Visual Causal Scene Refinement for Video Question Answering

Yushen Wei
Sun-Yat-Sen University
weiysh8@mail2.sysu.edu.cn

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

Hong Yan
Sun-Yat-Sen University
yanh36@mail2.sysu.edu.cn

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

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

Abstract

Existing methods for video question answering (VideoQA) often suffer from spurious correlations between different modalities, leading to a failure in identifying the dominant visual evidence and the intended question. Moreover, these methods function as black boxes, making it difficult to interpret the visual scene during the QA process. In this paper, to discover critical video segments and frames that serve as the visual causal scene for generating reliable answers, we present a causal analysis of VideoQA and propose a framework for cross-modal causal relational reasoning, named Visual Causal Scene Refinement (VCSR). Particularly, a set of causal front-door intervention operations is introduced to explicitly find the visual causal scenes at both segment and frame levels. Our VCSR involves two essential modules: i) the Question-Guided Refiner (QGR) module, which refines consecutive video frames guided by the question semantics to obtain more representative segment features for causal front-door intervention; ii) the Causal Scene Separator (CSS) module, which discovers a collection of visual causal and non-causal scenes based on the visual-linguistic causal relevance and estimates the causal effect of the scene-separating intervention in a contrastive learning manner. Extensive experiments on the NExT-QA, Causal-VidQA, and MSRVTT-QA datasets demonstrate the superiority of our VCSR in discovering visual causal scene and achieving robust video question answering.

1. Introduction

Video question answering [27, 28] is a challenging task requiring machines to understand and interpret complex visual scenes to answer natural language questions about the content of a given video. Since videos have good potential to understand event temporality, causality, and dynam-
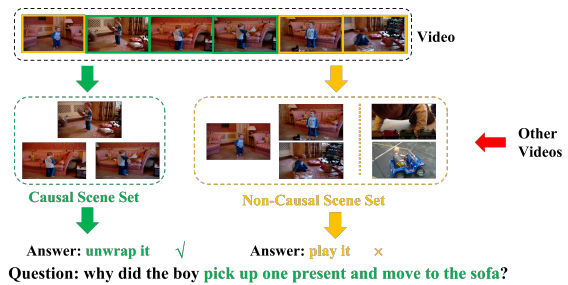
(a) Explanation of VideoQA based on causal scene sets.
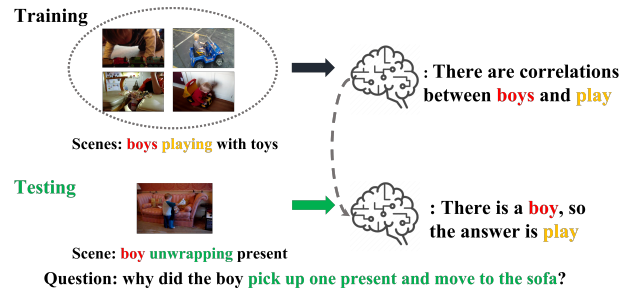
(b) Spurious correlations of visual contents in VideoQA tasks.

Figure 1. An example of causal explanation of VideoQA and visual spurious correlations. (a) illustrates the explanation of the model-predicted answer through a causal scene set, and (b) shows how the spurious correlation caused by visual confounders affects the prediction of the model.

ics [40, 41], we focus on discovering question-critical visual causal scenes and achieving robust video question answering. Our task aims to fully comprehend the richer multi-modal event space and answer the given question in a causality-aware way. To achieve innovative architecture, several studies have explored VideoQA's multi-modal nature, including enhancing vision-language alignment [22, 46] and reconsidering the structure of visual input [27, 63]. Most of the existing VideoQA methods

[30, 27, 46] use recurrent neural networks (RNNs) [52], attention mechanisms [55] or Graph Convolutional Networks [26] for relation reasoning between visual and linguistic modalities. Although achieving promising results, the current video question answering methods suffer from two common limitations.

First, the black-box nature of existing VideoQA models remains a significant challenge, as they lack transparency in their prediction process and offer little insight into the key visual cues used to answer questions about the video [8, 51]. Specifically, it is difficult to explicitly discover the dominant visual segments or frames that the model focuses on to answer the question about the video. This lack of interpretability raises concerns about the robustness and reliability of the model, particularly in safety and security applications. To improve the interpretability of VideoQA models, it is crucial to identify a subset of visual scenes, referred to as "causal scenes", that serve as evidence to support the answering process in a way that is interpretable to humans [51]. For instance, Figure 1(a) shows that the causal scene set contains the boy's question-related action, which can serve as the dominant visual causal scene that provides an intuitive explanation for why the model gives the answer "unwrap it". In contrast, the non-causal visual scene set includes question-irrelevant scenes that cannot faithfully reveal the correct question answering process.

Second, most of the existing video question answering models capture spurious visual correlations rather than the true causal structure, which leads to an unreliable reasoning process [45, 60, 32, 42]. For instance, frequently co-occurring visual concepts, such as those illustrated in Figure 1(b), can be visual confounders (C). These confounders lead to a "visual bias" denoting the strong correlations between visual features and answers. In the training set shown in Figure 1(b), the co-occurrence of the concepts "boy" and "play" dominates, which could lead the predictor to learn the spurious correlation between the two without considering the boy's action (i.e., causal positive scene P) to understand what the boy actually did. Consequently, there are significant differences in visual correlations between the training and testing sets, and memorizing strong visual priors can limit the reasoning ability of video question answering models. To mitigate visual spurious correlations, this paper takes a causal perspective on VideoQA by partitioning visual scenes into two parts: 1) causal positive scene P, which contains question-critical information, and 2) non-causal scene N, which is irrelevant to the answer. Thus, we scrutinize that the non-causal scene N is spuriously correlated with the answer A.

To address the aforementioned limitations, we propose the Visual Causal Scene Refinement (VCSR) framework to explicitly discover the visual causal scenes through causal front-door interventions. To obtain representative seg-

ment features for front-door intervention, we introduce the Question-Guided Refiner (QGR) module that refines consecutive video frames based on the question semantics. To identify visual causal and non-causal scenes, we propose the Causal Scene Separator (CSS) module based on the visual-linguistic causal relevance and estimates the causal effect of the scene-separating intervention through contrastive learning. Extensive experiments on the NExT-QA, Causal-VidQA, and MSRVTT-QA datasets demonstrate the superiority of VCSR over the state-of-the-art methods. Our main contributions are summarized as:

- We propose the Visual Causal Scene Refinement (VCSR) framework, to explicitly discover true causal visual scenes from the perspective of causal front-door intervention. To the best of our knowledge, we are the first to discover visual causal scenes for video question answering.

- We build the Causal Scene Separator (CSS) module that learns to discover a collection of visual causal and non-causal scenes based on the visual-linguistic causal relevance and estimates the causal effect of the scene-separating intervention in a contrastive learning manner.

- We introduce the Question-Guided Refiner (QGR) module that refines consecutive video frames guided by the question semantics to obtain more representative segment features for causal front-door intervention.

2. Related Work

2.1. Video Question Answering

Compared with image-based visual question answering [4, 68, 3], video question answering is much more challenging due to the additional temporal dimension. To address the VideoQA problem, the model must capture spatial-temporal and visual-linguistic relationships to infer the answer. To explore relational reasoning in VideoQA, Xu et al. [65] proposed an attention mechanism to exploit the appearance and motion knowledge with the question as guidance. Jang et al. [20, 19] proposed a dual-LSTM-based method with both spatial and temporal attention, which used a large-scale VideoQA dataset named TGIF-QA. Later, some hierarchical attention and co-attention-based methods [30, 12, 23] were proposed to learn appearance-motion and question-related multi-modal interactions. Le et al. [27] proposed the hierarchical conditional relation network (HCRN) to construct sophisticated structures for representation and reasoning over videos. Jiang et al. [22] introduced the heterogeneous graph alignment (HGA) nework that aligns the inter- and intra-modality in-

formation for cross-modal reasoning. Huang et al. [18] proposed a location-aware graph convolutional network to reason over detected objects. Lei et al. [28] employed sparse sampling to build a transformer-based model named CLIP-BERT, which achieved end-to-end video-and-language understanding. Liu et al. [34] proposed the hierarchical visual-semantic relational reasoning (HAIR) framework to perform hierarchical relational reasoning. However, these previous works tend to capture cross-modal spurious correlations within the videos and neglect interpreting the visual scene during the QA process. In contrast, we propose the Visual Causal Scene Refinement (VCSR) architecture to explicitly refine the visual causal scenes temporally.

2.2. Visual Causality Learning

Compared to conventional debiasing techniques [59], causal inference [47, 66, 42] has shown potential in mitigating spurious correlations [5] and disentangling model effects [6] to achieve better generalization. Counterfactual and causal inference are gaining increasing attention in several computer vision tasks, including visual explanations [16, 57], scene graph generation [9, 54], image recognition [58, 60], video analysis [38, 37, 39, 13, 25, 44], and vision-language tasks [1, 45, 67, 35, 36, 10]. Specifically, Tang et al. [53], Zhang et al. [69], Wang et al. [58], and Qi et al. [48] computed the direct causal effect and mitigated the bias based on observable confounders. Counterfactual based solutions are also effective. For example, Agarwal et al. [2] proposed a counterfactual sample synthesising method based on GAN [15]. Chen et al. [8] replaced critical objects and critical words with a mask token and reassigned an answer to synthesize counterfactual QA pairs. Apart from sample synthesising, Niu et al. [45] developed a counterfactual VQA framework that reduces multi-modality bias by using a causality approach named Natural Indirect Effect and Total Direct Effect to eliminate the mediator effect. Li et al. [32] proposed an Invariant Grounding for VideoQA (IGV) to force models to shield the answering process from the negative influence of spurious correlations. Li et al. [31] introduced a self-interpretable VideoQA framework named Equivariant and Invariant Grounding VideoQA (EIGV). Liu et al. [35] proposed a Cross-Modal Causal RelatIonal Reasoning (CMCIR) model for disentangling the visual and linguistic spurious correlations. Differently, our VCSR aims for visual causal scene discovery, which requires fine-grained understanding of spatial-temporal and visual-linguistic causal dependencies. Moreover, our VCSR explicitly finds the question-critical visual scenes temporally through front-door causal interventions.
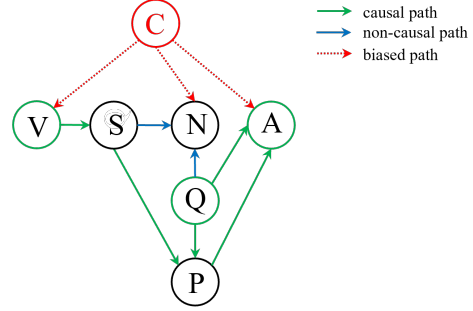

Figure 2. The Structured Causal Model (SCM) of VideoQA. V, Q and A denote video, question and answer respectively. C is the visual confounder, S denotes the refined video scenes, P and N are causal positive and negative visual scenes. Green flows: the causal path of VideoQA (the front-door path). Blue flows: the non-causal path. Red flows: biased VideoQA caused by the confounders (the back-door path).

3. Methodology

3.1. VideoQA in Causal Perspective

To discover visual causal scenes for VideoQA task, we employ Pearl's structural causal model (SCM) [47] to model the causal effect between video-question pairs and the answer, as shown in Figure 2. The variables V, Q, A are defined as the video, question, and answer. S is refined video scene set which can be divided into causal positive scene set P and negative scene set N. The front-door paths $V \rightarrow S \rightarrow P \rightarrow A$, $Q \rightarrow P \rightarrow A$, $Q \rightarrow A$ represent the true causal effects of VideoQA. These paths are involved in the reasoning process of watching the video, finding question-related scenes, and answering the question. However, the visual confounder C introduces a backdoor path $V \leftarrow C \rightarrow A$, which creates a spurious correlation between the video and answer. Unfortunately, visual domains have complex data biases, and it can be difficult to distinguish between different types of confounders. As a result, the visual confounder C cannot be observed. Since the causal positive scenes P completely mediates all causal effects from V to A, to address this issue and achieve the true visual causal effect of $V \rightarrow S \rightarrow P \rightarrow A$, we propose a causal front-door intervention by treating P as the mediator. The front-door intervention could be formulated as:

$$P(A|do(V), Q) = \sum_p P(p|do(V = v))P(A|do(P = p), Q)$$

$$= \sum_p \sum_s P(p|s)P(s|do(V = v))P(A|do(P = p), Q)$$

$$= \sum_p \sum_s P(p|s)P(s|v) \sum_{v'} \sum_{s'} P(A|p, s', Q)P(s'|v')P(v')$$
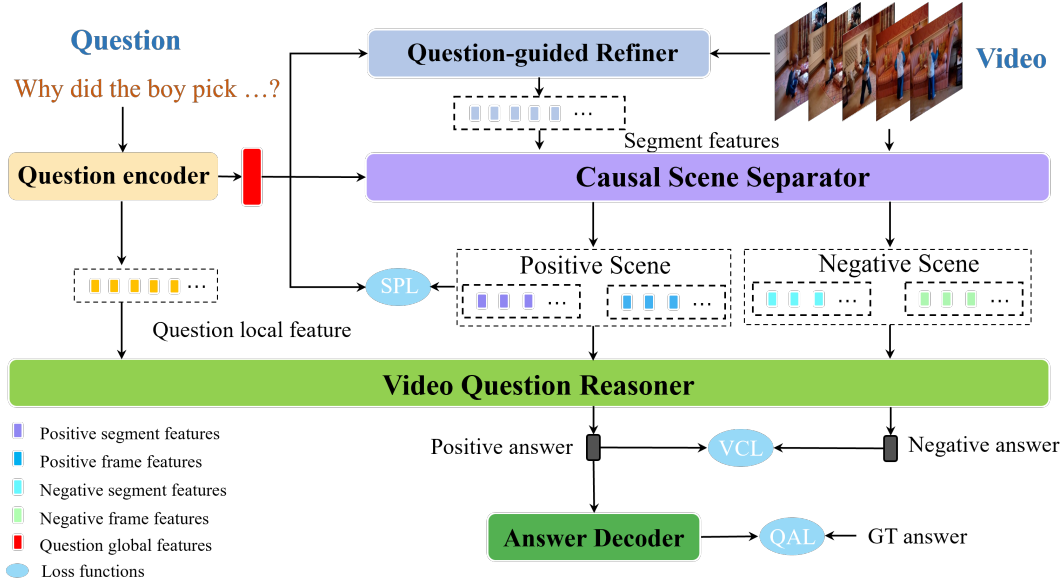
$$\tag{1}$$

Figure 3. An overview of our Visual Causal Scene Refinement (VCSR) framework. The Question-Guided Refiner (QGR) encodes consecutive video frames guided by the question semantics to obtain representative segment features for causal front-door intervention. Then, the Causal Scene Separator (CSS) learns to construct a collection of visual causal and non-causal scenes based on the visual-linguistic causal relevance and estimates the causal effect of the scene-separating intervention in a contrastive learning manner. Finally, the Video Question Reasoner (VQR) computes the answer embedding with positive and negative video features. (SPL: Semantic Preserving Loss, VCL: Visual Contrastive Loss, QAL: Question Answering Loss)

where $do(\cdot)$ is the *do*-operator indicating the intervention operation, and $P(s|do(V = v)) = P(s|v)$ because there is only front-door path between V and S, v' and s' denotes intervened videos and segment sets after $do(P = p)$. Since the total scene set S is determined given a video, we could eliminate s and s' from the eq.1:

$$P(A|do(V), Q) = \sum_p P(p|v) \sum_{v'} P(A|p, v', Q)P(v')$$
(2)

This is the front-door adjustment on causal path $V \rightarrow S \rightarrow P \rightarrow A$. After the intervention, we could eliminate the non-causal effect of biased back-door path $V \leftarrow C \rightarrow A$, making the model focus on the real causal effect. In section.3.2, we proposed the implementation of the front-door intervention eq.2.

3.2. Overall Causal Model Architecture

To implement the front-door intervention, we propose: 1) a QGR (question-guided refiner) to construct the total scene set S from video frames; 2) a CSS (causal scene separator) to model the causal positive scene distribution $P(p|v)$ in eq.2, and a multi-modal transformer to parameterize the expectation of $P(A|p, v', Q)$; and 3) leverage a contrastive learning-based training objective to handle the causal intervention. The overall framework of our VCSR is illustrated in Figure.3.
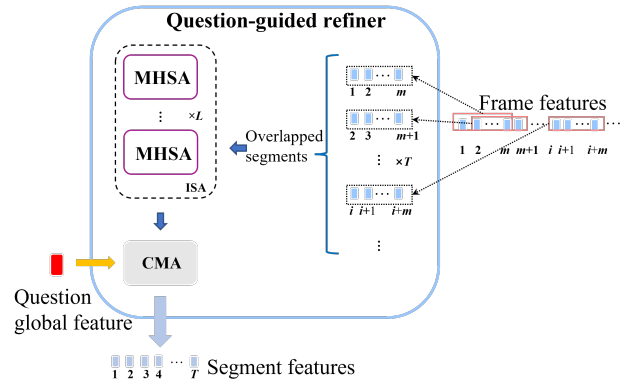

Figure 4. The Question-Guided Refiner (QGR) module. The frame features are grouped into T overlapped segments, then pass the In-segment Self Attention (ISA) module which contains L layers of in-segment Multi-head Self Attention (MHSA), and finally question-guided Cross-modal Attention (CMA) aggregates frames in the same segment.

3.3. Question-Guided Refiner

As shown in Figure.2 and eq.1, the causal effect of video V on the answer A comes through the total scene set S. To construct the scene set S from video segments, we designed Question-Guided Refiner (QGR) module to refine consecutive video frames by leveraging question semantics and obtaining more representative segment-level features for

causal front-door intervention. Firstly, a pre-trained BERT model [11] is employed to extract the question features from raw question texts. Next, the question features are encoded by a single-layer transformer encoder. The global representation of the question is denoted as the [CLS] features $q_g \in \mathbb{R}^d$, while the concatenation of other output features represents the local question, denoted as q_l.

Given the original video v, we sparsely sample N frames and utilize a pre-trained CLIP[49] encoder to extract the frame features $\mathbf{F_a} = \{f_1, f_2, ..., f_N\}$, where $f_n \in \mathbb{R}^d$, and d denotes the dimension of the frame feature. Then, we combine m adjacent frames to form a segment and obtain T overlapping segments $S = \{s_1, s_2, ..., s_T\}$, where $s_t \in \mathbb{R}^{m \times d}$ denotes the frame features in a single segment, as Figure.4 shows, each adjacent segments share $m-1$ overlapping frames. To mix the features within each segment, we employ an In-segment attention module (ISA), which is a transformer with l-layer multi-head self-attention module:

$$s'_t = [f'_{t,1}, f'_{t,2}, ..., f'_{t,m}] = \text{MHSA}^{(l)}(s_t + PE(s_t)) \quad (3)$$

where MHSA denotes the multi-head self-attention module, PE is the positional embedding and $f'_{i,j}$ is the j-th frame feature in the i-th segment.

The QGR module refines the frame features within the same segment to aggregate them temporally within the segment. To enhance the integration of feature aggregation with the VideoQA task, we incorporate a global question representation q_g to guide our refining process. We begin by utilizing a cross-modal attention (CMA) module to obtain attention scores, which implicitly reflect the relevance of frames to the QA task. We then aggregate the frame features to refine the segment-level features using the attention scores obtained from the CMA module:

$$Q = f_q(q_g), \ K = f_s(s'_t), \ V = s'_t \quad (4)$$

$$s^*_t = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (5)$$

in which f_q and f_s are linear projection layers, and s^*_t is the t-th segment feature after refining. Then, T segment features are concatenated as the refined segments for the next causal scene separation step: $S^* = \{s^*_1, s^*_2, ..., s^*_T\}$, as shown in Figure.4.

3.4. Causal Scene Separator

To construct a collection of causal scenes related to the question in a video(i.e., the positive scene P), we propose a Causal Scene Separator (CSS) that identifies segments and frames with higher causal relevance to the question, as depicted in Figure.5. The Causal Scene Separator comprises two modules: a causal segment generator and a causal frame filter.

Causal segment generator. The causal segment generator aims to generate sets of causal positive and negative segments for forming causal scenes. For positive segments, it initially computes the attention scores of the refined segments S^* and the global question features q_g using the cross-modal attention (CMA) module:

$$a_s = \text{Softmax}(g_q(q) \cdot g_s(S^*)^T) \quad (6)$$

where g_q and g_s are linear layers. Then, we leverage Gumbel-Softmax to generate a discrete selection mask for capturing the causal content:

$$s^i_p = S^* \text{Gumbel-Softmax}(a_s)^T \quad (7)$$

We repeat the selection process for k times to obtain the causal positive segment set of size k, denoted as $S_p = \{s^1_p, s^2_p, ..., s^k_p\}$. Other segments with attentive probability lower than a threshold τ and segments from the segment pool, including segments from other videos, form a negative candidate set. A subset of the candidate set is sampled as the causal negative segment set $S_n = \{s^1_n, s^2_n, ..., s^k_n\}$.

Causal frame filter. Besides segment features, question-related frames in a segment can complement the causal scenarios since some causal scenes may only contain a few or even one single frame. As shown in Figure.5(b), the causal frame filter first aligns the positive and negative segment sets with frames to obtain frames that belong to corresponding segments. Then, a selector similar to the one used in the segment filter in Figure.5(a) chooses a single frame for each segment to construct the causal positive frame set $F_p = \{f^1_p, f^2_p, ..., f^k_p\}$ and the causal negative frame set $F_n = \{f^1_n, f^2_n, ..., f^k_n\}$. The causal segment sets and the causal frame sets combine to form causal scenes. Formally, we have causal positive scene sets $C_p = \{S_p, F_p\}$ and causal negative scene sets $C_n = \{S_n, F_n\}$.

Segment-frame semantic preserving loss. To preserve specific semantics of segment and frame features, we propose a novel segment-frame semantic preserving loss. This loss is based on the assumption that if a single frame is sufficient to answer a question, then the video segment that contains that frame should also be sufficient for the same question. The above assumption described that the positive video segment should be relatively more important in answering the question. We estimate the relative importance of two types of visual contents with their cosine similarity to the question representation:

$$I = [I^i_f, I^i_s] = \text{Softmax}([\text{sim}(q_g, f^i_p), \text{sim}(q_g, s^i_p)]) \quad (8)$$

where $\text{sim}(\cdot)$ refers to the cosine similarity, $[,]$ means concatenation, I^i_f and I^i_s are the relative importance of i-th positive frame and segment. We then introduce hinge loss to model the relative importance constraint:

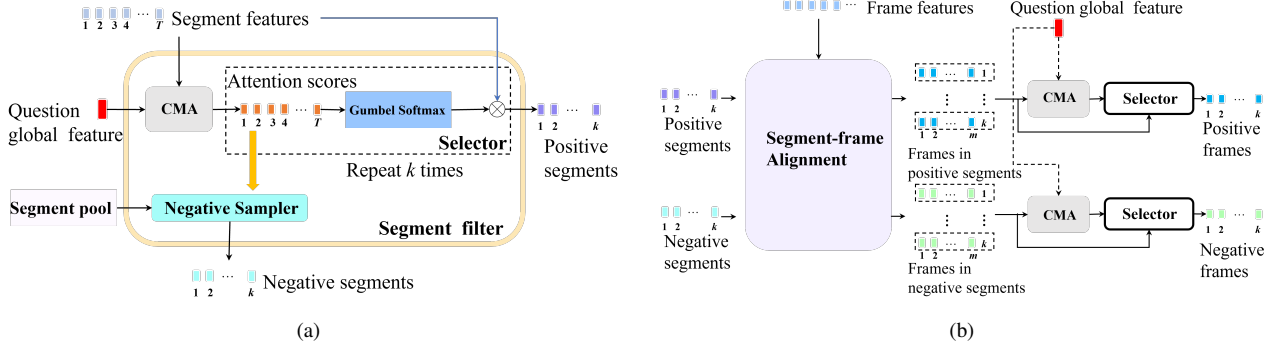$$\mathcal{L}_{SP} = \sum_{i=1}^{k} \max((I^i_f - I^i_s), 0) \quad (9)$$

Figure 5. The internal structure of Causal Scene Separator. (a) Causal segment generator. This module selects the possible causal positive segments in the video and generates negative segments by sampling segments from the segment pool. (b) Causal frame filter. Given positive and negative segments, the causal frame filter aligns segments with their respective frames, and then selects the most question-relevant frame for each positive and negative segment.
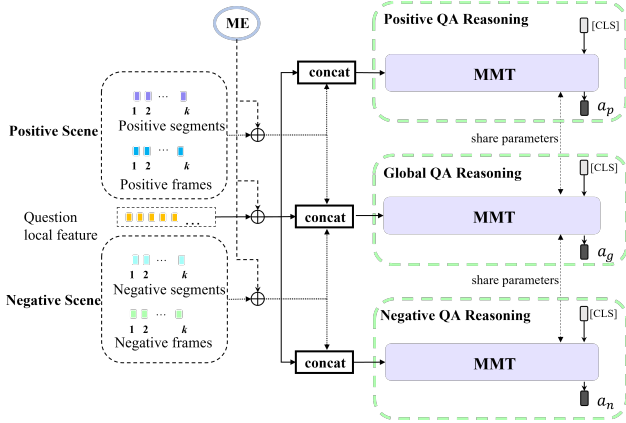

Figure 6. The multi-modal transformer (MMT) reasoner reasons the positive answer, negative answer, and global answer when given different causal scenes.

3.5. Video Question Reasoning

Given the causal scene sets $C_g = [C_p, C_n]$ (i.e., the intervened scene set s' in eq.1, by fixing the positive scene set, we implement $do(P = p)$), we leverage contrastive learning to model the reasoning about causal interventions based on scene separating. As shown in Figure.6, we derive answer representations by feeding the multi-modal transformer (MMT) reasoner with positive, negative, and global causal scenes:

$$a_p = MMT(ME(C_p), ME(q_l)) \qquad (10)$$

$$a_n = MMT(ME(C_n), ME(q_l)) \qquad (11)$$

$$a_g = MMT(ME(C_g), ME(q_l)) \qquad (12)$$

where ME is modality embedding module ,a_p and a_n are answer contrastive counterparts ,and a_g act as the contrastive anchor.

Visual contrastive loss. To estimate the causal effect of the scene-separating intervention, we introduce InfoNCE loss to construct a contrastive objective as follows:

$$\mathcal{L}_{VC} = -\log \frac{e^{a_p^T \cdot a_g}}{e^{a_p^T \cdot a_g} + \sum_{i=1}^{\mathcal{N}} e^{a_p^T \cdot a_g^i}} \qquad (13)$$

where $\mathcal{N}$ is the number of negative answers, those answers are obtained by feeding the QA reasoner with different sampling subsets of negative scenes.

3.6. Answer Prediction

For multi-choice QA settings, the local question representation q_l is derived by feeding the concatenation of questions and answer candidates to the question encoder. And the answer prediction is given by the positive part of answer representations:

$$\widetilde{a} = \arg\max(F(a_p)) \qquad (14)$$

where F is a set of linear projections that $F = \{f_a\}_{a=1}^{|\mathcal{A}|}$, $\mathcal{A}$ is the set of answer candidates, $f_a \in \mathbb{R}^{d \times 1}$ denotes the final linear head for each question candidates.

As for the open-ended QA setting, the formulation of the final answer prediction is:

$$\widetilde{a} = \arg\max(f_o(a_p)) \qquad (15)$$

in which $f_o \in \mathbb{R}^{d \times |\mathcal{A}|}$ is a fully-connected layer, and $|\mathcal{A}|$ denotes the length of answer dictionary.

Question answering loss. The question-answering loss is the cross entropy loss between the predicted answer $\widetilde{a}$ and the ground truth answer a_{gt}:

$$\mathcal{L}_{QA} = \text{CrossEntropy}(\widetilde{a}, a_{gt}) \qquad (16)$$

Methods	Visual backbones	Val				Test			
		Causal	Temporal	Descriptive	Acc.	Causal	Temporal	Descriptive	Acc.
EVQA[4]	ResNet + ResNeXt	42.64	46.34	45.82	44.24	43.27	46.93	45.62	44.92
STVQA[20]	ResNet + ResNeXt	44.76	49.26	55.86	47.94	45.51	47.57	54.59	47.64
CoMem[14]	ResNet + ResNeXt	45.22	49.07	55.34	48.04	45.85	50.02	54.38	48.54
HME[12]	ResNet + ResNeXt	46.18	48.20	58.30	48.72	46.76	48.89	57.37	49.16
HCRN[27]	ResNet + ResNeXt	45.91	49.26	53.67	48.20	47.07	49.27	54.02	48.89
HGA[22]	ResNet + ResNeXt	46.26	50.74	59.33	49.74	48.13	49.08	57.79	50.01
IGV[32]	ResNet + ResNeXt	-	-	-	-	48.56	51.67	59.64	51.34
HQGA[62]	ResNet + ResNeXt + FasterRCNN	48.48	51.24	61.65	51.42	49.04	**52.28**	59.43	51.75
ATP[7]	CLIP	53.1	50.2	**66.8**	54.3	-	-	-	-
VGT[63]	ResNet + ResNeXt + FasterRCNN	52.28	<u>55.09</u>	<u>64.09</u>	<u>55.02</u>	51.62	51.94	**63.65**	53.68
EIGV[31]	ResNet + ResNeXt	-	-	-	-	-	-	-	<u>53.7</u>
VCSR-ResNet*	ResNet + ResNeXt	50.17	50.74	57.92	51.56	49.62	50.28	61.00	51.69
VCSR-ResNet	ResNet + ResNeXt	50.9	51.3	58.36	52.22	49.98	<u>51.98</u>	61.78	52.53
VCSR-CLIP*	CLIP	<u>53.13</u>	53.23	62.55	54.62	<u>52.00</u>	50.88	60.64	53.07
VCSR-CLIP	CLIP	**54.12**	**55.33**	63.06	**55.92**	**53.00**	51.52	<u>62.28</u>	**54.06**

Table 1. Comparison with state-of-the-art methods on NExT-QA dataset. The **best** and <u>second-best</u> results are highlighted. The "**VCSR-ResNet***" and "**VCSR-CLIP***" denote the VCSR models that do not incorporate QGR and CSS modules and are trained without contrastive learning objective $\mathcal{L}_{VC}$ and semantic preserving objective $\mathcal{L}_{SP}$.

3.7. Training objective

Our total training objective comprises three components: question-answering loss (See eq.16), visual contrastive loss (See eq.13), and segment-frame semantic preserving loss (See eq.9), the overall objective is achieved by aggregating the above three objectives:

$$\mathcal{L} = \mathcal{L}_{QA} + \alpha\mathcal{L}_{VC} + \beta\mathcal{L}_{SP} \qquad (17)$$

where α and β are hyper-parameters that control the contribution of sub-objectives.

4. Experiments

4.1. Datasets

We evaluate our VCSR model on three VideoQA benchmarks that evaluate the model's reasoning capacity from different aspects including temporality, causality, and commonsense: **NExT-QA**[61], **Causal-VidQA**[29] and **MSRVTT-QA**[65].

NExT-QA highlights the causal and temporal relations among objects in videos. It is a manually annotated multi-choice QA dataset targeting the explanation of video contents, especially causal and temporal reasoning. It contains 5,440 videos and 47,692 QA pairs, each QA pair comprises one question and five candidate answers.

Causal-VidQA emphasizes both evidence reasoning and commonsense reasoning in real-world actions. It is a multi-choice QA benchmark containing 107,600 QA pairs and 26,900 video clips. Questions in Causal-VidQA dataset are categorized into four question types: description, explanatory, prediction, and counterfactual. For prediction and counterfactual questions, Causal-VidQA proposed three types of reasoning tasks: question to answer ($Q \rightarrow A$), question to reason ($Q \rightarrow R$), and question to answer and reason ($Q \rightarrow AR$).

MSRVTT-QA focuses on the visual scene-sensing ability by asking the descriptive questions. It is an open-ended QA benchmark containing 10,000 trimmed video clips and 243,680 QA pairs, with challenges including description and recognition capabilities.

4.2. Implementation details

For each video, we uniformly sample 64 frames following [27], and extract the features using a pre-trained CLIP (ViT-L/14) encoder. For the questions, we obtain word embeddings using a pre-trained BERT model.

For Causal-VidQA dataset, we follow [29] to add the BERT representation with Faster-RCNN[50] extracted instance representation for a fair comparison. The model hidden dimension d is set to 512, the segment length m is set to 6, and the positive segment number k is set to 4 for each dataset. The number of MHSA layers L in QGR is set to 2, and the MMT in the video question reasoner is implemented by a 3-layer transformer. The number of heads of all multi-head attention modules is set to 8. The training process is optimized by the AdamW[43] optimizer with the learning rate $lr = 1e-5$, $\beta_1 = 0.9$, $\beta_2 = 0.99$, and weight decay of 0. The hyper-parameter α is set to 0.0125 and β is set to 0.04. The training progress is carried out for 50 epochs, and the learning rate is halved if the validation accuracy does not improve after 5 epochs.

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

Table 1 presents a comparison of our VCSR methods with state-of-the-art (SOTA) methods on the NExT-QA dataset. The results demonstrate that our VCSR achieves superior performance on both the validation set and test set. Notably, our VCSR excels in *Causal* question splits, with an accuracy improvement of 1.02% and 1.38% in the validation set and test set, respectively, indicating a stronger causal relational reasoning ability. Additionally, our VCSR

Methods	Acc_E	Acc_D	Acc_P			Acc_C			Acc
			$Q \rightarrow A$	$Q \rightarrow R$	$Q \rightarrow AR$	$Q \rightarrow A$	$Q \rightarrow R$	$Q \rightarrow AR$	
EVQA[4]	60.95	63.73	45.68	46.40	27.19	48.96	51.46	30.19	45.51
CoMem[14]	62.79	64.08	51.00	50.36	31.41	51.61	53.10	32.55	47.71
HME[12]	61.45	63.36	50.29	47.56	28.92	50.38	51.65	30.93	46.16
HCRN[27]	61.61	65.35	51.74	51.26	32.57	51.57	53.44	32.66	48.05
HGA[22]	63.51	65.67	49.36	50.62	32.22	52.44	55.85	34.28	48.92
B2A[46]	62.92	**66.21**	48.96	50.22	31.15	53.27	**56.27**	**35.16**	49.11
VCSR-CLIP*	64.91	65.00	57.69	54.74	36.74	52.26	53.14	32.27	49.73
VCSR-CLIP	65.41(+0.5)	65.98(+0.98)	60.88(+3.19)	58.54(+3.8)	41.24(+4.5)	53.38(+1.12)	54.37(+1.23)	34.06(+1.79)	51.67(+1.94)

Table 2. Comparison with state-of-the-art methods on Causal-VidQA dataset. (*E*: explanatory, *D*: descriptive, *P*: prediction, *C*: counterfactual, *Q*: question, *A*: answer, *R*: reason)

Methods	What	Who	How	Total
QueST[21]	27.9	45.6	83.0	34.6
HGA[22]	29.2	45.7	83.5	35.5
DualVGR[56]	29.4	45.5	79.7	35.5
HCRN[27]	-	-	-	35.6
QESAL[33]	30.7	46.0	82.4	36.7
B2A[46]	-	-	-	36.9
ClipBert[28]	-	-	-	37.4
ASTG[24]	31.1	48.5	83.1	37.6
IGV[32]	-	-	-	38.3
HQGA[62]	-	-	-	38.6
VCSR-CLIP	**31.9**	**51.0**	**85.0**	**38.9**

Table 3. Comparison with SOTAs on MSRVTT dataset.

Methods	NExT-QA Val				MSRVTT-QA
	Causal	Temporal	Descriptive	Total	
VCSR-CLIP*	53.13	53.23	62.55	54.62	38.5
VCSR-CLIP w/o QGR	52.78	56.08	60.49	55.04	38.7
VCSR-CLIP w/o CSS	52.78	54.40	63.35	55.06	38.7
VCSR-CLIP w/o $\mathcal{L}_{SP}$	53.43	54.34	63.19	55.24	38.5
VCSR-CLIP w/o $\mathcal{L}_{VC}$	53.36	54.28	63.06	55.16	38.8
VCSR-CLIP	54.12	55.33	63.06	55.92	38.9

Table 4. Ablation study on modules and training objectives.

achieves competitive performance for *Temporal* questions. This validates that our VCSR can effectively discover temporally sensitive visual scenes in videos. For *Descriptive* questions, our VCSR achieves lower performance than previous methods ATP and VGT. This is because VGT adopts object detection pipeline that makes visual scene sensing more fine-grained. And ATP preserves the most representative frame for each video clip at the cost of harming temporal reasoning ability. Although without object detection, our VCSR can outperform these two methods on more challenging problems *Causal* and *Temporal*. Moreover, we assess the generalization ability of our VCSR on different visual backbones. VCSR-ResNet[17] replaces the CLIP visual feature with the concatenation of ResNet-101[64] extracted appearance feature and ResNeXt-101 extracted motion feature. The results reveal that the introduction of causal scene intervention also enhances the performance of VCSR-ResNet, highlighting the effectiveness of causal scene intervention on different visual backbones.

To further evaluate the evidence reasoning and commonsense reasoning ability of our VCSR in real-world actions, we evaluate the VCSR on the Causal-VidQA dataset, as shown in Table.2. Our VCSR achieves a total accuracy of 51.67%, outperforming the state-of-the-art B2A [46] by 2.56%. Additionally, for predictive and counterfactual tasks, the introduction of causal intervention significantly promotes the performance of VCSR in answering predic-

tive and counterfactual questions, which require better reasoning capability. This highlights the effectiveness of cross-modal causal relational reasoning when addressing these types of questions.

To evaluate the visual scene-sensing ability of our VCSR, we evaluate our VCSR on open-ended descriptive QA dataset MSRVTT. In Table 3, we compare the performance of VCSR with the state-of-the-art methods on the MSRVTT dataset. The results show that VCSR has good overall performance on the open-ended dataset, particularly for question types "Who" and "How".

The experimental results in these three large-scale datasets demonstrate that our VCSR outperforms state-of-the-art methods in terms of comprehensive understanding of visual concepts, temporality, causality, and commonsense within videos. This validates that our VCSR generalizes well across different VideoQA benchmarks.

5. Ablation Studies

We conduct ablation studies to verify the effectiveness of (1) QGR and CSS module, (2) training objectives $\mathcal{L}_{SP}$ and $\mathcal{L}_{VC}$. All ablation studies are conducted on NExT-QA validation set and MSRVTT-QA dataset, the variants of our VCSR are listed as follows:

VCSR-CLIP*: the VCSR model without QGR and CSS modules and training without contrastive objective $\mathcal{L}_{VC}$ and semantic preserving objective $\mathcal{L}_{SP}$.

VCSR-CLIP w/o QGR: the VCSR model without QGR module, the segment features are obtained by mean-pooling frame features.

VCSR-CLIP w/o CSS: remove the CSS module from VCSR. Without scene separation, the whole scene set is fed to the reasoner. In this setting, the contrastive objective is naturally removed since the lack of counterparts.
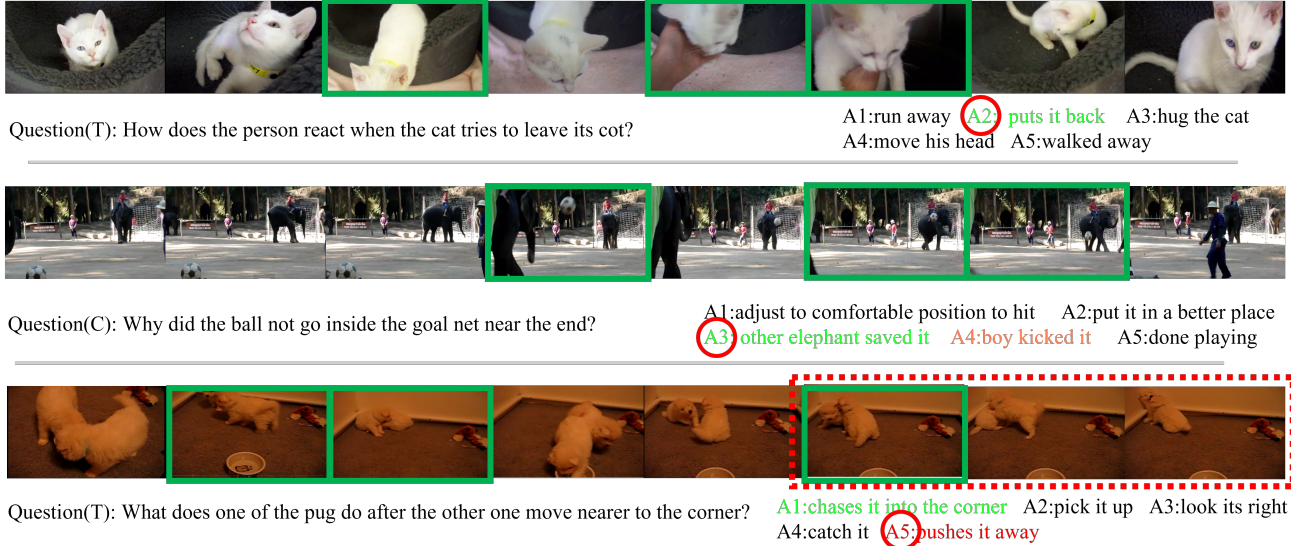
Figure 7. The visualization of proposed causal positive scenes on the NExT-QA dataset. For the first two questions, the proposed positive scenes cover critical video clips and the model predicts the correct answer accordingly. However, the model makes a mistake in the answer prediction for the last question, as it did not fully capture the entire critical scene set. The green boxes and answers represent the VCSR predicted rationales and answers, respectively, while the red circles indicate the ground truth answer. In the second example, the orange answer denotes the answer predicted by VCSR*, and in the last example, the red dashed box shows the rationale provided by a human.

VCSR-CLIP w/o $\mathcal{L}_{SP}$: Training without semantic preserving loss $\mathcal{L}_{SP}$.

VCSR-CLIP w/o $\mathcal{L}_{VC}$: Remove the contrastive objective $\mathcal{L}_{VC}$ from the total objective L, the answer prediction is predicted based on the positive answer embedding.

Table 4 presents the ablation results, indicating that all of the modules and objectives contribute to improving the total performance on both datasets. Specifically, on the validation set of NExT-QA, we observed that removing all modules and objectives would negatively affect the performance of the *Causal* split. Removing QGR, on the other hand, resulted in a decline in the performance of *Causal* and *Descriptive* splits but a boost in the performance of *Temporal* split. This is because the QGR module refining the question-related frames by weighing down other frames in a segment and leading to the partial loss of temporal information.

Moreover, we notice that removing CSS modules and both objectives has little effect on the performance of the *Descriptive* split. This could be due to the fact that our scene refinement and separating methods primarily focus on the temporal perspective, while descriptive questions depend more on spatial information.

6. Qualitative Results

To verify the ability of the VCSR in discovering visual causal scenes and visual-linguistic causal reasoning, we analyze correct and incorrect visualizations on the NExT-QA dataset. The results are presented in Figure 7. As shown in the figure, when answering the first two questions, the positive scene given by CSS could evidently explain the reason for choosing the correct answer (i.e., scenes of person putting the cat back to the cot and elephant saving the ball). This validates that the VCSR can reliably focus on the dominant visual scenes when making decisions. Specially, for the second question, we compare the answer predicted by VCSR and VCSR* and find that the VCSR* without causal intervention is affected by a spurious correlation between visual content "boy" and "ball", leading to the wrong answer of "boy kicked it". In our VCSR, we reduce such spurious correlation and pursue the true causality by adopting causal intervention, resulting in better dominant visual evidence and question intention. Moreover, we observe that when answering the last question, the CSS does not capture the entire causal scene set and thus predicts the wrong answer. This is probably caused by the similarity of the visual semantics of the pug's actions, which could be addressed with better visual backbones.

7. Conclusion

In this paper, we propose a cross-modal causal relational reasoning framework named VCSR for VideoQA, to explicitly discover the visual causal scenes through causal front-door interventions. From the perspective of causality, we model the causal effect between video-question pairs and the answer based on the structural causal model (SCM). To obtain representative segment features for front-door intervention, we introduce the Question-Guided Refiner (QGR)

module. To identify visual causal and non-causal scenes, we propose the Causal Scene Separator (CSS) module. Extensive experiments on three benchmarks demonstrate the superiority of VCSR over the state-of-the-art methods. We believe our work could shed light on exploring new boundaries of causal analysis in vision-language tasks. In the future, we will leverage object-centric causal relational inference to alleviate the spurious correlations.

References

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

[2] Vedika Agarwal, Rakshith Shetty, and Mario 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*, pages 9690–9698, 2020. 3

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

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2, 7, 8

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

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

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

[8] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020. 2, 3

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

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

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[12] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 2, 7, 8

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

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

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

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

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8

[18] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11021–11028, 2020. 3

[19] Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Video question answering with spatio-temporal reasoning. *International Journal of Computer Vision*, 127(10):1385–1412, 2019. 2

[20] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 2, 7

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

[22] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11109–11116, 2020. 1, 2, 7, 8

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

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

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

[26] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2

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

[28] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 1, 3, 8

[29] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21273–21282, 2022. 7

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

[31] Yicong Li, Xiang Wang, Junbin Xiao, and Tat-Seng Chua. Equivariant and invariant grounding for video question answering. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4714–4722, 2022. 3, 7

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

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

[34] Fei Liu, Jing Liu, Weining Wang, and Hanqing Lu. Hair: Hierarchical visual-semantic relational reasoning for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1698–1707, 2021. 3

[35] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *arXiv preprint arXiv:2207.12647*, 2022. 3

[36] Yang Liu, Guanbin Li, and Liang Lin. Causality-aware visual scene discovery for cross-modal question reasoning. *arXiv preprint arXiv:2304.08083*, 2023. 3

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

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

[39] Yang Liu, Zhaoyang Lu, Jing Li, Tao Yang, and Chao Yao. Deep image-to-video adaptation and fusion networks for action recognition. *IEEE Transactions on Image Processing*, 29:3168–3182, 2019. 3

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

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

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

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

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

[45] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021. 2, 3

[46] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge to answer: Structure-aware graph interaction network for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15526–15535, 2021. 1, 2, 8

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

[48] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10860–10869, 2020. 3

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 7

[51] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2662–2670, 2017. 2

[52] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. *Advances in Neural Information Processing Systems*, 2015:2440–2448, 2015. 2

[53] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33, 2020. 3

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

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[56] Jianyu Wang, Bing-Kun Bao, and Changsheng Xu. Dualvgr: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia*, 24:3369–3380, 2021. 8

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

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

[59] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *European Conference on computer vision*, pages 728–744. Springer, 2020. 3

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

[61] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9777–9786, 2021. 7

[62] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2804–2812, 2022. 7, 8

[63] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 39–58. Springer, 2022. 1, 7

[64] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 8

[65] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 2, 7

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

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

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

[69] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33, 2020. 3