Interventional Video Grounding with Dual Contrastive Learning

Guoshun Nan1,* Rui Qiao1,* Yao Xiao2,† Jun Liu3,‡ Sicong Leng1 Hao Zhang4,5 Wei Lu1,‡
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., \( P(Y \mid X) \). 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 \( P(Y \mid do(X)) \). 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 [34, 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 $L_{vv}$ and $L_{qv}$. $L_s$ and $L_c$ 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 $L_{vv}$ and $L_{qv}$. 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 $L_s$ and $L_c$ 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 $V = \{v_i\}_{i=1}^T$ and $Q = \{w_i\}_{i=1}^N$ be a given untrimmed video and textual query, respectively, where $f_i$ and $w_i$ denote the $t$-th frame and $i$-th word of the inputs, and $T$ and $N$ denote the total number of frames and tokens. We use feature extractor, eg., C3D [66], to obtain the video features $V = \{v_i\}_{i=1}^T \in \mathbb{R}^{T \times d_v}$, where $T$ indicates the number of extracted features from the video, $v_i$ 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 $T^s$ and $T^e$, we convert the boundaries $T^s$ and $T^e$ to the index of the video features to facilitate our learning process. Let $T$ denote the video duration, the start and end time index of the moment can be expressed as $I_s = T \times T^s/T$ and $I_e = T \times T^e/T$, respectively. For the query $Q$, we also use pre-trained word embeddings, eg., Glove [55], to obtain high-dimensional word features $Q = \{w_i\}_{i=1}^N \in \mathbb{R}^{N \times d_w}$, where $w_i$ is the $i$-th word feature of $d_w$ dimensions.
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 \( Q' \) and \( V' \), 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 \( q \) as the query representation pooled from \( Q' \), and \( I_{\theta}^{qv} \) as the MI between \( q \) and \( V' \) with parameter \( \theta \). 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 \( V' \) into two parts, i.e., \( V'_a \) as positive samples which denote the features that reside within a target moment, and \( V'_b \) as negative samples that denote the features are located outside of the target segment. The MI \( I_{\theta}^{qv} \) can be estimated by:

\[
I_{\theta}^{qv}(q, V') := E_{V_a}[sp(C(q, V'))] - E_{V_b}[sp(C(q, V'))]
\]

(1)

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

\[
L_{qv} = -I_{\theta}^{qv}
\]

(2)

3.3.2VV Contrastive Learning

Figure 3 (b) depicts the details of our VV-CL module. We denote \( I_{\theta}^{vv(s)}(v'_s, V') \) and \( I_{\theta}^{vv(e)}(v'_e, V') \) as the MIs between the start and end boundaries of the video and the other clips, respectively, where \( v'_s \) and \( v'_e \) 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 \( L_{vv} \) can be expressed as follows.

\[
L_{vv} = -I_{\theta}^{vv(s)} - I_{\theta}^{vv(e)}
\]

(3)

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

\[
\hat{\theta} = \arg \max_{\theta} I_{\theta}^{qv} + I_{\theta}^{vv(s)} + I_{\theta}^{vv(e)}
\]

(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 \( V', Q' \) as the inputs. The output of the fusion module \( X \) can be expressed as:

\[
X = F F N(CQA(V', Q'))
\]

(5)

where \( F F N \) refers to a single feed-forward layer.

3.5. Intervenational 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
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:

\[
P(S, E|do(Q, V)) = \sum_z P(S, E|do(Q, V), z)p(z|do(Q, V)) = \sum_z P(S, E|do(Q, V), z)p(z) \quad (7)
\]

\[
= \sum_z P(S, E|do(X), z)p(z)
\]

### 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 RmOIE [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 \( embed \):

\[
z = g(embed(z)) \quad (9)
\]

while making sure \( dim(z) = dim(x) \), where \( x \in col(X) \) 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:

\[
P(S, E|do(X = X)) = \sum_z P(S, E|X \oplus z)p(z)
\]

\[
\approx P(S, E)\sum_z (X \oplus z)p(z) \quad (10)
\]
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 $L_s$ (resp. $L_e$).

### 3.6. Training Objectives

The overall training loss $L$ can be computed by:

$$L = \alpha L_{vq} + \beta L_{vv} + L_s + L_e$$

(11)

where $\alpha$ and $\beta$ 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 = $\mu$" 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 $\mu$ 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 $\alpha$ and $\beta$ 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 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. SMRL [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
to VSLNet, our IVG-DCL achieves comparable results on IoU=0.7, and beats it on other settings with significant improvements, e.g., 9.23s 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 dataset is much smaller than the previous two datasets. One possible reason is that the performance of our model is comparable with the previous models.

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 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 combining 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
it becomes the VSLBase model.

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

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

4.7. Sensitivity on $\alpha$ and $\beta$

To understand how the loss weights $\alpha$ and $\beta$ 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 $\alpha$ and $\beta$ 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.

<table>
<thead>
<tr>
<th>${\alpha, \beta}$</th>
<th>IoU=0.3</th>
<th>IoU=0.5</th>
<th>IoU=0.7</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>${0.1, 0.01}$</td>
<td>67.63</td>
<td>50.24</td>
<td>32.88</td>
<td>48.02</td>
</tr>
<tr>
<td>${1, 1}$</td>
<td>62.93</td>
<td>43.49</td>
<td>26.08</td>
<td>43.65</td>
</tr>
<tr>
<td>${0.5, 0.1}$</td>
<td>64.44</td>
<td>49.97</td>
<td>31.45</td>
<td>46.94</td>
</tr>
<tr>
<td>${0.1, 0.5}$</td>
<td>63.23</td>
<td>44.68</td>
<td>27.37</td>
<td>44.08</td>
</tr>
<tr>
<td>${1.5, 1}$</td>
<td>64.33</td>
<td>44.78</td>
<td>27.04</td>
<td>44.89</td>
</tr>
</tbody>
</table>

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

We would like to thank the anonymous reviewers for their helpful comments. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-RP-2019-012), and partially supported by SUTD Project (No: PIE-SGP-AI2020-02), the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Fund (Project No: A18A1b0045 and No: A18A2b0046). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore and AI Singapore.
References


[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
