Synthesizing Counterfactual Samples for Overcoming Moment Biases in Temporal Video Grounding

Mingliang Zhai[1], Chuanhao Li[1], Chenchen Jing[1], and Yuwei Wu[1,2(✉)]

[1] Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing, China
[2] Shenzhen MSU-BIT University, Shenzhen, China
wuyuwei@bit.edu.cn

Abstract. Moment bias is a critical issue in temporal video grounding (TVG), where models often exploit superficial correlations between language queries and moment locations as shortcuts to predict temporal boundaries. In this paper, we propose a model-agnostic counterfactual samples synthesizing method to overcome moment biases by endowing TVG models with sensitivity to linguistic and visual variations. The models with sensitivity sufficiently utilize linguistic information and focus on important video clips rather than fixed patterns, therefore are not dominated by moment biases. Specifically, we synthesize counterfactual samples by masking important words in queries or deleting important frames in videos for training TVG models. During training, we penalize the model if it makes similar predictions on counterfactual samples and original samples to encourage the model to perceive linguistic and visual variations. Experiment results on two datasets (*i.e.*, Charades-CD and ActivityNet-CD) demonstrate the effectiveness of our method.

Keywords: Moment biases · Counterfactual samples · Temporal video grounding

1 Introduction

Temporal video grounding (TVG) is to locate the moment that best matches a language query in an untrimmed video. Given a query "person they start to take some medicine with a spoon" and a corresponding video, models are required to locate a temporal boundary in the video that best matches the query as shown in the original sample of Fig. 1. Recent work [11,14,20] reveals that most TVG models rely on superficial correlations (*i.e.*, moment biases) between language queries and moment locations to infer the temporal boundary. A TVG model that is dominated by moment biases usually utilizes fixed patterns to infer temporal boundaries and is insensitive to linguistic and visual variations. We consider that making TVG models sensitive to linguistic and visual variations can alleviate the influences of moment biases on the models.

S. Yu et al. (Eds.): PRCV 2022, LNCS 13534, pp. 436–448, 2022.
https://doi.org/10.1007/978-3-031-18907-4_34

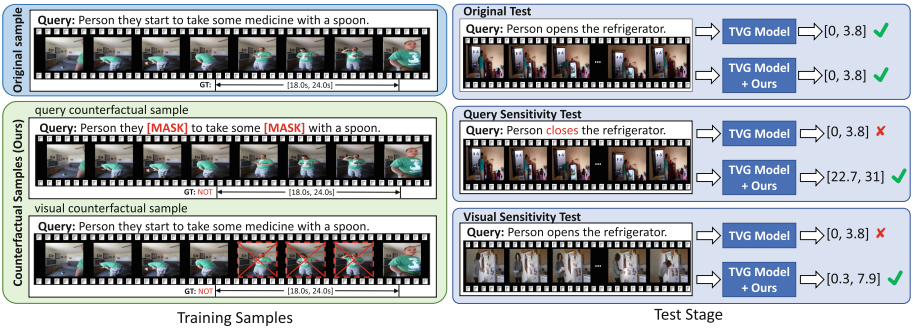

Fig. 1. We synthesize counterfactual samples to train TVG models for endowing them with two sensitivities. (a) Query sensitivity: the model should be sensitive to linguistic variations (*e.g.*, after replacing the important word "opens" with "closes", the predicted boundaries of two queries should be different). (b) Visual sensitivity: the model should be sensitive to the visual content variations.

In this paper, we propose a model-agnostic method to synthesize counterfactual samples by masking important words or deleting important frames, for alleviating moment biases in TVG. Our method serves as a plug-and-play component to endow various types of the TVG model with sensitivity, including proposal-based methods or proposal-free methods. The important words/frames refer to the word/frame that has high contributions to infer boundaries. As shown in Fig. 1, our method consists of two different types of sample synthesizing strategies. For each original training sample, we synthesize a query counterfactual sample and a visual counterfactual sample, both of which consist of a counterfactual Query-Video (QV) pair and corresponding boundaries. By training with synthesized samples, TVG models are encouraged to perceive boundary changes caused by masking words or deleting frames, thus being sensitive to linguistic and visual variations. However, assigning new boundaries to counterfactual QV pairs is non-trivial, because the moment matching the language query of the counterfactual QV pair may not exist. To this end, we introduce a difference maximization (DM) loss to maximize the differences between the model's predicted boundaries for counterfactual QV pairs and the ground-truth boundaries of original samples, which avoid assigning pseudo boundaries for counterfactual QV pairs. The idea behind the DM loss is to provide the counterfactual sample with the boundary difference from the original sample. Extensive experiments on the Charades-CD and ActivityNet-CD datasets demonstrate the effectiveness of our method.

The main contribution of this paper can be summarized as follows: (1) We propose a model-agnostic method, which synthesizes counterfactual samples to make TVG models sensitive to language queries and video moments for overcoming moment biases. (2) We introduce a difference maximization loss that maximizes the differences between the predicted boundaries for counterfactual

QV pairs and the ground-truth boundaries of original samples, to avoid assigning pseudo boundaries for counterfactual QV pairs.

2 Related Work

2.1 Temporal Video Grounding

Given an untrimmed video and a language query, temporal video grounding aims to locate the start and end time of the video segment that best matches the given query. Existing supervised methods can be mainly categorized into two groups: **(1) Proposal-based methods** [1,5,8,10] localize the target segment via generating video segment proposals. They use a boundary predictor to compute a score for each proposal. Ideally, a proposal gets a higher score if it is closer to the ground-truth moment. Then the proposal with the highest score is selected as the boundary. Candidate proposals are obtained by using temporal sliding windows or an anchor-based strategy. If the proposals are generated by an anchor-based strategy, the score is computed based on the multi-modal snippet feature sequence by applying multi-scale anchors in the boundary predictor. **(2) Proposal-free methods** [2,3,6,9] do not generate proposals. They use a regressor or a span predictor as a boundary predictor. Specifically, the regression-based predictor aims to regress the start and end timestamps after interacting the whole video with the query without pre-defining proposals. Existing models achieve promising performance, but they may suffer from moment biases. In contrast, we propose a model-agnostic counterfactual sample synthesizing method to alleviate the influences of moment biases on TVG models.

2.2 Moment Biases

Due to the uneven distribution of the dataset, the model relies on the superficial correlation (*i.e.*, moment biases) between query and moment annotations when making predictions. Recent work tries to overcome the influence of moment biases. Yuan *et al.* [11] first propose that the TVG model usually captures moment biases, and it is difficult to accurately evaluate the level of the model with existing datasets and metrics, so they propose new metrics and benchmarks to accurately evaluate the model. Zhang *et al.* [13] exploit a video-only branch and a query-only branch to capture the distributional bias of video and query, respectively, forcing the model to learn cross-modal interaction information. Liu *et al.* [12] first align the given video-query pair by a cross-modal graph convolutional network, and then utilize a memory module to record the cross-modal shared semantic features in the domain-specific persistent memory. Yang *et al.* [14] disentangle moment representations with location factors to infer crucial features of visual content and then apply to intervene causally on the disentangled multi-modal inputs based on back-door adjustment, which forces the model to fairly incorporate each possible location of the target into consideration. These methods focus on creating delicate models to directly reduce the

influence of moment biases. Differently, we synthesize counterfactual samples to endow TVG models with sensitivity to linguistic and visual variations without changing the network structure for overcoming moment biases.

2.3 Counterfactual Sample Synthesis

In vision-and-language, there are some counterfactual sample synthesis methods for improving the robustness of models. Zhang *et al.* [23] propose counterfactual contrastive learning paradigm to build contrastive training between positive and negative samples in weakly-supervised temporal video grounding and obtain negative samples by perturbing the feature of the feature layer, interaction layer, and relation layer. Hirota *et al.* [24] investigate the effectiveness of text representations for image understanding in VQA. In particular, they delves into the use of synthesized samples on language-only representations including counterfactual samples. Chen *et al.* [25] train the VQA model using a counterfactual sample synthesis training scheme to reduce language biases. These methods validate that counterfactual sample synthesis methods can improve the robustness of vision-and-language models. Unlike them, we are the first to explore the effectiveness of counterfactual sample synthesis for reducing moment biases in TVG.

3 Methodology

3.1 Problem Formulation

Given an untrimmed video $V = \{v_i\}_{i=0}^{n_v-1}$, where v_i denotes i-th frame in a video and n_v is the total number of frames, and a sentence $S = \{s_i\}_{i=0}^{n_s-1}$ as a language query, where s_i denotes i-th word among the sentence, and n_s is the number of words. $[t^s, t^e]$ denotes ground-truth moment. The video V is encoded into visual features $\mathbf{V} = \{\boldsymbol{v}_i\}_{i=0}^{n_v-1} \in \mathbb{R}^{n_v \times d_v}$ with a pre-trained feature extractor. The query Q is encoded into query features $\mathbf{Q} = \{\boldsymbol{w}_i\}_{i=0}^{n_q-1} \in \mathbb{R}^{n_q \times d_q}$ with a pre-trained model. The TVG task aims to localize the start and end timestamps $[t^s, t^e]$ of a specific segment in video $\mathbf{V}$, which refers to the corresponding semantic of query $\mathbf{Q}$.

3.2 Preliminaries

We utilize Grad-CAM [15] to obtain the contribution of each object to the model's prediction. Specifically, the network obtains the feature layer A and the predicted value y through forwarding propagation. We back-propagate y to obtain the gradient A' of the feature layer A, and then calculate $\alpha_k = \frac{1}{N}\sum_k A'^k$ to get the importance of A^k, where k denotes the index of the channel. Finally, we calculate $L_c = \mathrm{ReLU}(\sum_k \alpha_k A^k)$ to derive the contribution of each participant.

3.3 Synthesizing Counterfactual Samples

We propose a method for synthesizing counterfactual samples for training. This method consists of two different types of sample synthesis strategies. As shown in Fig. 2, for an original QV pair $< V, Q >$ organized by a video and a query, we synthesize a visual counterfactual sample $< V^*, Q >$ and a query counterfactual sample $< V, Q^* >$ by directly modifying features respectively. First of all, we use pre-trained models to extract the feature of the video and the query. Secondly, the model's predictions are obtained through forwarding propagation. Thirdly, we use Grad-CAM to obtain features of counterfactual samples by back-propagation without updating parameters. When the V or Q changes, the boundary should also be changed accordingly usually. However, it is hard to know what the changed boundaries should be, so we use the loss function to reflect the change of the boundary (*i.e.*, the value of the loss function is inversely proportional to the prediction accuracy). Finally, we feed synthesized sample features into TVG models which can encourage the model to understand language queries and visual content sufficiently for overcoming moment biases in TVG.

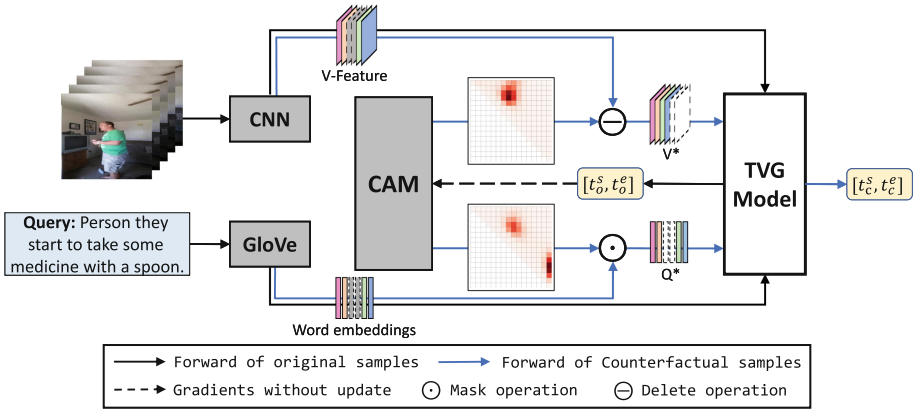

Fig. 2. Overall the TVG model with our method. Given a QV pair, we first encode their features and feed these features into the TVG model to obtain the predicted boundary. Then, we utilize Grad-CAM to obtain attention maps of the target layer. For obtaining counterfactual queries/videos, we use the attention maps to mask/delete the feature of the queries/videos. At last, we use counterfactual samples to train the TVG model.

3.3.1 Query Counterfactual Samples

We use a pre-trained model to obtain query features on which subsequent operations are based. The detailed processes of synthesizing query counterfactual samples are divided into two parts:

Calculating and Masking Important Words in the Query. For the model's prediction, a few words in a query can have a significant influence on the model's decision. We utilize Grad-CAM to derive the contribution of each object for results, and calculate the contribution heat map $I_c^q = [c_i]_{i=0}^{n_q-1}$ of each word to the result by

$$c_i(b, w_i) = C_i(P_{tvg}(b|V, Q), w_i) := \sum(\bigtriangledown_{w_i} P_{tvg}(b|V, Q)), \tag{1}$$

where $P_{tvg}(b|V, Q)$ is the predicted probability of boundary b given a video V and a query Q, and w_i is i-th word. The contributions of word w_i to boundary b are larger if the contribution value $c_i(b, w_i)$ is higher. We calculate $a^q = \text{Softmax}(I_c^q)$ and top-k word with the highest contribution whose

$$\boldsymbol{d}^q = \text{Top}_k(\text{argsort}[a_i^q]) \in \mathbb{R}^{k\times 1}, s.t. \sum_{i=0}^{k} \boldsymbol{d}^q \geqslant \epsilon, \tag{2}$$

then mask these word embeddings as

$$Q^* = Q \odot \boldsymbol{d}^q, \tag{3}$$

where $\odot$ is hadamard product. So far, we use Q^* and original video features V to form query counterfactual pairs $< V, Q^* >$.

Assigning Moment Annotations. We should assign the moment annotations for the query counterfactual sample $< V, Q^* >$. It is difficult to know the location of the boundary after the query has changed. So we design a difference maximization loss function $\mathcal{L}_{DM}$ to reflect the change of the boundary. In other words, the closer the model prediction result is to the original ground-truth boundary B, the greater the loss value, and vice versa. Specifically, we propose a difference maximization (DM) loss to maximize the differences between the model's predicted boundaries for counterfactual QV pairs and the ground-truth boundaries of original samples by

$$\mathcal{L}_{DM} = \frac{1}{N} \sum_{i}^{N} y_i \log(1 - P_i(b)) + (1 - y_i) \log(P_i(b)). \tag{4}$$

3.3.2 Visual Counterfactual Samples

For videos, we use a pre-trained model to encode the video. Similar to Eq. 1, we calculate the important frame by

$$c_i(b, v_i) = C_i(P_{TVG}(b|V, Q), v_i) := \sum(\bigtriangledown_{v_i} P_{TVG}(b|V, Q)). \tag{5}$$

Also, we use Grad-CAM to get the contribution image $I_c^v = [c_i]_{i=0}^{n_v-1}$ of each frame feature, then we calculate $a^v = \text{Softmax}(I_c^v)$, and delete the important frame as a new query V^*. The position where the sum of the top-k values $\boldsymbol{d}^q = \text{Top}_k(\text{argsort}[a_i^q]) \in \mathbb{R}^{k\times 1}$ from the obtained contribution image is greater

than or equal to the threshold ϵ (*i.e.*, $\sum_{i=0}^{k} \boldsymbol{d}^q \geqslant \epsilon$)is the important part. A counterfactual video is synthesized by deleting the original video as

$$V^* = V \ominus \boldsymbol{d}^v = \text{Concat}(\text{del}\{V[\boldsymbol{d}^v]\}, \mathbf{0}_{n_v - k}), \tag{6}$$

where del is delete operation, and $\mathbf{0}_{n_v-k}$ is $n_v - k$ dimension zero vector. We use V^* and original queries V to form visual counterfactual pairs $< V^*, Q >$ by doing so. The processing of moment annotations is the same as in Sect. 3.3.1. Although we can assign a boundary for $< V^*, Q >$, we still utilize $\mathcal{L}_{DM}$ to reflect the change of the boundary for unifying the way of assigning boundaries to two kinds of counterfactual pairs.

3.4 Training Process

By the above, we can get visual counterfactual samples $< V^*, Q >$ and query counterfactual samples $< V, Q^* >$. During the training process, all kinds of samples including original samples both participate in training. Specifically, when training with original samples, we directly utilize

$$\mathcal{L}_{CE} = \frac{1}{N} \sum_{i=1}^{N} y_i \log(P_i(B|Q,V)) + (1 - y_i) \log(1 - P_i(B|Q,V)) \tag{7}$$

to train the TVG model. For training with counterfactual samples, we first utilize $\mathcal{L}_{CE}$ to obtain the gradient of the target layer, thus synthesize counterfactual samples, and use $\mathcal{L}_{DM}$ to train the TVG model. The overall training process is as follows:

1. Training the TVG model using original samples to get the baseline model by minimizing the loss $\mathcal{L}_{CE}$.
2. Calculating gradients of the target layer of the loss $\mathcal{L}_{CE}$ without updating parameters to get a contribution of each object for synthesizing counterfactual samples $< V^*, Q >$ and $< V, Q^* >$.
3. Choosing a sample among the original samples $< V, Q >$, visual counterfactual samples $< V^*, Q >$, and query counterfactual samples $< V, Q^* >$ with the same probability.
4. Training a TVG model using counterfactual samples by minimizing the loss $\mathcal{L}_{DM}$.
5. Going back to step 2 until the model converges or the stopping condition is met.

At the testing stage, we directly predict boundaries through the TVG model without synthesizing counterfactual samples.

4 Experiments

4.1 Datasets and Metric

Datasets. We conduct experiments on Charades-CD and ActivityNet-CD datasets by Yuan *et al.* [11] proposed. These two datasets are reorganized for each split based on the Charades-STA [21] and ActivityNet Caption [22] datasets. Each dataset is re-splitted into four sets: training set, validation set, independent-identical-distribution (IID) test set, and out-of-distribution (OOD) test set. All samples in the training set, validation set, and iid-test set satisfy independently identical distribution, and the samples in the ood-test set are out-of-distribution. Charades-CD has 4,564 videos and 11,071 sample pairs in the training set, 333 videos and 859 sample pairs in the validation set, 333 videos and 823 sample pairs in the iid-test set, 1442 videos and 3375 sample pairs in the ood-test set. ActivityNet-CD has 10,984 videos and 51,414 sample pairs in the training set, 746 videos and 3,521 sample pairs in the validation set, 746 videos and 3,443 sample pairs in the iid-test set, 2,450 videos and 13,578 sample pairs in the ood-test set.

Table 1. Performance on Charades-CD and ActivityNet-CD dataset of different TVG models.

(a) Charades-CD								
Split	Model	dR@1,IoU@m			dR@5,IoU@m			mIoU
		m = 0.5	m = 0.7	m = 0.9	m = 0.5	m = 0.7	m = 0.9	
iid	2D-TAN	35.60	21.39	5.83	75.70	43.62	9.84	35.70
	2D-TAN + ours	**35.27**	**16.88**	**4.19**	**57.04**	**29.92**	**6.05**	**34.71**
	VSLNet	50.30	31.23	9.36	63.79	46.42	15.31	46.99
	VSLNet + ours	**50.06**	**30.62**	**7.65**	**65.01**	**45.20**	**13.61**	**44.98**
ood	2D-TAN	24.91	10.38	2.16	61.83	25.68	3.91	27.99
	2D-TAN + ours	**28.16**	**11.88**	**3.14**	**64.39**	**26.48**	**4.40**	**32.83**
	VSLNet	39.11	21.51	5.45	58.13	39.38	11.73	39.78
	VSLNet + ours	**45.90**	**26.47**	**7.74**	**62.96**	**45.26**	**13.88**	**45.92**
(b) ActivityNet-CD								
Split	Model	dR@1,IoU@m			dR@5,IoU@m			mIoU
		m = 0.5	m = 0.7	m = 0.9	m = 0.5	m = 0.7	m = 0.9	
iid	2D-TAN	34.51	18.03	4.91	65.83	37.55	8.92	34.26
	2D-TAN + ours	**32.12**	**16.10**	**4.15**	**62.62**	**33.94**	**7.90**	**32.09**
	VSLNet	35.91	23.08	9.94	50.62	35.00	13.93	37.68
	VSLNet + ours	**35.10**	**22.50**	**9.35**	**49.32**	**33.31**	**14.19**	**36.95**
ood	2D-TAN	18.90	9.37	2.50	43.54	24.36	5.25	22.72
	2D-TAN + ours	**19.68**	**10.32**	**2.76**	**44.41**	**25.10**	**5.13**	**23.58**
	VSLNet	17.89	9.71	3.16	32.38	18.12	5.22	22.57
	VSLNet + ours	**19.85**	**10.24**	**3.69**	**31.18**	**18.99**	**5.05**	**23.88**

Metric. We use the metric $d\text{R@}n, \text{IoU@}m$ [11]. This metric can better evaluate the performance of models for the current biased datasets. The metric adds a limiting factor to $\text{R@}n, \text{IoU@}m$ denoted as

$$d\text{R@}n, \text{IoU@}m = \frac{1}{N_q} \sum_i r(n, m, q_i) \cdot (1 - abs(p_i^s - g_i^s)) \cdot (1 - abs(p_i^e - g_i^e)), \quad (8)$$

Table 2. The preformance of unimodal models and 2D-TAN on Charades-CD dataset.

Model	dR@1,IoU@m		dR@5,IoU@m		mIoU
	m = 0.5	m = 0.7	m = 0.5	m = 0.7	
2D-TAN	24.91	10.38	61.83	25.68	27.99
2D-TAN (w/o video)	19.31	7.07	61.94	24.11	23.77
2D-TAN (w/o query)	13.11	4.38	43.27	13.66	15.80

where $p_i^{s/e}$ is the start/end time of the model prediction, and $g_i^{s/e}$ is the start/end time of the ground-truth moment. For each query q_i, $r(n, m, q_i) = 1$ if at least one of the top-n predicted moments has an IoU larger than threshold m with the ground-truth boundary, otherwise $r(n, m, q_i) = 0$. The total number of all samples is N_q.

4.2 Implementation Details

We utilize the 300d GloVe [16] vectors to initialize the words in the query. For the video, we use the pre-trained VGG feature [18] for Charades-CD and the C3D feature [17] for ACticityNet-CD. We add our method to 2D-TAN [8] and VSLNet [9]. For hyperparameters, we follow [8,9], and use the last convolutional layer of the feature fusion module as the target layer of Grad-CAM and the threshold ϵ is 0.8. All the experiments are conducted with the Adam optimizer [19] for 20 epochs with learning rate initialized as 5×10^{-4}. We train all models on two GTX 1080ti GPUs with PyTorch1.7.

4.3 Comparisons

The performances of 2D-TAN and VSLNet that are equipped with our method are significantly improved on Charades-CD and ActivityNet-CD ood-splitted. As shown in Table 1, VSLNet with our method obtains 6.14% gains in "mIoU" and 4.96% gains in "dR@1, IoU@0.7". Our method brings 4.84% gains in "mIoU" and 1.5% gains in "dR@1, IoU@0.7" for 2D-TAN.

Our method prevents the model from exploiting moment bias, thus the model performance drops on the iid-test set. On the contrary, the performance of the model on the ood-test set has been significantly improved. We know that Zhang *et al.* [13] also use VSLNet as the baseline, but they replace the stacked LSTMs with stacked transformer blocks, so it is difficult to make a fair comparison. Our improvement on the mIoU metric is better than theirs, and on other metrics is comparable performance.

4.4 Ablation Studies and Analyze

Moment Biases in Baseline Models. We first study the performance of baseline models on the Charades-CD and ActivityNet-CD datasets. We train

Table 3. Ablation studies on Charades-CD dataset. "(w/o Q)" denotes only using visual counterfactual samples for training. "(w/o V)" denotes only using query counterfactual samples for training. Our method adds both of the above samples to the training.

Model	dR@1,IoU@m			dR@5,IoU@m			mIoU
	m = 0.5	m = 0.7	m = 0.9	m = 0.5	m = 0.7	m = 0.9	
2D-TAN	23.19	9.64	2.19	50.24	21.47	3.56	26.69
2D-TAN + ours (w/o Q)	26.60	9.85	2.38	54.42	23.25	3.20	29.73
2D-TAN + ours (w/o V)	26.85	11.47	2.84	54.62	26.16	3.41	30.44
2D-TAN + ours	**28.16**	**11.88**	**3.14**	**54.39**	**26.48**	**4.40**	**31.83**
VSLNet	39.11	21.51	5.45	58.13	39.38	11.73	39.78
VSLNet + ours (w/o Q)	40.12	21.34	5.56	59.99	41.21	11.96	40.22
VSLNet + ours (w/o V)	45.23	25.97	7.59	60.18	45.10	13.06	44.78
VSLNet + ours	**45.90**	**26.47**	**7.74**	**62.96**	**45.26**	**13.88**	**45.92**

2D-TAN, 2D-TAN(w/o query), and 2D-TAN(w/o video) respectively, and the results are summarized in Table 2. It is observed that 2D-TAN(w/o query) has a large gap with 2D-TAN, and 2D-TAN(w/o video) has a small gap with 2D-TAN. For example, under the $mIoU$ metric, 2D-TAN(w/o video) drops by 4.22%, while 2D-TAN(w/o query) drops by 12.19%. According to the results, we conclude that the model relies on moment biases to make predictions.

Improving Sensitivity. We combine our method with two baseline models and test them on the Charades-CD and ACticityNet-CD datasets, and the results are listed in Table 3. All strategies can improve the performance of the baseline model on the ood-test set to a certain extent, which indicates that our method is beneficial to the prediction of the model on the ood-test set. For example, after adding our method to 2D-TAN, the performance of dR@1, IoU@0.7 metric is improved by 2.24%, and the mIoU metric is improved by 5.14%. In addition, the performance improvement of the "w/o V" strategy is better than that of the "w/o Q" strategy.

4.5 Visualizations

In this section, we present some visualizations. As shown in Fig. 3, we can intuitively observe the improvement of TVG models by our method. The figure on the left shows the model gives more accurate predictions for the same sample after adding our method to the model. The right shows the case of moment bias in the training set, which the original 2D-TAN relies on for prediction. Specifically, the boundary corresponding to most sentences containing the word "opens" is in the early 20% of the video moment, and the original 2D-TAN also predicts in the first 20% of the video moment, but this is a wrong boundary. The 2D-TAN using our method overcomes this moment bias and makes correct inferences. In summary, our method not only enables TVG to overcome moment biases but also better generalizes to out-of-distribution samples.

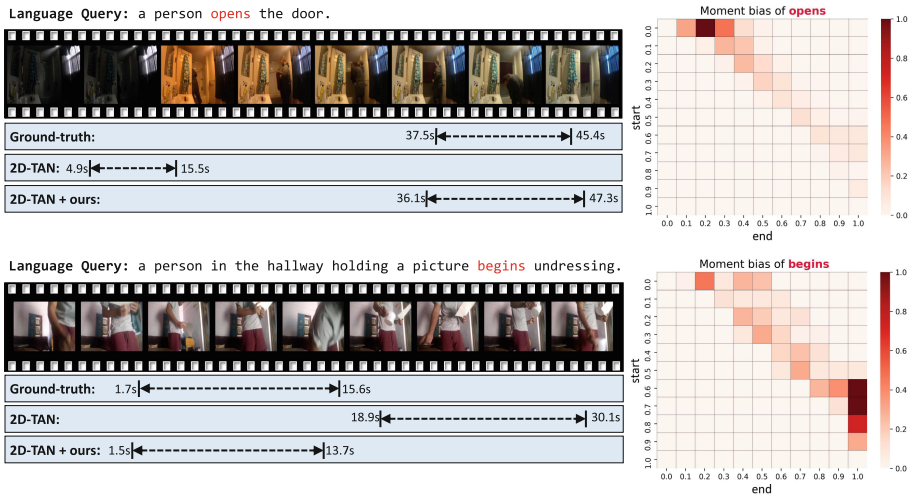

Fig. 3. Visualizations of qualitative examples. The left depicts the localized results of the two models in the ood-test set. The right shows superficial correlations between language queries and moment locations in the training set.

5 Conclusion

In this paper, we have presented a model-agnostic method for synthesizing counterfactual samples. Our method masks/deletes important parts of queries/videos, so that the model can focus on important words/frames rather than several strongly biased words/frames. Furthermore, our method can effectively improve the performance of TVG models and makes the model sensitive to linguistic and visual variations. Extensive experiments demonstrate our method helps overcome moment biases in TVG and improves models' performance on ood-test sets.

Acknowledgements. This work was supported by the Natural Science Foundation of China (NSFC) under Grants No. 62172041 and No. 62176021.

References

1. Wang, Z., Wang, L., Wu, T., et al.: Negative sample matters: a renaissance of metric learning for temporal grounding. arXiv preprint arXiv:2109.04872 (2021)
2. Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10810–10819 (2020)
3. Li, M., Wang, T., Zhang, H., et al.: End-to-End modeling via information tree for one-shot natural language spatial video grounding. arXiv preprint arXiv:2203.08013 (2022)

4. Krishna, R., Hata, K., Ren, F., et al.: Dense-captioning events in videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 706–715 (2017)
5. Liu, D., Qu, X., Zhou, P., et al.: Exploring motion and appearance information for temporal sentence grounding. arXiv preprint arXiv:2201.00457 (2022)
6. Li, J., Xie, J., Qian, L., et al.: Compositional temporal grounding with structured variational cross-graph correspondence learning. arXiv preprint arXiv:2203.13049 (2022)
7. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: temporal activity localization via language query. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5267–5275 (2017)
8. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 12870–12877 (2020)
9. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 6543–6554. https://doi.org/10.18653/v1/2020.acl-main.585
10. Yuan, Y., Ma, L., Wang, J., et al.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
11. Yuan, Y., Lan, X., Wang, X., et al.: A closer look at temporal sentence grounding in videos: dataset and metric. In: Proceedings of the 2nd International Workshop on Human-centric Multimedia Analysis, pp. 13–21 (2021)
12. Liu, D., Qu, X., Di, X., Cheng, Y., Xu Xu, Z., Zhou, P.: Memory-guided semantic learning network for temporal sentence grounding. In: AAAI (2022)
13. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Towards debiasing temporal sentence grounding in video. arXiv preprint arXiv:2111.04321 (2021)
14. Yang, X., Feng, F., Ji, W., Wang, M., Chua, T.-S.: Deconfounded video moment retrieval with causal intervention. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1–10 (2021)
15. Selvaraju, R.R., Cogswell, M., Das, A., et al.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
16. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543 (2014)
17. Tran, D., Bourdev, L., Fergus, R., et al.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR
19. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (2015)
20. Yang, X., Feng, F., Ji, W., et al.: Deconfounded video moment retrieval with causal intervention. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1–10 (2021)
21. Gao, J., Sun, C., Yang, Z., et al.: Tall: temporal activity localization via language query. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5267–5275 (2017)

22. Caba Heilbron, F., Escorcia, V., Ghanem, B., et al.: Activitynet: a large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 961–970 (2015)
23. Zhang, Z., Zhao, Z., Lin, Z., et al.: Counterfactual contrastive learning for weakly-supervised vision-language grounding. Adv. Neural Inf. Process. Syst. **33**, 18123–18134 (2020)
24. Hirota, Y., Garcia, N., Otani, M., et al.: Visual question answering with textual representations for images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3154–3157 (2021)
25. Chen, L., Yan, X., Xiao, J., et al.: Counterfactual samples synthesizing for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10800–10809 (2020)