Selective Query-guided Debiasing for Video Corpus Moment Retrieval

Sunjae Yoon[1], Ji Woo Hong[1], Eunseop Yoon[1], Dahyun Kim, Junyeong Kim[2], Hee Suk Yoon[1], and Chang D. Yoo[1]*

[1] Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
[2] Chung-Ang University, Seoul 06974, Republic of Korea
sunjae.yoon@kaist.ac.rk, junyeongkim@cau.ac.kr, cd_yoo@kaist.ac.kr

Abstract. Video moment retrieval (VMR) aims to localize target moments in untrimmed videos pertinent to a given textual query. Existing retrieval systems tend to rely on retrieval bias as a shortcut and thus, fail to sufficiently learn multi-modal interactions between query and video. This retrieval bias stems from learning frequent co-occurrence patterns between query and moments, which spuriously correlate objects (e.g., a pencil) referred in the query with moments (e.g., scene of writing with a pencil) where the objects frequently appear in the video, such that they converge into biased moment predictions. Although recent debiasing methods have focused on removing this retrieval bias, we argue that these biased predictions sometimes should be preserved because there are many queries where biased predictions are rather helpful. To conjugate this retrieval bias, we propose a Selective Query-guided Debiasing network (SQuiDNet), which incorporates the following two main properties: (1) Biased Moment Retrieval that intentionally uncovers the biased moments inherent in objects of the query and (2) Selective Query-guided Debiasing that performs selective debiasing guided by the meaning of the query. Our experimental results on three moment retrieval benchmarks (i.e., TVR, ActivityNet, DiDeMo) show the effectiveness of SQuiDNet and qualitative analysis shows improved interpretability.

Keywords: video moment retrieval, retrieval bias, selective debiasing

1 Introduction

Video streaming services (e.g., YouTube, Netflix) have rapidly grown these days, which promotes the development of video search technologies. As one of these video search technologies, video moment retrieval (VMR) [8,1] serves as essential building block to underpin many frontier interactive AI systems, including video/image captioning [31,26], video/image question answering [23,14] and visual dialog [3]. VMR aims to localize temporal moments of video pertinent to textual query. Recently, the growing interest in video searching drove this VMR to perform in a more general format of retrieval, referred to as video corpus

[*] Corresponding Author

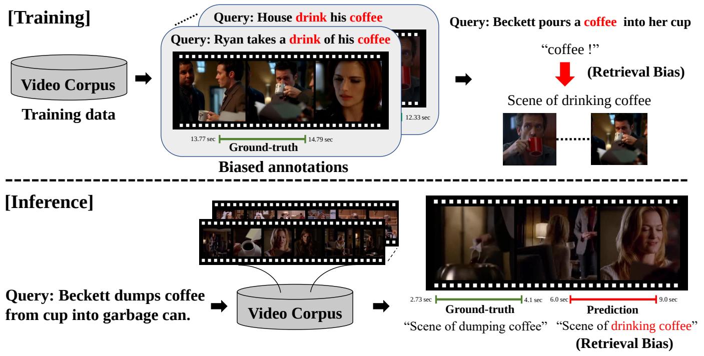
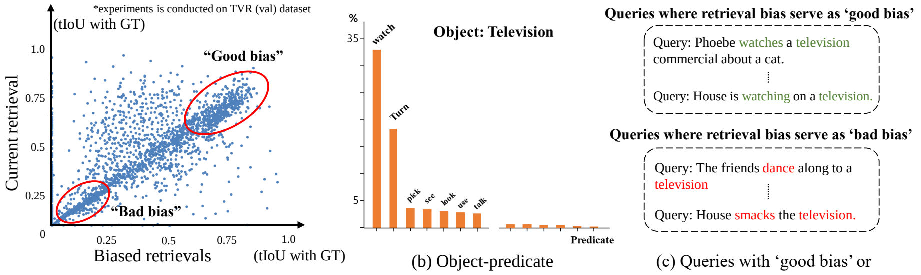


Fig. 1: VCMR training and inference. The biased annotations in training dataset make retrieval bias, which causes biased moment prediction in the inference.

moment retrieval (VCMR) [15]. VCMR also aims to localize a moment like VMR, but the search spaces extend to a ‘video corpus’ composed of the large number of videos. Therefore, given query, VCMR conducts two sub-tasks: (1) identifying relevant video in the video corpus and (2) localizing a moment in the identified video. Despite this respectful effort to generalize video retrieval, the VCMR systems still suffer from dependence on retrieval bias, which hinders the system from accurately learning multi-modal interactions. Figure 1 gives an example of incorrect moment predictions due to the retrieval bias. Given query as “Beckett dumps coffee from cup into garbage can” in inference time, current retrieval systems make incorrect moment prediction with the scene of ‘drinking a coffee’. This is because annotations of training dataset include many co-occurrences between the object word ‘coffee’ in query and the scene of ‘drinking,’ which leads to biased moment prediction referred to as *retrieval bias*. This retrieval bias constrains an object (e.g., coffee) to specific scene (e.g., scene of drinking), thus the other scenes related to that object word lose chance to be searched. Recent debiasing methods [19,30] have focused on removing or mitigating this retrieval bias as they assume the bias degrades retrievals. However, we argue that these biased predictions sometimes should be preserved because there are many queries where biased prediction is rather helpful, such that selective debiasing is required.

Our experimental studies in Figure 2 prove that retrieval bias can also be ‘good’. Figure 2-(a) presents a temporal intersection of union (tIoU) scores for all queries between moment predictions and ground-truth. The predictions are from two retrieval models: (1) the current best performance model and (2) the biased retrieval model. The biased retrieval model is intended to predict biased moments for given queries. To implement this, we simply build a toy model and give it ‘nouns in the query’ as inputs instead of ‘full query sentence’. This induces a deficiency of the original query’s meaning and leads the model to depend on predicting moments where those nouns are mainly used. After that, we represent these two retrieval models’ predictions in a joint-plot of tIoU scores, where it is



(a) tIoU score joint-plot between biased model and current model

(b) Object-predicate co-occurrence distribution for all quries (object: television)

(c) Queries with 'good bias' or 'bad bias' from retrieval bias ('television'-'watching')

Fig. 2: (a) All predictions' tIoU score joint plot between biased model and current model shows correlations between two models, (b) object ('television')-predicate co-occurrence distribution for all queries shows predominant predicate word ('watch'), (c) exemplifies queries where the retrieval bias ('television'-'scene of watching television') serves as 'good bias' or 'bad bias' from statistics in (b).

noted that the plot shows positive correlations. This correlation stands out strong in predictions of low and high tIoU scores, which tells that the current retrieval is both harmed and helped by the retrieval bias. Therefore, retrieval bias includes both *good bias* and *bad bias*. Then, what distinguishes good bias and bad bias? Figure 2(b) give example of our insight on this. We investigate predicates that appear together with a specific word (e.g., television) in all queries and identify that one or two predicates (e.g., watch, turn) are predominantly bound with that word. From these, we have knowledge that query sentences including object word and its most co-occurrent predicate should benefit from retrieval bias (i.e., good bias) because there are also many corresponding scenes (e.g., scene of watching television), but queries with other predicates would be degraded (i.e., bad bias) by this retrieval bias. Figure 2(c) shows query samples where the retrieval bias (i.e., television-scene of watching) serves as 'good bias' or 'bad bias'.

Intrigued by these two characteristics of retrieval bias, we propose a Selective Query-guided Debiasing network (SQuiDNet), which incorporates the following two main properties: (1) Biased Moment Retrieval (BMR) that intentionally uncovers the retrieval bias inherent in objects of the query and (2) Selective Query-guided Debiasing (SQuiD) that performs selective debiasing via disentangling 'good' and 'bad' retrieval bias according to the meaning of the query. In the overall pipeline, we first prepare two moment retrieval models: (1) Naive Moment Retrieval (NMR) and (2) Biased Moment Retrieval (BMR), where both predict the start-end-time of the moment pertinent to their input queries. The NMR is trained under the original purpose of VCMR, so it takes a video and query pair as inputs and sufficiently learns video-language alignment. However, the BMR is trained under the motivation of learning retrieval bias, so it takes a video and 'object words' in the query instead of a full query sentence. These words lose the contextual meaning of the original query, which makes the BMR difficult to properly learn a vision-language alignment, and rather, depend on the shortcut of memorizing spurious correlations that link given words to specific scenes. Based on these two retrievals, SQuiD decides whether the biased prediction of BMR is 'good bias' or 'bad bias' for the prediction of NMR via understanding the query meaning. Here, we introduce two technical contributions on how the SQuiD decides good or bad: (1) Co-occurrence table and (2) Learnable confounder. Our experimental results show state-of-the-art performances and enhanced interpretability.

2 Related Work

2.1 Video Moment Retrieval

The video moment retrieval (VMR) is the task of finding a moment pertinent to a given natural language query. The first attempts [9,11] have been made to localize the moment by giving multi-modal feature interaction between query and video. Previous VMR has focused on constructing modules that can help understand the contextual meaning of the query, including re-captioning [29] and temporal convolution [32]. With the success of natural language models [25,17], recent VMR systems are also interested in utilizing attention-based multi-modal interaction for the vision-language task. Zhang et al. [35] conjugate question answering attention model into VMR as multi-modal span-based QA by treating the video as a text passage and target moment as the answer span. Wang et al. [27] perform multi-levels of cross-modal attention coupled with content-boundary moment interaction for accurate localization of moment. Henceforth, there have been other efforts to perform a general format of video moment retrieval [6,15,16], which finds pertinent moments from a video corpus composed of multiple videos. For this general VMR, Zhang et al. [33] suggested a hierarchical multi-modal encoder, which learns video and moment-level alignment for video corpus moment retrieval. Zhang et al. [34] utilized multi-level contrastive learning to refine the alignment of text with video corpus, which enhances representation learning while keeping the video and text encoding separate for efficiency. To advance forward general format of retrieval, we present another vulnerability as "biased retrieval" in VMR and propose novel framework of debiasing to counter the retrieval bias.

2.2 Causal Reasoning in Vision-Language

Merged with natural language processing [5,22], many high-level vision-language tasks have been introduced, including video/image captioning [31,26], video moment retrieval [8,1], and video/image question answering [14,23]. Causal reasoning has recently contributed to another growth of these high-level tasks via giving ability to reason causal effects between vision and language modalities. Wang *et al.* [28] first introduced observational bias in visual representation and proposed to screen out confounding effect from the bias. Recent question answering systems [21,20] have also utilized this causal reasoning to eliminate language bias in question and answer. For the moment retrieval, there have been efforts to remove the spurious correlation for correct retrieval [19,30]. In this respect, we also uncover the retrieval bias, but furthermore, perform sensible debiasing by conjugating the bias in either positive or negative way.

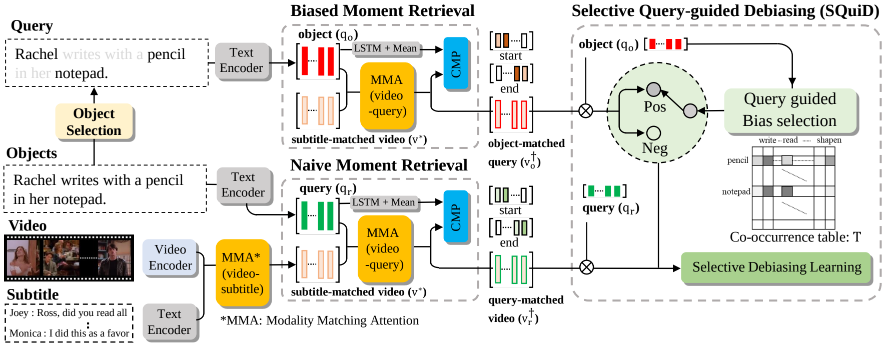

Biased Moment Retrieval
Selective Query-guided Debiasing (SQuiD)
Query
Rachel writes with a pencil in her notepad.
Text Encoder
object (q_o)
LSTM + Mean
CMP
MMA (video -query)
subtitle-matched video (v^*)
start
end
object-matched query ($v_o^\dagger$)
object (q_o)
Pos
Neg
Query guided Bias selection
write read shapen
pencil
notepad
Co-occurrence table: T
Object Selection
Objects
Rachel writes with a pencil in her notepad.
Text Encoder
Naive Moment Retrieval
query (q_r)
LSTM + Mean
CMP
MMA (video -query)
subtitle-matched video (v^*)
start
end
query-matched video ($v_r^\dagger$)
query (q_r)
Selective Debiasing Learning
Video
Video Encoder
MMA* (video-subtitle)
Subtitle
Joey : Ross, did you read all
Monica : I did this as a favor
Text Encoder
*MMA: Modality Matching Attention

Fig. 3: SQuiDNet is composed of 3 modules: (a) BMR which reveals biased retrieval, (b) NMR which performs accurate retrieval, (c) SQuiD which removes bad biases from accurate retrieval of NMR subject to the meaning of query.

3 Method

3.1 Selective Query-guided Debiasing Network

Figure 3 illustrates Selective Query-guided Debiasing Network (SQuiDNet). SQuiDNet prepares two moment retrievals under different motivations, where Naive Moment Retrieval (NMR) aims to perform accurate moment retrieval, while Biased Moment Retrieval (BMR) aims to explicitly reveal the retrieval bias in the training dataset. Following, Selective Query-guided Debiasing (SQuiD) conjugates the biased prediction of BMR to selectively debias NMR. Subject to the contextual meaning of the query, SQuiD decides positive or negative use of retrieval bias for contrastive learning between NMR and BMR. To this, we present two technical contributions to the decision rule in SQuiD: (1) Co-occurrence table and (2) Learnable confounder.

3.2 Input Representations

SQuiDNet takes single pair of video (i.e., video, subtitle) and query sentence as inputs, and training is performed under temporal boundary (i.e., start-end time) annotations. In inference, only video corpus and query are given, SQuiDNet predicts start-end time of moment pertinent to the query from the video corpus.

Video Representation. We use 2D and 3D feature extractors for video encoder. For 2D features, we use ResNet-101 [10] pre-trained on ImageNet [4], and for 3D features, we use SlowFast [7] pre-trained on Kinetics [12]. By concatenating the 2D and 3D features, 4352-dimensional features $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{N_\mathbf{v}}$ are used for video frame embedding, where $N_\mathbf{v}$ is number of frames in a video. With d-dimensional embedder $\delta_\mathbf{v}$, final video features v are embedded on top of layer normalization

LN [2] and positional encoding PE [25] as follows:

$$v = \mathrm{LN}(\delta_{\mathbf{v}}(\mathbf{V}) + \mathrm{PE}(\mathbf{V})) \in \mathbb{R}^{N_{\mathbf{v}} \times d}. \tag{1}$$

Text Representation. For text encoder, we use contextualized token embedding from pre-trained RoBERTa [17]. Here, we are given textual modalities as subtitle $\mathbf{S} = \{\mathbf{sub}(i)\}_i^{N_{\mathbf{s}}}$ and query $\mathbf{q}$, where $N_{\mathbf{s}}$ is the number of subtitles in a video. We first tokenize all the words in subtitles and query into 5072-dimensional word tokens, so that $\mathbf{W}_{\mathbf{sub}(i)} = \{\mathbf{w}^j_{\mathbf{sub}(i)}\}_{j=1}^{L_{s_i}}$ is word tokens in subtitle $\mathbf{sub}(i)$, where L_{s_i} is number of words in that subtitle. $\mathbf{W}_{\mathbf{q}} = \{\mathbf{w}^j_{\mathbf{q}}\}_{j=1}^{L_{\mathbf{q}}}$ is word token in query $\mathbf{q}$, where $L_{\mathbf{q}}$ is number of words in that query. As like the video feature v, final subtitle s_i and query q_r features are embedded by d-dimensional embedder $\delta_{\mathbf{t}}$:

$$s_i = \mathrm{LN}(\delta_{\mathbf{t}}(\mathbf{W}_{\mathbf{sub}_i}) + \mathrm{PE}(\mathbf{W}_{\mathbf{sub}_i}) \in \mathbb{R}^{L_{s_i} \times d}, \tag{2}$$

$$q_r = \mathrm{LN}(\delta_{\mathbf{t}}(\mathbf{W}_{\mathbf{q}}) + \mathrm{PE}(\mathbf{W}_{\mathbf{q}})) \in \mathbb{R}^{L_q \times d}, \tag{3}$$

Modality Matching Attention As shown in Figure 3, to give multi-modal interactions among input modalities (i.e., video-subtitle, video-query), we define Modality Matching Attention (MMA) founded on multi-layer attention in Transformer [24]. MMA takes video and text as inputs and produces text-matched video features. For mathematical definition of MMA, we first define d-dimensional input video feature $x = [x_1, \cdots, x_n] \in \mathbb{R}^{n \times d}$ and text features $y = [y_1, \cdots, y_m] \in \mathbb{R}^{m \times d}$, where n, m is the number of video frames and words in the text. To give interactions between x and y, we construct z by concatenating x and y along the frame and word axis, and perform self-attention on z. Here, we also add fixed token embedding $t_{<x>} \in \mathbb{R}^{n \times d}$ and $t_{<y>} \in \mathbb{R}^{m \times d}$ on x and y, so that the Transformer identifies the heterogeneity between x and y as follows:

$$x = x + t_{<x>}, y = y + t_{<y>}, \tag{4}$$

$$z = [x||y] \in \mathbb{R}^{l \times d}, \tag{5}$$

$$z^\star = \text{Self-Attention}(z) \in \mathbb{R}^{l \times d}, \tag{6}$$

$$x^\star = \mathrm{LN}(z^\star[: n] + x) \in \mathbb{R}^{n \times d}, \tag{7}$$

$$\mathrm{MMA}(x, y) = x^\star, \tag{8}$$

where $[\cdot||\cdot]$ is concatenation and $l = n + m$ is the number of frames and words. $[:]$ denotes slicing operation along the l axis, such that we take video features $z^\star[: n]$ in $z^\star$ as text-matched video features $x^\star$. Therefore, MMA produces $x^\star \in \mathbb{R}^{n \times d}$ comprehending language semantics in y. Henceforth, we introduce MMA into two types of video-text matching: (1) video-subtitle matching and (2) video-query matching for following two retrieval models (i.e., NMR and BMR).

3.3 Biased Moment Retrieval and Naive Moment Retrieval

Naive Moment Retrieval (NMR) is designed for the original purpose of moment retrieval, but Biased Moment Retrieval (BMR) aims at revealing retrieval bias.

Here, shown in Figure 3, the beauty of our proposed BMR is its model-agnostic manner, following the identical structure of NMR. In fact, NMR can be any model that performs moment retrieval (refer to experiments in Table 1), and BMR serves to remove bias inherent in that NMR. The only difference is that BMR takes object word features in query as inputs instead of full query sentence features. As these object words lose the contextual meaning of original query, BMR can only depend on the object words to find the video moment, causing it to prioritize the moment that commonly appears together with that object. To give mathematical definitions of NMR and BMR, we provide general formulations that can have variants of input text (i.e., query or object words). However, implementations of them should be independent, as they are trained for different purposes. Below, the NMR and BMR are performed in the following process: (1) video-subtitle matching, (2) video-query matching, and (3) conditional moment prediction.

Video-subtitle Matching Video frames and their subtitle appearing at the same time share common contextual semantics. Motivated by video-subtitle matching in [16], we also introduce MMA on video frames and their shared subtitles to give multi-modal interactions among them. For the inputs of MMA, we first reorganize video frames feature v as video clips $\mathbf{c} = \{c_i\}_{i=1}^{N_s}$ via collecting frames sharing single subtitle, where c_i collects video frames that share i-th subtitle s_i. N_s is the number of clips corresponding to the number of subtitles.

$$c_i^{\star} = \mathrm{MMA}(c_i, s_i), \tag{9}$$

thus $c_i^{\star}$ represents i-th subtitle-matched video clip. For the following video-query matching, we perform reunion of all clips $v^{\star} = c_1^{\star} \cup \cdots \cup c_{N_s}^{\star}$ to reconstruct original frames and define $v^{\star} \in \mathbb{R}^{N_v \times d}$ as subtitle-matched video feature.

Video-query Matching As shown in Figure 3, the $v^{\star}$ is utilized in MMA of two models (i.e., NMR, BMR) for video-query matching. But, for the input query, BMR utilizes object words instead of query sentence in order to learn retrieval bias. To this, we use nouns from the query for object words as they manly contain objects. Thus, we identify the part of speech (POS) of all words in query and sample noun words using natural language toolkit [18] like below:

$$\mathbf{W_o} = \mathrm{Noun}(\mathbf{W_q}), \tag{10}$$

$$q_o = \mathrm{LN}(\delta_{\mathbf{t}}(\mathbf{W_o}) + \mathrm{PE}(\mathbf{W_o})) \in \mathbb{R}^{L_{q_o} \times d}, \tag{11}$$

where Noun$(\cdot)$ denotes noun-filtering operation using POS tagger. The object words features $q_o \in \mathbb{R}^{L_{q_o} \times d}$ are also embedded from $\mathbf{W}_o$ like equation (3). The L_{q_o} is number of objects in query. Finally, we prepare query feature q_r and object feature q_o for video-query matching in NMR and BMR. Here, we define $q_x \in \{q_r, q_o\}$ for general formulation of two models in following MMA. The video-query matching is performed with q_x and subtitle-matched video $v^{\star}$:

$$v^{\star\star} = \mathrm{MMA}(v^{\star}, q_x), \tag{12}$$

where $v^{\star\star}$ is query-matched video, redefined as $v_x^{\dagger} = v^{\star\star}$ for two cases in q_x. Thus, $v_x^{\dagger} \in \mathbb{R}^{N_v \times d}$ is our final video features for moment prediction with the query q_x.

Conditional Moment Prediction We predict the start-time and end-time of moment for moment prediction, where we introduce conditional moment prediction (CMP) under our motivation that one prediction (e.g., start) can give causal information to the other prediction (e.g., end) rather then predicting these two independently. In details, given query feature q_x and video feature $v_x^{\dagger}$, CMP first, predicts start-time of moment t_{st}. In here, we use query sentence feature $\mathbf{q}_x = \text{MeanPool}(\text{LSTM}(q_x)) \in \mathbb{R}^{d \times 1}$ with lstm and mean-pooling over word axis to compute video-query similarities $v_x^{\dagger}\mathbf{q}_x \in \mathbb{R}^{N_v \times 1}$ in:

$$P(t_{st}|v_x^{\dagger}, q_x) = \text{Softmax}(\text{Conv1D}_{st}(v_x^{\dagger}\mathbf{q}_x)) \in \mathbb{R}^{N_v \times 1}, \tag{13}$$

where Conv1D_{st} is 1D convolution layer to embed start-time information. After that, we predict end-time t_{ed} with this prior start-time information I_{st} below:

$$I_{st} = \sigma(\text{Conv1D}_{st}(v_x^{\dagger}\mathbf{q}_x)) \in \mathbb{R}^{N_v \times 1}, \tag{14}$$

$$P(t_{ed}|v_x^{\dagger}, q_x) = \text{Softmax}(\text{Conv1D}_{ed}(v_x^{\dagger}\mathbf{q}_x + \alpha I_{st})) \in \mathbb{R}^{N_v \times 1}, \tag{15}$$

where $\sigma(\cdot)$ is nonlinear function like ReLU and $\alpha \in \mathbb{R}^1$ is learnable scalar. These two predictions $P(t_{st}|v_x^{\dagger}, q_x)$ and $P(t_{ed}|v_x^{\dagger}, q_x)$ are trained from ground-truth start-end labels (i.e., g_{st}, g_{ed}) using cross-entropy loss $CE(\cdot, \cdot)$ as follows:

$$\mathcal{L}_x = CE(g_{st}, P(t_{st}|v_x^{\dagger}, \mathbf{q}_x)) + CE(g_{ed}, P(t_{ed}|v_x^{\dagger}, \mathbf{q}_x)). \tag{16}$$

Depending on subscript $x \in \{r, o\}$, BMR performs biased training from $\mathcal{L}_o$ and NMR performs retrieval training from $\mathcal{L}_r$. Following SQuiD promotes selective debiasing NMR by conjugating retrieval bias in BMR.

3.4 Selective Query-guided Debiasing

Selective Query-guided Debiasing (SQuiD) is proposed to debias moment retrieval of NMR using biased retrieval from BMR. SQuiD introduces contrastive learning to promote unbiased learning of NMR and biased learning of BMR, by contrasting the prediction of NMR as positive and BMR as negative. However, biased predictions of BMR often should be positive for NMR, depending on the meaning of the query. For example, when given a query as “person drinks a coffee.”, BMR also finds the scene of “drinking coffee” in spite of input object words “person” and “coffee” due to spurious correlation between “coffee ” and “drinking”. SQuiD needs to be sensible in determining whether to use retrieval bias as negative or positive according to the given query. Therefore, our technical contribution is to introduce 2 different decision rules for SQuiD: (1) Co-occurrence table and (2) Learnable confounder.

Co-occurrence Table. Since we cannot directly know all the spurious correlations causing retrieval bias between ‘objects’ and ‘scenes’, we approximate these by referring to statistics of all query sentences. We assume that the ‘predicate’ in query would describe the ‘scene’ in the video, so based on top-K (e.g., K=100) frequent objects and predicates in training queries, we count the co-occurrence of predicates in query sentence for every object words. This counting builds Co-occurrence table $T_d \in \mathbb{R}^{\mathrm{K}\times\mathrm{K}}$ of co-occurrence between object and predicate (Co-occurrence table is illustrated in SQuiD of Figure 3). The row in table T_d holds the co-occurrence frequency of the predicates for a specific object. Figure 2(b) shows one of the co-occurrence distributions when the object "television" is given. To determine the biased prediction of BMR as negative or positive for contrastive learning with NMR, SQuiD utilizes the prior knowledge on predominant "object-predicate" pairs in the Co-occurrence table. For input object words in BMR, SQuiD identifies top-n (e.g., n=10) predominant predicates in the Co-occurrence table. If the top-n predicates appear in the original query sentence, SQuiD determines the prediction of BMR as positive instead of negative. For the selective debiasing learning, we used hinged loss based on video-query similarity $v_x^{\dagger}\mathbf{q}_x \in \mathbb{R}^{N_v}$ between NMR and BMR as follows:

$$\mathcal{L}^{n}_{hinge} = \max[0, \varDelta_{\mathbf{n}} - \max(v_r^{\dagger}\mathbf{q}_r) + \max(v_o^{\dagger}\mathbf{q}_o)], \tag{17}$$

$$\mathcal{L}^{p}_{hinge} = \max[0, \varDelta_{\mathbf{p}} - \max(v_r^{\dagger}\mathbf{q}_r) - \max(v_o^{\dagger}\mathbf{q}_o)]], \tag{18}$$

where $\mathcal{L}^{n}_{hinge}$ denotes the retrieval of BMR as negative and $\mathcal{L}^{p}_{hinge}$ denotes that as positive. SQuiD's decision is to select one of them. $\varDelta_{\mathbf{n}} = 0.2$ and $\varDelta_{\mathbf{p}} = 0.4$ is used and here, we give more margin for $\varDelta_{\mathbf{p}}$ to promote learning of positive.

Learnable Confounder. The Co-occurrence table can be a discrete approximation of retrieval bias as it assumes predefined predicates for selective debiasing. For better approximation, we introduce learnable confounder $\mathbf{Z} \in \mathbb{R}^{\mathbf{K}\times d}$ that can learn object-scene spurious correlation, where it consists of $\mathbf{K}$ (e.g., K = 100) confounders with d-dimensional learnable parameters. Assuming that predicate words sufficiently contain the contextual meaning of video scenes, we predict spuriously correlated *predicate* feature $\mathbf{Y}_B$ from object words q_o and the confounder $\mathbf{Z}$. If the $\mathbf{Y}_B$ is similar to the predicate feature $\mathbf{Y}_C$ of the original query used in NMR, it means that predicate $\mathbf{Y}_B$ obtained from objects word and $\mathbf{Y}_C$ in given query have similar contextual meaning, thus, in this case, the retrieval of BMR should be used as positive.

For above, we needs to pretrain $\mathbf{Z}$ to learn spurious correlations between objects and predicates, so that generated *predicate* $\mathbf{Y}_B$ is biased predicate of the object. To train $\mathbf{Z}$, we regress $\mathbf{Y}_B$ as $\mathbf{Y}_C$, which means $\mathbf{Z}$ is trained to generate predicate features $\mathbf{Y}_B$ that commonly appears together with given object words in a query. To this, we first prepare mean-pooled objects feature $\mathbf{q}_o = \mathrm{MeanPool}(\mathrm{LSTM}(q_o)) \in \mathbb{R}^{1\times d}$ over word axis. The $\mathbf{q}_o$ and confounder $\mathbf{Z}$ performs dot-product attention to make $\mathbf{Y}_B$, which regresses predicates feature $\mathbf{Y}_C$ in original query. To get $\mathbf{Y}_C$, we sample predicate words $\mathbf{W_p} = \mathrm{Pred}(\mathbf{W_q})$ in original query and embed predicate words feature q_p, which is the same process in equation (10,11) and Pred(·) denotes predicate-filtering.

$$\mathbf{Y}_B = \mathrm{Softmax}((\mathbf{q_o} W_{\mathbf{o}})(\mathbf{Z} W_{\mathbf{z}})^T)\mathbf{Z} \in \mathbb{R}^d, \tag{19}$$

$$\mathbf{Y}^{\star}_C = \mathrm{MeanPool}(\mathrm{LSTM}(q_p)) \in \mathbb{R}^d, \tag{20}$$

$$\mathcal{L}_{\mathbf{z}} = ||\mathbf{Y}^{\star}_C - \mathbf{Y}_B||^2_2 \tag{21}$$

Table 1: Performances for video corpus moment retrieval on TVR (test-public), ActivityNet and DiDeMo. ⋆: reconstruction-based results, N: NMR, B: BMR

Method	TVR			ActivityNet			DiDeMo (+ASR)		
	tIoU=0.7			tIoU=0.7			tIoU=0.7		
	R@1	R@10	R@100	R@1	R@10	R@100	R@1	R@10	R@100
XML [15]	3.32	13.41	30.52	-	-	-	1.74*	8.31*	27.63*
HERO [16]	6.21	19.34	36.66	1.19*	6.33*	16.41*	1.59*	9.12*	29.23*
HAMMER [33]	5.13	11.38	16.71	1.74	8.75	19.08	-	-	-
ReLoCLNet [34]	4.15	14.06	32.42	1.82	6.91	18.33	-	-	-
SQuiDNet (N)	4.09	12.30	28.31	1.62	7.82	18.53	1.73	9.84	30.14
SQuiDNet (N[16], B)	8.34	28.03	35.45	3.02	10.23	22.14	2.62	10.28	31.11
SQuiDNet (N, B)	**10.09**	**31.22**	**46.05**	**4.43**	**12.81**	**26.54**	**3.52**	**12.93**	**34.03**

Table 2: Performances for single video moment retrieval (SVMR) on TVR (val) and ActivityNet and video retrieval (VR) on TVR (val).

Method	TVR				Method	ActivityNet	
	SVMR		VR			SVMR	
	R@1,tIoU=μ		-			R@1,tIoU=μ	
	μ=0.5	μ=0.7	R@1	R@10		μ=0.5	μ=0.7
XML [15]	31.11	13.89	16.54	50.41	VSLNet [35]	43.22	26.16
HERO [16]	-	4.02	30.11	62.69	IVG [19]	43.84	27.1
ReLoCLNet [34]	31.88	15.04	22.13	57.25	SMIN [27]	48.46	30.34
SQuiDNet (N,B)	**41.31**	**24.74**	**31.61**	**65.32**	**SQuiDNet**	**49.53**	**31.25**

where $W_{\mathbf{o}}, W_{\mathbf{z}} \in \mathbb{R}^{d\times d}$ are embedding matrices and $\mathbf{Y}^\star_C$ is fixed mean-pooled predicate features, which is target for L2 loss regression $\mathcal{L}_{\mathbf{z}}$. After pretraining confounders $\mathbf{Z}$, SQuiD computes cosine similarity $r = \text{cosine}(\mathbf{Y}_B, \mathbf{Y}_C)$. If r is lager than 0, the retrieval from BMR is used as postive, otherwire as negative:

$$\mathcal{L}^D = \begin{cases} \mathcal{L}^n_{hinge} & \text{if } r \leq 0 \\ \mathcal{L}^p_{hinge} & \text{if } r > 0. \end{cases} \tag{22}$$

4 Experimental Results

4.1 Dataset

We validate SQuiDNet on three moment retrieval benchmarks as follows:

TV show Retrieval. TV show Retrieval (TVR) [15] is composed of 6 TV shows across 3 genres: sitcoms, medical and crime dramas, which includes 109K queries from 21.8K multi-character interactive videos with subtitles. Each video is about 60-90 seconds in length. The TVR is split into 80% train, 10% val, 5% test-private, 5% test-public. The test-public is prepared for official challenge.

ActivityNet. ActivityNet Captions [13] includes 20k videos with 100k query descriptions. 10k videos are given for training and 5k for validation (val_1), where the average length of all videos is 117 seconds, and the average length of queries is 14.8 words. We train our SQuiDNet and evaluate on the val_1 split.

DiDeMo. The Distinct Describable Moments (DiDeMo) [1] contains 10k videos under diverse scenarios. To mitigate complexity, most videos are about 30-seconds and uniformly divided into 5-seconds segments, thus a single video contains 21 possible segments. DiDeMo is split into 80% train, 10% val, and 10% test.

4.2 Experimental Details

Evaluation Metric. We perform three retrieval tasks: (1) video retrieval (VR), (2) single video moment retrieval (SVMR), and (3) video corpus moment retrieval (VCMR). VR is video-level retrieval, evaluating the number of correct predictions of video, where VR measures video-query similarities of all videos to select the highest one. SVMR is moment-level retrieval in given video, evaluating the degree of overlap between predicted moments and ground-truth. VCMR is moment-level retrieval in video corpus, thus we evaluate the incidences where: (1) the predicted video matches the ground-truth video; and (2) the predicted moment has high overlap with the ground-truth moment. SQuiDNet predicts top-n (n=10) videos first, and performs moment retrieval on them. Average recall at K (R@K) over query is used as the evaluation metric, where temporal Intersection over Union (tIoU) measures the overlap between predicted moment and the ground-truth.

4.3 Results on Benchmarks

Table 1 summarizes the best performances reported in XML [15], HERO [16], HAMMER [33], ReLoCLNet [34] on TVR, ActivityNet and DiDeMo[3]. SQuiDNet outperforms previous state-of-the-art performance. We also validate naive model without BMR, which shows large performance gap between full model of SQuiDNet, explaining the effectiveness of selective debiasing learning. As SQuiDNet is conducted on model-agnostic manner, we replace NMR with the HERO baseline from their public code, which also shows improvement from original HERO. SQuiDNet assumes subtitle as inputs, so we can utilize audio speech recognition (ASR) for DiDeMo, which is available in [16]. We also validate results without subtitle on DiDeMo by applying video feature v instead of subtitle-matched video feature $v^\star$ in equation (9). This gives slight performance drop -0.36%/-0.74%/-1.82% from full model of SQuiDNet, which explains that grouping video frames based on subtitles benefits understanding of contextual scenes in video. Table 2 summarizes the results of two sub-tasks of VCMR: (1) SVMR and (2) VR on TVR and AtivityNet. SQuiDNet also shows large performance gain on the SVMR and VR, which explains that selective debiasing is effective at both moment-level and video-level. Although SQuiDNet is assumed to use subtitle, it also shows gain without subtitles in SVMR on ActivityNet.

[3] please refer to Related Work and the papers for their detailed descriptions

Table 3: Ablation on SQuiDNet variants for VCMR on TVR (val)

Model variants	tIoU=0.7 R@1
Full SQuiDNet	8.52
w/o BMR	4.62
w/o CMP	8.17
w/ (All negative)	6.41
w/ (All positive)	6.72
w/ (Co-occurrence table)	7.91
w/ (Learnable confounder)	8.52

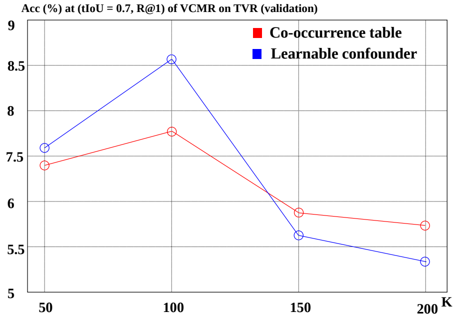
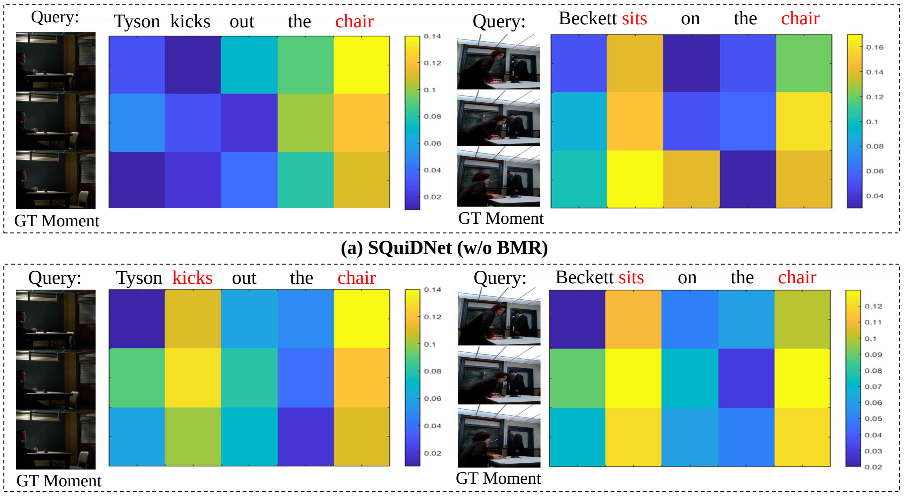


Fig. 4: Accuracy according to top-k objects for Co-occurrence table and variational k for Learnable confounder

4.4 Ablation Study

Table 3 presents ablation studies of our proposed components on SQuiDNet. The first section reports full model SQuiDNet for VCMR on TVR validation set. The second section shows ablative performance without Biased Moment Retrieval (BMR) and conditional moment prediction (CMP). Large performance drop is shown without BMR, which gives two interpretations: (1) retrieval datasets contain many spurious correlations and (2) selective debiasing of video moment retrieval is non-trivial. The performance gain of CMP is not as effective as that of BMR, however, it actually contributes on learning efficiency by promoting early convergence of training loss. We consider this reason to be that CMP narrows the search space with prior knowledge of start-time. Third section presents performance comparison from variants of SQuiD. To be confident of variants of SQuiD's decision rule, we first conduct ablations of giving all retrievals from BMR as negative or positive for the contrastive learning with CMR. The result shows that positive use of biased retrieval is more effective than negative, which explains why SQuiD needs discernment on selecting biased retrieval. Our proposed decision rules (i.e., Co-occurrence table, Learnable confounder) give effectiveness via selective debiasing. The Learnable confounder is more effective, but it needs additional work to train the confouder **Z**.

Figure 4 presents performances of VCMR according to the hyper-parameter **K** in Co-occurrence table and Learnable confounders. For the Co-occurrence table, **K** is the number of top-k objects, and for Learnable confounder, **K** is the number of confounders. Co-occurrence table utilizes statistics in training queries for approximating confounders while Learnable confounder learns confounder from object-predicate spurious correlations under our designed proxy learning with loss in equation (21), where both have similar curve. However, Learnable confounder has higher best-performance. We speculate Learnable confounder has superior control over confounders that cannot be defined in deterministic way.



(a) SQuiDNet (w/o BMR)

(b) SQuiDNet (w/ BMR)

Fig. 5: Visualization of word-level query-video similarities in GT moment. Upper box is results from SQuiDNet trained without BMR and lower box is results from SQuiDNet with BMR. It can be observed that the BMR enables the network to learn the uncommon predicate "kicks" of the object "chair" while also strengthens the learning of the spuriously correlated predicate "sits."

4.5 Qualitative Results

Figure 5 shows the word-level query-video similarities when GT moment is given to SQuiDNet when two queries are given as "Tyson kicks out the chair" and "Beckett sits on the chair." Figure 5(a) represents the similarity distributions from SQuiDNet trained without BMR, where they show high similarity in word "sit" and low in "kick." However in Figure 5(b), the results with BMR show high similarities in both words "sit" and "kick." This explains that when one object word "chair" is given, the system without debiasing can understand the spuriously correlated predicate word "sit" but failed to learn the uncommon predicate word like "kick." In this respect, debiasing allows learning of object words' various connections with other predicate words. Furthermore, it can be observed that the system with debiasing also strengthens the understanding of the spuriously correlated predicate word by having higher similarities in more accurate moments.

Figure 6 presents moment predictions of the two models: NMR and BMR, where red box is prediction from NMR and blue box from BMR, while green box is ground-truth moment. In the right of the Figure, SQuiD's decision is shown on whether to use retrieval bias as positive or negative. When the query is "Robin rides a bicycle through a subway train car," both BMR and NMR predict the scene of person riding a bike in the video, where the SQuiD decides the prediction of BMR as positive retrieval bias. But, for the query "Barney sits on a chair and

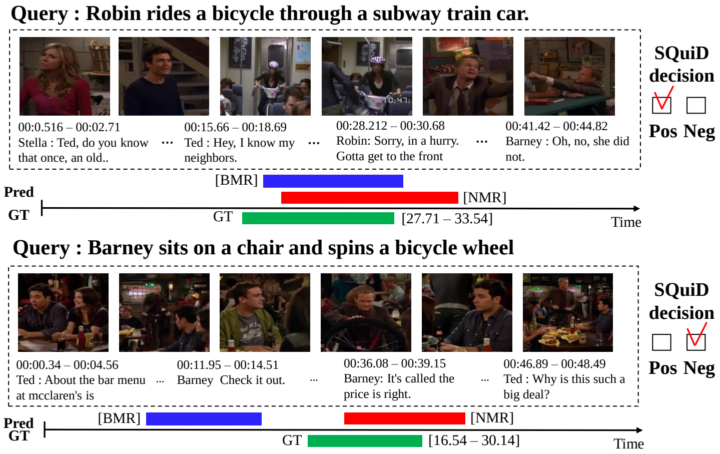

Query : Robin rides a bicycle through a subway train car.
00:0.516 – 00:02.71
Stella : Ted, do you know that once, an old..
…
00:15.66 – 00:18.69
Ted : Hey, I know my neighbors.
…
00:28.212 – 00:30.68
Robin: Sorry, in a hurry. Gotta get to the front
…
00:41.42 – 00:44.82
Barney : Oh, no, she did not.
SQuiD decision
☑ □
Pos Neg
[BMR]
[NMR]
Pred
GT
GT
[27.71 – 33.54]
Time
Query : Barney sits on a chair and spins a bicycle wheel
00:00.34 – 00:04.56
Ted : About the bar menu at mcclaren's is
...
00:11.95 – 00:14.51
Barney Check it out.
...
00:36.08 – 00:39.15
Barney: It's called the price is right.
...
00:46.89 – 00:48.49
Ted : Why is this such a big deal?
SQuiD decision
□ ☑
Pos Neg
[BMR]
[NMR]
Pred
GT
GT
[16.54 – 30.14]
Time

Fig. 6: Visualization of moment prediction on NMR and BMR, where the SQuiD decision is represented for using retrieval bias from BMR as positive or negative.

spins a bicycle wheel," BMR still predicts the scene of person riding a bike, which shows the retrieval bias between "bicycle" and "riding". Here, to counter this retrieval bias, SQuiD decides the prediction of BMR as negative bias, such that NMR is trained to recede the bias prediction of BMR by contrastive learning.

5 Conclusion

This paper considers Selective Query-guided Debiasing Network for video moment retrieval. Although recent debiasing methods have focused on only removing retrieval bias, it sometimes should be preserved because there are many queries where biased predictions are rather helpful. To conjugate this retrieval bias, SQuiDNet incorporates the following two main properties: (1) Biased Moment Retrieval that intentionally uncovers the biased moments inherent in objects of the query and (2) Selective Query-guided Debiasing that performs selective debiasing guided by the meaning of the query. Our experimental results on three moment retrieval benchmarks (TVR, ActivityNet, DiDeMo) show effectiveness of SQuiDNet, while qualitative analysis shows improved interpretability.

Acknowledgemnet

This work was partly supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No. 2021-0-01381, Development of Causal AI through Video Understanding) and partly supported by Institute of Information communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics)

References

1. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE international conference on computer vision. pp. 5803–5812 (2017)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 326–335 (2017)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Escorcia, V., Soldan, M., Sivic, J., Ghanem, B., Russell, B.: Temporal localization of moments in video collections with natural language. arXiv preprint arXiv:1907.12763 (2019)
7. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6202–6211 (2019)
8. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
9. Gao, J., Sun, C., Yang, Z., Nevatia, R.: TALL: temporal activity localization via language query. In: IEEE International Conference on Computer Vision, ICCV 2017. pp. 5277–5285. IEEE Computer Society (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.C.: Localizing moments in video with natural language. In: IEEE International Conference on Computer Vision, ICCV 2017. pp. 5804–5813. IEEE Computer Society (2017)
12. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
13. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017)
14. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. arXiv preprint arXiv:1809.01696 (2018)
15. Lei, J., Yu, L., Berg, T.L., Bansal, M.: Tvr: A large-scale dataset for video-subtitle moment retrieval. arXiv preprint arXiv:2001.09099 (2020)
16. Li, L., Chen, Y.C., Cheng, Y., Gan, Z., Yu, L., Liu, J.: Hero: Hierarchical encoder for video+ language omni-representation pre-training. arXiv preprint arXiv:2005.00200 (2020)
17. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

18. Loper, E., Bird, S.: Nltk: The natural language toolkit. arXiv preprint cs/0205028 (2002)
19. Nan, G., Qiao, R., Xiao, Y., Liu, J., Leng, S., Zhang, H., Lu, W.: Interventional video grounding with dual contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2765–2775 (2021)
20. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12700–12710 (2021)
21. Qi, J., Niu, Y., Huang, J., Zhang, H.: Two causal principles for improving visual dialog. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10860–10869 (2020)
22. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
23. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4631–4640 (2016)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NIPS) (2017)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
26. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence - video to text. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
27. Wang, H., Zha, Z.J., Li, L., Liu, D., Luo, J.: Structured multi-level interaction network for video moment localization via language query. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7026–7035 (2021)
28. Wang, T., Huang, J., Zhang, H., Sun, Q.: Visual commonsense r-cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10760–10770 (2020)
29. Xu, H., He, K., Sigal, L., Sclaroff, S., Saenko, K.: Text-to-clip video retrieval with early fusion and re-captioning. ArXiv **abs/1804.05113** (2018)
30. Yang, X., Feng, F., Ji, W., Wang, M., Chua, T.S.: Deconfounded video moment retrieval with causal intervention. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1–10 (2021)
31. Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W.: Video paragraph captioning using hierarchical recurrent neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4584–4593 (2016)
32. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. arXiv preprint arXiv:1910.14303 (2019)
33. Zhang, B., Hu, H., Lee, J., Zhao, M., Chammas, S., Jain, V., Ie, E., Sha, F.: A hierarchical multi-modal encoder for moment localization in video corpus. arXiv preprint arXiv:2011.09046 (2020)
34. Zhang, H., Sun, A., Jing, W., Nan, G., Zhen, L., Zhou, J.T., Goh, R.S.M.: Video corpus moment retrieval with contrastive learning. arXiv preprint arXiv:2105.06247 (2021)
35. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. arXiv preprint arXiv:2004.13931 (2020)