

Counterfactual Cross-modality Reasoning for Weakly Supervised Video Moment Localization

Zezhong Lv
zezhonglv0306@gmail.com
Gaoling School of Artificial
Intelligence, Renmin University of
China
Beijing Key Laboratory of Big Data
Management and Analysis Methods
Beijing, China

Bing Su*
subingats@gmail.com
Gaoling School of Artificial
Intelligence, Renmin University of
China
Beijing Key Laboratory of Big Data
Management and Analysis Methods
Beijing, China

Ji-Rong Wen
jrwen@ruc.edu.cn
Gaoling School of Artificial
Intelligence, Renmin University of
China
Beijing Key Laboratory of Big Data
Management and Analysis Methods
Beijing, China

ABSTRACT

Video moment localization aims to retrieve the target segment of an untrimmed video according to the natural language query. Weakly supervised methods gains attention recently, as the precise temporal location of the target segment is not always available. However, one of the greatest challenges encountered by the weakly supervised method is implied in the mismatch between the video and language induced by the coarse temporal annotations. To refine the vision-language alignment, recent works contrast the cross-modality similarities driven by reconstructing masked queries between positive and negative video proposals. However, the reconstruction may be influenced by the latent spurious correlation between the unmasked and the masked parts, which distorts the restoring process and further degrades the efficacy of contrastive learning since the masked words are not completely reconstructed from the cross-modality knowledge. In this paper, we discover and mitigate this spurious correlation through a novel proposed counterfactual cross-modality reasoning method. Specifically, we first formulate query reconstruction as an aggregated causal effect of cross-modality and query knowledge. Then by introducing counterfactual cross-modality knowledge into this aggregation, the spurious impact of the unmasked part contributing to the reconstruction is explicitly modeled. Finally, by suppressing the unimodal effect of masked query, we can rectify the reconstructions of video proposals to perform reasonable contrastive learning. Extensive experimental evaluations demonstrate the effectiveness of our proposed method. The code is available at <https://github.com/sLdZ0306/CCR>.

CCS CONCEPTS

• Information systems → Video search.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0108-5/23/10...\$15.00

<https://doi.org/10.1145/3581783.3612495>

KEYWORDS

video moment localization; cross-modal retrieval

ACM Reference Format:

Zezhong Lv, Bing Su, and Ji-Rong Wen. 2023. Counterfactual Cross-modality Reasoning for Weakly Supervised Video Moment Localization. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, October 29–November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3581783.3612495>

1 INTRODUCTION

The widespread use of the internet and mobile devices has led to an unprecedented surge in multimedia content consumption, especially videos, due to their stronger capacity for information expression compared to other media forms. As a result, video understanding has become a crucial challenge in computer vision, encompassing a vast range of research topics, such as video highlight detection [12, 32] and temporal action localization [3, 9, 18]. While these tasks primarily focus on single video modality, video moment localization [8, 13] is increasingly becoming a central research topic in modeling the relationship between video and natural language, which is largely motivated by the requirements of the cross-modality application scenarios. Concretely, given a natural language query, video moment localization aims at localizing the start and end boundaries of the target video segment from an untrimmed video according to the semantic contents of the query. Fully supervised video moment localization methods have achieved promising retrieval performance, which are trained with the exact temporal labels (*i.e.* the start and end times of the target moment) provided in the given datasets. However, these fine-grained temporal labels are not always accessible in many application scenarios, and manually annotating such labels are expensive and time-consuming. Besides, the performance of a fully supervised method heavily relies on the quality of the ground truth labels. Nevertheless, it has been reported that there is a significant man-made temporal bias existing in the distributions of the boundary labels in the widely used benchmarks of video moment localization [15].

Weakly supervised methods are recently [37, 38, 40] proposed to fix these issues, which are trained with the matched pairs of a whole untrimmed video and its language query and do not require the temporal labels during training anymore. Due to the lacking of concrete temporal labels, there are no indications of the precise semantic alignment between the query and its corresponding segment in the given video. Therefore, weakly supervised methods can only learn the relationship between vision and language modalities

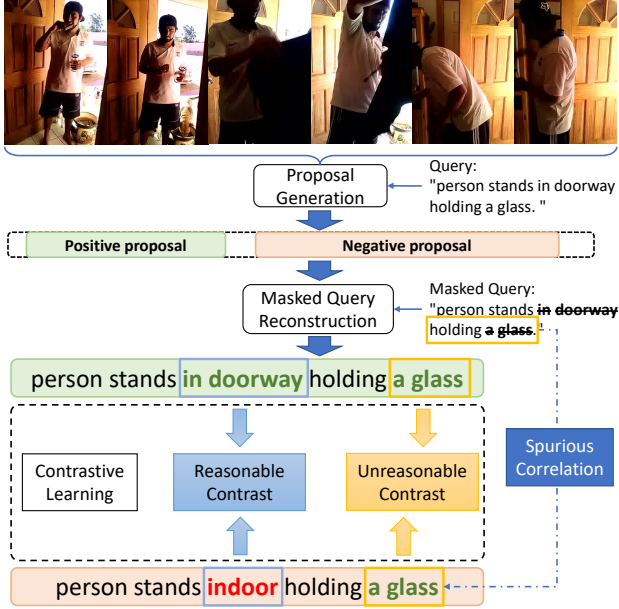


Figure 1: A general query reconstruction based weakly supervised video moment localization framework. The reconstruction based on positive and negative proposals is contrasted to learn the cross-modality alignment. However, the spurious correlation between un-masked and masked words confuses the contrastive learning process. e.g. the prediction for the masked “indoor” can differ between positive and negative proposals due to changes in the person’s location. In contrast, the word “glass” may still be correctly predicted even if it does not appear in the negative proposal. This abnormal correct prediction is due to a certain pattern learned from the language rather than cross-modality knowledge, which leads to an unreasonable contrast.

from a more coarse perspective, *i.e.* video-query pairs. Because there is no fine-grained vision-language alignment provided in the training process, the cross-modality interactions can be confused by the latent mismatch caused by the coarse relationship. One way to build the connection from such coarsely matched pairs is the mask-reconstruction technique [20], which partially masks one sample and learns to reconstruct the masked part from the other sample. This methodology requires no fine-grained annotations to find and model the corresponding information shared between the two modalities. Concretely, recent state-of-the-art weakly supervised video moment localization methods [40, 41] apply this technique through learning a cross-modality fusion model to predict the masked query based on the video moment candidates, and using the accuracy of the prediction as the measurement of the similarity to apply contrastive learning.

As illustrated in Figure 1, one of the common end-to-end schemes utilized by existing weakly supervised methods can be summarized as follows: (1) Generating several candidate proposals from the untrimmed video based on the cross-modality features, and extracting their features; (2) By interacting the proposal features with the masked query, the reconstructed query for each proposal is obtained, while the reconstruction loss is taken as the proxy of the alignment score between video proposal and query; These two

steps enable the model to discover the fine-grained cross-modality alignment with no temporal labels; (3) Applying intra-sample contrastive learning between the reconstruction errors of the positive and negative proposals to train the model. By comparing the reconstruction-driven alignment scores obtained by the positive and negative proposals respectively, the model is forced to recognize the difference between the well-aligned video segment and the irrelevant ones with respect to the given query, and thus generates promising proposals as the final localization results during inference. The core idea of these methods relies on that a larger similarity between the video proposal and the query implies the masked query can be reconstructed easier according to the video proposal, and vice versa.

However, when training the cross-modality interaction module on biased queries in the dataset, which illustrate certain patterns of word combinations, the model learns the spurious correlations implied by these combinations. For example, the non-uniform joint distribution of the noun and predicate in queries reported in [33] will make the model tend to predict the masked word based on the combination of regular word pairs. As illustrated in Figure 1, given a query “person stands in doorway holding a glass”, when we mask the word “glass” and reconstruct it, the answer could be predicted directly based on the language knowledge rather than the cross-modality fusion model because of the highly statistic correlation between the word “glass” and “holding” implying by dataset. In other words, the difficulty of the query reconstruction is significantly lower because the masked words can be approximately reconstructed even if the mismatching video moment proposal is fed into the cross-modality model. Thus, these spurious correlations will distort the similarity measurement generated by the cross-modality model, which is the fundamental of this paradigm to correlate vision and language information in semantic space, and further degrade the performance of contrastive learning.

We propose to discover and mitigate these spurious correlations through Counterfactual Cross-modality Reasoning (CCR). Specifically, we disentangle the total causal effect on the reconstruction of the masked query as an aggregation of two individual branches, the main branch and the side branch, which model the vision-language cross-modality knowledge and query language knowledge, respectively. The main branch is applied to predict the original query based on the interaction of the information embedded in both the video and masked query, while the side branch is utilized to measure the contribution of the un-masked tokens in masked query for the reconstruction. By applying a counterfactual cross-modality knowledge in the aggregation of two branches, the unimodal impact of the masked query is extracted. Because the reconstruction of the masked query should be performed mainly by the cross-modality interaction between the video and masked query, we weaken the contribution of the uni-modal masked query in the reconstruction by directly removing it from the final prediction.

We summarize our major contributions as follows: (1) We propose a novel method, counterfactual cross-modality reasoning, for weakly supervised video moment localization, aimed at discovering and mitigating potential spurious correlations between different words in the query. (2) We formulate the query reconstruction task in weakly supervised video moment localization from a causal reasoning perspective, and disentangle the causal effect on the

prediction into main-branch and side-branch, which encode the cross-modality and query knowledge respectively. Through aggregating a counterfactual cross-modality knowledge with the query, the unimodal effect contributed by the masked query is explicitly modeled. By suppressing this spurious effect, the prediction of the original query is rectified to rely more on the cross-modality knowledge rather than the un-masked words in the query, and thus the efficacy of contrastive learning is directly promoted. (3) We evaluate the proposed CCR on ActivityNet Captions and Charades-STA benchmark datasets. Experimental results show that our CCR significantly outperforms the state-of-the-art baseline.

2 RELATED WORK

Fully supervised video moment localization. Video moment localization is formulated in a fully supervised setting in early studies [8, 16, 34–36], which is trained based on the exact temporal boundaries of video moment corresponding to each query. Existing methods can be divided into two categories, which utilize anchor-based and anchor-free paradigms respectively, according to whether they need anchors during training. Anchor-free methods [16, 35] predict the probability of the temporal boundary for each frame within the given video based on the cross-modality fusion feature in a one-stage manner. While anchor-based methods [8, 34, 36] firstly generate a set of video proposals, and then train an interaction model to predict the similarities between these proposals and the given query. In fully supervised setting, the proposal generation processing can be treated as an additional supervision signal compared to anchor-free methods, and thus the performance of anchor-based methods usually surpasses that of anchor-free methods. However, it is expensive and sometimes unreliable to manually annotate the precise temporal boundary [15], which indeed restricts the generalization performance of fully supervised methods.

Weakly supervised video moment localization. To increase the scalability of video moment localization in real life practice, weakly supervised methods [17] are introduced with no requirement of the boundary label. Because there is no precise temporal location of the target moment, most of the existing weakly supervised methods follow anchor-based paradigm to train their models based on the generated proposals, [5, 14, 17, 28] propose to generate video proposals by utilizing sliding temporal windows strategies. [20] firstly introduces the self-supervised reconstruction of the masked query to connect the information between video and query for weakly supervised video moment localization. The reconstruction loss is utilized as the similarity measurement between the proposals and query, where the rationale lies in that the visual content inside the matched video proposal should be more helpful comparing to the mismatched ones. However, generating proposal through an enumerate manner is unreasonable because they are irrelevant with neither video nor query semantic content. Besides, in order to cover more potential video moment, they have to increase the number of proposals, which cause huge computational cost. Recent works propose to generate video proposals utilizing learn-based methods, [40, 41] apply a multi-model transformer to fuse the video and query, which outputs the parameters of a series of temporal weights shaped like Gaussian distributions as the positive proposals and the corresponding negative proposals obtained in a heuristic method. After that, they feed the temporal proposals

along with masked query into the transformer to reconstruct the masked words and apply contrastive learning to contrast between the similarities between the positive and negative proposals. However, the central step, which is the reconstruction of the masked query, can be turbulent because of the spurious correlation between the masked words and the un-masked ones. To tackle this issue, we propose Counterfactual Cross-modality Reasoning (CCR) to decouple the causal effect on the prediction of the masked words into the effect of cross-modality fusion knowledge and query knowledge, which are indicated as the main-branch and side-branch respectively, and thus mitigate the spurious correlation inside the query by suppressing the contribution of the side-branch.

Causal reasoning in multi-model learning. Causal reasoning has shown its capability of resolving the ubiquitous biased training set and spurious correlation issues in multi-model fields including video corpus moment retrieval [33] and video question answering [2]. [2, 23] try to reduce the unimodal biases in video question answering through modeling the statistical regularities between the question and the answer, which have similar key idea to our CCR. However, our CCR is proposed to calibrate the cross-modality contrastive learning process by rectifying the masked query reconstruction, while [2, 23] are designed for de-biased answer making. [38] designs a two-stage paradigm started with generating proposals through coarse contrastive learning between different video-query pairs, and then develops three memory bank based heuristical transformations to apply counterfactual contrastive learning on the generated proposals within a mini-batch to tackle weakly supervised video moment localization. Nevertheless, heuristically replacing and perturbing the proposal features is not sufficient and reliable for a reasonable counterfactual situation because it heavily relies on the positive and negative proposals generated in the first inter-sample based contrastive learning. By comparison, our CCR is proposed to facilitate the alignment between fine-grained intra-video proposals and the given query by erasing the spurious correlation hidden in the masked query reconstruction process.

3 METHOD

3.1 Preliminary

Given a natural language query W and a video V , video moment localization task aims to localize a video moment \hat{V} corresponding to the semantic information of the query W . To localize this segment, a prediction model $\eta_{\theta}(\hat{V}|V, W)$ parameterized by θ is trained to minimize the distance between \hat{V} and the ground truth segment \tilde{V} according to the assessment metric such as intersection-over-union.

However, the ground truth \tilde{V} is unavailable in the weakly supervised setting. To refine the video-level annotated query to the corresponding target segment level, video proposals are generated as the candidates to semantically match with the query. To model the semantic alignment between the query and video proposals, the query is partially masked, and the model is trained to learn cross-modality fusion knowledge to reconstruct the original query. This is an efficient methodology to build fine-grained alignment in weakly supervised settings [20]. Following the scheme illustrated in Figure 1, we elaborate on the procedure of query reconstruction-based weakly supervised video moment retrieval as follows: (1) Given a video V its corresponding query W , positive video proposals $S^P = \{S_i^P | i = 1, \dots, N^P\}$ and negative video

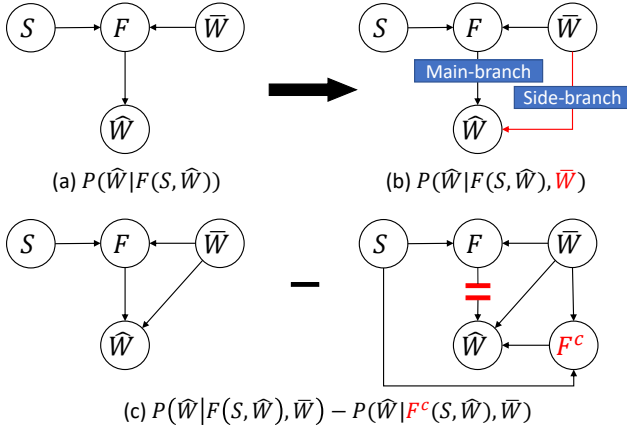


Figure 2: S , \bar{W} , and \hat{W} indicate the proposal feature, masked query, and reconstructed query, we omit the superscripts $* \in \{p, n\}$ of S in SCM; F is the cross-modality fusion knowledge. (a) Conventional query reconstruction SCM based on the cross-modality fusion between video and masked query. (b) Modified causal graph decouples the effect on reconstruction into the main-branch and the side-branch, which encode the cross-modal and the unimodal impact of the masked query respectively. (c) Removing the spurious correlation between the masked query and its reconstruction by introducing the counterfactual cross-modality knowledge F^c .

proposals $S^n = \{S_j^n | j = 1, \dots, N^n\}$ are obtained by interacting the video and query, where N_p and N_n are the numbers of positive and negative proposals, respectively. (2) For each proposal $S_i^p \in S^p$ and $S_j^n \in S^n$, model learns to reconstruct the original query as \hat{W}_i^p, \hat{W}_j^n based on the masked query \bar{W} respectively. (3) Reconstruction losses $L_i^p = \text{Loss}(\hat{W}_i^p, W)$ and $L_j^n = \text{Loss}(\hat{W}_j^n, W)$ are utilized to indicate the similarities between the video proposals and the query, where a lower loss implies a higher alignment degree between the proposal and query, and vice versa. (4) Perform contrastive learning between the reconstruction losses of positive and negative proposals $\langle S^p, S^n \rangle$ to train the model. (5) During inference, S^n is neglected, and the reconstruction loss L_i^p for each $S_i^p \in S^p$ is calculated. The proposal with the lowest L_i^p is output as the final localization. From the scheme above we can find that, query reconstruction is the core step that not only semantically connects vision and language but also serves as the measurement during evaluation. One of the key factors in establishing semantic alignment between video proposals and queries is to use contrastive learning to increase the discrepancy between the reconstruction losses of positive and negative proposals. However, the spurious correlation between the masked query and its reconstruction leads to an invalid contrast between positive and negative video proposals, which directly perturbs the cross-modality alignment.

3.2 Revisit Masked Query Reconstruction in Causality View

To model and further mitigate this spurious correlation, we propose to revisit the cross-modality fusion based query reconstruction task from the perspective of the Structured Causal Model (SCM) [11].

Without loss of generality, we can simplify the notation of each video proposal by omitting all subscripts and just denoting it as S . We introduce the cross-modality knowledge as F and formulate the Structured Causal Model (SCM) of the conventional query reconstruction methodology in Figure 2 (a), as follows

$$P(\hat{W}|F(S, \bar{W})), \quad (1)$$

where the cross-modality fusion knowledge is the only causal of the query reconstruction, as the $\{S, \bar{W}\} \rightarrow F \rightarrow \hat{W}$ path indicated in Figure 2 (a).

However, the SCM in Figure 2 (a) neglects the spurious correlation between the masked query \bar{W} and the final prediction. Therefore, from a perspective of causality, we modify the conventional SCM by adding a causal connection from \bar{W} directly to \hat{W} to model this spurious correlation, as shown in Figure 2 (b). We note that this causal effect is formulated individually from the existing cross-modality fusion knowledge F because the original query can be reconstructed only by the unimodal impact of the masked query, which is induced by certain patterns learned by the model. Based on the upgraded SCM, we can reformulate the masked query reconstruction as the result of the aggregated effect of two branches

$$P(\hat{W}|F(S, \bar{W}), \bar{W}), \quad (2)$$

where $F(S, \bar{W})$ and \bar{W} indicate the causal effects of main-branch $F \rightarrow \hat{W}$ and side-branch $\bar{W} \rightarrow \hat{W}$ respectively. To explicitly model the spurious correlation implying in the side-branch $\bar{W} \rightarrow \hat{W}$, we cut off the potential impact from the main-branch $F \rightarrow \hat{W}$ by applying a counterfactual cross-modality knowledge F^c which does not provide any useful information for establishing semantic interaction between the video and query, and hence, does not assist in the reconstruction of the original query. As illustrated in Figure 2 (c), by replacing the cross-modality effect $F(S, \bar{W})$ in Equation (2) by $F^c(S, \bar{W})$, we obtain the counterfactual reconstruction as

$$P(\hat{W}|F^c(S, \bar{W}), \bar{W}), \quad (3)$$

which is known as the total indirect effect in causality [24]. Following this, we can finally eliminate the spurious correlation between the masked query and original query reconstruction as

$$P(\hat{W}|F(S, \bar{W}), \bar{W}) - P(\hat{W}|F^c(S, \bar{W}), \bar{W}), \quad (4)$$

as shown in Figure 2 (c), where the effect on reconstruction is attributed solely to the interaction F between video proposal and query. Therefore, our proposed methodology can effectively capture the true causal relationship between the reconstruction and cross-modality knowledge.

It is important to note that reducing the spurious correlation induced by masked query solely by reconstructing queries from the video proposal is not feasible. This is due to the presence of redundant visual content in video proposal, and masked query plays a crucial attentional role in establishing cross-modality alignments.

3.3 Counterfactual Cross-modality Reasoning

In this subsection, we present a detailed description of the proposed Counterfactual Cross-modality Reasoning (CCR), as illustrated in Figure 3. The central idea of CCR is to mitigate the spurious correlation between the masked query and its reconstruction. To achieve

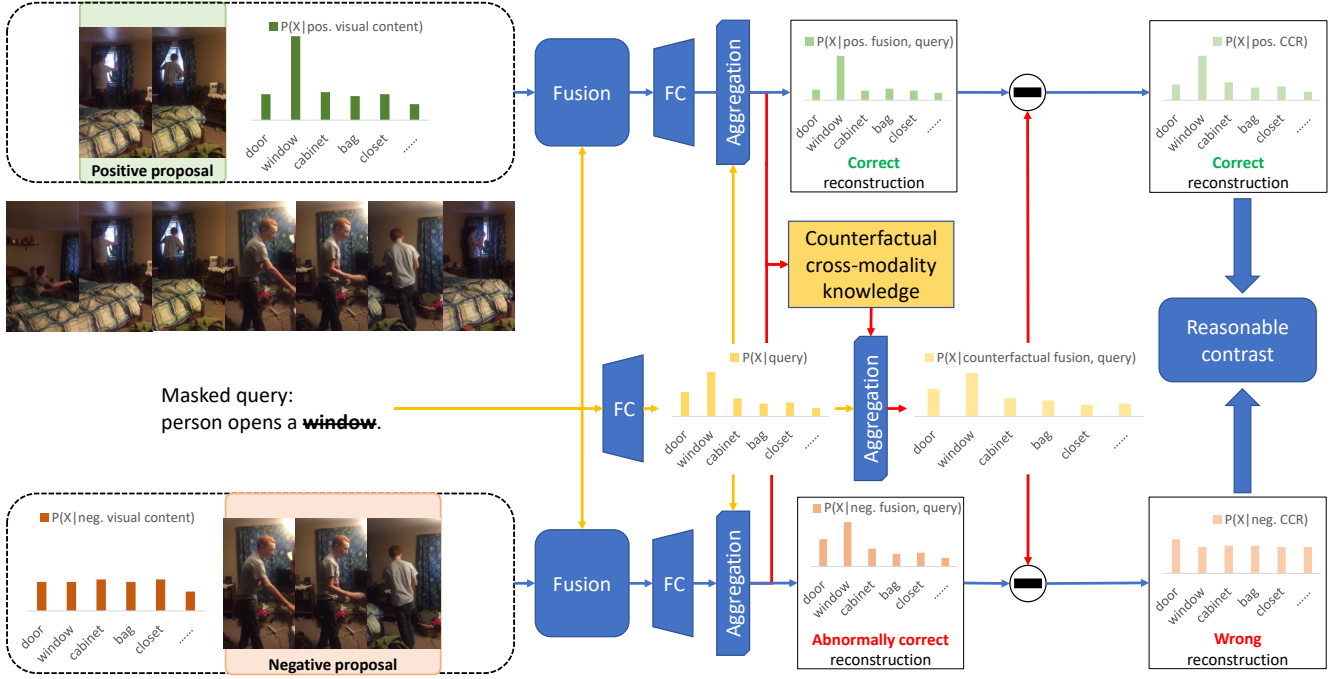


Figure 3: Overview of our proposed Counterfactual Cross-modality Reasoning (CCR) scheme. The main-branches associated with the positive and negative proposals, which are indicated in the blue connections, reconstruct the original query correctly and abnormally correctly, respectively. The reason behind the abnormal reconstruction is the spurious correlation between the masked query and its reconstruction. *e.g.* because of the biased distribution of the co-occurrence of words “open” and “person”, the masked “window” can be easily reconstructed only based on the certain pattern of query even though “window” is not correlated with the visual content in the negative proposal $P(X|neg. visual content)$. This spurious correlation between the masked words “window” and the remaining ones, which is noted as the side-branch in the yellow connections, is modeled as $P(X|query)$, and is aggregated with the counterfactual cross-modality knowledge to obtain the total effect of the masked query $P(X|counterfactual fusion, query)$. Finally, by suppressing this spurious correlation in both the reconstructions of positive and negative proposals, a reasonable contrast can be applied between the rectified prediction $P(X|pos. CCR)$ and $P(X|neg. CCR)$.

this, we propose to decouple the total effect of masked query reconstruction as main-branch and side-branch respectively, and reconstruct the original query by combining these two branches. Given video proposal feature S and masked query feature \bar{q} , we model the main-branch in Equation (1) as

$$P(\hat{W}|F(S, \bar{W})) : \phi_q = \pi(\chi(S, \bar{q})), \quad (5)$$

where ϕ_q is the reconstruction logit produced by cross-modality knowledge, $\chi(\cdot)$ indicates a cross-modality interaction module, and $\pi(\cdot)$ is a fully connected layer, which projects the fusion feature from latent space to word embedding space. Meanwhile, we model the side-branch highlighted in Figure 2 (b) as the impact of the masked query on the reconstruction as

$$P(\hat{W}|\bar{W}) : \psi_q = \pi(\bar{q}), \quad (6)$$

where ψ_q denotes the reconstruction logit only generated by the masked query \bar{q} . Hence we can reconstruct the original query by combining the prediction logits of these two branches as

$$\begin{aligned} P(\hat{W}|F(S, \bar{W}), \bar{W}) : \hat{q} &= \rho(\phi_q, \psi_q) \\ &= \rho(\pi(\chi(S, \bar{q})), \pi(\bar{q})), \end{aligned} \quad (7)$$

where \hat{q} is the final reconstruction logit, and $\rho(\cdot)$ is an aggregation function.

To better isolate the impact of the side-branch on original query reconstruction, we propose a counterfactual approach that cuts off the effect of the main-branch. This ensures that the cross-modality knowledge contributes nothing to the reconstruction of the original query, allowing us to focus solely on the contribution of the side-branch. To create a counterfactual main-branch, we force the cross-modality knowledge to predict the original query randomly, regardless of the input video proposal and masked query. As a result, the final reconstruction obtained by aggregating the predictions generated by cross-modality knowledge and masked query will solely rely on the latter. To achieve this, we modify the cross-modality prediction ϕ_q in Equation (7) into a uniform logit in this counterfactual situation, which is parameterized by a learnable scalar μ [23]. By aggregating the logits produced by μ and masked query, we obtain the counterfactual reconstruction logit \hat{q}^c as

$$P(\hat{W}|F^c(S, \bar{W}), \bar{W}) : \hat{q}^c = \rho(\mu, \pi(\bar{q})), \quad (8)$$

where μ solely impacts the absolute value of the logit \hat{q}^c , but it does not alter the relative value that represents the final reconstruction.

Based on the reconstruction \hat{q} and its counterfactual version \hat{q}^c , which is altered to rely solely on the masked query by incorporating counterfactual cross-modality knowledge into the reconstruction process, we propose to mitigate the spurious correlation between the masked query and its reconstruction by removing the unimodal

effect as

$$\begin{aligned} \hat{W} &= \text{Softmax}(\hat{q} - \hat{q}^c) \\ &= \text{Softmax}\left(\rho(\pi(\chi(S, \bar{q})), \pi(\bar{q})) - \rho(\mu, \pi(\bar{q}))\right). \end{aligned} \quad (9)$$

To prevent a trivial solution of Equation (9), we use a non-linear aggregation function

$$\rho(x, y) = x \odot \text{Sigmoid}(y) \quad (10)$$

to combine the effects of the main-branch and side-branch.

3.4 Training and inference

We embed the proposed CCR into an off-the-shelf query reconstruction based contrastive learning scheme [41]. To begin with, a multi-modal transformer is applied to interact the video with the query as the implementation of $\chi(\cdot)$. Then, for each video-query pair, the transformer generates n positive proposal $S^p = \{S_i^p | i = 1, \dots, N^p\}$ and their corresponding $2n$ intra-video negative proposals $S^n = \{S_j^{n_k} | j = 1 \dots N^p, k = 1, 2\}$, and the whole video is treated as the reference proposal S^r . The diversity of the positive proposals is ensured by a penalization term [19]

$$\ell_{div} = \|\Omega\Omega^\top - \lambda I\|_F^2, \quad (11)$$

where λ is a hyperparameter, and $\Omega = \text{cat}[\omega_1^p; \dots; \omega_{N^p}^p]$ where ω_i^p is the temporal weight of S_i^p [41].

For each triplet $\langle S_i^p, S_i^{n_1}, S_i^{n_2} \rangle$ and S^r , by omitting the subscript i , the reconstruction losses of positive, negative and reference proposals, which are denoted as ℓ_c^p , ℓ_c^r , $\ell_c^{n_1}$, and $\ell_c^{n_2}$ respectively, are rectified through our proposed CCR to minimize the losses of the spurious correlation mitigated reconstruction \hat{W} in Equation (9) and the aggregated logit \hat{q} in Equation (7) corresponding to all the proposals as

$$\begin{aligned} \ell_c^* &= CE(\hat{W}^*, W) + CE(\text{Softmax}(\hat{q}^*), W) \\ * &\in \{p, r, n_1, n_2\}, \end{aligned} \quad (12)$$

where $CE(\cdot)$ is the cross entropy loss. Thus, the counterfactual intra-video contrastive loss [41] for each video-query pair is obtained as

$$\begin{aligned} \ell_c &= \max(0, \alpha_p + \ell_c^p - \ell_c^r) \\ &+ \max(0, \alpha_n + \ell_c^p - \ell_c^{n_1}) + \max(0, \alpha_n + \ell_c^p - \ell_c^{n_2}), \end{aligned} \quad (13)$$

where α_p and α_n are hyperparameters.

Meanwhile, we train the fully connected layer π to minimize the reconstruction error given the masked query as

$$\ell_q = CE(\text{Softmax}(\psi_q), W). \quad (14)$$

Then we optimize the cross-modality module χ and projection layer π with respect to

$$\ell = \ell_c + \ell_q + \ell_{div}. \quad (15)$$

Additionally, μ , which provides the counterfactual cross-modality knowledge, is optimized individually from χ and π as

$$\begin{aligned} \ell_{kl}^* &= KL(\text{Softmax}(\hat{q}^*) | \text{Softmax}(\hat{q}^c)), * \in \{p, r, n_1, n_2\} \\ \ell_{kl} &= \sum_{* \in \{p, r, n_1, n_2\}} \ell_{kl}^*, \end{aligned} \quad (16)$$

for all the proposals to minimize the Kullback-Leibler divergence between \hat{q}^* and its counterfactual version \hat{q}^c to prevent the rectified reconstruction from being dominated by one of them [23].

Algorithm 1: Counterfactual Cross-modality Reasoning (CCR) for each video-query pair in dataset

Data: positive proposal features S^p , negative proposal features S^n , masked query \bar{q} , query W , weight matrix of positive proposals Ω

Result: cross-modality fusion module χ , prediction layer π , uniform logit μ

```

1 while  $S^p \in S^p$  and  $S^{n_1}, S^{n_2} \in S^n$  do
2    $\psi_q \leftarrow \pi(\bar{q})$ ;
3    $\hat{q}^c \leftarrow \rho(\mu, \psi_q)$ ;
4   for  $* \in \{p, n_1, n_2, r\}$  do
5     calculate  $\phi_q^*, \hat{q}^*, \hat{W}^*$  w.r.t. Equation (5), (6), and (9);
6   end
7   if training then
8      $\ell_{div} \leftarrow \|\Omega\Omega^\top - \lambda I\|_F^2$ ;
9      $\ell_q \leftarrow CE(\text{Softmax}(\psi_q), W)$ ;
10    calculate  $\ell_c$  w.r.t. Equation (12) and (13);
11    update  $\chi, \pi$  w.r.t.  $\ell_q, \ell_{div}$  and  $\ell_c$ ;
12    calculate  $\ell_{kl}$  w.r.t. Equation (16);
13    update  $\mu$  w.r.t.  $\ell_{kl}$ ;
14  else
15    return the best proposal in  $S^p$  according to
      vote-strategy [41];
16  end
17 end
```

During inference, given an untrimmed video V and its corresponding query W , the multi-modal fusion module $\chi(\cdot)$ first encode them to generate the set of positive proposals $S^p = \{S_i^p | i = 1, \dots, N^p\}$. Following the experiment setting in [41], a vote-based strategy [42] is utilized to select the best proposal as the output. More specifically, for each positive proposal, we compute its Intersection over Union (IoU) with the other $N_p - 1$ positive proposals, and the sum of IoUs represents the number of votes it receives. Ultimately, we select the positive proposal with the highest number of votes as the final prediction. The overall training and inference procedures are presented in Algorithm 1.

4 EXPERIMENTS

4.1 Implementation Details

In this paper, we measure the effectiveness of moment localization using temporal Intersection over Union (IoU), which is the ratio between the temporal overlap and union of the segment predicted by the model and the ground truth moment. Specifically, we use “ $R@a, mIoU$ ” to evaluate localization performance, which is the average IoU of the a predictions with the lowest reconstruction loss based on the rectified reconstruction $\hat{q}^p - \hat{q}^c$. Additionally, we use “ $R@a, IoU = b$ ” as an evaluation metric to further assess performance, which means there is at least one predicted moment with a temporal IoU larger than b among the top a predictions. We reproduce the CPL [41] as our baseline on one NVIDIA GeForce RTX 3090 GPU, and follow all the hyperparameter settings provided by their official repository [39] to ensure a fair comparison.

Table 1: Comparison of $mIoU$ on Charades-STA and ActivityNet Captions datasets. The best result for each metric is displayed in bold, and the second-best result is marked in red. CPL-R denotes the result reproduced based on the official CPL repository and is used to replace the original CPL result in the top two rankings.

Methods	Charades-STA		ActivityNet Captions	
	$R@1, mIoU$	$R@5, mIoU$	$R@1, mIoU$	$R@5, mIoU$
WS-DEC [6]	-	-	28.23	-
CTF [5]	27.3	-	32.20	-
WSLLN [10]	-	-	32.20	-
WSRA [7]	31.00	-	-	-
VCA [29]	38.49	-	33.15	-
LCNet [31]	38.94	-	34.29	-
CPL [41]	43.48	-	-	-
CPL-R[39]	43.50	67.70	35.71	43.78
Ours	44.66	67.86	36.69	53.37

4.2 Datasets

We evaluate our proposed CCR on two widely used benchmark datasets, Charades-STA [25] and ActivityNet-Captions [1].

Charades-STA. The Charades-STA dataset [25] consists of 16,128 video-query pairs generated by 6,672 videos, with an average video duration of 29.96 seconds. Following [41], we trained our model on the training set, which contains 12,408 video-query pairs, and evaluated the performance on the test set, which contains 3,720 video-query pairs.

ActivityNet-Captions. The ActivityNet-Captions dataset [1] contains 19,209 videos with an average duration of 117.6 seconds. Following [41], we split the dataset into training, validation, and test sets, which contain 37,417, 17,505, and 17,031 video-query pairs, respectively.

4.3 Comparison with state-of-the-arts

We compare the performance of our proposed CCR with state-of-the-art methods on Charades-STA and ActivityNet Captions using $R@a, mIoU$ and $R@a, IoU = b$, where $a \in \{1, 2\}$ and $b \in \{0.1, 0.3, 0.5, 0.7\}$. The results are presented in Table 1, Table 2, and Table 3, respectively. Because there are significant differences between the performances of our reproduction and CPL [41], we additionally include the results of our reproduction as CPL-R for comparison. Directly comparing CRM [14] with other methods, including ours, is unfair because CRM requires multiple queries that appear sequentially in the video for training, and hence we have not highlighted its results in the top two rankings.

We compare the average temporal IoU between our proposed CCR and the existing methods in Table 1. Our proposed CCR outperforms the current state-of-the-art method CPL in all evaluation metrics, demonstrating a remarkable improvement of over 9% in the $R@5, mIoU$ metric on the ActivityNet Captions dataset. In Table 2, our CCR overall surpasses both CPL and CPL-R according to $R@1$ metrics, with an average absolute gain of about 2%. For $R@5$ metrics, our CCR outperformed CPL overall. On ActivityNet Captions dataset, we surpass the baseline and outperform it by approximately 3% and 9% on average for $R@1$ and $R@5$ metrics,

Table 2: Performance Comparison on Charades-STA. The best result for each metric is displayed in bold, and the second-best result is marked in red. CPL-R denotes the result reproduced based on the official CPL repository and is used to replace the original CPL result in the top two rankings. The results of CRM (indicated as CRM^\dagger) are not included in the top two rankings because it needs paragraph-video annotations during training.

Methods	$R@1, IoU =$			$R@5, IoU =$		
	0.3	0.5	0.7	0.3	0.5	0.7
TGA [22]	32.14	19.94	8.84	86.58	65.52	33.51
SCN [20]	42.96	23.58	9.97	95.56	71.8	38.87
CTF [5]	39.8	27.3	12.9	-	-	-
WSTAN [28]	43.39	29.35	12.28	93.04	76.13	41.53
BAR [30]	44.97	27.04	12.23	-	-	-
WSRA [7]	50.13	31.20	11.01	86.75	70.50	39.02
VLANet [21]	45.24	31.83	14.17	95.7	82.85	33.09
LoGAN [27]	48.04	31.74	13.71	89.01	72.17	37.58
MARN [26]	48.55	31.94	14.81	90.70	70.00	37.40
CCL [38]	-	33.21	15.68	-	73.50	41.87
CRM^\dagger [14]	53.66	34.76	16.37	-	-	-
CNM[40]	60.39	35.43	15.45	-	-	-
LCNet[31]	59.60	39.19	18.87	94.78	80.56	45.24
RTBPN[37]	60.04	32.36	13.24	97.48	71.85	41.18
VCA[29]	58.58	38.13	19.57	98.08	78.75	37.75
CPL[41]	65.99	49.05	22.61	96.99	84.71	52.37
CPL-R[39]	66.53	49.43	22.36	96.80	84.20	52.18
Ours	68.59	50.79	23.75	96.85	84.48	52.44

respectively. The significant gain achieved by CCR on the ActivityNet Captions dataset is due to the more variational visual content, which increases the frequency of negative proposals producing abnormal correct reconstructions as illustrated in Figure 1 and Figure 3. Our CCR is designed to mitigate this issue, leading to its superior performance on this dataset. Our CCR achieves comparable performance to other methods, and on average performs better than LCNet and VCA on $mIoU$ in Table 1, despite their state-of-the-art performance on $R@5, IoU = 0.3, 0.5$ metrics.

4.4 Ablation Studies

Generation of counterfactual cross-modality knowledge. The only additional parameter in our proposed method compared to the baseline is the counterfactual cross-modality knowledge μ . In addition to the uniform distribution presented in Section 3, we also explore two other possible generation methods for U , referred to as “Average” and “Random selected”, and evaluate their effectiveness on the Charades-STA dataset. As illustrated in Table 5, our experiments reveal that the replacement within mini-batch is insufficient to generate counterfactual cross-modality knowledge in this scenario. In this situation, the uniform prediction strategy outperforms the other two methods for generating μ .

Aggregation of main-branch and side-branch. As discussed in Section 3, we non-linearly aggregate the effects of main-branch and side-branch as

$$\rho(\phi_q, \psi_q) = \phi_q \odot \text{Sigmoid}(\psi_q), \quad (17)$$

Table 3: Performance comparison on ActivityNet-Captions. The best result for each metric is displayed in bold, and the second-best result is marked in red. CPL-R denotes the result reproduced based on the official CPL repository and is used to replace the original CPL result in the top two rankings. The results of CRM (indicated as CRM^\dagger) are not included in the top two rankings because it needs paragraph-video annotations during training.

Methods	$R@1, IoU =$			$R@5, IoU =$		
	0.1	0.3	0.5	0.1	0.3	0.5
CTF [5]	74.2	44.3	23.6	-	-	-
EC-SL [4]	68.48	44.29	24.16	-	-	-
MARN [26]	-	47.01	29.95	-	72.02	57.49
SCN [20]	71.48	47.23	29.22	90.88	71.56	55.69
BAR [30]	-	49.03	30.73	-	-	-
RTBPN [37]	73.73	49.77	29.63	93.89	79.89	60.56
CCL[38]	-	50.12	31.07	-	77.36	61.29
WSTAN[28]	79.78	52.45	30.01	93.15	79.38	63.42
CRM^\dagger [14]	81.61	55.26	32.19	-	-	-
CNM[40]	78.13	55.68	33.33	-	-	-
VCA [29]	67.96	50.45	31.00	92.14	71.79	53.83
LCNet[31]	78.58	48.49	26.33	93.95	82.51	62.66
CPL[41]	82.55	55.73	31.37	87.24	63.05	43.13
CPL-R[39]	78.13	51.19	28.19	88.23	62.16	40.04
Ours	80.32	53.21	30.39	91.44	71.97	56.50

Table 4: Ablation study on different manners of generating counterfactual cross-modality knowledge. The best result is indicated in bold. “Average” means that μ is set as the average prediction of main-branch in a mini-batch, and “Random selected” denotes that μ is randomly selected from the predictions within a mini-batch.

Counterfactual cross-modality knowledge μ	$R@1$			
	$mIoU$	$IoU = 0.3$	$IoU = 0.5$	$IoU = 0.7$
Baseline	43.50	66.53	49.43	22.36
Average	43.69	66.14	49.02	22.95
Random selected	44.01	67.51	49.87	23.02
Uniform	44.66	68.59	50.79	23.75

which allows us to obtain the total impact on query reconstruction. We also implemented ρ as another heuristic non-linear summation and learnable projection network to evaluate their corresponding performance, as presented in Table 4. However, both the learning-based and non-linear summation methods achieved lower performance compared to the aggregation applied in Equation (17).

Qualitative results. We provide a qualitative example in Figure 4 to further illustrate the effectiveness of our proposed CCR. In this video, a person first places a laptop on a table, as described in the query, then lies on a sofa, and finally watches TV while holding a remote. The original reconstruction generated by the positive proposal is the same as that of negative proposals, which includes “a laptop”. Therefore, the contrast between positive and negative proposals is invalid. Our proposed CCR rectifies the reconstructions to predict the wrong answer “used on” for the masked words,

Table 5: Ablation study on different aggregation manners of the main-branch and side-branch. The best result is indicated in bold. $Proj(cat[x; y])$ presents we first concatenate x and y along the embedding dimension, then utilize a learnable linear layer to project it back to the original space.

Branch aggregation $\rho(x, y)$	$R@1$			
	$mIoU$	$IoU = 0.3$	$IoU = 0.5$	$IoU = 0.7$
Baseline	43.50	66.53	49.43	22.36
$Proj(cat[x; y])$	44.30	67.89	50.76	23.65
$Sigmoid(x + y)$	44.21	68.13	49.22	22.37
$x \odot Sigmoid(y)$	44.66	68.59	50.79	23.75

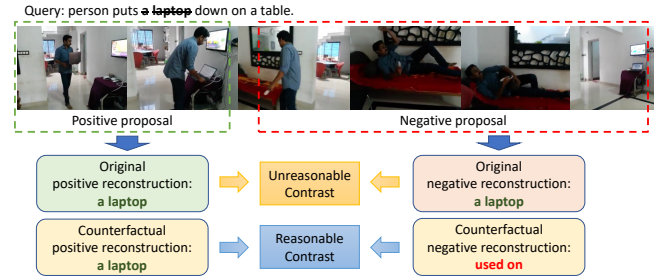


Figure 4: A qualitative example that highlights the contrasts between reconstructions with and without CCR. Even though there’s no visual content pertaining to the “laptop” masked in the query of the negative proposal, the words “a laptop” were still successfully reconstructed, rendering the contrast between it and the positive proposal invalid. Utilizing our proposed CCR, the reconstructions are rectified to produce correct and incorrect outcomes for positive and negative proposals, respectively.

enabling reasonable contrastive learning and further improving the cross-modality alignment.

5 CONCLUSION

In this paper, we introduce a novel Counterfactual Cross-modality Reasoning (CCR) method, which addresses the challenge of weakly supervised video moment localization. We focus on the problem of unreasonable contrastive learning, which arises due to the spurious correlation between masked and unmasked query words. This issue is commonly overlooked by current state-of-the-art query reconstruction based methods.

To overcome this problem, we first model the impact on query reconstruction as a combination of cross-modality driven main-branch and query-driven side-branch. We then extract the spurious correlation induced by the unimodal impact by applying counterfactual cross-modality knowledge during the aggregation process. Finally, we address the problem of spurious correlation by removing it from the reconstructions of positive and negative proposals, enabling reasonable contrastive learning.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China No. 61976206 and No. 61832017, Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, Beijing Academy of Artificial Intelligence (BAAI), the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China 21XNLG05, and Public Computing Cloud, Renmin University of China.

REFERENCES

- [1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Nibbles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 961–970.
- [2] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems* 32 (2019).
- [3] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. 2018. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1130–1139.
- [4] Shaoxiang Chen and Yu-Gang Jiang. 2021. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8425–8435.
- [5] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. 2020. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308* (2020).
- [6] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. *Advances in Neural Information Processing Systems* 31 (2018).
- [7] Zhiyuan Fang, Shu Kong, Zhe Wang, Charles Fowlkes, and Yezhou Yang. 2020. Weak supervision and referring attention for temporal-textual association learning. *arXiv preprint arXiv:2006.11747* (2020).
- [8] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*. 5267–5275.
- [9] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. 2017. Turn tap: Temporal unit regression network for temporal action proposals. In *Proceedings of the IEEE International Conference on Computer Vision*. 3628–3636.
- [10] Mingfei Gao, Richard Socher, and Caiming Xiong. 2020. Weakly Supervised Natural Language Localization Networks. US Patent App. 16/531,343.
- [11] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.
- [12] Zhaoyu Guo, Zhou Zhao, Wei Jin, Wang Dazhou, Liu Ruitao, and Jun Yu. 2021. TaoHighlight: Commodity-Aware Multi-modal Video Highlight Detection in E-Commerce. *IEEE Transactions on Multimedia* (2021).
- [13] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing Moments in Video with Temporal Language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- [14] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. 2021. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7199–7208.
- [15] Xiaohan Lan, Yitian Yuan, Xin Wang, Zhi Wang, and Wenwu Zhu. 2023. A survey on temporal sentence grounding in videos. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 2 (2023), 1–33.
- [16] Kun Li, Dan Guo, and Meng Wang. 2021. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1902–1910.
- [17] Siyang Li, Xiangxin Zhu, Qin Huang, Hao Xu, and C-C Jay Kuo. 2017. Multiple instance curriculum learning for weakly supervised object detection. *arXiv preprint arXiv:1711.09191* (2017).
- [18] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. 2021. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3320–3329.
- [19] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).
- [20] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11539–11546.
- [21] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. 2020. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII* 16. Springer, 156–171.
- [22] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. 2019. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11592–11601.
- [23] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12700–12710.
- [24] Donald B Rubin. 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics* (1978), 34–58.
- [25] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*. Springer, 510–526.
- [26] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. 2020. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048* (2020).
- [27] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. 2021. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2083–2092.
- [28] Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li. 2021. Weakly supervised temporal adjacent network for language grounding. *IEEE Transactions on Multimedia* 24 (2021), 3276–3286.
- [29] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. 2021. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1459–1468.
- [30] Jie Wu, Guanlin Li, Xiaoguang Han, and Liang Lin. 2020. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1283–1291.
- [31] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. 2021. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE Transactions on Image Processing* 30 (2021), 3252–3262.
- [32] Qinghao Ye, Xiyue Shen, Yuan Gao, Zirui Wang, Qi Bi, Ping Li, and Guang Yang. 2021. Temporal Cue Guided Video Highlight Detection With Low-Rank Audio-Visual Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7950–7959.
- [33] Sunjae Yoon, Ji Woo Hong, Eunseop Yoon, Dahyun Kim, Junyeong Kim, Hee Suk Yoon, and Chang D Yoo. 2022. Selective Query-Guided Debiasing for Video Corpus Moment Retrieval. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*. Springer, 185–200.
- [34] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems* 32 (2019).
- [35] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based Localizing Network for Natural Language Video Localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6543–6554.
- [36] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12870–12877.
- [37] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jiemin Zhu, and Xiuqiang He. 2020. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4098–4106.
- [38] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems* 33 (2020), 18123–18134.
- [39] Minghang Zheng. [n.d.]. *CPL: Weakly Supervised Temporal Sentence Grounding with Gaussian-based Contrastive Proposal Learning*. <https://github.com/minghangz/cpl> (2023, Feb 20).
- [40] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. 2022. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3517–3525.
- [41] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. 2022. Weakly Supervised Temporal Sentence Grounding with Gaussian-based Contrastive Proposal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15555–15564.
- [42] Zhi-Hua Zhou and Zhi-Hua Zhou. 2021. *Ensemble learning*. Springer.