our VISA can connect "push up" and "exercise ball" to the words "push up (with) ball".

Figure 4. Qualitative examples of our VISA and VLSNet. The red boxes represent the ground-truth.

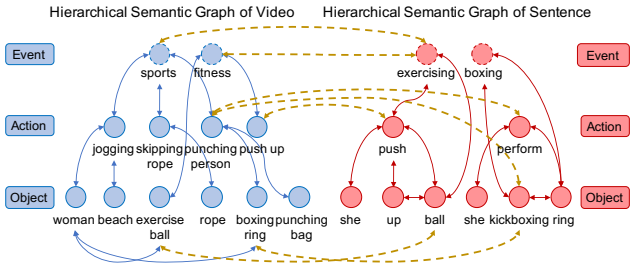

Figure 5. Visualization of the learned hierarchical semantic graph.

6. Conclusions

In this paper, we introduce a new task, compositional temporal grounding to systematically evaluate the compositional generalizability of temporal grounding models. We conduct in-depth analyses on SOTA methods, and find they lack of compositional generalizability. We then introduce a novel VISA framework that learns fine-grained semantic correspondence between video and language in three semantic hierarchies. Experiments illustrate significant improvement of our VISA on compositional generalizability.

Limitations and Futuer work. We observe some failure cases that VISA cannot discriminate subtle semantics of adverbs, *e.g.*, "fly close" to "fly away". We expect future research to utilize the new benchmarks to make progress on fine-grained semantics grounding, thus achieving compositional generalization.

Acknowledgment. This work has been supported in part by National Key Research and Development Program of China (2018AAA0101900), Zhejiang NSF (LR21F020004), Key Research and Development Program of Zhejiang Province, China (No.2021C01013), Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, Chinese Knowledge Center of Engineering Science and Technology (CKCEST), Zhejiang University iFLYTEK Joint Research Center. The author from UCSC is not supported by any of

the projects above. We thank all the reviewers for valuable comments.

References

[1] Jacob Andreas. Good-enough compositional data augmentation. *arXiv preprint arXiv:1904.09545*, 2019. 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 5

[3] 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. 1

[4] Noam Chomsky. *Syntactic structures*. De Gruyter Mouton, 2009. 1

[5] Yadong Ding, Yu Wu, Chengyue Huang, Siliang Tang, Fei Wu, Yi Yang, Wenwu Zhu, and Yueting Zhuang. Nap: Neural architecture search with pruning. *Neurocomputing*, 2022. 2

[6] Yadong Ding, Yu Wu, Chengyue Huang, Siliang Tang, Yi Yang, Longhui Wei, Yueting Zhuang, and Qi Tian. Learning to learn by jointly optimizing neural architecture and weights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[7] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. *arXiv preprint arXiv:1812.03849*, 2018. 6

[8] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016. 1

[9] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988. 1

[10] 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. 1, 2, 3

[11] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*, 2018. 3

[12] Jonathan Gordon, David Lopez-Paz, Marco Baroni, and Diane Bouchacourt. Permutation equivariant models for compositional generalization in language. In *International Conference on Learning Representations*, 2019. 2

[13] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 3

[14] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 3

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5

[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. 1, 2, 3

[17] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR, 2018. 2

[18] Brenden M Lake. Compositional generalization through meta sequence-to-sequence learning. *arXiv preprint arXiv:1906.05381*, 2019. 2

[19] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 2

[21] Juncheng Li, Siliang Tang, Fei Wu, and Yueting Zhuang. Walking with mind: Mental imagery enhanced embodied qa. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1211–1219, 2019. 2

[22] Juncheng Li, Siliang Tang, Linchao Zhu, Haochen Shi, Xuanwen Huang, Fei Wu, Yi Yang, and Yueting Zhuang. Adaptive hierarchical graph reasoning with semantic coherence for video-and-language inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1867–1877, 2021. 3

[23] Juncheng Li, Xin Wang, Siliang Tang, Haizhou Shi, Fei Wu, Yueting Zhuang, and William Yang Wang. Unsupervised reinforcement learning of transferable meta-skills for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12123–12132, 2020. 2

[24] Mengze Li, Ming Kong, Kun Kuang, Qiang Zhu, and Fei Wu. Multi-task attribute-fusion model for fine-grained image recognition. In *Optoelectronic Imaging and Multimedia Technology VII*, 2020. 2

[25] Mengze Li, Kun Kuang, Qiang Zhu, Xiaohong Chen, Qing Guo, and Fei Wu. Ib-m: A flexible framework to align an interpretable model and a black-box model. In *BIBM*, 2020. 2

[26] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Wenming Tan, Jin Wang, Peng Wang, Shiliang Pu, and Fei Wu. End-to-end modeling via information tree for one-shot natural language spatial video grounding. 2

[27] 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. 6

[28] Shugao Ma, Leonid Sigal, and Stan Sclaroff. Learning activity progression in lstms for activity detection and early detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1942–1950, 2016. 1

[29] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 5222–5230, 2021. 3

[30] 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*, pages 1049–1059, 2020. 3

[31] Richard Montague et al. Universal grammar. *1974*, pages 222–46, 1970. 1

[32] 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. 2, 6

[33] Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. Compositional generalization in image captioning. *arXiv preprint arXiv:1909.04402*, 2019. 3

[34] Maxwell I Nye, Armando Solar-Lezama, Joshua B Tenenbaum, and Brenden M Lake. Learning compositional rules via neural program synthesis. *arXiv preprint arXiv:2003.05562*, 2020. 2

[35] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. *arXiv preprint arXiv:2009.00325*, 2020. 2

[36] 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. 4

[37] Jake Russin, Jason Jo, Randall C O'Reilly, and Yoshua Bengio. Compositional generalization in a deep seq2seq model by separating syntax and semantics. *arXiv preprint arXiv:1904.09708*, 2019.

[38] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1961–1970, 2016. 1

[39] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491, 2015. 5

[40] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 450–459, 2019. 3

[41] 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. 6

[42] Yitian Yuan, Xiaohan Lan, Long Chen, Wei Liu, Xin Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Datasets and metrics. *arXiv preprint arXiv:2101.09028*, 2021. 2

[43] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *arXiv preprint arXiv:1910.14303*, 2019.

[44] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166, 2019. 2

[45] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1247–1257, 2019. 2

[46] 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. 2, 6

[47] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, 2020. 2

[48] 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. 2, 6

[49] Shengyu Zhang, Ziqi Tan, Zhou Zhao, Jin Yu, Kun Kuang, Tan Jiang, Jingren Zhou, Hongxia Yang, and Fei Wu. Comprehensive information integration modeling framework for video titling. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, 2020. 2

[50] Shengyu Zhang, Dong Yao, Zhou Zhao, Tat-Seng Chua, and Fei Wu. Causerec: Counterfactual user sequence synthesis for sequential recommendation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, 2021. 2

[51] 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. *arXiv preprint arXiv:2112.06558*, 2021. 3

[52] Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. Consensus graph representation learning for better grounded image captioning. In *Proc 35 AAAI Conf on Artificial Intelligence*, 2021. 3

[53] Wenqiao Zhang, Xin Eric Wang, Siliang Tang, Haizhou Shi, Haochen Shi, Jun Xiao, Yueting Zhuang, and William Yang Wang. Relational graph learning for grounded video description generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3807–3828, 2020. 2

[54] Wenqiao Zhang, Lei Zhu, James Hallinan, Andrew Makmur, Shengyu Zhang, Qingpeng Cai, and Beng Chin Ooi. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[55] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 3