Discovering the Real Association: Multimodal Causal Reasoning in Video Question Answering

Chuanqi Zang[1]✉, Hanqing Wang[1], Mingtao Pei[1], Wei Liang[1,2]✉*
[1]School of Computer Science and Technology, Beijing Institute of Technology
[2]Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing

Abstract

Video Question Answering (VideoQA) is challenging as it requires capturing accurate correlations between modalities from redundant information. Recent methods focus on the explicit challenges of the task, e.g. multimodal feature extraction, video-text alignment and fusion. Their frameworks reason the answer relying on statistical evidence causes, which ignores potential bias in the multimodal data. In our work, we investigate relational structure from a causal representation perspective on multimodal data and propose a novel inference framework. For visual data, question-irrelevant objects may establish simple matching associations with the answer. For textual data, the model prefers the local phrase semantics which may deviate from the global semantics in long sentences. Therefore, to enhance the generalization of the model, we discover the real association by explicitly capturing visual features that are causally related to the question semantics and weakening the impact of local language semantics on question answering. The experimental results on two large causal VideoQA datasets verify that our proposed framework 1) improves the accuracy of the existing VideoQA backbone, 2) demonstrates robustness on complex scenes and questions. The code will be released at `https://github.com/Chuanqi-Zang/Discovering-the-Real-Association`.

1. Introduction

Video Question Answering (VideoQA) aims to understand visual information and describe it in language question-answer format, which is a natural cognitive capability for humans. Computer Vision (CV) and Natural Language Processing (NLP) as the base models for VideoQA have shown significant progress due to the successful application of deep learning, such as action classification [25], object detection [10], instance segmentation [31], and large-scale pre-trained language model [6, 19]. These tremendous advances in basic applications fuel the confidence in fine-grained multimodal analysis and reasoning, not only feature extraction, but also fine-grained general causality estimation, which is critical for a robust cognitive system.

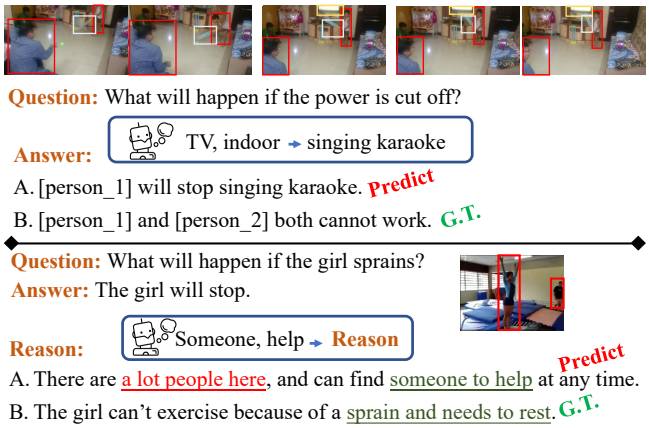

Figure 1. Two samples in VideoQA dataset. They exhibit spurious reasoning processes of B2A [26] that rely on statistical patterns, including visually spurious object-relationship associations (top) and textually unilateral semantic representations (bottom).

Recent VideoQA methods usually explore multimodal relational knowledge by sophisticated structured architecture, such as memory-augmented model [9], hierarchical model [20], topological model [18], and transformer-based model [37]. Although experiments validate their feature fusion capabilities, we find that these methods concentrate on statistical association based on multimodal data, ignoring the real stable association. They usually use a generally constrained approach with Empirical Risk Minimization (ERM), which tends to over-rely on the co-occurrence bias of repeated objects and words in the collected observational data, and bypasses the impact of complete semantics at the sentence level. This mechanism reduces the robustness of the model on new data, even in the test set which has similar distributions to the training set.

For example, as shown in Fig. 1 (top), two people are playing "cricket" indoors, and there are other objects in the room, including a TV that is on. If relying on the statistical relationship, the model may be confused by the two im-

*Wei Liang is the corresponding author.

portant visual factors of "indoor" and "television", and misjudge the concerned event, that is, "sing karaoke". Based on the clues provided by the question, the concerned event is related to the action that is taking place, e.g. "two men playing cricket". This requires the model to accurately judge the objects involved in the event and infer the answer. In Fig. 1 (bottom), we show the predicted deviation caused by statistical relationships in the text. When inferring the reasons for answer selection, existing models intensively rely on correlations between local words and video content, e.g. "someone; help", ignoring unreasonable inferences from other parts of the sentence, e.g. "a lot".

From these observations, we summarize two causal challenges for the VideoQA task. 1) **Irrelevant objects confounder.** The visual information related to the question is usually causally related to finite objects in the video. When other objects or background are considered, they are subject to data bias and become confounders, misleading the model to select the negative candidate. 2) **Keywords confounder.** The semantics expressed by textual information is represented by the overall sentence. Some long sentences contain partially sensible keywords. When just focusing on local keywords semantics, the model falls into spurious causal inferences, reducing the robustness of the model.

To address the above causal challenges and improve the robustness of the model, we propose a Multimodal Causal Reasoning (MCR) framework for video-text data. In this framework, causal features and confounding features are decoupled separately in visual and textual modes through two training strategies. To explicitly decouple the influence of different objects and scenes, MCR extracts fine-gained object-level appearance features and motion dynamics by spatial Region of Interest (ROI) Align [13] on global visual features. Among them, causally related objects are selected based on the correlation between object features and question semantics. In addition to the visual feature, we also model object coordinate information, category information, and global object interaction information to provide spatio-temporal relation representations for accurate causal attribute classification. For textual confounders, we adopt a strategy to reduce the impact of keywords on causality. MCR relies on the correlation between word encoding and question-visual co-embedding to select keywords which have a crucial impact on the prediction results. These keywords provide negative representations for successive deductive answers. Therefore, we combine these keywords with other candidate answers to generate difficult negative samples to improve the recognition ability of the model. During training, visual intervention and textual intervention are iteratively optimized. Multimodal causal relationships are gradually established which improves the robustness and reusability of the model.

We summarize our contributions as: (1) We discover two new types of causal challenges for both visual data and textual data. (2) We propose an object-level causal relationship extraction strategy to establish the real association between objects and language semantics, and a keyword broadcasting strategy to cut off the spurious influence of local textual information. (3) We achieve state-of-the-art performance on two latest large causal VideoQA datasets.

2. Related Work

Video Question Answering (VideoQA). In early works, the long-term dependency features in video and text were extracted by RNN-based modules, and then were fused by element-wise multiplication [42, 43] with attention mechanism [35]. Considering the implicit interaction in multimodal data, Jiang et al. [18] proposed a heterogeneous graph convolution-based network for crossmodal fusion. Park et al. [26] enhanced crossmodal graphs by a bridged visual-to-visual interactions structure. Huang et al. [15] improved graph interaction reasoning at the fine-grained object level. Dang et al. [5] explored the symbol-like manipulable reasoning by a hierarchically nested spatio-temporal graph. Benefited from the pre-trained language-based transformer [6] and video-text-alignment transformer [37, 41], current works [24, 38] can fine-tune the pre-trained model and show remarkable feature extraction ability and cross-model aligning ability. However, these methods ignore the potential bias distribution in the data. While improving the ability of feature extraction and alignment, they introduce confounders that lead to poor generalization.

Causal learning. Except for representation learning, recent work found that causal reasoning in data is meaningful for VideoQA. In the synthetic dataset [39], inspired by neural-symbolic method [40] for the ImageQA task, some work [4, 7, 39] explicitly represent the appearance information and physical information of each object in the scene. To investigate the causal modeling ability of models in real life, recent datasets [21, 36, 39] asked questions about behavior causality in addition to descriptive questions. For some common-sense causal knowledge that is not represented in the videos, Chadha et al. [3] proposed a knowledge base as additional guidance. To find clues of causal associations, Xu et al. [36] proposed to dynamically select the frame from the past or future. The same in frame level, Li et al. [22] explored invariant learning [1] to distinguish question-irrelevant scenes.

In real-life scenes, videos contain redundant object information. Since lacking human annotations, frequently-occurring objects are introduced as confounders, which bring more complex spurious guidance for video understanding than rough frame-level background snippets. In addition, in texts involving causal reasoning, some phrases are also prone to become confounders. This paper will unify the analysis of causal inference on multimodal data.

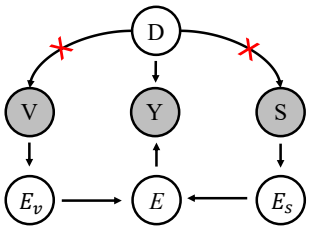

Figure 2. Causal formulation of VideoQA. Gray variables are observed data. V is a specific video and S is a specific sentence. D represents the data domain determined by the collected dataset. It causes the co-occurring of objects in video V, various words in sentence S, and answer label Y. The ideal causal inference is that model extracts video event features E_v from videos and text event features E_s from words. Then it integrates different modal event representations and filters out the event of interest E for answer selection Y. Here, we provide the backdoor adjustment from D to V and from D to S, which produces our MCR architecture.

3. Method

3.1. Formulation of VideoQA

Causal Preliminaries. We formalize cross-modal causality for the VideoQA task via a Structure Causal Model [27]. In Fig. 2, we represent the key factors in the VideoQA task as components in the causal graph, and represent the relationships between factors as connecting links that are built by current models [18,23,26,33] as default, including data domain D, video V, sentence S, label Y, visual event features E_v, and textual event features E_s, alignment events E.

$\boldsymbol{V \rightarrow E_v \rightarrow E \leftarrow E_s \leftarrow S}$ denotes feature extraction and alignment of multimodal data. E_v and E_s represent the complete events feature extracted from the video and sentence, respectively. In visual domain, pre-trained feature extraction models extract different objects and interactions as events. In textual domain, pre-trained models represent linguistic concepts in stages as events according to punctuation and connectives. $E_v \rightarrow E \leftarrow E_s$ represents multimodal event selection and alignment. The deep model references language events to select video events of interest. Then it aligns event representations and infers a unified event representation E.

$\boldsymbol{E \rightarrow Y}$ represents the direct causal effect from the unified feature of the concerned event to the label, which is an ideal way that remains invariant in other data D with different data distributions. For example, "play cricket" is causally related to cricket bat and swing, and irrelevant to "indoor" and "outdoor".

$\boldsymbol{V \leftarrow D \rightarrow Y}$ denotes that the data domain D as a confounder provides a spurious shortcut from video V to label Y in the visual domain. Specifically, the data domain D represents the statistical relationships contained in the limited collected video data to represent the co-occurrence of objects and labels. In the top example in Fig. 1, "TV" and "indoor" provide the shortcut connection that is identified as "karaoke" because of lacking human interaction event.

$\boldsymbol{S \leftarrow D \rightarrow Y}$ denotes that the data domain D also establishes spurious associations between sentence S and label Y. This is reflected in the fact that the model is easy to establish the association between labels and local concerned expressions of sentences while ignoring the overall expressions. Especially in multiple-choice questions, the model shows a stronger preference for the local "key" representation. For example, in the below example in Fig. 1, the model selects the locally correct answer, "someone" and "help".

Causal Intervention. The well-known backdoor adjustment [27] helps in eliminating spurious correlations, resulting in better generalization for the model. We present the true causality from V and S to Y as $P(Y|do(V,\ S))$, where $do()$ denotes the interventional operation. $P(Y|do(V,\ S))$ can cut off the link $D \rightarrow V$ and $D \rightarrow S$ to block this backdoor path by changing the original training data into new data $\mathcal{T} = \{\tau_1, ..., \tau_h\}$. Each of them donates a confounder stratum (total h), guiding the model to find invariants from the confounder. The backdoor adjustment can be represented as:

$$\begin{aligned} P\left(Y|\ do\left(V,\ S\right)\right) &= \sum_{\tau \in \mathcal{T}} P(Y|V,S,\tau)P(\tau) \\ &= \sum_{\tau_v \in \mathcal{T}_v} P\left(Y|\ V,\tau_v\right) P\left(\tau_v\right) \\ &+ \sum_{\tau_s \in \mathcal{T}_s} P\left(Y|S,\tau_s\right) P\left(\tau_s\right) \end{aligned} \tag{1}$$

where $P(Y|V,S,\tau)$ denotes the prediction in each new data split τ. $P(\tau)$ denotes confounder selection probability for a specific video and sentence, calculated by $P\left(\tau\right) := \frac{1}{h}$. We calculate probability estimates separately by two individual architectures for video and text. In this way, the interaction cuts off the confounder effect in $\boldsymbol{V \leftarrow D \rightarrow Y}$ and $\boldsymbol{S \leftarrow D \rightarrow Y}$ as shown in Fig. 2. Take video data as an example, traversing all the confounders for a video is expensive. When the video number in the dataset is $\mathcal{M}$, Eq. 1 expands the training data from $\mathcal{M}$ to $\mathcal{M}^2$ within one epoch with additional memory consumption. Therefore, to balance the scale of confounder set $\mathcal{T}_v$ and training speed, we find the confounder by our MCR and combine it with original data V every epoch. After K training epochs, we can approximate the visual part of the adjustment equation, i.e. $\sum_{\tau_v \in \mathcal{T}_v} P\left(Y|\ V,\tau_v\right) P\left(\tau_v\right) \approx \sum_{k=1}^{K} \sum_{\tau_v \in \mathcal{T}_v{}^k} P\left(Y|\ V,\tau_v\right) P\left(\tau_v\right)$.

Preliminary. Given a video I, VideoQA aims to understand acting events by asking questions q and predict the correct answer $\widetilde{a}$ from a answer candidate set $A = \{a_z\}_{z=1}^{Z}$. Z is the candidate number. The process is generally formu-

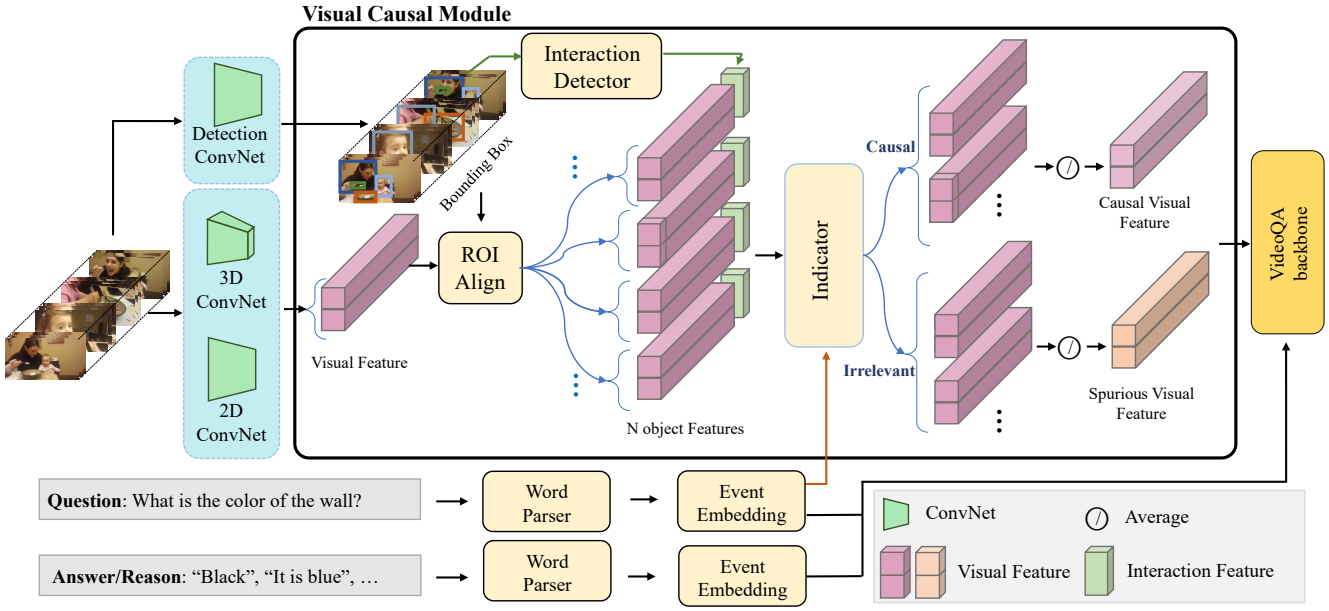

Figure 3. Visual Causal Module. Before the appearance information and motion information are sent to the backbone for question answering, we propose a visual causal module to find object-level causal features and irrelevant features in video data. The image-level visual features are first divided into object-level visual features by the detected bounding box. The category information, location information and interaction relationship of each object are modeled by Interaction Detector. According to the question prompt, the Indicator judges whether the features of different objects are causally related or irrelevant. The uniform representation of causal and irrelevant features are used for further prediction in the VideoQA backbone.

lated as follows:

$$\widetilde{a} = \underset{a \in A}{\operatorname{argmax}} \mathcal{F}_\theta\left(a|q,\ I\right), \tag{2}$$

where θ represents the set of model parameters of VideoQA function $\mathcal{F}$, which maps a pair of video and question into the same feature domain to find the answer. In our work, we do not change the VideoQA function $\mathcal{F}$. In a causal view, we intervene video I and answer candidate A.

3.2. Visual Causal Module

To answer the question with accurate causally related features, we propose a visual causal module that disentangles the confounder and causal factors at the object level. As shown in Fig. 3, the video with shape $W \times H \times L$ is encoded by 2D ConvNet, 3D ConvNet into a list of appearance feature and motion feature with shape $\frac{W}{16} \times \frac{H}{16} \times \frac{L}{2}$. They are uniformly represented by frame feature sequence $\mathcal{V} = \{v_t\}_{t=1}^{\frac{L}{2}}$, where t means timestamp. The object-level grounding is detected by pre-trained Detection ConvNet and donated as the bounding box $\mathcal{B} = \{b_{n,t}\}_{n=1,t=1}^{N,\frac{L}{2}}$, where N is the detected object number. We use Region of Interest (ROI) Align [13] to extract object visual features $\mathcal{O} = \{o_{n,t}\}_{n=1,t=1}^{N,\frac{L}{2}}$, expressed as $\mathcal{O} = ROIAlign\left(\mathcal{V}|\mathcal{B}\right)$.

Since the scene in the real video is ever-changing, some objects in the video may suddenly disappear or appear, making it difficult to obtain accurate instance tracking data. Therefore, we extract unified features for the same category of targets, e.g. human, chairs. When objects of this category do not appear in part of the consecutive frames, we fill the feature with 0 value.

For each object feature, we employ Multilayer Perceptrons (MLP) and Long short-term memory (LSTM) [12] to encode the temporal embedding for visual feature:

$$\mathcal{O}^g = LSTM(MLP(\mathcal{O})), \tag{3}$$

where $\mathcal{O}^g = \{o_n^g\}_{n=1}^N \in \mathbb{R}^d$ is the global N object representations, and d is the hidden dimension of LSTM. In addition to the visual encoding, we also model the bounding box position correlations $\mathcal{B}$, object category C, and spatio-temporal inter-object and intra-object interactions in an Interaction Detecting stream. They are implemented by Non-local Net [30] with MLP:

$$L^g = NonL(MLP(MLP(\mathcal{B}); MLP(C))), \tag{4}$$

where ; represents channel concatenation. $L^g = \{l_n^g\}_{n=1}^N \in \mathbb{R}^d$ means detected interaction feature. The outputs of two feature extraction streams are then fused by channel concatenation: $\mathcal{O}^g = [\mathcal{O}^g; L^g]$. All objects are causal or irrelevant candidates. Their identities are determined by the language semantics of the question. Therefore, given a question sentence, we apply Word Parser and Event Embedding

to get the global question representation q^g. The Event Embedding is implemented by LSTM:

$$q^g = LSTM(Parser(q)), \tag{5}$$

where Parser uses a pre-trained GloVe model [28] or BERT model [6]. LSTM is implemented by the question embedding model of the VideoQA backbone. With the specific question, causally related objects are usually a small part of object set $\mathcal{O}^g$ and are interconnected internally. Reasoning about their relationship to unrelated objects is redundant and prone to spurious associations. Therefore, inspired by [22] ,we use hard segmentation [22] in Indicator to explicitly divide object features into causally related features and irrelevant features, rather than soft GCN inference or attention-based models. Specifically, the question embedding q^g and objects feature $\mathcal{O}^g$ are mapped into the same space by MLP. The Indicator computes a score for each object feature and uses Gumbel-Softmax [17] to classify the causal attribute labels of objects:

$$\begin{aligned} \mathcal{O}^c &= GS(MLP(q^g) \cdot (MLP(\mathcal{O}^g)^T)) \circ \mathcal{O}, \\ \mathcal{O}^{\bar{c}} &= \mathcal{O} - \mathcal{O}^c, \end{aligned} \tag{6}$$

where GS means Gumbel-Softmax and $\cdot$ means matrix multiplication. $\circ$ means multiply by index. $\mathcal{O}^c = \{{o^c}_n\}_{n^c=1}^{N^c}, \mathcal{O}^{\bar{c}} = \{{o^{\bar{c}}}_{n^{\bar{c}}}\}_{n^{\bar{c}}=1}^{N^{\bar{c}}}$ are causal object features set and irrelevant features set, respectively, where the corner mark of length is omitted for clarity. We average the features in these sets to obtain causal visual features and irrelevant visual features, respectively: $\mathcal{V}^c = \frac{1}{N^c}\sum \mathcal{O}^c$, $\mathcal{V}^{\bar{c}} = \frac{1}{N^{\bar{c}}}\sum \mathcal{O}^{\bar{c}}$.

According to the deduction of Eq. 1, in the visual part, the video V and the average sampling confounder are combined to predict the label. In model designing, the video V can be present by causal object features $\mathcal{V}^c$. The confounder is from the confounder of another example in the same mini-batch, represented by:

$$\hat{\mathcal{V}} = \frac{1}{N^c + N'^{\bar{c}}} \sum (O^c + O'^{\bar{c}}), \tag{7}$$

where $\hat{\mathcal{V}}$ is the blended features. $O'^{\bar{c}}$ is the confounder of another example. $N'^{\bar{c}}$ is the object number of $O'^{\bar{c}}$. Both $\mathcal{V}$, $\mathcal{V}^c$, $\mathcal{V}^{\bar{c}}$, and $\hat{\mathcal{V}}$ are sent to the VideoQA backbone for training the visual causal module. The visual module can capture causal relationships between object features and labels, providing a stable association for multimodal data.

3.3. Textual Causal Module

For textual causal inference, prediction is often plagued by correlations between local language semantics and visual features. Therefore, the key to getting rid of text confounders is to enhance the model's ability to recognize

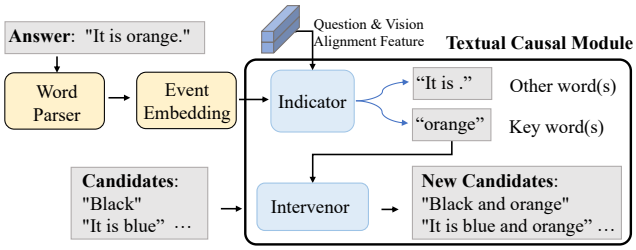

Figure 4. Textual Causal Module. Keywords are selected as confounders from the prediction results by the indicator. They are sent to the intervenor to generate new candidates.

such confusing samples which are implemented by a textual causal module.

In the VideoQA base model, the visual information and the question have been aligned, containing the events information related to the answer. Therefore, we use the alignment feature $f_{qv} = \mathcal{F}(\mathcal{V}^c,\ q^g)$ to retrieve the keywords in the answer. As shown in Fig. 4, the same with question embedding, we encode the answer by Word Parser and Event Embedding:

$$a^g, a^l = LSTM(Parser(a)), \tag{8}$$

where a^g is the global representation of the answer, and a^l is the local representation of each word. The same with visual Indicator, we select the keyword by the relation score between a^g and f_{qv}:

$$a^{\bar{c}} = GS(MLP(f_{qv}) \cdot (MLP(a^g))^T) \circ a^l, \tag{9}$$

where $a^{\bar{c}}$ represents the keywords, which are potential confounders in this question. $a^c = a^l - a^{\bar{c}}$ represents other words.

According to the textual part of Eq. 1, the label is predicted by the sentence and confounder. Due to the correlation between sentences and videos, we look for confounders of sentences from the same video, especially the predicted answer by VideoQA backbone which contains factors that increase the score. Similar to visual intervention, we pick confounders once per epoch.

Confounders are used to augment the hard sample of candidates that changes the textual distributions. The model joint original candidates $A = \{a_z\}_{z=1}^{Z}$ with $a^{\bar{c}}$ by inserting $a^{\bar{c}}$ at the beginning or end of a sentence and generate a new candidates set $\hat{A} = \{\hat{a}_z\}_{z=1}^{Z}$. In the training process, we randomly replace the negative example from the new candidates set $\hat{A}$.

3.4. Training

In this section, we introduce our training targets and training pipeline for multimodal data. The algorithm can be found in Supplementary. First of all, to ensure the domain consistency of training sets and test sets, we train

the VideoQA backbone with raw data $(\mathcal{V}, \mathrm{A})$, which also guarantees that the training process satisfies the pre-defined causal formulation:

$$\mathcal{L}_o = \mathrm{XE}\left(\mathcal{F}\left((\mathcal{V}, \mathrm{q}), (\mathcal{V}, \mathrm{A})\right), \mathrm{a}\right), \tag{10}$$

where XE is the cross-entropy loss. $\mathcal{F}$ calculates the dot-product between question-related visual features and answer-related visual features as the predicted results. At test time, we use the candidate with the highest score as the predicted result as in Eq.2.

Visual Module Loss. Visual module loss aims to capture causal object features and non-causal object features for further robust prediction. In the training stage, the causal visual features have the same ability for question answering, which is restricted by the cross-entropy loss: $\mathcal{L}_{\mathrm{c}}^{\mathrm{v}}$. In contrast, non-causal objects are irrelevant to the question and should be unbiased toward answer candidates: $\mathcal{L}_{\bar{c}}^{\mathrm{v}}$. The blended data $\hat{\mathcal{V}}$ are supposed to obtain the same effect with causal feature: $\mathcal{L}_{\hat{o}}^{\mathrm{v}}$. The formulation of visual module loss shows as follows:

$$\begin{aligned}
\mathcal{L}_{\mathrm{c}}^{\mathrm{v}} &= \mathrm{XE}\left(\mathcal{F}\left((\mathcal{V}^{\mathrm{c}}, \mathrm{q}), (\mathcal{V}^{\mathrm{c}}, \mathrm{A})\right), \mathrm{a}\right) \\
\mathcal{L}_{\bar{c}}^{\mathrm{v}} &= \mathrm{MSE}\left(\mathcal{F}\left((\mathcal{V}^{\bar{c}}, \mathrm{q}), (\mathcal{V}^{\bar{c}}, \mathrm{A})\right), \mathrm{avg}\right) \\
\mathcal{L}_{\hat{o}}^{\mathrm{v}} &= \mathrm{XE}(\mathcal{F}((\hat{\mathcal{V}}, q), (\hat{\mathcal{V}}, A)), a), \\
\mathcal{L}_{\mathrm{v}} &= \mathcal{L}_{\mathrm{c}}^{\mathrm{v}} + \mathcal{L}_{\bar{c}}^{\mathrm{v}} + \mathcal{L}_{\hat{o}}^{\mathrm{v}},
\end{aligned} \tag{11}$$

where avg is the average score of all candidates, $avg = \mathcal{F}\left((\mathcal{V}, q), (\mathcal{V}, A)\right)$. MSE means Mean Square Error for non-causal objects to get a neutral score.

Textual Module Loss. The text module loss aims to find keywords from the prediction results that directly affect the model selection and reduce the sensitivity of the model to keywords. Intuitively, keywords have the ability for prediction, while other words are difficult to judge the results. We use cross-entropy loss for keywords $\mathcal{L}_{\bar{c}}^{\mathrm{s}}$ and MSE for other words $\mathcal{L}_{\mathrm{c}}^{\mathrm{s}}$. The blended answer is restricted by the cross-entropy loss $\mathcal{L}_{\hat{o}}^{\mathrm{s}}$. The formulation of textual module loss is:

$$\begin{aligned}
\mathcal{L}_{\bar{c}}^{\mathrm{s}} &= \mathrm{XE}\left(\mathcal{F}\left((\mathcal{V}, \mathrm{q}), (\mathcal{V}, a^{\bar{\mathrm{c}}})\right), \mathrm{a}\right) \\
\mathcal{L}_{\mathrm{c}}^{\mathrm{s}} &= \mathrm{MSE}\left(\mathcal{F}\left((\mathcal{V}, \mathrm{q}), (\mathcal{V}, a^{\mathrm{c}})\right), avg\right) \\
\mathcal{L}_{\hat{o}}^{\mathrm{s}} &= \mathrm{XE}(\mathcal{F}((\mathcal{V}, q), (\mathcal{V}, \hat{A})), a) \\
\mathcal{L}_{\mathrm{s}} &= \mathcal{L}_{\mathrm{c}}^{\mathrm{s}} + \mathcal{L}_{\bar{c}}^{\mathrm{s}} + \mathcal{L}_{\hat{o}}^{\mathrm{s}}.
\end{aligned} \tag{12}$$

The overall loss of our model can be expressed as:

$$\mathcal{L} = \mathcal{L}_o + \lambda_1 \mathcal{L}_{\mathrm{v}} + \lambda_2 \mathcal{L}_{\mathrm{s}}, \tag{13}$$

where λ_1 and λ_2 are the hyper-parameters to balance the visual loss and textual loss.

Intervene Pipeline. We show the Intervene pipeline in Fig. 5. The textual intervention and the visual intervention are accomplished in two steps. Visual causal loss is computed with Eq. 11 and used to obtain causally related object features and irrelevant object features. They are used to generate blended visual data with an intervenor, which is implemented by Eq. 7. Textual causal loss is computed with Eq. 12 and used to obtain keywords. Keywords are sent to textual interveners to generate blended textual data. We execute K_2 epochs textual interventions after executing K_1 epochs visual interventions for the visual causal module and the textual causal module.

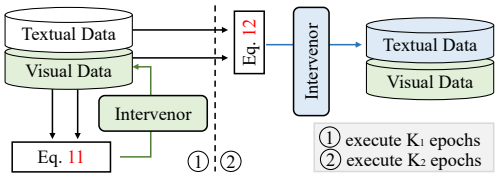

Figure 5. Causal intervention pipeline.

4. Experiment

4.1. Dataset

We evaluate our Multimodal Causal Reasoning (MCR) framework on two recent large causal-related VideoQA datasets: Causal-VidQA [21] and NExT-QA [32]. In these datasets, models need to answer not only simple descriptive or statistical questions but also implicit global evidence reasoning. The collected video data is close to the real open scene, including a variety of objects and interactive actions. Answer candidates are sentences with different expression structures and are not limited to a single sentence.

The Causal-VidQA dataset selects 26,900 video clips from Kinetics-700 [2] and asks 107,600 questions, including description, explanation, prediction, and counterfactual questions. Each question has 5 answers, with 5 more reasons in predictions and counterfactuals. For question-answer accuracies, we adopt the previous causal VideoQA evaluation metrics [21] that report accuracy for each question type, as well as the accuracy of the consistency between answers and reasons.

NExT-QA dataset contains 5,440 videos from VidOR [29] and proposes 47,692 questions for the multi-choice task, including description, explanation, and temporal reasoning questions. We also report accuracy for each question type.

4.2. Experimental Settings

We use a unified feature extraction method for both datasets. Each video with varying length is divided into 8 clips, each containing 16 frames. For the object-level motion feature, we uniformly sample from the frame-level appearance features and clip-level motion features by ROI Align [13]. The bounding boxes in each frame are extracted by pre-trained Mask R-CNN [13]. The appearance features and motion features are extracted by ResNet-101 [14] and

Table 1. Comparison of accuracy with state-of-the-art methods on Causal-VidQA dataset. D means description question. E means explanation question. P means prediction question. C means counterfactual question. QA means to answer the answer. QR means to answer the reason. QAR means the answer and reason are both accurate.

Method	Text Feature	Acc-D	Acc-E	Acc-P			Acc-C			Acc
				QA	QR	QAR	QA	QR	QAR	
HME [8]	GloVe	47.25	43.80	41.02	42.53	23.25	35.29	34.19	15.34	32.41
HCRN [20]	GloVe	58.89	53.53	43.14	45.07	26.17	43.69	43.47	22.75	40.33
HGA [18]	GloVe	60.32	55.02	46.55	47.21	28.53	44.00	44.04	23.63	41.88
B2A [26]	GloVe	61.29	56.43	46.82	48.17	30.01	45.12	44.99	25.29	43.26
IGV [22]+B2A	GloVe	59.24	48.41	45.70	47.32	28.20	38.99	40.97	18.84	38.67
MCR+B2A	GloVe	**66.72**	**61.26**	**50.46**	**52.17**	**32.13**	**52.17**	**51.50**	**31.91**	**48.01**
HME [8]	BERT	63..36	61.45	50.29	47.56	28.92	50.38	. 51.65	30.93	46.16
HCRN [20]	BERT	65.35	61.61	51.74	51.26	32.57	51.57	53.44	32.66	48.05
HGA [18]	BERT	65.67	63.51	49.36	50.62	32.22	52.44	55.85	34.28	48.92
B2A [26]	BERT	66.21	62.92	48.96	50.22	31.15	**53.27**	**56.27**	**35.16**	49.11
IGV [22]+B2A	BERT	65.92	62.13	52.77	53.47	35.00	50.67	52.29	31.22	48.57
MCR+HCRN	BERT	**68.68**	**65.97**	55.44	**58.18**	37.63	52.24	52.39	31.17	50.86
MCR+B2A	BERT	67.47	65.59	**56.46**	56.42	**37.82**	52.39	54.08	33.38	**51.06**

3D ResNeXt-101 [34] with the pre-trained model, respectively. For textual data, word token representation for questions and answers is provided by GloVe [28] and BERT [6] respectively, which greatly affects the performance of models [21]. MLP is implemented by fully connected layers, followed by Batch Normalization [16] and ReLU [11]. λ_1 and λ_2 are both set as 1. K_1 and K_2 are set as 1 and 3, respectively. During training, we use Adam optimizer with the initial learning rate of 1e-4 and halve the learning rate in every 5 epochs. The batch size is set as 128 in Causal-VidQA dataset and 64 in NExT-QA dataset.

4.3. Comparision with State-of-the-Arts

Results on Causal-QA. Table 1 presents the results of four state-of-the-art baseline methods and one causal method on the Causal-VidQA dataset, including HME [8], HCRN [20], HGA [18], B2A [26], and IGV [22]. Compared to all existing VideoQA methods, MCR achieves the best performance across almost all types of questions. It is worth noting that MCR improves accuracy across all question types when using GloVe as the text model, which has an average performance improvement of 4.75% over B2A (48.01% vs. 43.26%). Compared with the IGV, MCR improves by 9.44% (48.01% vs. 38.67%). Our result is even comparable to the backbones of using BERT as a language model. When using BERT as the text model and B2A as the VideoQA backbone, MCR achieves an accuracy improvement of 1.85% on average. In counterfactual questions, it is difficult to answer the content of the associative question with limited knowledge learning from the dataset. Involving additional commonsense knowledge may be helpful. Experimentally, in the absence of commonsense knowledge, VideoQA backbones that rely on data bias can also achieve good performance for counterfactual questions.

Table 2. Accuracies on NExT-VidQA of different architectures.

Models	Causal	Temp	Descrip	All
HME [8]	46.76	48.89	57.37	49.16
HCRN [20]	47.07	49.27	54.02	48.82
HGA [18]	48.13	49.08	57.79	50.01
B2A [26]	47.37	49.01	58.3	49.60
IGV [22]	48.56	51.67	59.64	51.34
MCR+B2A	47.3	50.25	61.26	50.42
MCR+HGA	**49.19**	**51.98**	**62.29**	**52.35**

Our proposed framework is in parallel with the VideoQA backbone that helps it discover the causal reason. To verify the effectiveness of the parallel strategy, we combine the MCR with different backbones, including HCRN [20] and B2A [26]. We can see that our method has remarkable effects on existing methods. This is due to previous methods being confused by confounders in multimodal data, resulting in poor generalization. Our MCR can effectively alleviate the biased modeling of these methods.

Results on NExT-QA. For further comparison, we evaluate our method on the NExT-QA dataset and report the evaluation results in Table 2. Compared with the baselines, our proposed MCR achieves the top performance in all question types. Specifically, when using B2A as the VideoQA backbone, our MCR is able to improve 0.82% on average accuracy. When combined with HGA, our MCR surpasses the previous backbone and causal method by clear margins, e.g. 2.34% and 1.01% higher than HGA and IGV, respectively. These validate that our multimodal causal reasoning indeed improves the robustness and generalization of baseline on causal reasoning dataset.

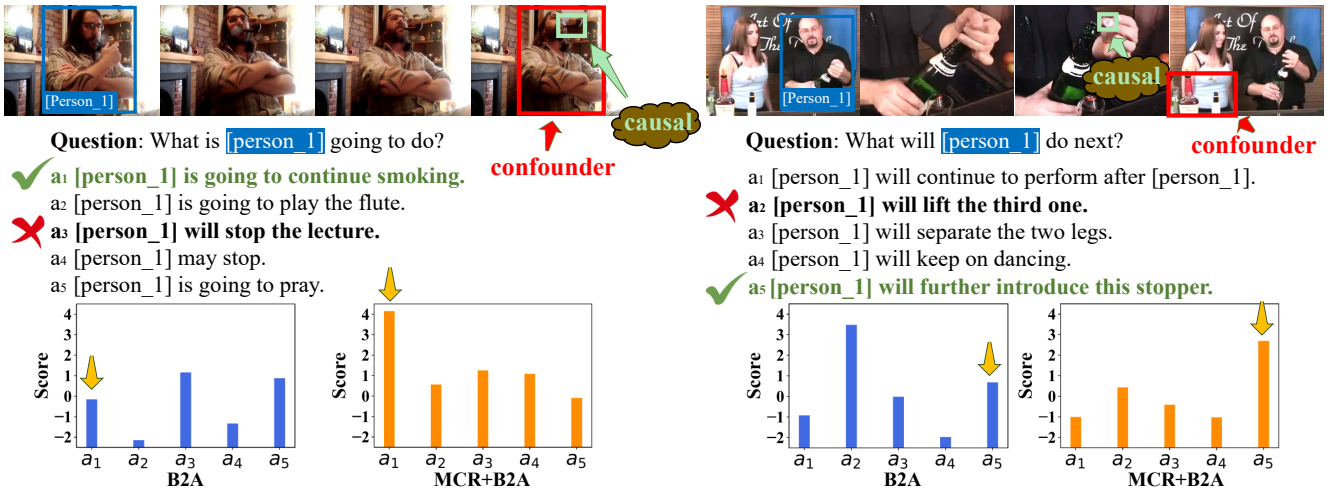

Figure 6. Qualitative comparison on Causal dataset validation split. The green bold answer denotes the ground-truth one. The yellow arrows indicate that our method can improve the score of predictions by finding the causal association.

4.4. Ablation Study

Table 3. Performance comparisons of different variants on Causal-QA dataset. "MCRv" and "MCRt" denote the visual causal module of MCR and the textual module of MCR, respectively.

Method	Acc-D	Acc-E	Acc-P	Acc-C	Acc
B2A	66.21	62.92	31.15	**35.16**	49.11
B2A+MCRv	67.76	65.38	35.55	32.95	50.41
B2A+MCRt	68.57	64.38	33.73	31.61	49.57
B2A+MCR	**68.72**	**65.01**	**36.92**	33.21	**50.96**

We analyze the effectiveness of different modules of MCR in Table 3. When adopting the visual causal module of MCR, we can achieve comparable improvements (1.30% on average), indicating that the visual indicator can change the co-occurrence relationship between irrelevant objects and labels, which improves the robustness of the model to visual data. With only MCR's text module, B2A improves by an average of 0.46%. This is because we eliminate the misleading of local language semantics attention. When adopting both causal modules for multimodal data, B2A+MCR achieves the best performance.

Table 4. The influence of hype-parameters of training pipeline on Causal-VideoQA Dataset.

k_1/k_2	1	3	5	1/3	1/5
Acc	46.65	**50.58**	49.58	48.23	46.37

Hyper-parameters. In this paper, we propose a causal intervention pipeline for multimodal data. Here, we conduct ablation studies on the hyper-parameters setting in Tab. 4. When we interact textual data every 3 epochs and video data every 1 epoch, MCR performs best on average accuracy. More Hyper-parameters about λ_1, λ_2 and the ablation study of single loss function are shown in Supplementary.

4.5. Qualitative Analysis

Fig. 6 shows the qualitative comparison between our MCR and B2A. Our MCR can reduce the scores of confusing candidates and enhance the confidence for the accurate answer. In the example on the left, B2A chooses the third answer because both the body pose of the human and the environment are related to "lecture". Our approach explicitly helps B2A reduce the influence of irrelevant factors and find the causal motion "smoking". In the right example, B2A is perturbed by the correlation between "third" in the text and the three bottles on the table in the video. According to the causal object selection, MCR effectively arrives at a lower score for it. These examples demonstrate that our model can perform real and generalizable reasoning.

5. Conclusion

In this paper, we revisit causal effects in multimodal data and propose a causal prediction architecture to model the causal association between video and text for the VideoQA task. Compared with previous methods, MCR can modify the distribution of data according to the backdoor adjustment and improve the robustness of the model. Considering limitations, our method intervenes in the textual data by word insertion. Some post-intervention examples express eccentric sentence structures that are easy to distinguish. Intervening the textual data with a reasonable text generator would be a reasonable future work. Besides, MCR cannot be directly adapted to the Video Story QA task in which most videos are human-human interactions rather than human-object interactions. Enhancing interventional operations for human instances may be another future work.

Acknowledgement This research is supported by the National Natural Science Foundation of China (NSFC) (No. 61972038).

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[3] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. iperceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. *arXiv preprint arXiv:2011.07735*, 2020.

[4] Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B Tenenbaum, and Chuang Gan. Grounding physical concepts of objects and events through dynamic visual reasoning. *arXiv preprint arXiv:2103.16564*, 2021.

[5] Long Hoang Dang, Thao Minh Le, Vuong Le, and Truyen Tran. Hierarchical object-oriented spatio-temporal reasoning for video question answering. *arXiv preprint arXiv:2106.13432*, 2021.

[6] 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.

[7] Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, and Chuang Gan. Dynamic visual reasoning by learning differentiable physics models from video and language. *Advances In Neural Information Processing Systems*, 34:887–899, 2021.

[8] 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.

[9] 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.

[10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[11] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[12] Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[14] 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.

[15] 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.

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

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

[18] 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.

[19] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Yuhang Liu, Wei Wei, Daowan Peng, and Feida Zhu. Declaration-based prompt tuning for visual question answering. *arXiv preprint arXiv:2205.02456*, 2022.

[25] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[26] 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.

[27] Judea Pearl. *Causality*. Cambridge university press, 2009.

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

[29] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287, 2019.

[30] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[31] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022.

[32] 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.

[33] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. AAAI, 2022.

[34] 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.

[35] 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.

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

[37] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021.

[38] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155*, 2022.

[39] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

[40] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018.

[41] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021.

[42] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[43] Zhou Zhao, Jinghao Lin, Xinghua Jiang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical dual-level attention network learning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1050–1058, 2017.