Learning Sample Importance for Cross-Scenario Video Temporal Grounding

Peijun Bao
peijunbao@pku.edu.cn
Wangxuan Institute of Computer Technology, Peking University, Beijing, China
Nanyang Technological University, Singapore

Yadong Mu*
myd@pku.edu.cn
Wangxuan Institute of Computer Technology, Peking University, Beijing, China

ABSTRACT

The task of temporal grounding aims to locate video moment in an untrimmed video, with a given sentence query. This paper for the first time investigates some superficial biases that are specific to the temporal grounding task, and proposes a novel targeted solution. Most alarmingly, we observe that existing temporal ground models heavily rely on some biases (*e.g.*, high preference on frequent concepts or certain temporal intervals) in the visual modal. This leads to inferior performance when generalizing the model in cross-scenario test setting. To this end, we propose a novel method called Debiased Temporal Language Localizer (Debias-TLL) to prevent the model from naively memorizing the biases and enforce it to ground the query sentence based on true inter-modal relationship. Debias-TLL simultaneously trains two models. By our design, a large discrepancy of these two models' predictions when judging a sample reveals higher probability of being a biased sample. Harnessing the informative discrepancy, we devise a data re-weighing scheme for mitigating the data biases. We evaluate the proposed model in cross-scenario temporal grounding, where the train / test data are heterogeneously sourced. Experiments show large-margin superiority of the proposed method in comparison with state-of-the-art competitors.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**; **Visual content-based indexing and retrieval**.

KEYWORDS

Deep neural network, visual temporal grounding, debias

ACM Reference Format:
Peijun Bao and Yadong Mu. 2022. Learning Sample Importance for Cross-Scenario Video Temporal Grounding. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3512527.3531403

1 INTRODUCTION

Given a sentence query and an untrimmed video, the goal of temporal grounding [1, 6] is to localize video moment described by the sentence query. In recent years, a list of promising models [7, 9, 15, 21, 23, 25, 26, 28, 30, 31] have been designed to tackle this task. Despite remarkable research progress, we empirically find that these models are heavily affected by some superficial bias of the data, leading to inferior generalization performance on cross-scenario testing data. In one of our pilot experiments, we take some well-trained state-of-the-art temporal grounding model and zero the feature vector of all testing queries. This boils down to using only the visual information in the temporal grounding. Surprisingly, the performance under such a setting is comparable to many models that normally read both queries and videos during testing. It also significantly outstrips random guess.

To make the point more clear, let us provide some concrete empirical observations. We have identified two sorts of dominant biases that a model can exploit for over-fitting the training scenarios, namely the visual content bias and temporal interval bias. In detail, a few visual concepts and temporal intervals are more frequently queried by the sentence than others in the training dataset. Although the issue of dominant biases have been previously reported as language bias in visual-question answering tasks [3, 5, 20], the preference of visual contents and temporal intervals is specific to the temporal grounding task and our report here is the first. For instance, the concept "run" is largely frequent than "sit" in the benchmark of ActivityNet. Likewise, certain temporal intervals are more likely to be grounded. As illustrated in Figure 1, the interval $[0.20, 0.60]$ statistically has more annotations than $[0.60, 0.80]$. If above biases were sufficiently strong, a fully uni-modal input (such as zeroing query's features) can still achieve good performance in this multi-modal task. However, when generalizing the learned model into other unseen scenarios, these superficial biases between video moments and ground-truth may disappear, which adversely impacts the cross-scenario performance.

To this end, we propose a novel method called Debiased Temporal Language Localizer (Debias-TLL) to prevent the model from naively learning the video moment bias and enforce it to ground sentence in the video. Our key idea is to simultaneously train two twined models, with one of them aiming to learn video moment bias from the data and further to debias the other model. The models have an identical backbone. One of them reads only the video input, and the other normally has access to the full video-query input. The first model is designed to learn the video moment bias and predict the localization results only from visual modality. As illustrated in Figure 2, we then use the prediction of the first model to reweigh

[*] Corresponding author.

ICMR '22, June 27–30, 2022, Newark, NJ, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9238-9/22/06…$15.00
https://doi.org/10.1145/3512527.3531403

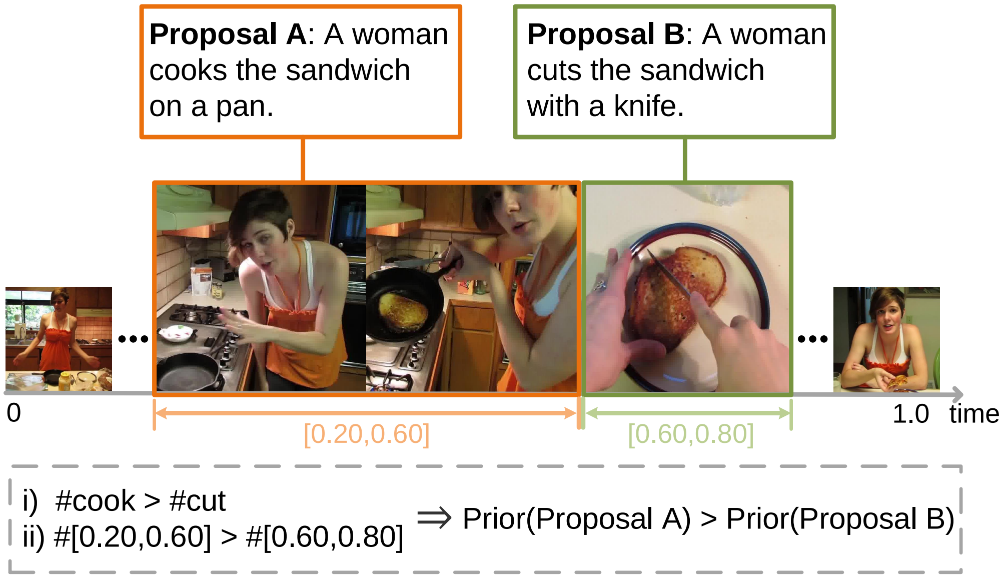

Proposal A: A woman cooks the sandwich on a pan.
Proposal B: A woman cuts the sandwich with a knife.
0
1.0 time
[0.20,0.60]
[0.60,0.80]
i) #cook > #cut
ii) #[0.20,0.60] > #[0.60,0.80]
⇒ Prior(Proposal A) > Prior(Proposal B)

Figure 1: Illustration of video uni-modal bias in ActivityNet. Even without knowing the sentence query, the prior of video moment proposal A to be grounded is larger than B, mainly due to two facts in the annotations: i) the visual concept "cook" contained in A are much more frequently queried than the concept "cut" in B; ii) the temporal interval $[0.2, 0.6]$ of A appears more frequently annotated than B. # denotes the frequency.

the importance of training samples for the second model and adjust the loss function accordingly. During this process, those training samples with high probability to being biased is suppressed. In this way, the training data is adaptively re-weighed in order to mitigate video moment bias in the second model. At the inference stage, we drop the first model and only use the second one for final prediction.

Note that the weakness of video modality biased model cannot be reflected by existing standard evaluation process because the training and testing data share a similar distribution of video moment correlation. To fairly evaluate the model, we propose a novel cross-scenario setting for the video temporal grounding task. In specific, we conduct the training and evaluation processes across two data distributions where video moment correlation cannot transfer from one data distribution to another. Under such settings, a model which makes prediction utilizing video moment bias would fail to perform well on the testing data.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to investigate the video moment bias in the video temporal grounding task, which adversely affects the generalization ability of the model. Two specific sorts of video moment biases in the data (visual content bias and temporal interval bias) are studied.
- We propose a novel two-model-based methods to re-weigh the training data via learning sample importance and delineate the video moment bias for a temporal grounding model.
- A cross-scenario evaluation setting is proposed to reveal the weakness of video-moment biased temporal grounding. Our proposed method beats the state-of-the-art competitors with a clear margin under the cross-scenario settings.

2 RELATED WORK

2.1 Temporal Grounding

The task of temporal grounding in video is recently introduced by [1, 6]. It aims to determine the start and end time of the video moment described by a sentence query. A moment context network is proposed in [1] to effectively localize sentence query in the video by integrating local and global video features over time. And Gao [6] devises a cross-modal localizer to regress action boundary for candidate video clips. A semantic matching reinforcement learning framework is developed in [27] to select the frame sequence and associate the sentence query with video content in a matching-based manner. [17, 18] leverage multi-head attention mechanism to retrieve the crucial part of visual features or query contents.

Several recent works [2, 16, 21, 22, 25, 29, 30] propose to capture long-range semantic dependencies in video context and closely integrate cross-modal representation through graph convolution or non-local modules. And [23, 31, 32] further exploit syntactic structure of natural language queries and decompose the sentence query as multiple components for fine-grained temporal reasoning.

2.2 Unbiased Cross-Modal Understanding

Some recent works [3, 5, 8, 20] study language bias in visual question answering (VQA) caused by answer prior. [5] proposes a model-agnostic counterfactual samples-synthesizing training scheme to reduce the language biases, which generates numerous counterfactual training samples by masking critical objects in images or words

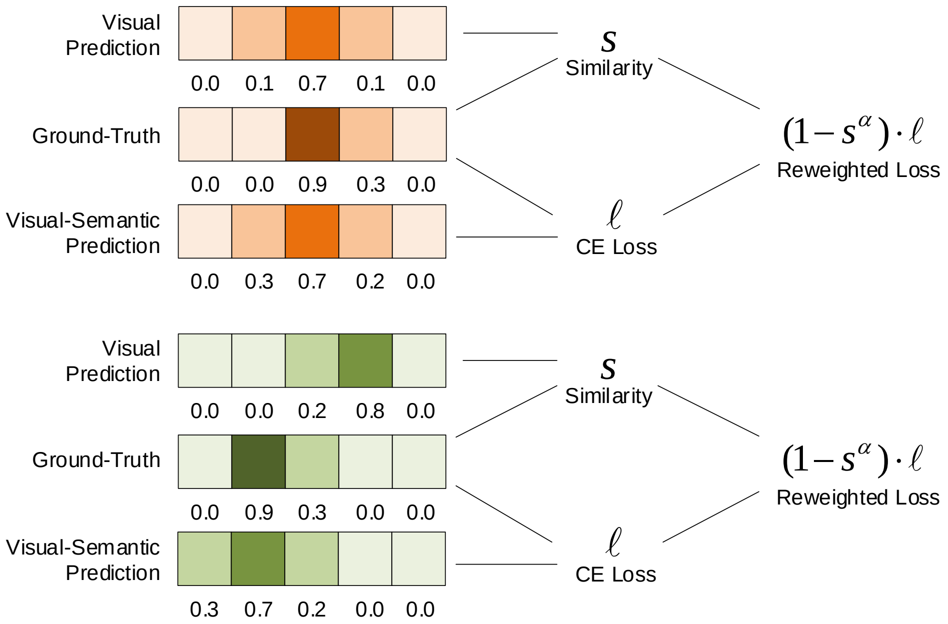


Figure 2: Sample importance reweighing. The visual localizer adjusts the loss function for the visual-semantic localizer, *i.e.*, suppressing the importance of training sample with high relevance to video moment bias (upper figure) and augmenting the weight of the irrelevant one (lower figure).

in questions. [20] introduces a question-only model, and then pose training as an adversarial game which discourages the VQA model from capturing language biases in its question.

Recently, [24] studies the predicate bias on the task of scene graph generation from image. And [19] investigates the dialog history bias in the visual dialog and proposes two causal principles for improving the quality of visual dialog. To the best of our knowledge, the video moment bias is specific to the temporal grounding task and is never explored before.

3 METHODS

3.1 Problem Formulation

Given an untrimmed video V and a sentence description S, the goal of temporal grounding is to localize temporal moments T described by the sentence. More specifically, the video is presented as a sequence of frames $V = \{v_i\}_{i=1}^{L_V}$ where v_i is the feature of i-th frame and L_V is the frame number of the video. The sentence description S is presented as $S = \{s_i\}_{i=1}^{L_S}$ where s_i represents i-th word in the sentence and L_S denotes the total number of words. The temporal moment T is defined by the start and end time points of the moment in the video.

3.2 Debiased Temporal Language Localizer

As illustrated in Figure 3, our proposed model consists of four main components: a language encoder, a video encoder, a visual-semantic localizer, and a visual localizer. This section will elaborate on the details of each component.

3.2.1 Video Encoder. Given an untrimmed video, the goal of the video encoder is to extract visual representations of video moment proposals from the raw input frames. We first segment the input video into small clips. Each of these clips consists of T frames. A fixed-interval sampling is performed to obtain N video clips. Then we apply a pretrained I3D Network [4] for each sampled video clip to extract a sequence of spatio-temporal features $V = \{v_i\}_{i=1}^{N}$.

Based on these I3D features, then we can generate visual feature embeddings for moment proposals. In more detail, to get the visual feature embedding for a moment proposal (a, b) with start point at a and end point at b, a boundary-matching (BM) operation [14] is applied over all I3D features covered by this proposal:

$$\tilde{f}^{V_{ab}} = \mathrm{BM}(\{v_i\}_{i=a}^{b}). \tag{1}$$

Through a series of convolutional operations and bilinear sampling, the boundary-matching operation can generate proposal-level features from basic clip-level features. More implementation details of the boundary-matching operation are omitted here and can be found in [14]. Finally, we pass the proposal-level feature $\tilde{f}^{V_{ab}}$ through a fully-connected layer to obtain the final feature embedding $f^{V_{ab}} \in \mathbb{R}^{d^V}$ for the moment proposal (a, b). Note that the final visual feature $f^{V_{ab}}$ extracts spatial-temporal patterns from raw video frames and summarizes the visual information for the moment proposal.

3.2.2 Language Encoder. To effectively retrieve the video moment of interest described by the natural language, we exploit a language encoder to extract the language feature embedding f^S from the input sentence query S. Instead of encoding each word with a one-hot vector or learning word embeddings from scratch, we

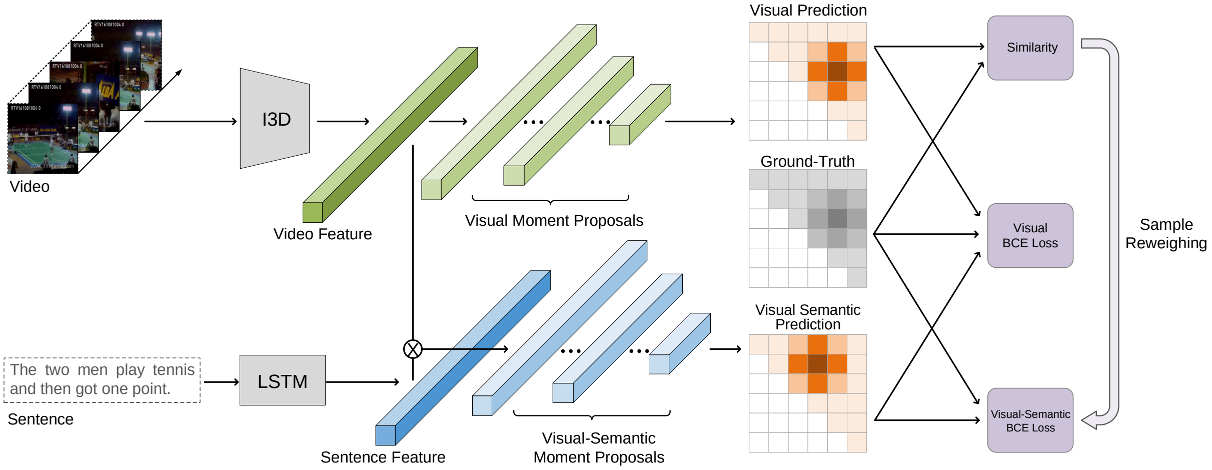


Figure 3: Our proposed model consists of four main components: a language encoder, a video encoder, a visual localizer, and a visual-semantic localizer. The visual localizer learns the video moment bias from the data with a single input of video modality. Then it adjusts the loss function for the visual-semantic localizer, *i.e.*, suppressing the importance of training sample with high relevance to video moment biases.

leverage word embeddings pretrained on large-scale datasets of text documents. In more specific, we firstly encode each word s_i in S into Glove word embedding [11] as w_i. Then we feed the sequence of word embedding $\{w_i\}_{i=1}^{L_S}$ to an LSTM [10] and pass the last hidden state of LSTM to a single fully-connected layer. The feature embeddings for the sentence query is then extracted as $f^S \in \mathbb{R}^{d^S}$.

3.2.3 Visual Localizer and Visual-Semantic Localizer. We design two twined models i.e. visual localizer and visual-semantic localizer, with the visual localizer aiming to learn video moment bias from the data and further to debias the visual-semantic localizer. The visual localizer reads only the video input, and the visual-semantic localizer normally has access to the full video-query input.

The visual-semantic localizer first constructs visual-semantic features of moment proposals for the sentence query and then localizes the described moments. In specific, we firstly compute the video moment feature $f^{V_{ab}}$ for all possible moment proposals (i.e. $1 \leq a \leq b \leq N$) according to Eq. (1). Then the features of the visual modality and language modality are fused to construct visual-semantic features for each moment proposals. To interact the language feature f^S with video moment feature $f^{V_{ab}}$, f^S is first multiplied with video moment clip feature $f^{V_{ab}}$. The fused feature $\hat{M}_{ab}$ is then normalized with its $\mathcal{L}_2$ norm to obtain the final visual-semantic feature M_{ab}, formulated as

$$\begin{aligned} \hat{M}_{ab} &= f^{V_{ab}} \odot f^S, \\ M_{ab} &= \hat{M}_{ab} / ||\hat{M}_{ab}||_2, \end{aligned} \tag{2}$$

where $\odot$ is the element-wise product.

Then the visual-semantic features $\{M_{ab}\}$ are passed to a fully-connected layer and a sigmoid layer, which generates the visual-semantic score map $\{p_{ab}\}$. Each value $p_{a,b}$ in the visual-semantic score map denotes the predicted matching score of the temporal moment (a, b) for the sentence query. And the maximum of the score map p corresponds to the final grounding result.

The visual localizer directly guesses the most interested moments based on the visual feature of moment proposals $\{f^{V_{ab}}\}$ without the input of sentence query. Specifically, the video moment features $\{f^{V_{ab}}\}$ are directly passed to a fully-connected layer and sigmoid layer to generate a visual score map $\{p'_{ab}\}$, which represents the predicted prior of video moment (a, b) to be grounded.

3.3 Sample Importance Reweighing

Each training sample consists of an input video V, a sentence query S, and the temporal annotation T associated with the query. In the training phase, based on the temporal annotations, we need to determine which temporal moment in the temporal-sentence score map and visual score map corresponds to the ground-truth result. Instead of using a hard label, a soft label is assigned to each temporal moment based on its overlap with the annotations. In more detail, for each moment proposal in the score map, we first compute the IoU score IoU_{ab} between its temporal boundary (a, b) and the annotation T. Then we can assign a soft ground truth label gt_{ab} to the proposal (a, b):

$$gt_{ab} = \begin{cases} 0 & IoU_{ab} \leq \mu_{min}, \\ \frac{IoU_{ab} - \mu_{min}}{\mu_{max} - \mu_{min}} & \mu_{min} < IoU_{ab} < \mu_{max}, \\ 1 & IoU_{ab} \geq \mu_{max}, \end{cases} \tag{3}$$

where μ_{min} and μ_{max} are two hyperparameters to determine the soft label distribution.

The visual localizer's goal is to learn the video moment bias and predict the localization results only from visual modality. For each training sample, the visual localizer can be trained with a binary cross entropy loss, which is defined as:

$$\mathcal{L}_v = -\sum_{(a,b) \in C} \mathrm{gt}_{ab} \log(p'_{ab}) + (1 - \mathrm{gt}_{ab}) \log(1 - p'_{ab}), \tag{4}$$

where $C = \{(a, b) | 1 \leq a \leq b \leq N\}$ is the set of all valid moment proposal boundaries and p'_{ab} is the prediction output of visual localizer.

To train the visual-semantic localizer, previous works commonly train it with binary cross entropy loss similar to Eq. 4 as

$$\mathcal{L}_{vs} = -\sum_{(a,b)\in C} \mathrm{gt}_{ab}\log(p_{ab}) + (1-\mathrm{gt}_{ab})\log(1-p_{ab}), \quad (5)$$

where p_{ab} is the prediction of visual-semantic localizer.

Due to the superficial bias between video moments and ground-truth, the temporal grounding model trained with the cross entropy loss tends to simply exploit the video modality to make a prediction, rather than jointly understand both video and language as claimed before. To this end, we use the prediction output p'_{ab} of the visual localizer to reweigh the importance of training sample and adjust the loss function $\mathcal{L}_{vs}$ for the the visual-semantic localizer accordingly. Specifically, we first compute the cosine similarity s of visual localizer prediction $p' = \{p'_{ab}\}$ and ground-truth $\mathrm{gt} = \{\mathrm{gt}_{ab}\}$ as

$$s = \frac{p' \cdot \mathrm{gt}}{||p'||_2||\mathrm{gt}||_2} \quad (6)$$

Then the adjusted loss $\mathcal{L}'_{vs}$ for visual-semantic localizer is reweighted by s as follows

$$\mathcal{L}'_{vs} = (1 - s^{\alpha}) \cdot \mathcal{L}_{vs}, \quad (7)$$

where α is a hyper-parameter to control the weight decay. Intuitively, high value of s implies large probability of inferring the ground truth merely from the visual modal, implying tight relevance to the video moment biases. Following this intuition, Eq. 7 suppresses their sample weight.

Last, we define the total loss function for Debias-TLL as:

$$\mathcal{L}_t = \mathcal{L}_v + \mathcal{L}'_{vs}, \quad (8)$$

which consists of the loss $\mathcal{L}_v$ for visual localizer and the adjusted loss $\mathcal{L}'_{vs}$ for visual-semantic localizer.

With the final loss function $\mathcal{L}_t$, Debias-TLL can be trained in an end-to-end manner to mitigate video moment bias. At the inference stage, we drop the visual localizer and only use the visual-semantic localizer.

4 EXPERIMENT

4.1 Dataset

ActivityNet Captions. It consists of 19,209 untrimmed videos with the annotation of sentence description and moment boundary. The contents of the videos are diverse. It is originally built for dense-captioning events [13] and lately introduced for temporal grounding. It is the largest existing dataset in the field of temporal grounding. There are 37,417, 17,505, and 17,031 moment-sentence pairs in the training, validation and testing set, respectively.

Charades-STA. It contains 9,848 videos of daily indoors activities. It is originally designed for action recognition and localization. Gao et al. [6] extend the temporal annotation (*i.e.*, labeling the start and end time of moments) of this dataset with language descriptions and name it as Charades-STA. There are 3,720 moment-sentence pairs in the testing set.

DiDeMo. It was recently proposed in [9], specially for natural language moment retrieval in open-world videos. DiDeMo contains 10,464 videos with 4,021 annotated moment query pairs in the testing set.

4.2 Evaluation Metrics

The evaluation metric for temporal grounding is known to be "Recall@N,IoU=θ ". For each sentence query, we first calculate the Intersection over Union (IoU) between a grounded temporal segment and the ground truth. "Recall@N,IoU=θ " represents the percentage of top N grounded temporal segments that have at least one segment with higher IoU than θ. Following previous works [28, 30], we report the results as $N \in \{1, 5\}$ with $\theta \in \{0.5, 0.7\}$ on ActivityNet Captions, Charades-STA and DiDeMo dataset.

4.3 Baseline Methods

We compare our methods with several state-of-the-art methods listed as followings: **CTRL** [6]: Cross-model Temporal Regression Localizer. **PFGA** [21]: Proposal-free Temporal Moment Localization using Guided Attention. **SCDM** [28]: Semantic Conditioned Dynamic Modulation. **2D-TAN** [30]: 2D Temporal Adjacent Networks. We further consider following methods: **random**: randomly select the moment proposals. **TLL**: the model with identical archetecture to Debias-TLL, but trained by commonly used binary cross entropy loss.

4.4 Implementation Details

We use pretrained CNN [4] as previous methods to extract I3D video features on all datasets. Glove word embeddings [11] pretrained on Common Crawl are utilized to represent words in the sentence. A three layer LSTM is applied to word-embeddings to obtain the sentence representation. We set the channel numbers of sentence feature and video proposal feature d^S, d^V to 512. And the number of sampled clips N is set to 32. For BM operations in the video encoder, the sampling number for the moment proposal is set to 32.

We adopt Adam [12] with learning rate of 1×10^{-4}, a momentum of 0.9 and batch size of 16 for training. The hyperparameter α in Eq. 7 is set to be 1.0. During inference, we choose the moment proposals with the highest confidence score as the final result for the sentence query. If we require to retrieve multiple temporal moments for each sentence query (*i.e.*, for R@5), non-maximum suppression (NMS) with a threshold of 0.4 is applied to remove redundant candidates.

4.5 Analysis of Video Moment Biases

To show the video moment bias in the temporal grounding model, in this experiment, we mask all the words of the sentence input and evaluate existing models with single video input (marked as "video-only") on the ActivityNet Captions. The results are summarized in Table 1, where all of these methods achieve better performance than the random method with large margins. This shows that the model can heavily exploit the superficial correlation between video moment and ground-truth to provide correct localization results.

Here we further identify two sorts of specific biases (visual content bias and temporal interval bias) that the model can exploit to localize the target moment when ignoring the sentence query. Figure 4a presents the frequencies of action concepts in all the sentence queries in both training / testing data of the ActivityNet Captions. The action concept distribution follows a long-tail distribution, and some actions concept are much more frequently queried than others. We further illustrate the top frequency of action concepts in

Table 1: Performance evaluation results on the ActivityNetCap.

Input	Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
	random	13.99	4.69	44.69	17.64
video & query	CTRL	29.01	10.34	59.17	37.54
	PFGA	33.04	19.26	-	-
	SCDM	36.75	19.86	64.99	41.53
	2D-TAN	44.51	26.54	77.13	61.96
	TLL	44.24	27.01	75.22	60.23
video-only	PFGA	21.69	12.56	-	-
	SCDM	23.84	12.93	51.66	32.36
	2D-TAN	27.56	13.93	61.65	36.78
	TLL	28.10	13.96	59.07	36.25

Table 2: Cross-scenario performance of video-only model on the AcNet2Charades and AcNet2DiDeMo.

Dataset	Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
Charades	random	11.88	3.76	46.64	16.88
	video-only	6.68	0.41	56.68	25.55
DiDemo	random	8.36	2.59	37.53	11.79
	video-only	4.87	1.22	38.74	16.54

Fig. 4b, showing that training and testing data share a similar distribution. A similar conclusion also exists in the object concepts in the sentence query. Furthermore, we illustrate the frequency distribution and the top frequency of temporal intervals of the queried moments in Figure 4c and 4d. Like the video content, the temporal intervals also follow long-tail distributions shared by the training and testing data. The temporal information of the moment can be captured by temporal grounding model with temporal context modeling using recurrent neural networks, non-local blocks, etc. This means video moment proposals that contain certain video concepts and temporal intervals are more likely to be localized as positive, and the model can then infer the localization results only according to the video, irrespective of the sentence query.

4.6 Cross-Scenario Evaluation

To quantify the effect of all aforementioned task-specific biases, we evaluate under a cross-scenario setting for temporal grounding in video. The training and evaluation processes are conducted on two different distributions where video moment biases cannot be transferred from one data distribution to another. Specifically, we train the model on the ActivityNet Captions considering its large-scale data amount and the diversity of scenes and activities. Then we test the model on the Charades-STA and DiDeMo (dubbed as AcNet2Charades and AcNet2DiDeMo respectively). As shown in Table 2, under the cross-scenario setting, the "video-only" TLL model fails to perform well on the testing data and is even inferior to the random guess model on the metrics of R1.

4.7 Performance Comparison

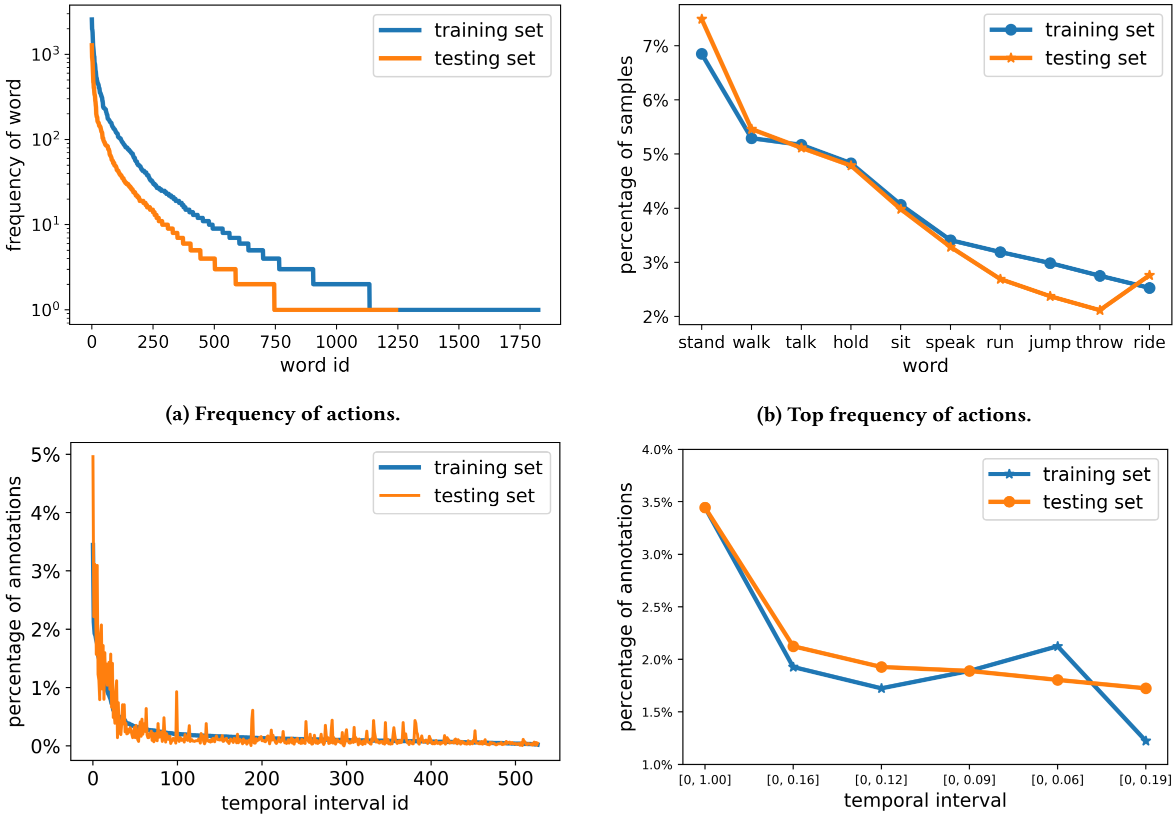
The results of the proposed Debias-TLL and the baselines on ACNet2Charades and ACNet2DiDeMo are summarized in Table 3 and 4 respectively. Our algorithm outperforms all the competing methods with a clear margin. It is noticeable that the proposed technique surpasses the state-of-the-art performances by 8.64% and 4.08% points in terms of R1@0.5 and R1@0.7 metrics, respectively. This verifies the effectiveness of the sample importance reweighing in cross-scenario temporal grounding. The prevailing solutions for temporal

Table 3: Performance evaluation results on the AcNet2Charades.

Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
random	11.88	3.76	46.64	16.88
video-only	6.68	0.41	56.68	25.55
PFGA	5.75	1.53	-	-
SCDM	15.91	6.19	54.04	30.39
2D-TAN	15.81	6.30	59.06	31.53
Debias-TLL	**21.45**	**10.38**	**62.34**	**32.90**

Table 4: Performance evaluation results on the AcNet2DiDeMo.

Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
random	8.36	2.59	37.53	11.79
video-only	4.87	1.22	38.74	16.54
PFGA	6.24	2.01	-	-
SCDM	10.88	4.34	43.30	18.40
2D-TAN	12.50	5.50	44.88	20.73
Debias-TLL	**13.11**	**7.70**	**44.98**	**21.32**



(a) Frequency of actions.

(b) Top frequency of actions.

(c) Frequency of intervals.

(d) Top frequency of intervals.

Figure 4: Video moment bias analysis on ActivityNet Captions

grounding can be grouped into two categories i.e. top-down and bottom-up approaches. We note that the top-down method PFGA achieves inferior results than the top-down methods SCDM and 2D-TAN, which suggests the superiority of top-down design compared to the bottom-up one under the cross-scenario setting. We suspect that this is because the bottom-up approach directly predicts each frame's probabilities as ground-truth interval boundary and more easy to overfit to the temporal intervals bias.

4.8 Ablation Study

Impact of Importance Reweighing. To study the impact of sample importance reweighing in Debias-TLL, we substitute the two-model based adjusted loss Eq. 7 with the commonly used binary cross entropy loss (marked as TLL), and train the model on ActivityNet Captions. The evaluation results on Chrades-STA and DiDeMo are listed in Table 5. Without the sample importance reweighing, the TLL model gets inferior results than Debias-TLL on both datasets, verifying the effectiveness of the proposed technique under the setting of cross-scenario temporal grounding.

Impact of Hyperparameter α. The hyperparameter α plays a key role in controlling the magnitude of sample importance reweighing. And when $\alpha \longrightarrow \infty$, $1 - s^{\alpha}$ approximates 1 and then the adjusted loss Eq. 7 approximates to binary cross entropy loss Eq. 5. An appropriate setting of α requires to rebalance the heavily video moment biased data. Figure 5 illustrates the performance R1@0.7 and R5@0.7 of the Debias-TLL model on AcNet2Charades with α set to 0.25 to 1.0. We found that the performances increase until $\alpha = 1.0$ and then decrease afterward. This shows that $\alpha = 1$ is a proper selection to balance the distribution of training samples and achieve satisfactory results. Note that even when $\alpha = 1.75$, the performance is still much superior to the TTL baseline without sample importance reweighing.

5 CONCLUSION

In this paper, we show that temporal grounding models are heavily affected by video moment bias of the data, limiting the generalization performance on cross-scenario testing data. To prevent the model from naively memorizing the biases and enforce it to ground the query sentence based on true cross-modal understanding, we propose a novel Debiased Temporal Language Localizer with a two-model based data-reweighing mechanism. Experiments show large-margin superiority of the proposed method in comparison with state-of-the-art competitors in cross-scenario temporal grounding.

Acknowledgement: This work is supported by Science and Technology Innovation 2030 - New Generation Artificial Intelligence of

Table 5: Ablation study on AcNet2Charades and AcNet2DiDeMo.

Dataset	Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
Charades	TLL	14.76	6.12	60.41	31.89
	Debias-TLL	**21.45**	**10.38**	**62.34**	**32.90**
DiDeMo	TLL	10.25	5.13	44.73	19.49
	Debias-TLL	**13.11**	**7.70**	**44.98**	**21.32**

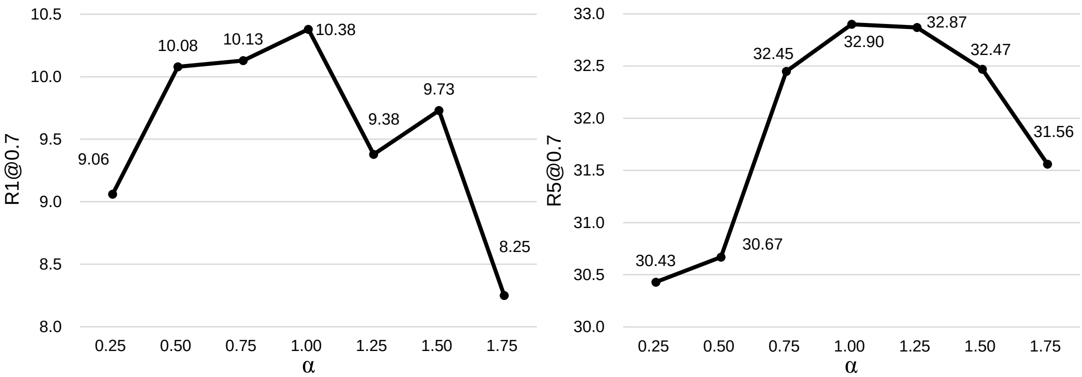


Figure 5: Impact of α on the AcNet2Charades.

China (2020AAA0104401) and Beijing Natural Science Foundation (Z190001).

REFERENCES

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *ICCV*.

[2] Peijun Bao, Qian Zheng, and Yadong Mu. 2021. Dense Events Grounding in Video. In *AAAI*.

[3] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. In *NeurIPS*.

[4] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

[5] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual Samples Synthesizing for Robust Visual Question Answering. (2020).

[6] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *ICCV*.

[7] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. 2019. Excl: Extractive clip localization using natural language descriptions. *arXiv preprint arXiv:1904.02755* (2019).

[8] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. 2021. Greedy gradient ensemble for robust visual question answering. In *ICCV*.

[9] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing moments in video with temporal language. In *EMNLP*.

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997).

[11] RichardSocher Jeffrey Pennington and ChristopherD Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

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

[13] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *ICCV*.

[14] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*.

[15] Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. 2018. Temporal modular networks for retrieving complex compositional activities in videos. In *ECCV*.

[16] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021. Context-aware Biaffine Localizing Network for Temporal Sentence Grounding. In *CVPR*.

[17] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *SIGIR*.

[18] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal moment localization in videos. In *ACM MM*.

[19] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two Causal Principles for Improving Visual Dialog. In *CVPR*.

[20] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *NeurIPS*.

[21] Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Basura Fernando, Hongdong Li, and Stephen Gould. 2021. DORi: Discovering Object Relationships for Moment Localization of a Natural Language Query in a Video. In *WACV*.

[22] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. 2021. VLG-Net: Video-language graph matching network for video grounding. In *ICCV*.

[23] Jonathan C Stroud, Ryan McCaffrey, Rada Mihalcea, Jia Deng, and Olga Russakovsky. 2019. Compositional Temporal Visual Grounding of Natural Language Event Descriptions. *arXiv preprint arXiv:1912.02256* (2019).

[24] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *CVPR*.

[25] Jingwen Wang, Lin Ma, and Wenhao Jiang. 2020. Temporally Grounding Language Queries in Videos by Contextual Boundary-aware Prediction. In *AAAI*.

[26] Weining Wang, Yan Huang, and Liang Wang. 2019. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*.

[27] Weining Wang, Yan Huang, and Liang Wang. 2019. Language-driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model. In *CVPR*.

[28] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In *NeurIPS*.

[29] 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 *CVPR*.

[30] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *AAAI*.

[31] Songyang Zhang, Jinsong Su, and Jiebo Luo. 2019. Exploiting Temporal Relationships in Video Moment Localization with Natural Language. In *ACM MM*.

[32] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*.