

-Supplementary Materials- Counterfactually Augmented Event Matching for De-biased Temporal Sentence Grounding

1 The Experimental Details

In this section, we provide more details about our experiments, including the statistics of the four benchmark datasets, adopted evaluation metrics, and experimental settings.

Dataset. We evaluate our proposed CAEM models on the *Charades-CD* and *ActivityNet-CD* [11] datasets, which are repartitioned to evaluate the performance and generalizing ability of TSG models. Moreover, to demonstrate the effectiveness of our proposed CAEM method, we also generalize it on the *Charades-CG* and *ActivityNet-CG* [5] datasets, which contain normal test queries where all words are seen in the training set (denoted as *Trivial*) and generalized test queries with unseen words in the training set (denoted as *Novel*). Overall, we summarize the detailed statistics of the four benchmark datasets in Table 1.

Table 1: Statistics of the evaluated datasets, where N_{train} , N_{eval} , $N_{\text{test-iid}}$, and $N_{\text{test-ood}}$ are the number of events in the train, val, test-iid, and test-ood splits respectively.

Dataset	Domain	OOD Type	N_{train}	N_{eval}	$N_{\text{test-iid}}$	$N_{\text{test-ood}}$
Charades-CD	Indoors	Vision	11071	859	823	3375
ActivityNet-CD	Open	Vision	51415	3521	3443	13578
Charades-CG	Indoors	Text	8281	-	3096	703
ActivityNet-CG	Open	Text	36724	-	3944	15712

Evaluation Metrics. As we mentioned in our formal text, existing D-TSG methods [2, 6, 7, 11] employed different metrics to evaluate their model performance on the D-TSG task. For fair comparisons, we adopt two kinds of evaluation metrics during the testing stage, *i.e.*, the conventional recall metrics and discount recall metrics proposed by [11].

Following the previous methods [2, 6, 7, 9], we first employ the conventional recall metrics: We consider a predicted candidate is correct if it has the IoU greater than a threshold, which is denoted as Recall@Top-k , $\text{IoU} = m$ and abbreviated as $\text{R@}k$, $\text{IoU@}m$, where K is the top range number of ranked generated candidates and m is the threshold. Specifically, for each query q_i , it first calculates the Intersection-over-Union (IoU) between the predicted moment and its groundtruth, and this metric is formally defined as:

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

where $r(k, m, q_i) = 1$ if there is at least one of top- k predicted moments of query q_i having an IoU larger than threshold m , otherwise it equals to 0. N_q is the total number of all queries. In our method, for all four datasets, k is all set to 1, while m is set to $\{0.5, 0.7\}$.

However, in [11], Yuan *et al.* noted that the conventional evaluation metric, *i.e.*, $\text{R@}K, \text{IoU} = m$, is unreliable under small thresholds. To alleviate this issue, they proposed a more challenging metrics termed discounted recall metrics to calibrate the value by considering the "temporal distance" between the predicted and ground-truth moments, which can be formulated as follows:

$$\begin{aligned} \text{dR@}n, \text{IoU@}m &= \frac{1}{N_q} \sum_i r(k, m, q_i) \cdot \alpha_i^s \cdot \alpha_i^e, \\ \alpha_i^s &= 1 - \text{abs}(p_i^s - g_i^s), \alpha_i^e = 1 - \text{abs}(p_i^e - g_i^e), \end{aligned} \quad (2)$$

where $p^{s,e}$ and $g^{s,e}$ represent the starting and ending timestamps of predictions and ground truth. When the predicted and ground-truth moments are very close to each other, the discount ratio $\alpha_i^{s,e}$ will be close to 1, *i.e.*, and the new metric can degrade to conventional recall metrics with exactly accurate predictions. Otherwise, even if the IoU threshold condition is met, the score will still be discounted. It helps to alleviate the inflating recall scores under small IoU thresholds. As these newly proposed metrics could evaluate the model generalization on OOD samples by discounting the normal recall metrics for suppressing the performance of speculation methods that over-rely on moments annotation biases, we follow recent state-of-the-art DTSG methods [4, 9–11] to furtherly test the OOD generalization of our proposed CAEM method.

Experimental Settings. Following the previous methods [7, 9, 11, 12], we adopt the off-the-shelf video features that are extracted by pre-trained 3D CNN backbones [1, 8]. The visual features and textual embeddings are projected into 256 dimensions before sending to vision-language transformers, and the hidden dimension of transformers is also set to 256. As for the hyperparameters, we set the random ratio ϕ in the Temporal Counterfactual Augmentation to 0.5. The balance factor λ for calculating the final matching score in the Counterfact-Adaptive Framework is set to 0.6. The temporal factor in L_{SCL} and L_{CCL} is set to 0.1. the training stage, we employ the Adam optimizer [3] is employed to update the parameters with the learning rate set to 4×10^{-4} on a single Nvidia A6000 with 64 batch size.

2 Training and Testing Details

During the training stage, all the key components of CAEM method, *i.e.*, Temporal Counterfactual Augmentation (TCA), Event-Query Matching model (EQM), and Counterfact-Adaptive Framework (CAF), are enabled. However, during the inference stage, the TCA will be disabled. The factual and counterfactual EQMs are employed to process the same video-query pairs, and the predicted results will be integrated to generate a ranked moment sequence. We denote the training procedure of the proposed CAEM method in Algo. 1.

Algorithm 1: Our proposed CAEM method

Data: Untrimmed video V , text query Q
Result: Predicted Event Moment
Initialize pre-trained video feature extractors M_v and text feature extractors M_q ;
Initialize factual and counterfactual models EQM and EQM*;
if *isTraining* **then**
 Extracting visual and textual features:
 $F_v \leftarrow M_v(V)$, $F_q \leftarrow M_q(Q)$;
 if $l/l_V < 0.5$ **then**
 $F_v^* \leftarrow \text{Delaying}(F_v, \rho_1, \phi)$ counterfactual annotations: $l^* \leftarrow (t_s + \rho_1, t_e + \rho_1)$
 else
 Selecte irrelevant event \tilde{E} from other videos;
 $F_v^* \leftarrow \text{Inserting}(F_v, \tilde{E}, \rho_2, \phi)$;
 counterfactual annotations: $l^* \leftarrow (t_s + \rho_2, t_e + \rho_2)$;
 Calculate loss for factuals:
 $L_{SCL}, L_{LOC} \leftarrow \text{EQM}(F_v, F_q, l)$;
 Calculate loss for counterfactuals:
 $L_{SCL}^*, L_{LOC}^* \leftarrow \text{EQM}^*(F_v^*, F_q, l^*)$;
 Calculate L_{CCL} ;
 Backward;
else
 Calculating factual matching scores: $s \leftarrow \text{EQM}(F_v, F_q)$;
 Calculating counterfactual matching scores:
 $s^* \leftarrow \text{EQM}^*(F_v, F_q)$;
 Integrating matching scores: $s_{\text{total}} = s + \lambda s^*$;
 Rank moments according to s_{total} in descending;
 Return the first moment;

3 Additional Further Analysis

In this section, we conduct more ablation studies on key model components and training objectives on the ActivityNet-CD dataset. Moreover, to further demonstrate the effectiveness of our CAEM method, we combine our proposed TCA and CAF modules with the previous 2DTAN model and observe the achieved improvements.

Table 2: Experimental results of ablation studies on the ActivityNet-CD dataset.

SCL	TCA	CAF	dR@1, IoU=0.3		dR@1, IoU=0.5		dR@1, IoU=0.7	
			IID	OOD	IID	OOD	IID	OOD
✓	✗	✗	53.36	37.20	44.97	25.16	33.08	14.24
✗	✓	✗	52.73	36.00	44.46	24.37	32.09	13.92
✓	✓	✗	52.60	36.75	43.83	25.23	32.30	14.26
✗	✓	✓	53.02	36.87	44.88	25.27	32.75	14.34
✓	✓	✓	54.16	38.35	46.07	26.42	33.92	14.80

Ablation Studies on ActivityNet-CD dataset. We conduct more ablation studies on the three key modules, *i.e.*, Temporal Counterfactual Augmentation (TCA), Counterfact-Adaptive Framework (CAF), and the Semantic Consistency Learning (SCL) in the fundamental structure Query-Event Matching Model of our method. The

experimental results on the ActivityNet-CD dataset are reported in Table 2. We can observe that the three key components consistently show boosted performance which is similar to the results on the Charades-CD dataset. These outcomes affirm again the effectiveness of our approach, highlighting the pivotal role played by TCA, CAF, and SCL in enhancing OOD generalization of TSG performance across diverse video datasets.

Table 3: Analysis with respect to the training objectives in Event-Query Matching Model on the ActivityNet-CD dataset.

L_{rlm}	L_{mlm}	L_{cl}	dR@1, IoU=0.3		dR@1, IoU=0.5		dR@1, IoU=0.7	
			IID	OOD	IID	OOD	IID	OOD
✗	✗	✗	45.37	29.92	36.43	18.06	24.17	9.17
✓	✓	✗	53.02	36.87	44.88	25.27	32.75	14.34
✓	✓	✗	54.09	37.28	45.48	25.91	33.19	14.68
✓	✗	✓	53.32	38.09	45.46	25.86	33.25	14.60
✓	✓	✓	54.16	38.35	46.07	26.42	33.92	14.80

Additional Analysis on Training Objectives. We also conduct more analysis concerning the training objectives in EQM on the ActivityNet-CD dataset. According to the experimental results listed in Table 3, we can see that both IID and OOD drop dramatically without L_{rlm} , which is consistent with the experimental results on the Charades-CD dataset. The reason is the L_{rlm} aims to teach our model to learn relative temporal locations between these fixed temporal locations and ground truth. Event-level multimodal representations are learned with the compositions of vision-aware tokens, which have fixed mapping temporal locations determined manually. Hence the L_{rlm} performs an essential role in precisely rectifying the temporal locations.

Analysis on Model Transferability. We also demonstrate the model transferability of our proposed CAEM method by combining two key modules, *i.e.*, TCA and CAF modules, with previous TSG models. Specifically, we employ the typical TSG method 2DTAN [12], and devise two extended models 2DTAN + TCA and 2DTAN + TCA + CAF that are facilitated with our modules. The experiments are conducted on Charades-CD and ActivityNet-CD datasets under the same experimental settings. According to the results illustrated in Fig. 2, we can observe that our proposed TCA and CAF modules effectively boost the performance of the previous TSG method 2DTAN on both IID and OOD test splits. By comparing the experimental results of the three models, we can see the performance is increasing with TCA and CAF being introduced gradually. Particularly, facilitated with the two modules, the 2DTAN method achieves remarkable improvements on the OOD test samples, outperforming previous results by a large margin. It proves again that overcoming the limitations of OOD generalization in the TSG task with counterfactual data augmentation and consistency rule is reasonable.

Combination with Proposal-free Method. We have validated the transferability of our proposed key components by combining our method with 2DTAN. This typical proposal-based method selects predictions by evaluating the matching scores of event-query joint representations. Considering that the proposal-free methods

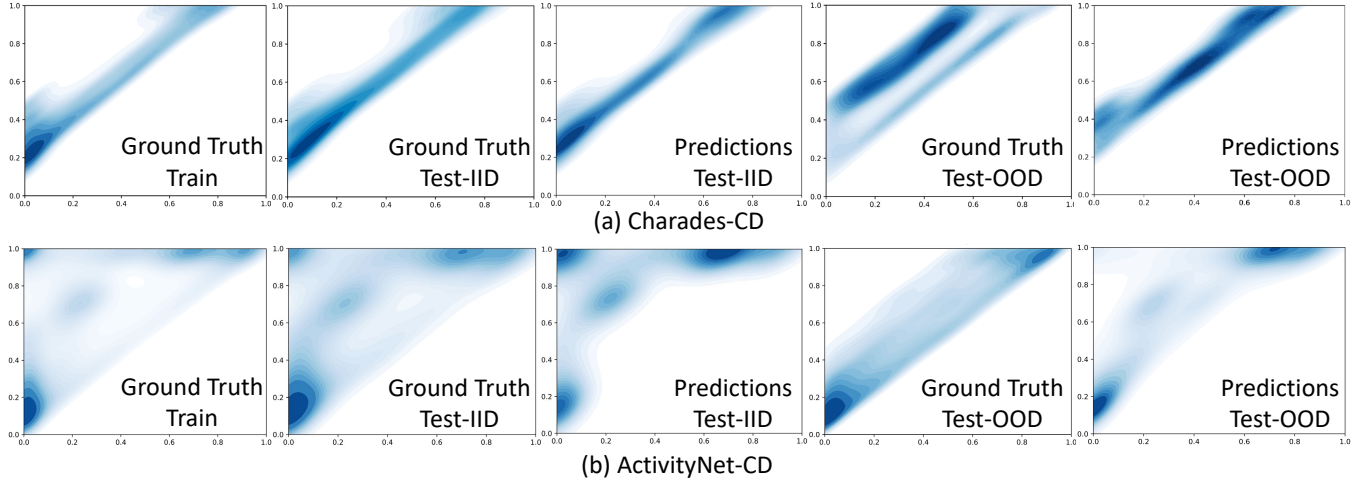


Figure 1: Visualizations of temporal distributions of all train, test-iid, and test-ood samples from (a) ActivityNet-CD and (b) Charades-CD datasets and corresponding predictions of our proposed CAEM method. Note that the X and Y axes in each subfigure represent normalized starting and ending timestamps respectively.

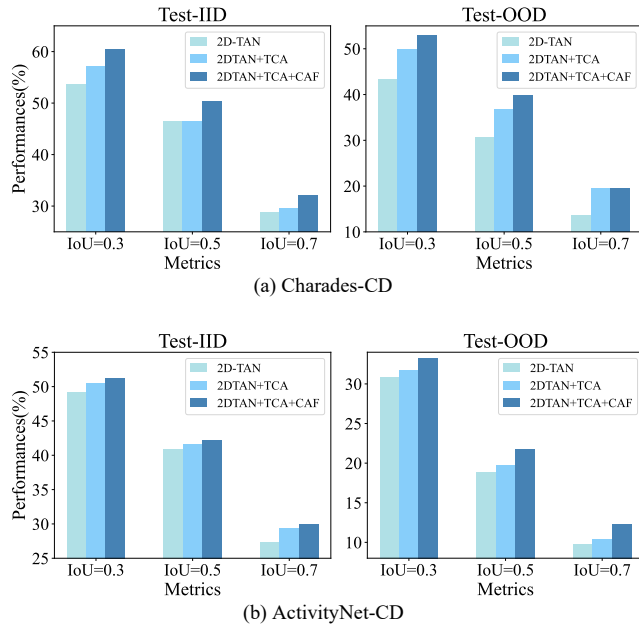


Figure 2: Analysis of model transferability. Note that we adopt the discounted recall metrics which is the same as experiments in ablation studies in formal text.

are also proposed to address the conventional TSG challenge, we combine our method based on a proposal-free paradigm to explore its effectiveness further. Here, we report the experimental results based on a proposal-free method VSLNet, and make comparisons with the SOTA BSSARD method. We can observe that our method consistently boosts the performance of VSLNet on both two datasets and outperforms the counterpart BSSARD method. However, it can be observed that our complete CAEM shows better results. One

probable reason is that our proposed two modules are deliberately designed for the pipeline illustrated in Fig. 1(b) in the formal paper.

Table 4: Additional experimental results of combining our proposed method with the proposal-free VSLNet baseline.

Methods	ActivityNet-CD-OOD		Charades-CD-OOD	
	R1@0.5	R1@0.7	R1@0.5	R1@0.7
VSLNet (ACL'20)	25.40	13.51	43.08	22.52
VSLNet+BSSARD (AAAI'24)	27.02	14.93	47.20	27.17
VSLNet+TCA+CAF (Ours)	27.41	15.04	49.21	29.10
CAEM (Ours)	28.98	15.54	54.42	28.29

Visualizations of Temporal Distributions. In this part, we visualize the temporal distributions of all predictions generated by our CAEM method on both two datasets in Fig. 1, including IID and OOD test samples. By comparing the ground truth and predictions of our CAEM method, we come to the following conclusions: (1) Our method shows significant OOD generalization ability against training bias. According to the visualizations of our predictions on OOD test samples of the two datasets, we can see the centers with high density have been shifted compared with the training split, which demonstrated that our method avoids overfitting the training bias that leads to degenerated generalization. (2) However, we also note that the kernels of predictions in high density are partially overlapped only with the kernels of ground truth. It means our method could localize the temporal locations roughly better still has limitations on predicting the event duration. Such observations also inspire us for future work on the DTSG task.

4 Additional Qualitative Analysis

Finally, we visualize more OOD test cases from the Charades-CD and ActivityNet-CD datasets in Fig. 3. Following the settings adopted in the formal text, we also make fair comparisons with

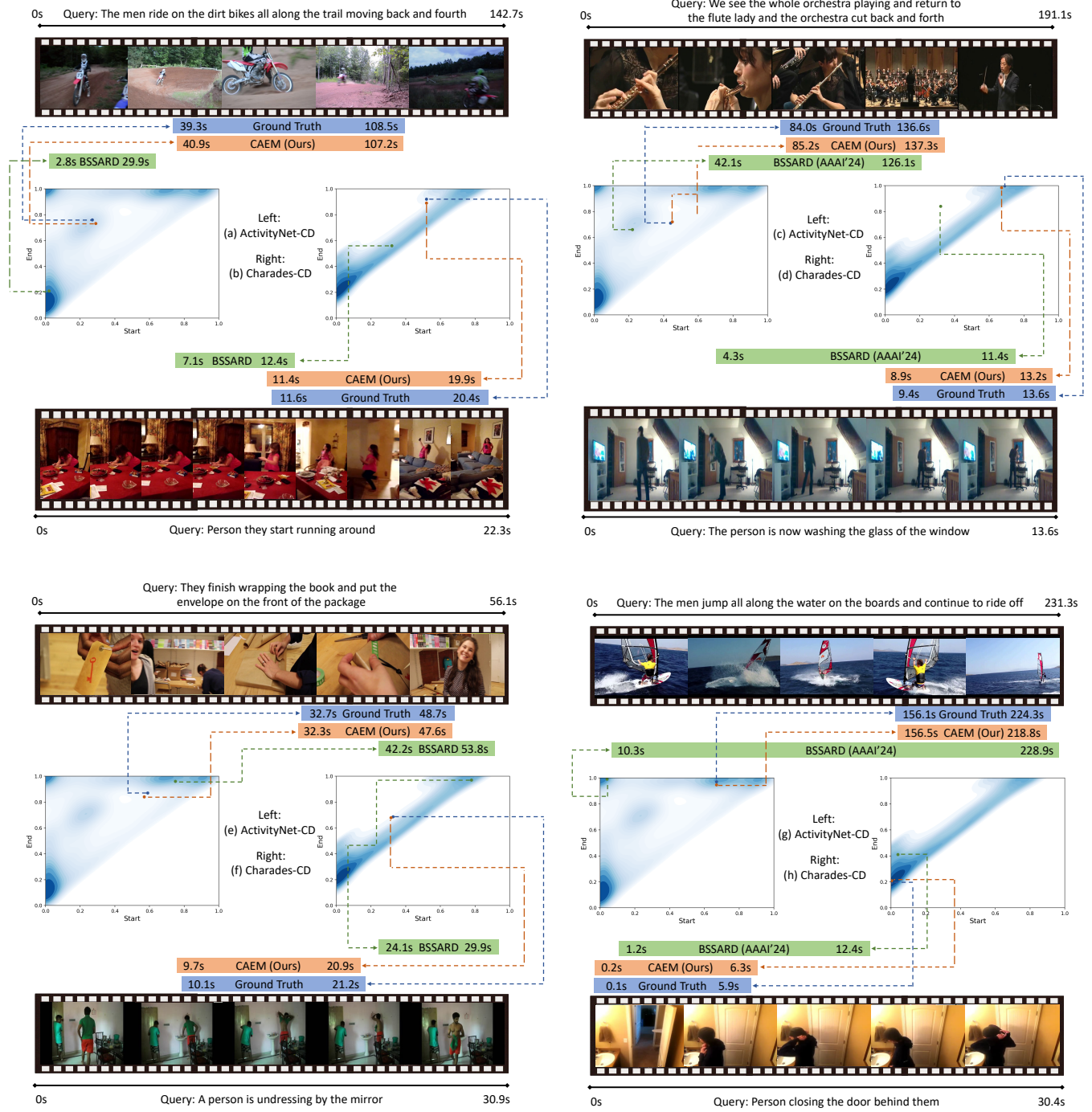


Figure 3: Additional visualizations of grounding results of OOD test samples from ActivityNet-CD (left) and Charades-CD (right) datasets and corresponding distributions of training sets.

recent state-of-the-art method BSSARD [7]. By observing the visualization results, we can see that our proposed CAEM method precisely localizes the target event that is the most relevant to the given text query. It proves the effectiveness of our solution again, which

introduces counterfactual data augmentation and consistency rule into event-query matching to achieve better generalization.

References

- [1] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4724–4733.
- [2] Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. 2022. Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In *Proceedings of the European conference on computer vision*. 130–147.
- [3] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.
- [4] Xiaohan Lan, Yitian Yuan, Hong Chen, Xin Wang, Zequn Jie, Lin Ma, Zhi Wang, and Wenwu Zhu. 2023. Curriculum multi-negative augmentation for debiased video grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1213–1221.
- [5] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. 2022. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3032–3041.
- [6] Daizong Liu, Xiaoye Qu, and Wei Hu. 2022. Reducing the vision and language bias for temporal sentence grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4092–4101.
- [7] Zhaobo Qi, Yibo Yuan, Xiaowen Ruan, Shuhui Wang, Weigang Zhang, and Qingming Huang. 2024. Bias-Conflict Sample Synthesis and Adversarial Removal Debias Strategy for Temporal Sentence Grounding in Video. *arXiv preprint arXiv:2401.07567* (2024).
- [8] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision*. 4489–4497.
- [9] Xin Wang, Zihao Wu, Hong Chen, Xiaohan Lan, and Wenwu Zhu. 2023. Mixup-Augmented Temporally Debiased Video Grounding with Content-Location Disentanglement. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4450–4459.
- [10] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. 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.
- [11] 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.
- [12] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *AAAI Conference on Artificial Intelligence*. 12870–12877.