Mixup-Augmented Temporally Debiased Video Grounding with Content-Location Disentanglement

Xin Wang*†
Department of Computer Science and Technology, BNRist, Tsinghua University
xin_wang@tsinghua.edu.cn

Zihao Wu*
Department of Computer Science and Technology, Tsinghua University
wuzh22@mails.tsinghua.edu.cn

Hong Chen
Department of Computer Science and Technology, Tsinghua University
h-chen20@mails.tsinghua.edu.cn

Xiaohan Lan
Department of Computer Science and Technology, Tsinghua University
lanxh20@tsinghua.org.cn

Wenwu Zhu†
Department of Computer Science and Technology, BNRist, Tsinghua University
wwzhu@tsinghua.edu.cn

ABSTRACT

Video Grounding (VG), has drawn widespread attention over the past few years, and numerous studies have been devoted to improving performance on various VG benchmarks. Nevertheless, the label annotation procedures in VG produce imbalanced query-moment-label distributions in the datasets, which severely deteriorate the learning model's capability of truly understanding the video contents. Existing works on debiased VG either focus on adjusting the learning model or conducting video-level augmentation, failing to handle the temporal bias issue caused by imbalanced query-moment-label distributions. In this paper, we propose a Disentangled Feature Mixup (DFM) framework for debiased VG, which is capable of performing unbiased grounding to tackle the temporal bias issue. Specifically, a feature-mixup augmentation strategy is designed to generate new (text, location) pairs with diverse temporal distributions via jointly augmenting the representation of text queries and the location labels. This strategy encourages making prediction based on more diverse data samples with balanced query-moment-label distributions. Furthermore, we also design a content-location disentanglement module to disentangle the representations of the temporal information and content information in videos, which is able to remove the spurious effect of temporal biases on video representation. Given that our proposed DFM framework conducts feature-level augmentation and disentanglement, it is model-agnostic and can be applied to most baselines simply yet effectively. Extensive experiments show that our proposed DFM framework is able to significantly outperform baseline models in various metrics under both independent identical distribution (i.i.d.) and out-of-distribution (o.o.d.) scenes, especially in scenarios with annotation distribution changes.

*Equal Contributions.
†Corresponding Authors. BNRist is the abbreviation of Beijing National Research Center for Information Science and Technology.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612401

Query:
A woman pours water into a cup, takes a sip and opens the window.

Figure 1: An example of Video Grounding (VG)

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Video Grounding (VG), Disentanglement, Debiasing Learning

ACM Reference Format:
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 (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612401

1 INTRODUCTION

Video Grounding (VG), has received an increasing amount of attention in recent years [8, 20, 25, 27, 57–59, 64, 67, 69]. Specifically, given a descriptive natural language sentence, the goal of the video sentence grounding task is to retrieve a video moment semantically matching the sentence query from the untrimmed video [37]. As shown in Figure 1, given a query "A woman pours water into a cup, takes a sip and opens the window.", VG aims to return a moment (3.2s-16.7s) which contains the semantics indicated by the query.

Nevertheless, the label annotation procedures in VG produce imbalanced query-moment-label distributions in the datasets, which severely deteriorate the learning model's capability of truly understanding the video contents. With these imbalanced query-moment-label distributions, current VG approaches may resort to "shortcuts" solutions without obtaining genuine comprehension of the videos

and texts, but rather achieving good results via overfitting the temporal biases caused by the imbalanced distributions, which has been previously discussed [1, 39, 46, 60, 61]. When the training and test sets of VG datasets are independently and identically distributed (i.i.d.) and have obvious biases (e.g., long duration moments that have higher predictability with a higher IoU), VG models may heavily overfit the biases in groundtruth moment annotations, leading to inadequate utilization of multimodal inputs. Consequently, VG approaches suffer from overfitting and spurious correlation due to distribution biases in existing VG datasets.

There exist several works on debiased VG [35, 36, 46, 47, 60, 61, 70], where they either focus on adjusting the learning models [36, 60] or conducting video-level data augmentations [35]. As such, these existing works fail to handle the temporal bias issue caused by imbalanced query-moment-label distributions.

Therefore, we study temporally debiased video grounding through moment-level augmentation, which poses two challenges:

(1) It is non-trivial to augment positive moment-level samples to conduct unbiased training for temporally debiased VG.
(2) It is unclear how to eliminate the spurious correlations between multimodal input and the target moment location, which may lead to the temporal biases.

To solve these challenges, in this paper, we propose a Disentangled Feature Mixup (DFM) framework for debiased VG, which is capable of performing unbiased grounding to tackle the temporal bias issue. In concrete, to obtain positive moment-level samples, we design a feature mixup augmentation strategy which can jointly augment the representation of text queries and the location labels to produce new (text, location) pairs with diverse temporal distributions for video moments. Therefore, the proposed feature-mixup augmentation is able to enrich the sample space with more diversity, thus encouraging the model to make predictions based on more diverse data samples with balanced query-moment-label distributions. To eliminate the spurious correlations between multimodal inputs and target moment location, we further develop a content-location disentanglement module to separate the latent representations of the temporal information and content information in videos via reconstruction and information bottleneck. Thus, we are able to remove the spurious effects from temporal information through forcing the model to focus more on the true relation between multimodal input and the target moment location. Given the feature-level augmentation and disentanglement, our proposed DFM framework is model-agnostic and can be applied to most baseline models simply yet effectively. Extensive experiments show that our proposed DFM framework achieves superior performances over state-of-the-art baselines on several datasets, especially in scenarios with annotation distribution changes.

To summarize, this paper makes the following contributions.

- We propose a Disentangled Feature Mixup (DFM) framework, which is able to perform unbiased video grounding to eliminate the temporal biases.
- We design a feature-mixup augmentation strategy which can produce many new (text, location) pairs with diverse temporal distributions for video moments in order to conduct unbiased training.
- We develop a content-location disentanglement module to separate the representations of the temporal information and content information in videos so that spurious correlation between multimodal input and target moment location can be eliminated.
- We conduct extensive experiments on several datasets to demonstrate the superiority of our DFM framework over existing baselines for VG tasks.

2 RELATED WORK

Video Grounding. On the one hand, many promising research works have emerged, which have continuously improved model performances for VG [8, 20, 25, 27, 57–59, 64, 67, 69]. On the other hand, some recent studies [46, 61] point out significant distribution biases in existing VG datasets. In particular, Yuan et al. [61] reorganize the two benchmark datasets to create two different test sets (i.i.d. and o.o.d.). Otani et al. [46] further propose two alternative evaluation metrics to handle subjective bias and mislabeling in the VG dataset. Similarly, Yuan et al.[61] also propose a new metric, discounted-R@n, IoU@m to alleviate the bias inherent in the evaluation metric (R@n, IoU@m) for VG. Yang et al. [60] propose a debiased cross-modal matching network to eliminate the confounding effect of temporal location for VG. Lan et al. [36] later propose a causality-based multi-branch de-biasing (MDD) framework for VG to remove the effects caused by multiple confounders and help the model better match the semantics between queries and clips. In addition, Zhou et al. [70] focus on another type of bias in the VG task, i.e., the single style of annotation, through proposing DeNet with a debiasing mechanism to produce diverse and reasonable predictions. Nan et al. [47] propose a method for approximating the distribution of potential confusion sets based on causal inference to eliminate the selection biases introduced by the datasets. To the best of our knowledge, our method is the first to augment the biased temporal annotations from the *moment-level* and explicitly eliminates the spurious effect of location variables simultaneously.

Disentangled Representation Learning. Disentangled representation learning aims to identify and disentangle the underlying explanatory factors [2]. In general, variational methods are widely applied for disentangled representation over images. β-VAE [22] demonstrates that disentanglement can emerge once the KL divergence term in the VAE [31] objective is aggressively penalized. In particular, Kingma and Welling [31] propose to utilize Bayesian posterior inference and variational estimation to learn the controllable factors hidden in the observed data. Higgins et al. [22] propose $\beta - VAE$ by setting a weight β for the KL divergence to improve representation disentanglement learned in the observed data while sacrificing mutual information between input data and latent representations. Later approaches separate the information bottleneck term [51, 52] and the total correlation term, and achieve a greater level of disentanglement [4, 9, 30]. Other works either design an attentive architecture to learn aspect matrix for word embeddings [21] or utilize methods based on triplets to learn aspect representations from sentences where each aspect has a separate encoder [26]. The majority of the existing efforts are from the field of computer vision [3, 10, 12–14, 22, 23, 28, 32, 33, 44, 71]. Disentangled representation learning on relational data, such as graph-structured

data, has not been explored until recently [38, 42, 56, 68]. Besides the relational data on graph, disentangled representation learning for recommendation has also drawn research attentions from the community [6, 43, 54, 55].

Mixup. Data Augmentation is a strategy for increasing the diversity of training instances without explicitly collecting new data [15]. As a data augmentation method, *mixup* was first introduced in the computer vision community by Zhang et al. [65]. The *mixup* training strategy is simple to implement and has only a small computational overhead, but greatly improves generalization errors on state-of-the-art models, and also enhances robustness for corrupted labels, and adversarial samples, and stabilizes the training of generative adversarial networks(GANs) [65]. Given the original sample, Cheng et al. [11] first construct an adversarial sample, and then applied two types of samples with the mixup strategies. Sun et al. [50] propose Mixup-Transformer, which combines mixup with a Transformer-based pre-training structure to validate its performance on a text classification dataset. Mixup in NLP can effectively improve data augmentation performance and has significant effects in avoiding model overfitting [7, 45, 65]. In the multimodal community, Li et al. [40] adopt mixup as a baseline augmentation method which serves as an equivariant intervention to the training video through applying a linear interpolation on the causal scene, question, and answer. Researchers have also proposed variants of mixup for speech [29] and a nonlinear hybrid scheme [18], etc.

3 METHOD

In this section, we present the details of our proposed Disentangled Feature Mixup (DFM) framework, covering notations and problem formulation, mixup augmentation, content-location disentanglement as well as learning objectives.

3.1 Notations and Problem Formulation

As shown in the example in Figure 1, given an uncropped video V and a sentence S as a query, the VG task is to retrieve one or more consecutive clips forming the moment M which semantically matches the query. Specifically, the query sentence is denoted as $S = \{s_i\}_{i=0}^{l^S-1}$, where s_i represents a word in the sentence and l^S is the total number of words. The input video consists of a sequence of frames, i.e., $V = \{x_i\}_{i=0}^{l^V-1}$, where x_i represents a frame in the video, and l^V is the total number of frames. The output moment M starts from frame x_{M_s} and ends with frame x_{M_e}, providing the semantics that match the input sentence S.

The given video V is first divided into a set of video clip anchors with different time lengths, i.e., $\{C_i\}_{i=0}^{N-1}$ where N is the number of candidate clips. Each candidate clip corresponds to a temporal position, i.e., the starting and ending timestamp, and several consecutive clips may form one moment. The correlation score between each candidate moment M and the ground truth moment M^*, measured in terms of temporal *Intersection over Union* (IoU) between M and M^*, is used as the supervision signal for training. A cross-modal function $\mathcal{F}$ will be learned to map each <query, moment> pair to a real value, indicating their correlation scores.

Temporal Biases in VG. The moments are actually located via the temporal coordinate of (start time, end time) within the given video. The temporal biases refer to the phenomenon that it is possible to correctly guess the (start time, end time) coordinate simply from the words or sentences of the query without truly understanding the real meaning of the query and the content of the video. This temporal biases are caused by the spurious correlation between the multimodal input and the target moment location.

3.2 Mixup Augmentation

We innovatively conduct *mixup* augmentation to text queries and corresponding temporal IoU labels, which simultaneously enhances the representation of text queries [19] as well as generate diverse data by mixing samples with different temporal locations. We employ two mixup strategies, i) Word Mixup which performs sample interpolation in the word embedding space, ii) Sentence Mixup which performs sample interpolation on the latent space projected through the sentence encoder.

Word Mixup. Word Mixup operates on the embedding of words. All sentences are first zero-padded to the same length, and then interpolated for each dimension of each word in the sentence. Given a piece of text, e.g., a sentence with N words, it can be represented as a matrix $B \in R^{N\times d}$. Each row w of the matrix corresponds to a word, denoted as B_w, which is represented as the word embedding table provided by the learned word or as a randomly generated d-dimension vector. Formally, consider a pair of samples $(B^i; y^i)$ and $(B^j; y^j)$, where B^i and B^j denote the embedding vectors of the input sentence pairs, and y^i and y^j denote the sample class labels represented using one-hot vectors. Then, for the w^{th} word in the sentence, the process of sample linear interpolation in the word embedding space can be expressed as follows,

$$\tilde{B}_w^{ij} = \lambda B_w^i + (1-\lambda) B_w^j, \quad \tilde{y}^{ij} = \lambda y^i + (1-\lambda) y^j. \tag{1}$$

The samples $(\tilde{B}_w^{ij}, \tilde{y}^{ij})$ generated by mixup serve as the subsequent training samples.

Sentence Mixup. Sentence Mixup operates on the latent space projected through the sentence encoder. The latent embeddings of two sentences with the same dimension are generated by the encoder (e.g., LSTM), followed by a subsequent linear interpolation on the sentence embeddings. Specifically, f denotes the sentence encoder, which encodes a pair of sentences B^i and B^j to sentence embeddings $f(B^i)$ and $f(B^j)$, with each dimension k of the latent sentence embedding being mixed, as shown in Eq.(2),

$$\tilde{B}_{\{k\}}^{ij} = \lambda f(B^i)_{\{k\}} + (1-\lambda) f(B^j)_{\{k\}}, \tilde{y}^{ij} = \lambda y^i + (1-\lambda) y^j. \tag{2}$$

Finally, the embedding vector $\tilde{B}_{\{k\}}^{ij}$ is passed to the softmax layer to generate the predictive distribution over the possible target classes.

Previous works on VG use each (sentence, moment) pair to learn multimodal representations and predict matching relationships separately during training, via utilization of the corresponding temporal locations as supervision. These models tend to be strapped in a limited and specific temporal location solution space, i.e., fitting the labeling bias of temporal locations in the dataset. In contrast, conducting *mixup* strategy to textual queries and their corresponding temporal IoU labels can enable predicting (sentence, moment) pairs at various locations in the same video, thus reducing empirical risk [65]. In general, our utilization of mixup strategy in VG can encourage the model to learn cross-modal matching relationships without temporal biases in a larger solution space, benefiting in improvement of the generalization ability.

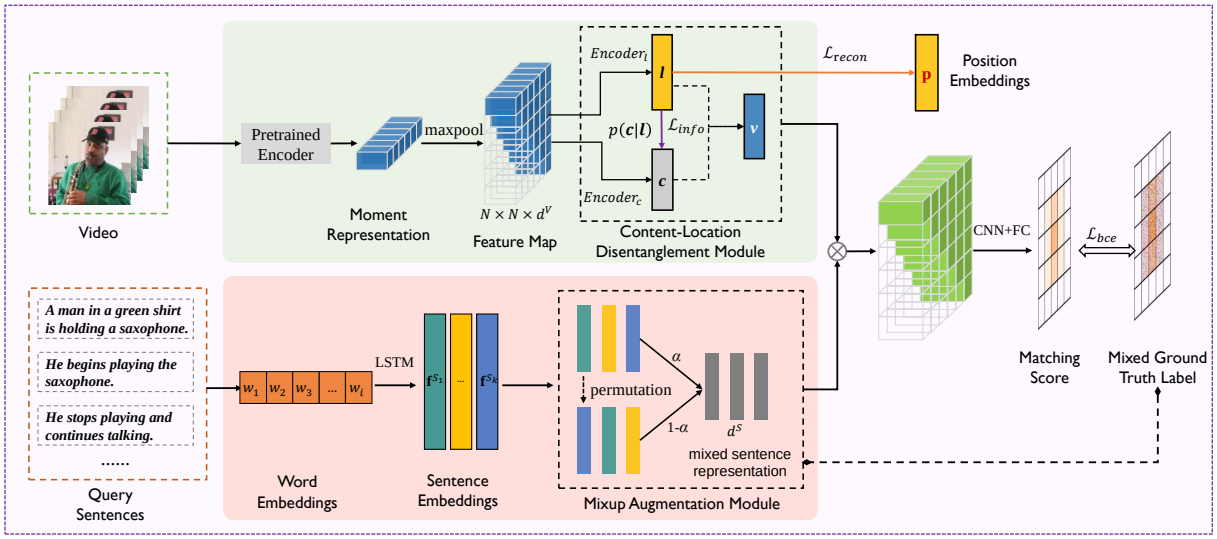

Figure 2: The overall structure of our proposed Disentangled Feature Mixup (DFM) framework for debiased VG. We conduct the content-location disentanglement and mixup augmentation in an end-to-end paradigm. The mixup augmentation module jointly augments the representation of text queries and the location labels to incorporate rich context information with less biased location reliance. The content-location disentanglement module removes the spurious effect of temporal location and enhance the visual representation.

3.3 Content-Location Disentanglement

It is very likely that the supervision signals lying in the 2D coordinate map may introduce specific location information during the model training process, leading to the temporal contextual bias between moment and its location that may influence the clip representations.

To address this issue, we propose to disentangle video representations $\mathbf{v}$ into mutually independent potential variables, i.e., content representation $\mathbf{c} \in \mathbb{R}^d$ and location representation $\mathbf{l} \in \mathbb{R}^d$, through disentangled representation learning as follows,

$$\mathbf{c} = Encoder_c(\mathbf{v}), \quad \mathbf{l} = Encoder_l(\mathbf{v}), \tag{3}$$

where $Encoder_c$ and $Encoder_v$ can be implemented using any deep architectures such as multi-layer perceptron (MLP). Specifically, we achieve the disentanglement of content and location representations by resorting to i) reconstruction constraints and ii) information constraints.

Reconstruction Constraints. In disentangled representation learning, reconstruction loss is crucial for ensuring that the representations learned by the model are semantically meaningful. For VG, there exists an inherent timestamp (x_i, x_j) that can be naturally used as a supervised signal of the reconstructed temporal location representation $\mathbf{l}$. The reconstructing loss function $\mathcal{L}_{recon}(\mathbf{l}, \mathbf{p})$ encourages temporal position representation $\mathbf{l}$ to approximate the real temporal location representation $\mathbf{p}$, where $\mathbf{p} \in \mathbb{R}^d$ is the intractable position embedding vector [31] obtained from the video representations (similar to position embedding in Transformer). For simplicity, we use L_2 distance for the reconstruction loss as follows,

$$\mathcal{L}_{recon}(\mathbf{l}, \mathbf{p}) = \|\mathbf{l} - \mathbf{p}\|_2. \tag{4}$$

Information Constraints. In order to achieve content-location disentanglement, content representation $\mathbf{c}$ and location representation $\mathbf{l}$ from the video representation $\mathbf{v}$ are supposed to satisfy the independence constraint after being extracted from the video representation, i.e., $\mathbf{c}$ and $\mathbf{l}$ should be invariant and contain the least information from each other. The information loss function $\mathcal{L}_{info}(\mathbf{c}, \mathbf{l})$ encourages the content representation c and location representations l to be independent with each other in the hidden space. This independence can be realized via minimizing the mutual information $\mathbf{MI}(\mathbf{c}; \mathbf{l})$ between content c and location l. Meanwhile, we maximize $\mathbf{MI}(\mathbf{c}; \mathbf{v})$ to ensure that the content embedding c sufficiently encapsulates information from the video v. Therefore, our overall information constraint objective is as follows,

$$\begin{aligned}\mathcal{L}_{info}(\mathbf{c}, \mathbf{l}) &= \mathbf{MI}(\mathbf{c}; \mathbf{l}) - \mathbf{MI}(\mathbf{c}; \mathbf{v}) \\ &= D_{\mathrm{KL}}\left(p(\mathbf{c}, \mathbf{l}) || p(\mathbf{c})p(\mathbf{l})\right) - D_{\mathrm{KL}}\left(p(\mathbf{c}, \mathbf{v}) || p(\mathbf{c})p(\mathbf{v})\right) \\ &= \mathbb{E}_{p(\mathbf{c},\mathbf{l})}\left[\log \frac{P(\mathbf{c}, \mathbf{l})}{P(\mathbf{c})P(\mathbf{l})}\right] - \mathbb{E}_{p(\mathbf{c},\mathbf{v})}\left[\log \frac{P(\mathbf{c}, \mathbf{v})}{P(\mathbf{c})P(\mathbf{v})}\right].\end{aligned} \tag{5}$$

However, the mutual information is usually difficult to calculate in practice. To estimate mutual information $\mathbf{MI}(\mathbf{x}, \mathbf{y})$, the variational distribution $q(\mathbf{x}|\mathbf{y})$ is introduced as follows,

$$\mathbf{MI}(\mathbf{x}, \mathbf{y}) \geq \mathbf{H}(\mathbf{x}) + \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\log q(\mathbf{x}|\mathbf{y})\right],$$

where $H(\mathbf{x}) = \mathbb{E}_{p(\mathbf{x})}\left[-\log p(\mathbf{x})\right]$ is the entropy of variable $\mathbf{x}$. Thus, we can get an upper bound of Eq.(5):

$$\mathcal{L}_{info} \leq \mathbf{MI}(\mathbf{c}; \mathbf{l}) - \left[\mathbf{H}(\mathbf{v}) + \mathbb{E}_{p(\mathbf{c},\mathbf{v})}\left[\log q_\phi(\mathbf{v}|\mathbf{c})\right]\right]. \tag{6}$$

Note that $\mathbf{H}(\mathbf{v})$ is constant, we only need to minimize

$$\tilde{\mathcal{L}}_{info} = \mathbf{MI}(\mathbf{c}; \mathbf{l}) - \mathbb{E}_{p(\mathbf{c},\mathbf{v})}\left[\log q_\phi(\mathbf{v}|\mathbf{c})\right], \tag{7}$$

where the content representation $\mathbf{c}$ and location representation $\mathbf{l}$ are expected to be independent by minimizing mutual information $\mathbf{MI}(\mathbf{c}; \mathbf{l})$, while the content representation $\mathbf{c}$ should maximize the log-likelihood $\mathbb{E}_{p(\mathbf{c},\mathbf{v})}\left[\log q_\phi(\mathbf{v}|\mathbf{c})\right]$ to contain sufficient information from the video $\mathbf{v}$.

Algorithm 1 The DFM framework for debiased VG.

input: a series of sentence queries, video clip pairs and their correlational scores $\mathcal{S}$, $\mathcal{M}$, $\mathcal{Y}$, the given video V
output: the predicted matching score map P
$\mathbf{f}^V = \text{I3D}\,(V)$.
▷ Extract video features with I3D or C3D algorithms
for m_i in $\mathcal{M}$ **do**
  $a = m_i$.start_index.
  $b = m_i$.end_index.
  $\mathbf{f}^{Vp}_{a,b} = \text{maxpool}\left(\mathbf{f}^V_a,\ \mathbf{f}^V_{a+1}, \ldots,\ \mathbf{f}^V_b\right)$.
▷ Get video clip features by max pooling operation
$\mathbf{F}^V_{2D} = \text{temporal}\,(\mathbf{f}^{Vp})$, where $\mathbf{F}^V_{2D}\,[a,b,:] = \mathbf{f}^{Vp}_{a,b}$.
▷ Organize the video features into a 2D feature map
$\mathbf{f}^c,\ \mathbf{f}^l = \text{Disentangle}_v\left(\mathbf{F}^V_{2D}\right)$.
for s_i in $\mathcal{S}$ **do**
  $w_i = \text{GloVe}(s_i)$.
▷ Get the embedding of each word in the query sentence
$\mathbf{f}^S = \text{LSTM}\left(\ \{w_i\}_{i=0}^{l^S-1}\right)$
▷ Input word embeddings into LSTM to obtain query sentence representations
$\mathbf{f}^S_{mixup} = \text{mixup}(\mathbf{f}^S, y)$
▷ Mix the representations of query sentences from the same video, and the corresponding IoU scores are used as labels
$\mathbf{F}^{fuse} = \mathbf{f}^c \bigotimes \mathbf{f}^S_{mixup}$
▷ Fuse representations of two modal with Hadamard products
$\mathbf{F}^{match} = \text{TAN}\left(\mathbf{F}^{fuse}\right)$
▷ Obtain 2D matching feature maps
$P = \text{Sigmoid}\left(\text{FC}\left(\mathbf{F}^{match}\right)\right)$
▷ Obtain matching score by FC layer and sigmoid function
return P

3.4 Learning Objectives

To train the proposed DFM framework, we employ the scaled temporal IoU as the supervised signal, which is subject to two thresholds t_{min} and t_{max},

$$y_i = \begin{cases} 0 & o_i \le t_{min} \\ \dfrac{o_i - t_{min}}{t_{max} - t_{min}} & t_{min} < o_i < t_{max} \\ 1 & o_i \ge t_{max} \end{cases}, \tag{8}$$

where o_i is the temporal intersection ratio between a candidate moment and the true ground truth moment.

The binary cross-entropy loss function (BCE) is used, which can be expressed as follows,

$$\mathcal{L}_{bce} = \frac{1}{C}\sum_{i=1}^{C} y_i \log p_i + (1 - y_i)\log(1 - p_i)\,, \tag{9}$$

where p_i denotes the predicted match score of a video clip.

In summary, a linear combination of three losses is employed to train our proposed DFM framework as follows,

$$\mathcal{L} = \mathcal{L}_{\text{bce}} + \lambda_1 \mathcal{L}_{recon} + \lambda_2 \mathcal{L}_{info}\,, \tag{10}$$

where λ_1 and λ_2 are hyperparameters that control the reconstruction loss $\mathcal{L}_{recon}$ and the information constraint loss $\mathcal{L}_{info}$ in the disentanglement module, respectively.

3.5 Disentangled Feature Mixup for Debiasing

The overall framework of our proposed Disentangled Feature Mixup framework for debiasing VG is shown in Figure 2. The input uncropped video is first passed through a pre-trained convolutional neural network to extract features $\mathbf{f}^V \in \mathbb{R}^{d^V}$, where d^V represents the dimensionality of the video representation. We specifically use C3D [53] and I3D [5] as the video understanding model to extract the video clip features, the process of which is described in detail in Section 4. The extracted video clip features are processed by max pooling operation and temporal position arrangement to obtain a two-dimensional temporal feature map as the final video representation $\mathbf{F}^V_{2D} \in \mathbb{R}^{N\times N\times d^V}$. The position on the feature map (i, j) represents of a candidate moment which starts at $i\tau$ and ends at $(j+1)\tau$. The two-dimensional temporal feature map is then fed into the content-location disentanglement module, which disentangles the video representation into content representation and location representation via reconstruction and information constraints. The input query sentence S is firstly passed to the pre-training word vector model (GloVe model [48]) to obtain the word embedding of each word $\{w_i\}_{i=0}^{l^S-1}$. The word embedding vectors are then fed sequentially into the LSTM network to obtain the representations of the corresponding sentences $\mathbf{f}^S \in \mathbb{R}^{d^S}$, where d^S represents the dimensionality of the sentence representations. The extracted representations encode the linguistic structure of the query sentence, thus describing the corresponding target video clip.

Subsequently, based on the two mixup strategies, either word embeddings (Word Mixup) or sentence embeddings (Sentence Mixup) are fed into the mixup augmentation module to mix multiple query embeddings with IoU scores accordingly. The mixup augmentation encourages the model to make predictions for queries involving different locations, thus reducing temporal biases and enhancing the generalization of the model. In this process, the query embeddings for mixup are ensured to come from the same video and are matched with their corresponding moments for cross-modal prediction. The content representations from the disentanglement and the query representations after the mixup augmentation are normalized and fused across modalities by Hadamard product operations to obtain a two-dimensional feature map. Temporal Adjacent Network (TAN) [67] is then employed to enable the two-dimensional fused feature map to obtain matching results through a convolutional neural network, and the output of the network maintains the same shape as the input fused feature map by complementary zeros.

Finally, the matching scores of candidate moments with a given query sentence are predicted based on a two-dimensional temporal graph. The output features of the temporal adjacency network are passed through a fully connected layer and a Sigmoid function to generate a two-dimensional score graph. Each value on the score graph p_i represents the match score between the candidate moment and the query sentence, and the maximum value corresponds to the best matched moment. Algorithm 1 presents the implementation details of our proposed DFM framework.

4 EMPIRICAL EXPERIMENTS

We conduct experiments to test our proposed DFM framework on popular datasets covering diverse video scenes, different data distributions, with multiple evaluation metrics.

4.1 Experimental Setup

Datasets. Three popular datasets, TACoS [49], ActitvityNet Captions [34], and Charades-STA [16], are used for the experiments. In addition, following existing literature [61], we also adopt the reorganized datasets based on ActivityNet Captions and Charades-STA, which are denoted as ActivityNet-CD and Charades-CD, with varying distributions. Specifically, each dataset is repartitioned into four groups, i.e., the training set, the validation set, the i.i.d. test set, and the o.o.d. test set. All samples in the training, validation, and i.i.d. test sets follow independent identical distributions, and the samples in the o.o.d. test set are in an out-of-distribution setting. Clearly, the performance gap between the i.i.d. test set and the o.o.d. test set can be used to effectively assess the generalization ability of the model. Further details of the data repartitioning process are described below.

Feature Extraction. The feature extraction of query text for VG is performed by the GloVe word embedding model [48]. A 300-dimensional word vector pre-trained on a corpus of 840B size by the official open source project is used in our experiment. For each word in the query sentence S, its word embedding is obtained as w_i, and then the word embedding vector $\{w_i\}_{i=0}^{l^S-1}$ is fed into the LSTM network in turn to obtain the representation of sentence S.

To represent videos, we use features extracted from the C3D model [53] for the TACoS and ActivityNet Captions datasets, and the I3D model [5] for the Charades-STA dataset.

Metrics. There are two common types of metrics for VG tasks, which are introduced in [16] for the first time. One of those is *mIoU* (i.e., average IoU), which simply takes the average of the temporal IoU of all test samples to evaluate the performance. Another commonly used metric is *R@n, IoU@m* [24]. For sample i, if there exists a moment with a temporal IoU exceeding m among top n retrieval moments, then $r(n, m, q_i) = 1$. Otherwise, $r(n, m, q_i) = 0$. *R@n, IoU@m* is the percentage of positive samples over all samples:

$$\text{R@n,IoU@m} = \frac{1}{N_q}\sum_i r(n, m, q_i), \tag{11}$$

where it is common practice to set $n \in \{1, 5, 10\}$ and $m \in \{0.3, 0.5, 0.7\}$. Usually, when the model uses a proposal-free (or anchor-free) approach, $n = 1$.

For different types of experiments, we use *R@n, IoU@m* and *mIoU* scores. Due to the difficulty of VG, previous works often choose an IoU threshold m of 0.1 or 0.3 to evaluate the prediction results, while such metrics often do not credibly and accurately reflect the true performance of the model in some cases [61]. So in our experiments we primarily choose $m \geq 0.5$ in *R@n, IoU@m* to achieve a more rigorous performance test. We also report the results with the unbiased metric *dR@n, IoU@m* [61] that further discounts the recall values of *R@n, IoU@m* based on temporal distances.

4.2 Model Performance

We compare DFM with several state-of-the-art models over the above benchmark datasets, including CTRL [16], ACRN [41], ABLR [63], SCDM [62], DRN [17], 2D-TAN [67] and VSLNet [66]. Our results show that DFM achieves the best performance on three benchmark datasets with different criteria in various scenarios. Notably, on the TACoS dataset, ActivityNet-CD dataset, and Charades-CD dataset, our DFM framework significantly outperforms the state-of-the-art models to a large extent. In addition, DFM also outperforms the top-ranked methods on both the i.i.d. and o.o.d. distributed test sets. This verifies that the proposed model can predict the query-moment matching patterns more accurately and obtain a better matching ability with better robustness and generalization.

Table 1: Overall performance (%) comparisons with other VG models on TACoS dataset (the best results are in BOLD, second in both BOLD and underline and third in underline).

Models	R@1			mIoU
	IoU@0.3	IoU@0.5	IoU@0.7	
CTRL [16]	18.32	13.30	-	-
ACRN [41]	19.52	14.62	-	-
ABLR [63]	18.90	9.30	-	13.40
SCDM [62]	26.11	21.17	-	-
DRN [64]	-	23.17	-	-
2D-TAN [67]	**37.29**	21.94	11.20	23.97
VSLNet [66]	29.61	24.27	**20.03**	24.11
DFM (+2D-TAN)	**40.04**	**28.57**	14.77	**27.35**
DFM (+VSLNet)	33.85	**29.40**	**22.36**	**27.29**

Results Analysis. We evaluate baseline models and the DFM framework on the TACoS dataset, respectively, to verify the effectiveness of the proposed framework on the dataset with original distributions. We use *R@n, IoU@m*, and *mIoU* as evaluation metrics. The experimental results are shown in Table 1. As can be observed from the experimental data, our proposed DFM achieves the best performance in almost all metrics, regardless of the values of n and m. This proves that the model obtains great improvement on the representation learning and cross-modal matching ability for VG.

Further, we conduct evaluations on the ActivityNet-CD and Charades-CD datasets with reorganized distributions [61] to verify the effectiveness of the proposed method on the i.i.d. and o.o.d. datasets, i.e., the ability in removing data bias. The overall results are shown in Table 2. We observe that the proposed DFM framework achieves excellent experimental performance on both datasets reorganized for the o.o.d. problem, ranking first in almost all metrics. DFM achieves superior results on both i.i.d. and o.o.d. test data, outperforming baseline approaches to a large extent. In particular, DFM is able to beat the baselines comprehensively on the o.o.d. test set, indicating its robustness in o.o.d. environment, and validating its capability of generalization as well as being resilient to the bias of temporal location annotations.

The results show that our proposed DFM framework plays an important role in enhancing the model's out-of-distribution generalization ability and its performance with respect to debiased VG.

Comparison with Existing Debiased Models. We investigate and compare DFM with several existing debiasing methods in detail, as shown in Table 2. A simple bias-based baseline model (Bias-based) [61] is employed as a comparative approach. The experimental results of the comparative analysis show that our proposed DFM framework achieves quite competitive results compared with various existing debiasing models, and achieves optimal or suboptimal performance in various evaluation metrics on the ActivityNet-CD

Table 2: Overall performance (%) comparisons with other VG models on Changing-Distribution datasets (the best results are in BOLD, second in both BOLD and underline and third in underline).

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 [16]	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 [41]	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 [63]	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 [62]	58.14	52.38	47.36	41.60	30.79	22.22	46.44	31.56	35.15	19.14	22.04	9.31
TSP-PRL [57]	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 [67]	53.71	43.45	46.48	30.77	28.76	13.73	49.18	30.86	40.87	18.86	27.36	9.77
VSLNet [66]	55.51	48.08	47.60	32.72	29.88	19.61	49.47	30.90	39.86	19.57	26.45	11.14
DRN [64]	51.35	40.45	41.91	30.43	26.74	15.91	48.92	36.86	39.27	**25.15**	25.71	**14.33**
TCN-DCM [60]	-	-	52.50	40.51	**35.28**	21.02	-	-	42.15	20.86	29.69	11.07
MDD [36]	-	-	52.78	40.39	34.71	**22.70**	-	-	43.63	20.80	**31.44**	11.66
Multi-NA [35]	64.21	52.21	53.82	39.86	34.47	21.38	49.91	32.32	41.67	20.78	28.82	11.03
DFM (+2D-TAN)	**64.52**	**55.10**	**56.38**	**44.01**	34.87	22.28	**58.84**	**40.27**	**45.92**	24.32	**32.18**	**12.72**
DFM (+VSLNet)	**64.50**	**56.49**	**57.97**	**41.65**	**35.37**	**23.34**	**57.21**	**38.82**	**46.05**	**25.27**	30.17	12.55

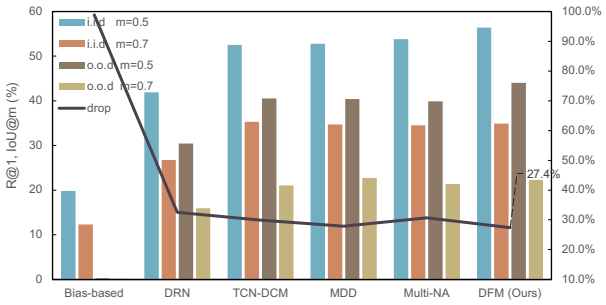

Figure 3: Comparison with other debiased VG models.

and Charades-CD datasets. This demonstrates the effectiveness and superiority of our proposed framework, and verifies that the mixup augmentation and content-location disentanglement modules achieve strong generalization ability at a low cost.

In addition, as shown in Figure 3, the *drop* metric indicates the decrease of model's performance from the i.i.d. test set to the o.o.d. test set, and is calculated as $drop = (iid\_score - ood\_score)/iid\_score$, which reflects the model's representation learning and matching performance for VG. If the model only fits the data distribution without learning the multimodal representation and matching patterns of query-video clip pairs well, it cannot generalize well to out-of-distribution data, and the performance reflected in the o.o.d. test set will be significantly degraded. Our proposed DFM also performs best on this metric, demonstrating that the DFM is effective at handling bias and learning essential cross-modal representations and matching patterns.

4.3 Ablation Studies

In this section, we evaluate the impact of different factors on the performance of the proposed DFM framework. We adopt the implementation based on 2D-TAN [67] to perform ablation studies.

Hyperparameter Analysis. Different implementations of the mixup augmentation module, i.e., *word mixup* and *sentence mixup*, tend to fuse and augment text embedding in different ways. As

Table 3: Ablation study on mixup type and size on the TACoS dataset.

mixup type	mixup size	R@1		R@5	
		IoU@0.5	IoU@0.7	IoU@0.5	IoU@0.7
-	32	21.91	11.20	44.06	22.44
word	32	22.06	11.22	45.24	23.10
	16	26.99	14.27	48.01	25.49
	8	28.15	14.76	48.30	25.97
sentence	32	22.51	11.84	45.46	24.32
	16	24.52	12.67	46.69	25.57
	8	**28.57**	**14.77**	**49.36**	**26.59**

such, different mixup augmentation strategies may generate text representation structures with different characteristics, imposing different effects on the model performances. Meanwhile, since the mixup augmentation is performed in each batch during the training process, too large batch size will lead to redundant mixing results, thus slowing down convergence of the model, while too small batch size will lead to insufficient number of samples available for augmentation, thus limiting the generalization of representations. We conduct ablation experiments on both implementations over the TACoS dataset, and the experimental results are presented in Table 3. The experimental results demonstrate that ii) the sentence mixup yields better performance than word mixup for DFM, and ii) using a relatively small mixup size (e.g., 8) during training also yields better results. This is consistent with the intuitive common sense, since sentence embeddings tend to be more representative of the semantics hidden in the entire query statement than word embeddings. Given that the sentence embeddings are better matched with the corresponding video clip, the sentence mixup augmentation will result in better representation quality.

Mixup hyperparameter α and the number of mixup augmentation *rounds* are two hyperparameters that need to be balanced. The original work on *mixup* [65] recommends to choose $\alpha \in [0.1, 0.4]$. However, mixup implicitly controls the complexity of the model,

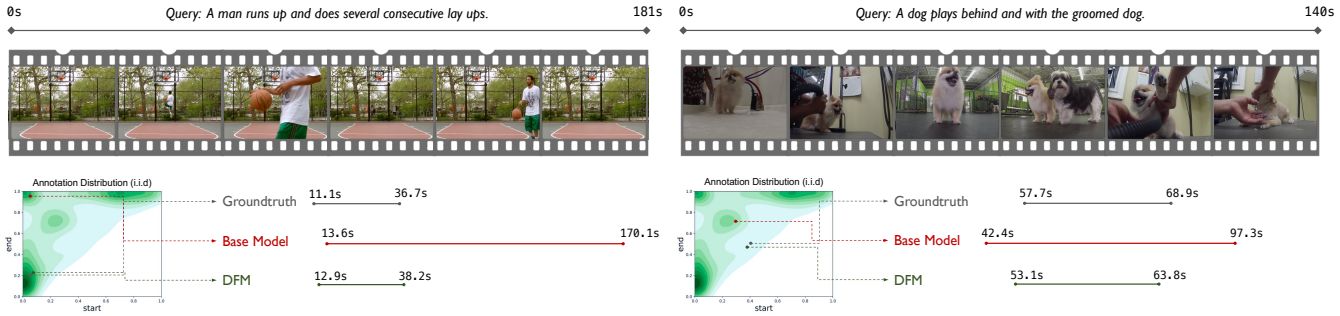

Figure 4: Qualitative analysis of VG cases from the *test-ood* set of ActivityNet-CD dataset.

Table 4: Ablation study on mixup α and *rounds* on the Charades-CD dataset.

mixup α	mixup *rounds*	dR@1,IoU@0.5		dR@1,IoU@0.7	
		i.i.d.	o.o.d.	i.i.d.	o.o.d.
0.2	1	53.46	43.21	31.11	19.16
1.3	1	54.07	42.91	33.78	19.31
8	1	53.22	40.18	31.59	17.14
0.2	3	51.76	40.57	33.17	19.07
1.3	3	54.68	40.87	33.29	20.08
8	3	**56.38**	**44.01**	**34.87**	**22.28**

where a larger α helps to enhance the generalizability of the model, while running the risk of underfitting and degrading the model performance. The number of mixup *rounds* refers to the number of batches that yield augmented samples. Smaller mixup *rounds* may not yield sufficiently rich samples for the model to learn more generalized representations, yet larger mixup *rounds* increase computational cost and induce the model to learn redundant features. Therefore, we investigate the ablation study of these two parameters on the Charades-CD dataset for o.o.d. problem in Table 4. The experimental results show that there is indeed a trade-off in choosing the hyperparameter α and *rounds*. A relatively larger α will eventually produce better performance, but a too large α exceeding a certain threshold will degrade the model performance, which also applies to mixup *rounds*. We finally select α = 8, *rounds* = 3 on the Charades-STA dataset.

Table 5: Performance comparison of DFM w/o mixup augmentation module and content-location disentanglement module on the Charades-CD dataset.

Model settings	dR@1,IoU@0.5		dR@1,IoU@0.7	
	i.i.d.	o.o.d.	i.i.d.	o.o.d.
base	46.48	30.77	28.76	13.73
base+mixup	54.73	40.74	33.05	19.96
base+mixup+disentangle	**56.38**	**44.01**	**34.87**	**22.28**

Module Analysis. In order to verify the contributions of each component of our proposed DFM framework, we conduct ablation experiments on each module of DFM, i.e., mixup augmentation module and content-location disentanglement module. As shown in Table 5, with the addition of the mixup augmentation module, the model achieves a considerable performance improvement. After further disentangling the representations of video content and temporal location, the model eliminates the location bias and achieves even better performance. The results show that both modules of DFM are coordinated and compatible, and contribute to achieving significant improvement on the representation learning and generalization performance of the VG model.

4.4 Qualitative Evaluation

We report the qualitative results of VG cases that are relatively difficult to ground (*c.f.* Figure 4). We observe that in some cases where the video content is indistinguishable throughout or the target moment location is infrequently spotted, the base model may tend to exploit the temporal distribution bias to return an unreliable prediction. Under these circumstances, our proposed DFM is able to utilize comprehensive and intrinsic representations within a larger context, and is less affected by the biased factors, thus achieving better performance compared to the base model.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we study temporally biased video grounding via feature-level mixup augmentation and content-location disentanglement. we propose a Disentangled Feature Mixup (DFM) framework for debiased VG, which is capable of performing unbiased grounding to tackle the temporal bias issue. Our proposed DFM framework conducts feature-level augmentation and disentanglement, capable of being applied to most baselines simply yet effectively. Experiments show that the proposed method achieves SOTA results in various metrics under i.i.d. and o.o.d. scenes. We conclude that our proposed DFM framework is superior over existing baslines both quantitatively and qualitatively.

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. 62222209, 62250008, 62102222), Beijing National Research Center for Information Science and Technology Grant No. BNR2023RC01003, BNR2023TD03006, and Beijing Key Lab of Networked Multimedia.

REFERENCES

[1] Peijun Bao and Yadong Mu. 2022. Learning Sample Importance for Cross-Scenario Video Temporal Grounding. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*. 322–329.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[3] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. 2018. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[4] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. 2018. Understanding disentangling in $beta$-VAE. *arXiv preprint arXiv:1804.03599* (2018).

[5] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[6] Hong Chen, Yudong Chen, Xin Wang, Ruobing Xie, Rui Wang, Feng Xia, and Wenwu Zhu. 2021. Curriculum disentangled recommendation with noisy multi-feedback. *Advances in Neural Information Processing Systems* 34 (2021), 26924–26936.

[7] Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2147–2157.

[8] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. 2020. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10551–10558.

[9] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. 2018. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*. 2610–2620.

[10] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*. 2172–2180.

[11] Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. AdvAug: Robust Adversarial Augmentation for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5961–5970.

[12] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. 2016. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648* (2016).

[13] Emilien Dupont. 2018. Learning disentangled joint continuous and discrete representations. In *Advances in Neural Information Processing Systems*. 710–720.

[14] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. 2018. Neural scene representation and rendering. *Science* 360, 6394 (2018), 1204–1210.

[15] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. (2021), 968–988.

[16] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: Temporal Activity Localization via Language Query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 5267–5275.

[17] Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind Picard. 2021. DISSECT: Disentangled Simultaneous Explanations via Concept Traversals. In *International Conference on Learning Representations*.

[18] Hongyu Guo. 2020. Nonlinear mixup: Out-of-manifold data augmentation for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4044–4051.

[19] Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941* (2019).

[20] Meera Hahn, Asim Kadav, James M Rehg, and Hans Peter Graf. 2019. Tripping through time: Efficient localization of activities in videos. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*.

[21] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[22] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, Vol. 3.

[23] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. 2018. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*. 517–526.

[24] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4555–4564.

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

[26] Sarthak Jain, Edward Banner, Jan-Willem van de Meent, Iain Marshall, and Byron C Wallace. 2018. Learning Disentangled Representations of Texts with Application to Biomedical Abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4683–4693.

[27] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. 2019. Cross-modal video moment retrieval with spatial and language-temporal attention. In *Proceedings of the 2019 on international conference on multimedia retrieval*. 217–225.

[28] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. 2017. Variational deep embedding: an unsupervised and generative approach to clustering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 1965–1972.

[29] Amit Jindal, Narayanan Elavathur Ranganatha, Aniket Didolkar, Arijit Ghosh Chowdhury, Di Jin, Ramit Sawhney, and Rajiv Ratn Shah. 2020. SpeechMix — Augmenting Deep Sound Recognition Using Hidden Space Interpolations. In *Proc. Interspeech 2020*. 861–865. https://doi.org/10.21437/Interspeech.2020-3147

[30] Hyunjik Kim and Andriy Mnih. 2018. Disentangling by Factorising. In *International Conference on Machine Learning*. 2654–2663.

[31] Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*.

[32] Nikos Komodakis and Spyros Gidaris. 2018. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*.

[33] Adam Kosiorek, Hyunjik Kim, Yee Whye Teh, and Ingmar Posner. 2018. Sequential attend, infer, repeat: Generative modelling of moving objects. In *Advances in Neural Information Processing Systems*. 8606–8616.

[34] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*. 706–715.

[35] 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*.

[36] Xiaohan Lan, Yitian Yuan, Xin Wang, Long Chen, Zhi Wang, Lin Ma, and Wenwu Zhu. 2022. A closer look at debiased temporal sentence grounding in videos: Dataset, metric, and approach. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022).

[37] Xiaohan Lan, Yitian Yuan, Xin Wang, Zhi Wang, and Wenwu Zhu. 2021. A survey on temporal sentence grounding in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2021).

[38] Haoyang Li, Xin Wang, Ziwei Zhang, Zehuan Yuan, Hang Li, and Wenwu Zhu. 2021. Disentangled contrastive learning on graphs. *Advances in Neural Information Processing Systems* 34 (2021), 21872–21884.

[39] 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.

[40] Yicong Li, Xiang Wang, Junbin Xiao, and Tat-Seng Chua. 2022. Equivariant and invariant grounding for video question answering. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4714–4722.

[41] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 15–24.

[42] Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. 2019. Disentangled graph convolutional networks. In *International conference on machine learning*. PMLR, 4212–4221.

[43] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. *Advances in neural information processing systems* 32 (2019).

[44] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 99–108.

[45] Vukosi Marivate and Tshephisho Sefara. 2020. Improving short text classification through global augmentation methods. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 385–399.

[46] Esa Rahtu Mayu Otani, Yuta Nakahima and Janne Heikkilä. 2020. Uncovering Hidden Challenges in Query-Based Video Moment Retrieval. In *The British Machine Vision Conference (BMVC)*.

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

[48] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

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

[50] Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, S Yu Philip, and Lifang He. 2020. Mixup-Transformer: Dynamic Data Augmentation for NLP Tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*. 3436–3440.

[51] N TISHBY. 1999. The Information Bottleneck Method. In *Proc. of the 37th Allerton Conference on Communication and Computation, 1999*.

[52] Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*. IEEE, 1–5.

[53] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[54] Xin Wang, Hong Chen, Yuwei Zhou, Jianxin Ma, and Wenwu Zhu. 2022. Disentangled representation learning for recommendation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 408–424.

[55] Xin Wang, Hong Chen, and Wenwu Zhu. 2021. Multimodal disentangled representation for recommendation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[56] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. 2020. Disentangled graph collaborative filtering. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 1001–1010.

[57] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. 2020. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12386–12393.

[58] Shaoning Xiao, Long Chen, Jian Shao, Yueting Zhuang, and Jun Xiao. 2021. Natural Language Video Localization with Learnable Moment Proposals. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 4008–4017.

[59] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2986–2994.

[60] 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.

[61] 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.

[62] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems* 32 (2019).

[63] Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9159–9166.

[64] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10287–10296.

[65] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

[66] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based Localizing Network for Natural Language Video Localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6543–6554.

[67] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12870–12877.

[68] Yin Zhang, Ziwei Zhu, Yun He, and James Caverlee. 2020. Content-collaborative disentanglement representation learning for enhanced recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 43–52.

[69] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems* 33 (2020), 18123–18134.

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

[71] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. 2018. Visual object networks: Image generation with disentangled 3D representations. In *Advances in Neural Information Processing Systems*. 118–129.