Cross-Modal Causal Relational Reasoning for Event-Level Visual Question Answering

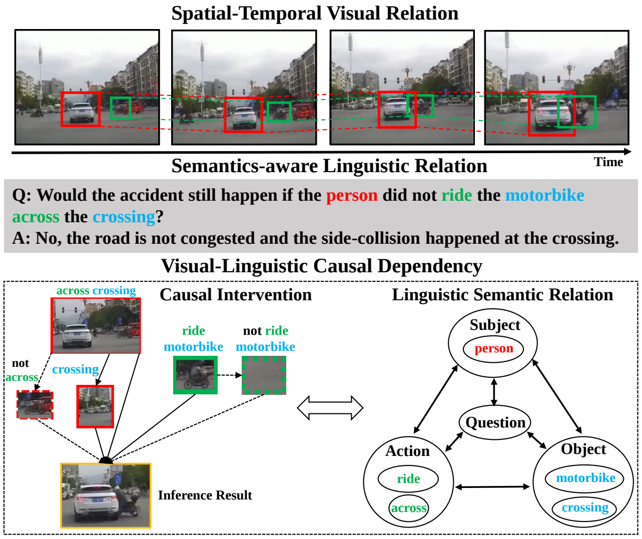
Yang Liu, *Member, IEEE*, Guanbin Li, *Member, IEEE*, and Liang Lin, *Senior Member, IEEE*

Abstract—Existing visual question answering methods often suffer from cross-modal spurious correlations and oversimplified event-level reasoning processes that fail to capture event temporality, causality, and dynamics spanning over the video. In this work, to address the task of event-level visual question answering, we propose a framework for cross-modal causal relational reasoning. In particular, a set of causal intervention operations is introduced to discover the underlying causal structures across visual and linguistic modalities. Our framework, named **C**ross-**M**odal **C**ausal Relat**I**onal **R**easoning (CMCIR), involves three modules: i) Causality-aware Visual-Linguistic Reasoning (CVLR) module for collaboratively disentangling the visual and linguistic spurious correlations via front-door and back-door causal interventions; ii) Spatial-Temporal Transformer (STT) module for capturing the fine-grained interactions between visual and linguistic semantics; iii) Visual-Linguistic Feature Fusion (VLFF) module for learning the global semantic-aware visual-linguistic representations adaptively. Extensive experiments on four event-level datasets demonstrate the superiority of our CMCIR in discovering visual-linguistic causal structures and achieving robust event-level visual question answering. The datasets, code, and models are available at https://github.com/HCPLab-SYSU/CMCIR.

Index Terms—Visual Question Answering, Causal Inference, Cross-Modal Reasoning, Video Event Understanding.

1 Introduction

With the rapid development of deep learning [1], event understanding [2] has become a prominent research topic in video analysis [3], [4], [5] because videos have good potential to go beyond image-level understanding (scenes, people, objects, activities, etc.) to understand event temporality, causality, and dynamics. Accurate and efficient cognition and reasoning over complex events are extremely important in video-language understanding. Since natural language can potentially describe a richer event space [6] that facilitates deeper event understanding, we focus on the complex (temporal, causal) event-level visual question answering task in a cross-modal (visual, linguistic) setting. Our task aims to fully comprehend the richer multi-modal event space and answer the given question in a causality-aware way. To achieve event-level visual question answering [7], [8], [9], the model needs to have a 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. Therefore, robust and reliable multi-modal relation reasoning is essential in event-level visual question answering. Actually, understanding events in multi-modal visual-linguistic context is a long-standing challenge. Existing visual question answering methods [10], [11], [12], [13] use recurrent neural networks (RNNs) [14], attention mechanisms [15] or Graph Convolutional Networks [16] for relation reasoning between visual and linguistic modalities. Although achieving promising results, these methods suffer from two common limitations.

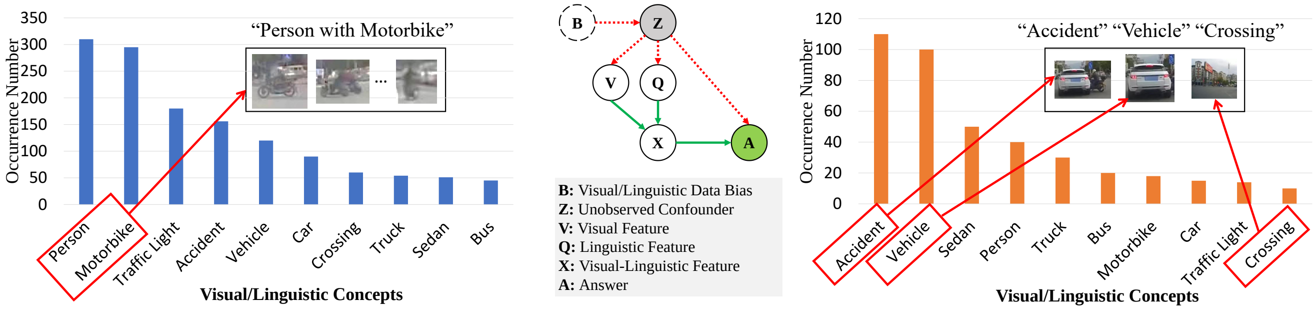


Q: Would the accident still happen if the person did not ride the motorbike across the crossing?
A: No, the road is not congested and the side-collision happened at the crossing.

Fig. 1. An example of an event-level counterfactual visual question answering task. The counterfactual inference is to obtain the outcome of certain hypothesis that does not occur in the visual scene. To infer the causality-aware answer, the model is required to explore the visual-linguistic causal dependencies and spatial-temporal relation.

First, existing visual question answering methods usually focus on simple events that do not require a deep understanding of causality, temporal relations, and linguistic interactions, and tend to overlook more challenging events. In Fig. 1, given a video and an associated question, a typical human reasoning process involves first memorizing the relevant objects and their interactions in each video frame (e.g., car runs on the road, person rides a motorbike across a crossing), then deriving the corresponding answer based on this memorized video content. However, the event-level counterfactual visual question answering task in Fig. 1 requires the outcome of certain hypotheses (e.g., "the person

- *Y. Liu, G. Li and L. Lin are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, and also with the GuangDong Province Key Laboratory of Information Security Technology. E-mail: liuy856@mail.sysu.edu.cn; liguanbin@mail.sysu.edu.cn; linliang@ieee.org. (Corresponding author: Liang Lin.)*



(a) Training Set **(b) Structured Causal Model** **(c) Testing Set**

Fig. 2. The example (Fig. 1) shows why the visual question answering model without causal reasoning tends to learn spurious correlations. (a) A training dataset constructed with visual and linguistic biases where the concepts "person" and "motorbike" frequently appear. (b) The structured causal model (SCM) shows how the confounder induces the spurious correlation in event-level visual question answering. The green path denotes the unbiased visual question answering (the true causal effect). The red path is the biased visual question answering caused by the confounders (the back-door path). (c) As a result, if we provide some samples where the "vehicle" concept is highly related to the "accident" to reason how the accident actually happens, the model does not really exploit the true question intention and dominant visual evidence to infer the answer.

did not ride the motorbike across the crossing") that do not occur in the given video. Simply correlating relevant visual contents cannot 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 must conduct hierarchical relational reasoning and fully explore the causality, logic, and spatial-temporal dynamic structures of the visual and linguistic content. This involves performing causal intervention to discover the true causal structure that facilitates answering the question truthfully based on the imagined visual evidence and the correct question intention. However, the multi-level interaction and causal relations between the language and spatial-temporal structure of complex multi-modal events are not fully explored.

Second, the current visual question answering models tend to capture spurious linguistic or visual correlations introduced by the confounders rather than the true causal structure and causality-aware multi-modal representations, leading to an unreliable reasoning process [17], [18], [19], [20]. Fig. 2 shows that some frequently appearing concepts in linguistic and visual modalities can be considered as the confounders. The "linguistic bias" represents strong correlations between questions and answers, while the "visual bias" represents the strong correlations between certain key visual features and answers. For example, the training dataset is built with visual and linguistic biases, where the concepts "person" and "motorbike" frequently appear (Fig. 2). Such biased dataset entails two causal effects: visual and linguistic biases B leads to confounder Z, which then affects the visual feature V, question feature Q, visual-linguistic feature X, and the answer A. Thus, we can draw two causal links to describe these causal effects: $Z \rightarrow \{V, Q\} \rightarrow X$ and $Z \rightarrow A$. If we want to learn the true causal effect $\{V, Q\} \rightarrow X \rightarrow A$ while using the biased dataset to train this model (Fig. 2 (a)), the 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). Therefore, the model may learn the spurious correlation between "motorbike" and "person" without considering the "vehicle" concept (i.e., exploit the true question intention and dominant visual evidence) to reason how the accident occurred. Since the potential visual and linguistic correlations are complicated in complex events, there are significant differences in visual and linguistic biases between the training and testing sets. To mitigate the dataset bias, causal inference [21] has shown promising performance in scene graph generation [22], image classification [23] and image question answering [17], [24]. However, directly applying existing causal methods to the event-level visual question answering task may yield unsatisfactory results due to the unobservable confounder in the visual domain and the complex interaction between visual and linguistic content.

To mitigate the aforementioned limitations, this paper proposes a framework named **C**ross-**M**odal **C**ausal Relat**I**onal **R**easoning (CMCIR) for event-level VQA. The proposed Causality-aware Visual-Linguistic Reasoning (CVLR) module addresses the confounders' bias and uncovers causal structures in visual and linguistic modalities through front-door and back-door causal interventions. To address the unobservable confounder in the visual modality, a Local-Global Causal Attention Module (LGCAM) is proposed, which uses attention to aggregate local and global visual representations causality-awarely. Additionally, a back-door intervention module is designed to discover the causal effect within the linguistic modality. A Spatial-Temporal Transformer (STT) is introduced to model the multi-modal interaction between appearance-motion and language representations, containing Question-Appearance (QA), Question-Motion (QM), Appearance-Semantics (AS), and Motion-Semantics (MS) modules. Finally, a novel Visual-Linguistic Feature Fusion (VLFF) module is proposed to adaptively fuse causality-aware visual and linguistic features. Experimental results on various datasets show that CMCIR outperforms state-of-the-art methods. The main contributions of the paper can be summarized as follows:

- We propose a causality-aware event-level visual question answering framework named **C**ross-**M**odal **C**ausal Relat**I**onal **R**easoning (CMCIR), to discover true causal structures via causal intervention on the integration of visual and linguistic modalities and achieve robust event-level visual question answering performance. To the best of our knowledge, we are the first to discover cross-modal causal structures for the event-level visual question answering task.

- We introduce a linguistic back-door causal intervention module guided by linguistic semantic relations to mitigate the spurious biases and uncover the causal dependencies within the linguistic modality. To disentangle the visual spurious correlations, we propose a **L**ocal-**G**lobal **C**ausal **A**ttention **M**odule (LGCAM) that aggregates the local and global visual representations by front-door causal intervention.
- We construct a **S**patial-**T**emporal **T**ransformer (STT) that models the multi-modal co-occurrence interactions between the visual and linguistic knowledge, to discover the fine-grained interactions among linguistic semantics, spatial, and temporal representations.
- To adaptively fuse the causality-aware visual and linguistic features, we introduce a **V**isual-**L**inguistic **F**eature **F**usion (VLFF) module that leverages the hierarchical linguistic semantic relations to learn the global semantic-aware visual-linguistic features.
- Extensive experiments on SUTD-TrafficQA, TGIF-QA, MSVD-QA, and MSRVTT-QA datasets show the effectiveness of our CMCIR for discovering visual-linguistic causal structures and achieving promising event-level visual question answering performance.

2 Related Works

2.1 Visual Question Answering

Compared to image-based visual question answering (i.e., ImageQA) [25], [26], [27], event-level visual question answering (i.e., VideoQA) is much more challenging due to the extra temporal dimension. To solve the VideoQA problem, the model needs to capture spatial-temporal and visual-linguistic relations to infer the answer. To explore relational reasoning in VideoQA, Xu et al. [28] proposed an attention mechanism to exploit the appearance and motion knowledge with the question as a guidance. Jang et al. [29], [30] released a large-scale VideoQA dataset named TGIF-QA and proposed a dual-LSTM based method with both spatial and temporal attention. Later on, some hierarchical attention and co-attention based methods [11], [31], [32] are proposed to learn appearance-motion and question-related multi-modal interactions. Le et al. [12] proposed hierarchical conditional relation network (HCRN) to construct sophisticated structures for representation and reasoning over videos. Jiang et al. [33] introduced the heterogeneous graph alignment (HGA) nework that aligns the inter- and intra-modality information for cross-modal reasoning. Huang et al. [10] proposed a location-aware graph convolutional network to reason over detected objects. Lei et al. [34] employed sparse sampling to build a transformer-based model named CLIPBERT and achieve end-to-end video-and-language understanding. Liu et al. [35] proposed a hierarchical visual-semantic relational reasoning (HAIR) framework to perform hierarchical relational reasoning.

Unlike the works that focus on relatively simple events like movie, TV-show or synthetic videos, our CMCIR framework focuses on complex event-level visual question answering and performs cross-modal causal relational reasoning on the spatial-temporal and linguistic content. The only existing work for event-level urban visual question answering is Eclipse [36], which built an event-level urban traffic visual question answering dataset and proposed an efficient glimpse network to achieve computation-efficient and reliable video reasoning. Different from the Eclipse that focuses on the exploration of the efficient and dynamic reasoning in urban traffic events, our work aims to uncover the causal structures behind the visual-linguistic modalities and models the interaction between the appearance-motion and language knowledge in a causality-aware manner. In addition, these previous works tend to capture spurious linguistic or visual correlations within the videos, while we build a Causality-aware Visual-Linguistic Reasoning (CVLR) module to mitigate the bias caused by confounders and uncover the causal structures for the integration of complex event-level visual and linguistic modalities.

2.2 Relational Reasoning for Event Understanding

Besides VideoQA, relational reasoning has been explored in other event understanding tasks, such as action recognition [37], [38], [39] and spatial-temporal grounding [40]. To recognize and localize actions, Girdhar et al. [41] introduced a transformer-style architecture to aggregate features from the spatiotemporal context around the person. For action detection, Huang et al. [42] introduced a dynamic graph module to model object-object interactions in video actions. Ma et al. [43] utilized an LSTM to model interactions between arbitrary subgroups of objects. Mavroudi et al. [44] built a symbolic graph using action categories. Pan et al. [45] 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 textual query, Nan et al. [46] 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. [47] proposed a causal framework to learn the deconfounded object-relevant association for robust video object grounding. However, these methods only perform relational reasoning over visual modality and neglects potential causal structures from linguistic semantic relation, resulting in incomplete and unreliable understanding of visual-linguistic content. Additionally, our CMCIR conducts causality-aware spatial-temporal relational reasoning to uncover the causal structure for visual-linguistic modality and utilizes hierarchical semantic knowledge for spatial-temporal relational reasoning.

2.3 Causal Inference in Visual Representation Learning

Compared to the conventional debiasing techniques [48], causal inference [21], [49], [50] shows its potential in mitigating spurious correlations [51] and disentangling model effects [52] for better generalization. Counterfactual and causal inference have attracted increasing attention in several computer vision tasks, including visual explanations [53], [54], scene graph generation [22], [55], image recognition [19], [24], video analysis [46], [56], [57], and vision-language tasks [17], [18], [58], [59], [60]. Specifically, Tang et al. [61], Zhang et al. [62], Wang et al. [24], and Qi et al. [63] computed the direct causal effect and mitigated the bias based on observable confounders. Counterfactual

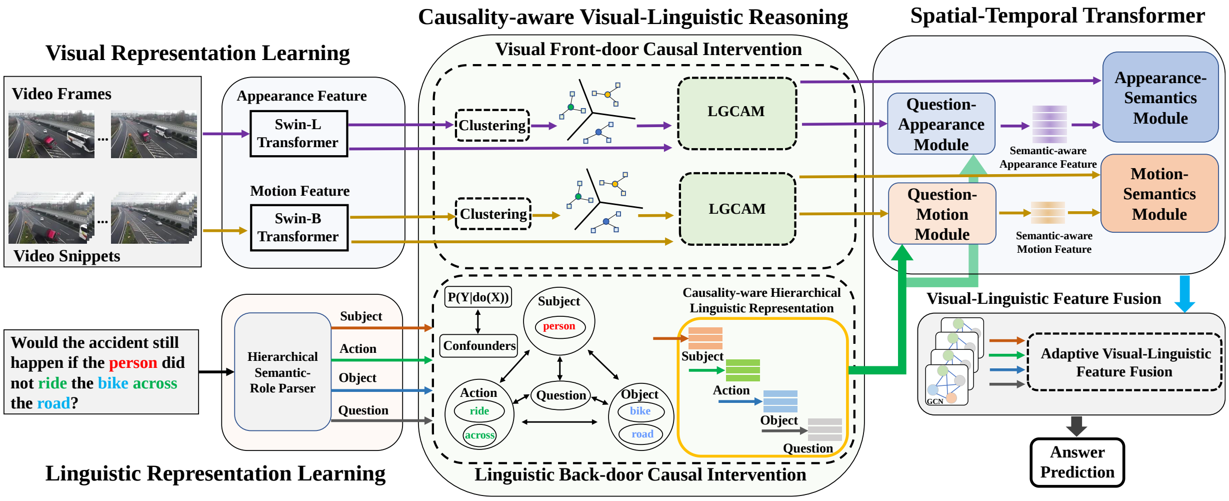

Visual Representation Learning
Video Frames
Video Snippets
Appearance Feature
Swin-L Transformer
Motion Feature
Swin-B Transformer
Causality-aware Visual-Linguistic Reasoning
Visual Front-door Causal Intervention
Clustering
LGCAM
Clustering
LGCAM
Spatial-Temporal Transformer
Question-Appearance Module
Appearance-Semantics Module
Semantic-aware Appearance Feature
Question-Motion Module
Motion-Semantics Module
Semantic-aware Motion Feature
Would the accident still happen if the person did not ride the bike across the road?
Hierarchical Semantic-Role Parser
Subject
Action
Object
Question
P(Y|do(X))
Confounders
Subject
person
Action
ride
across
Question
Object
bike
road
Causality-ware Hierarchical Linguistic Representation
Subject
Action
Object
Question
Linguistic Representation Learning
Linguistic Back-door Causal Intervention
Visual-Linguistic Feature Fusion
GCN
Adaptive Visual-Linguistic Feature Fusion
Answer Prediction

Fig. 3. Overview of CMCIR. The Linguistic Representation Learning (LRL) aims to parse the question into relation-centered tuples (subject, action, object) and then learns the hierarchical linguistic representations. The Causality-aware Visual-Linguistic Reasoning (CVLR) contains a visual front-door causal intervention module and a linguistic back-door causal intervention module. The visual front-door causal intervention module contains the Local-Global Causal Attention Module (LGCAM) that aggregates the local and global appearance and motion representations in a causality-aware way. The linguistic back-door causal intervention module models the linguistic confounder set from the perspective of semantic roles and de-confounds the language bias based on a structured causal model (SCM). Based on the causality-aware visual and linguistic representations, the Spatial-Temporal Transformer (STT) models the interaction between the appearance-motion and language knowledge in a coarse-to-fine manner. Finally, the Visual-Linguistic Feature Fusion (VLFF) module applies semantic graph guided adaptive feature fusion to obtain the multi-modal output.

based solutions are also effective, for example, Agarwal et al. [64] proposed a counterfactual sample synthesising method based on GAN [65]. Chen et al. [66] tried to replace critical objects and critical words with a mask token and reassigned a answer to synthesis counterfactual QA pairs. Apart from sample synthesising, Niu et al. [17] developed a counterfactual VQA framework that reduce multi modality bias by using causality approach named Natural Indirect Effect and Total Direct Effect to eliminate the mediator effect. Li et al. [20] proposed an Invariant Grounding for VideoQA (IGV) to force the VideoQA models to shield the answering process from the negative influence of spurious correlations. Liu et al. [59] introduced Visual Causality Discovery (VCD) architecture to find question-critical scene temporally and disentangle the visual spurious correlations by the front-door causal intervention.

However, most of the existing causal visual tasks are relatively simple without considering more challenging tasks such as video understanding and event-level visual question answering. Although some recent works CVL [58], Counterfactual VQA [17], CATT [18], IGV [20] and VCD [59] focused on visual question answering tasks, they adopted structured causal model (SCM) to eliminate either the linguistic or visual bias without considering cross-modal causality discovery. Different from previous methods, our CMCIR aims for event-level visual question answering that requires fine-grained understanding of spatial-temporal visual relation, linguistic semantic relation, and visual-linguistic causal dependency. Moreover, our Causality-aware Visual-Linguistic Reasoning (CVLR) applies front-door and back-door causal intervention modules to discover cross-modal causal structures.

3 Methodology

The framework of the CMCIR is shown in Fig. 3, which is an event-level visual question answering architecture. In this section, we present the detailed implementations of CMCIR.

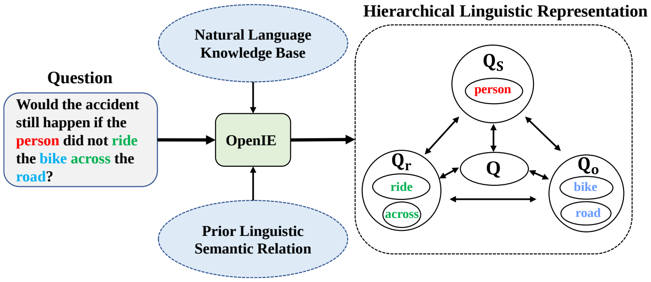

Hierarchical Linguistic Representation
Natural Language Knowledge Base
Question
Would the accident still happen if the person did not ride the bike across the road?
OpenIE
Prior Linguistic Semantic Relation
Q_S
person
Q_r
ride
across
Q
Q_o
bike
road

Fig. 4. The proposed Hierarchical Semantic-Role Parser (HSRP) parses the question into verb-centered relation tuples (subject, action, object).

3.1 Visual Representation Learning

The goal of event-level visual question answering is to deduce an answer $\tilde{a}$ from a video $\mathcal{V}$ with a given question q. The answer $\tilde{a}$ can be found in an answer space $\mathcal{A}$ which is a predefined set of possible answers for open-ended questions or a list of answer candidates for multiple-choice questions. 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 feature: frame-wise appearance feature vectors $F_i^a = \{f_{i,j}^a | f_{i,j}^a \in \mathbb{R}^{1536}, j = 1, \ldots, T\}$ and motion feature vector at clip level $f_i^m \in \mathbb{R}^{1024}$. In our experiments, Swin-L [67] is used to extract the frame-level appearance features F^a and Video Swin-B [68] is applied to extract the clip-level motion features F^m. Then, we use a linear feature transformation layer to map F^a and F^m into the same d-dimensional feature space. Thus, we have $f_{i,j}^a, f_i^m \in \mathbb{R}^d$.

3.2 Linguistic Representation Learning

From the perspective of linguistic semantic relations, a question usually contains the vocabulary of a subject, an action, and an object, since most videos can be described as "somebody doing something". Therefore, we propose an efficient approach to approximate the confounder set distribution from the perspective of natural language. Specifically, we

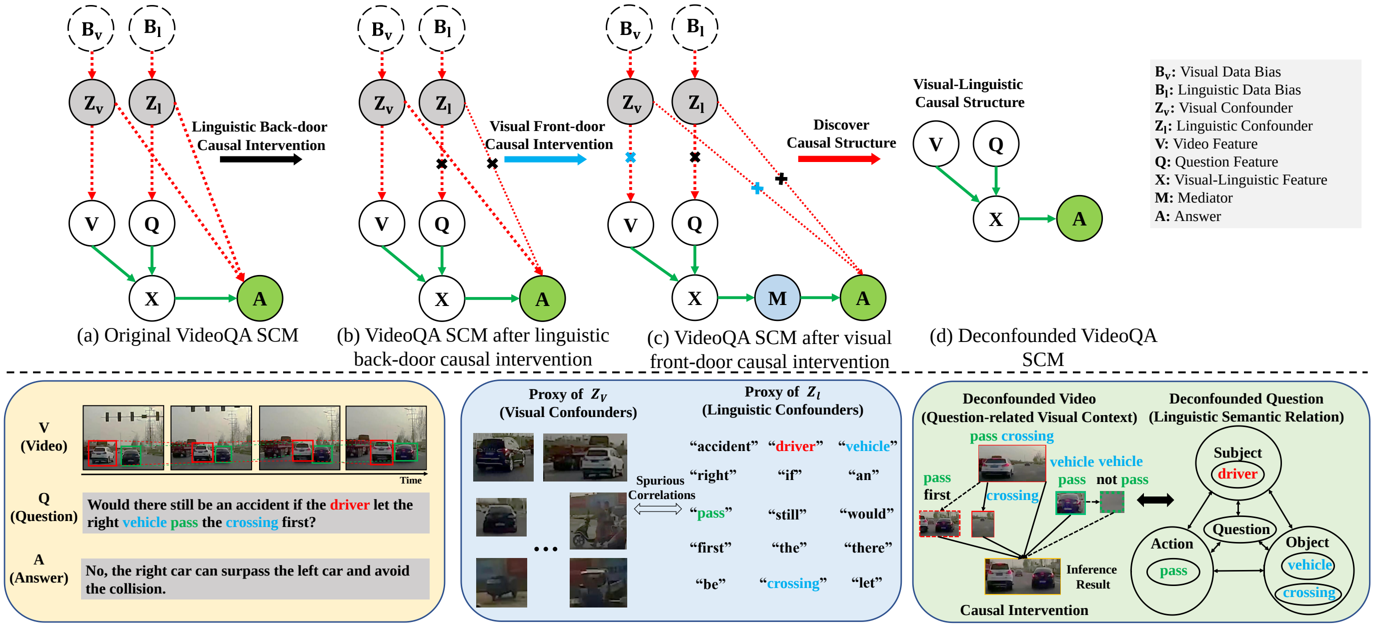

B_v
B_l
Z_v
Z_l
V
Q
X
A
Linguistic Back-door Causal Intervention
Visual Front-door Causal Intervention
M
Discover Causal Structure
Visual-Linguistic Causal Structure
(a) Original VideoQA SCM
(b) VideoQA SCM after linguistic back-door causal intervention
(c) VideoQA SCM after visual front-door causal intervention
(d) Deconfounded VideoQA SCM
B_v: Visual Data Bias
B_l: Linguistic Data Bias
Z_v: Visual Confounder
Z_l: Linguistic Confounder
V: Video Feature
Q: Question Feature
X: Visual-Linguistic Feature
M: Mediator
A: Answer
V (Video)
Time
Q (Question)
Would there still be an accident if the driver let the right vehicle pass the crossing first?
A (Answer)
No, the right car can surpass the left car and avoid the collision.
Proxy of Z_V (Visual Confounders)
Proxy of Z_l (Linguistic Confounders)
Spurious Correlations
"accident" "driver" "vehicle"
"right" "if" "an"
"pass" "still" "would"
"first" "the" "there"
"be" "crossing" "let"
Deconfounded Video (Question-related Visual Context)
Deconfounded Question (Linguistic Semantic Relation)
pass crossing
vehicle vehicle
pass first
crossing
pass not pass
Inference Result
Causal Intervention
Subject
driver
Question
Action
pass
Object
vehicle
crossing

Fig. 5. The proposed causal graph of visual-linguistic causal intervention. The green path represents the unbiased visual question answering, which is the true causal effect. The red path shows the biased visual question answering caused by the confounders, also known as the back-door path. The bottom part of the figure provides an intuitive explanation of a real VideoQA sample using visual-linguistic causal intervention.

build a Hierarchical Semantic-Role Parser (HSRP) to parse the question into verb-centered relation tuples (subject, action, object) and construct three sets of vocabulary accordingly. The verb-centered relation tuples are subsets of the words of the original question around the key words subject, action, and object. The HSRP is based on the state-of-the-art Open Information Extraction (OpenIE) model [69], which discovers linguistic semantic relations from a large-scale natural language knowledge base, as shown in Fig. 4. For the whole question Q, subject Q_s, action Q_r, object Q_o, and answer candidates A, each word is respectively embedded into a vector of 300 dimensions by adopting pre-trained GloVe [70] word embedding, which is further mapped into a d-dimensional space using linear transformation. Then, we represent the corresponding question and answer semantics as $Q=\{q_1,q_2,\cdots,q_L\}$, $Q_s=\{q_1^s,q_2^s,\cdots,q_{L_s}^s\}$, $Q_r=\{q_1^r,q_2^r,\cdots,q_{L_r}^r\}$, $Q_o=\{q_1^o,q_2^o,\cdots,q_{L_o}^o\}$, $A=\{a_1,a_2,\cdots,a_{L_a}\}$, where L, L_s, L_r, L_o, L_a indicate the length of Q, Q_s, Q_r, Q_o, and A.

To obtain contextual linguistic representations that aggregate dynamic long-range temporal dependencies from multiple time-steps, a BERT [71] model is employed to encode Q, Q_s, Q_r, Q_o, and the answer A, respectively. Finally, the updated representations for the question, question tuples, and answer candidates can be written as:

$$\begin{aligned} &Q=\{q_i|q_i\in\mathbb{R}^d\}_{i=1}^{L},\quad Q_s=\{q_i^s|q_i^s\in\mathbb{R}^d\}_{i=1}^{L_s},\\ &Q_r=\{q_i^r|q_i^r\in\mathbb{R}^d\}_{i=1}^{L_r},\quad Q_o=\{q_i^o|q_i^o\in\mathbb{R}^d\}_{i=1}^{L_o} \end{aligned} \tag{1}$$

and

$$A=\{a_i|a_i\in\mathbb{R}^d\}_{i=1}^{L_a} \tag{2}$$

3.3 Causality-aware Visual-Linguistic Reasoning

For visual-linguistic question reasoning with spatial-temporal data, we employ Pearl's structural causal model (SCM) [21] to model the causal effect between video-question pairs and the answer, as shown in Fig. 5 (a). The nodes are variables and edges are causal relations. Conventional VQA methods only learn: $\{V,Q\}\rightarrow X\rightarrow A$, which learn the ambiguous statistics-based association $P(A|V,Q)$. They ignore the spurious association brought by the confounder, while our method address these problems in a causal framework and propose a fundamental solution. In the following, we detail the rationale behind our causal graph. The bottom part of Fig. 5 presents the high-level explanation of the visual-linguistic causal intervention. Here, we provide the detailed interpretation for some subgraphs.

$\{B_v,B_l\}\rightarrow\{Z_v,Z_l\}\rightarrow\{V,Q\}$. The visual and linguistic confounders Z_v and Z_l (probably an imbalanced distribution of the dataset caused by data sampling biases B_v and B_l) may lead to spurious correlations between videos and certain words. The *do*-operation on $\{V,Q\}$ can enforce their values and cuts off the direct dependency between $\{V,Q\}$ and their parents Z_v and Z_l (Fig. 5 (b) and (c)).

$\{B_v,B_l\}\rightarrow\{Z_v,Z_l\}\rightarrow A$. Since Z_v and Z_l are the visual and linguistic confounders for the dataset, we must also have Z_v and Z_l connected to prediction A via directed paths excluding $\{V,Q\}$. This ensures the consideration of the confounding impact from Z_v and Z_l to A.

$A\leftarrow\{Z_v,Z_l\}\rightarrow\{V,Q\}\rightarrow X$. There are two back-door paths where confounders Z_v and Z_l affect the video V and question Q respectively, and ultimately affect answer A, leading the model to learn the spurious association. As discussed before, if we had successfully cut off the path $\{Z_v,Z_l\}\nrightarrow\{V,Q\}\rightarrow X\rightarrow A$, $\{V,Q\}$ and A are deconfounded and the model can learn the true causal effect $\{V,Q\}\rightarrow X\rightarrow A$.

To train a video question answering model that learns the true causal effect $\{V,Q\}\rightarrow X\rightarrow A$: the model should reason the answer A from the video V and the question Q instead of exploiting the spurious correlations induced by the confounders Z_v and Z_l (i.e., overexploiting the co-occurrence between the visual and linguistic concepts). For example, since the answer to the question "What the color of the vehicle involved in the accident?" is "white" in most cases, the model will easily learn the spurious correlation between the concepts "vehicle" and "white". Conventional visual-linguistic question reasoning models usually focus on correlations between video and question by directly learning $P(A|V,Q)$ without considering the confounders Z_v and Z_l. Thus, when given an accident video of black vehicle, the model still predicts answer "white" with strong confidence. In our SCM, the non-interventional prediction can be expressed using Bayes rule as:

$$P(A|V,Q) = \sum_{z} P(A|V,Q,z)P(z|V,Q) \tag{3}$$

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 unblocked back-door path $\{V,Q\} \leftarrow Z \rightarrow A$. 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 the spurious correlation is eliminated. In this way, $\{V,Q\}$ and A are deconfounded and the model can learn the true causal effect $\{V,Q\} \rightarrow X \rightarrow A$. Actually, there are two techniques to calculate $P(A|do(V,Q))$, which are the back-door and front-door adjustments [21], [72], respectively. The back-door adjustment is effective when the confounder is observable. However, for the visual-linguistic question reasoning, the confounder in visual and linguistic modalities is not always observable. Thus, we propose both back-door and front-door causal intervention modules to discover the causal structure and disentangle the linguistic and visual biases based on their characteristics.

3.3.1 Linguistic Back-door Causal Intervention

For linguistic modality, the confounder set Z_l caused by selection bias cannot be observed directly due to the unavailability of the sampling process. Due to the existence of linguistic confounders, existing approaches that mainly rely on the entire question representations tend to capture spurious linguistic correlations and ignore semantic roles embedded in questions. To mitigate the bias caused by confounders and uncover the causal structure behind the linguistic modality, we design a back-door adjustment strategy that approximates the confounder set distribution from the perspective of linguistic semantic relations. Based on the linguistic representation learning in Section 3.2, our latent confounder set is approximated based on the verb-centered relation roles for the whole question, subject-related question, action-related question, object-related question Q, Q_s, Q_r, Q_o. Blocking the back-door path $B_l \rightarrow Z_l \rightarrow Q$ makes Q have a fair opportunity to incorporate causality-aware factors for prediction (as shown in Fig. 5 (b)). The back-door adjustment calculates the interventional distribution $P(A|V,do(Q))$:

$$\begin{aligned} P(A|V,do(Q)) &= \sum_{z_l} P(A|V,do(Q),z_l)P(z_l|V,do(Q)) \\ &\approx \sum_{z_l} P(A|V,do(Q),z_l)P(z_l) \end{aligned} \tag{4}$$

To implement the theoretical and imaginative intervention in Eq. (4), we approximate the confounder set Z_l to a set of verb-centered relation vocabularies $Z_l = [z_1, z_2, z_3, z_4] = [Q, Q_s, Q_r, Q_o]$. We compute the prior probability $P(z_l)$ in Eq. (4) for verb-centered relation phrases z in each set z_1, z_2, z_3, z_4 based on the dataset statistics:

$$P(z_l) = \frac{|z_l|}{\sum_{j \in z_l^i} |j|}, \ \forall z_l \in z_l^i, \ i = 1, \cdots, 4 \tag{5}$$

where z_l^i is one of the four verb-centered relation vocabulary sets, $|z_l|$ is the number of samples in z_l, and $|j|$ is the number of occurrences of the phrase j. The representation of z_l is calculated in a similar way as in Eq. (1). Since $P(A|V,do(Q))$ is calculated by softmax, we apply Normalized Weighted Geometric Mean (NWGM) [73] to Eq. (4) to approximate the deconfounded prediction:

$$\begin{aligned} P(A|V,do(Q)) &= \sum_{z_l} P(A|V,\text{concat}(Q,z_l))P(z_l) \\ &\approx P(A|\sum_{z_l}(V,\text{concat}(Q,z_l))P(z_l)) \end{aligned} \tag{6}$$

where concat$(\cdot)$ represents vector concatenation. According to Eq. (6), each element of the causality-aware hierarchical linguistic representation $Q^h = \{Q, Q_s, Q_r, Q_o\}$ needs to be integrated into the QA inference phase using Eq. (6), which is essentially a weighted sum of the occurrences of the values of the linguistic confounders in the dataset.

3.3.2 Visual Front-door Causal Intervention

As shown in Eq. (4), the back-door adjustment requires us to determine what the confounder is in advance. However, in visual domains, data biases are complex and it is hard to know and disentangle different types of confounders. Existing approaches usually define the confounders as the average of visual features [19], [24]. Actually, the average features may not properly describe a certain confounder especially for complex heterogeneous spatial-temporal data. Fortunately, the front-door adjustment gives a feasible way to calculate $P(A|do(V),Q)$ when we cannot explicitly represent the confounder. As shown in Fig. 5 (c), to apply the front-door adjustment, an additional mediator M should be inserted between X and A to construct a front-door path $V \rightarrow X \rightarrow M \rightarrow A$ to transmit knowledge. For visual-linguistic question reasoning task, an attention-based model will select a few regions from the video V based on the question Q to predict the answer A, where m denotes the selected knowledge from mediator M:

$$P(A|V,Q) = \sum_{m} P(M=m|V,Q)P(A|M=m) \tag{7}$$

Then, the answer predictor can be represented by two parts: a feature extractor $V \rightarrow X \rightarrow M$ and a answer predictor $M \rightarrow A$. Thus, the interventional probability $P(A|do(V),Q)$ can be represented as:

$$P(A|do(V),Q) = \sum_{m} P(M=m|do(V),Q)P(A|do(M=m)) \tag{8}$$

Next, we discuss the above feature extractor $V \rightarrow X \rightarrow M$ and answer predictor $M \rightarrow A$, respectively.

Feature Extractor $V \rightarrow X \rightarrow M$. As shown in Fig. 5 (c), for the causal link $V \rightarrow X \rightarrow M$, the back-door path between V and M: $X \leftarrow V \leftarrow Z_v \rightarrow M \rightarrow A$ is already

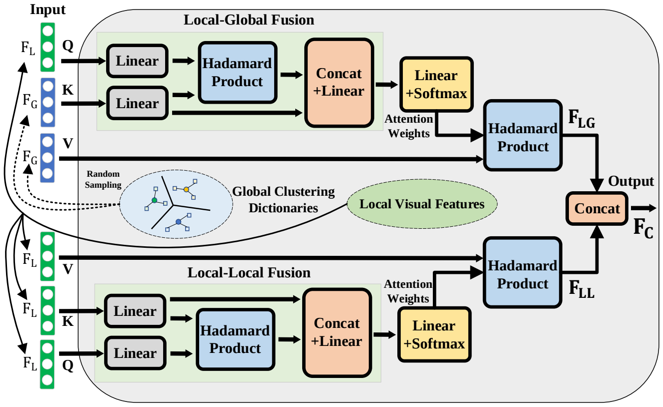


Fig. 6. The structure of Local-Global Causal Attention Module (LGCAM), which jointly estimates $\hat{M}$ and $\hat{V}$ in an unified attention module.

blocked. Thus, the interventional probability is equal to the conditional one:

$$P(M=m|do(V),Q)=P(M=m|V,Q) \tag{9}$$

Answer Predictor $M \rightarrow A$. To realize $P(A|do(M=m))$, we can cut off $M \leftarrow X$ to block the back-door path $M \leftarrow X \leftarrow V \leftarrow Z_v \rightarrow A$:

$$P(A|do(M=m))=\sum_v P(V=v)P(A|V=v,M=m) \tag{10}$$

To sum up, by applying Eq. (9) and Eq. (10) into Eq. (8), we can calculate the true causal effect between V and A:

$$\begin{aligned} &P(A|do(V),Q)= \\ &\sum_m P(M=m|V,Q)\sum_v P(V=v)P(A|V=v,M=m) \end{aligned} \tag{11}$$

To implement visual front-door causal intervention Eq. (11) 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 transformed into classification formulations:

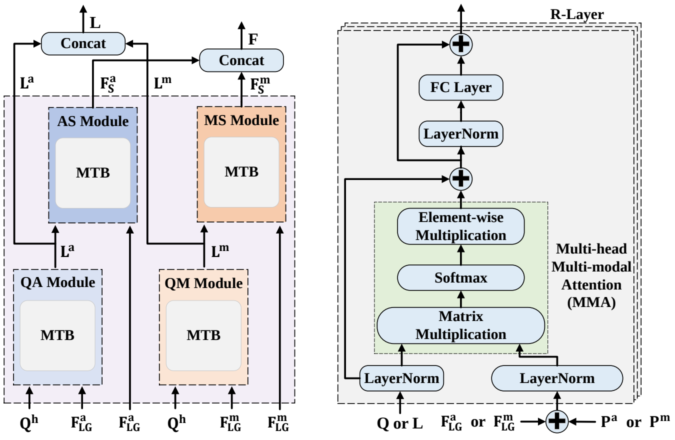
$$P(A|V,M)=\text{Softmax}[g(M,V)] \tag{12}$$

From Eq. (11), we can see that both V and M are required to be sampled and fed into the network to complete $P(A|do(V),Q)$. However, the cost of forwarding all the samples is high. To tackle this problem, we apply the Normalized Weighted Geometric Mean (NWGM) [73] to incorporate the outer sampling into the feature level, thereby requiring only one forward pass of the "absorbed input" in the network, as shown in Eq. (13):

$$\begin{aligned} &P(A|do(V),Q)\approx \text{Softmax}[g(\hat{M},\hat{V})] \\ &=\text{Softmax}\Big[g(\sum_m P(M=m|f(V))m,\sum_v P(V=v|h(V))v)\Big] \end{aligned} \tag{13}$$

where $\hat{M}$ and $\hat{V}$ denote the estimations of M and V, $h(\cdot)$ and $f(\cdot)$ denote the network mapping functions.

Actually, $\hat{M}$ is essentially an in-sample sampling process where m denotes the selected knowledge from the current input sample V, and $\hat{V}$ is essentially a cross-sample sampling process since it comes from other samples. Therefore, both $\hat{M}$ and $\hat{V}$ can be calculated by attention networks [18]. Specifically, we propose a novel Local-Global Causal Attention Module (LGCAM) that jointly estimates $\hat{M}$ and $\hat{V}$ in a unified attention module to increase the representation ability of the causality-aware visual features. $\hat{M}$ can be calculated by learning local-local visual feature F_{LL}, $\hat{V}$ can be calculated by learning local-global visual feature F_{LG}. Here, we use the computation of F_{LG} as an example to clarify our LGCAM, as shown in the upper part of Fig. 6.



(a) (b)

Fig. 7. Illustration of the (a) Spatial-Temporal Transformer (STT), and the (b) Multi-modal Transformer Block (MTB) in the STT.

Specifically, we first calculate $F_L=f(V)$ and $F_G=h(V)$ and use them as the input of the LGCAM, 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 inputs and computes local-global visual feature F_{LG} by conditioning global visual feature F_G on the local visual feature F_L. The output of the LGCAM is denoted as F_{LG}, which is given by:

$$\begin{aligned} &\textbf{Input}: Q=F_L, K=F_G, V=F_G \\ &\textbf{Local-Global Fusion}: H=[W_V V, W_Q Q\odot W_K K] \\ &\textbf{Activation Mapping}: H'=\text{GELU}(W_H H+b_H) \\ &\textbf{Attention Weights}: \alpha=\text{Softmax}(W_{H'}H'+b_{H'}) \\ &\textbf{Output}: F_{LG}=\alpha\odot F_G \end{aligned} \tag{14}$$

where $[.,.]$ denotes a concatenation operation, $\odot$ represents the Hadamard product, W_Q, W_K, W_V, $W_{H'}$ represent the weights of the linear layers, b_H and $b_{H'}$ denote the biases of the linear layers. From Fig. 3, the visual front-door causal intervention module has two branches for appearance and motion features. Therefore, the F_{LG} has two variants, F_{LG}^a for the appearance branch, and F_{LG}^m for the motion branch.

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 Spatial-Temporal Transformer

After performing linguistic and visual causal intervention, we need to conduct visual-linguistic relation modeling and feature fusion. However, existing vision-and-language transformers typically neglect the multi-level and fine-grained interaction between text and appearance-motion information, which is crucial for the event-level visual question answering task. Therefore, we propose a Spatial-Temporal Transformer (STT) that includes four submodules, namely Question-Appearance (QA), Question-Motion (QM), Appearance-Semantics (AS) and Motion-Semantics (MS), as depicted in Fig. 7 (a), to uncover the fine-grained interactions between linguistic and spatial-temporal representations. The QA (QM) module consists of an R-layer Multi-modal Transformer Block (MTB) (Fig. 7 (b)) for multi-modal interaction between the question and the appearance (motion) features. Similarly, the AS (MS) uses the MTB to deduce the appearance (motion) information given the question semantics.

The QA and AM modules aim to develop a comprehensive comprehension of the question concerning the visual appearance and motion content, respectively. For QA and QM modules, the input of MTB are $Q^h = \{Q, Q_s, Q_r, Q_o\}$ obtained from section 3.3.1 and F^a_C, F^m_C obtained from section 3.3.2, respectively. To maintain the positional information of the video sequence, the appearance feature F^a_C and motion feature F^m_C 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 $F^a_C = [F^a_C, P^a]$, $F^m_C = [F^m_C, P^m]$, Q^a, and Q^m, the multi-modal output for QA and QM are computed as:

$$\begin{aligned}
\hat{Q}^a_r &= U^a_r + \sigma^a(\mathrm{LN}(U^a_r)) \\
\hat{Q}^m_r &= U^m_r + \sigma^m(\mathrm{LN}(U^m_r)) \\
U^a_r &= \mathrm{LN}(\hat{Q}^a_{r-1}) + \mathrm{MMA}^a(\hat{Q}^a_{r-1}, F^a_C) \\
U^m_r &= \mathrm{LN}(\hat{Q}^m_{r-1}) + \mathrm{MMA}^m(\hat{Q}^m_{r-1}, F^m_C)
\end{aligned} \tag{15}$$

where $\hat{Q}^a_0 = Q^h$, $\hat{Q}^m_0 = Q^h$, U^a_r and U^m_r are the intermediate features at the r-th layer of the MTB. $\mathrm{LN}(\cdot)$ denotes the layer normalization operation and $\sigma^a(\cdot)$ and $\sigma^m(\cdot)$ denote the linear projections. $\mathrm{MMA}(\cdot)$ is the Multi-head Multi-modal Attention layer. We denote the output semantics-aware appearance and motion features of QA and MA as $L^a = \hat{Q}^a = \hat{Q}^a_R$ and $L^m = \hat{Q}^m = \hat{Q}^m_R$, respectively.

Since an essential step of VideoQA is to infer the visual information within the appearance-motion features given the question semantics, we propose the Appearance-Semantics (AS) and Motion-Semantics (MS) modules to infer the visual information from the interactions between the linguistic semantics and the spatial-temporal representations, with a similar architecture to the Multi-modal Transformer Block (MTB). Given the semantics-aware appearance and motion features L^a and L^m, we use the AS and MS to discover visual information to answer the question based on the spatial-temporal visual representations, respectively.

Similar to Eq. (15), given the visual appearance and motion features $\hat{F}^a_{LG}$, $\hat{F}^m_{LG}$ and question semantics L^a, L^m, the multi-modal output for AS and MS are computed as:

$$\begin{aligned}
\hat{L}^a_r &= U^a_r + \sigma^a(\mathrm{LN}(U^a_r)) \\
\hat{L}^m_r &= U^m_r + \sigma^m(\mathrm{LN}(U^m_r)) \\
U^a_r &= \mathrm{LN}(F^a_{C,r-1}) + \mathrm{MMA}^a(F^a_{C,r-1}, L^a) \\
U^m_r &= \mathrm{LN}(F^m_{C,r-1}) + \mathrm{MMA}^m(F^m_{C,r-1}, L^m)
\end{aligned} \tag{16}$$

where the MTB has $r = 1, 2, \ldots, R$ layers, and $F^a_{C,0} = F^a_C$, $F^m_{C,0} = F^m_C$. The output visual clues of QA and MA are denoted as $F^a_s = \hat{L}^a_R$ and $F^m_s = \hat{L}^m_R$, respectively. Then, the output of the AS and MS is concatenated to make the final visual output $F = [F^a_s, F^m_s] \in \mathbb{R}^{2d}$. 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 Visual-Linguistic Feature Fusion

According to Eq. (6) in section 3.4.1, each item of the causality-aware hierarchical linguistic representation $Q^h = \{Q, Q_s, Q_r, Q_o\}$ is required to conduct the QA prediction process respectively, and then integrate their results by their semantic relations. Thus, for Q, Q_s, Q_r, Q_o, their respective visual and linguistic outputs of the STT model are denoted as F, F_s, F_r, F_o and L, L_s, L_r, L_o, respectively. Specifically, a semantic graph is constructed, and the representation of the graph nodes is denoted as $L_g = \{L, L_s, L_r, L_o\}$, as shown in Fig. 8. The feature vectors in L_g are treated as the nodes. According to the hierarchical linguistic semantic relations among Q, Q_s, Q_r and Q_o learned by the HSRP, we build the fully-connected edges and then perform g-layer semantic graph convolutional (GCN) [16] embedding:

$$L^e_g = \mathrm{GCN}(L_g) = \{L^e, L^e_s, L^e_r, L^e_o\} \tag{17}$$

where $\mathrm{GCN}(\cdot)$ denotes the g-layer graph convolutions.

As the linguistic features from different semantic roles are correlated, we have built an adaptive linguistic feature fusion module that receives features from different semantic roles and learns a global context embedding. This embedding is then used to recalibrate the input features from different semantic roles, as shown in Fig. 8. The linguistic features of nodes learned from semantic GCN are denoted as $\{L^e_1, L^e_2, L^e_3, L^e_4\} = \{L^e, L^e_s, L^e_r, L^e_o\}$, where $L^e_k \in \mathbb{R}^{2d}(k = 1, \cdots, 4)$. To leverage the correlation among linguistic features, we concatenate them and obtain joint representations G^k_u for each semantic role L^e_k by passing them through a fully-connected layer:

$$G^k_u = W^k_s[L^e_1, L^e_2, L^e_3, L^e_4] + b^k_s, \ \ k = 1, \cdots, 4 \tag{18}$$

where $[\cdot, \cdot]$ denotes the concatenation operation, $G^k_u \in \mathbb{R}^{d_u}$ denotes the joint representation, W^k_s and b^k_s are weights and bias of the fully-connected layer. We choose $d_u = d$ to restrict the model capacity and increase its generalization ability. To utilize the global context information aggregated in the joint representations G^k_u, we predict an excitation signal for them via a fully-connected layer:

$$E^k = W^k_e G^k_u + b^k_e, \ \ k = 1, \cdots, 4 \tag{19}$$

where W^k_e and b^k_e are the weights and biases of the fully-connected layer. After obtaining the excitation signal $E^k \in \mathbb{R}^c$, we use it to adaptively recalibrate the input feature L^e_k by a simple gating mechanism:

$$\widetilde{L}^e_k = \delta(E^k) \odot L^e_k \tag{20}$$

where $\odot$ is a channel-wise product operation for each element in the channel dimension, and $\delta(\cdot)$ is the ReLU function. In this way, we can allow the features of one semantic role to recalibrate the features of another semantic

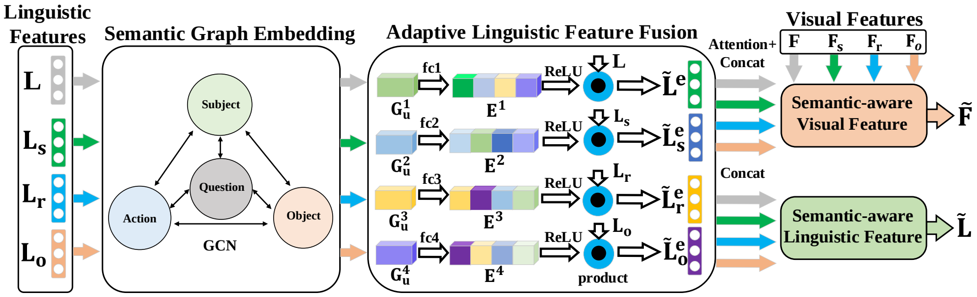


Fig. 8. Illustration of the Visual-Linguistic Feature Fusion (VLFF) module, which leverages the hierarchical linguistic semantic relations to learn the global semantic-aware visual-linguistic features, and finally fuses the causality-aware visual and linguistic features adaptively.

role while preserving the correlation among different semantic roles. Then, these refined linguistic feature vectors $\{\widetilde{L}^e, \widetilde{L}_s^e, \widetilde{L}_r^e, \widetilde{L}_o^e\}$ are concatenated to form the final semantic-ware linguistic feature $\widetilde{L} = [\widetilde{L}^e, \widetilde{L}_s^e, \widetilde{L}_r^e, \widetilde{L}_o^e] \in \mathbb{R}^{4d}$.

To obtain the semantic-aware visual feature, we compute the visual feature $\widetilde{F}_k$ by individually conditioning each semantic role from the visual features $\{F_1, F_2, F_3, F_4\} = \{F, F_s, F_r, F_o\}$ to each semantic role from the refined linguistic features $\{\widetilde{L}_1^e, \widetilde{L}_2^e, \widetilde{L}_3^e, \widetilde{L}_4^e\} = \{\widetilde{L}^e, \widetilde{L}_s^e, \widetilde{L}_r^e, \widetilde{L}_o^e\}$ using the same operation as in [12]. For each semantic role k ($k = 1, 2, 3, 4$), the weighted semantic-aware visual feature is:

$$\begin{aligned} I_k &= \mathrm{ELU}(W_k^I [W_k^f F_k, W_k^f F_k \odot W_k^l \widetilde{L}_k^e] + b_k^I) \\ \widetilde{F}_k &= \mathrm{Softmax}(W_k^{I'} I_k + b_k^{I'}) \odot F_k \end{aligned} \tag{21}$$

Then, these semantic-aware visual features $\widetilde{F}_k$ ($k = 1, \cdots, 4$) are concatenated to form the final semantic-aware visual feature $\widetilde{F} = [\widetilde{F}_1, \widetilde{F}_2, \widetilde{F}_3, \widetilde{F}_4] \in \mathbb{R}^{4d}$. Finally, we infer the answer based on the semantic-aware visual feature $\widetilde{F}$ and linguistic feature $\widetilde{L}$. Specifically, we apply different answer decoders [12] depending on the visual question reasoning tasks, which are divided into three types: open-ended, multi-choice, and counting.

4 Experiments

In this section, we conduct extensive experiments to evaluate the performance of our CMCIR model. To verify the effectiveness of CMCIR and its components, we compare CMCIR with state-of-the-art methods and conduct ablation studies. Then, we conduct parameter sensitivity analysis to evaluate how the hyper-parameters of CMCIR affect the performance. We further show some visualization analysis to validate the ability of causal reasoning of CMCIR.

4.1 Datasets

In this paper, we evaluate our CMCIR on the event-level urban dataset SUTD-TrafficQA [36] and three benchmark real-world datasets TGIF-QA [29], MSVD-QA [28], and MSRVTT-QA [28]. The detailed descriptions of these datasets are as follows:

SUTD-TrafficQA. This dataset consists of 62,535 QA pairs and 10,090 videos collected from traffic scenes. There are six challenging reasoning tasks including basic understanding, event forecasting, reverse reasoning, counterfactual inference, introspection and attribution analysis. The basic understanding task is to perceive and understand traffic scenarios at the basic level. The event forecasting task is to infer future events based on observed videos, and the forecasting questions query about the outcome of the current situation. The reverse reasoning task is to ask about the events that have happened before the start of a video. The counterfactual inference task queries the consequent outcomes of certain hypothesis that do not occur. The introspection task is to test if models can provide preventive advice that could have been taken to avoid traffic accidents. The attribution task seeks the explanation about the causes of traffic events and infer the underlying factors.

	Video	QA pairs	Count	Action	Transition	FrameQA
Train	62,846	139,414	26,843	20,475	52,704	39,392
Test	9,575	25,751	3,554	2,274	6,232	13,691
Total	71,741	165,165	30,397	22,749	58,936	53,083

TABLE 1
Statistics of the TGIF-QA dataset.

	Video	QA pairs	What	Who	How	When	Where
Train	1,200	30,933	19,485	10,479	736	161	72
Val	250	6,415	3,995	2,168	185	51	16
Test	520	13,157	8,149	4,552	370	58	28
Total	1,970	50,505	31,629	17,199	1,291	270	116

TABLE 2
Statistics of the MSVD-QA dataset.

	Video	QA pairs	What	Who	How	When	Where
Train	6,513	158,581	108,792	43,592	4,067	1,626	504
Val	497	12,278	8,337	3,439	344	106	52
Test	2,990	72,821	49,869	20,385	1,640	677	250
Total	10,000	243,680	166,998	67,416	6,051	2,409	806

TABLE 3
Statistics of the MSRVTT-QA dataset.

TGIF-QA. This dataset has 165K QA pairs collected from 72K animated GIFs. It has four tasks: repetition count, repeating action, state transition, and frame QA. Repetition count is a counting task that requires a model to count the number of repetitions of an action. Repetition action and state transition are multiple-choice tasks with 5 optional answers. FrameQA is an open-ended task with a predefined answer set, which can be answered from a single video frame. Table 1 shows the statistics of the TGIF-QA dataset.

MSVD-QA. This dataset is created from the Microsoft Research Video Description Corpus [81], which is widely used in the video captioning task. It consists of 50,505 algorithm-generated question-answer pairs and 1,970 trimmed video clips. Each video lasts approximately 10 seconds. It contains five questions types: What, Who, How, When, and Where. The statistics of the MSVD-QA dataset are presented in Table 2.

MSRVTT-QA. This larger dataset contains more complex scenes constructed from the MSRVTT [82]. It contains 10,000 trimmed video clips of approximately 15 seconds

Method	Question Type						
	Basic (4759)	Attribution (348)	Introspection (482)	Counterfactual (302)	Forecasting (166)	Reverse (565)	All (6622)
VIS+LSTM [74]	-	-	-	-	-	-	29.91
I3D+LSTM [75]	-	-	-	-	-	-	33.21
BERT-VQA [76]	-	-	-	-	-	-	33.68
TVQA [77]	-	-	-	-	-	-	35.16
VQAC† [78]	34.02	49.43	34.44	39.74	38.55	49.73	36.00
MASN† [79]	33.83	50.86	34.23	41.06	41.57	50.80	36.03
DualVGR† [80]	33.91	50.57	33.40	41.39	41.57	50.62	36.07
HCRN [12]	-	-	-	-	-	-	36.49
HCRN† [12]	34.17	50.29	33.40	40.73	44.58	50.09	36.26
Eclipse [36]	-	-	-	-	-	-	37.05
CMCIR (ours)	**36.10** (+1.93)	**52.59** (+1.73)	**38.38** (+3.94)	**46.03** (+4.64)	**48.80** (+4.22)	**52.21** (+1.41)	**38.58** (+1.53)

TABLE 4
Results on SUTD-TrafficQA dataset. '†' indicates the result re-implemented by the officially code. The **best** and second-best results are highlighted.

each. A total of 243,680 question-answer pairs contained in this dataset are automatically generated by the NLP algorithm. The dataset contains five question types: What, Who, How, When, and Where. The statistics of the MSRVTT-QA dataset are presented in Table 3.

4.2 Implementation Details

For fair comparisons with other methods, we follow [12] to divide the videos into 8 clips for the SUTD-TrafficQA and TGIF-QA datasets, and 24 clips for the MSVD-QA and MSRVTT-QA datasets that contain long videos. The Swin-L [67] pretrained on ImageNet-22K dataset is used to extract the frame-level appearance features, and the video Swin-B [83] pretrained on Kinetics-600 is applied to extract the clip-level motion features. For the question, we adopt the pre-trained 300-dimensional GloVe [70] 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. The number of GCN layers g is set to 1 in the semantic graph embedding. In the training process, we train the model using the Adam optimizer with an initial learning rate 2e-4, 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. The dropout rate is set to 0.15 to prevent overfitting. All experiments are terminated after 50 epochs. We implement our model by PyTorch with an NVIDIA RTX 3090 GPU. For multi-choice and open-ended tasks, we use the accuracy to evaluate the performance of our model. For the counting task in TGIF-QA dataset, we adopt the Mean Squared Error (MSE) between the predicted answer and the right answer.

4.3 Comparison With State-of-the-Art Methods

4.3.1 Results on SUTD-TrafficQA Dataset

Since the splits of six reasoning tasks are not provided by the original SUTD-TrafficQA dataset [36], we divide the SUTD-TrafficQA dataset into six reasoning tasks according to the question types. The overall accuracy and the accuracy of each reasoning types are reported.

The results in Table 4 demonstrate that our CMCIR achieves the best performance for six reasoning tasks including basic understanding, event forecasting, reverse reasoning, counterfactual inference, introspection and attribution analysis. Specifically, the CMCIR improves the best state-of-the-art method Eclipse [36] by 1.53% for all reasoning tasks. Compared with the re-implemented methods VQAC†, MASN†, DualVGR†, and HCRN†, our CMCIR performs better than these methods in all six tasks by a significant margin. For example, compared with HCRN†, our CMCIR improves the accuracy by 1.93% for basic understanding, 2.30% for attribution analysis, 4.98% for introspection, 5.30% for counterfactual inference, 4.22% for event forecasting, 2.12% for reverse reasoning, and 2.32% for all tasks. It is obvious that our method improves three types of questions the most: introspection, counterfactual inference, and event forecasting. The introspection task is to test if models can provide preventive advice that could have been taken to prevent traffic accidents. The event forecasting task is to infer future events based on observed videos, and the forecasting questions inquire about the outcome of the current situation. The counterfactual inference task queries the consequent outcomes of certain hypotheses that did not occur. All of these three question types require causal relational reasoning among the causal, logic, and spatial-temporal structures of the visual and linguistic content. This validates that our CMCIR can model multi-level interaction and causal relations between the language and spatial-temporal structure of the event-level urban data.

4.3.2 Results on Other Benchmark Datasets

To evaluate the generalization ability of CMCIR on other event-level datasets, we conduct experiments on TGIF-QA, MSVD-QA, and MSRVTT-QA datasets and compare our model with the state-of-the-art methods. The comparison results on the TGIF-QA dataset are presented in Table 5. We can see that our CMCIR achieves the best performance for the *Action* and *FrameQA* tasks. Additionally, our CMCIR also achieves relatively high performance for the *Transition* and *Count* tasks. Specifically, the CMCIR improves the best performing method HAIR [35] by 0.3% for the *Action* task, 2.1% for the *FrameQA* task. For the *Transition* task, the CMCIR also outperforms other comparison methods except CASSG [84] and Bridge2Answer [13]. For the *Count* task, our CMCIR also achieves a competitive MSE loss value.

Method	Task Type			
	Action↑	Transition↑	FrameQA↑	Count↓
ST-VQA [29]	62.9	69.4	49.5	4.32
Co-Mem [85]	68.2	74.3	51.5	4.10
PSAC [11]	70.4	76.9	55.7	4.27
HME [31]	73.9	77.8	53.8	4.02
GMIN [86]	73.0	81.7	57.5	4.16
L-GCN [10]	74.3	81.1	56.3	3.95
HCRN [12]	75.0	81.4	55.9	3.82
HGA [33]	75.4	81.0	55.1	4.09
QueST [87]	75.9	81.0	59.7	4.19
Bridge2Answer [13]	75.9	82.6	57.5	**3.71**
QESAL [88]	76.1	82.0	57.8	3.95
ASTG [89]	76.3	82.1	61.2	3.78
CASSG [84]	77.6	**83.7**	58.7	3.83
HAIR [35]	77.8	82.3	60.2	3.88
CMCIR (ours)	**78.1**	82.4	**62.3**	3.83

TABLE 5
Comparison with state-of-the-art methods on TGIF-QA dataset.

Method	Question Type					
	What	Who	How	When	Where	All
	(8,149)	(4,552)	(370)	(58)	(28)	(13,157)
Co-Mem [85]	19.6	48.7	81.6	74.1	31.7	31.7
AMU [28]	20.6	47.5	83.5	72.4	**53.6**	32.0
HME [31]	22.4	50.1	73.0	70.7	42.9	33.7
HRA [90]	-	-	-	-	-	34.4
HGA [33]	23.5	50.4	83.0	72.4	46.4	34.7
GMIN [86]	24.8	49.9	84.1	75.9	**53.6**	35.4
QueST [87]	24.5	52.9	79.1	72.4	50.0	36.1
HCRN [12]	-	-	-	-	-	36.1
CASSG [84]	24.9	52.7	**84.4**	74.1	**53.6**	36.5
QESAL [88]	25.8	51.7	83.0	72.4	50.0	36.6
Bridge2Answer [13]	-	-	-	-	-	37.2
HAIR [35]	-	-	-	-	-	37.5
VQAC [78]	26.9	53.6	-	-	-	37.8
MASN [79]	-	-	-	-	-	38.0
HRNAT [91]	-	-	-	-	-	38.2
ASTG [89]	26.3	55.3	82.4	72.4	50.0	38.2
DualVGR [80]	28.6	53.8	80.0	70.6	46.4	39.0
CMCIR (ours)	**33.1**	**58.9**	84.3	**77.5**	42.8	**43.7**

TABLE 6
Comparison with state-of-the-art methods on MSVD-QA dataset.

Method	Question Type					
	What	Who	How	When	Where	All
	(49,869)	(20,385)	(1,640)	(677)	(250)	(72,821)
Co-Mem [85]	23.9	42.5	74.1	69.0	42.9	31.9
AMU [28]	26.2	43.0	80.2	72.5	30.0	32.5
HME [31]	26.5	43.6	82.4	76.0	28.6	33.0
QueST [87]	27.9	45.6	83.0	75.7	31.6	34.6
HRA [90]	-	-	-	-	-	35.0
MASN [79]	-	-	-	-	-	35.2
HRNAT [91]	-	-	-	-	-	35.3
HGA [33]	29.2	45.7	83.5	75.2	34.0	35.5
DualVGR [80]	29.4	45.5	79.7	76.6	36.4	35.5
HCRN [12]	-	-	-	-	-	35.6
VQAC [78]	29.1	46.5	-	-	-	35.7
CASSG [84]	29.8	46.3	**84.9**	75.2	35.6	36.1
GMIN [86]	30.2	45.4	84.1	74.9	**43.2**	36.1
QESAL [88]	30.7	46.0	82.4	76.1	41.6	36.7
Bridge2Answer [13]	-	-	-	-	-	36.9
HAIR [35]	-	-	-	-	-	36.9
ClipBERT [34]	-	-	-	-	-	37.4
ASTG [89]	31.1	48.5	83.1	77.7	38.0	37.6
CMCIR (ours)	**32.2**	**50.2**	82.3	**78.4**	38.0	**38.9**

TABLE 7
Comparison with state-of-the-art methods on MSRVTT-QA dataset.

Table 6 shows the comparison results on the MSVD-QA dataset. From the results, we can see that our CMCIR outperforms nearly all the state-of-the-art comparison methods by a significant margin. For example, our CMCIR achieves the best overall accuracy of 43.7%, which leads to 4.7% improvement over the best performing method DualVGR [80]. For *What*, *Who*, and *When* types, the CMCIR significantly outperforms all the comparison methods. Although GMIN [86] and CASSG [84] perform marginally better than our CMCIR for *How* and *Where* types, our CMCIR performs significantly better than GMIN for What ($+8.3\%$), Who ($+9.0\%$), When ($+1.6\%$), and the overall ($+8.3\%$) tasks.

Table 7 shows the comparison results for the MSRVTT-QA dataset. It can be observed that our CMCIR outperforms the best performing method ASTG [89], with the highest accuracy of 38.9%. For *What*, *Who*, and *When* question types, the CMCIR performs the best compared to all the previous state-of-the-art methods. Although CASSG [84] and GMIN [86] achieve better accuracies than our CMCIR for *How* and *Where* question types respectively, our CMCIR achieves a significantly performance improvement over these two methods for other question types.

In Table 6 and Table 7, our method achieves lower performance than previous best method when the question types are *How* and *Where*. It can be seen from Table 6 and Table 7 that the number of *How* and *Where* samples are much smaller than that of the other question types. Due to the existence of data bias in these two datasets, the model tends to learn spurious correlation from other question types. This may lead to the performance degradation when testing on these two question types. Nonetheless, we can still obtain promising performance for question type *When*, which also has limited samples. This validates that our CMCIR indeed mitigate the spurious correlations for most of the question types including *What*, *Who*, and *When*.

The experimental results in Table 5-7 show that our CMCIR outperforms state-of-the-art methods on three large-scale benchmark event-level datasets. This validates that our CMCIR generalizes well across different event-level datasets, including urban traffic and real-world scenes. Our CMCIR achieves more promising performance than existing relational reasoning methods like HGA, QueST, GMIN, Bridge2Answer, QESAL, ASTG, PGAT, HAIR and CASSG, which validates that our CMCIR has good potential to model multi-level interaction and causal relations between the language and spatial-temporal structure of videos. The main reason for good generalization across different datasets is that our CMCIR can mitigate both the visual and linguistic biases through front-door and back-door causal intervention modules. Due to the strong multi-modal relational reasoning ability of the CMCIR, we can disentangle the spurious correlations within visual-linguistic modality and achieve robust spatial-temporal relational reasoning.

Comparing the average improvement across different datasets, we notice that CMCIR achieves the best improvement on SUTD-TrafficQA (+1.53%), MSVD-QA (+4.7%) while relatively moderate gains on TGIF-QA (+0.3%∼0.9%) and MSRVTT-QA (+1.3%). The reason for such discrepancy is that SUTD-TrafficQA and MSVD-QA are relatively small in size, which constrains the reasoning ability of the backbone models by limiting their exposure to training instances. As a comparison, SUTD-TrafficQA is four times smaller than MSRVTT-QA in terms of QA pairs (60K vs 243K), while MSVD-QA is five times smaller than MSRVTT-QA in terms of QA pairs (43K vs 243K). However, such deficiency caters to the focal point of our CMCIR, which

Datasets	CMCIR w/o HSRP	CMCIR w/o LBCI	CMCIR w/o VFCI	CMCIR w/o CVLR	CMCIR w/o SGE	CMCIR w/o ALFF	CMCIR
SUTD	37.65	37.71	37.68	37.42	37.93	37.84	**38.58**
TGIF (Action)	75.4	75.1	75.5	75.0	75.4	75.2	**78.1**
TGIF (Transition)	81.2	81.3	80.6	80.4	81.0	81.2	**82.4**
TGIF (FrameQA)	62.0	61.9	61.6	61.2	61.3	61.1	**62.3**
TGIF (Count)	4.03	3.89	4.10	4.05	3.91	4.12	**3.83**
MSVD	42.4	42.7	42.2	42.0	42.9	42.5	**43.7**
MSRVTT	38.5	38.3	38.1	38.0	38.2	38.4	**38.9**

TABLE 8
Ablation study on SUTD-TrafficQA, TGIF-QA, MSVD-QA, and MSRVTT-QA datasets.

develops better in a less generalized situation, thus leading to more preferable growth on MSVD-QA. This validates that our causality-aware visual-linguistic representation has good generalization ability.

4.4 Ablation Studies

We further conduct ablation experiments using the following variants of CMCIR to verify the contributions of the components designed in out method.

- CMCIR w/o HSRP: we remove the Hierarchical Semantic-Role Parser (HSRP), which parses the question into verb-centered relation tuples (subject, relation, object). The CMCIR model only uses the original question as the linguistic representation.
- CMCIR w/o LBCI: we remove the Linguistic Back-door Causal Intervention (LBCI) module. The CVLR module only contains visual front-door causal intervention (VFCI) module.
- CMCIR w/o VFCI: we remove the Visual Front-door Causal Intervention (VFCI) module. The CVLR module only contains linguistic back-door causal intervention (LBCI) module.
- CMCIR w/o CVLR: we remove the Causality-aware Visual-Linguistic Reasoning (CVLR) module. The CMCIR model combines the visual and linguistic representations using spatial-temporal transformer (STT) and visual-linguistic feature fusion modules.
- CMCIR w/o SGE: we remove the Semantic Graph Embedding (SGE) module when conducting visual-linguistic feature fusion. The linguistic features are directly used for adaptive linguistic feature fusion.
- CMCIR w/o ALFF: we remove the Adaptive Linguistic Feature Fusion (ALFF) module when conducting visual-linguistic feature fusion. The semantic graph embedded linguistic features are directly used to fused with the visual features.

Table 8 shows the evaluation results of the ablation study on SUTD-TrafficQA, TGIF-QA, MSVD-QA, and MSRVTT-QA datasets. It can be observed that our CMCIR achieves the best performance compared to the six variants across all datasets and tasks. Without HSRP, the performance drops significantly due to the lack of the hierarchical linguistic feature representation. This shows that our proposed hierarchical semantic-role parser indeed increase the representation ability of question semantics. To be noticed, the performance of CMCIR w/o LBCI, CMCIR w/o VFCI, and CMCIR w/o CVLR are all lower than that of the CMCIR. This validates that both the linguistic back-door and visual front-door causal interventions contribute to discover the causal structures and learn the causality-aware visual-linguistic representations, and thus improve the model performance. For CMCIR w/o SGE and CMCIR w/o ALFF, their performance are higher than that of the CMCIR w/o LBCI, CMCIR w/o VFCI, and CMCIR w/o CVLR, but lower than that of our CMCIR, which indicates effectiveness of semantic graph embedding and adaptive linguistic feature fusion that leverages the hierarchical linguistic semantic relations as the guidance to adaptively learn the global semantic-aware visual-linguistic representations. With all the components, our CMCIR performs the best because all these components are beneficial and work collaboratively to achieve robust event-level visual question answering.

4.5 Parameter Sensitivity

To evaluate how the hyper-parameters of CMCIR affect the performance, we report the results of different values of the heads h of the Multi-head Multi-modal Attention (MMA) module, the layers r of Multi-modal Transformer Block (MTB), and GCN layers g in the semantic graph embedding. Moreover, the dimension of hidden states d is also analyzed. The results for the SUTD-TrafficQA, TGIF-QA, MSVD-QA, and MSRVTT-QA datasets are shown in Table 9. We can see that the performance of CMCIR with 8 MMA heads performs the best across all datasets and tasks compared to CMCIR with fewer MMA heads. This indicates that more heads can facilitate the MMA module to employ more perspectives to explore the relations between different modalities. For MTB layers, the optimal layer numbers are different for different datasets. The performance of the CMCIR is the best when the number of MTB layers is 3 on the SUTD-TrafficQA dataset, 8 on TGIF-QA dataset, 5 on the MSVD-QA dataset, and 6 on the MSRVTT-QA dataset. For GCN layers, we can see that more GCN layers will increase the amount of learnable parameters and thus make model converge more difficultly. Since one GCN layer can achieve the best performance, we choose one-layer GCN. For the dimension of hidden states, we can see that 512 is the best dimensionality of hidden states of the VLICR model due to its good compromise between feature representation ability and model complexity.

To validate whether our CMCIR can generalize to different visual appearance and motion features, we evaluate the performance of the CMCIR on the SUTD-TrafficQA, MSVD-QA and MSRVTT-QA datasets using different visual appearance and motion features, as shown in Table 10. The best performing comparison methods on the SUTD-TrafficQA, MSVD-QA and MSRVTT-QA datasets are also shown in Table 10. It can be observed that when using Swin-L and Video Swin-B as the visual and motion features, our CMCIR can achieves the state-of-the-art performance compared with other methods. In our experiments, visual appearance features are the pool5 output of ResNet-101 [92] and visual motion features are derived by ResNetXt-101 [93], [94]. When using ResNet-101 and ResNetXt-101 as the visual and motion features, our CMCIR can also achieve competitive accuracy on SUTD-TrafficQA, MSVD-QA and MSRVTT-QA datasets. For SUTD-TrafficQA dataset, the performance of using ResNet and ResNetXt is 38.10%, which is

		SUTD-TrafficQA	TGIF-QA (Action)	TGIF-QA (Transisition)	TGIF-QA (FrameQA)	TGIF-QA (Count)	MSVD-QA	MSRVTT-QA
MMA Heads	1	37.83	75.8	80.7	61.2	3.92	42.3	38.5
	2	38.17	75.7	79.7	60.6	3.96	42.0	38.5
	4	37.51	75.8	79.2	61.1	3.93	42.2	38.3
	8	**38.58**	**78.1**	**82.4**	**62.3**	**3.83**	**43.2**	**38.9**
MTB Layers	1	37.81	74.5	79.4	60.3	4.26	42.9	38.7
	2	37.98	74.8	80.4	61.0	4.20	42.8	38.2
	3	**38.58**	75.1	80.1	61.0	4.03	43.0	38.4
	4	37.84	76.6	80.2	61.6	3.96	42.6	38.7
	5	37.63	75.5	80.6	61.0	3.94	**43.7**	38.7
	6	37.73	76.2	80.8	61.4	4.12	43.2	**38.9**
	7	37.73	75.4	80.3	61.2	3.98	43.1	38.3
	8	37.58	**78.1**	**82.4**	**62.3**	**3.83**	42.8	38.6
GCN Layers	1	**38.58**	**78.1**	**82.4**	**62.3**	**3.83**	**43.2**	**38.9**
	2	37.84	74.9	80.3	61.0	4.07	41.8	38.3
	3	37.58	74.7	80.3	60.8	4.03	42.1	38.4
Dimension	256	37.60	73.9	79.9	61.0	3.96	42.8	38.8
	512	**38.58**	**78.1**	**82.4**	**62.3**	**3.83**	**43.2**	**38.9**
	768	37.74	75.0	80.0	62.2	3.90	42.8	38.0

TABLE 9
Performance of CMCIR with different values of MMA heads, MTB layers, GCN layers, and hidden state dimension on the SUTD-TrafficQA, TGIF-QA, MSVD-QA, and MSRVTT-QA datasets.

	Method	Appearance	Motion	Accuracy
SUTD-QA	Eclipse [36]	ResNet-101	MobileNetV2	37.05
	Ours	Swin-L	Video Swin-B	**38.58** (+1.54)
	Ours	ResNet-101	ResNetXt-101	<u>38.10</u> (+1.05)
MSVD-QA	DualVGR [80]	ResNet-101	ResNetXt-101	39.0
	Ours	Swin-L	Video Swin-B	**43.7** (+4.70)
	Ours	ResNet-101	ResNetXt-101	<u>40.3</u> (+1.30)
MSRVTT-QA	HCRN [12]	ResNet-101	ResNeXt-101	35.6
	Ours	Swin-L	Video Swin-B	**38.9** (+3.30)
	Ours	ResNet-101	ResNeXt-101	<u>37.0</u> (+1.40)

TABLE 10
Performance of CMCIR with different visual appearance and motion features on SUTD-TrafficQA, MSVD-QA, and MSRVTT-QA datasets.

Models	SUTD-TrafficQA	MSVD-QA	MSRVTT-QA
Co-Mem [85]	35.10	34.6	35.3
Co-Mem [85]+ CVLR	**37.12** (+2.02)	**40.7** (+6.1)	**38.0** (+2.7)
HGA [33]	35.81	35.4	36.1
HGA [33]+ CVLR	**37.23** (+1.42)	**41.9** (+6.5)	**38.2** (+2.1)
HCRN [12]	36.49	36.1	35.6
HCRN [12]+ CVLR	**37.54** (+1.05)	**42.2** (+6.1)	**37.8** (+2.2)
Our Backbone	37.42	42.0	38.0
Our Backbone + CVLR	**38.58** (+1.16)	**43.7** (+1.7)	**38.9** (+0.9)

TABLE 11
The CVLR module is applied to different existing non-causal models.

the also the best accuracy among all the comparison methods (Table 4). For the MSVD-QA dataset, the performance of using ResNet-101 and ResNetXt-101 is 40.3%, which also outperforms other comparison methods (Table 6). For the MSRVTT-QA dataset, the performance of using ResNet-101 and ResNetXt-101 is 37.0%, which also achieves competitive performance compared to other comparison methods (Table 6). These results validates that our CMCIR generalizes well across different visual appearance and motion features due to the learned causality-aware visual-linguistic representations. More importantly, the performance improvement of our CMCIR is mainly attributed to our elaborately designed visual-linguistic causal reasoning model.

4.6 The Evidence of Reducing Spurious Correlations

Actually, the process of building VideoQA datasets will introduce undesirable spurious correlations rather than the overarching reality [46]. Therefore, we can assume that all our evaluation datasets contain spurious correlations. To validate the effectiveness of the CVLR module in reducing spurious correlations in non-causal frameworks, we apply the CVLR to three state-of-the-art models Co-Mem [85], HGA [33] and HCRN [12]. Since our CVLR is orthogonal to the backbone, we can insert the CVLR directly after the feature extraction layers of these models, which is the same as our CMCIR. As shown in Table 11, our CVLR brings each backbone a sharp gain across all benchmark datasets (+0.9%∼6.5%), which evidences its model-agnostic property. Nevertheless, we notice that the improvements fluctuate across the backbones. As a comparison, on MSVD-QA and MSRVTT-QA benchmarks, CVLR acquires more favorable gains with backbones Co-Mem, HGA and HCRN than it does with our backbone. This is because the fine-grained interactions between linguistic semantics and spatial-temporal representations empower our backbone with robustness, especially to questions of the descriptive type on MSVD-QA and MSRVTT-QA benchmarks. Therefore, it achieves stronger backbone performances on benchmarks that focus on the descriptive question (i.e., MSVD-QA and MSRVTT-QA), which, in turn, accounts for the contribution of CVLR to some extent, thus makes improvement of our backbone less remarkable. In contrast, when it comes to the causal and temporal question (i.e., SUTD-TrafficQA), the CVLR shows equivalent improvements on all four backbones (+1.05%∼2.02%). These results validate that our CVLR is effective in capturing the causality and reducing the spurious correlations across different models.

4.7 Qualitative Results

To verify the ability of the CMCIR in robust spatial-temporal relational reasoning, we aim to gain insight into its visual-linguistic causal reasoning capabilities by inspecting some correct and failure examples from the SUTD-TrafficQA dataset and show the visualization results in Fig. 9. We show how our model conducts robust spatial-temporal relational reasoning and reduces spurious correlations.

Reliable reasoning. As shown in Fig. 9 (a), there exists an ambiguity problem where the dominant visual regions of the accident may be distracted by other visual concepts (i.e. different cars/vehicles on the road). In our CMCIR, we learn the question-relevant visual-linguistic association by

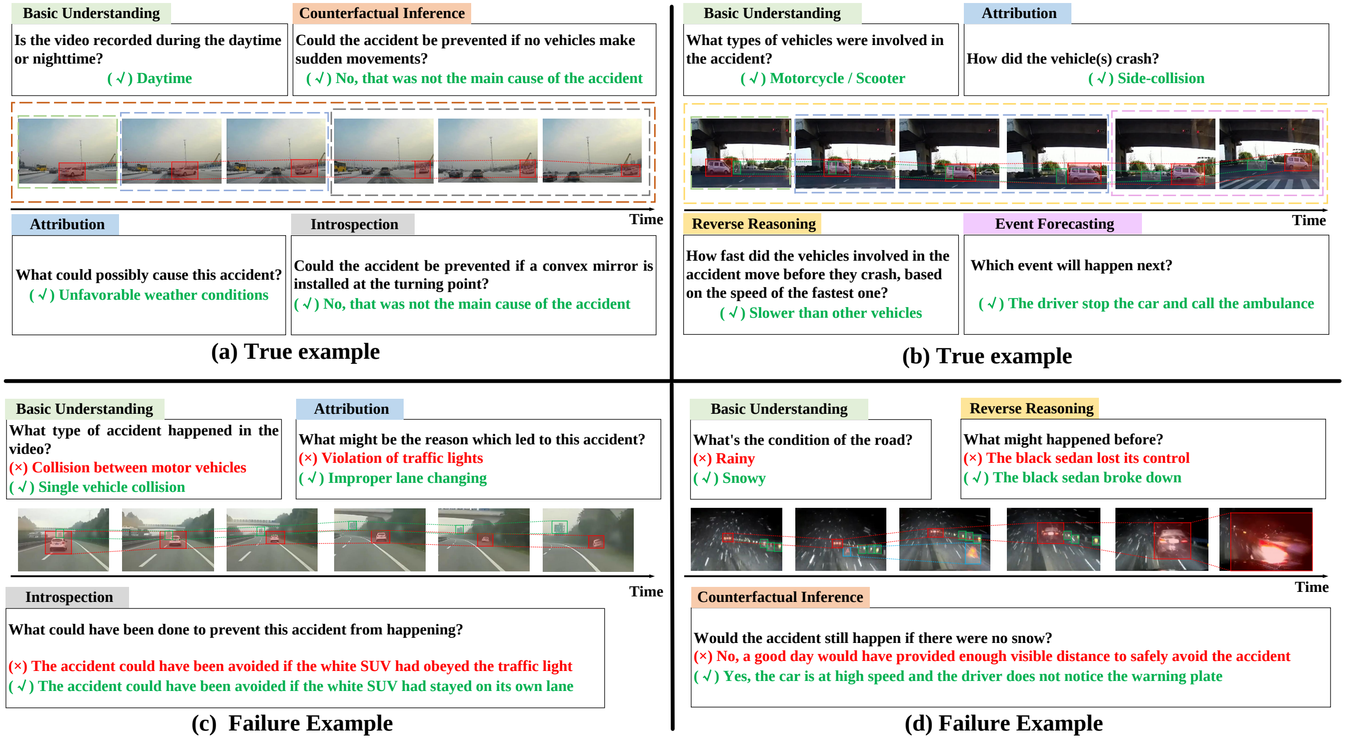

Basic Understanding
Is the video recorded during the daytime or nighttime?
(√) Daytime
Counterfactual Inference
Could the accident be prevented if no vehicles make sudden movements?
(√) No, that was not the main cause of the accident
Time
Attribution
What could possibly cause this accident?
(√) Unfavorable weather conditions
Introspection
Could the accident be prevented if a convex mirror is installed at the turning point?
(√) No, that was not the main cause of the accident
(a) True example
Basic Understanding
What types of vehicles were involved in the accident?
(√) Motorcycle / Scooter
Attribution
How did the vehicle(s) crash?
(√) Side-collision
Time
Reverse Reasoning
How fast did the vehicles involved in the accident move before they crash, based on the speed of the fastest one?
(√) Slower than other vehicles
Event Forecasting
Which event will happen next?
(√) The driver stop the car and call the ambulance
(b) True example
Basic Understanding
What type of accident happened in the video?
(×) Collision between motor vehicles
(√) Single vehicle collision
Attribution
What might be the reason which led to this accident?
(×) Violation of traffic lights
(√) Improper lane changing
Time
Introspection
What could have been done to prevent this accident from happening?
(×) The accident could have been avoided if the white SUV had obeyed the traffic light
(√) The accident could have been avoided if the white SUV had stayed on its own lane
(c) Failure Example
Basic Understanding
What's the condition of the road?
(×) Rainy
(√) Snowy
Reverse Reasoning
What might happened before?
(×) The black sedan lost its control
(√) The black sedan broke down
Time
Counterfactual Inference
Would the accident still happen if there were no snow?
(×) No, a good day would have provided enough visible distance to safely avoid the accident
(√) Yes, the car is at high speed and the driver does not notice the warning plate
(d) Failure Example

Fig. 9. Visualization of visual-linguitic causal reasoning examples from the SUTD-TrafficQA dataset. Each video is accompanied by several question types that contain spurious correlations. The color windows in the videos denote the concentrated visual concepts for the inference.

causal relational learning, thus mitigating such ambiguity in our inference results where video-question-answer triplets exhibit a strong correlation between the dominant spatial-temporal scenes and the question semantics. This validates that the CMCIR can reliably focus on the correct visual regions when making decisions.

Reducing spurious correlation. In Fig. 9 (b), we present a case reflecting the spurious correlation, where the visual regions of "van" are spuriously correlated with associated with the "sedan", due to their frequent co-occurrences. In other words, the model without explicitly considering reducing spurious correlations (e.g., Co-Mem [85], HGA [33] and HCRN [12]) will hesitate when encountering the visual concepts of "van" and "motorbike" with regard to region-object correspondence. In our CMCIR, we reduce such spurious correlation and pursue the true causality by adopting visual-linguistic causal intervention, resulting in better dominant visual evidence and question intention.

Generalization ability. From Fig. 9 (a)-(b), we can see that the CMCIR can generalize well across different question types. which shows that the CMCIR is sensitive to questions and can effectively capture the dominant spatial-temporal content in the videos by conducting robust and reliable spatial-temporal relational reasoning.

Introspective and counterfactual learning. For challenging question types, such as introspection and counterfactual inference, the CMCIR model can accurately determine whether the attended scene reflects the logic behind the answer. This verifies that the CMCIR can fully explore the causal, logical, and spatial-temporal structures of the visual and linguistic content, due to its promising ability to perform robust visual-linguistic causal reasoning that disentangles visual-linguistic spurious correlations.

Additional failure cases. Moreover, we provide failure examples in Fig. 9 (c)-(d) to gain further insights into the limitations of our method. In Fig. 9 (c), our model mistakenly correlates the visual concept "suv" with the green "traffic plate" when conducting visual-linguistic reasoning. This is because the visual region of "traffic plate" looks like the "truck", while only the white "suv" exists in the video. In Fig. 9 (d), it is difficult to distinguish between "rainy" and "snowy" due to their similar visual appearance in the video. Additionally, the "reflective stripes" along the road are mistakenly considered as the dominant visual concepts. As our CMCIR model lacks an explicit object detection pipeline, some visually ambiguous concepts are challenging to determine. Moreover, without external prior knowledge of traffic rules, some questions such as "how to prevent the accident" and "the cause of the accident" are difficult to answer. One possible solution may be to incorporate object detection and external knowledge of traffic rules into our method, which we will explore in our future work.

5 Conclusion

We propose an event-level visual question answering framework named Cross-Modal Causal RelatIonal Reasoning (CMCIR), to mitigate the spurious correlations and discover the causal structures for visual-linguistic modality. To uncover causal structures for visual and linguistic modalities, we propose a Causality-aware Visual-Linguistic Reasoning (CVLR) module, which leverages front-door and back-door causal interventions to disentangle the spurious correlations between visual and linguistic modalities. Extensive experiments on the event-level urban dataset SUTD-TrafficQA and three benchmark real-world datasets TGIF-QA, MSVD-QA, and MSRVTT-QA demonstrate the effectiveness of CMCIR in discovering visual-linguistic causal structures and achieving robust event-level visual question answering. Unlike previous methods that simply eliminate either the linguistic or visual bias without considering cross-modal causality discovery, we apply front-door and back-door causal intervention modules to discover cross-modal causal structures.

We believe our work could shed light on exploring new boundaries of causal analysis in vision-language tasks (**Causal-VLReasoning**[1]). In the future, we will further explore more comprehensive causal discovery methods to discover the question-critical scene elements in event-level visual question answering, particularly in the temporal aspect. By further exploiting the fine-grained temporal consistency in videos, we may achieve a model that pursues better causality. Additionally, we can leverage object-level causal relational inference to alleviate the spurious correlations from object-centric entities. Besides, we will incorporate external expert knowledge into our intervention process. Moreover, due to the inherent unobservable nature of properties, how to quantitatively analyze spurious correlations within datasets remains a challenging problem. Thus, we will discover more intuitive and reasonable metrics to compare the effectiveness of different methods in reducing spurious correlations.

Acknowledgments

This work is supported in part by the National Key R&D Program of China under Grant No.2021ZD0111601, in part by the National Natural Science Foundation of China under Grant No.62002395 and No.61976250, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No.2023A1515011530, No.2021A1515012311 and No.2020B1515020048, and in part by the Guangzhou Science and Technology Planning Project under Grant No. 2023A04J2030.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[2] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 706–715.

[3] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 803–818.

[4] Y. Liu, K. Wang, G. Li, and L. Lin, "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 5573–5588, 2021.

[5] Y. Liu, K. Wang, L. Liu, H. Lan, and L. Lin, "Tcgl: Temporal contrastive graph for self-supervised video representation learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 1978–1993, 2022.

[6] S. Buch, C. Eyzaguirre, A. Gaidon, J. Wu, L. Fei-Fei, and J. C. Niebles, "Revisiting the" video" in video-language understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2917–2927.

[7] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual dialog," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 326–335.

[8] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. 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*, 2018, pp. 3674–3683.

[9] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6720–6731.

[10] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, and C. Gan, "Location-aware graph convolutional networks for video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 021–11 028.

[11] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan, "Beyond rnns: Positional self-attention with co-attention for video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8658–8665.

[12] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9972–9981.

[13] J. Park, J. Lee, and K. 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*, 2021, pp. 15 526–15 535.

[14] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-to-end memory networks," *Advances in Neural Information Processing Systems*, vol. 2015, pp. 2440–2448, 2015.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[16] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[17] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, "Counterfactual vqa: A cause-effect look at language bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 700–12 710.

[18] X. Yang, H. Zhang, G. Qi, and J. Cai, "Causal attention for vision-language tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9847–9857.

[19] T. Wang, C. Zhou, Q. Sun, and H. Zhang, "Causal attention for unbiased visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3091–3100.

[20] Y. Li, X. Wang, J. Xiao, W. Ji, and T.-S. Chua, "Invariant grounding for video question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2928–2937.

[21] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[22] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3716–3725.

[23] Z. Yue, H. Zhang, Q. Sun, and X.-S. Hua, "Interventional few-shot learning," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[24] T. Wang, J. Huang, H. Zhang, and Q. Sun, "Visual commonsense r-cnn," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 760–10 770.

[25] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[26] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[27] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. 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*, 2018, pp. 6077–6086.

[28] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1645–1653.

[1] Our visual-linguistic causal learning open-source framework https://github.com/HCPLab-SYSU/Causal-VLReasoning.

[29] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2758–2766.

[30] Y. Jang, Y. Song, C. D. Kim, Y. Yu, Y. Kim, and G. Kim, "Video question answering with spatio-temporal reasoning," *International Journal of Computer Vision*, vol. 127, no. 10, pp. 1385–1412, 2019.

[31] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1999–2007.

[32] C. JiayinCai, C. Shi, L. Li, Y. Cheng, and Y. Shan, "Feature augmented memory with global attention network for videoqa," in *IJCAI*, 2020, pp. 998–1004.

[33] P. Jiang and Y. Han, "Reasoning with heterogeneous graph alignment for video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 109–11 116.

[34] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. 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*, 2021, pp. 7331–7341.

[35] F. Liu, J. Liu, W. Wang, and H. Lu, "Hair: Hierarchical visual-semantic relational reasoning for video question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1698–1707.

[36] L. Xu, H. Huang, and J. 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*, 2021, pp. 9878–9888.

[37] Y. Liu, Z. Lu, J. Li, T. Yang, and C. Yao, "Global temporal representation based cnns for infrared action recognition," *IEEE Signal Processing Letters*, vol. 25, no. 6, pp. 848–852, 2018.

[38] Y. Liu, Z. Lu, J. Li, and T. Yang, "Hierarchically learned view-invariant representations for cross-view action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2416–2430, 2018.

[39] Y. Liu, Z. Lu, J. Li, T. Yang, and C. Yao, "Deep image-to-video adaptation and fusion networks for action recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 3168–3182, 2019.

[40] Y. Zhu, Y. Zhang, L. Liu, Y. Liu, G. Li, M. Mao, and L. Lin, "Hybrid-order representation learning for electricity theft detection," *IEEE Transactions on Industrial Informatics*, 2022.

[41] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 244–253.

[42] H. Huang, L. Zhou, W. Zhang, J. J. Corso, and C. Xu, "Dynamic graph modules for modeling object-object interactions in activity recognition," in *British Machine Vision Conference*, 2019.

[43] C.-Y. Ma, A. Kadav, I. Melvin, Z. Kira, G. AlRegib, and H. P. Graf, "Attend and interact: Higher-order object interactions for video understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6790–6800.

[44] E. Mavroudi, B. B. Haro, and R. Vidal, "Representation learning on visual-symbolic graphs for video understanding," in *European Conference on Computer Vision*. Springer, 2020, pp. 71–90.

[45] J. Pan, S. Chen, M. Z. Shou, Y. Liu, J. Shao, and H. Li, "Actor-context-actor relation network for spatio-temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 464–474.

[46] G. Nan, R. Qiao, Y. Xiao, J. Liu, S. Leng, H. Zhang, and W. Lu, "Interventional video grounding with dual contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2765–2775.

[47] W. Wang, J. Gao, and C. Xu, "Weakly-supervised video object grounding via causal intervention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[48] T. Wang, Y. Li, B. Kang, J. Li, J. Liew, S. Tang, S. Hoi, and J. Feng, "The devil is in classification: A simple framework for long-tail instance segmentation," in *European Conference on computer vision*. Springer, 2020, pp. 728–744.

[49] X. Yang, H. Zhang, and J. Cai, "Deconfounded image captioning: A causal retrospect," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[50] Y. Liu, Y.-S. Wei, H. Yan, G.-B. Li, and L. Lin, "Causal reasoning meets visual representation learning: A prospective study," *Machine Intelligence Research*, pp. 1–27, 2022.

[51] E. Bareinboim and J. Pearl, "Controlling selection bias in causal inference," in *Artificial Intelligence and Statistics*. PMLR, 2012, pp. 100–108.

[52] 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.

[53] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2376–2384.

[54] P. Wang and N. Vasconcelos, "Scout: Self-aware discriminant counterfactual explanations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8981–8990.

[55] L. Chen, H. Zhang, J. Xiao, X. He, S. Pu, and S.-F. Chang, "Counterfactual critic multi-agent training for scene graph generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4613–4623.

[56] Z. Fang, S. Kong, C. Fowlkes, and Y. Yang, "Modularized textual grounding for counterfactual resilience," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6378–6388.

[57] A. Kanehira, K. Takemoto, S. Inayoshi, and T. Harada, "Multimodal explanations by predicting counterfactuality in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8594–8602.

[58] E. Abbasnejad, D. Teney, A. Parvaneh, J. Shi, and A. v. d. Hengel, "Counterfactual vision and language learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 044–10 054.

[59] Y. Liu, G. Li, and L. Lin, "Causality-aware visual scene discovery for cross-modal question reasoning," *arXiv preprint arXiv:2304.08083*, 2023.

[60] W. Chen, Y. Liu, C. Wang, G. Li, J. Zhu, and L. Lin, "Visual-linguistic causal intervention for radiology report generation," *arXiv preprint arXiv:2303.09117*, 2023.

[61] K. Tang, J. Huang, and H. Zhang, "Long-tailed classification by keeping the good and removing the bad momentum causal effect," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[62] D. Zhang, H. Zhang, J. Tang, X.-S. Hua, and Q. Sun, "Causal intervention for weakly-supervised semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[63] J. Qi, Y. Niu, J. Huang, and H. Zhang, "Two causal principles for improving visual dialog," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 860–10 869.

[64] V. Agarwal, R. Shetty, and M. Fritz, "Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9690–9698.

[65] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[66] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, "Counterfactual samples synthesizing for robust visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 800–10 809.

[67] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[68] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 3202–3211.

[69] G. Stanovsky, J. Michael, L. Zettlemoyer, and I. Dagan, "Supervised open information extraction," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 885–895.

[70] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[71] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[72] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic books, 2018.

[73] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[74] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," *Advances in neural information processing systems*, vol. 28, pp. 2953–2961, 2015.

[75] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[76] Z. Yang, N. Garcia, C. Chu, M. Otani, Y. Nakashima, and H. Takemura, "Bert representations for video question answering," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1556–1565.

[77] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "Tvqa: Localized, compositional video question answering," *arXiv preprint arXiv:1809.01696*, 2018.

[78] N. Kim, S. J. Ha, and J.-W. Kang, "Video question answering using language-guided deep compressed-domain video feature," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1708–1717.

[79] A. Seo, G.-C. Kang, J. Park, and B.-T. Zhang, "Attend what you need: Motion-appearance synergistic 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*, 2021, pp. 6167–6177.

[80] J. Wang, B. Bao, and C. Xu, "Dualvgr: A dual-visual graph reasoning unit for video question answering," *IEEE Transactions on Multimedia*, 2021.

[81] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 190–200.

[82] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[83] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019.

[84] Y. Liu, X. Zhang, F. Huang, B. Zhang, and Z. Li, "Cross-attentional spatio-temporal semantic graph networks for video question answering," *IEEE Transactions on Image Processing*, vol. 31, pp. 1684–1696, 2022.

[85] J. Gao, R. Ge, K. Chen, and R. Nevatia, "Motion-appearance co-memory networks for video question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6576–6585.

[86] M. Gu, Z. Zhao, W. Jin, R. Hong, and F. Wu, "Graph-based multi-interaction network for video question answering," *IEEE Transactions on Image Processing*, vol. 30, pp. 2758–2770, 2021.

[87] J. Jiang, Z. Chen, H. Lin, X. Zhao, and Y. Gao, "Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 101–11 108.

[88] F. Liu, J. Liu, R. Hong, and H. Lu, "Question-guided erasing-based spatiotemporal attention learning for video question answering," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[89] W. Jin, Z. Zhao, X. Cao, J. Zhu, X. He, and Y. Zhuang, "Adaptive spatio-temporal graph enhanced vision-language representation for video qa," *IEEE Transactions on Image Processing*, vol. 30, pp. 5477–5489, 2021.

[90] M. I. H. Chowdhury, K. Nguyen, S. Sridharan, and C. Fookes, "Hierarchical relational attention for video question answering," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 599–603.

[91] L. Gao, Y. Lei, P. Zeng, J. Song, M. Wang, and H. T. Shen, "Hierarchical representation network with auxiliary tasks for video captioning and video question answering," *IEEE Transactions on Image Processing*, 2022.

[92] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[93] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[94] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

Yang Liu (M'21) is currently a research associate professor working at the School of Computer Science and Engineering, Sun Yat-sen University. He received his Ph.D. degree from Xidian University in 2019. His current research interests include multi-modal cognitive reasoning and causal relation discovery. He has authorized and co-authorized more than 20 papers in top-tier academic journals and conferences. He has been serving as a reviewer for numerous academic journals and conferences such as IEEE TPAMI, TIP, TNNLS, TMM, TCSVT, CVPR, ICCV, ECCV and AAAI.

Guanbin Li (M'15) is currently an associate professor in School of Computer Science and Engineering, Sun Yat-Sen University. He received his PhD degree from the University of Hong Kong in 2016. His current research interests include computer vision, image processing, and deep learning. He is a recipient of ICCV 2019 Best Paper Nomination Award. He has authorized and co-authorized on more than 100 papers in top-tier academic journals and conferences. He serves as an area chair for the conference of VISAPP. He has been serving as a reviewer for numerous academic journals and conferences such as TPAMI, IJCV, TIP, TMM, TCyb, CVPR, ICCV, ECCV and NeurIPS.

Liang Lin (M'09, SM'15) is a Full Professor of computer science at Sun Yat-sen University. He served as the Executive Director and Distinguished Scientist of SenseTime Group from 2016 to 2018, leading the R&D teams for cutting-edge technology transferring. He has authored or co-authored more than 200 papers in leading academic journals and conferences, and his papers have been cited by more than 26,000 times. He is an associate editor of IEEE Trans.Neural Networks and Learning Systems and IEEE Trans. Multimedia, and served as Area Chairs for numerous conferences such as CVPR, ICCV, SIGKDD and AAAI. He is the recipient of numerous awards and honors including Wu Wen-Jun Artificial Intelligence Award, the First Prize of China Society of Image and Graphics, ICCV Best Paper Nomination in 2019, Annual Best Paper Award by Pattern Recognition (Elsevier) in 2018, Best Paper Dimond Award in IEEE ICME 2017, Google Faculty Award in 2012. His supervised PhD students received ACM China Doctoral Dissertation Award, CCF Best Doctoral Dissertation and CAAI Best Doctoral Dissertation. He is a Fellow of IET/IAPR.