Reverse Distribution Based Video Moment Retrieval for Effective Bias Elimination

Lingdu Kong[1], Xiaochun Yang[1*], Tieying Li[1], Bin Wang[1,2,3], Xiangmin Zhou[4]

[1]Northeastern University, China
[2]National Frontiers Science Center for Industrial Intelligence and Systems Optimization, China
[3]Key Laboratory of Data Analytics and Optimization for Smart Industry (Northeastern University), Ministry of Education, China
[4]RMIT University, Australia
2210707@stu.neu.edu.cn, {yangxc, binwang}@mail.neu.edu.cn, tieying@stumail.neu.edu.cn, xiangmin.zhou@rmit.edu.au

Abstract

Video Moment Retrieval (VMR) aims to identify a temporal segment in an untrimmed video that best matches a given textual query. Bias in VMR is a critical issue, where the model achieves favorable results even if disregarding the video input. Existing evaluation methods, such as Resplitting, have attempted to address bias by creating out-of-distribution (OOD) datasets. However, these methods provide an incomplete definition of bias and do not quantify bias. To this end, we provide a comprehensive definition of bias in VMR, encompassing both data bias and model bias. Besides, our evaluation metrics can analyze the magnitude of these biases better. To address both data and model biases comprehensively, we introduce Reverse Distribution based VMR (ReDis-VMR). This novel approach dynamically generates datasets with inverse distributions tailored to different models based on Gaussian kernel estimation. As a result, it enables a more accurate evaluation of model performance. Building on ReDis-VMR, we further propose the Dynamic Expandable Adjustment (DEA) pipeline. DEA incrementally expands the model structure to enhance its focus on video and text features, and it incorporates a fair loss to minimize the influence of concentrated data distributions. The experiments on bias ratio demonstrate that our ReDis method achieves state-of-the-art performance in bias elimination, while the results on moment retrieval confirm the effectiveness of our DEA framework across three evaluation methods, two datasets, and three baselines.

Code — https://github.com/NoobKLD/ReDis-VMR

Introduction

Video Moment Retrieval (VMR), as a downstream task in video content understanding, has recently garnered significant attention from researchers (Gao et al. 2017; Hendricks et al. 2017; Liu et al. 2018; Lu et al. 2019; Hahn et al. 2020; Wang, Huang, and Wang 2019). The objective of VMR is to identify a temporal segment in an untrimmed video that best matches a textual query. Researchers typically focus on designing models to obtain improved fusion representations of video and text (Zhang et al. 2020b; Cui et al. 2022; Zhang et al. 2020a; Gao et al. 2017; Hendricks et al. 2017; Liu et al. 2018; Lu et al. 2019). It is worth noting that the biases in VMR are often overlooked. Due to the similarity of concentrated distributions of temporal segments in existing training and testing datasets, models can achieve favorable results even when overlooking video inputs (Otani et al. 2020).

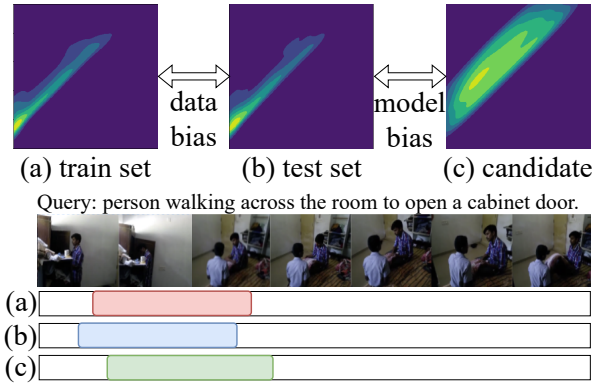


Figure 1: An example of data bias and model bias: (a) Distribution of the train set and its highest confidence segments. (b) Distribution of the test set and the ground truth. (c) Model's candidates and their highest confidence segments.

Existing methods (Yang et al. 2021; Liu, Qu, and Hu 2022; Wang et al. 2023; Yin et al. 2023) primarily verify the generalization ability of models without bias by evaluating on resplitted datasets (Yuan et al. 2021). The resplitting method modifies the training and test data based on the distribution of the original training set and test set, transforming them into datasets with out-of-distribution (OOD) distributions. This resplitting method primarily addresses data bias, which refers to the phenomenon where a model achieves favorable results due to the similarity between the distributions of the training set and the test set. However, in addition to data bias, we identified the presence of model bias, which is equally important and warrants attention. Model bias refers to the similarity between the distribution of the model's candidate set and the distribution of the test set, leading to a predictive tendency in the model. As a result, the model can achieve favorable results even without training. Nevertheless, a VMR model should be capable of predicting any temporal segment without such predictive tendencies.

*Corresponding author.

As shown in Figure 1(a) and (b), the distributions of the training set and the test set in existing datasets (Gao et al. 2017; Krishna et al. 2017) are significantly similar, which constitutes data bias. As illustrated in Figure 1(b) and (c), existing models often design different candidates for different datasets. These multiple candidates form a distribution. Since this distribution often closely resembles that of the test set, the model can achieve favorable results even without training. We define this phenomenon as model bias. Moreover, previous methods have not provided specific evaluation metrics to measure the magnitude of both biases. To quantify data bias and model bias, we have designed an evaluation metric. This metric measures the residual bias by calculating the KL divergence between the distributions of the training set, test set, and model's candidate set. In analyzing the data bias and model bias, we found that the existing resplitting method (Yuan et al. 2021) is ineffective in mitigating data bias and also fails to address the problem of model bias. Therefore, following the resplitting method's setting, there is an urgent need for a method to modify the dataset eliminating both data and model biases to validate the real performance of VMR.

To meet this need, we propose Reverse Distribution based VMR (ReDis-VMR), which dynamically generates data with inverse distributions according to the characteristics of different model designs. ReDis-VMR utilizes prediction distribution on the train set to obtain a reverse distributed dataset of the original concentrated distributed dataset. The combination of the two datasets becomes a uniformly distributed dataset, which can eliminate the concentrated distribution for this specific model design. Based on this effective evaluation setting of ReDis-VMR, we propose the Dynamic Expandable Adjustment (DEA) pipeline. In DEA, when the model converges, the dynamic model structure is incrementally expanded, which enhances the model's perception of video features. Meanwhile, to reduce the impact of concentrated data distribution and avoid predictive tendencies, we design a fair loss that encourages the predicted confidence distribution of our DEA to be a uniform distribution.

Our contributions can be summarized as follows:

- We are the first to provide a comprehensive definition of bias in VMR, encompassing both data bias and model bias. The necessity of eliminating these biases is emphasized, along with the introduction of evaluation metrics for quantifying them.
- We propose ReDis, which effectively eliminates both model bias and data bias, allowing for the validation of model performance without bias. In addition, DEA enhances the focus on video and text features through its dynamic model structure. A fair loss objective is also introduced, designed to counteract the reliance on concentrated data distribution.
- The ReDis method achieves state-of-the-art results in eliminating both data bias and model bias. Additionally, the DEA framework excels in experimental results across three validation methods and three baselines, demonstrating its broad applicability.

Related Works

Video Moment Retrieval. VMR aims to identify temporal segments in videos that best match textual queries. Based on different supervision information, methods can be categorized into two types: fully-supervised methods (Zhang et al. 2021; Yang and Wu 2022; Huang et al. 2022) and weakly-supervised methods (Mithun, Paul, and Roy-Chowdhury 2019; Lin et al. 2019; Wu et al. 2020; Zhang et al. 2020c; Ma et al. 2020). Additionally, based on different model structures, they can be divided into three categories: two-stage methods (Gao et al. 2017; Hendricks et al. 2017), end-to-end methods (Liu et al. 2018; Lu et al. 2019; Zhang et al. 2019), and reinforcement learning-based methods (Hahn et al. 2020; Wang, Huang, and Wang 2019). two-stage methods, end-to-end methods, and reinforcement learning-based methods.

Biases in Video Moment Retrieval. In VMR, there exists a particular phenomenon where the model, during its inference process after learning, often disregards one or several input modalities and directly yields correct results. Otani et al. (Otani et al. 2020) pointed out this issue of biases and analyzed datasets and some SOTA methods (Zhang et al. 2020b; Yuan et al. 2019). They found that on Charades (Gao et al. 2017) and ActivityNet (Krishna et al. 2017) datasets, temporal segments corresponding to sentences with the same vocabulary are concentrated. This leads the model to easily predict these concentrated segments, ignoring the video input, and achieving good performance metrics.

Subsequently, researchers also began to focus on the biases problem in VMR. Existing solutions can mainly be classified into three categories: directly modifying data distribution or model designs (Zhang et al. 2022; Wang et al. 2022; Lan et al. 2023; Yang et al. 2021; Liu, Qu, and Hu 2022; Wang et al. 2023; Yin et al. 2023; Yoon et al. 2022), modifying performance metrics (Togashi et al. 2022; Zhou et al. 2021), and weak supervision (Zheng et al. 2022; Cui et al. 2022; Ji et al. 2023). In methods that modify data distribution or model designs, Yuan et al. (Yuan et al. 2021) created a set of OOD datasets to evaluate the model's generalization ability under OOD conditions. Hao et al. (Hao et al. 2022) disrupted global video representations to address the bias problem without sacrificing accuracy. Lan et al. (Lan et al. 2023) balanced sample distribution by adding negative samples. Yang et al. (Yang et al. 2021) addressed bias problems by eliminating confounding effects of video segments. Liu et al. (Liu, Qu, and Hu 2022) devised three bias representations and moved the primary representation away from the video representation. Wang et al. (Wang et al. 2023) designed a hybrid sentence representation to enhance the model's generalization. Yin et al. (Yin et al. 2023) emphasized the importance of enhancing action text representations to address biases. Modifying performance metrics (Togashi et al. 2022; Zhou et al. 2021) measured the model's predictive ability without biases using new evaluation metrics. Weak supervision approaches (Zheng et al. 2022; Cui et al. 2022; Ji et al. 2023), utilized less supervisory information, which means reducing biases in supervisory information. However, these methods either overlook the impact

of data on bias or ignore the influence of the model on bias. Additionally, they do not provide evaluation metrics to quantitatively assess bias.

Bias Analysis and Quantification

Problem Formulation

Given a video feature V and a query text feature T, the purpose of VMR is to obtain the most relevant video segment to this text. The start and end timestamps of the video segment are denoted by (s, e). Ideally, a VMR model should utilize multi-modal representations (V, T) to predict the positions of temporal segments, denoted by $(V, T) \xrightarrow{M} (s, e)$, where M represents the trained model. The predicted values of s and e can be rational only when the multimodal representations of V and T are jointly considered.

However, in practice, the distribution of (s, e) in the original dataset, denoted as D_0, tends to be concentrated. During training, the model often achieves a low loss by predicting temporal segments within the concentrated distribution. In other words, it disregards multi-modal features (V, T) and directly predicts the temporal segments s and e based on the distribution D_0, denoted by $D_0 \xrightarrow{M} (s, e)$. We can observe that biases are influenced by two causes: the distribution of the dataset D_0 and the design of the model M. **Data bias**: bias caused by D_0 primarily arises from the similarity between the distributions of the training set and the test set. **Model bias**: bias caused by M is mainly due to the similarity between the distribution of the model's candidate set and the test set.

Bias Quantification

To quantify the relationship between D_0 and M in data and model biases, we need to define the computation methods for the distributions of the training set, test set, and model's candidate set.

Let $X = \{x_i \mid x_i = (s_i, e_i),\ i = 1, 2, \ldots, n\}$ denote the set of temporal segments. The training set and test set in D_0, as well as the candidate set in M, are denoted as X_{train}, X_{test} and X_{can}, respectively. We compute the distributions of the training set X_{train} and test set X_{test} using Gaussian kernel density estimation:

$$g(x) = \frac{1}{2n\pi h^2} \sum_{i=1}^{n} \exp\left(-\frac{\|x - x_i\|^2}{2h^2}\right), \tag{1}$$

where n is the number of samples and h is the bandwidth, which is set using Silverman's rule of thumb, denoted as $h = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-1/(d+4)}\sigma$. The distributions of the training set and test set can be represented by $g_{\text{train}}(x)$ and $g_{\text{test}}(x)$, respectively.

We define the distribution of the model's candidate set as the expected score for each candidate when samples are uniformly and randomly distributed. This expectation reflects the model's predictive tendency in the absence of training. The the ith candidate's expected score is given by: $E_i = \sum_{j=1}^{n} \text{IoU}(x_i, x_j)$ where IoU() represents the intersection-over-union metric. The expected score E_i for each candidate x_i is the accumulation of scores for other candidates. Finally, we estimate the distribution of the candidate set using a weighted Gaussian kernel. The formula is as follows:

$$\hat{g}(x) = \frac{1}{\sum_{i=1}^{n} E_i} \sum_{i=1}^{n} E_i \frac{1}{2\pi h^2} \exp\left(-\frac{\|x - x_i\|^2}{2h^2}\right). \tag{2}$$

We can obtain the distribution of the candidate set $\hat{g}_{\text{can}}$. Kullback-Leibler (KL) divergence is used to calculate the additional information required by the training set g_{train} and the candidate set $\hat{g}_{\text{can}}$ relative to the test set g_{test}. The greater this additional information, the lower the similarity between g_{train} and $\hat{g}_{\text{can}}$ relative to g_{test}, indicating less bias. Let X_{all} represent the set of all possible candidate segments of model f. The formula of KL divergence is as follows:

$$\text{KL}_{\text{data}}(g_{\text{test}} \| g_{\text{train}}) = \sum_{x \in X_{\text{all}}} g_{\text{test}}(x) \log \frac{g_{\text{test}}(x)}{g_{\text{train}}(x)}, \tag{3}$$

$$\text{KL}_{\text{model}}(g_{\text{test}} \| \hat{g}_{\text{can}}) = \sum_{x \in X_{\text{all}}} g_{\text{test}}(x) \log \frac{g_{\text{test}}(x)}{\hat{g}_{\text{can}}(x)}. \tag{4}$$

Finally, we normalize this value to a range of 0 to 1 to represent the proportion of bias present. A higher proportion indicates a greater residual bias, while a lower proportion indicates less bias. The formula for the bias ratio of model M on dataset D_0 is as follows:

$$\text{B}_{\text{data}}(D_0, M) = \frac{1}{1 + \text{KL}_{\text{data}}}, \text{B}_{\text{model}}(D_0, M) = \frac{1}{1 + \text{KL}_{\text{model}}}, \tag{5}$$

$$\text{BiasRatio}(D_0, M) = \frac{1}{2}(\text{B}_{\text{data}}(D_0, M) + \text{B}_{\text{model}}(D_0, m)). \tag{6}$$

The symbols B_{data} and B_{model} represent the data bias ratio and model bias ratio, respectively. We take the average of these two values as the final bias ratio.

Analysis of Existing Debias Methods

(Otani et al. 2020) posits that due to the similar temporal segment distributions between the training and test sets in the existing dataset (Gao et al. 2017; Krishna et al. 2017), models can predict the correct temporal segments by exploiting the distribution, rather than the actual video content. They summarize this phenomenon as bias.

The resplitting method (Yuan et al. 2021) re-partitions the dataset based on the existing distribution, making the training and test sets OOD datasets. They assert that validating on this re-partitioned dataset effectively verifies the model's performance without bias. However, both (Otani et al. 2020; Yuan et al. 2021) only address the causes of bias from a data perspective and do not provide metrics to quantify bias. This lack of comprehensive bias quantification and mitigation leads to inefficiencies in existing methods.

To highlight these inefficiencies, we calculated the bias ratios for various models under the traditional dataset distribution, the resplitted dataset distribution, and the ReDis dataset distribution. As shown in Figure 2, we have calculated the bias ratios for various models under the traditional,

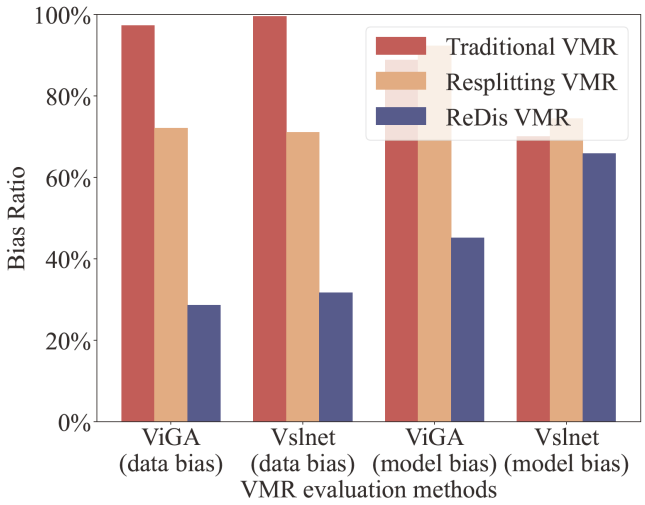


Figure 2: Data and model bias ratios for three VMR evaluation methods under ViGA and Vlsnet in Charades dataset.

resplitted and ReDis dataset distribution. Compared to Traditional VMR, Resplitting VMR is ineffective at eliminating data bias and actually increases the model bias ratio. In contrast, both data bias and model bias show a significant decrease in the ReDis-VMR evaluation settings.

Method

In this section, to effectively validate the true performance of the VMR model without the influence of data and model biases, we introduce the Dynamic Reverse Data Generation module (module$_{\text{RDG}}$) in the ReDis-VMR evaluation setting. Subsequently, we propose Dynamic Expandable Adjustment, enhancing the model's utilization of video and text representations. Finally, we introduce the learning objectives of DEA, including a fair loss designed to mitigate the model's exploitation of data distribution.

Dynamic Reverse Data Generation

To validate the true performance of the model, module$_{\text{RDG}}$ is employed during the model training process to dynamically generate temporal segments data with a reverse distribution, derived from the original concentrated distribution. These segments are then integrated into the test process.

To begin with, we train the model on the original dataset until convergence and then make predictions on the train set. The predictions of temporal segments are denoted as $\hat{X} = \{\hat{x}_i \mid \hat{x}_i = (\hat{s}_i, \hat{e}_i), i = 1, 2, \ldots, n\}$. This sequence contains the concentrated distributed temporal segments predicted by this model design due to biases. To simplify the illustration, all $\hat{s}_i$ and $\hat{e}_i$ are normalized to the range between 0 and 1. Based on this sequence, we can estimate a density function of the model's temporal segment data distribution using a Gaussian kernel:

$$F(s,e) = \frac{1}{Nh^2} \sum_{i=1}^{l} \sum_{j=i+1}^{l} w_{ij} K\left(\frac{s - \frac{i}{l}}{h}, \frac{e - \frac{j}{l}}{h}\right), \quad (7)$$

where $K(s,e)$ is the Gaussian kernel function, defined as $K(s,e) = \frac{1}{2\pi} \exp(-\frac{s^2+e^2}{2})$. l is the length of the video feature. w_{ij} represents the weight of the temporal segment candidate $(\frac{i}{l}, \frac{j}{l})$, and it is designed to prevent the subsequent reverse distribution from generating too many long temporal segments, which are easy to achieve high IoU. We calculate the weight of each temporal segment candidate by using its sum of IoU with all temporal segments of the predicted sequence:

$$w_{ij} = \sum_{i=1}^{n} \text{IoU}((\hat{s}_i, \hat{e}_i), (\frac{i}{l}, \frac{j}{l})). \quad (8)$$

Next, we need to derive its reverse density function for generating new data. We achieve this by subtracting the maximum probability density from the function and then dividing it by the integral within this region, yielding the normalized density function, as shown in the formula:

$$G(s,e) = \max_{0 \le t \le u \le 1} F(t,u) - F(s,e), \quad (9)$$

$$T(s,e) = \frac{G(s,e)}{\int_0^1 \int_0^e G(s,e)\, ds\, de}. \quad (10)$$

Finally, we obtain $T(s,e)$, which is the probability density function. This density function $T(s,e)$ represents the reverse distribution of $F(s,e)$. Based on this probability density function, we can randomly sample temporal segments to generate a new dataset. To generate new video features, we simply adopt the mainstream approach (Zhang et al. 2020b, 2019; Cui et al. 2022; Zhang et al. 2020a) that uses the fixed sampling to correct the video features. Specifically, we divide the original video features into three segments based on the generated start and end times. Then, we apply fixed sampling to each segment to obtain new video features. The original methods also performed fixed sampling on the original video input. It is reasonable to assume that the model can accept scaled video features.

Dynamic Expandable Adjustment

We have observed that many model designs tend to exploit model bias to achieve better evaluation metrics, rather than inferring temporal segments using essential video and text representations. Therefore, in this section, we propose a Dynamic Expandable Adjustment pipeline that places greater emphasis on video and text representations.

To achieve this, we deconstruct existing model structures. We decompose existing models into two parts: a model structure focused on fusion representations and a model structure focused on temporal segment distribution. Thus, we adopt a dynamically changing model structure, as illustrated in Figure 3. We define the model structure that focuses on representation as $\Phi_{r_{old}}$. During the training process, when new data are generated, a new model structure $\Phi_{r_{new}}$ is created for the model focusing on representation. The model structure that emphasizes data distribution is denoted as Φ_d. We can formulate the traditional inference process for the model as:

$$(s,e) = \Phi_d(\Phi_{r_{old}}(V,T)). \quad (11)$$

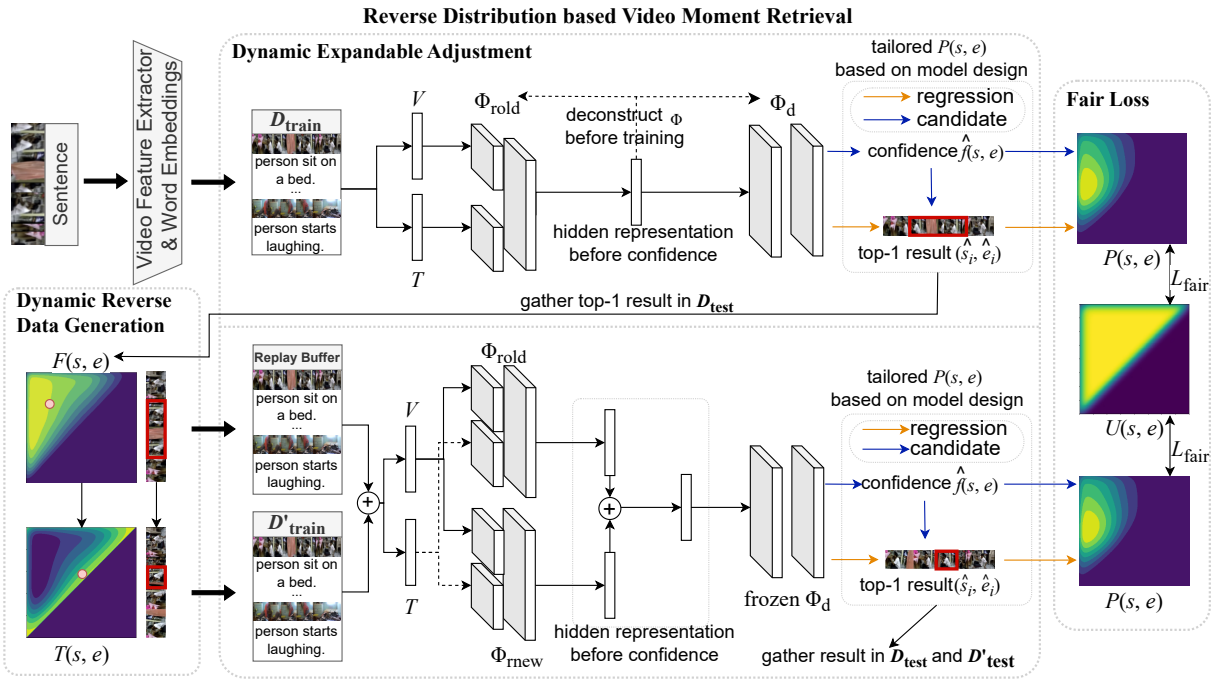


Figure 3: The ReDis-VMR evaluation setting and the Dynamic Expandable Adjustment pipeline. After the training of the model on the original dataset, the module$_{\text{RDG}}$ of ReDis-VMR generates a training dataset and a testing dataset with a reverse distribution. During the training and testing processes on this reverse dataset, our proposed Dynamic Expandable Adjustment involves adding and freezing operations to the existing model structure. Additionally, a fair loss is employed to make the model insensitive to the dataset distribution.

After enhancing the model structure that emphasizes representation, the inference process for the new model becomes:

$$(s, e) = \Phi_d(W[\Phi_{r_{old}}(V, T), \Phi_{r_{new}}(V, T)]), \tag{12}$$

where W represents the weights of the constructed fully connected layer. In the model training process, for the model structure Φ_d to be overlooked, we do not generate a new model structure and freeze Φ_d from updating parameters. Thus, the impact of concentrated distribution on the overall model is reduced.

Fair Loss

To further mitigate the model's dependency on the concentrated distribution of temporal segments in predictions, we employ the fair loss function L_{fair}. The fundamental idea behind this loss function is to constrain the density function of the model's final predictions using a uniform distribution. This loss function enhances the diversity of the model's prediction results. We utilize the KL divergence to impose constraints on the two density functions, as shown in the formula:

$$L_{\text{fair}}(P||U) = \sum_{s,e} P(s,e) \log\left(\frac{P(s,e)}{U(s,e)}\right), \tag{13}$$

where $P(s,e)$ represents the confidence predicted by the model, and $U(s,e)$ denotes the uniform distribution under the model's design. However, as analyzed in Section , the definitions of $P(s,e)$ and $U(s,e)$ should be tailored differently for various model designs. For instance, in regression models where the model directly and separately predicts start and end times, we need to transform the generated temporal segments into density distributions. Therefore, the formulas for $P(s,e)$ and $U(s,e)$ are:

$$\begin{aligned} P(s,e) &= \tfrac{1}{2\pi} \exp\left(-\tfrac{\|s-s_1\|^2+\|e-e_1\|^2}{2}\right), \\ U(s,e) &= \tfrac{1}{\int_s^L \int_0^s dsde}, 0 \le s \le e \le L, \end{aligned} \tag{14}$$

where (s_1, e_1) represents the predicted temporal segment result, and L is the length of the video. In discrete candidate set models (such as 2d-map and anchor models), assuming the candidate set is $(s_1, e_1), (s_2, e_2), \ldots, (s_n, e_n)$, we need to normalize both $P(s,e)$ and $U(s,e)$ as distributions. Thus, their formulas are:

$$\begin{aligned} P(s,e) &= \tfrac{\hat{f}(s,e)}{\sum_{i=1}^{n} \hat{f}(s_i,e_i)}, 1 \le i \le n, \\ U(s,e) &= \tfrac{1}{n}, 1 \le i \le n, \end{aligned} \tag{15}$$

where $\hat{f}(s,e)$ represents the confidence obtained by the model for each candidate. Depending on the specific model definitions of $P(s,e)$ and $U(s,e)$, we combine loss function L_{fair} with the original model's loss function L_{base}. Taking the candidate set approach's binary cross entropy loss as an

Evaluation	*Model*	*Charades* (Gao et al. 2017)					*ActivityNet* (Krishna et al. 2017)				
		Bias Ratio	R1@0.3	R1@0.5	R1@0.7	mIoU	Bias Ratio	R1@0.3	R1@0.5	R1@0.7	mIoU
Traditional VMR	ViGA	93.04	71.21	45.05	20.27	44.57	94.91	59.61	35.79	16.96	40.12
	ViGA+DEA		**71.53**	**45.35**	**20.99**	**44.68**		**59.70**	**37.34**	**17.95**	**40.50**
	Vslnet	84.77	64.30	47.31	30.19	45.15	78.80	55.32	39.24	24.05	41.27
	Vslnet+DEA		**67.04**	**48.76**	**30.22**	**46.88**		**57.61**	**41.02**	**25.71**	**42.04**
	Shuffle	63.58	66.34	50.55	34.26	46.81	88.49	60.56	44.58	27.52	44.28
	Shuffle+DEA		**70.22**	**55.28**	**36.62**	**51.04**		**60.92**	**45.53**	**28.21**	**45.18**
Resplitting VMR	ViGA	82.15	64.92	41.22	21.21	42.42	81.25	41.72	24.85	10.96	29.86
	ViGA+DEA		**66.99**	**41.48**	**22.04**	**43.95**		**52.25**	**26.89**	**11.33**	**35.50**
	Vslnet	72.77	54.61	34.10	17.87	36.34	66.70	38.30	20.03	10.29	28.18
	Vslnet+DEA		**62.19**	**41.48**	**21.24**	**40.80**		**40.04**	**21.44**	**11.21**	**29.13**
	Shuffle	34.75	64.95	46.67	27.08	44.30	70.94	42.08	24.57	13.21	30.45
	Shuffle+DEA		**66.43**	**48.74**	**27.53**	**45.11**		**42.93**	**25.10**	**13.43**	**30.99**
ReDis-VMR	ViGA	**36.83**	33.76	17.38	5.98	21.94	**79.59**	36.93	16.75	7.93	27.87
	ViGA+DEA		**44.56**	**26.28**	**10.95**	**29.61**		**40.97**	**22.41**	**10.55**	**29.66**
	Vslnet	**50.78**	35.20	8.94	3.55	26.76	**62.21**	37.59	20.83	6.77	25.58
	Vslnet+DEA		**48.76**	**30.58**	**17.45**	**35.13**		**41.85**	**25.59**	**14.00**	**31.02**
	Shuffle	**31.52**	42.14	25.85	11.76	27.81	**53.98**	42.30	26.18	12.69	31.84
	Shuffle+DEA		**59.85**	**43.40**	**25.48**	**41.66**		**52.35**	**36.30**	**20.72**	**37.09**

Table 1: Performance comparison with existing methods under Traditional VMR, Resplitting VMR and ReDis-VMR settings.

example, it is represented as follows:

$$\begin{gathered} L = L_{\text{base}} + \lambda L_{\text{fair}}, \\ L_{\text{base}} = \frac{1}{n}\sum_{i=1}^{n} y(s_i, e_i)\log \hat{f}(s_i, e_i) + \\ (1 - y(s_i, e_i))\log(1 - \hat{f}(s_i, e_i)), \end{gathered} \tag{16}$$

where λ is a hyper-parameter that controls the balance between L_{base} and L_{fair}, balancing their contributions to the overall loss , and $y(s_i, e_i)$ represents the IoU value between the ith candidate set and the ground truth.

Experiments

Expertment Settings

Datasets. Charades-STA (Gao et al. 2017): The collection comprises daily indoor activity videos, sourced from the Charades dataset, totaling $6,672$ videos with $16,128$ annotations and $11,767$ moments.

ActivityNet Captions (Krishna et al. 2017): With $20,000$ videos spanning various fields, this dataset contains an average of 3.65 temporally localized sentences per video.

Evaluation Metrics. We utilize the evaluation metrics "Rn@m" proposed in (Gao et al. 2017). "Rn" signifies the top-n retrieved moment results, while "m" denotes that the Intersection over Union (IoU) threshold is set to m. When combined, "Rn@m" indicates the percentage of queries with at least one result having an IoU greater than m among the top-n results. Additionally, "mIoU" represents the average IoU of all top-1 retrieved results. Additionally, we utilize the proposed bias ratio to quantify the residual bias.

Comparison Results

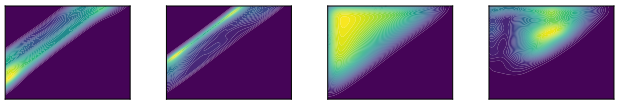
Table 1 compares the bias ratios of Traditional VMR, Resplitting VMR, and ReDis-VMR across three models: ViGA (Cui et al. 2022), Vslnet (Zhang et al. 2020a) and Shuffle (Hao et al. 2022). It also shows the IoU performance of models with and without DEA across three baselines. Compared to the four settings used by (Yang et al. 2021) and (Ji et al. 2023), we employed nine different settings, which sufficiently validate the generalization and effectiveness of our DEA. The lowest bias ratio and the best IoU performance are highlighted in bold. The the best IoU performances under lowest bias ratio is underlined. Based on these results, we have made the following observations: **(1)** Our proposed ReDis-VMR achieves state-of-the-art bias ratios, indicating its strongest capability in bias elimination. Both Traditional VMR and Resplitting VMR exhibit high bias ratios across all three models, demonstrating that the phenomenon of obtaining correct results while disregarding video input is prevalent in these methods. **(2)** Among the three methods, ViGA shows the greatest reduction in bias ratio. This is primarily because the candidate set distribution designed by ViGA is very similar to the test set distribution, resulting in high model bias and allowing ViGA to achieve high metrics even without video input. ReDis-VMR effectively counteracts this issue, providing a more accurate assessment of model performance. Comparatively, the bias ratio for the Charades dataset decreases more significantly than that for the ActivityNet dataset, indicating that the data bias in Charades is greater than in ActivityNet. **(3)** In the comparison between models with and without DEA, configurations with DEA consistently achieve better results across three validation methods, two datasets, and three models. The improvement brought by DEA is more pronounced with lower bias ratios, demonstrating that DEA effectively utilizes video input to predict temporal segments rather than relying on bias to achieve misleading performance.

Ablation Study

In this section, we conducted ablation experiments on components of DEA. Table 2 present the ablation studies on Charades and ActivityNet datasets, respectively. In these tables, "dynamic," "replay," and "fair" represent dynamic model structure, replay mechanism, and fair loss function, respectively. We observed that when using only one component, dynamic model structure achieved significant improvements over the baseline on both datasets. However, when increas-

Variants			*Charades-STA*			
dynamic	replay	fair	R1@0.3	R1@0.5	R1@0.7	mIoU
			33.76	17.38	5.98	21.94
		√	34.23	16.71	5.78	21.94
√			44.26	25.44	**10.99**	28.87
√		√	**45.71**	25.44	10.23	29.40
√	√		44.64	25.55	10.93	28.89
√	√	√	44.56	**26.28**	10.95	**29.61**
Variants			*ActivityNet Captions*			
dynamic	replay	fair	R1@0.3	R1@0.5	R1@0.7	mIoU
			36.93	16.75	7.93	27.87
		√	37.58	20.07	8.28	26.82
√			40.81	21.28	9.37	29.34
√		√	40.93	21.90	9.89	29.59
√	√		40.83	21.18	9.48	29.33
√	√	√	**40.97**	**22.41**	**10.55**	**29.66**

Table 2: The ablation studies of our ViGA+DEA based on Charades-STA and ActivityNet Captions.



(a) Test set (b) Predictions (c) $F(s, e)$ (d) Generated

Figure 4: Effectiveness Evaluation of the module$_{\text{RDG}}$ under the ViGA on the Charades dataset.

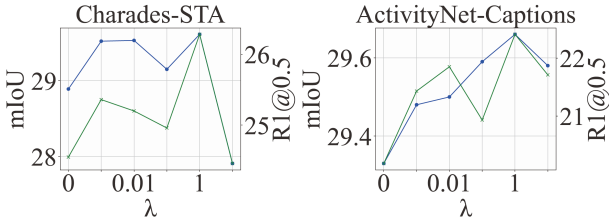


Figure 5: Effectiveness Evaluation of the Fair Loss under the ViGA baseline on the Charades dataset. The blue line represents mIoU, and the green line represents R1@0.5.

ing the number of components to two, improvements were consistently observed across all combinations. This demonstrates the robustness of our pipeline, as minor modifications did not lead to significant accuracy losses.

Effectiveness of module$_{\text{RDG}}$. To validate the effectiveness of the module$_{\text{RDG}}$, we conducted an analysis of the generation process under the ViGA baseline on the Charades dataset. Figure 4(a) illustrates the distribution of the Charades test set. Figure 4(b) shows the distribution of model's predictions. Figure 4(c) represents the $F(s, e)$ function obtained through Gaussian kernel density estimation. Figure 4(d) represents the distribution of generated reverse distributed data. These results demonstrate that the design of our module$_{\text{RDG}}$ effectively mitigates the impact of biases that may mislead the model into obtaining favorable results.

Effectiveness of Fair Loss. Figure 5 illustrates the variation of mIoU with respect to λ. It can be observed that

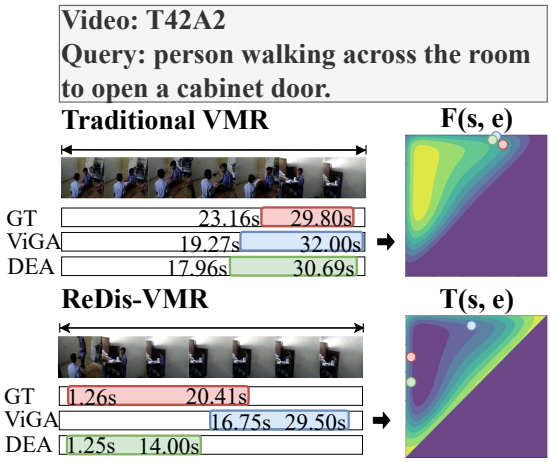


Figure 6: Case study of DEA on the original and generated test split of Charades-STA dataset. Our DEA exhibits good performance even in reverse data.

Evaluation	*Model*	*Bias Ratio*	R1@0.5	*Model*	*Bias Ratio*	R1@0.5
Traditional	CLIP	88.55	75.65	DETR	73.38	54.84
	DEA		**77.72**	DEA		**58.02**
ReDis	CLIP	**57.22**	44.56	DETR	**63.95**	34.73
	DEA		**47.41**	DEA		**37.26**

Table 3: Performance comparison on CLIP-based method with HiREST dataset (Zala et al. 2023) and DETR with QVHighlight dataset (Lei, Berg, and Bansal 2021)

when λ is between 0.001 and 1, the obtained mIoU evaluation metrics are better than when λ is 0. This indicates that the loss function can effectively reduce the impact of concentration distribution, enabling it to achieve good evaluation metrics in the ReDis-VMR setting. Additionally, we found that when λ is 1, the model achieves the best evaluation metrics.

Conclusion

In this paper, we have introduced an effective approach to addressing bias in VMR by providing a comprehensive definition that includes both data bias and model bias. We emphasize the importance of mitigating these biases and propose evaluation metrics for their quantification. Our proposed ReDis-VMR evaluation method effectively eliminates both biases, enabling an unbiased evaluation of model performance. Additionally, we present the DEA, which enhances the model's focus on video and text features while minimizing the impact of concentrated data distributions through a fair loss. Experimental results demonstrate that our ReDis-VMR method sets a new state-of-the-art in bias elimination, and our DEA framework shows robust generalizability across multiple settings. In the future, we will continue to explore more effective methods for eliminating bias in VMR, with the goal of achieving a more accurate assessment of a model's video understanding capabilities.

Acknowledgments

The work is partially supported by the National Key Research and Development Program of China (2024YFF0617702), the National Natural Science Foundation of China (Nos. U22A2025, 62072088, 62232007, U23A20309, 61991404), and Liaoning Provincial Science and Technology Plan Project - Key R&D Department of Science and Technology (No. 2023JH2/101300182), and 111 Project (No. B16009).

References

Cui, R.; Qian, T.; Peng, P.; Daskalaki, E.; Chen, J.; Guo, X.; Sun, H.; and Jiang, Y.-G. 2022. Video Moment Retrieval from Text Queries via Single Frame Annotation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1033–1043. Madrid Spain: ACM. ISBN 978-1-4503-8732-3.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. TALL: Temporal Activity Localization via Language Query. *2017 IEEE International Conference on Computer Vision (ICCV)*, 5277–5285.

Hahn, M.; Kadav, A.; Rehg, J. M.; and Graf, H. P. 2020. 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*. BMVA Press.

Hao, J.; Sun, H.; Ren, P.; Wang, J.; Qi, Q.; and Liao, J. 2022. Can Shuffling Video Benefit Temporal Bias Problem: A Novel Training Framework for Temporal Grounding. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, 130–147. Springer.

Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. C. 2017. Localizing Moments in Video with Natural Language. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 5804–5813. IEEE Computer Society.

Huang, J.; Jin, H.; Gong, S.; and Liu, Y. 2022. Video activity localisation with uncertainties in temporal boundary. In *European Conference on Computer Vision*, 724–740. Springer.

Ji, W.; Liang, R.; Liao, L.; Fei, H.; and Feng, F. 2023. Partial Annotation-based Video Moment Retrieval via Iterative Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4330–4339. Ottawa ON Canada: ACM. ISBN 9798400701085.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Niebles, J. C. 2017. Dense-Captioning Events in Videos. *2017 IEEE International Conference on Computer Vision (ICCV)*, 706–715.

Lan, X.; Yuan, Y.; Chen, H.; Wang, X.; Jie, Z.; Ma, L.; Wang, Z.; and Zhu, W. 2023. Curriculum Multi-Negative Augmentation for Debiased Video Grounding. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 1213–1221. AAAI Press.

Lei, J.; Berg, T. L.; and Bansal, M. 2021. QVHighlights: Detecting Moments and Highlights in Videos via Natural Language Queries. *CoRR*, abs/2107.09609.

Lin, Z.; Zhao, Z.; Zhang, Z.; Wang, Q.; and Liu, H. 2019. Weakly-Supervised Video Moment Retrieval via Semantic Completion Network. In *AAAI Conference on Artificial Intelligence*.

Liu, D.; Qu, X.; and Hu, W. 2022. Reducing the Vision and Language Bias for Temporal Sentence Grounding. *Proceedings of the 30th ACM International Conference on Multimedia*.

Liu, M.; Wang, X.; Nie, L.; He, X.; Chen, B.; and Chua, T.-S. 2018. Attentive Moment Retrieval in Videos. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 15–24. Ann Arbor MI USA: ACM. ISBN 978-1-4503-5657-2.

Lu, C.; Chen, L.; Tan, C.; Li, X.; and Xiao, J. 2019. DEBUG: A Dense Bottom-Up Grounding Approach for Natural Language Video Localization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5144–5153. Hong Kong, China: Association for Computational Linguistics.

Ma, M.; Yoon, S.; Kim, J.; Lee, Y.; Kang, S.; and Yoo, C. D. 2020. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 156–171. Springer.

Mithun, N. C.; Paul, S.; and Roy-Chowdhury, A. K. 2019. Weakly Supervised Video Moment Retrieval From Text Queries. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11584–11593.

Otani, M.; Nakashima, Y.; Rahtu, E.; and Heikkilä, J. 2020. Uncovering Hidden Challenges in Query-Based Video Moment Retrieval. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press.

Togashi, R.; Otani, M.; Nakashima, Y.; Rahtu, E.; Heikkila, J.; and Sakai, T. 2022. AxIoU: An Axiomatically Justified Measure for Video Moment Retrieval. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21044–21053. New Orleans, LA, USA: IEEE. ISBN 978-1-66546-946-3.

Wang, W.; Huang, Y.; and Wang, L. 2019. Language-Driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 334–343. Long Beach, CA, USA: IEEE. ISBN 978-1-72813-293-8.

Wang, X.; Wu, Z.; Chen, H.; Lan, X.; and Zhu, W. 2023. Mixup-Augmented Temporally Debiased Video Grounding with Content-Location Disentanglement. In *Proceedings of the 31st ACM International Conference on Mul-*

timedia, 4450–4459. Ottawa ON Canada: ACM. ISBN 9798400701085.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022. Negative Sample Matters: A Renaissance of Metric Learning for Temporal Grounding. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, 2613–2623. AAAI Press.

Wu, J.; Li, G.; Han, X.; and Lin, L. 2020. Reinforcement Learning for Weakly Supervised Temporal Grounding of Natural Language in Untrimmed Videos. *Proceedings of the 28th ACM International Conference on Multimedia*.

Yang, S.; and Wu, X. 2022. Entity-aware and Motion-aware Transformers for Language-driven Action Localization in Videos. In *International Joint Conference on Artificial Intelligence*.

Yang, X.; Feng, F.; Ji, W.; Wang, M.; and Chua, T.-S. 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. Virtual Event Canada: ACM. ISBN 978-1-4503-8037-9.

Yin, J.; Li, L.; Zhang, J.; Yan, C.; Zhang, L.; and Zhu, Z. 2023. Reducing Intrinsic and Extrinsic Data Biases for Moment Localization with Natural Language. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4584–4594. Ottawa ON Canada: ACM. ISBN 9798400701085.

Yoon, S.; Hong, J. W.; Yoon, E.; Kim, D.; Kim, J.; Yoon, H. S.; and Yoo, C. D. 2022. Selective Query-Guided Debiasing for Video Corpus Moment Retrieval. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, 185–200. Springer.

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. Virtual Event China: ACM. ISBN 978-1-4503-8671-5.

Yuan, Y.; Ma, L.; Wang, J.; Liu, W.; and Zhu, W. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44: 2725–2741.

Zala, A.; Cho, J.; Kottur, S.; Chen, X.; Oguz, B.; Mehdad, Y.; and Bansal, M. 2023. Hierarchical Video-Moment Retrieval and Step-Captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 23056–23065. IEEE.

Zhang, H.; Sun, A.; Jing, W.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021. Parallel Attention Network with Sequence Matching for Video Grounding. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, 776–790. Association for Computational Linguistics.

Zhang, H.; Sun, A.; Jing, W.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2022. Natural Language Video Localization: A Revisit in Span-Based Question Answering Framework. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8): 4252–4266.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2020a. Span-based Localizing Network for Natural Language Video Localization. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 6543–6554. Association for Computational Linguistics.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020b. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 12870–12877. AAAI Press.

Zhang, Z.; Lin, Z.; Zhao, Z.; and Xiao, Z. 2019. Cross-Modal Interaction Networks for Query-Based Moment Retrieval in Videos. In Piwowarski, B.; Chevalier, M.; Gaussier, É.; Maarek, Y.; Nie, J.; and Scholer, F., eds., *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, 655–664. ACM.

Zhang, Z.; Lin, Z.; Zhao, Z.; Zhu, J.; and He, X. 2020c. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4098–4106.

Zheng, M.; Huang, Y.; Chen, Q.; Peng, Y.; and Liu, Y. 2022. Weakly Supervised Temporal Sentence Grounding with Gaussian-based Contrastive Proposal Learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15534–15543. New Orleans, LA, USA: IEEE. ISBN 978-1-66546-946-3.

Zhou, H.; Zhang, C.; Luo, Y.; Chen, Y.; and Hu, C. 2021. Embracing Uncertainty: Decoupling and De-Bias for Robust Temporal Grounding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 8445–8454. Computer Vision Foundation / IEEE.