Interventional Video Grounding with Dual Contrastive Learning

Guoshun Nan[1,*]   Rui Qiao[1,*]   Yao Xiao[2,†]   Jun Liu [3,‡]   Sicong Leng[1]   Hao Zhang[4,5]   Wei Lu[1,‡]

[1] StatNLP Research Group, Singapore University of Technology and Design, [2] Shanghai Jiao Tong University, China

[3] Information Systems Technology and Design, Singapore University of Technology and Design, Singapore

[4] School of Computer Science and Engineering, Nanyang Technological University, Singapore

[5] Institute of High Performance Computing, A*STAR, Singapore

{guoshun_nan,rui_qiao}@sutd.edu.sg, 119033910058@sjtu.edu.cn, jun_liu@sutd.edu.sg

sicong_leng@mymail.sutd.edu.sg, zhang_hao@ihpc.a-star.edu.sg, luwei@sutd.edu.sg

Abstract

Video grounding aims to localize a moment from an untrimmed video for a given textual query. Existing approaches focus more on the alignment of visual and language stimuli with various likelihood-based matching or regression strategies, i.e., $P(Y|X)$. Consequently, these models may suffer from spurious correlations between the language and video features due to the selection bias of the dataset. 1) To uncover the causality behind the model and data, we first propose a novel paradigm from the perspective of the causal inference, i.e., interventional video grounding (IVG) that leverages backdoor adjustment to deconfound the selection bias based on structured causal model (SCM) and do-calculus $P(Y|do(X))$. Then, we present a simple yet effective method to approximate the unobserved confounder as it cannot be directly sampled from the dataset. 2) Meanwhile, we introduce a dual contrastive learning approach (DCL) to better align the text and video by maximizing the mutual information (MI) between query and video clips, and the MI between start/end frames of a target moment and the others within a video to learn more informative visual representations. Experiments on three standard benchmarks show the effectiveness of our approaches.

1. Introduction

Video grounding [3, 24], which aims to automatically locate the temporal boundaries of the target video span for a given textual description, is a challenging multimedia information retrieval task due to the flexibility and complexity of text descriptions and video content. It has been widely used

---

*Equally contributed

†Work done during internship at StatNLP group, SUTD
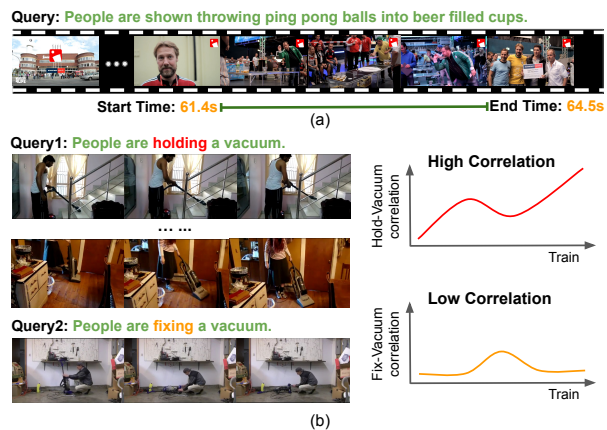
‡Corresponding Authors

Figure 1: (a) Illustration of video grounding. (b) Spurious correlations between object "people" and "vacuum" and the activity "people are holding a vacuum" in the Charades-TA [24] dataset.

in many applications, such as video question answering [40] and video summarization [74]. As shown in Figure 1 (a), the query "People are shown throwing ping pong balls into beer-filled cups" involves two actions "shown" and "throwing", one role "people", and three objects "ping pong balls", "beer" and "cups", which will be located in the video with a start time (61.4s) and an end time (64.5s). To retrieve the most relevant segment, a model needs to well understand the complex interactions among these actions and objects from both language and video context, and then properly align the two sides semantic information for a prediction.

Extensive works have been proposed for the aforementioned challenging video grounding task, which can be mainly categorized into three groups: 1) ranking methods that typically count on a two-stage propose-and-rank pipeline [24, 3, 87], or attention-based localization approach [82] to find out the target video span among candidates with the best matching score. 2) regression meth-

ods [46, 86, 50] that directly predict the start and end time of the target moment to avoid the heavy computations for a large number of candidates. 3) reinforcement learning [70, 30, 11] approaches that dynamically filter out a sequence of video frames conditioned on the given textual query and finally outputs temporal boundaries. The above studies keep pushing the boundary of state-of-the-art performance for the moment localization, and have made it possible to benefit the downstream applications [40, 23, 74].

Despite the enormous success of the current neural models, we argue that these approaches may suffer from the spurious correlations between textual and visual features due to the selection bias of the dataset. As shown in Figure 1 (b), a dataset involves many relevant training instances for some queries that contain the action description word "hold" and two objects "people" and "vacuum'. Meanwhile, there are a very limited number of queries with the action "fix" and two objects "people" and "vacuum'. We can draw a conclusion that the tokens "people" and "vacuum' are highly correlated to the video moments relevant to "people are holding a vacuum", and have low correlations to the video moments "people are fixing a vacuum". Hence, a query that contains the description "people are fixing a vacuum" may be incorrectly located to the video segment "people are holding a vacuum" with high probability. We observe that such biased distributions commonly exist in many datasets, such as Charades-STA [24] and TACoS [57].

Many methods [47, 18, 69, 90] attempt to address the above issue. However, these re-sampling [18] and re-weighting [47] methods are mainly designed for the classification tasks, and it can be difficult for them to be applied to temporal localization. One may also consider that pre-trained knowledge is immune to the issue as they are learned from large-scale datasets. However, as highlighted by Zhang *et al.* [84], the pre-trained models, such as BERT [20] and ResNET [32], may still suffer from the biased issue. From a very different direction, causal inference [54] based on structured causal model (SCM) [54] and potential outcomes [58] have recently shown great promise, achieving state-of-the-art performance on many applications such as scene graph generation [65], data clustering [71], image classification [84], and visual question answering [2, 56]. Despite these success, directly applying these causal methods to the video grounding task may not yield good results, due to the more complex interactions for moment retrieval compared with the image-based bi-modal alignment.

To address the above problem, this paper presents a novel paradigm named interventional video grounding (IVG) based on Pear's SCM [54] from the perspective of causal inference. Theoretically, SCM uses the graphical formalism to treat nodes as random variables and directed edges as the direct causal dependencies between variables. We borrow the idea of the backdoor adjustment and *do*-calculus theory

$P(Y|do(X))$ [54] to deconfound the aforementioned spurious correlations. The main challenge here is to get the unobserved confounders Z that influences both bi-modal representations and the predictions, leading to the unexpected correlations between language and video features by only learning from the likelihood $P(Y|X)$. Previous studies on image-based tasks treat the latent confounder as the prior knowledge or dataset distribution, and approximate them with static probabilities [56, 68] of the image objects from the dataset, or the probabilities predicted by pre-trained classifiers [84]. We propose a simple yet effective method to approximate the prior $P(Z)$ and then obtain $P(Y|do(X))$. Meanwhile, we introduce a dual contrastive learning method (DCL) to learn more informative visual representations by maximizing the MI between query and video clips to better align the bi-modal features, and the MI between the start/end time of the target moment and other clips in the video. With the above two key components IVG and DCL, our proposed IVG-DCL can alleviate the confounder bias and learn high-quality representations for the challenging video grounding task. Experiments on three public datasets show the effectiveness of our proposed IVG-DCL. Specifically, our main contributions are:

- We propose IVG, a novel model for video grounding by introducing causal interventions $P(Y|do(X))$ to mitigate the spurious correlations from the dataset. We also present a novel approximation method for the latent confounder based on SCM.

- We propose a dual contrastive learning method DCL based on MI maximization to learn more informative feature representations in an unsupervised manner.

- We conduct quantitative and qualitative analyses on three benchmark datasets and show interesting findings based on our observations.

2. Related Work

2.1. Video Grounding:

The task of video grounding or moment localization [3, 24] aims to retrieve video segments for the given textual queries. Previous works can be mainly categorized into three groups. 1) **The ranking methods** [24, 3, 33, 73, 43, 44, 76, 25, 87] rely on bi-modal matching mechanisms to obtain the target moment that has the best score. Typical works [24, 3, 76, 87] in this direction resort to a two-stage propose-and-rank pipeline. These approaches highly rely on the quality of proposals and may suffer from heavy computation cost. 2) **Regression approaches** [46, 86, 50] that regress the visual boundaries without matching scores. There are also some works [14, 26, 89] that frame the video grounding as a question answering problem [4] and treat the

target moments as the answering span. 3) **Reinforcement learning methods** [70, 30, 11] that progressively localize the targeting moment by treating the problem of temporal grounding as a sequential decision-making process. There also exist some other studies [22, 78, 21, 23, 79, 42, 49, 92] for the relevant vision-language tasks. Unlike these previous approaches, we advance the video grounding in the perspective of causal inference by mitigating the spurious correlation between video and language features.

2.2. Causal Inference:

Compared to the conventional debiasing techniques [90, 69], causal inference [54, 58, 27] shows its potential in alleviating the spurious correlations [6], disentangling the desired model effects [9], and modularizing reusable features for better generalization [53]. It has been widely adopted in many areas, including image classification [12, 45], imitation learning [19], visual question answering (QA) [15, 1, 51], zero-shot visual recognition [83], long-tailed image recognition and segmentation [65, 64, 88], stable prediction [37], policy evaluation [94], and treatment effect estimation [38]. Specifically, Tang *et al.* [64] and Zhang *et al.* [88] rely on backdoor adjustment [54] to compute the direct causal effect and mitigate the bias introduced by the confounders. Most similar to our work are DIC [80] and CVL [1] which are proposed for image captions and image-based visual QA, and both of them adopt SCM to eliminate the bias for vision-language tasks. The key difference between our work and the previous ones is: our SCM is proposed for video-based moment localization and the latent confounder approximation considers roles, actions and objects in a moment, while DIC and CVL are designed for static image-based vision-language tasks.

2.3. Contrastive Learning and Mutual Information

Contrastive learning [28, 75, 93, 48, 31, 17] methods are proposed to learn representations by contrasting positive pairs against negative pairs. Some prior works consider to maximize the mutual information (MI) between latent representations [34, 5]. MI [8, 35] quantifies the "amount of information" achieved for one random variable through observing the other random variable. There are many estimators [7, 52] to calculate the lower bounds for MI, which have been proven to be effective for unsupervised representation learning [34, 67, 63]. Our DCL is mainly inspired by the Deep InfoMax [34], which maximizes the MI between input data and learned high-level representations. The main differences between our work and the prior studies: 1) our DCL discriminates between positive samples and negative samples based on the temporal moment. 2) our dual contrastive learning module achieves two goals, i.e., guiding the encoder to learn better video representation, as well as better alignment between text and video.
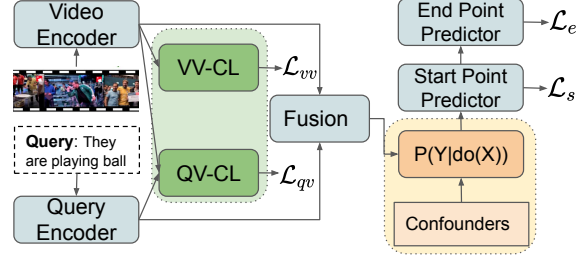

Figure 2: The architecture of our IVG-DCL. VV-CL and QV-CL refer to two contrastive modules with losses expressed as $\mathcal{L}_{vv}$ and $\mathcal{L}_{qv}$. $\mathcal{L}_s$ and $\mathcal{L}_e$ denote the cross-entropy losses for predicting the boundary of the target span.

3. Model

3.1. Overview

Figure 2 depicts the architecture of our IVG-DCL. 1) Given an input query and video, the two encoders output contextualized visual and textual representations respectively. 2) Then, these representations will be fed into two contrastive modules VV-CL and QV-CL respectively to learn high-quality representations with two contrastive losses $\mathcal{L}_{vv}$ and $\mathcal{L}_{qv}$. 3) The output of two feature encoders are fed to a fusion module with a context-query attention mechanism to capture the cross-modal interactions between visual and textual features. 4) Next, to mitigate the spurious correlations between textual and visual features, we use causal interventions $P(Y|do(X))$ with event as surrogate confounders to learn representations. 5) Finally, two losses $\mathcal{L}_s$ and $\mathcal{L}_e$ for the start and end boundaries are introduced. We use the multi-task learning paradigm to train the model with the above four different losses.

3.2. Feature Encoder:

Feature Extraction: Let $V = \{f_t\}_{t=1}^{\hat{T}}$ and $Q = \{w_i\}_{i=1}^{N}$ be a given untrimmed video and textual query, respectively, where f_t and w_i denote the t-th frame and i-th word of the inputs, and $\hat{T}$ and N denote the total number of frames and tokens. We use feature extractor, eg., C3D [66], to obtain the video features $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{T} \in \mathbb{R}^{T \times d_v}$, where T indicates the number of extracted features from the video, $\mathbf{v}_i$ represents the i-th video feature, and d_v refers to the dimension of the video features. As we aim to learn to predict the boundaries of the target moment, which can be denoted as $\mathcal{T}^s$ and $\mathcal{T}^e$, we convert the boundaries $\mathcal{T}^s$ and $\mathcal{T}^e$ to the index of the video features to facilitate our learning process. Let $\mathcal{T}$ denote the video duration, the start and end time index of the moment can be expressed as $I_s = T \times \mathcal{T}^s/\mathcal{T}$ and $I_e = T \times \mathcal{T}^e/\mathcal{T}$, respectively. For the query Q, we also use pre-trained word embeddings, eg., Glove [55], to obtain high-dimensional word features $\mathbf{Q} = \{\mathbf{w}_i\}_{i=1}^{N} \in \mathbb{R}^{N \times d_w}$, where $\mathbf{w}_i$ is the i-th word feature of d_w dimensions.
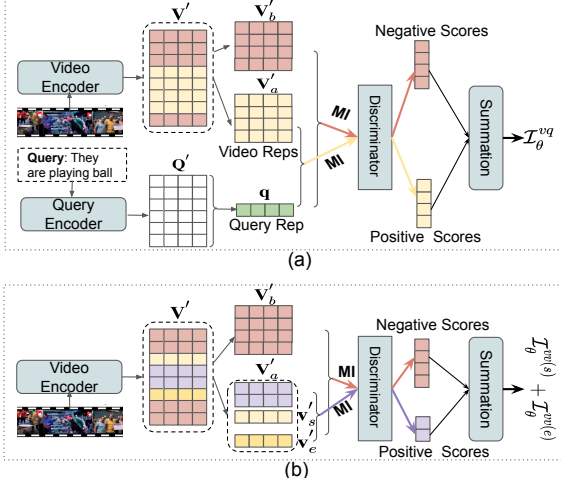
Figure 3: Our `DCL`. $\mathbf{V}'$ and $\mathbf{Q}'$ refer to contextualized video and query representations, respectively. $\mathbf{V}'_a$ and $\mathbf{V}'_b$ represent the positive and negative samples in $\mathbf{V}'$, respectively. (a) our `VQ-CL` module. We generate $\mathbf{q}$ by max-pooling on $\mathbf{Q}'$. (b) our `VV-CL` module. $\mathbf{v}'_s$ and $\mathbf{v}'_e$ refer to the video representations of start and end index of a target moment. The representations (color purple) indicate the other features in the target moment except for starting and ending point.

Encoders: We use two linear layers to project the video features $\mathbf{V}$ and query features $\mathbf{Q}$ to the same dimension d. Then we refer to the previous works [82, 89] and use four convolutions layers, a multi-head attention layer, and a feed-forward layer to generate contextualized representations $\mathbf{Q}' = \{\mathbf{w}'_i\}_{i=1}^{T} \in \mathbb{R}^{N \times d_v}$ and $\mathbf{V}' = \{\mathbf{v}'_i\}_{i=1}^{T} \in \mathbb{R}^{T \times d_v}$. The visual and textual encoders share the same parameters.

3.3. Contrastive Learning Module

Contrastive loss [34] is able to measure the similarities between sample pairs in the representation space, and it can be served as an unsupervised objective function for training the encoder networks with discriminative positive and negative samples. To guide the encoder to better align the textual and video representations $\mathbf{Q}'$ and $\mathbf{V}'$, we treat the video clips that reside in the boundaries of the target moment as the positive samples, and the ones that are outside of the boundaries as the negative samples. And then we use a discriminative approach based on mutual information (MI) maximization [67] to compute the contrastive loss. In information theory, the mutual information refers to the measurement of mutual dependence between two random variables and quantifies the *amount of information* obtained about one variable by observing the other variable.

3.3.1 VQ Contrastive Learning

Figure 3 (a) presents our `VQ-CL` module. We denote $\mathbf{q}$ as the query representation pooled from $\mathbf{Q}'$, and $\mathcal{I}_\theta^{vq}$ as the MI

between $\mathbf{q}$ and $\mathbf{V}'$ with parameter θ. However, MI estimation is generally intractable for continuous and random variables. We therefore alternatively maximize the value over lower bound estimators of MI by Jensen-Shannon MI estimator [34]. We divide the video representation $\mathbf{V}'$ into two parts, i.e., $\mathbf{V}'_a$ as positive samples which denote the features that reside within a target moment, and $\mathbf{V}'_b$ as negative samples that denote the features are located outside of the target segment. The MI $\mathcal{I}_\theta^{vq}$ can be estimated by:

$$\mathcal{I}_\theta^{vq}(\mathbf{q}, \mathbf{V}') := E_{\mathbf{V}'_a}[sp(\mathcal{C}(\mathbf{q}, \mathbf{V}')] \\ - E_{\mathbf{V}'_b}[sp(\mathcal{C}(\mathbf{q}, \mathbf{V}'))] \tag{1}$$

where $\mathcal{C} : \mathbb{R}^{d_v} \times \mathbb{R}^{d_v} \to \mathbb{R}$ refers to the MI discriminator, and $sp(z) = log(1 + e^z)$ is the softplus function. Then we can get the contrastive loss $\mathcal{L}_{vq}$ as follows.

$$\mathcal{L}_{vq} = -\mathcal{I}_\theta^{vq} \tag{2}$$

3.3.2 VV Contrastive Learning

Figure 3(b) depicts the details of our `VV-CL` module. We denote $\mathcal{I}_\theta^{vv(s)}(\mathbf{v}'_s, \mathbf{V}'))$ and $\mathcal{I}_\theta^{vv(e)}(\mathbf{v}'_e, \mathbf{V}'))$ as the MIs between the start and end boundaries of the video and the other clips, respectively, where $\mathbf{v}'_s$ and $\mathbf{v}'_e$ refer to the video representations of the start and end boundaries of the target moment. We follow the same algorithm in Eq. 1 to achieve the results. The loss $\mathcal{L}_{vv}$ can be expressed as follows.

$$\mathcal{L}_{vv} = -\mathcal{I}_\theta^{vv(s)} - \mathcal{I}_\theta^{vv(e)} \tag{3}$$

Therefore, we are able to train a better encoder with the proposed `DCL` by maximizing the MIs as follows.

$$\hat{\theta} = \arg\max_\theta \mathcal{I}_\theta^{vq} + \mathcal{I}_\theta^{vv(s)} + \mathcal{I}_\theta^{vv(e)} \tag{4}$$

3.4. Fusion Module

We leverage context-query attention (CQA) [59] and follow VSLNet [89] to capture the bi-modal interaction between video and text features. We denote CQA as the function for the interaction, which takes the $\mathbf{V}'$, $\mathbf{Q}'$ as the inputs. The output of the fusion module $\mathbf{X}$ can be expressed as:

$$\mathbf{X} = FFN(CQA(\mathbf{V}', \mathbf{Q}')) \tag{5}$$

where FFN refers to a single feed-forward layer.

3.5. Interventional Causal Model

3.5.1 Structural Causal Model

In the case of video grounding, we employ Pearl's SCM [54] to obtain the causal effect between a query and the target moment as illustrated in Figure 4 (a), where the nodes are variables and edges are causal relations. We can justify
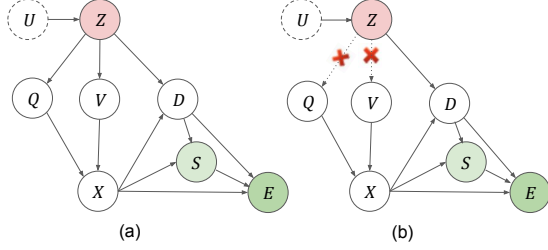
Figure 4: The proposed interventional video grounding (`IVG`) based on structured causal model (SCM). U: selection bias for generating a video grounding dataset. Z: the unobserved confounders that may lead to spurious correlations. Q: textual query. V: untrimmed video. X: the multi-modal representations. D: the mediator. S and E: the start and the end boundaries of the target moment.

the correctness of the graph by detailed interpretation for the subgraphs as follows:

$U \rightarrow Z \rightarrow \{V, Q\}$. An unknown confounder Z (probably imbalanced distribution of dataset caused by sampling bias U) may lead to spurious correlations between video clips and certain words. The do-operation on $\{V, Q\}$ is able to enforce their values and cuts off the direct dependency between $\{V, Q\}$ and their parent Z (Figure 4 (b)).

$Z \rightarrow D \rightarrow \{S, E\}$. Since Z is the set of confounders for the dataset, we must also have Z connected to prediction $\{S, E\}$ via directed paths excluding $\{V, Q\}$. This ensures the consideration of confounding impact from Z to $\{S, E\}$.

$X \rightarrow D \leftarrow Z$. D is a mediator that contains the specified representation of Z considering the presence of X. The adoption of D is a more flexible notation compared to directly have the edge $Z \rightarrow S$ as it permits additional effect of X towards the computation of $\{S, E\}$ through D. This allows us to compare a variety of model architectures under the same causal framework. Besides, unlike previous models [88] that emphasize the mediation effect of X through D, we believe that the main constituent of D should still come from Z, despite having only an approximation.

3.5.2 Causal Intervention

Conventional video grounding models, which are based on correlations between query and video, directly learn $P(S, E|Q, V)$ without considering the confounder set Z. In our SCM, the non-interventional prediction can be expressed using Bayes rule as:

$$P(S, E|Q, V) = \sum_z P(S, E|Q, V, z)P(z|Q, V) \quad (6)$$

However, the above objective learns not only the main direct correlation from $\{Q, V\} \rightarrow X \rightarrow \{S, E\}$ but also the spurious one from the unblocked backdoor path $\{Q, V\} \leftarrow Z \rightarrow D \rightarrow \{S, E\}$. An intervention on $\{Q, V\}$, denoted as

$do(Q, V)$, forces their value and removes the dependency from their parent Z. Therefore, we can adopt do-calculus and backdoor adjustment [54] to alleviate the backdoor correlation and compute the interventional objective:

$$
\begin{aligned}
P(S, &E|do(Q, V)) \\
&= \sum_z P(S, E|do(Q, V), z)p(z|do(Q, V)) \\
&= \sum_z P(S, E|do(Q, V), z)p(z) \\
&= \sum_z P(S, E|do(X), z)p(z)
\end{aligned}
\quad (7)
$$

3.5.3 Latent Confounder Approximation

Unfortunately, the confounder set Z caused by selection bias cannot be observed directly in our case, due to the unavailability of the sampling process. To estimate its distribution, existing approaches [84] mainly rely on image representations from pre-trained classifier, which is expensive and ignores actions embedded in video sequence. Here we propose a novel and efficient approach to approximate the confounder set distribution from the perspective of natural language. Our surrogate confounder set includes the vocabulary of roles, actions and objects extracted from captions, since most video clips are described as "somebody does something". With this in mind, we use the state-of-the-art Open Information Extraction model RnnOIE [62] to extract verb-centered relation tuples (subject, verb, object) from caption texts and construct three sets of vocabulary accordingly. We also compute the prior probability for phrases z in each set based on the dataset statistics:

$$p(z) = \frac{\#z}{\sum_{i \in C} \#i}, \forall z \in C \quad (8)$$

where $\#i$ is the number of occurrences of the phrase i and C is one of the three vocabulary sets. Equipped with approximated Z and its prior distribution, we also propose to model the representation of Z by linear projection g and a word embedding layer $embed$:

$$\mathbf{z} = g(embed(z)) \quad (9)$$

while making sure $dim(\mathbf{z}) = dim(\mathbf{x})$, where $\mathbf{x} \in col(\mathbf{X})$ and col denotes the column space of a matrix and dim indicates the dimension of the vector. Since $P(S, E|do(X))$ is calculated using softmax, we can apply Normalized Weighted Geometric Mean (NWGM)[77] and the deconfounded prediction can be approximated by:

$$
\begin{aligned}
P(S, E|do(X = \mathbf{X})) &= \sum_z P(S, E|\mathbf{X} \oplus \mathbf{z})p(z) \\
&\approx P(S, E| \sum_z (\mathbf{X} \oplus \mathbf{z})p(z))
\end{aligned}
\quad (10)
$$

where $\oplus$ is the broadcast add operator that adds a vector to each vector on the columns of the matrix. Finally, we can achieve the start and end index I_s and I_e of the target moment by the above deconfounded prediction, and also simply use cross entropy to calculate the loss for S (resp. E) as $\mathcal{L}_s$ (resp. $\mathcal{L}_e$).

3.6. Training Objectives

The overall training loss $\mathcal{L}$ can be computed by:

$$\mathcal{L} = \alpha \mathcal{L}_{vq} + \beta \mathcal{L}_{vv} + \mathcal{L}_s + \mathcal{L}_e \tag{11}$$

where α and β are weights for the dual contrastive losses. During the inference stage, the DCL will be ignored as it only facilitates the representation learning and requires the moment annotation to identify the contrastive samples.

4. Experiments

4.1. Dataset

TACoS is constructed from MPII Cooking Composite Activities dataset [57]. We follow the same split of [24] for fair comparisons, where 10,146, 4,589 and 4,083 instances are used for training, validation, and test, respectively. Each video has 148 queries on average.

Charades-STA is a benchmark for the video grounding, which is generated by [24] based on Charades dataset [60] mainly for indoor activities, with 12,408 and 3,720 moment annotations for the training and test, respectively.

ActivityNet Caption involves about 20K videos with 100K queries from the ActivityNet [10]. We refer to the split used in [86], with 37,421 moments for annotations for training, and 17,505 moments for testing. For fair comparisons, we follow [82, 89] for training, evaluation and testing.

4.2. Experimental Settings

Metrics: We follow [24, 89, 86] to use "R@n, IoU = μ" as the evaluation metrics, which denotes the percentage of testing samples that have at least one correct result. A correct result indicates that intersection over IoU with ground truth is larger than μ in top-n retrieved moments.

Settings: We re-implement the VSLBase [89] in Pytorch and use it as our backbone network. We follow the previous works [89, 91] to use the same pre-trained video features [66] and 300-dimension word embedding from Glove [55]. The loss weights α and β are configured as 0.1 and 0.01 respectively.

4.3. Model Zoo

Ranking methods rely on multi-modal matching architectures to obtain the target moment with the highest confidence score, including 2D-TAN [91], MAN [87], etc. Among them, 2D-TAN [91] relies on a temporal adjacent

Type	Model	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
RL	SM-RL [70]	-	24.36	11.17	-
	RWM [30]	-	36.70	-	-
Ranking	CTRL [24]	-	23.63	8.89	-
	ACRN [43]	-	20.26	7.64	-
	SAP [16]	-	27.42	13.36	-
	MAC [25]	-	30.48	12.20	-
	QSPN [76]	54.70	35.60	15.80	-
	2D-TAN [91]	-	39.70	23.31	-
	MAN [87]	-	46.53	22.72	-
Regression	DEBUG [46]	54.95	37.39	17.69	36.34
	DRN [86]	-	45.40	26.40	-
	VSLBase [89]	61.72	40.97	24.14	42.11
	VSLNet [89]	64.30	47.31	30.19	45.15
	Ours	**67.63**	**50.24**	**32.88**	**48.02**

Table 1: Performance comparisons on the Charades-STA dataset. Note that we do not finetune the feature extractor.

network to localize the target moment. MAN [87] captures moment-wise temporal relations as a structured graph and devise an adjustment network to find out the best candidate. **Regression models** directly regress the moment boundaries to avoid heavy computations, including ABLR [82], DEBUG [46], DRN [86], and VSLNet[89], etc. DRN [86] relies on a dense regression network to improve video grounding accuracy by regressing the frame distances to the starting (ending) frame. VSLNet [89] obtains the start and end indexes of the target span based on a QA framework.
Reinforcement learning (RL) methods progressively localize the moment boundaries for a given query. SM-RL [70] and RWM [30] treat the problem of temporal grounding as a sequential decision-making process, which naturally can be resolved by the RL paradigm [85].

4.4. Performance Comparisons

The results of our method on three benchmark datasets are given in the Table 1, Table 2, and Table 3, respectively. As shown in Table 1, our proposed IVG-DCL consistently outperforms the baselines under various settings and achieves the new state-of-the-art performance on the Charades-STA dataset. Compared with VSLNet, our model achieves 3.33 points improvement in IoU=0.3 and 2.87 points measured by mIoU. Our model significantly outperforms a regression baseline VSLBase by 9.27 over IoU = 0.5 and 6.09 in terms of mIoU. The results indicate the superiority of our proposed IVG-DCL in predicting more accurate temporal moment.

Table 2 summarizes the comparisons on the TACoS and it shows that our IVG-DCL performs best among all baselines under IoU=0.1, IoU=0.3, IoU=0.5 and mIoU. For example, our IVG-DCL model outperforms a strong baseline 2D-TAN and the recently proposed DRN by 3.75 and 5.9 in IoU=0.5. The results further confirm the superiority of our approach in capturing the complex interactions of roles, actions, and objects in the video grounding task. Compared

Type	Model	IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
RL	SM-RL [70]	26.51	20.25	15.95	-	-
	TripNet [29]	-	23.95	19.17	-	-
Ranking	ROLE [44]	20.37	15.38	9.94	-	-
	MCN [3]	14.42	-	5.58	-	-
	CTRL [24]	24.32	18.32	13.30	-	-
	ACRN [43]	24.22	19.52	14.62	-	-
	QSPN [76]	25.31	20.15	15.23	-	-
	MAC [25]	31.64	24.17	20.01	-	-
	SAP [16]	31.15	-	18.24	-	-
	TGN [13]	41.87	21.77	18.90	-	-
	2D-TAN [91]	47.59	37.29	25.32	-	-
Regression	SLTA [36]	23.13	17.07	11.92	-	-
	VAL [61]	25.74	19.76	14.74	-	-
	ABLR [82]	34.70	19.50	9.40	-	-
	DEBUG [46]	-	23.45	11.72	-	16.03
	DRN [86]	-	-	23.17	-	-
	VSLBase [89]	-	23.59	20.40	16.65	20.10
	VSLNet [89]	-	29.61	24.27	**20.03**	24.11
	Ours	**49.36**	**38.84**	**29.07**	19.05	**28.26**

Table 2: Performance comparisons on the TACoS dataset.

Type	Model	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
RL	TripNet [29]	48.42	32.19	12.93	-
	RWM [30]	-	36.90	-	-
Ranking	MCN [3]	39.35	21.36	6.43	-
	TGN [13]	43.81	27.93	-	-
	CTRL [24]	47.43	29.01	10.34	-
	ACRN [43]	49.70	31.67	11.25	-
	QSPN [76]	52.13	33.26	13.43	-
	ABLR [82]	55.67	36.79	-	36.99
	2D-TAN [91]	59.45	**44.51**	26.54	-
Regression	DRN [86]	45.45	24.36	-	-
	DEBUG [46]	55.91	39.72	-	39.51
	VSLBase [89]	58.18	39.52	23.21	40.56
	VSLNet [89]	63.16	43.22	26.16	43.19
	Ours	**63.22**	43.84	**27.10**	**44.21**

Table 3: Performance comparisons on ActivityNet Caption.

Causal Model	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
FWA	60.67	40.99	22.85	41.70
CWA	64.35	48.60	30.30	45.63
VCR-CNN	65.62	47.58	28.47	45.94
IVG(Ours)	67.63	50.24	32.88	48.02

Table 4: Comparisons with different causal models on the Charades-STA dataset.

Causal Model	IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
FWA	48.34	36.87	26.19	16.17	26.47
CWA	49.59	36.37	27.07	17.62	27.18
VCR-CNN	47.91	36.87	27.27	16.97	26.76
IVG(Ours)	50.99	38.79	28.89	18.25	28.35

Table 5: Results of different causal models on the TACoS.

to VSLNet, our `IVG-DCL` achieves comparable results on IoU=0.7, and beats it on other settings with significant improvements, e.g., 9.23 gains over IoU=0.3.

Table 3 reports the results on the ActivityNet Caption. We observe that our model achieves state-of-the-art results in most settings. However, the performance gain on this dataset is much smaller than the previous two datasets. One possible reason is that the activities on the ActivityNet Caption are much more complex with more roles, actions, and objects, and the target moment duration is much longer. For example, the average duration of the target moment is 36.18s [89], while the ones on Charades-STA and TACoS are 8.22s and 5.45s, respectively.

4.5. Comparisons with Causal Models

We also adapt four causal models from the image-based tasks and compare them with our proposed `IVG-DCL`, including feature-wise adjustment (FWA) [84], class-wise adjustment (CWA) [84] and VCR-CNN [68]. The detailed descriptions are given as follows. 1) FWA divides the final feature representation to disjoint subsets with respect to the

dimensions and treats the confounder as feature subset selector. 2) Primarily designed for image classification, CWA uses image classes as data stratum to approximate the confounder distribution. Since we entirely rely on pre-trained video features in our video grounding task, we, therefore, adjust this technique for object phrases and compute average word embedding of the object phrases as "class" representation. 3) VCR-CNN is similar to class-wise adjustment but adds an attention-based method to combine the representation between X and the confounding variable. As shown in Table 4 and Table 5, our proposed causal model outperforms these baselines on both Charades-STA and TACoS under various video grounding settings. We believe that the improvement stems from our novel latent confounder approximation method, which considers the roles, actions, as well as objects. These factors are important clues to distinguish activities. The comparisons confirm our hypothesis that such a novel design is useful for the video grounding task. Whereas the previous image-based approaches only consider objects in SCMs.

4.6. Ablation Study

The ablation study is given in Table 6 to illustrate the contributions of each component of our model. As shown, removing the proposed `IVG` module will result in 4 points of performance drop on average, indicating that the causal model plays the key role to the performance gain. We also observed that the removal of any contrastive learning module, i.e., `QV-CL` or `VV-CL` will decrease the performance by 1 point on average. While the removal of `DCL` will bring about 2 points of performance drop on average. This indicates that both contrastive modules contribute to the temporal localization, and combing these two will bring more gains. It is not surprising that removing both `IVG` and `DCL` will sharply decrease by 10 points on average, as essentially
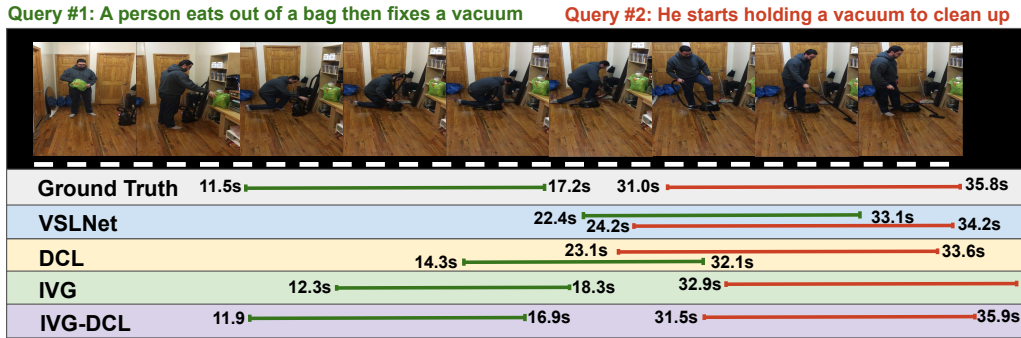
Figure 5: A case study on the Charades-STA dataset to demonstrate the capability of our model in mitigating the spurious correlations between textual and video features.

it becomes the VSLBase model.

Model	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
Full Model	67.63	50.24	32.88	48.02
w/o IVG	64.70	47.60	30.80	45.34
w/o QV-CL	66.30	49.31	31.19	47.15
w/o VV-CL	66.75	49.16	32.17	47.24
w/o DCL	65.21	48.76	31.93	46.94
w/o IVG+DCL	61.72	40.97	24.12	42.11

Table 6: Ablation study on the Charades-STA dataset.

4.7. Sensitivity on α and β

To understand how the loss weights α and β influence the performance, we manually set different values to the two hyper-parameters. Figure 7 reports the results. We found that the best combination is $\{\alpha = 0.1, \beta = 0.01\}$. These weights indicate the importance of each loss in our multi-task learning paradigm for temporal localization. We found that increasing the α and β will lead to performance drop, and this aligns with our hypothesis that the IVG plays more important role compared with DCL. We also observed that the combination $\{\alpha = 1, \beta = 1\}$ will even decrease the performance by 5 points on average. This implies that directly integrating unsupervised representation learning to SCMs may not yield good performance.

$\{\alpha, \beta\}$	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
$\{0.1, 0.01\}$	67.63	50.24	32.88	48.02
$\{1, 1\}$	62.93	43.49	26.08	43.65
$\{0.5, 0.1\}$	64.44	49.97	31.45	46.09
$\{0.1, 0.5\}$	63.23	44.68	27.37	44.08
$\{1.5, 1\}$	64.33	44.78	27.04	44.89

Table 7: Sensitivity analysis on the Charades-STA dataset.

4.8. Qualitative Analysis

We show a case in Figure 5 from Charades-STA dataset to demonstrate how our proposed ICV-DCL alleviates the

spurious correlation between text and video features. As there are a lot more relevant training instances for query #2 "he starts holding a vacuum to clean up" compared to query #1 "a person fixes a vacuum" (208 instances vs. 35 instances) in the training set, there will exist unexpected high correlations between vacuum-cleaning queries and vacuum-related video moments. The conventional VSLNet that is trained only based on the correlations will tend to incorrectly localize the query #1 to the query #2 related segments. We observe that our IVG-DCL outputs more accurate retrieval for both two queries as it is able to better distinguish between "holding" and "fixing". We believe such improvement mainly stems from the proposed causal model.

5. Conclusion

This paper proposes IVG-DCL, a novel model which aims to eliminate the spurious correlations between query and video features based on causal inference. Experiments on three standard benchmarks show the effectiveness of our proposed method. In the future, we plan to extend our approach to more unbiased vision-language tasks that involve dialogues [41, 72, 39] and question answering [4, 81, 15].

Acknowledgments

References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *CVPR*, 2020. 3

[2] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *CVPR*, 2020. 2

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 1, 2, 7

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *CVPR*, 2015. 2, 8

[5] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, 2019. 3

[6] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *AISTATS*, 2012. 3

[7] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018. 3

[8] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 1995. 3

[9] Michel Besserve, Arash Mehrjou, Rémy Sun, and Bernhard Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *ICLR*, 2019. 3

[10] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 6

[11] Da Cao, Yawen Zeng, Meng Liu, Xiangnan He, Meng Wang, and Zheng Qin. Strong: Spatio-temporal reinforcement learning for cross-modal video moment localization. In *ACMM*, 2020. 2, 3

[12] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. *arXiv preprint arXiv:1412.2309*, 2014. 3

[13] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, 2018. 7

[14] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. In *AAAI*, 2019. 2

[15] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, 2020. 3, 8

[16] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *AAAI*, 2019. 6, 7

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 3

[18] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 2

[19] Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *NeurIPS*, 2019. 3

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[21] Hehe Fan and Yi Yang. Person tube retrieval via language description. In *AAAI*, 2020. 3

[22] Hehe Fan, Linchao Zhu, Yi Yang, and Fei Wu. Recurrent attention network with reinforced generator for visual dialog. *ACM Transactions on Multimedia Computing, Communications, and Applications TOMM*, 16(3):78:1–78:16, 2020. 3

[23] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 2, 3

[24] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *CVPR*, 2017. 1, 2, 6, 7

[25] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *WACV*, 2019. 2, 6, 7

[26] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander G Hauptmann. Excl: Extractive clip localization using natural language descriptions. In *ACL*, 2019. 2

[27] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020. 3

[28] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 3

[29] Meera Hahn, Asim Kadav, James M Rehg, and Hans Peter Graf. Tripping through time: Efficient localization of activities in videos. *arXiv preprint arXiv:1904.09936*, 2019. 7

[30] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*, 2019. 2, 3, 6, 7

[31] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[33] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. In *EMNLP*, 2018. 2

[34] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2018. 3, 4

[35] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000. 3

[36] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. Cross-modal video moment retrieval with spatial and language-temporal attention. In *ICMR*, 2019. 7

[37] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *SIGKDD*, 2018. 3

[38] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Yashen Wang, Fei Wu, and Shiqiang Yang. Treatment effect estimation via differentiated confounder balancing and regression. *TKDD*, 2019. 3

[39] Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. Multi-modal transformer networks for end-to-end video-grounded dialogue systems. In *ACL*, 2019. 8

[40] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 1, 2

[41] Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *ACL*, 2018. 8

[42] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, 2020. 3

[43] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *SIGIR*, 2018. 2, 6, 7

[44] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *ACMMM*, 2018. 2, 7

[45] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *CVPR*, 2017. 3

[46] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. Debug: A dense bottom-up grounding approach for natural language video localization. In *EMNLP*, pages 5147–5156, 2019. 2, 6, 7

[47] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 2

[48] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020. 3

[49] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, 2019. 3

[50] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, 2020. 2

[51] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, 2021. 3

[52] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[53] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *ICML*, 2018. 3

[54] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 2, 3, 4, 5

[55] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 3, 6

[56] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *CVPR*, 2020. 2

[57] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities, 2012. 2, 6

[58] Donald B Rubin. Essential concepts of causal inference: a remarkable history and an intriguing future. *Biostatistics & Epidemiology*, 3(1):140–155, 2019. 2, 3

[59] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *ICLR*, 2016. 4

[60] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 6

[61] Xiaomeng Song and Yahong Han. Val: Visual-attention action localizer. In *PCM*, 2018. 7

[62] Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and I. Dagan. Supervised open information extraction. In *NAACL-HLT*, 2018. 5

[63] Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *ICLR*, 2019. 3

[64] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020. 3

[65] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020. 2, 3

[66] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 3, 6

[67] Petar Velickovic, William Fedus, William L Hamilton, Pietro Lio, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *ICLR*, 2019. 3, 4

[68] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *CVPR*, 2020. 2, 7

[69] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *ECCV*, 2020. 2, 3

[70] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, 2019. 2, 3, 6, 7

[71] Xiao Wang, Shaohua Fan, Kun Kuang, Chuan Shi, Jiawei Liu, and Bai Wang. Decorrelated clustering with data selection bias. In *IJCAI*, 2020. 2

[72] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *ACL*, pages 5635–5649, 2019. 8

[73] Aming Wu and Yahong Han. Multi-modal circulant fusion for video-to-language and backward. In *IJCAI*, 2018. 2

[74] Jiaxin Wu, Sheng-hua Zhong, and Yan Liu. Dynamic graph convolutional network for multi-video summarization. *Pattern Recognition*, 107:107382, 2020. 1, 2

[75] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 3

[76] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, 2019. 2, 6, 7

[77] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, R. Salakhutdinov, R. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *ArXiv*, abs/1502.03044, 2015. 5

[78] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. Tree-augmented cross-modal encoding for complex-query video retrieval. In *SigIR*, 2020. 3

[79] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *ACM MM*, pages 1939–1947, 2020. 3

[80] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*, 2020. 3

[81] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 8

[82] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, 2019. 1, 4, 6, 7

[83] Zhongqi Yue, Tan Wang, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Counterfactual zero-shot and open-set visual recognition. *CVPR*, 2021. 3

[84] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In *NeurIPS*, 2020. 2, 5, 7

[85] Sangdoo Yun, Jongwon Choi, Youngjoon Yoo, Kimin Yun, and Jin Young Choi. Action-decision networks for visual tracking with deep reinforcement learning. In *CVPR*, 2017. 6

[86] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *CVPR*, 2020. 2, 6, 7

[87] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, 2019. 1, 2, 6

[88] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *NeurIPS*, 2020. 3, 5

[89] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *ACL*, 2020. 2, 4, 6, 7

[90] Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. In *NAACL*, 2019. 2, 3

[91] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, 2020. 6, 7

[92] Z. Zhang, Z. Zhao, Zhijie Lin, J. Zhu, and X. He. Counterfactual contrastive learning for weakly-supervised vision-language grounding. In *NeurIPS*, 2020. 3

[93] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *ICCV*, 2019. 3

[94] Hao Zou, Kun Kuang, Boqi Chen, Peixuan Chen, and Peng Cui. Focused context balancing for robust offline policy evaluation. In *SIGKDD*, 2019. 3