Causality-aware Visual Scene Discovery for Cross-Modal Question Reasoning

Yang Liu
Sun-Yat-Sen University
liuy856@mail.sysu.edu.cn

Guanbin Li
Sun-Yat-Sen University
liguanbin@mail.sysu.edu.cn

Liang Lin
Sun-Yat-Sen University
linliang@ieee.org

Abstract

Existing visual question reasoning methods usually fail to explicitly discover the inherent causal mechanism and ignore the complex event-level understanding that requires jointly modeling cross-modal event temporality and causality. In this paper, we propose an event-level visual question reasoning framework named Cross-Modal Question Reasoning (CMQR), to explicitly discover temporal causal structure and mitigate visual spurious correlation by causal intervention. To explicitly discover visual causal structure, the Visual Causality Discovery (VCD) architecture is proposed to find question-critical scene temporally and disentangle the visual spurious correlations by attention-based front-door causal intervention module named Local-Global Causal Attention Module (LGCAM). To align the fine-grained interactions between linguistic semantics and spatial-temporal representations, we build an Interactive Visual-Linguistic Transformer (IVLT) that builds the multi-modal co-occurrence interactions between visual and linguistic content. Extensive experiments on four datasets demonstrate the superiority of CMQR for discovering visual causal structures and achieving robust question reasoning.

1. Introduction

Event understanding [33, 37, 75] has become a prominent research topic in video analysis because videos [44, 45, 46] have good potential to go beyond image-level understanding (scenes, people, objects, activities, etc.) to understand event temporality and causality. Accurate and efficient cognition and reasoning over complex events is extremely important in video-language understanding. Since the expressivity of natural language can potentially describe a richer event space [7] that facilitates the deeper event understanding, in this paper, we focus on complex (temporal, causal) event-level visual question reasoning task, which aims to fully understand richer multi-modal event space and answer the given question in a causality-aware way. To achieve event-level visual question reasoning [12, 3],

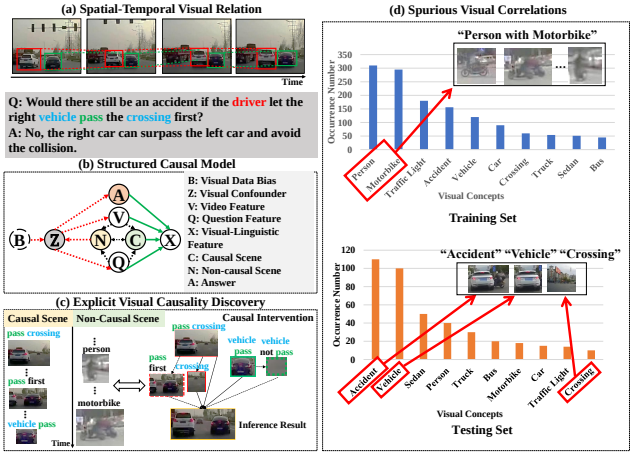


Figure 1. An example of event-level counterfactual visual question reasoning task and its structured causal model (SCM). The counterfactual inference is to obtain the outcome of certain hypothesis that do not occur. The SCM shows how the confounder induces the spurious correlation. The green path is the unbiased visual question reasoning. The red path is the biased one.

the model is required to achieve fine-grained understanding of video and language content involving various complex relations such as spatial-temporal visual relation, linguistic semantic relation, and visual-linguistic causal dependency. Most of the existing visual question reasoning methods [38, 35, 55] use recurrent neural networks (RNNs) [60], attention mechanisms [62] or Graph Convolutional Networks [32] for relation reasoning between visual and linguistic modalities. Although achieving promising results, the current visual question reasoning methods suffer from the following two common limitations.

First, the existing visual question reasoning methods usually focus on relatively simple events where temporal understanding and causality discovery are simply not required to perform well, and ignore more complex and challenging events that require in-depth understanding of the causality, spatial-temporal dynamics, and linguistic relations. As shown in Fig. 1 (a), the event-level counterfactual visual question reasoning task requires the outcome of

certain hypothesis that does not occur (e.g. "the driver let the right vehicle pass the crossing first") in the given video. If we just simply correlate relevant visual contents, we cannot to get the right inference result without discovering the hidden spatial-temporal and causal dependencies. To accurately reason about the imagined events under the counterfactual condition, the model is required to not only conduct relational reasoning in a hierarchical way but also fully explore the causality, logic, and spatial-temporal dynamic structures of the visual and linguistic content. However, the multi-level interaction and causal relations between the language and spatial-temporal structure of the complex multi-modal events is not fully explored in current methods.

Second, most of the visual question reasoning models tend to capture the spurious visual correlations rather than the true causal structure, which leads to an unreliable reasoning process [53, 67, 39, 47]. As shown in the SCM from Fig. 1 (b), we can consider some frequently appearing visual concepts as the visual confounders. The "visual bias" denotes the strong correlations between visual features and answers. For example, the concepts "person" and "motorbike" are dominant in training set (Fig. 1 (c) and (d)) and thus the predictor may learn the spurious correlation between the "person" with the "motorbike" without looking at the collision region (causal scene C) to reason how actually the accident happens. As a result, it limits the reasoning ability of visual-linguistic question reasoning models if they memorize the strong visual prior. Taking a causal look at VideoQA, we partition the visual scenes into two parts: 1) causal scene C, which holds the question-critical information, 2) non-causal scene N, which is irrelevant to the answer. Thus, we scrutinize that the non-causal scene N is also spuriously correlated with the answer A.

Such biased dataset entails two causal effects: the visual bias B and non-causal scene N lead to the confounder Z, and then affects the visual feature V, causal scene C, question feature Q, visual-linguistic feature X, and the answer A. Here, we can draw two causal links to describe these causal effects: $Z \rightarrow \{V, Q\} \rightarrow C \rightarrow X$ and $Z \rightarrow A$. If we want to learn the true causal effect $\{V, Q\} \rightarrow C \rightarrow X \rightarrow A$ while employing the biased dataset to train this model (Fig. 1 (d)), this model may simply correlate the concepts "person" and "motorbike", i.e., through $Z \rightarrow \{V, Q\} \rightarrow X$, and then use this biased knowledge to infer the answer, i.e., through $Z \rightarrow A$. In this way, this model learns the spurious correlation between $\{V, Q\}$ and A through the backdoor path $A \leftarrow Z \rightarrow \{V, Q\} \rightarrow X$ induced by the confounder Z, as shown in Fig. 2 (b). Due to the existence of unobservable visual confounders and complex visual-linguistic interaction, the model learns the spurious correlation without exploiting the true question intention and dominant visual evidence to achieve robust question reasoning.

To address the aforementioned limitations, we propose an event-level visual question reasoning framework named Cross-Modal Question Reasoning (CMQR). To explicitly uncover the visual causal structure, we propose a Visual Causality Discovery (VCD) architecture that learns to find the temporal causal scenes for the given question semantics and mitigates the unobservable visual spurious correlations by an attention-based causal front-door intervention module named Local-Global Causal Attention Module (LGCAM). To align the multi-modal interaction between the appearance-motion and language representations, we build an Interactive Visual-Linguistic Transformer (IVLT). Experiments on SUTD-TrafficQA, TGIF-QA, MSVD-QA, and MSRVTT-QA datasets show the advantages of our CMQR over the state-of-the-art methods. The main contributions can be summarized as follows:

- We propose a causality-aware event-level visual question reasoning framework named Cross-Modal Question Reasoning (CMQR), to discover cross-modal causal structures via causal interventions and achieve robust visual question reasoning and answering.
- We introduce the Visual Causality Discovery (VCD) architecture that learns to find the temporal causal scenes for a given question and mitigates the unobservable visual spurious correlations by an attention-based causal front-door intervention module named Local-Global Causal Attention Module (LGCAM).
- We construct an Interactive Visual-Linguistic Transformer (IVLT) to align and discover the multi-modal co-occurrence interactions between linguistic semantics and spatial-temporal visual concepts.

2. Related Works

2.1. Visual Question Reasoning

Compared with the image-based visual question reasoning [4, 74, 2], event-level visual question reasoning (VideoQA) is much more challenging due to the existence of temporal dimension. To accomplish this task, the model needs to capture spatial-temporal and visual-linguistic relations. To explore relational reasoning, Xu et al. [69] proposed an attention mechanism to exploit the appearance and motion knowledge with the question as a guidance. Later on, some hierarchical attention and co-attention based methods [38, 14, 28, 35, 27, 22, 36, 41] are proposed to learn appearance-motion and question-related multi-modal interactions. Le et al. [35] proposed hierarchical conditional relation network (HCRN) to construct sophisticated structures for representation and reasoning over videos. Lei et al. [36] employed sparse sampling to build a transformer-based model named CLIPBERT and achieve video-and-language understanding. However, previous works tend

to implicitly capture the spurious visual-linguistic correlations, while we build a Visual Causality Discovery (VCD) architecture to explicitly uncover the visual causal structure.

2.2. Relational Reasoning for Event Understanding

Besides VideoQA, relational reasoning has been explored in other event understanding tasks [43, 42, 18, 23, 50]. For example, Pan et al. [54] designed a high-order actor-context-actor relation network to realize indirect relation reasoning for spatial-temporal action localization. To localize a moment from videos for a given query, Nan et al. [51] introduced a dual contrastive learning approach to align the text and video by maximizing the mutual information between semantics and video clips. Wang et al. [68] learned the deconfounded object-relevant association for robust video object grounding. However, these methods only perform relational reasoning over visual modality and neglects the potential causal structures from linguistic semantic relation, resulting in incomplete understanding of visual-linguistic content. Additionally, our CMQR conducts causality-aware spatial-temporal relational reasoning to uncover the causal structure for visual-linguistic modality, and utilizes hierarchical semantic knowledge.

2.3. Causal Inference in Visual Learning

Compared with conventional debiasing techniques [66], causal inference [56, 58, 72, 47] shows its potential in mitigating the spurious correlations [5] and disentangling the desired model effects [6] for better generalization. Counterfactual and causal inference have attracted increasing attention in visual explanations [19, 21, 64], scene graph generation [9, 61], image recognition [65, 67], video analysis [15, 30, 51], and vision-language tasks [8, 34, 1, 53, 73, 39, 47, 10]. However, most of the existing visual tasks are relatively simple. Although some recent works CVL [1], Counterfactual VQA [53], CATT [73], and IGV [39] focused on visual question reasoning tasks, they adopted structured causal model (SCM) to eliminate either the linguistic or visual bias without considering explicit cross-modal causality discovery. Differently, our CMQR aims for event-level visual question reasoning that requires fine-grained understanding of spatial-temporal and visual-linguistic causal dependency. Moreover, our Visual Causality Discovery (VCD) explicitly finds question-critic visual scene and applies front-door causal interventions to discover visual causal structure.

3. Methodology

The CMQR is an event-level visual question reasoning architecture, as shown in Fig. 2. In this section, we present the detailed implementations of CMQR.

3.1. Visual Representation Learning

The goal of event-level visual question reasoning is to deduce an answer $\tilde{a}$ from a video $\mathcal{V}$ with a given question q. The video $\mathcal{V}$ of L frames is divided into N equal clips. Each clip of C_i of length $T = \lfloor L/N \rfloor$ is presented by two types of visual features: frame-wise appearance feature vectors $F_i^a = \{f_{i,j}^a | f_{i,j}^a, j = 1, \ldots, T\}$ and motion feature vector at clip level f_i^m. In our experiments, the vision-language transformer with frozen parameters XCLIP [52] (other visual backbones are evaluated in Table 7) is used to extract the frame-level appearance features F^a and the clip-level motion features F^m. Then, we use a linear layer to map F^a and F^m into the same d-dimensional feature space.

3.2. Linguistic Representation Learning

Each word of the question is respectively embedded into a vector of 300 dimension by Glove [57] word embedding, which is further mapped into a d-dimensional space using linear transformation. Then, we represent the corresponding question and answer semantics as $Q = \{q_1, q_2, \cdots, q_L\}$, $A = \{a_1, a_2, \cdots, a_{L_a}\}$, where L, L_a indicate the length of Q and A, respectively. To obtain contextual linguistic representations that aggregates dynamic long-range temporal dependencies from multiple time-steps, a Bert [13] model is employed to encode Q and the answer A, respectively. Finally, the updated representations for the question and answer candidates can be written as:

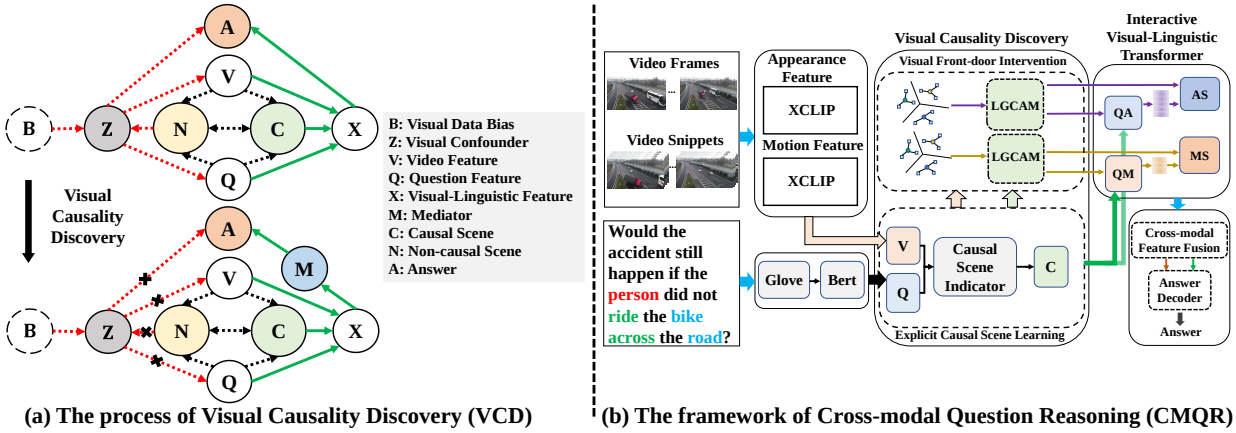
$$\begin{aligned} Q &= \{q_i | q_i \in \mathbb{R}^d\}_{i=1}^{L} \\ A &= \{a_i | a_i \in \mathbb{R}^d\}_{i=1}^{L_a} \end{aligned} \tag{1}$$

3.3. Visual Causality Discovery

For visual-linguistic question reasoning, we employ Pearl's structural causal model (SCM) [56] to model the causal effect between video-question pairs (V, Q), causal scene C, non-causal scene N, and the answer A, as shown in Fig. 2 (a). We hope to train a video question answering model to the learn the true causal effect $\{V, Q\} \rightarrow C \rightarrow X \rightarrow A$: the model should reason the answer A from video feature V, causal-scene C and question feature Q instead of exploiting the non-causal scene N and spurious correlations induced by the confounders Z (i.e., the existence of non-causal scene and overexploiting the co-occurrence between visual concepts and answer). In our SCM, the non-interventional prediction can be expressed by Bayes rule:

$$P(A|V, Q) = \sum_z P(A|V, Q, z) P(z|V, Q) \tag{2}$$

However, the above objective learns not only the main direct correlation from $\{V, Q\} \rightarrow X \rightarrow A$ but also the spurious one from the back-door path $\{V, Q\} \leftarrow Z \rightarrow A$.



(a) The process of Visual Causality Discovery (VCD) **(b) The framework of Cross-modal Question Reasoning (CMQR)**

Figure 2. The process of Visual Causality Discovery (VCD) ((a)) and framework of Cross-modal Question Reasoning ((b)). The VCD explicitly finds question-critic visual scene and applies front-door causal intervention to discover visual causal structure. The Interactive Visual-Linguistic Transformer (IVLT) models the interaction between the appearance-motion and language knowledge. The green path is the unbiased question reasoning. The red path is the biased one. The black path is the division of causal and non-causal visual scenes.

An intervention on $\{V, Q\}$ is denoted as $do(V, Q)$, which cuts off the link $\{V, Q\} \leftarrow Z$ to block the back-door path $\{V, Q\} \leftarrow Z \rightarrow A$ and eliminate the spurious correlation. In this way, $\{V, Q\}$ and A are deconfounded and the model can learn the true causal effect $\{V, Q\} \rightarrow C \rightarrow X \rightarrow A$.

3.3.1 Explicit Causal Scene Learning

Inspired by the fact that only part of the visual scenes are critical to answering the question, we split the video V into causal scene C and non-causal scene N (see the black path in Fig. 2 (a)). Specifically, given the causal scene C and question Q, we assume that the answer A is determined, regardless the variations of the non-causal scene N: $A \perp N|C, Q$, where $\perp$ denotes the probabilistic independence. Therefore, we build an explicit causal scene learning (ECSL) module to estimate the value of C.

For a video-question pair (v, q), we encode video instance v as a sequence of K visual clips. The ECSL module aims to estimate the causal scene $\hat{c}$ according to the question q. Concretely, we first construct a cross-modal attention module to indicate the probability of each visual clip belongs to causal scene ($p_{\hat{c}} \in \mathbb{R}^K$):

$$p_{\hat{c}} = \text{softmax}(G_v^1(v) \cdot G_q^1(q)^\top) \tag{3}$$

where G_v^1 and G_q^1 are fully connected layers to align cross-modal representations. However, the soft mask makes $\hat{c}$ overlap. To achieve a differentiable selection on attentive probabilities and compute the selector vector $S \in \mathbb{R}^K$ on the attention score over each clip (i.e., $p_{\hat{c},i}, i \in K$), we employ Gumbel-Softmax [24] and estimate $\hat{c}$ as:

$$\hat{c} = \text{Gumbel-Softmax}(p_{\hat{c},i}) \cdot v \tag{4}$$

For a video-question pair (v, q), we obtain the original video v and causal scene $\hat{c}$. According to Eq. (2), we pair original video v and causal scene $\hat{c}$ with q to synthesizes two instances: (v, q) and $(\hat{c}, q)$. Then, we feed these two instances into visual front-door causal intervention (VFCI) module to deconfound $\{V, Q\}$ and A.

3.3.2 Visual Front-door Causal Intervention

In visual domains, it is hard to explicitly represent confounders due to complex data biases. Fortunately, the front-door adjustment give a feasible way to calculate $P(A|do(V), Q)$. In Fig. 2 (a), an additional mediator M can be inserted between X and A to construct the front-door path $\{V, Q\} \rightarrow X \rightarrow M \rightarrow A$ for transmitting knowledge. For visual question reasoning, an attention-based model $P(A|V, Q) = \sum_m P(M = m|V, Q)P(A|M = m)$ will select a few regions from the original video V and causal scene C based on the question Q to predict the answer A, where m denotes the selected knowledge from M. Thus, the answer predictor can be represented by two parts: two feature extractors $V \rightarrow X \rightarrow M$, $C \rightarrow X \rightarrow M$, and an answer predictor $M \rightarrow A$. In the following, we take the visual interventional probability $P(A|do(V), Q)$ for original video V as an example (the $P(A|do(C), Q)$ for causal scene C is implemented in the same way):

$$\begin{aligned} &P(A|do(V), Q) = \\ &= \sum_m P(M = m|do(V), Q)P(A|do(M = m)) \\ &= \sum_m P(M = m|V, Q) \sum_v P(V = v)P(A|V = v, M = m) \end{aligned} \tag{5}$$

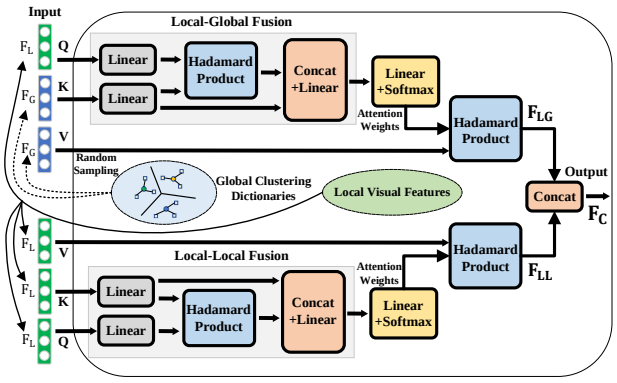


Figure 3. The Local-Global Causal Attention Module (LGCAM).

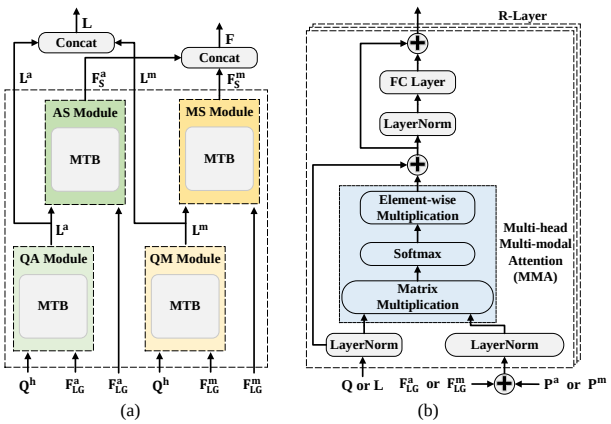


Figure 4. Illustration of the (a) Interactive Visual-Linguistic Transformer (IVLT), and (b) Multi-modal Transformer Block (MTB).

To implement visual front-door causal intervention Eq. (5) in a deep learning framework, we parameterize the $P(A|V,M)$ as a network $g(\cdot)$ followed by a softmax layer since most of visual-linguistic tasks are classification formulations, and then apply Normalized Weighted Geometric Mean (NWGM) [70] to reduce computational cost:

$$
\begin{aligned}
&P(A|do(V),Q) \approx \text{Softmax}[g(\hat{M},\hat{V})] \\
&= \text{Softmax}\Big[g(\sum_{m} P(M=m|f(V))m, \sum_{v} P(V=v|h(V))v)\Big]
\end{aligned} \tag{6}
$$

where $\hat{M}$ and $\hat{V}$ denote the estimations of M and V, $h(\cdot)$ and $f(\cdot)$ denote the network mappings. The derivation details from Eq. (5)-(6) is given in the Appendix B.

Actually, $\hat{M}$ is essentially an in-sample sampling process where m denotes the selected knowledge from the current input sample V, $\hat{V}$ is essentially a cross-sample sampling process since it comes from the other samples. Therefore, both $\hat{M}$ and $\hat{V}$ can be calculated by attention networks [73].

Therefore, we propose a Local-Global Causal Attention Module (LGCAM) that jointly estimates $\hat{M}$ and $\hat{V}$ to increase the representation ability of the causality-aware visual features. $\hat{M}$ can be learned by local-local visual feature F_{LL}, $\hat{V}$ can be learned by local-global visual feature F_{LG}. Here, we take the computation of F_{LG} as the example to clarify our LGCAM, as shown in the upper part of Fig. 3. Specifically, we firstly calculate $F_L = f(V)$ and $F_G = h(V)$ and use them as the input, where $f(\cdot)$ denotes the visual feature extractor (frame-wise appearance feature or motion feature) followed by a query embedding function, and $h(\cdot)$ denotes the K-means based visual feature selector from the whole training samples followed by a query embedding function. Thus, F_L represents the visual feature of the current input sample (local visual feature) and F_G represents the global visual feature. The F_G is obtained by randomly sampling from the whole clustering dictionaries with the same size as F_L. The LGCAM takes F_L and F_G as the inputs and computes local-global visual feature F_{LG} by conditioning global visual feature F_G to the local visual feature F_L. The output of the LGCAM is denoted as F_{LG}:

$$
\begin{aligned}
&\textbf{Input} : Q = F_L, K = F_G, V = F_G \\
&\textbf{Local-Global} : H = [W_V V, W_Q Q \odot W_K K] \\
&\textbf{Activation Mapping} : H' = \text{GELU}(W_H H + b_H) \\
&\textbf{Attention Weights} : \alpha = \text{Softmax}(W_{H'} H' + b_{H'}) \\
&\textbf{Output} : F_{LG} = \alpha \odot F_G
\end{aligned} \tag{7}
$$

where $[.,.]$ denotes concatenation operation, $\odot$ is the Hadamard product, W_Q, W_K, W_V, $W_{H'}$ denote the weights of linear layers, b_H and $b_{H'}$ denote the biases of linear layers. From Fig. 2 (b), the visual front-door causal intervention module has two branches for appearance and motion features. Therefore, the F_{LG} has two variants, one for appearance branch F^a_{LG}, and the other for motion branch F^m_{LG}. The F_{LL} can be computed similarly as F_{LG} when setting $Q = K = V = F_L$. Finally, the F_{LG} and F_{LL} are concatenated $F_C = [F_{LG}, F_{LL}]$ for estimating $P(A|do(V),Q)$.

3.4. Interactive Visual-Linguistic Transformer

To align the fine-grained interactions between linguistic semantics and spatial-temporal representations, we build an Interactive Visual-Linguistic Transformer (IVLT) that contains four sub-modules, namely Question-Appearance (QA), Question-Motion (QM), Appearance-Semantics (AS) and Motion-Semantics (MS), as shown in Fig. 4 (a). The QA (QM) module consists of an R-layer Multi-modal Transformer Block (MTB) (Fig. 4 (b)) for multi-modal interaction between the question and the appearance (motion) features. Similarly, the AS (MS) uses the MTB to infer the appearance (motion) information given the questions.

For QA and QM modules, the input of MTB is question representation Q obtained from section 3.2 and causality-aware visual representations F_C^a, F_C^m obtained from section 3.3.2, respectively. To maintain the positional information of the video sequence, the appearance feature F_C^a and motion feature F_C^m are firstly added with the learned positional embeddings P^a and P^m, respectively. Thus, for $r = 1, 2, \ldots, R$ layers of the MTB, with the input $F_C^a = [F_C^a, P^a]$, $F_C^m = [F_C^m, P^m]$, Q^a, and Q^m, the multi-modal output for QA and QM are computed as:

$$\begin{aligned}
\hat{Q}_r^a &= U_r^a + \sigma^a(\text{LN}(U_r^a)) \\
\hat{Q}_r^m &= U_r^m + \sigma^m(\text{LN}(U_r^m)) \\
U_r^a &= \text{LN}(\hat{Q}_{r-1}^a) + \text{MMA}^a(\hat{Q}_{r-1}^a, F_C^a) \\
U_r^m &= \text{LN}(\hat{Q}_{r-1}^m) + \text{MMA}^m(\hat{Q}_{r-1}^m, F_C^m)
\end{aligned} \tag{8}$$

where $\hat{Q}_0^a = Q^h, \hat{Q}_0^m = Q^h$, U_r^a and U_r^m are the intermediate feature at r-th layer of the MTB. $\text{LN}(\cdot)$ denotes the layer normalization operation and $\sigma^a(\cdot)$ and $\sigma^m(\cdot)$ denote the two-layer linear projections with GELU activation. $\text{MMA}(\cdot)$ is the Multi-head Multi-modal Attention layer. We denote the output semantics-aware appearance and motion features of QA and MA as $L^a = \hat{Q}^a = \hat{Q}_R^a$ and $L^m = \hat{Q}^m = \hat{Q}_R^m$, respectively.

Similar to Eq. (8), given the visual appearance and motion feature F_{LG}^a, F_{LG}^m and question semantics L^a, L^m, the multi-modal output for AS and MS are computed as:

$$\begin{aligned}
\hat{L}_r^a &= U_r^a + \sigma^a(\text{LN}(U_r^a)) \\
\hat{L}_r^m &= U_r^m + \sigma^m(\text{LN}(U_r^m)) \\
U_r^a &= \text{LN}(F_{C,r-1}^a) + \text{MMA}^a(F_{C,r-1}^a, L^a) \\
U_r^m &= \text{LN}(F_{C,r-1}^m) + \text{MMA}^m(F_{C,r-1}^m, L^m)
\end{aligned} \tag{9}$$

where $F_{C,0}^a = F_C^a$, $F_{C,0}^m = F_C^m$. The output visual clues of QA and MA are denoted as $F_s^a = \hat{L}_R^a$ and $F_s^m = \hat{L}_R^m$, respectively. Then, the output of the AS and MS are concatenated to make the final visual output $F = [F_s^a, F_s^m] \in \mathbb{R}^{2d}$. The output of the QA and QM are concatenated to make the final question semantics output $L = [L^a, L^m] \in \mathbb{R}^{2d}$.

3.5. Cross-modal Feature Fusion and Training

For (v, q) and $(\hat{c}, q)$, their visual and linguistic outputs of the IVLT model are denoted as F, F_c and L, L_c, respectively. Inspired by the adaptive feature fusion [46] (refer to Appendix A), we obtain the refined linguistic feature vectors $\{\widetilde{L}, \widetilde{L}_c\}$, which are then concatenated to form the final semantic-ware linguistic feature $\widetilde{L} = [\widetilde{L}, \widetilde{L}_c] \in \mathbb{R}^{2d}$.

To obtain the semantic-aware visual feature, we compute the visual feature $\widetilde{F}_k$ by individually conditioning each instance from visual features $\{F_1, F_2\} = \{F, F_c\}$ to each instance from refined linguistic features $\{\widetilde{L}_1, \widetilde{L}_2\} = \{\widetilde{L}, \widetilde{L}_c\}$ using the same operation as [35]. Then, these semantic-aware visual features $\widetilde{F}_k$ $(k = 1, 2)$ are concatenated to form the final semantic-aware visual feature $\widetilde{F} \in \mathbb{R}^{2d}$.

Finally, we apply different answer decoders [35] (i.e., open-ended, multi-choice, and counting) to (v, q) and $(\hat{c}, q)$ and obtain original prediction and causal prediction losses:

$$\begin{aligned}
\mathcal{L}_o &= \text{XE}(\text{CMQR}(v, q), a) \\
\mathcal{L}_c &= \text{XE}(\text{CMQR}(\hat{c}, q), a)
\end{aligned} \tag{10}$$

where XE denotes the cross-entropy loss, a is the ground-truth answer, CMQR denotes our proposed framework. Furthermore, to make the predictions of original and causal scene consistent, we apply KL-divergence between the predictions of (v, q) and $(\hat{c}, q)$:

$$\mathcal{L}_a = \text{KL}(\text{CMQR}(v, q), \text{CMQR}(\hat{c}, q)) \tag{11}$$

Finally, the learning objective of our CMQR is:

$$\mathcal{L}_{\text{CMQR}} = \mathcal{L}_o + \lambda_c \mathcal{L}_c + \lambda_a \mathcal{L}_a \tag{12}$$

4. Experiments

4.1. Datasets

In this paper, we evaluate our CMQR on four VideoQA datasets. **SUTD-TrafficQA** [71] consists of 62,535 QA pairs and 10,090 traffic videos. There are six challenging reasoning tasks including basic understanding, event forecasting, reverse reasoning, counterfactual inference, introspection and attribution analysis. **TGIF-QA** [25] has 165K QA pairs collected from 72K animated GIFs. It has four tasks: repetition count, repeating action, state transition, and frame QA. **MSVD-QA** [69] contains 50,505 algorithm-generated question-answer pairs and 1,970 trimmed video clips. **MSRVTT-QA** [69] contains 10,000 trimmed video clips and 243,680 question-answer pairs. More details of these datasets are given in Appendix D.

4.2. Implementation Details

For fair comparisons, we follow [35] to divide the videos into 8 clips for all datasets. The XCLIP [49] with ViT-L/14 pretrained on Kinetics-600 dataset is used to extract the appearance and motion features. For the question, we adopt the pre-trained 300-dimensional Glove [57] word embeddings to initialize the word features in the sentence. For parameter settings, we set the dimension d of hidden layer to 512. For the Multi-modal Transformer Block (MTB), the number of layers r is set to 3 for SUTD-TrafficQA, 8 for TGIF-QA, 5 for MSVD-QA, and 6 for MSRVTT-QA. The number of attentional heads H is set to 8. The dictionary is initialized by applying K-means over the whole visual features from the whole training set to get 512 clusters and is updated during end-to-end training. We train the model using the Adam optimizer with an initial learning rate 2e-4,

Method	Basic	Attri.	Intro.	Counter.	Fore.	Rev.	All
VQAC† [31]	34.02	49.43	34.44	39.74	38.55	49.73	36.00
MASN† [59]	33.83	50.86	34.23	41.06	41.57	50.80	36.03
DualVGR† [63]	33.91	50.57	33.40	41.39	41.57	50.62	36.07
HCRN [35]	-	-	-	-	-	-	36.49
HCRN† [35]	34.17	50.29	33.40	40.73	44.58	50.09	36.26
Eclipse [71]	-	-	-	-	-	-	37.05
IGV† [39]	-	-	-	-	-	-	37.71
CMQR (ours)	**36.10**	**52.59**	**38.38**	**46.03**	**48.80**	**58.05**	**38.63**

Table 1. Results on SUTD-TrafficQA dataset. '†' indicates the result re-implemented by the officially code.

Method	Action↑	Transition↑	FrameQA↑	Count↓
ST-VQA [25]	62.9	69.4	49.5	4.32
Co-Mem [16]	68.2	74.3	51.5	4.10
PSAC [38]	70.4	76.9	55.7	4.27
HME [14]	73.9	77.8	53.8	4.02
GMIN [20]	73.0	81.7	57.5	4.16
L-GCN [22]	74.3	81.1	56.3	3.95
HCRN [35]	75.0	81.4	55.9	3.82
HGA [27]	75.4	81.0	55.1	4.09
QueST [26]	75.9	81.0	59.7	4.19
Bridge [55]	75.9	82.6	57.5	**3.71**
QESAL[40]	76.1	82.0	57.8	3.95
ASTG [29]	76.3	82.1	61.2	3.78
CASSG [48]	77.6	**83.7**	58.7	3.83
HAIR [41]	77.8	82.3	60.2	3.88
CMQR (ours)	**78.1**	82.4	**62.3**	3.83

Table 2. Comparison with state-of-the-art methods on TGIF-QA.

a momentum 0.9, and a weight decay 0. The learning rate reduces by half when the loss stops decreasing after every 5 epochs. The batch size is set to 64. All experiments are terminated after 50 epochs. λ_c and λ_a are all set to 0.1.

4.3. Comparison with State-of-the-art Methods

Since the splits of six reasoning tasks are not provided by the original SUTD-TrafficQA dataset [71], we divide the SUTD-TrafficQA dataset into six reasoning tasks according to the question types. The results in Table 1 demonstrate that our CMQR achieves the best performance for six reasoning tasks including basic understanding, event forecasting, reverse reasoning, counterfactual inference, introspection and attribution analysis. Specifically, the CMQR improves the state-of-the-art methods Eclipse [71] and IGV [39] by 1.58% and 0.92%. Compared with the re-implemented methods VQAC†, MASN†, DualVGR†, HCRN† and IGV†, our CMQR outperforms these methods for introspection and counterfactual inference tasks that require causal relational reasoning among the causal, logic, and spatial-temporal structures of the visual and linguistic content. These results show that our CMQR has strong ability in modeling multi-level interaction and causal relations between the language and spatial-temporal structure.

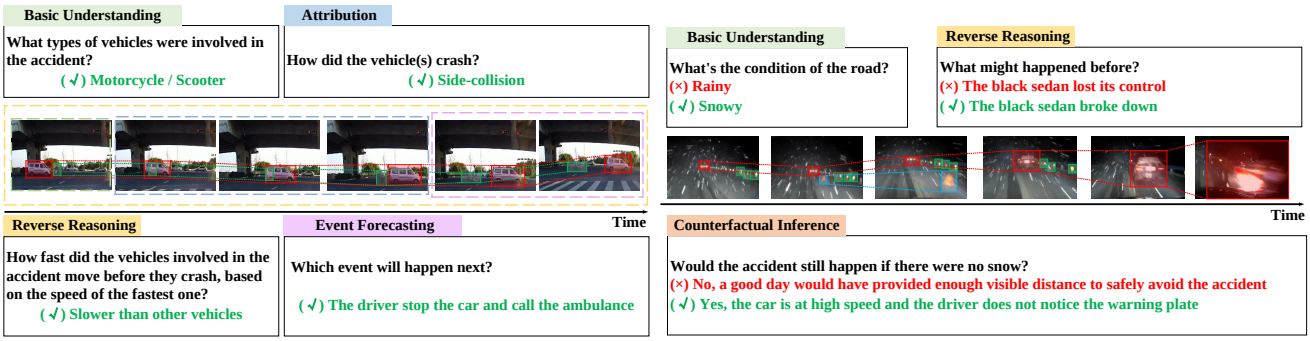
To evaluate the generalization ability of CMQR on other event-level datasets, we conduct experiments on TGIF-QA, MSVD-QA, and MSRVTT-QA datasets, as shown in Table 2-4. From Table 2, we can see that our CMQR achieves the best performance for *Action* and *FrameQA* tasks. Additionally, our CMQR also achieves relatively high performance for *Transition* and *Count* tasks. For the *Transition* task, the CMQR also outperforms nearly all comparison methods. For the *Count* task, we also achieve a competitive MSE value. From Table 3, our CMQR outperforms all the comparison methods by a significant margin. For *What*, *Who*, and *When* types, the CMQR outperforms all the comparison methods. It can be observed in Table 4 that our CMQR performs better than the best performing method IGV [39]. For *What*, *Who*, and *When* question types, the CMQR performs the best. In Table 3-4, we achieve lower performance than previous method for *How* and *Where* questions. Actually, the number of *How* and *Where* samples are much smaller than that of the other question types. Due to the existence of data bias, the model tends to learn spurious correlation from other question types, which leads to the performance degradation. Nonetheless, we can still obtain promising performance for *When*, which also has limited samples. This validates that our CMQR indeed mitigate the spurious correlations for most of the question types. Moreover, our CMQR can generalize well across different datasets and has good potential to model multi-level interaction and causal relations between the language and spatial-temporal structure. The main reasons for good generalization is that our CMQR can mitigate the visual bias by explicit causal scene learning and front-door causal intervention.

Method	What	Who	How	When	Where	All
HGA [27]	23.5	50.4	83.0	72.4	46.4	34.7
GMIN [20]	24.8	49.9	84.1	75.9	**53.6**	35.4
QueST [26]	24.5	52.9	79.1	72.4	50.0	36.1
HCRN [35]	-	-	-	-	-	36.1
CASSG [48]	24.9	52.7	**84.4**	74.1	**53.6**	36.5
QESAL[40]	25.8	51.7	83.0	72.4	50.0	36.6
Bridge [55]	-	-	-	-	-	37.2
HAIR [41]	-	-	-	-	-	37.5
VQAC [31]	26.9	53.6	-	-	-	37.8
MASN [59]	-	-	-	-	-	38.0
HRNAT [17]	-	-	-	-	-	38.2
ASTG [29]	26.3	55.3	82.4	72.4	50.0	38.2
DualVGR [63]	28.6	53.8	80.0	70.6	46.4	39.0
IGV [39]	-	-	-	-	-	40.8
CMQR (Ours)	**37.0**	**59.9**	81.0	**75.8**	46.4	**46.4**

Table 3. Comparison with state-of-the-art methods on MSVD-QA.

4.4. Ablation Studies

We further conduct ablation experiments to verify the contributions of five essential components: 1) Explicit Causal Scene Learning (ECSL), 2) Visual Front-door Causal Intervention (VFCI), 3) Visual Causality Discovery (VCD), 4) Visual Causality Discovery (VCD), and Interactive Visual-Linguistic Transformer (IVLT). From Ta-



(a) True example **(b) Failure Example**

Figure 5. Visualization of visual-linguistic causal reasoning example on SUTD-TrafficQA dataset. The color windows denotes the concentrated causal scenes for the corresponding question types.

Method	What	Who	How	When	Where	All
QueST [26]	27.9	45.6	83.0	75.7	31.6	34.6
HRA [11]	-	-	-	-	-	35.0
MASN [59]	-	-	-	-	-	35.2
HRNAT [17]	-	-	-	-	-	35.3
HGA [27]	29.2	45.7	83.5	75.2	34.0	35.5
DualVGR [63]	29.4	45.5	79.7	76.6	36.4	35.5
HCRN [35]	-	-	-	-	-	35.6
VQAC [31]	29.1	46.5	-	-	-	35.7
CASSG [48]	29.8	46.3	**84.9**	75.2	35.6	36.1
GMIN [20]	30.2	45.4	84.1	74.9	**43.2**	36.1
QESAL[40]	30.7	46.0	82.4	76.1	41.6	36.7
Bridge [55]	-	-	-	-	-	36.9
HAIR [41]	-	-	-	-	-	36.9
ClipBert [36]	-	-	-	-	-	37.4
ASTG [29]	31.1	48.5	83.1	77.7	38.0	37.6
IGV [39]	-	-	-	-	-	38.3
CMQR (ours)	**32.2**	**50.2**	82.3	**78.4**	38.0	**38.9**

Table 4. Comparison with state-of-the-art methods on MSRVTT.

Datasets	CMQR w/o ECSL	CMQR w/o VFCI	CMQR w/o VCD	CMQR w/o IVLT	CMQR
SUTD	37.68	37.42	37.28	37.75	**38.63**
MSVD	44.9	44.7	43.6	44.8	**46.4**
MSRVTT	38.1	38.0	37.5	37.7	**38.9**

Table 5. Ablation study on three datasets.

Models	SUTD-TrafficQA	MSVD-QA	MSRVTT-QA
Co-Mem [16]	35.10	34.6	35.3
Co-Mem [16]+ VCD	**37.12** (+2.02)	**40.7** (+6.1)	**38.0** (+2.7)
HGA [27]	35.81	35.4	36.1
HGA [27]+ VCD	**37.23** (+1.42)	**41.9** (+6.5)	**38.2** (+2.1)
HCRN [35]	36.49	36.1	35.6
HCRN [35]+ VCD	**37.54** (+1.05)	**42.2** (+6.1)	**37.8** (+2.2)
Our Backbone	37.42	44.7	38.0
Our Backbone + VCD	**38.63** (+1.21)	**46.4** (+1.9)	**38.9** (+0.9)

Table 6. The VCD module is applied to non-causal models.

ble 5, our CMQR achieves the best performance across all datasets and tasks. Without ECSL, the performance drops significantly due to the lack of the causal scene. This shows that the ECSL indeed find the causal scene that facilitate question reasoning. The performance of CMQR w/o ECSL, CMQR w/o VFCI are all lower than that of the CMQR. This validates that both the causal scene and visual front-door causal intervention are indispensable and contribute to discover the causal structures, and thus improve the model performance. The performance of CMQR w/o IVLT is higher than that of CMQR w/o VCD shows that visual and linguistic causal intervention modules contribute more than the IVLT due to the existence of cross-modal bias. With all components, our CMQR performs the best because all components contribute to our CMQR.

To validate the effectiveness of our causal module VCD in non-causal models, we apply the VCD to three state-of-the-art models Co-Mem [16], HGA [27] and HCRN [35]. As shown in Table 6, our VCD brings each backbone model a sharp gain across all benchmark datasets (+0.9%∼6.5%), which evidences its model-agnostic property. To be noticed, for the causal and temporal questions (i.e., SUTD-TrafficQA), our VCD shows equivalent improvements on all four backbones (+1.05%∼2.02%). These results validate that our VCD is effective in capturing the causality and reducing the spurious correlations across different models.

To validate whether our CMQR could generalize to different visual appearance and motion features, we evaluate the performance of the CMQR using different visual backbones, as shown in Table 7. These results validates that our CMQR generalizes well across both vision-language transformers and CNN backbones due to the learned causality-ware visual-linguistic representations. More importantly, the performance improvement of our CMQR is mainly attributed to our visual causality discovery model.

4.5. Parameter and Visualization Analysis

From Table 8, we can see that 8 MMA heads performs the best because more heads can facilitate the MMA module employ more perspectives between different modalities.

	Method	Appearance	Motion	Accuracy
SUTD-QA	Eclipse [71]	ResNet-101	MobileNetV2	37.05
	Ours	XCLIP	XCLIP	**38.63** (+1.59)
	Ours	Swin-L	Video Swin-B	38.58 (+1.54)
	Ours	ResNet-101	ResNetXt-101	38.10 (+1.05)
MSVD-QA	DualVGR [63]	ResNet-101	ResNetXt-101	39.0
	Ours	XCLIP	XCLIP	**46.4** (+7.40)
	Ours	Swin-L	Video Swin-B	43.7 (+4.70)
	Ours	ResNet-101	ResNetXt-101	40.3 (+1.30)
MSRVTT-QA	HCRN [35]	ResNet-101	ResNeXt-101	35.6
	Ours	XCLIP	XCLIP	**38.9**(+3.30)
	Ours	Swin-L	Video Swin-B	38.6 (+3.00)
	Ours	ResNet-101	ResNeXt-101	37.0 (+1.40)

Table 7. Performance with different visual appearance and motion features on SUTD-TrafficQA, MSVD, and MSRVTT datasets.

		SUTD	TGIF (A)	TGIF (T)	TGIF (F)	TGIF (C)	MSVD	MSRVTT
MMA Heads	1	37.83	75.8	80.7	61.2	3.92	45.0	38.5
	2	38.17	75.7	79.7	60.6	3.96	44.7	38.5
	4	37.51	75.8	79.2	61.1	3.93	44.9	38.3
	8	**38.63**	**78.1**	**82.4**	**62.3**	**3.83**	**46.4**	**38.9**
MTB Layers	3	**38.63**	75.1	80.1	61.0	4.03	45.7	38.4
	4	37.84	76.6	80.2	61.6	3.96	45.3	38.7
	5	37.63	75.5	80.6	61.0	3.94	**46.4**	38.7
	6	37.73	76.2	80.8	61.4	4.12	45.9	**38.9**
	7	37.73	75.4	80.3	61.2	3.98	45.8	38.3
	8	37.58	**78.1**	**82.4**	**62.3**	**3.83**	45.5	38.6
Dimension	256	37.60	73.9	79.9	61.0	3.96	45.5	38.8
	512	**38.63**	**78.1**	**82.4**	**62.3**	**3.83**	**46.4**	**38.9**
	768	37.74	75.0	80.0	62.2	3.90	45.5	38.0

Table 8. Performance of CMQR with different MMA heads, MTB layers, and hidden state dimension on four datasets.

For MTB layers, the optimal layer numbers are different for different datasets. For the dimension of hidden states, 512 is the best dimensionality of hidden states of the CMQR model due to its good compromise between the feature representation ability and model complexity.

To verify the ability of the CMQR in robust spatial-temporal relational reasoning, we inspect some correct and failure examples in Fig. 5. The example in Fig. 5 (a) exhibits a strong correlation between the dominant spatial-temporal scenes and the question semantics, which validates that the CMQR is question-sensitive to effectively capture the dominant spatial-temporal content. In Fig. 5 (b), it is hard to discriminate "rainy" and "snowy" due to the similar visual appearance in the video. And the "reflective stripes" along the road are mistakenly considered as the dominant visual concepts. Since our CMQR model contains no explicit object detection pipeline, some ambiguity visual concepts are challenging to be determined. More visualization results and analysis are given in Appendix E.

5. Conclusion

In this paper, we propose an event-level visual question reasoning framework named Cross-Modal Question Reasoning (CMQR), to explicitly discover cross-modal causal structures. To explicitly discover visual causal structure, we propose a Visual Causality Discovery (VCD) architecture that learns to discover temporal question-critical scenes and mitigate the visual spurious correlations by front-door causal intervention. To align the fine-grained interactions between linguistic semantics and spatial-temporal visual concepts, we build an Interactive Visual-Linguistic Transformer (IVLT). Extensive experiments on four datasets well demonstrate the effectiveness of our CMQR for discovering visual causal structure and achieving robust event-level visual question reasoning. We believe our work could inspire more causal reasoning methods in vision-language tasks.

References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10044–10054, 2020. 3

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018. 1

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[5] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108. PMLR, 2012. 3

[6] M Besserve, A Mehrjou, R Sun, and B Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *Eighth International Conference on Learning Representations (ICLR 2020)*, 2020. 3

[7] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2917–2927, 2022. 1

[8] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4162–4170, 2020. 3

[9] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. Counterfactual critic multi-agent training for scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4613–4623, 2019. 3

[10] Weixing Chen, Yang Liu, Ce Wang, Guanbin Li, Jiarui Zhu, and Liang Lin. Visual-linguistic causal intervention for radiology report generation. *arXiv preprint arXiv:2303.09117*, 2023. 3

[11] Muhammad Iqbal Hasan Chowdhury, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Hierarchical relational attention for video question answering. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 599–603. IEEE, 2018. 8

[12] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017. 1

[13] 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. 3

[14] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 2, 7

[15] Zhiyuan Fang, Shu Kong, Charless Fowlkes, and Yezhou Yang. Modularized textual grounding for counterfactual resilience. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6378–6388, 2019. 3

[16] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585, 2018. 7, 8

[17] Lianli Gao, Yu Lei, Pengpeng Zeng, Jingkuan Song, Meng Wang, and Heng Tao Shen. Hierarchical representation network with auxiliary tasks for video captioning and video question answering. *IEEE Transactions on Image Processing*, 2022. 7, 8

[18] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019. 3

[19] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019. 3

[20] Mao Gu, Zhou Zhao, Weike Jin, Richang Hong, and Fei Wu. Graph-based multi-interaction network for video question answering. *IEEE Transactions on Image Processing*, 30:2758–2770, 2021. 7, 8

[21] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 264–279, 2018. 3

[22] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11021–11028, 2020. 2, 7

[23] Hao Huang, Luowei Zhou, Wei Zhang, Jason J Corso, and Chenliang Xu. Dynamic graph modules for modeling object-object interactions in activity recognition. In *British Machine Vision Conference*, 2019. 3

[24] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4

[25] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 6, 7

[26] Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11101–11108, 2020. 7, 8

[27] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11109–11116, 2020. 2, 7, 8

[28] C JiayinCai, Cheng Shi, Lei Li, Yangyang Cheng, and Ying Shan. Feature augmented memory with global attention network for videoqa. In *IJCAI*, pages 998–1004, 2020. 2

[29] Weike Jin, Zhou Zhao, Xiaochun Cao, Jieming Zhu, Xiuqiang He, and Yueting Zhuang. Adaptive spatio-temporal graph enhanced vision-language representation for video qa. *IEEE Transactions on Image Processing*, 30:5477–5489, 2021. 7, 8

[30] Atsushi Kanehira, Kentaro Takemoto, Sho Inayoshi, and Tatsuya Harada. Multimodal explanations by predicting counterfactuality in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8594–8602, 2019. 3

[31] Nayoung Kim, Seong Jong Ha, and Je-Won Kang. Video question answering using language-guided deep compressed-domain video feature. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1708–1717, 2021. 7, 8

[32] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 1

[33] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 1

[34] Hung LE, Doyen SAHOO, and Nancy F CHEN. Choi. multimodal transformer networks for end-to-end video-grounded dialogue systems.(2019). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,*

Florence, Italy, 2019 July 28-August, volume 2, pages 5612–5623. 3

[35] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020. 1, 2, 6, 7, 8, 9

[36] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 2, 8

[37] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8769–8784, 2020. 1

[38] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665, 2019. 1, 2, 7

[39] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2928–2937, 2022. 2, 3, 7, 8

[40] Fei Liu, Jing Liu, Richang Hong, and Hanqing Lu. Question-guided erasing-based spatiotemporal attention learning for video question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 7, 8

[41] Fei Liu, Jing Liu, Weining Wang, and Hanqing Lu. Hair: Hierarchical visual-semantic relational reasoning for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1698–1707, 2021. 2, 7, 8

[42] Yang Liu, Zhaoyang Lu, Jing Li, and Tao Yang. Hierarchically learned view-invariant representations for cross-view action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2416–2430, 2018. 3

[43] Yang Liu, Zhaoyang Lu, Jing Li, Tao Yang, and Chao Yao. Global temporal representation based cnns for infrared action recognition. *IEEE Signal Processing Letters*, 25(6):848–852, 2018. 3

[44] Yang Liu, Zhaoyang Lu, Jing Li, Tao Yang, and Chao Yao. Deep image-to-video adaptation and fusion networks for action recognition. *IEEE Transactions on Image Processing*, 29:3168–3182, 2019. 1

[45] Yang Liu, Keze Wang, Guanbin Li, and Liang Lin. Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition. *IEEE Transactions on Image Processing*, 30:5573–5588, 2021. 1

[46] Yang Liu, Keze Wang, Lingbo Liu, Haoyuan Lan, and Liang Lin. Tcgl: Temporal contrastive graph for self-supervised video representation learning. *IEEE Transactions on Image Processing*, 31:1978–1993, 2022. 1, 6

[47] Yang Liu, Yu-Shen Wei, Hong Yan, Guan-Bin Li, and Liang Lin. Causal reasoning meets visual representation learning: A prospective study. *Machine Intelligence Research*, pages 1–27, 2022. 2, 3

[48] Yun Liu, Xiaoming Zhang, Feiran Huang, Bo Zhang, and Zhoujun Li. Cross-attentional spatio-temporal semantic graph networks for video question answering. *IEEE Transactions on Image Processing*, 31:1684–1696, 2022. 7, 8

[49] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 6

[50] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6790–6800, 2018. 3

[51] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2765–2775, 2021. 3

[52] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 1–18. Springer, 2022. 3

[53] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021. 2, 3

[54] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 464–474, 2021. 3

[55] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge to answer: Structure-aware graph interaction network for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15526–15535, 2021. 1, 7, 8

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

[57] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3, 6

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

[59] Ahjeong Seo, Gi-Cheon Kang, Joonhan Park, and Byoung-Tak Zhang. Attend what you need: Motion-appearance syn-

ergistic networks for video question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6167–6177, 2021. 7, 8

[60] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. *Advances in Neural Information Processing Systems*, 2015:2440–2448, 2015. 1

[61] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. 3

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[63] Jianyu Wang, Bingkun Bao, and Changsheng Xu. Dualvgr: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia*, 2021. 7, 8, 9

[64] Pei Wang and Nuno Vasconcelos. Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8981–8990, 2020. 3

[65] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020. 3

[66] 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 *European Conference on computer vision*, pages 728–744. Springer, 2020. 3

[67] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021. 2, 3

[68] Wei Wang, Junyu Gao, and Changsheng Xu. Weakly-supervised video object grounding via causal intervention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[69] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 2, 6

[70] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 5

[71] Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9878–9888, 2021. 6, 7, 9

[72] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[73] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9847–9857, 2021. 3, 5

[74] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016. 2

[75] Yuying Zhu, Yang Zhang, Lingbo Liu, Yang Liu, Guanbin Li, Mingzhi Mao, and Liang Lin. Hybrid-order representation learning for electricity theft detection. *IEEE Transactions on Industrial Informatics*, 2022. 1