Curriculum Multi-Negative Augmentation for Debiased Video Grounding

Xiaohan Lan[1], Yitian Yuan[2], Hong Chen[1], Xin Wang[1*], Zequn Jie[2], Lin Ma[2], Zhi Wang[1], Wenwu Zhu[1*]

[1]Tsinghua University
[2]Meituan Inc.
{lanxh20,h-chen20}@mails.tsinghua.edu.cn, yuanyitian@foxmail.com, {xin_wang,wwzhu}@tsinghua.edu.cn, {zequn.nus,forest.linma}@gmail.com, wangzhi@sz.tsinghua.edu.cn

Abstract

Video Grounding (VG) aims to locate the desired segment from a video given a sentence query. Recent studies have found that current VG models are prone to over-rely the groundtruth moment annotation distribution biases in the training set. To discourage the standard VG model's behavior of exploiting such temporal annotation biases and improve the model generalization ability, we propose multiple negative augmentations in a hierarchical way, including cross-video augmentations from clip-/video-level, and self-shuffled augmentations with masks. These augmentations can effectively diversify the data distribution so that the model can make more reasonable predictions instead of merely fitting the temporal biases. However, directly adopting such data augmentation strategy may inevitably carry some noise shown in our cases, since not all of the handcrafted augmentations are semantically irrelevant to the groundtruth video. To further denoise and improve the grounding accuracy, we design a multi-stage curriculum strategy to adaptively train the standard VG model from easy to hard negative augmentations. Experiments on newly collected Charades-CD and ActivityNet-CD datasets demonstrate our proposed strategy can improve the performance of the base model on both i.i.d and o.o.d scenarios.

Introduction

Video grounding (VG), one of the video-and-language tasks that seeks to determine the start and end timestamps of a natural language-described segment (also named moment) from an untrimmed video, has attracted increasing interests of the multimedia community over the past few years (Gao et al. 2017; Lan et al. 2021). As shown in Figure 1(a), taking a video and a descriptive sentence as inputs, a VG model needs to return the temporal locations of the target moment referring to the sentence query. Compared to other natural language-based video understanding tasks like video question answering and video captioning, the grounding accuracy in VG is a more intuitive metric to evaluate the video-sentence matching from the level of semantics. Therefore, VG has been widely investigated these years and achieved promising grounding results by various state-of-the-art (SOTA) deep models.

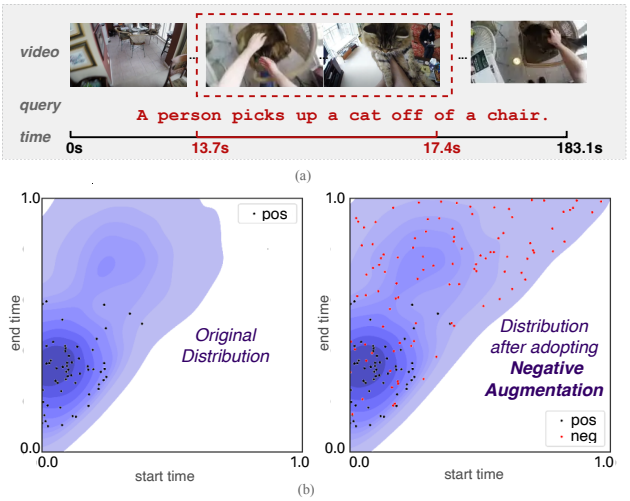


Figure 1: (a): VG task definition. Given a sentence, a semantically corresponding video segment should be temporally located in the video. (b): Conceptual illustration of our data augmentation-based debiasing strategy for VG. Our designed negative samples (shown in red dots) are able to balance the biased moment annotation distribution of positive samples (shown in black dots) with fair negative treatment to moment candidates at more diverse locations.

However, recent studies (Otani et al. 2020; Yuan et al. 2021) start to doubt whether the performance gains of SOTA VG models come from more powerful cross-modal semantic matching ability rather than exploiting some kind of *shortcuts*, *e.g.*, the temporal annotation distribution biases in the training set. Except some early two-stage VG models (Gao et al. 2017; Hendricks et al. 2017) that individually match the pre-cut video moment candidates with the query, the majority of advanced VG methods (Chen et al. 2018; Yuan et al. 2019; Zhang et al. 2020) resort to modeling the temporal relations within the video by encoding the video units (clips or moments) in the whole video context/sequence, which will embed the temporal location information in the video clip/-moment features and these temporal location clues will inevitably influence the model learning procedure. When the model gets no clues on cross-modal matching, or only understands one side of these two modalities, it may leverage

*Corresponding authors

the location priors to 'take a guess' of the target moment. If the temporally annotated labels in the training set are unbalanced with obvious distribution biases, it is easy for those models that implicitly embed the temporal signals and exploit such biases to perform temporal predictions. To better demonstrate the above problem, Otani *et al*. 2020 firstly report the dataset bias issue by conducting groups of experiments, and Yuan *et al*. 2021 further design new evaluation protocols to assess the generalization ability of SOTA models.

To discourage the standard VG model's behavior of exploiting such temporal annotation biases and improve the model generalization ability, we present a simple yet effective Curriculum Multi-Negative Augmentation (NA) framework. Our motivation of designing such a curriculum Multi-NA framework comes from the following observation: even given a sentence query and a video semantically unrelated to it as inputs, a biased model would still unreasonably produce higher grounding confidence scores for those video clips/-moments whose temporal positions appear more frequently among the training set's ground-truth annotations. We hope to debias the model by forcing it to fairly output low confidence scores for these video units that have no semantic relevance to the sentence query, and therefore make the model focus more on learning semantic matching relationships between videos and sentences rather than fitting temporal location biases in the dataset. Therefore, we propose multiple negative augmentations in a hierarchical way to enrich the training set's temporal label distribution.

Specifically, our proposed Multi-NA strategy mainly focuses on creating some 'pseudo' negative video samples that have no or weak semantic relations with the given sentence query, which includes: (i) *clip-level cross-video NA samples* composed by the clips randomly picked up from other videos, (ii) *video-level cross-video NA samples* that directly replace themselves with other videos, and (iii) *self-shuffled NA samples with masks* obtained by shuffling the positive video itself and setting a proportion of positions to zero at feature level. As illustrated in Figure 1(b), these NA samples added in the model training procedure can balance the moment annotation distribution, and our objective function will encourage the model to give lower grounding confidence scores to those video moments which have weak or no semantic relevance to the sentence query. Therefore, these augmentations can effectively diversify the temporal annotation distribution and help the model make more reasonable predictions rather than fit the temporal biases.

However, not all of the above negative video samples are equally semantic-irrelevant to the positive video (as well as the corresponding sentence query). Actually, the difficulty of distinguishing the negative attributes of our created 'pseudo' video samples is from easy to hard. For example, both the temporal continuity and semantic relevance of the 'pseudo' videos from our first clip-level cross-video NA strategy are broken, and it is easy for the model to determine such videos as negative. However, the synthesized videos from the third self-shuffled NA strategy are harder to distinguish, since the clips in these videos are from the positive videos and they are still highly semantically relevant to the sentence query. The shuffling and masking operation can only break temporal continuity and logic in the original video sequence. Therefore, if we input the hard negative samples generated from the third NA strategy in the early training stage, it will inevitably confuse the model and bring some noise. Inspired by the automatic denoising characteristics (Wang, Chen, and Zhu 2022) of curriculum learning, in our scenario, we design a multi-stage curriculum strategy to adaptively train the VG model with negative augmentations gradually from easy (clean) to hard (noisy) samples, which guides the model to better optima with cleaner gradients in the early training process (Bengio et al. 2009), so that the VG model will not get confused by those noisy data.

To prove the effectiveness of our method, we conduct experiments on the newly collected Charades-CD and ActivityNet-CD datasets (Yuan et al. 2021). The experimental results show that our proposed Curriculum Multi-NA strategy can effectively debias the base model with significant improvements on both i.i.d and o.o.d scenarios. Our contribution can be summarized as follows[1]:

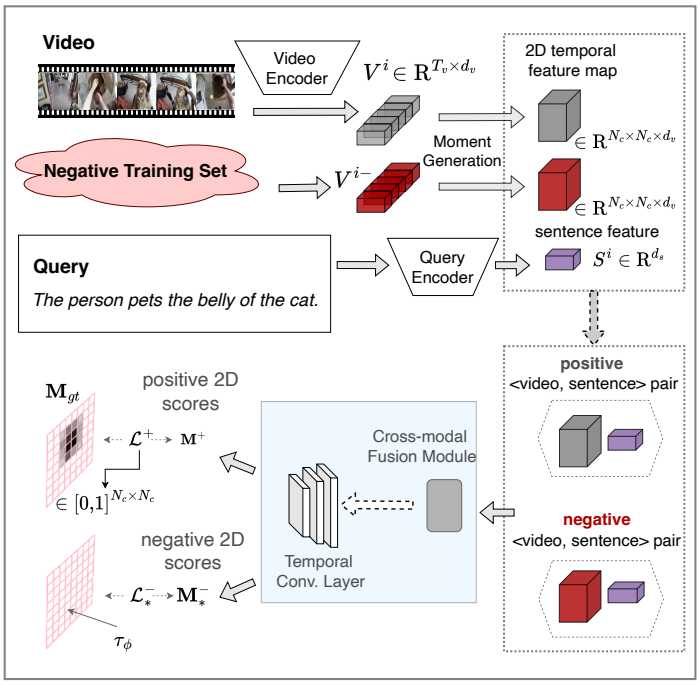
- We hierarchically propose multiple VG-specific negative augmentations (Multi-NA) for the debiased video grounding problem, from the perspective of enriching the temporal label distribution.
- We propose a multi-stage curriculum VG training strategy for the hierarchically augmented samples, to alleviate the impact of noise contained in the negative augmentations.
- Experimental results show that our proposed Curriculum Multi-NA strategy can effectively debias the base model with significant improvements on both i.i.d and o.o.d scenarios of Charades-CD and ActivityNet-CD datasets.

Related Work

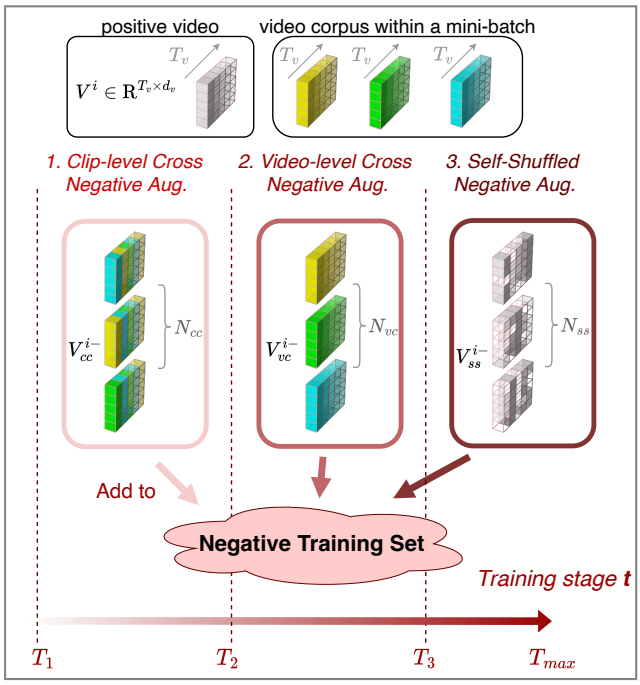
Video Grounding

Video grounding aims to retrieve a desired video segment (or moment) from a given video according to a sentence query. Therefore, a VG model needs to model the cross-modal relation and find the semantic correspondence between the visual and natural language inputs. Since the task was proposed by Gao *et al*. 2017, a variety of deep VG models have attempted to capture the cross-modal semantic matching relationships and predict the target moment locations more and more accurately and efficiently. These VG models can be basically grouped into two categories, *i.e*., proposal-based and proposal-free methods. The proposal-based methods aim to obtain the matching scores of segment candidate proposals and choose the proposals with top scores as predictions (Gao et al. 2017; Hendricks et al. 2017; Chen et al. 2018; Yuan et al. 2019; Zhang et al. 2020), while the proposal-free methods take the start (end) timestamps as supervision signals and directly output the target locations without the proposal generation process (Yuan, Mei, and Zhu 2019; Ghosh et al. 2019; Zeng et al. 2020). Since

[1]Our codes are available at https://github.com/rubylan/Curri-MultiNA



The Overall Grounding Pipeline



Multi-stage Curriculum Process with NA
(from *easy* to *hard*)

Figure 2: The overview of our proposed Curriculum Multi-NA framework. We choose a proposal-based method as the base model, add three kinds of negative data augmentations against the positive video to the whole training procedure, and encourage the proposals generated by those negative videos to fairly output low confidence scores (shown on the left). Since some of the augmented videos may have semantic relations with the original video, such noisy data would disturb the training at the early stage. Therefore, we adopt a multi-stage curriculum strategy to gradually add the negative samples from easy to hard for adaptive training (shown on the right). The curriculum strategy can effectively remove the noise brought by the relatively hard (noisy) negative samples and further improve the grounding accuracy.

the acquirement of moment-level temporal annotations is labor-intensive, a great number of weakly supervised approaches (Mithun, Paul, and Roy-Chowdhury 2019; Duan et al. 2018) that have no need to access such temporal annotations at the training stage have appeared. These weakly supervised methods mainly leverage the semantic alignment of groundtruth video-sentence pairs to yield accurate locations, for example, some of them (Song et al. 2020) attempt to utilize the reconstruction loss that is produced by generating captions for video moment proposals to supervise the training.

Temporal Annotation Biases in VG

The temporal bias issue has been initially proposed by Otani *et al*. 2020. Afterwards, various efforts attempt to address it from both improvements of benchmarks and models. For example, as for benchmarks, to disentangle the effects brought by temporally-biased datasets and evaluate the model generalization more reliably, Yuan *et al*. 2021 re-split the commonly used benchmark datasets to conduct the out-of-distribution test with a test set that has different groundtruth annotation distribution with the training set. Soldan *et al*. 2022 present a large-scale MAD dataset that is able to alleviate the identified temporal biases with considerably longer videos from movies and a much greater amount of sentence samples from audio descriptions in movies. As for models, various debiased approaches are specifically designed and proposed for alleviating the temporal bias issue in VG. Inspired by the visual question answering domain (Agrawal et al. 2018), some employ the uni-modal branch to capture the biases and integrate it into the standard branch in an ensemble way (Zhang et al. 2021). Some other methods attempt to view the VG problem from the perspective of causality and debias the base model with causal intervention (Yang et al. 2021; Lan et al. 2022; Bao and Mu 2022). To the best of our knowledge, our method is the first to adopt curriculum learning-based data augmentation for debiased video grounding, which is orthogonal to existing methods.

Negative Sample Augmentations in VG

Introducing additional negative samples as a way of data augmentation has been proven to be effective in many domains (Li et al. 2022a; Zhang et al. 2022a), which also benefits the problem of video grounding (Luo et al. 2021; Wang, Chen, and Jiang 2021; Nan et al. 2021; Ding et al. 2021;

Zheng et al. 2022; Li et al. 2022b), especially in the weakly supervised setting. It becomes a necessity to exploit the negative moment/query samples for contrastive learning-based training to better learn the cross-modal semantic alignment relations with no access to fine-grained temporal annotations. The negative augmentations in our method are fundamentally different from these previous studies in two folds: i) Objects to be augmented are different. The augmentation of previous works is mainly operated in moment-level (Nan et al. 2021; Ding et al. 2021) or query-level (Zheng et al. 2022), while our method generates the negative augmentation in video-level; ii) Levels of the model where the negative augmentations are projected and optimized are different: for previous works, the negative samples are accompanied with the positive samples to compose a triplet for contrastive training in the common feature space. However, in our work, the effects of negative augmentations will happen at the inception level near the output layer.

Curriculum Learning

The concept of curriculum learning (CL) is firstly proposed by Bengio *et al*. 2009. Inspired by the easy-to-hard learning paradigms of human beings, CL likewise trains the deep learning model on the easier subset of data at the beginning, and then gradually increases the difficulty of data subset, until the entire training dataset is reached. Recent studies have shown the power of CL (Zhou et al. 2022a) as a task-free training paradigm applying to a wide range of specific problems (Chen et al. 2021b; Zhou et al. 2022b; Zhang et al. 2022b), *e.g*., image classification (Gong et al. 2016; Guo et al. 2018), machine translation (Tay et al. 2019; Kumar et al. 2019), and recommendation (Chen et al. 2021a).

According to the systematic review on CL (Wang, Chen, and Zhu 2021), given the common scheme of CL, it should be properly instantiated into practices by defining a difficulty measurer and a training scheduler for a specific task. Both these two components could be either pre-defined (statically fixed before training), or automatic (*i.e*., dynamically adjusted during training). Furthermore, CL should be used either to *guide* the training towards better parametric-space regions in the way of model optimization, or to *denoise* through concentrating more on high-confidence areas of data distributions. The curriculum design in our framework is primarily out of the second motivation, *i.e*., due to the noisy data generated by the negative augmentation policies, we present a simple yet effective pre-defined multi-stage curriculum strategy to help denoise and improve the robustness and generalizability of the trained model. The CL-driven debiasing solution has also been successfully applied in VQA (Lao et al. 2021).

Methodology

The overview of our proposed Curriculum Multi-NA framework is shown in Figure 2. For clearer presentation, we will first take a quick review on the VG problem formulation and the base model which is of necessity but not our main contribution. Afterwards, the VG-specific augmentation methods for three kinds of negative samples are introduced followed with the overall optimization functions, which achieves the goal of debiasing. Then our multi-stage curriculum strategy is presented to further denoise the training with noisy augmented samples.

Problem Formulation

Given a video V and a sentence query S, a VG model will learn a projection function $f_\theta(V, S) = \tau_p$, which identifies the start and end timestamps in V semantically corresponding to sentence S. During the training stage, we utilize the groundtruth annotation τ_g to optimize the VG model.

Base Model

We follow the work by Zhang *et al*. 2020 as the whole grounding pipeline shown in the left part of Figure 2. The core design of their grounding method is to represent the candidate moments with a 2D temporal feature map $\in [0,1]^{N_c \times N_c \times d_v}$, where each position of 2D coordinates $(i, j),\ i, j \in \{1, 2, \cdots, N_c\}$ represents one candidate moment that starts at the i-th and ends at the j-th temporal unit. The 2D temporal map is able to model the adjacent relations of moments while it would also embed the position information unavoidably. After feeding the 2D moment features and the query feature into the cross-modal fusion module followed by the temporal convolution layer, we can obtain a 2D score map $\mathbf{M} \in [0,1]^{N_c \times N_c}$ that describes the moment-to-text semantic relevance.

Negative Augmentations

To alleviate the temporal bias issue, we provide three different ways to generate 'pseudo' negative video samples. The negative augmentations (NA) can obviously change the data distribution of the training set and effectively penalize the behaviours of biased prediction with the help of the non-matched query-video pairs. Therefore, those negative augmentations are able to force the VG model to truly understand the multimodal inputs and learn the cross-modal semantic alignment. In the following, these three negative augmentations against a given video-sentence pair (V^i, S^i) will be introduced.

Clip-level Cross-video NA. Suppose each video V is composed by a sequence of T_v video clips, for example, $V^i = \{c_1^i, c_2^i, ..., c_{T_v}^i\}$. To generate the clip-level cross-video NA sample V_{cc}^{i-} for V^i, we synthesize it in a clip-by-clip manner, *i.e*., randomly select a video V^{p_k} other than V^i in its mini-batch (with batch size B), and then take the k-th video clip $c_k^{p_k}$ in V^{p_k} as the k-th clip of V_{cc}^{i-}. As such, the synthesized negative 'pseudo' video can be formulated as:

$$V_{cc}^{i-} = \{c_k^{p_k}\}_{k=1}^{T_v}, \quad p_k \in \{1, 2, ..., B\} \text{ and } p_k \neq i. \quad (1)$$

Since V_{cc}^{i-} is composed of randomly selected clips from other videos, it is not able to represent a continuing sequence of interactive activities, *i.e*., the semantics behind it are limited.

Algorithm 1: Multi-stage Curriculum Process

Input: negative augmentation function $\mathcal{G}_{\{cc,vc,ss\}}(\cdot)$, the standard VG model f_θ, original training set S_o, augmented training set $\hat{S}_o$, training stage update time T_* and total training step number T_{max}.
Output: optimized model f_θ

```
Initialization: S_o ← {(v_*, s_*, τ_*)}, Ŝ_o ← ∅
for t = 1, ··· , T_max do
    {When training stage 1 is reached.}
    if t = T_1 then
        V_cc^- ← G_cc(V, N_cc^-) based on Equation (1)
        Ŝ_o ← Ŝ_o ∪ {(v_cc*^-, s_*, τ_φ)}
    end if
    {When training stage 2 is reached.}
    if t = T_2 then
        V_vc^- ← G_vc(V, N_vc^-) based on Equation (2)
        Ŝ_o ← Ŝ_o ∪_o {(v_vc*^-, s_*, τ_φ)}
    end if
    {When training stage 3 is reached.}
    if t = T_3 then
        V_ss^- ← G_ss(V, N_ss^-) based on Equation (3)
        Ŝ_o ← Ŝ_o ∪_o {(v_ss*^-, s_*, τ_φ)}
    end if
    {Update the gradients for optimization.}
    θ ← train(f_θ, λ, {S_o, Ŝ_o})
end for
```

Video-level Cross-video NA. The second negative augmentation is to randomly pick up another video V^k ($k \neq i$) in the mini-batch:

$$V_{vc}^{i-} = V^k, \quad k \neq i. \tag{2}$$

Compared to the clip-level cross-video augmented sample (V_{cc}^{i-}), the video-level negative sample V_{vc}^{i-} processes certain semantic information, which is more likely to be semantically relevant to the positive video V^i compared to the clip-level ones, thus more likely to bring noise for training.

Self-Shuffled NA. with Masks. The third NA strategy is to generate the negative video sample by the positive video itself. Firstly, the original positive video will be temporally shuffled in its clips. Then, we will generate a feature map mask, where some feature dimensions of some video clips will be randomly set to zero controlled by a mask ratio $\alpha \in (0, 1)$:

$$V_{ss}^{i-} = \text{Shuffle}(V^i) \odot \text{Mask}(\alpha). \tag{3}$$

The self-generated negative sample should be highly semantically similar with the original video in the feature space compared with those cross-video generated ones. However, its temporal continuity is totally damaged by the shuffling operation and the complete semantic information is partially hidden by subsequent masking so that this kind of synthesized sample can be viewed as negative as well.

Objective Function

The optimization objective consists of two parts, the loss for the original positive samples and the loss for our augmented ones.

For the original positive samples (v_*, s_*, τ_{g*}), we adopt the same loss function as Zhang *et al*. 2020:

$$\mathcal{L}^+ = \mathcal{L}_{bce}(\mathbf{M}^+, \mathbf{M}_{gt}), \tag{4}$$

where we compute the binary cross-entropy loss between the predicted 2D score $\mathbf{M}^+$ and the groundtruth $\mathbf{M}_{gt}$ of the scaled temporal IoU.

For all the negative samples (v_*^-, s_*, τ_ϕ), since the negative video v^- is semantically irrelevant to the sentence query s, we uniformly assign them with the *empty* groundtruth label, denoted as τ_ϕ, which means that our objective is to encourage the model to fairly predict the relevance scores of all candidate moments as zero. To achieve this, we design the negative loss function that can decrease the unreasonable relevance scores simply using the L1 distance loss:

$$\mathcal{L}_{\{cc,vc,ss\}}^- = \mathcal{L}_1(\mathbf{M}_{\{cc,vc,ss\}}^-, \tau_\phi), \tag{5}$$

Then we further adopt the hype-parameters λ_1, λ_2 and λ_3 to balance the losses obtained by positive and negative samples:

$$\mathcal{L} = \mathcal{L}^+ + \lambda_1 \mathcal{L}_{cc}^- + \lambda_2 \mathcal{L}_{vc}^- + \lambda_3 \mathcal{L}_{ss}^-, \tag{6}$$

and $\lambda_{\{1,2,3\}}$ is able to control the curriculum process as well.

Multi-stage Curriculum Strategy

As discussed above, these negative augmentations are hierarchically generated from different granularities (clip- or video-level) or sources (other videos or the positive video itself) so the maintained semantic relevance to the positive video differs as well. Generally, the self-generated video (*i.e*., V_{ss}^{i-}) should be more similar to the original video than the cross-video generated ones (*i.e*., $V_{\{cc,vc\}}^{i-}$) since it keeps the same domain scenes with the original one. As for those two kinds of cross-video generated videos, the complete video-level replacement from a single video has more chance to maintain the overlapped semantics with the original one than the clip-level synthesized one from multiple video sources.

When we train the model with enhancement of negative samples that share some semantics with the original video, it will get harder for the model to differentiate the outcomes of both negative and positive ones. Note that the labels for the positive ones are given by the groundtruth, but the labels for the negative ones are empty. Similar input semantics but totally different labels will confuse the model. Therefore, these negative augmentations unavoidably bring noise for training despite playing a positive role of diversifying the data distribution. Motivated by the power of curriculum learning for denoising, we design a multi-stage curriculum process to adaptively train the model by gradually adding those negative augmentations stage by stage. The detailed pipeline is described at Algorithm 1.

Experiments

Evaluation Protocols

To better validate whether our proposed method could alleviate the over-reliance on dataset biases, we adopt the eval-

Models	Charades-CD						ActivityNet-CD					
	dR@1,IoU=0.3		dR@1,IoU=0.5		dR@1,IoU=0.7		dR@1,IoU=0.3		dR@1,IoU=0.5		dR@1,IoU=0.7	
	i.i.d	o.o.d	i.i.d	o.o.d	i.i.d	o.o.d	i.i.d	o.o.d	i.i.d	o.o.d	i.i.d	o.o.d
CTRL	42.65	44.97	29.80	30.73	11.86	11.97	19.42	15.68	11.27	7.89	4.29	2.53
ACRN	47.50	44.69	31.77	30.03	12.93	11.89	20.06	16.06	11.57	7.58	4.41	2.48
ABLR	52.26	44.62	41.13	31.57	23.50	11.38	46.86	33.45	35.45	20.88	20.57	10.03
SCDM	58.14	**52.38**	47.36	**41.60**	30.79	**22.22**	46.44	31.56	35.15	19.14	22.04	9.31
DRN	51.35	40.45	41.91	30.43	26.74	15.91	48.92	**36.86**	39.27	**25.15**	25.71	**14.33**
TSP-PRL	46.44	31.93	35.43	19.37	17.01	6.20	44.93	29.61	33.93	16.63	19.50	7.43
2D-TAN	53.71	43.45	46.48	28.18	28.18	13.73	49.18	30.86	40.87	18.86	**28.95**	9.77
Ours	**64.21**	52.21	**53.82**	39.86	**34.47**	21.38	**49.91**	32.32	**41.67**	20.78	28.82	11.03

Table 1: Overall performance (%) comparisons with other VG models (best results are in bold and second in underline).

uation protocols proposed by Yuan *et al*. 2021 with two re-organized datasets, which can evaluate the model's generalization ability with out-of-distribution test using a test set (*i.e*., *test-ood*) that has a totally different moment annotation distribution against the *train*/*val*/*test-iid* sets. More dataset details are as follows:

Charades-CD. It is re-organized from Charades-STA dataset (Gao et al. 2017) with an average video length of 30 seconds. The numbers of videos in *train*/*val*/*test-iid*/*test-ood* splits are $4,564/333/333/1,442$, and the numbers of video-query pairs are $11,071/859/823/3,375$ respectively.

ActivityNet-CD. It is built upon ActivityNet Captions dataset (Krishna et al. 2017). The videos contain the daily activities and are around 180 seconds on average. The numbers of videos in *train*/*val*/*test-iid*/*test-ood* splits are $10,984/746/746/2,450$, and the numbers of video-query pairs are $51,415/3,521/3,443/13,578$ respectively.

Metrics. As for evaluation, we adopt the commonly used metric R@n,IoU=m (Gao et al. 2017). It returns the proportion of positive samples which have at least one moment out of top n retrieved moments whose temporal IoU score is larger than m with the groundtruth moment. We also report the results with the new metric of dR@n,IoU=m (Yuan et al. 2021) that further discounts the recall values of R@n,IoU=m based on temporal distances, which is more reliable under small IoU thresholds.

Implementation Details

As for the training strategy setting, we trained 30/20 (*i.e*., T_{max}) epochs for Charades-CD/ActivityNet-CD and report results of the epoch whose *test-iid* set performs the best with metric R@1,IoU=0.7. The batch sizes and learning rates were set to 64/32 and $0.0005/0.0001$, respectively. $\lambda_{\{1,2,3\}}$ in $\mathcal{L}$ were all set to 5.0 for Charades-CD, and set to 15.0 for ActivityNet-CD. We adaptively trained the model with the multi-stage curriculum process and set training stage update time T_1, T_2 and T_3 to 3/7/18 and 2/5/13, respectively.

As for the model architecture setting, to implement the Multi-NA strategy, we set the mask ratio α to 0.55 and the numbers of per-sample generated samples for each NA type (*i.e*., $N^{-}_{\{cc,vc,ss\}}$) to 1 on both datasets. Other hyper-parameters of the architecture (*e.g*., query hidden size, temporal convolutional layer number or kernel size) were the same as the original paper (Zhang et al. 2020). To fairly compare with other VG models, we also followed Yuan *et al*. (Yuan et al. 2021) to extract I3D (Carreira and Zisserman 2017) video features for Charades-CD and C3D (Tran et al. 2015) video features for Activity-CD and encode the sentence with GloVe (Pennington, Socher, and Manning 2014).

Experimental Results

We present our evaluation results on both *test-iid* and *test-ood* sets with the metrics dR@1,IoU=$\{0.3,0.5,0.7\}$ in Table 1. The results of other VG methods with the new benchmark are originally reported in Yuan *et al*. 2021, including CTRL (Gao et al. 2017), ACRN (Liu et al. 2018), ABLR (Yuan, Mei, and Zhu 2019), SCDM (Yuan et al. 2019), DRN (Zeng et al. 2020), TSP-PRL (Wu et al. 2020) and 2D-TAN (Zhang et al. 2020).

After comparing the grounding performance of our method with that of 2D-TAN, it can be concluded that our proposed Curriculum Multi-NA can improve the base model with significant performance gains. More observations based on the results of both datasets are shown as follows:

- Performance improvements upon the base model (2D-TAN) after adopting our proposed Curriculum Multi-NA are significant on both *i.i.d* and *o.o.d* scenarios of Charades-CD, *e.g*., with recall point gains of $6.29/7.65$ under the most challenging metric dR@1,IoU=0.7, which demonstrates that our NA strategy can effectively enrich the label space of the dataset for more accurate and reasonable grounding.
- The improvements upon the base model on ActivityNet dataset are not significant as much as Charades-CD, which may be due to that the videos from Activity-CD contain more complicated interactive activities, the benefits of distinguishing our synthesized negative videos is not adequate for cross-modal semantic matching under such a more challenging phenomenon.
- It is observed that our strategy achieves 1.96% (IID) and 10.18% (OOD) improvements on dR@1,IoU=0.5 over the base model and reduces the IID-OOD performance gap. Thus the model generalization ability increases accordingly after adopting our strategy.

No.	experimental settings				R@1,IoU=0.5		R@1,IoU=0.7		R@5,IoU=0.5		R@5,IoU=0.7	
	w/cc NA.	w/vc NA.	w/ss NA.	w/curriculum	i.i.d	o.o.d	i.i.d	o.o.d	i.i.d	o.o.d	i.i.d	o.o.d
1					49.33	39.50	26.73	17.82	85.05	79.42	54.56	40.78
2	✓				50.67	40.87	27.70	19.28	87.61	79.54	57.59	44.13
3		✓			57.23	41.96	33.54	20.55	89.55	81.88	59.17	45.34
4			✓		47.75	37.01	28.19	17.59	84.33	76.81	54.68	41.76
5	✓	✓	✓		55.77	43.39	35.60	21.26	89.43	**83.19**	**61.85**	44.69
6	✓	✓	✓	✓	**58.57**	**44.75**	**36.33**	**22.98**	**89.91**	83.07	59.66	**47.78**

Table 2: Evaluation results (%) on Charades-CD to investigate the effects of NA and curriculum strategies.

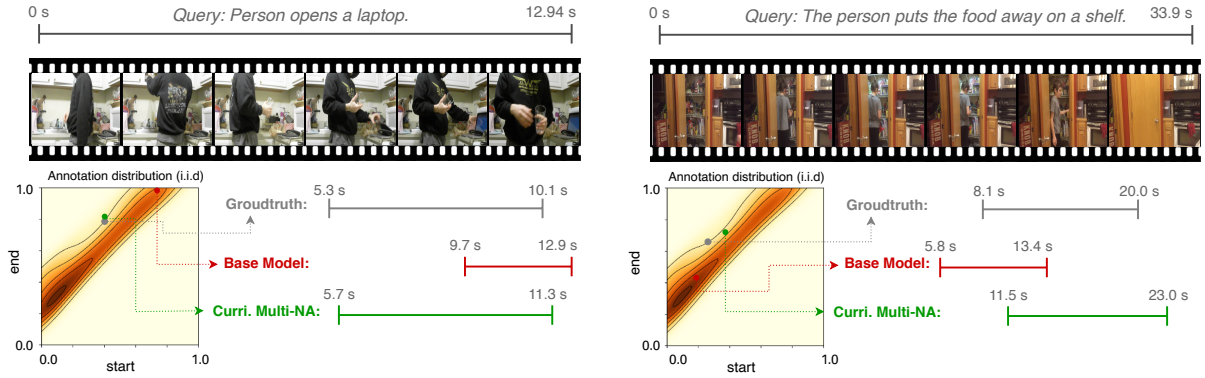


Figure 3: The qualitative analysis of VG samples from the *test-ood* set of Charades-CD.

We further compare our evaluation results with other VG models and find that even though the performance of the base model is behind SCDM and DRN, our method can achieve competitive results with SOTA models on both datasets. This observation proves that our debiasing strategy is able to improve the grounding results and make up shortcomings of the model design from another perspective.

Ablation Studies

To further validate the contributions of each component in our proposed Curriculum Multi-NA strategy, we conduct more studies and report the results in Table 2, based on which we have the following observations:

- It is observed that each type of NA (*c.f.*, No.2 – 4) can achieve better grounding results upon the base model (*c.f.*, No.1), which shows the proposed augmentation strategy for debiasing can effectively promote the grounding accuracy from the perspective of enriching the label space.
- It is worth noting that the only video-level augmentation works the best (*c.f.*, No.3) with the largest performance improvement compared to other two clip-level augmentations (*c.f.*, No.2 and No.4). The possible reason is that the video-level augmentation can maintain the temporal semantic information in video contents so distinguishing this kind of NA from the positive video is more useful for the model to capture the visual semantics.
- We can also observe that our designed curriculum strategy is able to further promote the grounding results comparing the evaluation results of No.5 and No.6, where No.5 fixes λ during the whole training process. The results show that the multi-stage curriculum strategy for denoising can further bring performance gains for more accurate grounding.

Qualitative Results

We report the qualitative results of VG samples whose temporal locations appear rarely in the training i.i.d set (*c.f.*, Figure 3). For these VG samples, the model is not able to exploit the location bias to accurately ground the sentence query. It can be shown that our Curriculum Multi-NA framework can achieve better grounding results compared to the base model. The base model seems to be affected by the annotation distribution biases, having the tendency to predict the locations from the area of higher density, while our proposed framework is able to perform the less biased prediction that has large IoU with the groundtruth moment.

Conclusion and Future Work

In this paper, we present a curriculum learning-driven data augmentation-based debiasing method to alleviate the temporal annotation bias issue in VG. The VG-specific data augmentation strategy can diversify its data distribution in the video-sentence label space while the proposed curriculum strategy can reduce the effects brought by the augmented noisy data samples. For the future work, we hope to explore the possibility of adopting such debiasing strategy in other related domains like video question answering or temporal action localization in videos.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China No. 2020AAA0106300, National Natural Science Foundation of China (No. 62250008, 62222209, 62102222, 62206149, 61872215), Shenzhen Science and Technology Program (Grant No. RCYX20200714114523079)

References

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 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.

Bao, P.; and Mu, Y. 2022. Learning Sample Importance for Cross-Scenario Video Temporal Grounding. *arXiv preprint arXiv:2201.02848*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 41–48.

Carreira, J.; and Zisserman, A. 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.

Chen, H.; Chen, Y.; Wang, X.; Xie, R.; Wang, R.; Xia, F.; and Zhu, W. 2021a. Curriculum Disentangled Recommendation with Noisy Multi-feedback. In *Proceedings of the International Conference on Neural Information Processing Systems*, 26924–26936.

Chen, J.; Chen, X.; Ma, L.; Jie, Z.; and Chua, T.-S. 2018. Temporally Grounding Natural Sentence in Video. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 162–171.

Chen, Y.; Wang, X.; Fan, M.; Huang, J.; Yang, S.; and Zhu, W. 2021b. Curriculum meta-learning for next POI recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2692–2702.

Ding, X.; Wang, N.; Zhang, S.; Cheng, D.; Li, X.; Huang, Z.; Tang, M.; and Gao, X. 2021. Support-set based cross-supervision for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11573–11582.

Duan, X.; Huang, W.; Gan, C.; Wang, J.; Zhu, W.; and Huang, J. 2018. Weakly Supervised Dense Event Captioning in Videos. In *Proceedings of the International Conference on Neural Information Processing Systems*, 3063–3073.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. TALL: Temporal Activity Localization via Language Query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5277–5285.

Ghosh, S.; Agarwal, A.; Parekh, Z.; and Hauptmann, A. 2019. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1984–1990.

Gong, C.; Tao, D.; Maybank, S. J.; Liu, W.; Kang, G.; and Yang, J. 2016. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing*, 25(7): 3249–3260.

Guo, S.; Huang, W.; Zhang, H.; Zhuang, C.; Dong, D.; Scott, M. R.; and Huang, D. 2018. Curriculumnet: Weakly supervised learning from large-scale web images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 135–150.

Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. C. 2017. Localizing Moments in Video with Natural Language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5804–5813.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Niebles, J. C. 2017. Dense-Captioning Events in Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 706–715.

Kumar, G.; Foster, G.; Cherry, C.; and Krikun, M. 2019. Reinforcement Learning based Curriculum Optimization for Neural Machine Translation. In *Proceedings of NAACL-HLT*, 2054–2061.

Lan, X.; Yuan, Y.; Wang, X.; Chen, L.; Wang, Z.; Ma, L.; and Zhu, W. 2022. A Closer Look at Debiased Temporal Sentence Grounding in Videos: Dataset, Metric, and Approach. *arXiv preprint arXiv:2203.05243*.

Lan, X.; Yuan, Y.; Wang, X.; Wang, Z.; and Zhu, W. 2021. A survey on temporal sentence grounding in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*.

Lao, M.; Guo, Y.; Liu, Y.; Chen, W.; Pu, N.; and Lew, M. S. 2021. From Superficial to Deep: Language Bias driven Curriculum Learning for Visual Question Answering. In *Proceedings of the ACM International Conference on Multimedia*, 3370–3379.

Li, J.; Yang, T.; Ji, W.; Wang, J.; and Cheng, L. 2022a. Exploring Denoised Cross-Video Contrast for Weakly-Supervised Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19914–19924.

Li, S.; Li, C.; Zheng, M.; and Liu, Y. 2022b. Phrase-level Prediction for Video Temporal Localization. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 360–368.

Liu, M.; Wang, X.; Nie, L.; He, X.; Chen, B.; and Chua, T. 2018. Attentive Moment Retrieval in Videos. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 15–24.

Luo, F.; Chen, S.; Chen, J.; Wu, Z.; and Jiang, Y.-G. 2021. Self-supervised learning for semi-supervised temporal language grounding. *arXiv preprint arXiv:2109.11475*.

Mithun, N. C.; Paul, S.; and Roy-Chowdhury, A. K. 2019. Weakly Supervised Video Moment Retrieval From Text Queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11592–11601.

Nan, G.; Qiao, R.; Xiao, Y.; Liu, J.; Leng, S.; Zhang, H.; and Lu, W. 2021. Interventional Video Grounding with Dual Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2765–2775.

Otani, M.; Nakashima, Y.; Rahtu, E.; and Heikkilä, J. 2020. Uncovering Hidden Challenges in Query-Based Video Moment Retrieval. In *The British Machine Vision Conference (BMVC)*.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1532–1543.

Soldan, M.; Pardo, A.; Alcázar, J. L.; Caba, F.; Zhao, C.; Giancola, S.; and Ghanem, B. 2022. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5026–5035.

Song, Y.; Wang, J.; Ma, L.; Yu, Z.; and Yu, J. 2020. Weakly-Supervised Multi-Level Attentional Reconstruction Network for Grounding Textual Queries in Videos. *ArXiv preprint*, abs/2003.07048.

Tay, Y.; Wang, S.; Luu, A. T.; Fu, J.; Phan, M. C.; Yuan, X.; Rao, J.; Hui, S. C.; and Zhang, A. 2019. Simple and Effective Curriculum Pointer-Generator Networks for Reading Comprehension over Long Narratives. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 4922–4931.

Tran, D.; Bourdev, L. D.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4489–4497.

Wang, X.; Chen, Y.; and Zhu, W. 2021. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, X.; Chen, Y.; and Zhu, W. 2022. A Survey on Curriculum Learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9): 4555–4576.

Wang, Z.; Chen, J.; and Jiang, Y.-G. 2021. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *Proceedings of the ACM International Conference on Multimedia*, 1459–1468.

Wu, J.; Li, G.; Liu, S.; and Lin, L. 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.

Yang, X.; Feng, F.; Ji, W.; Wang, M.; and Chua, T.-S. 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.

Yuan, Y.; Lan, X.; Wang, X.; Chen, L.; Wang, Z.; and Zhu, W. 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.

Yuan, Y.; Ma, L.; Wang, J.; Liu, W.; and Zhu, W. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In *Proceedings of the International Conference on Neural Information Processing Systems*, 534–544.

Yuan, Y.; Mei, T.; and Zhu, W. 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.

Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense Regression Network for Video Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10284–10293.

Zhang, C.; Yang, T.; Weng, J.; Cao, M.; Wang, J.; and Zou, Y. 2022a. Unsupervised pre-training for temporal action localization tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14031–14041.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2021. Towards debiasing temporal sentence grounding in video. *arXiv preprint arXiv:2111.04321*.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12870–12877.

Zhang, Z.; Zhang, Z.; Wang, X.; and Zhu, W. 2022b. Learning to solve travelling salesman problem with hardness-adaptive curriculum. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9136–9144.

Zheng, M.; Huang, Y.; Chen, Q.; Peng, Y.; and Liu, Y. 2022. Weakly Supervised Temporal Sentence Grounding With Gaussian-Based Contrastive Proposal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15555–15564.

Zhou, Y.; Chen, H.; Pan, Z.; Yan, C.; Lin, F.; Wang, X.; and Zhu, W. 2022a. CurML: A Curriculum Machine Learning Library. In *Proceedings of the ACM International Conference on Multimedia*, 7359–7363.

Zhou, Y.; Wang, X.; Chen, H.; Duan, X.; Guan, C.; and Zhu, W. 2022b. Curriculum-nas: Curriculum weight-sharing neural architecture search. In *Proceedings of the ACM International Conference on Multimedia*, 6792–6801.