A Closer Look at Debiased Temporal Sentence Grounding in Videos: Dataset, Metric, and Approach

XIAOHAN LAN, Tsinghua Shenzhen International Graduate School, China
YITIAN YUAN, Meituan, China
XIN WANG, Tsinghua University, China
LONG CHEN, Columbia University, USA
ZHI WANG, Tsinghua Shenzhen International Graduate School, China
LIN MA, Meituan, China
WENWU ZHU, Tsinghua University, China

Temporal Sentence Grounding in Videos (TSGV), which aims to ground a natural language sentence that indicates complex human activities in an untrimmed video, has drawn widespread attention over the past few years. However, recent studies have found that current benchmark datasets may have obvious moment annotation biases, enabling several simple baselines even without training to achieve **state-of-the-art (SOTA)** performance. In this paper, we take a closer look at existing evaluation protocols for TSGV, and find that both the prevailing dataset splits and evaluation metrics are the devils that lead to untrustworthy benchmarking. Therefore, we propose to re-organize the two widely-used datasets, making the ground-truth moment distributions different in the training and test splits, i.e., **out-of-distribution (OOD)** test. Meanwhile, we introduce a new evaluation metric "dR@n,IoU=m" that discounts the basic recall scores especially with small IoU thresholds, so as to alleviate the inflating evaluation caused by biased datasets with a large proportion of long ground-truth moments. New benchmarking results indicate that our proposed evaluation protocols can better monitor the research progress in TSGV. Furthermore, we propose a novel causality-based **Multi-branch Deconfounding Debiasing (MDD)** framework for unbiased moment prediction. Specifically, we design a multi-branch deconfounder to eliminate the effects caused by multiple confounders with causal intervention. In order to help the model better align the semantics between sentence queries and video moments, we enhance the representations during feature encoding. Specifically, for textual information, the query is parsed into several verb-centered phrases to obtain a more fine-grained textual feature. For visual information, the positional information has been decomposed from the moment features to enhance the representations of moments with diverse locations. Extensive experiments demonstrate that our proposed approach can achieve competitive results among existing SOTA approaches and outperform the base model with great gains.

This work was supported in part by the National Key Research and Development Program of China No. 2020AAA0106300 and National Natural Science Foundation of China (No. 62250008, 62222209, 62102222, 61872215).

Authors' addresses: X. Lan and Z. Wang, Tsinghua Shenzhen International Graduate School, University Town of Shenzhen, Nanshan District, Shenzhen 518055 China; emails: lanxh20@mails.tsinghua.edu.cn, wangzhi@sz.tsinghua.edu.cn; Y. Yuan and L. Ma, Meituan, Tower C, Hengjiweiye Building, No. 4 Wangjing East Road, Chaoyang District, Beijing, 100102 China; emails: yuanyitian@foxmail.com, forest.linma@gmail.com; X. Wang and W. Zhu (corresponding author), Tsinghua University, 30 Shuangqing Rd, Beijing, 100084 China; emails: {xin_wang, wwzhu}@tsinghua.edu.cn; L. Chen (corresponding author), Columbia University, 500 W 120th St. Mudd 510 New York, NY 10027 USA; email: zjuchenlong@gmail.com.

1551-6857/2023/07-ART218
https://doi.org/10.1145/3565573

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; *Natural language processing*; *Computer vision;*

Additional Key Words and Phrases: Temporal sentence grounding in videos, dataset bias, evaluation metric, dataset re-splitting, out-of-distribution test

ACM Reference format:
Xiaohan Lan, Yitian Yuan, Xin Wang, Long Chen, Zhi Wang, Lin Ma, and Wenwu Zhu. 2023. A Closer Look at Debiased Temporal Sentence Grounding in Videos: Dataset, Metric, and Approach. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 6, Article 218 (July 2023), 23 pages.
https://doi.org/10.1145/3565573

1 INTRODUCTION

Temporal Sentence Grounding in Videos (TSGV) has received increased attention in recent years. Specifically, given one descriptive sentence, the TSGV task aims to retrieve a video segment (i.e., moment) from an untrimmed video corresponding to the sentence query. For example, as shown in Figure 1(a), when the sentence describes a person pouring coffee into a cup in the dining room, the corresponding video segment (21.3s–30.7s) should be located. It can be observed that TSGV needs to understand both visual information in videos and textual information in sentences, which is an extremely challenging task in the multimedia community [33, 60].

In recent years, a number of approaches [5, 11, 20, 29, 45, 54, 57] have emerged to solve the TSGV problem. Although each newly proposed method can plausibly achieve better performance than the previous one, a recent study [32] reveals that current **state-of-the-art (SOTA)** methods may take shortcuts by fitting the ground-truth moment annotation distribution biases, without truly understanding the multimodal inputs. As shown in the Figure 1(b), the `Bias-based` approach, which samples a moment from the frequency statistics of the ground-truth moment annotations in the training set as prediction, can unexpectedly outperform several SOTA deep models on Charades-STA [11] dataset. This observation indicates that current benchmark datasets may have obvious biases in terms of moment location distribution, and it is hard to judge whether existing methods are merely fitting the biases or truly learning the semantic alignment relationship between the two modalities. Another characteristic of biased datasets is that they have a large proportion of long samples, e.g., 40% queries in the ActivityNet Captions dataset [25] refer to a moment occupying over 30% temporal ranges of the whole input video. Since prevailing metric for TSGV task is "R@n,IoU=m", i.e., the percentage of testing samples which have at least one of the top-n results with IoU larger than m, these overlong ground-truth moments can be hit easily especially with small IoU threshold m, resulting in untrustworthy evaluation results. As an extreme case, a simple baseline which directly returns the whole video as the prediction (c.f., the `PredictAll` baseline in Figure 1(c)) can still achieve a SOTA performance with the metric of "R@1,IoU=0.3".

Therefore, to disentangle the effect caused by the biases and alleviate the inflating evaluation, we propose to re-split the datasets and design a new metric. Specifically, we re-organize two widely-used datasets, i.e., Charades-STA and ActivityNet Captions and name them **Charades-CD** and **ActivityNet-CD** (**CD** means under **Changing Distribution**). For each dataset, besides the test set with the same distribution as the training set (test-iid), we also construct a test set with a completely different distribution of moment locations from the training set (dubbed as test-ood set), i.e., **Out-Of-Distribution (OOD)** test. As for metrics, we design a new evaluation metric "dR@n,IoU=m" that takes temporal distances between the predicted moment and ground-truth moment into consideration. The new metric can discount the basic recall scores especially under small IoU thresholds. So our proposed evaluation protocols (i.e., re-organized dataset splits and improved evaluation metric) are able to provide more trustworthy evaluation results for existing methods and figure out whether they just fit the moment annotation biases. Several representative

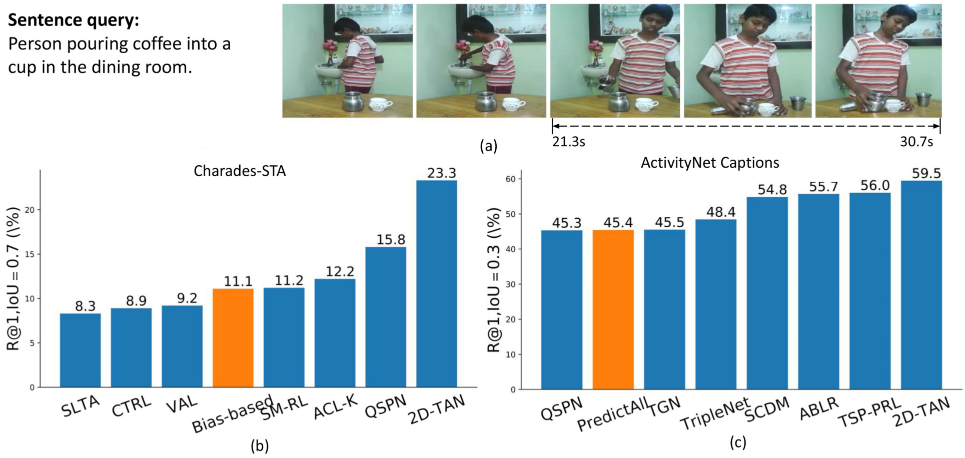


Fig. 1. (a): Given an untrimmed video and a sentence query, TSGV aims to localize the semantic-related moment with the start timestamp (21.3s) and end timestamp (30.7s). (b): The performance comparisons of some SOTA TSGV models with the Bias-based baseline (orange bar) on Charades-STA with evaluation metric "R@1,IoU=0.7". (c): The performance comparisons of some SOTA TSGV models with PredictAll baseline (orange bar) on ActivityNet Captions with evaluation metric "R@1,IoU=0.3".

TSGV methods are tested with such new evaluation protocols and one key finding is that the performances of the vast majority of these methods degrade significantly on test-ood set compared to test-iid set, which indicates the fact that current methods heavily rely on the biases for moment prediction. Therefore, how to effectively debias a TSGV model to make it truly focus on the semantic alignment between the two modalities becomes one primary issue to be addressed.

To reduce the effects of moment annotation biases, we further propose a novel causality-based **Multi-branch Deconfounding Debiasing** (MDD) framework. Specifically, by constructing a causal graph, we find that multiple confounders can lead to spurious correlations between the multimodal inputs (video moments and descriptive sentences) and final predicted matching scores. Thus, we present a multi-branch deconfounder to block the effects caused by the confounders with backdoor adjustment. Furthermore, we also enhance the representation capability for two modalities. For textual information, we exploit a semantic role labeling toolkit to parse the sentence into a three-layer semantic role tree, and a more fine-grained sentence feature is obtained by adopting hierarchical attention mechanism on the tree. For visual information, in order to discriminate video moments and distinguish different temporal relationships, a reconstruction loss function is created to enhance the video moment features. Extensive experiments demonstrate that the adopting of our debiasing strategy can significantly improve the grounding accuracy on both the test-iid and test-ood sets of two datasets.

This paper is a substantial extension of our ACM Multimedia HUMA Workshop paper [52], which won the best paper award. Compared with the previous version, we make several improvements:

- We propose a new Multi-branch Deconfounding Debiasing (MDD) framework for unbiased moment prediction. The proposed multi-branch deconfounder can simultaneously remove the spurious correlation between multimodal inputs and predicted scores caused by multiple confounders, avoiding the excessive abuse of dataset biases.
- We also enhance the visual and textual features for better cross-modal matching. The gated fine-grained feature extractor for queries and position reconstruction module for video moments can capture richer and more discriminative representations of these two modalities.

- We conduct extensive studies on MDD framework compared with existing SOTA models. The experimental results demonstrate that our approach can achieve competitive results among existing SOTA models and outperform the base model with great gains on both test-iid and test-ood sets.

2 RELATED WORK

2.1 Temporal Sentence Grounding in Videos

Existing TSGV methods can be summarized into four main categories:

Two-Stage Methods. Early methods commonly address the TSGV task in a two-stage manner. In particular, they first extract a large number of moment candidates via sliding window sampling strategy, and then either project the query and these candidates into a common space for subsequent cross-modal matching [20] or fuse the query feature and video moment features to predict the alignment score and refine the moments with position offset regression [11, 14, 23, 26, 27, 38, 47, 48]. To reduce the number of candidates for accelerating the localization process, Xu et al. [49] proposed QSPN, which filters unlikely video moments by injecting the textual feature into the early process of candidate generation.

End-to-End Methods. Other than adopting the two-stage framework that is inefficient due to the redundant computation with pre-segmented overlapping candidate moments, some studies start to address the TSGV task in an end-to-end pipeline [3, 5, 6, 28, 53–58]. TGN [5] adopts LSTM [21] to sequentially score a bunch of multi-scale moment candidates ended at each time step in one single pass. Instead of candidate moment scoring, ABLR [54] directly regresses the start and end timestamps of the predicted moments from the attention weights yielded by the multi-turn cross-modal interaction. It is worth noting that both TGN and ABLR use LSTM to process the video stream and some other TSGV frameworks adopt temporal convolutional networks as the solutions. MAN [56] employs a hierarchical convolutional network to encode the whole video stream, where the language features are integrated as its dynamic filters to address semantic misalignment. Yuan et al. [53] presented SCDM, which conducts a query semantics-guided feature normalization process among different temporal convolutional layers. Both MAN and SCDM encode the video sequence with 1D feature map which can naturally indicate the temporal locations and scales of different moments, while 2D-TAN [57] models the temporal relations between video moments with a 2D temporal map. The 2D temporal map can encode the temporally adjacent relations of diverse moments indicated by their 2D position coordinates. Thus, more discriminative moment representations can be learned for cross-modal matching.

RL-based Methods. Some recent works employ **Reinforcement Learning (RL)**-based frameworks, which formulate the TSGV task as a problem of sequential decision making, progressively adjusting the temporal boundaries of predicted moment [17, 18, 44, 45]. Specifically, He et al. [18] proposed the RL model that iteratively regulates current locations according to the learned policy. The policy network is implemented by a **recurrent neural network (RNN)** that outputs the probability distribution over its action space. Wang et al. [44] presented a **semantic matching RL (SM-RL)** model, which is also based on RNN. The SM-RL integrates visual semantic concepts into the video features to bridge the semantic gap between visual and textual information. TripNet [17] can efficiently localize the desired moment without watching the entire video, by making the agent learn how to intelligently move the candidate window around the video. Inspired by the coarse-to-fine human decision-making paradigm, Wu et al. [45] designed a tree-structured policy based RL model, where the root policy and leaf policy represent the coarse and fine decision-making steps respectively, to progressively regulate the predicted moment locations.

Weakly Supervised Methods. Since the annotation process for temporal boundaries of retrieved moments is labor-intensive and costly, some studies resort to address the TSGV problem with only the video-level descriptions available for training [10, 13, 29, 39, 40]. This kind of setting is dubbed as weakly supervised TSGV. TGA [29] learns a joint embedding network to align the text and video features, where the global visual features are obtained by weighted pooling according to the text-guided attentions. Duan et al. [10] established a cycle system that consists of the weakly supervised localization task and its dual problem (i.e., weakly supervised dense event captioning) and minimized the reconstruction error for training such a loop system. Huang et al. [22] presented a **cross-sentence relations mining (CRM)** method that explores the cross-sentence relations in the multi-sentence paragraph to improve the per-sentence grounding accuracy.

2.2 Biases in Temporal Sentence Grounding in Videos

Recently, there are many works that are related to uncovering some forms of biases in TSGV datasets [30, 32, 50, 59]. Otani et al. [32] revealed that the mainstream datasets have latent biases on ground-truth moment locations and current deep models are good at making use of them. Yang et al. [50] stated that it is the location variable that causes the spurious correlation between video moments and predicted scores as a confounder, and they further presented a deconfounded cross-modal matching network to remove the confounding effects of the moment location. However, these works either just point out the problem of dataset biases in TSGV without a solution or design a debiased model without careful thinking about current evaluation protocols.

Moreover, Zhou et al. [59] were devoted to dealing with another kind of bias, i.e., the single-style of annotations. The proposed DeNet with a debiasing mechanism can produce diverse yet plausible predictions. Nan et al. [30] proposed an approach to approximate the latent confounder set distribution based on the theory of causal inference to deconfound selection biases introduced by datasets (e.g., in datasets, it appears more often that a person is holding a vacuum cleaner than a person is repairing a vacuum cleaner). However, these two works [30, 59] cannot resolve the issue of moment annotation distribution biases in TSGV.

Different from the above relevant studies, we not only raise the location bias issue and design new evaluation protocols including re-organized datasets and more reliable metrics, but propose a new debiasing framework from the perspective of causality to resolve the problem as well.

2.3 Biases in Other Tasks

Besides TSGV, the dataset bias issue has been observed and addressed in many other multimedia tasks [1, 2, 9, 16, 31, 41, 51].

In **Visual Question Answering (VQA)**, due to the unbalanced distribution of answers, some models are able to give fairly good answers without understanding the visual contents. Thus, a new data split namely **VQA-CP (under Changing Priors)** [1] that alters the language prior distribution is proposed to evaluate the generalization ability of models. In VQA-CP dataset, the answer distribution for each question type in the test set is different from that in the training set. To avoid exploiting the language biases, some ensemble-based methods including fusion-based approaches [1, 2, 7–9, 15] and adversarial-based approaches [16, 35] have emerged.

However, these debiased VQA methods are not able to give a formal formulation of the bias. CF-VQA [31] creatively revisits the methods above from a causal perspective, formulating the language biases as the direct causal effect of questions on answers, and it further presents a novel counterfactual inference framework. The causality can provide good interpretability and theoretical support for debiasing strategies. Such causality-based debiasing idea has also inspired other fields [41, 51]. Tang et al. [41] proposed an unbiased method from biased training for **Scene Graph Generation (SGG)**. Specifically, after analyzing the causal graph, they attempt to remove the harmful bias by computing Natural Direct Effect with counterfactual causality.

Yang et al. [51] analyzed the hidden cause in image captioning and pointed out the confounder is the pre-training dataset. They further presented DICv1.0 framework with both front-door and back-door adjustment.

3 REVISITING EVALUATION PROTOCOLS

In this section, we perform a deep analysis on the limitations of current evaluation protocols including the benchmark datasets and metrics in Section 3.1. To address such limitations, we propose new and more trustworthy evaluation protocols in Section 3.2.

3.1 Analysis of Current Evaluation Protocols

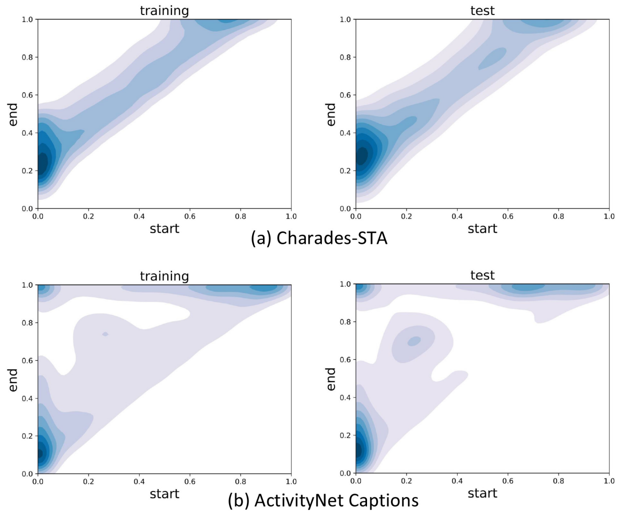
To figure out where the specific biases come from and why the metrics cause unreliable model evaluation, we thoroughly analyze the datasets and metrics that are commonly adopted in TSGV.

3.1.1 Datasets. In TSGV research communities, four public datasets are widely used for evaluation, i.e., **TACoS** [36], **DiDeMo** [20], **Charades-STA** [11] and **ActivityNet Captions** [25]. However, some of them have obvious and inherent shortcomings, e.g., the video scene is restricted into the kitchen domain in TACoS dataset, and the ground-truth moments are comprised of the five-second video segment units in DiDeMo dataset. Therefore, the remaining two datasets (i.e., Charades-STA and ActivityNet Captions) have become the mainstream datasets for TSGV evaluation [5, 17, 49, 53, 55, 57], which are also what we focus on.

Below are more details about Charades-STA and ActivityNet Captions datasets. **Charades-STA** [11] is built upon the original Charades dataset [37], focusing on those videos containing indoor daily activities. Its video length is around 30 seconds on average. The training/test splits are of 12,408/3,720 query-moment pairs. **ActivityNet Captions** [25] is extended from ActivityNet v1.3 dataset [19] for dense event captioning. The videos cover various complex human activities. Each video is annotated with multiple descriptive sentences and their corresponding temporal boundaries of video moments. Since the test split is withheld for the public competition challenge, the two accessible validation sets (i.e., “val 1”, “val 2”) are commonly merged as a test set for the TSGV evaluation. The training/test splits are of 37,421/34,536 query-moment pairs, respectively.

We visualize the joint distribution of normalized start and end points of the ground-truth moments (c.f. Figure 2) in both datasets. An obvious observation is that the distributions of training and test sets for each dataset are almost the same, in other words, these two sets follow the **independent and identical distribution (iid)**. We can also observe that each dataset has its own characteristics of the biased distribution. For Charades-STA, as we can see from Figure 2(a) that the vast majority of ground-truth moments are shorter than 0.5 (after normalization). The fact that the high-density parts concentrate on top-right and bottom-left corners indicates that moments are likely to be either at the beginning/end of the whole videos. For ActivityNet Captions (c.f., Figure 2(b)), there are mainly three types of ground-truth moments appearing more frequently: Short moment samples (≤0.3 after normalization) that start either at the beginning or end of the videos and overlong moment samples that nearly cover the whole length (top-left corner). The main reason for so many overlong samples in ActivityNet Captions is that this dataset is originally created for dense video captioning, which should be annotated with video-level captions. Table 1 shows more detailed statistics about these two datasets. In summary, both of these two datasets have strong biases of the ground-truth moment distribution. A simple baseline method that only exploits such biases may be able to achieve competitive results with SOTA models (c.f., Figure 1).

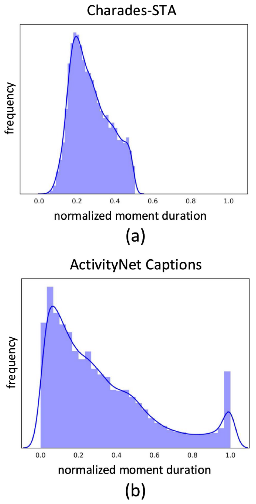
3.1.2 Evaluation Metrics. The commonly used evaluation metric for assessing the moment localization results in TSGV is “R@n,IoU=m” [11]. It measures the percentage of positive samples



(a) Charades-STA

(b) ActivityNet Captions

Fig. 2. The ground-truth moment annotation distributions of all query-moment pairs in Charades-STA and ActivityNet Captions. The deeper the color, the larger density in distributions.



(a)

(b)

Fig. 3. The histogram of the normalized ground-truth moment durations in Charades-STA and ActivityNet Captions.

Table 1. The Detailed Statistics of Datasets, Including the Number of Videos and Query-moment Pairs for Each Data Split

Dataset	Split	# Videos	# Pairs
Charades-STA	training	5,338	12,408
	test	1,334	3,720
ActivityNet Captions	training	10,009	37,421
	test	4,917	34,536
Charades-CD (Ours)	training	4,564	11,071
	val	333	859
	test-iid	333	823
	test-ood	1,442	3,375
ActivityNet-CD (Ours)	training	10,984	51,415
	val	746	3,521
	test-iid	746	3,443
	test-ood	2,450	13,578

out of all testing samples, which is formally defined as:

$$\text{R@}n\text{,IoU=}m = \frac{1}{N_q}\sum_i r(n, m, q_i), \tag{1}$$

where for each query q_i, $r(n, m, q_i) = 1$ if at least one of the top-n predicted moments has an **IoU (Intersection-over-Union)** larger than threshold m with the ground-truth moment, otherwise $r(n, m, q_i) = 0$. The total number of all samples is N_q.

Some existing works [5, 27, 49, 54, 57] report the metric scores with some small IoU thresholds like $m \in \{0.1, 0.3, 0.5\}$. However, such metrics with small IoU thresholds may overrate the model performance when datasets have obvious annotation biases. As shown in Figure 3 (b), for ActivityNet Captions, a substantial proportion of ground-truth moments occupy a long period of video duration. In statistics, 40%, 20%, and 10% of queries refer to a moment occupying over 30%, 50%, and 70% duration of the entire video, respectively. Such annotation biases can increase the chance of hitting the ground-truth moments when IoU thresholds are small. Taking an extreme case as an example, when the IoU threshold is 0.3, if the ground-truth moment is the entire video, any predictions with a duration longer than 0.3 can be seen as positive. Thus, the metric "R@n,IoU=m" with small m is unreliable for current biased annotated datasets.

3.2 New Evaluation Protocols

In order to overcome the shortcomings of current evaluation protocols, we come up with solutions for both the datasets and metrics. As for datasets with obvious annotation biases, we propose to re-organize them, deliberately changing the moment location distribution in the test set. As for unreliable evaluation metrics with small IoU thresholds, we design new metrics to rectify the overrating performance scores.

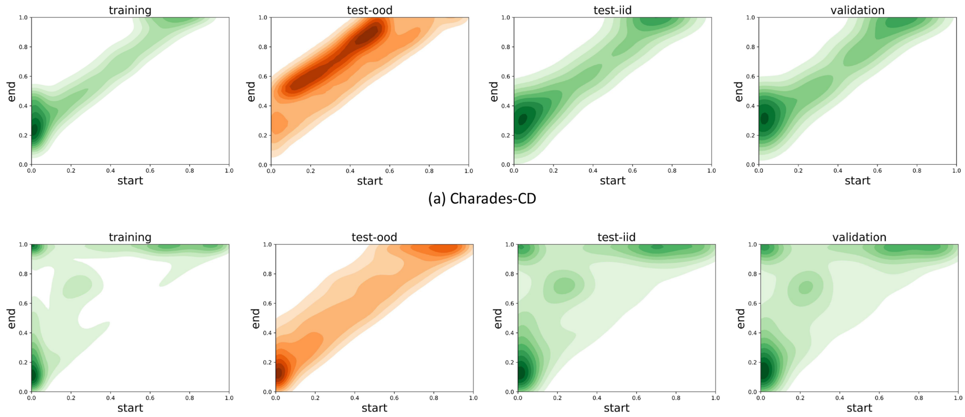
3.2.1 Dataset Re-splitting. We propose to re-organize the two datasets (i.e., Charades-STA and ActivityNet Captions), naming the re-organized ones as **Charades-CD** and **ActivityNet-CD** (CD means **C**hanging **D**istribution), respectively. To be specific, each dataset is re-split into four sets, i.e., **training**, **validation (val)**, **test-iid**, and **test-ood**. We make all samples from the training, val, and test-iid sets follow the independent and identical distribution, and make the samples of test-ood set out-of-distribution. Obviously, the performance gap between the test-iid set and test-ood set can effectively evaluate the generalization capability of the model. The following parts further describe the details during the process of data re-splitting.

Dataset Aggregation and Splitting. For each dataset, we collect all the query-moment pairs (samples) in the training and test sets, and use the Gaussian kernel density estimation to fit the moment annotation distribution as mentioned in Section 3.1 (c.f., Figure 2). Afterwards, we sort all the samples based on their probabilistic density values (from high to low), and take the lowest 20% samples as the preliminary test-ood set since the distribution is furthest different from that of the whole dataset. The remaining 80% samples are divided into the preliminary training set.

Conflicting Video Elimination. Since each video is associated with several sentence queries (samples), it is necessary to ensure that no video simultaneously appears in both the training and test sets. Thus, after obtaining the preliminary test-ood set, we check whether the videos of test-ood samples are also in the preliminary training set. If so, we move all samples (i.e., query-moment pairs) referring to the same video into the split with most of samples. In addition, to avoid the inflating performance of overlong predictions in ActivityNet-CD (c.f., the `PredictAll` baseline in Figure 1), we leave all samples with ground-truth moment occupying over 50% video duration in the training set.

After elimination of all conflicting videos, the final test-ood set occupying around 20% query-moment pairs of the entire dataset is obtained. Then, we randomly split the remaining samples (based on videos) into three groups for the collection of the training, val, and test-iid sets, which occupy around 70%, 5%, and 5% samples, respectively. More detailed statistics of the re-organized datasets should be found in Table 1.

New Split Analysis. Figure 4 depicts the ground-truth moment distributions of these two re-organized datasets. An obvious observation is that the annotation distributions of test-ood set (best expressed in orange) are significantly different from others while the distributions of other three sets (best expressed in green) are similar with those of original training/test splits (c.f., Figure 2). We investigate the difference of the proposed test-ood split for each of these two datasets: (1) For Charades-CD, the start points of the ground-truth moments are distributed more diversely, instead



(a) Charades-CD

(b) ActivityNet-CD

Fig. 4. (a) and (b) illustrate the ground-truth moment annotation distributions of each split in two re-organized datasets.

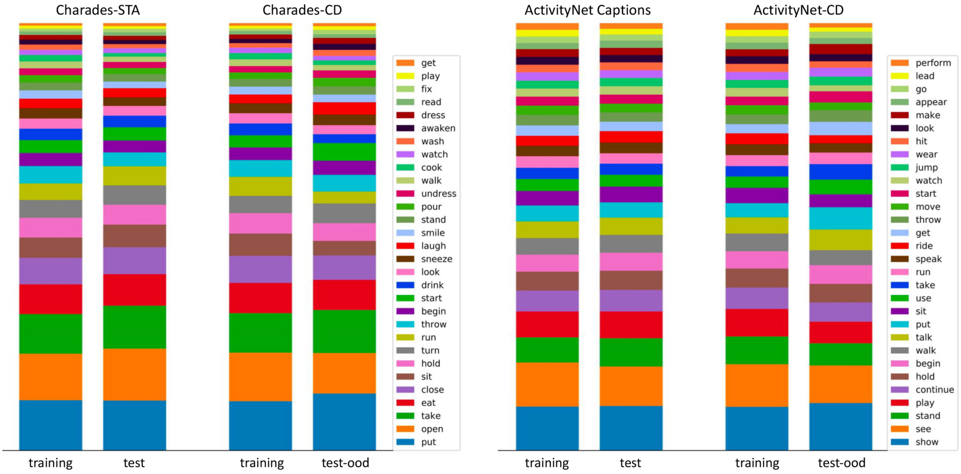
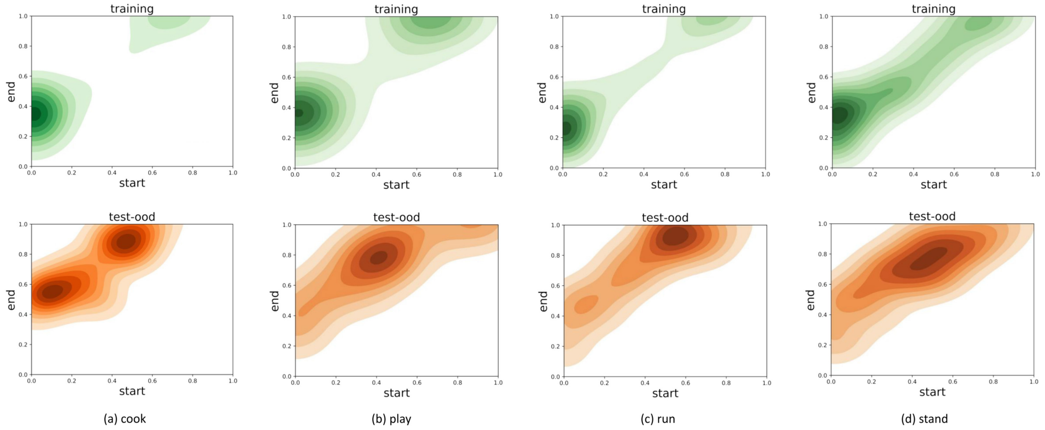


Fig. 5. Top-30 frequent actions in training/test splits for each dataset. The longer the bar, the more frequently the action appears.

of concentrating at the beginning of the videos. (2) For ActivityNet-CD, instead of concentrating in three corners, there are more samples locating in relatively central areas so that models will fail to perform well by merely exploiting the moment distribution biases.

We also investigate the action distribution in each of the original and re-organized datasets. We count the frequency of each verb occurring in the sentence queries of each split, which obviously forms a long-tail distribution. Then the top-30 frequent verbs are shown in Figure 5, with action coverage of 92.7% and 52.9% for Charades-CD and ActivityNet-CD, respectively. We can observe that the action distribution of new test-ood set is still similar with that of either original or re-organized training split for each datasets, which indicates that the OOD comes from each specific verb. As shown in Figure 6, for a given verb, the moment annotations of the training and test-ood sets are of significantly different distributions.



(a) cook (b) play (c) run (d) stand

Fig. 6. The moment annotation distributions of the query-indicated moments which contain a specific action (e.g., *cook*) in the training and test-ood sets of Charades-CD. The deeper the color, the larger the density in the distribution.

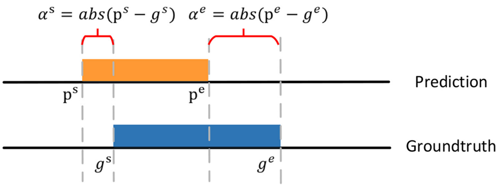


Fig. 7. An illustration of the proposed "dR@*n*,IoU=*m*" metric.

3.2.2 *Proposed Evaluation Metric.* As discussed in Section 3.1.2, the most prevailing evaluation metric — R@*n*,IoU=*m* — is untrustworthy under small threshold *m*. To alleviate this issue, as shown in Figure 7, we propose to calibrate the $r(n, m, q_i)$ value by considering the "temporal distance" between the predicted and ground-truth moments. Specifically, we propose a new metric discounted-R@*n*,IoU=*m*, denoted as "dR@*n*,IoU=*m*":

$$\text{dR@}n\text{,IoU=}m = \frac{1}{N_q}\sum_i r(n, m, q_i) \cdot \alpha_i^s \cdot \alpha_i^e, \tag{2}$$

where $\alpha_i^* = 1 - \text{abs}(p_i^* - g_i^*)$, and $\text{abs}(p_i^* - g_i^*)$ is the absolute distance between the boundaries of predicted and ground-truth moments. Both p_i^* and g_i^* are normalized to the range (0, 1) by dividing the video duration. When the predicted and ground-truth moments are very close to each other, the discount ratio α_i^* will be close to 1, i.e., the new metric can degrade to "R@*n*,IoU=*m*" with exactly accurate predictions. Otherwise, even the IoU threshold condition is met, the score $r(n, m, q_i)$ will still be discounted by α_i^*, which helps to alleviate the inflating recall scores under small IoU thresholds. With the proposed "dR@*n*,IoU=*m*" metric, those speculation methods which over-rely on moments annotation biases (e.g., long moments annotations in ActivityNet Captions) will not perform well.

4 PROPOSED DEBIASING APPROACH

To reduce the effects of moment annotation biases, we further propose a novel debiasing approach. The overall framework is shown in Figure 8. Basically, we add three key components to the base

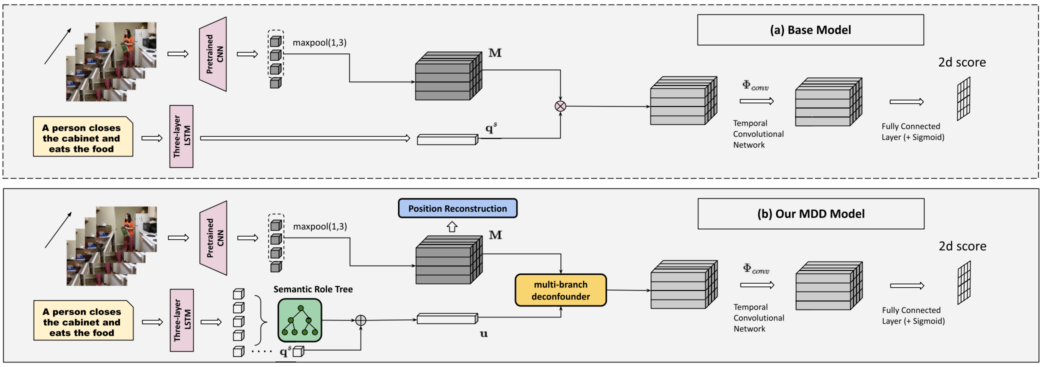


Fig. 8. The overall framework of the Multi-branch Deconfounding Debiasing (MDD) framework. Specifically, (a) briefly shows the pipeline of base 2D-TAN model [57], based on which, we develop three important components indicated by “semantic role tree”, “position recontruction” and “multi-branch decondounder”, yielding our proposed MDD model as shown in (b). More specifically, (i) we enhance the moment representations with position reconstruction module and (ii) parse the query into the semantic role tree to get more fine-grained textual features. (iii) During the multimodal fusion process, we adopt a multi-branch deconfounder to remove the effects caused by multiple confounders.

model for unbiased moment predictions. In this section, we firstly define the TSGV problem and illustrate how the base model works. Afterwards, each of the key components will be described in detail, along with ultimate learning objectives.

4.1 Problem Formulation

As shown in the example of Figure 1, a formal TSGV task takes a sentence query and an untrimmed video as inputs. The untrimmed video can be divided into multiple candidate moments. We let Q denote the sentence query and V denote the candidate video moments. For a proposal-based method which outputs the matching scores between the sentence query and each of candidate moments, a function $\mathcal{F}(Q, V)$ should be learned. The highest output score of the function indicates the best matching query-moment pair.

4.2 Base Model

Due to the superior performance of 2D-TAN [57] in recent public models, we adopt it as the base model for our unbiased temporal sentence grounding. The core idea of 2D-TAN is utilizing a 2D feature map to represent candidate moments of various lengths and locations, where one dimension depicts the start indices of moments and the other one represents the end.

More specifically, as shown in Figure 8(a), for the sentence query, it first embeds the words within the sentence query S via GloVe [34] to obtain the corresponding word vectors, and then the word vectors are fed into a three-layer LSTM [21], where the last hidden state denoted as $\mathbf{q}^s \in \mathbb{R}^{d^h}$ is used to encode the whole query. For the video sequence, it first segments the video into non-overlapping clips, then samples the clips to a fixed size. The features of sampled N^v video clips are extracted by a pre-trained CNN model and projected into the dimension of d^v, which can be denoted as $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{N^v}\}$. The moment feature $\mathbf{m}_{ij}$ $(1 \leq i \leq j \leq N^v)$ out of the 2D feature map $\mathbf{M} \in \mathbb{R}^{N^v \times N^v \times d^v}$ can be obtained by adopting max pooling strategy on clips $\{\mathbf{c}_i, \mathbf{c}_{i+1}, \ldots, \mathbf{c}_j\}$. Afterwards, the 2D feature map $\mathbf{M}$ is fused with the query feature $\mathbf{q}^s$ and fed into a temporal adjacent network to model the temporal relations of moments. Then it passes through a fully connected layer and a Sigmoid function to generate the final 2D matching score map.

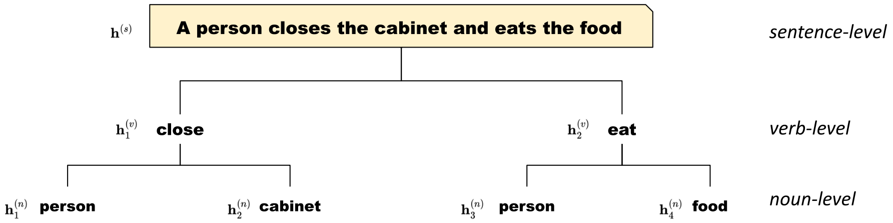

$\mathbf{h}^{(s)}$
A person closes the cabinet and eats the food
sentence-level
$\mathbf{h}_1^{(v)}$ close
$\mathbf{h}_2^{(v)}$ eat
verb-level
$\mathbf{h}_1^{(n)}$ person
$\mathbf{h}_2^{(n)}$ cabinet
$\mathbf{h}_3^{(n)}$ person
$\mathbf{h}_4^{(n)}$ food
noun-level

Fig. 9. The parsed three-layer semantic role tree, where the root represents the whole sentence, and each subtree below the root represents a phrase centered by a verb with relevant objects (nouns) as leaf nodes.

However, the inherent structure of 2D-TAN has natural advantages in exploiting location bias of datasets, since 2D feature map **M** is indexed by moment locations. Therefore, we propose to improve this base model from two aspects. On the one hand, due to the difficulties of semantic alignment between two modalities, the representation capability from each single modality should be enhanced. On the other hand, we attempt to debias the model from perspective of causality as causality-based methods have proven to be successful in debiasing from other fields.

4.3 Improvements on Feature Extraction

In order to improve the representation of both modalities, we propose to perform more detailed and effective feature extraction. Section 4.3.1 depicts the process of extracting more fine-grained query feature, instead of directly involving the global query feature $\mathbf{q}^s$ into subsequent multimodal fusion, a more fine-grained one denoted as $\mathbf{q}^u$ is obtained based on $\mathbf{q}^s$ and the sentence structure. Section 4.3.2 illustrates how to enhance the 2D moment representations **M** via position reconstruction, which aims to make it explicitly associated with the location attribute.

4.3.1 Gated Fine-grained Query Feature. In order to ultimately obtain the fine-grained query feature $\mathbf{q}^u$, one off-the-shelf toolkit [12, 46] is used to parse the sentence into a semantic role tree (c.f., Figure 9). By adopting hierarchical attention mechanism on the tree, we can get the phrase-level features $\{\mathbf{g}_k\}_1^{N_{verb}}$. Then the phrase-level features are aggregated to obtain the final fine-grained sentence representation.

Attention from Sentence-level to Verb-level. More specifically, we first initiate the representation $\mathbf{h} \in \mathbb{R}^{d^h}$ of each node with sequential outputs of the three-layer LSTM. Then we obtain the attention weights of verbs according to the root:

$$\gamma_k^{(v)} = \mathbf{W}_\gamma \left(\tanh\left(\mathbf{W}_{top}\mathbf{h}^{(s)}||\mathbf{W}_{down}\mathbf{h}_k^{(v)}\right)\right), \qquad k = 1, \ldots, N_{verb}, \tag{3}$$

where $\mathbf{W}_{top} \in \mathbb{R}^{d^h \times d^h}$, $\mathbf{W}_{down} \in \mathbb{R}^{d^h \times d^h}$ and $\mathbf{W}_\gamma \in \mathbb{R}^{1\times 2d^h}$ are learnable variables. $\mathbf{h}^{(s)} \in \mathbb{R}^{d^h}$ is the global feature of the whole sentence (i.e., $\mathbf{q}^s$) and $\mathbf{h}_k^{(v)} \in \mathbb{R}^{d^h}$ denotes the feature of the k-th verb node. || implies the concatenation operation.

Attention from Verb-level to Noun-level. Then we aggregate all the verb nodes to obtain the global verb representation $\widetilde{\mathbf{h}}^{(v)} \in \mathbb{R}^{d^h}$ as:

$$\begin{aligned} \widetilde{\mathbf{h}}^{(v)} &= \sum_{k=1}^{N_{verb}} \alpha_k^{(v)} \mathbf{h}_k^{(v)}, \\ \boldsymbol{\alpha}^{(\boldsymbol{v})} &= \mathrm{Softmax}(\boldsymbol{\gamma}^{(\boldsymbol{v})}). \end{aligned} \tag{4}$$

Afterwards, we use a similar attention module to obtain the attention weights of noun (leaf) nodes:

$$\gamma_l^{(n)} = \mathbf{W}_\gamma \left(\tanh\left(\mathbf{W}_{top}\widetilde{\mathbf{h}}^{(v)}||\mathbf{W}_{down}\mathbf{h}_l^{(n)}\right)\right), \qquad l = 1, \ldots, N_{noun}, \tag{5}$$

where $\mathbf{h}_l^{(n)}$ denotes the feature of the l-th noun node. It is worth noting that $\mathbf{W}_\gamma$, $\mathbf{W}_{top}$ and $\mathbf{W}_{down}$ are the sharing parameters with Equation (3).

Phrase-level Features. Then the phrase-level representation of each subtree $\mathbf{g}_k$ can be yielded by aggregating all nodes within the subtree based on the weights:

$$\begin{aligned}\boldsymbol{\beta} &= \text{Softmax}\left(\gamma_k^{(v)}, \gamma_{z_{k,1}}^{(n)}, \gamma_{z_{k,2}}^{(n)}, \ldots, \gamma_{z_{k,nv_k}}^{(n)}\right), \\ \mathbf{g}_k &= \beta_0\mathbf{h}_k^{(v)} + \sum_{j=1}^{nv_k}\beta_j\mathbf{h}_j^{(n)},\end{aligned} \tag{6}$$

where $z_{k,*}$ is a set of indices to enumerate all leaf nodes of subtree k.

Then all the subtree representations are aggregated to obtain the gating signal $\bar{\mathbf{g}}$, and finally the fine-grained sentence feature representation $\mathbf{q}^u$ is obtained as follows:

$$\begin{aligned}\bar{\mathbf{g}} &= \frac{1}{N_{verb}}\sum_{k=1}^{N_{verb}}\mathbf{g}_k, \\ \mathbf{q}^u &= \mathbf{q}^s + \mathbf{q}^s \odot \bar{\mathbf{g}}.\end{aligned} \tag{7}$$

4.3.2 Enhanced Moment Representation via Position Reconstruction. For visual information, in order to better discriminate video moments with unique position information, we attempt to decouple the positional feature from the video moment feature to enhance the moment representation. Specifically, we feed the 2D temporal moment feature map $\mathbf{M} \in \mathbb{R}^{N^v \times N^v \times d^v}$ into a fully-connected layer to obtain the learned 2D position embedding $\mathbf{M}_p \in \mathbb{R}^{N^v \times N^v \times d^p}$, and then we establish a reconstruction loss function to make $\mathbf{M}_p$ close to the 2D position encoding $\mathbf{M}_e \in \mathbb{R}^{N^v \times N^v \times d^p}$:

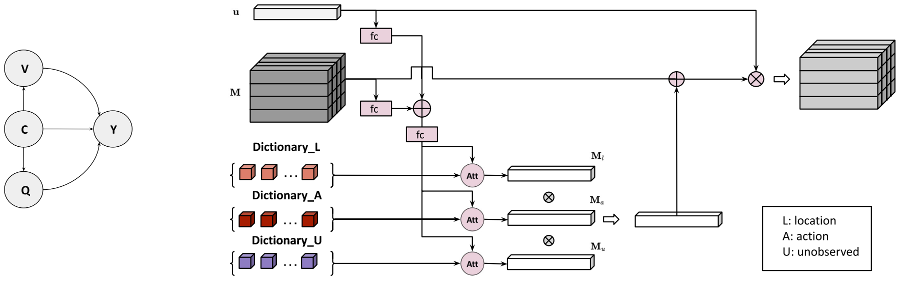
$$\begin{aligned}\mathbf{M}_p &= \tanh(\text{FC}(\mathbf{M})), \\ \mathcal{L}_{recon} &= ||\mathbf{M}_p - \mathbf{M}_e||_2.\end{aligned} \tag{8}$$

Here, $||\cdot||_2$ denotes L2-norm, 2D position encoding $\mathbf{M}_e$ is computed by sine and cosine functions of the different frequencies following [43] and d^p denotes the dimension of positional features.

4.4 Multi-branch Deconfounder

Analysis on Multiple Confounders. Inspired by the work [50], we leverage the structured causal model to analyze the underlying relations among all variables of this TSGV problem. The causal graph which is a **directed acyclic graph (DAG)** is shown in Figure 10(a), where the nodes denote the variables and the directed edges denote the relations between nodes. Q is the variable of query, V denotes the video moment and Y is the variable of predicted matching score. For those traditional TSGV models, they train a model to obtain the probabilities $P(Y|Q, V)$ that is conditioned on Q and V. However, there may exist a confounder C that has connections with both the multimodal inputs (i.e., V and Q) and output scores Y. The confounder is harmful since it causes spurious correlation between the inputs and outputs.

We further investigate the characteristics of TSGV task and find there may exist multiple confounders. Some of the confounders are observable, e.g., the location variable L [50]. Since the location information is naturally encoded in the moment representations while we can also use the moment location distribution priors shown in Figure 4 to perform moment predictions. Moreover, the action variable A could also be the confounder. The activity concepts implicitly exist in



(a) (b)

Fig. 10. (a) is the causal graph for one single confounder. (b) illustrates the pipeline of multi-branch deconfounder, the dictionary of one single confounder is aggregated by the weights based on the fusion of multimodal inputs.

the inputs of video moments and queries while the model could also predict the matching score with only the action label. For example, it can localize on a short moment at the beginning of the video when seeing action "open" based on the action-conditioned moment annotation distribution shown in Figure 6. Besides, some of the confounders (denoted as U) are not observable, such unobserved confounders should also be taken into consideration. Therefore, the do-calculus operation for intervening multiple confounders should be:

$$P(Y|do(Q,V)) = \sum_l P(l) \sum_a P(a) \sum_u P(u) \cdot P(Y|Q,V,l,a,u). \tag{9}$$

Here, we assume that all the confounder variables are independent of each other.

Implementation of Base Model. After obtaining the 2D temporal moment feature $\mathbf{M}$ and gated fine-grained query feature $\mathbf{u}$, the probabilities $P(Y|Q,V)$ without do-calculus can be learned by:

$$P(Y|Q,V) \approx \sigma(\mathbf{W}^T(\Phi_{conv}(\mathbf{u} \odot \mathbf{M}))). \tag{10}$$

Here, the moment features are fused with the broadcasting query feature via Hadamard product. Then such multimodal representations are fed into the temporal convolutional network Φ_{conv}, followed by a fully connected layer with learnable matrix $\mathbf{W}^T$ and the Sigmoid function $\sigma(\cdot)$ to get the final 2D temporal matching scores.

Implementation of Multi-branch Deconfounder. As shown in Figure 10(b), we consider getting three confounders L, A and U intervened as the multi-branch deconfounder. Each confounder is represented by a dictionary of enumerable elements. Specifically, we implement such intervention by adding a weighted embedding of all elements in the dictionary for each query-moment pair. More concretely, we assign the dictionary of location L with the 2D position encodings which is the same as the position reconstruction module (Section 4.3.2), and we initiate the dictionary of action A with the corresponding word embeddings of limited top-frequency action labels. The unobserved confounder U can be represented by learnable dictionary embeddings of a fixed size. In order to get all confounders intervened at the same time, the weighted representations of multiple confounders are subsequently fused by element-wise multiplication to achieve multi-branch de-confounding. $P(Y|do(Q,V))$ can be approximated as:

$$P(Y|do(Q,V)) \approx \sigma(\mathbf{W}^T(\Phi_{conv}(\mathbf{u} \odot (\mathbf{M} + \mathbf{M}_l \odot \mathbf{M}_a \odot \mathbf{M}_u)))), \tag{11}$$

where the effects of multiple confounders are implemented by integrating all the weighted 2D embedding $\mathbf{M}_k \in \mathbb{R}^{N^v \times N^v \times d^v}$, $k \in \{l, a, u\}$, and then adding such integrated embedding to $\mathbf{M}$ (c.f., Figure 10(b)). Each $\mathbf{M}_k$ with $k \in \{l, a, u\}$ denotes the effect of any confounder belonging to $\{L, A, U\}$, which is the weighted average of all elements within the dictionary $\mathbb{E}_k[h_{qv}(k)]$. $\mathbb{E}_k[h_{qv}(k)]$ can be computed with the multi-head attention module [43] whose query is the fusion of $\mathbf{M}$ and $\mathbf{q}^u$. In other words, the attention weight of each element within the dictionary is determined by each query-moment pair. Specifically, $\mathbf{M}_k$ can be defined as:

$$\begin{aligned} \mathbf{M}_k &= \mathbb{E}_k[h_{qv}(k)] = \text{Concat}(\widetilde{\mathbf{A}}_1, \ldots, \widetilde{\mathbf{A}}_H), \\ \widetilde{\mathbf{A}}_i &= \left[\text{Softmax}\left(\frac{\mathbf{Q}_i \mathbf{D}_{ki}^T}{\sqrt{d^H}}\right)\mathbf{D}_{vi}\right] \quad i = 1, \ldots, H, \end{aligned} \tag{12}$$

where H is the head number and $d^H = \frac{d^v}{H}$ is the dimension of each subspace. $\mathbf{D} \in \mathbb{R}^{N^k \times d^v}$ represents the dictionary containing N^k elements. And $\mathbf{D}_k = \mathbf{D}\mathbf{W}_1$, $\mathbf{D}_v = \mathbf{D}\mathbf{W}_2$ with learnable parameters $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d^v \times d^v}$. The query for multi-head attention is $\mathbf{Q} = \text{FC}_q(\text{FC}_u(\mathbf{q}^u) + \text{FC}_m(\mathbf{M}))$, where FC_q, FC_u, FC_m are all the fully connected layers with learnable parameters $\in \mathbb{R}^{d^v \times d^v}$. Note that $\mathbf{Q}$ is flatten to $\mathbb{R}^{L^q \times d^v}$ for subsequent computation, where $L^q = N^v \times N^v$. Then $\mathbf{D}_k$ is equally divided into H parts $\{\mathbf{K}_i\}_1^H \in \mathbb{R}^{N^k \times d^H}$ along the feature dimension, so do $\mathbf{D}_v$ and $\mathbf{Q}$.

4.5 Learning Objectives

Besides the reconstruction loss L_{recon}, we use the scaled ground-truth IoU in [57] as the binary cross entropy loss:

$$\mathcal{L}_{bce} = \sum_{i=1}^{N} y_i \log s_i + (1 - y_i) \log(1 - s_i), \tag{13}$$

where y_i is the scaled IoU score and s_i is the predicted matching score. The final learning objectives are defined as:

$$\mathcal{L} = \mathcal{L}_{bce} + \lambda \mathcal{L}_{recon}, \tag{14}$$

where λ is the hyperparameter.

5 EXPERIMENTS

In this section, we conduct a series of experiments to validate the effectiveness of new evaluation protocols and our proposed debiasing framework.

5.1 Implementation Details

For benchmarking existing methods, we used their open-sourced codes and claimed hyperparameters to train the models with our proposed data splits. The models were validated by the iid set and tested by both the test-iid and test-ood sets. For fair comparisons, we uniformly used pre-trained I3D [4] features for Charades-CD and C3D [42] features for ActivityNet-CD as video encoding. For query encoding, we used GloVe [34] to embed the words.

For our debiasing framework, we followed [57] to use three-layer uni-directional LSTM to sequentially encode the queries and adopted max-pooling strategy for moment feature extraction. For both datasets, all of the hidden sizes (i.e., d^v, d^h and d^p) were set to 512 and the number of sampled clips (i.e., N^v) was set to 16. The head number H was set to 4. The number of stacked convolutional layers for predicting matching scores was set to 4 with kernel size of 5. The sizes for the dictionaries of L, A and U were set to $256 (N^v \times N^v)$, 80 and 80 respectively. During the training process, batch sizes were set to 64 and 32 for Charades-CD and ActivityNet-CD, respectively, and hyperparameter λ was 1. During the inference stage, we set the **non maximum suppression threshold (NMS)** as 0.45. We used Adam optimizer [24] with learning rate of 1e−4.

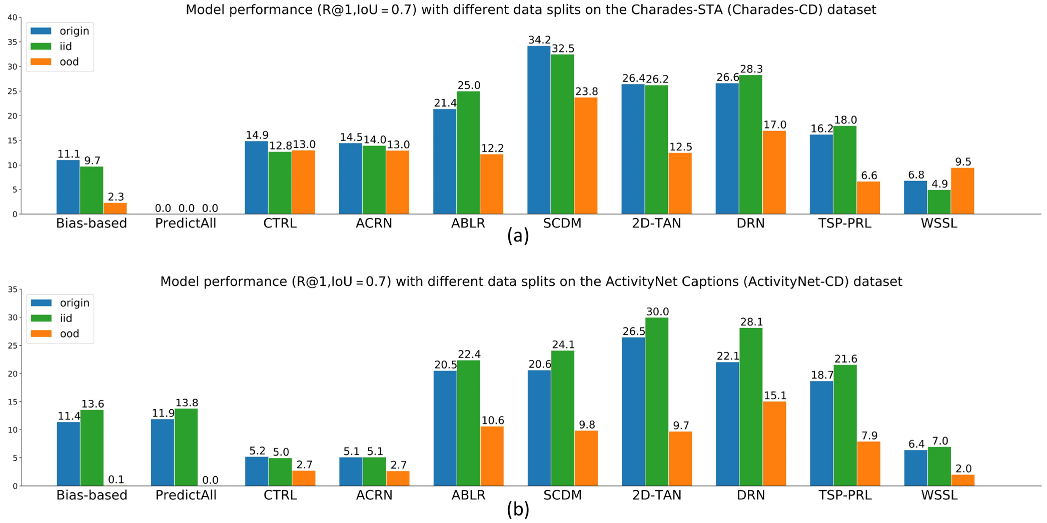


Fig. 11. Performances (%) of SOTA TSGV methods on the test set of original splits (Charades-STA and ActivityNet Captions) and test sets (test-iid and test-ood) of proposed splits (Charades-CD and ActivityNet-CD). We use metric R@1,IoU=0.7 in all cases.

5.2 Performance Comparisons on the Original and Proposed Data Splits

To evaluate the generalization ability of existing methods and demonstrate the difficulty of the newly proposed splits (i.e., Charades-CD and ActivityNet-CD), we compared the performance of two simple baselines and eight representative SOTA methods. In general, we can group all methods into following categories:

- *Non-deep methods*: Non-deep methods contain two simple baselines without training. The first one is the **Bias-based** method, which uses the Gaussian kernel density estimation to fit the moment annotation distribution, and randomly samples several locations based on the fitted distribution as the final moment predictions. The second one is the **PredictAll** method, which directly predicts the whole video as the final moment predictions.
- *Two-Stage methods*: Cross-modal Temporal Regression Localizer (**CTRL**) [11], and Attentive Cross-modal Retrieval Network (**ACRN**) [26].
- *End-to-End methods*: Attention-Based Location Regression (**ABLR**) [54], 2D Temporal Adjacent Network (**2D-TAN**) [57], Semantic Conditioned Dynamic Modulation (**SCDM**) [53], and Dense Regression Network (**DRN**) [55].
- *RL-based method*: Tree-Structured Policy based Progressive Reinforcement Learning (**TSP-PRL**) [45].
- *Weakly-supervised method*: Weakly-Supervised Sentence Localizer (**WSSL**) [10].

We report the performance of all mentioned TSGV methods with metric "R@1,IoU=0.7" in Figure 11. We can observe that almost all methods have a significant performance gap between the test-iid and test-ood sets, i.e., these methods are prone to over-relying on the moment annotation biases, and fail to generalize to the OOD test. Meanwhile, the evaluation results on the original test set and the proposed test-iid set are relatively close, which shows that the moment distribution of the test-iid set is similar to the majority of the whole dataset. More detailed experimental result analyses are provided in the following:

Non-deep Methods. The Bias-based method that only exploits the annotation biases of the training set is apparently unable to perform well after changing the moment annotation distributions in test splits. Statistically, its performance on ActivityNet-CD heavily degrades from 13.6% of the test-iid set to 0.1% of the test-ood set. As for the PredictAll method, since all the ground-truth moments in Charades-CD account for less than 50% range of the whole videos, simply taking the entire video as the prediction will inevitably lead to "R@1,IoU=0.7" of 0.0 on all test splits. The test samples in ActivityNet-CD are much longer, so the PredictAll method can achieve high results of 11.9% and 13.8% on the original test set and new test-iid set, respectively. However, the longer moments are excluded in the test-ood set, thus the performance decreases to 0.0 as well.

Two-Stage Methods. We find that the two-stage methods (i.e., CTRL and ACRN) are less sensitive to the domain gaps between the test-iid and test-ood sets. This is because they utilize a sliding-window strategy to obtain moment candidates, and match these moment candidates with each query individually. In this way, all moment candidates without specific positional attributes are treated equally, thus the moment annotation distribution has less effect on the evaluation results. It is observed that the performance of the test-iid and test-ood sets on Charades-CD are competing while the OOD performance presents a more obvious drop on ActivityNet-CD. The primary reason for this observation is that the moment candidates have more chance to hit the Charades-CD ground-truth moments, which take up a longer percentage of the entire videos (c.f., Figure 4(a)). Despite the less sensitivity against the annotation biases, the performances of these two-stage methods are still far behind those of the SOTA methods from other categories.

End-to-End Methods. For all tested end-to-end methods, we can observe common and significant performance drops on the test-ood set compared to the test-iid set with both two datasets. All of these methods have considerate thoughts about the temporal relations and contextual information of the whole video, since some queries may contain words indicating temporal locations and orders like "begin", "end", "first", "before" and "after", and some of them intend to model the temporal relations between moments. Unfortunately, although our test-ood split does not break any temporal relations, their OOD performances still drop significantly, which demonstrates that current TSGV methods fail to utilize the visual temporal relation or cross-modal interaction.

RL-based Method. The RL-based method (i.e., TSP-PRL) suffers from obvious OOD performance drops on the test-ood set as well. TSP-PRL adopts IoU between current predicted moment and the ground-truth at each step as the training reward, so the temporal annotations can directly affect the learning process. Therefore, the changing of temporal annotation distributions will inevitably cause the model performance degradation.

Weakly-supervised Method. The evaluation results of the weakly-supervised method WSSL are thought-provoking: it achieves higher performance on test-ood set compared to test-iid set in Charades-CD, but results of these splits in ActivityNet-CD are exactly the reverse. One key finding after investigating the grounding results is that the normalized (start, end) moment predictions of both two re-organized datasets converge on a few certain intervals (i.e., (0, 1), (0, 0.5), (0.5, 1)), which indicates that WSSL does not learn the semantic alignment between the videos and sentences at all. It only speculatively guesses several likely locations instead.

5.3 Our Proposed MDD Framework vs. SOTA Methods

We also compare our approach to the above methods with the new metrics "dR@1,IoU={0.5,0.7}". As shown in Table 2, our approach outperforms the base model 2D-TAN with a great gain and has comparable results with another debiasing method TCN-DCM [50].

For Charades-CD dataset, MDD achieves the best results on both iid-0.5 and ood-0.7 (iid/ood-*m* denotes "dR@1,IoU=*m*" for test-iid/ood set). The performance of MDD on iid-0.7 and ood-0.5 is slightly lower than the best with 0.57% and 1.21%, respectively. These observations indicate that the

Table 2. Performance Comparisons with dR@1,IoU=m (%) (The **BOLD** Number Indicates the Best Performance and the UNDERLINE Number Indicates the Second Best One)

	Charades-CD				ActivityNet-CD			
	test-iid		test-ood		test-iid		test-ood	
	m = 0.5	m = 0.7	m = 0.5	m = 0.7	m = 0.5	m = 0.7	m = 0.5	m = 0.7
Bias-based	16.87	9.34	5.04	2.21	19.81	12.27	0.26	0.11
PredictAll	0.00	0.00	0.06	0.00	20.05	12.45	0.00	0.00
CTRL [11]	29.80	11.86	30.73	11.97	11.27	4.29	7.89	2.53
ACRN [26]	31.77	12.93	30.03	11.89	11.57	4.41	7.58	2.48
ABLR [54]	41.13	23.50	31.57	11.38	35.45	20.57	20.88	10.03
2D-TAN [57]	46.48	28.76	28.18	13.73	40.87	28.95	18.86	9.77
SCDM [53]	47.36	30.79	**41.60**	22.22	35.15	22.04	19.14	9.31
DRN [55]	41.91	26.74	30.43	15.91	39.27	25.71	**25.15**	**14.33**
TSP-PRL [45]	35.43	17.01	19.37	6.20	33.93	19.50	16.63	7.43
WSSL [10]	14.06	4.27	23.67	8.27	17.20	6.16	7.17	1.82
TCN-DCM [50]	52.50	**35.28**	40.51	21.02	42.15	29.69	20.86	11.07
MDD (Ours)	**52.78**	34.71	40.39	**22.70**	**43.63**	**31.44**	20.80	11.66

Table 3. Effectiveness of Each Component in Our Proposed MDD on ActivityNet-CD with Metrics of dR@1,IoU=m (%) (EM: Enhanced Modalities, MC: Multi-branch Confounder ($avg_L * avg_U * avg_U$)

	w/old metric				w/new metric			
	test-iid		test-ood		test-iid		test-ood	
	m = 0.5	m = 0.7	m = 0.5	m = 0.7	m = 0.5	m = 0.7	m = 0.5	m = 0.7
base	46.35	31.25	21.36	10.37	40.87	28.95	18.86	9.77
base + EM	47.89	32.94	22.75	11.73	42.48	30.69	20.35	11.08
base + EM + MC	49.03	33.72	23.19	12.33	43.63	31.44	20.8	11.66

enhancement of textual and visual features and the causal intervention strategy via multi-branch deconfounder can effectively improve the performance and increase the robustness of moment prediction.

For ActivityNet-CD, the absolute gain of MDD against the base model (e.g., 2.76%/2.49% on iid-0.5/iid-0.7) is not as significant as Charades-CD (e.g., 6.30%/5.95% on iid-0.5/iid-0.7) since ActivityNet-CD is more challenging with diverse actions and complex scenarios. But MDD obviously surpasses all methods with test-iid set and get competitive results with test-ood set, which demonstrates that the fine-grained extraction module can better capture the relations of different objects within the queries, and the reconstruction module can obtain more discriminative moment features for further cross-modal matching. Notably the DRN model has achieved significantly great results on the test-ood set of ActivityNet-CD. One possible reason is that ActivityNet-CD has longer video lengths than Charades-CD, and the dense regression network is much more useful for the dataset of more sparse annotated positive frames. But its policy of densely regarding the frames within the groundtruth moment as positive can still make the model get influenced by the biased groundtruth moment annotations.

5.4 Ablation Studies for MDD Framework

5.4.1 Model Component Analysis. We investigate the effects of each component in our proposed MDD model, including the modality enhancement module and causality-based multi-branch deconfounder module. As shown in Table 3, the **base** model is implemented by 2D-TAN, and a visible gain can be observed in the **base + EM** model after improving the representations of two modalities as described in Section 4.3, since the modality enhancement operation does

Table 4. Performance Comparisons of Different Combinations of Confounders on Charades-CD with Metrics of dR@1,IoU=m (%)

	w/old metric				w/new metric			
	test-iid		test-ood		test-iid		test-ood	
	m = 0.5	m = 0.7	m = 0.5	m = 0.7	m = 0.5	m = 0.7	m = 0.5	m = 0.7
base	50.67	30.38	31.58	14.68	46.48	28.76	28.18	13.73
MDD-avg_L	54.56	36.70	44.22	23.28	50.44	34.81	39.64	21.78
MDD-avg_A	55.89	36.45	45.70	23.34	51.50	34.57	**40.85**	21.77
MDD-avg_U	57.11	35.72	43.95	22.60	52.68	33.98	39.37	21.16
MDD-$avg_L * avg_A$	57.11	36.57	42.76	21.56	52.75	34.79	38.26	20.16
MDD-$avg_L * avg_U$	56.38	**37.42**	44.07	22.42	52.18	**35.57**	39.51	21.01
MDD-$avg_A * avg_U$	56.38	37.30	44.66	24.08	51.91	35.39	39.99	22.50
MDD-$avg_L * avg_A * avg_U$	56.50	36.09	42.44	21.03	52.02	34.19	37.96	19.70
MDD-$avg_L * avg_U * avg_U$	**57.23**	36.57	**45.08**	**24.32**	**52.78**	34.71	40.39	**22.7**

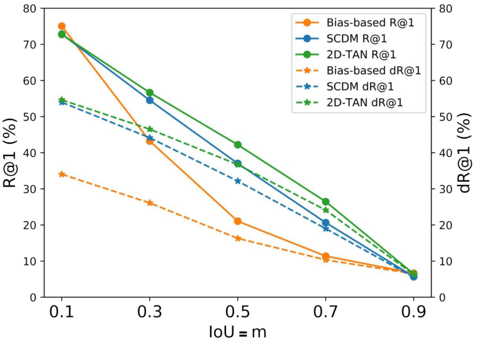


Fig. 12. Performance (%) comparisons of SOTA TSGV methods between original metric ("R@1,IoU=m") and proposed metric ("dR@1,IoU=m"). All results come from the test set of ActivityNet Captions.

enhance the representation power. And the **base + EM + MC** model which further includes the multi-branch confounder (Section 4.4) yields more improvement based on the **base + EM** model, proving the effectiveness of intervention of multiple confounders.

5.4.2 Analysis on Multi-Branch Deconfounder. As shown in Table 4, we further explore the impacts of different combinations of confounders to the model performance on the Charades-CD dataset. For example, MDD-$avg_L * avg_A * avg_U$ denotes the multi-branch deconfounder with combining three confounders including location L, action A and unobserved variable U. Firstly, we consider using only one variable as the confounder. It can be observed that the performance of MDD-avg_L is close to that of MDD-avg_A, and both of them can surpass the base model with a large gap. This observation demonstrates that introducing the intervention with any confounder (i.e., location, action, unobserved variables) can benefit the model and reduce the influence of the location bias. Then we attempt to increase the number of confounders and the performance gets higher as the amount increases. After many trials we find that the best case to introduce external intervention for unbiased temporal sentence grounding is using the combination of one location variable and two unobserved variables (i.e., MDD-$avg_L * avg_U * avg_U$) as multiple confounders.

5.5 Performance Gap Between R@1,IoU=m and dR@1,IoU=m

Figure 12 shows the performance gap between the old and new metric. When the IoU threshold is small, "dR@1,IoU@m" is much lower than "R@1,IoU@m", and the gap between them gradually decreases with the increase of IoU threshold. In other words, the performance scores can get discounted by the new metric more heavily under small IoU thresholds, which is able to avoid unreliable evaluation results. For example, it is observed that the simple bias-based method can beat some SOTA methods under the old metric with small IoU threshold of 0.1, but it cannot outperform others under the new one. This observation further proves the value brought by the proposal of new metrics.

These results further indicate that recall values under small IoU thresholds are untrustworthy and overrated. Although some moment predictions reach the IoU threshold, they still have a great discrepancy to the ground-truth moments. Instead, our proposed "dR@n,IoU=m" metric can discount the recall value based on the temporal distances between the predicted and ground-truth moment. When the moment prediction meets the larger IoU requirements, the discount effect will be weakened, i.e., the "dR@n,IoU=m" values and "R@n,IoU=m" values will be closer to each other. Therefore, our proposed "dR@n,IoU=m" metric is more stable on different IoU thresholds, and it can suppress some inflating results (such as Bias-based or PredictAll baselines) caused by the moment annotation biases in the datasets. Moreover, the results also reveal that it is more reliable to report the localization accuracy with large IoU thresholds.

6 CONCLUSION

In this paper, we take a closer look at mainstream benchmark datasets for temporal sentence grounding in videos and finds that there exists significant annotation bias, resulting in highly untrustworthy results for evaluating model performance. Therefore, we propose to re-split the datasets so that the location distribution of moment annotation in the training and test sets are different. To alleviate the inflating performance evaluation that is caused by biased datasets as well, we design a new metric to discount the scores considering the temporal distances. The re-organized datasets with the new metric can better monitor current research progress of TSGV.

In addition, we design a new debiasing framework to reduce the negative effect caused by the biases from two perspectives: one is to strengthen representations of two modalities, which makes the model easier to learn the semantic alignment between two modalities, and the other is to perform debiasing based on causality, which can both provide good theoretical support and achieve effective debiasing. Experiments show that the newly proposed approach can outperform the base model with a great gap and the evaluation results are also competitive with those of other SOTA models, laying a solid foundation for future research work. In the future, we will explore more debiasing strategies to increase the generalizability of the TSGV model in both data-level and model-level. We will also consider applying our benchmark design and debiasing strategy to other multimedia applications with untrustworthy benchmarks.

REFERENCES

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4971–4980.

[2] Rémi Cadène, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019. RUBi: Reducing unimodal biases for visual question answering. In *Proceedings of the International Conference on Neural Information Processing Systems*. 839–850.

[3] Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. 2021. On pursuit of designing multi-modal transformer for video grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[4] João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4724–4733.

[5] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 162–171.

[6] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. 2020. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 10551–10558.

[7] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10797–10806.

[8] Long Chen, Yuhang Zheng, Yulei Niu, Hanwang Zhang, and Jun Xiao. 2021. Counterfactual samples synthesizing and training for robust visual question answering. *ArXiv preprint* abs/2110.01013 (2021).

[9] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 4069–4082.

[10] Xuguang Duan, Wen-bing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. In *Proceedings of the International Conference on Neural Information Processing Systems*. 3063–3073.

[11] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5277–5285.

[12] Junyu Gao and Changsheng Xu. 2021. Fast video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[13] Mingfei Gao, Larry Davis, Richard Socher, and Caiming Xiong. 2019. WSLLN:Weakly supervised natural language localization networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1481–1487.

[14] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. 2019. MAC: Mining activity concepts for language-based temporal localization. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 245–253.

[15] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 878–892.

[16] Gabriel Grand and Yonatan Belinkov. 2019. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*. 1–13.

[17] Meera Hahn, Asim Kadav, James M. Rehg, and Hans Peter Graf. 2020. Tripping through time: Efficient localization of activities in videos. In *The British Machine Vision Conference (BMVC)*.

[18] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. 2019. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8393–8400.

[19] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A Large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 961–970.

[20] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5804–5813.

[21] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997), 1735–1780.

[22] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. 2021. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7199–7208.

[23] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. 2019. Cross-modal video moment retrieval with spatial and language-temporal attention. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. 217–225.

[24] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations, ICLR*.

[25] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 706–715.

[26] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 15–24.

[27] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal moment localization in videos. In *Proceedings of the ACM International Conference on Multimedia*. 843–851.

[28] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. 2019. DEBUG: A dense bottom-up grounding approach for natural language video localization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 5144–5153.

[29] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K. Roy-Chowdhury. 2019. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11592–11601.

[30] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. 2021. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2765–2775.

[31] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual VQA: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12700–12710.

[32] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. 2020. Uncovering hidden challenges in query-based video moment retrieval. In *The British Machine Vision Conference (BMVC)*.

[33] Yu-xin Peng, Wen-wu Zhu, Yao Zhao, Chang-sheng Xu, Qing-ming Huang, Han-qing Lu, Qing-hua Zheng, Tie-jun Huang, and Wen Gao. 2017. Cross-media analysis and reasoning: Advances and directions. *Frontiers of Information Technology & Electronic Engineering* (2017), 44–57.

[34] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

[35] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Proceedings of the International Conference on Neural Information Processing Systems*. 1548–1558.

[36] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics* 1 (2013), 25–36.

[37] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 510–526.

[38] Xiaomeng Song and Yahong Han. 2018. VAL: Visual-attention action localizer. In *Pacific Rim Conference on Multimedia*.

[39] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. 2020. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *ArXiv preprint* abs/2003.07048 (2020).

[40] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A. Plummer. 2019. wMAN: Weakly-supervised moment alignment network for text-based video segment retrieval. In *arXiv*.

[41] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3713–3722.

[42] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4489–4497.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the International Conference on Neural Information Processing Systems*. 5998–6008.

[44] Weining Wang, Yan Huang, and Liang Wang. 2019. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 334–343.

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

[46] Ziyue Wu, Junyu Gao, Shucheng Huang, and Changsheng Xu. 2021. Diving into the relations: Leveraging semantic and visual structures for video moment retrieval. In *IEEE International Conference on Multimedia and Expo*. 1–6.

[47] Shaoning Xiao, Long Chen, Jian Shao, Yueting Zhuang, and Jun Xiao. 2021. Natural language video localization with learnable moment proposals. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[48] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2986–2994.

[49] Huijuan Xu, Kun He, Bryan A. Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9062–9069.

[50] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–10.

[51] Xu Yang, Hanwang Zhang, and Jianfei Cai. 2021. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

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

[53] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *Proceedings of the International Conference on Neural Information Processing Systems*. 534–544.

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

[55] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10284–10293.

[56] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S. Davis. 2019. MAN: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1247–1257.

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

[58] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 655–664.

[59] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. 2021. Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8445–8454.

[60] Wenwu Zhu, Peng Cui, Zhi Wang, and Gang Hua. 2015. Multimedia big data computing. *IEEE Multimedia* (2015).

Received 23 February 2022; revised 3 July 2022; accepted 3 August 2022