

Equivariant and Invariant Grounding for Video Question Answering

Yicong Li¹, Xiang Wang^{2*}, Junbin Xiao¹, and Tat-Seng Chua¹

¹National University of Singapore, ²University of Science and Technology of China
liyicong@u.nus.edu,xiangwang1223@gmail.com,junbin@comp.nus.edu.sg,dcscts@nus.edu.sg

ABSTRACT

Video Question Answering (VideoQA) is the task of answering the natural language questions about a video. Producing an answer requires understanding the interplay across visual scenes in video and linguistic semantics in question. However, most leading VideoQA models work as black boxes, which make the visual-linguistic alignment behind the answering process obscure. Such black-box nature calls for visual explainability that reveals “What part of the video should the model look at to answer the question?” Only a few works present the visual explanations in a post-hoc fashion, which emulates the target model’s answering process via an additional method. Nonetheless, the emulation struggles to faithfully exhibit the visual-linguistic alignment during answering.

Instead of post-hoc explainability, we focus on intrinsic interpretability to make the answering process transparent. At its core is grounding the question-critical cues as the causal scene to yield answers, while rolling out the question-irrelevant information as the environment scene. Taking a causal look at VideoQA, we devise a self-interpretable framework, Equivariant and Invariant Grounding for Interpretable VideoQA (EIGV). Specifically, the equivariant grounding encourages the answering to be sensitive to the semantic changes in the causal scene and question; in contrast, the invariant grounding enforces the answering to be insensitive to the changes in the environment scene. By imposing them on the answering process, EIGV is able to distinguish the causal scene from the environment information, and explicitly present the visual-linguistic alignment. Extensive experiments on three benchmark datasets justify the superiority of EIGV in terms of accuracy and visual interpretability over the leading baselines. Our code is available at <https://github.com/yl3800/EIGV>.

CCS CONCEPTS

- Information systems → Question answering: Multimedia and multimodal retrieval.

KEYWORDS

Video Question Answering, Invariant Learning, Equivariant Learning, Interpretability

* Corresponding author. This research is supported by the Sea-NExT Joint Lab, and the CCCD Key Lab of Ministry of Culture and Tourism, USTC..



This work is licensed under a Creative Commons Attribution International 4.0 License.

ACM Reference Format:

Yicong Li, Xiang Wang, Junbin Xiao, and Tat-Seng Chua. 2022. Equivariant and Invariant Grounding for Video Question Answering. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Oct. 10–14, 2022, Lisboa, Portugal*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3503161.3548035>

1 INTRODUCTION

Video Question Answering (VideoQA) [54] is a keystone in interactive AI, such as vision-language navigation and communication systems. It aims to answer the natural language question based on the video content. Striving for the architecture novelty, many studies have been conducted on modeling VideoQA’s multi-modal nature, such as fostering the vision-language alignment [17, 26] and revisiting the visual input structure [6, 19]. However, existing VideoQA models usually operate as black boxes, which fail to exhibit the working mechanism behind the predictions and hardly exhibit “What knowledge should the model use to answer the question about the video?”. As a result, the black-box nature causes concern for the model’s reliability, especially in applications to safety and security.

The concern on the black-box nature calls for better transparency of VideoQA models. Here we focus on visual-explainability [4, 33], aiming to reveal “Which part of the video should the model look at to answer the question?”. It requires us to find a subset of visual scenes – rationale – that support the answering as evidence in way of human interpretation [33]. Taking Figure 1 as an example, when answering the question “What is the girl doing?”, the rationale should focus on the “girl-riding on-horse” scene in the first two clips. Towards this end, existing studies [8, 23, 41] dwell mainly on the paradigm of **post-hoc explainability** [32, 35], which distributes the predictive answer of the target model to the input visual features via an additional explainer method. They visualize the attention weights or gradient-like signals toward the visual features, and then identify a salient pattern as the rationale. However, post-hoc explainability has several major limitations: (1) It fails to make the target model intrinsically interpretable [34, 43, 51], only approximating the decision-making process of the model. As a result, the identified rationale cannot faithfully reveal how the model leverages the multi-modal information. (2) Such visual inspections are fragile against input perturbations, since some artifacts can be easily captured as explanations instead of genuine knowledge from the data [9, 12, 18, 38].

The limitations of post-hoc explainability inspire us to explore the paradigm of **intrinsic interpretability** [9, 34], which embeds a rationalization module into the model to make the decision-making process transparent. Surprisingly, the intrinsic interpretability of VideoQA models is until-now lacking. To fill the void, we draw on **causal theory** [27, 29] to formulate the interpretability task as

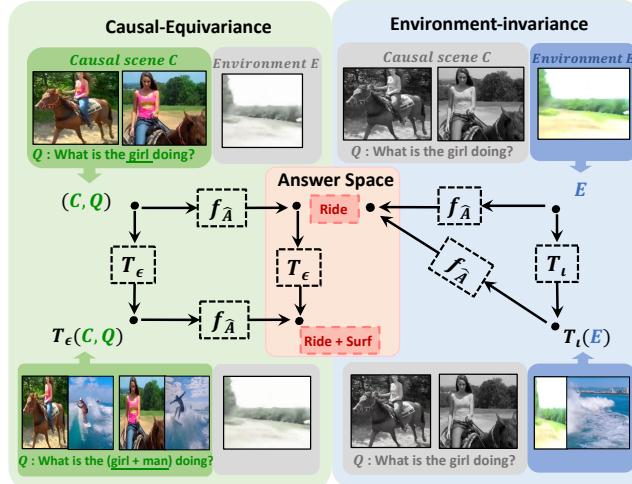


Figure 1: Illustration of equivariant and invariant grounding. The causal-equivariant principle (left) asks that the semantic change T_ϵ applied to the causal scene C and question Q should be faithfully reflected in the answer change. In contrast, the environment-invariant principle (right) outputs the same answer, regardless of changes T_i on the environment scene E . Here, $f_{\hat{A}}$ maps input to answer space.

disclosing “Which part of the video is critical/causal to answering the question?”. Concretely, we aim to identify the causal component of input video on-the-fly, which holds the question-response information and filters out the question-irrelevant cues. Following this essence, one straightforward realization is to ground the input video into two segments: (1) **causal scene**, which retains the question-critical visual content and sufficiently approaches the answer, thus naturally serving as the rationale; and (2) **environment scene**, which holds the question-irrelevant visual content and can be seen as the rationale’s complement.

However, discovering causal scene without the supervision of ground-truth rationale is challenging. With a causal look at the reasoning process (*cf.* Section 3.1), we argue that the crux of intrinsic interpretability is to amplify the connection between the causal scene and the answer, while blocking the non-causal effect of the environment scene. Following this line, we propose two principles to guide the grounding of the rationale:

- **Causal-Equivariance.** By “equivariance”, we mean that answering should be sensitive to the semantic changes on the causal scene and question (termed E-intervention), *e.g.*, any change on the causal scene and question should be faithfully reflected on the predicted answer. For example, in Figure 1, the “girl-riding on-horse” and “man-surfing in-ocean” scenes are the oracle rationales of “What is the girl doing?” and “What is the man doing?”, respectively. The intervention [22] applied on the input (*i.e.*, mixing the “girl-riding on-horse” and “man-surfing in-ocean” scene, and combining two questions as “What is the girl doing? What is the man doing?”) should set off an equivariant change in the answer (*i.e.*, changing from “Ride” to “Ride+Surf”).
- **Environment-Invariance.** By “invariance”, we mean that answering should be insensitive to the changes in the environment scene (termed I-intervention), conditioning on the causal scene

and question. Considering Figure 1 again, the intervention applied to the environment (*i.e.*, mixing the “meadow” and “ocean” scenes) implies no impact towards answering “What is the girl doing?”, reflecting a homogeneity in the answer space.

To impose these two principles for intrinsic interpretability, we propose a new framework, Equivariant and Invariant Grounding for Interpretable VideoQA (EIGV). EIGV equips the VideoQA backbone model with three additional modules: a grounding indicator, an intervener, and a disruptor. First, the grounding indicator learns to attend the causal scene based on the input question, while leaving the rest as the environment. Then, the intervener parameterizes the proposed principles to guide the grounding. Specifically, towards the causal-equivariance principle, it conducts the E-intervention on the causal scene and question — that is, mix them with the counterparts from another video-question pair — and encourages the predictive answer to be anticipated accordingly. Towards the environment-invariance principle, when leaving the causal scene and question untouched, it applies the I-intervention on the environment — that is, mix it with the environmental stratification of a memory bank — and enforces the predictive answer to be invariant. Moreover, we build an unified sight of two principles via the lens of contrastive learning. Concretely, on top of each intervened video-question pair, the disruptor constructs the positive views by disrupting the environment scene randomly, while creating the negative views by substituting the causal scene with random scenes. Training with these two principles allows the backbone model to distinguish the causal scene from the environmental cues, and hinge on the critical visual-linguistic alignment.

Briefly put, our contributions are:

- We propose EIGV, a model-agnostic VideoQA framework that distills the causal visual-linguistic alignment to generate answers in a self-interpretable manner.
- We investigate the soundness of grounding rationale by posing the equivariant-invariant principle on visual grounding.
- We justify the superiority of EIGV on three popular benchmark datasets (*i.e.*, MSVD-QA [50], MSRTT-QA [50], NExT-QA [48]) with extensive experiments, where our design outmatches the state-of-the-art models. Moreover, our EIGV is a model-agnostic framework that can be applied to different VideoQA models.

2 PRELIMINARIES

Here we provide a holistic view of VideoQA by summarizing a common paradigm throughout existing works. Specifically, we denote a variable and its deterministic value by upper-cased (*e.g.*, A) and lower-cased (*e.g.*, a) letters, respectively.

Modeling. Given the video V , the VideoQA model $f_{\hat{A}}(\cdot)$ answers the question Q by formulating the visual-linguistic alignment:

$$\hat{A} = f_{\hat{A}}(V, Q), \quad (1)$$

where \hat{A} is the predictive answer. Typically, $f_{\hat{A}}(\cdot)$ is a combination of two modules:

- Video-question encoder, which warps up the visual content and linguistic semantics via two encoders: (1) the video encoder capsules the video content by methods like hierarchical design [6, 19, 30], enhanced memory architecture [7, 8] and structural

- graph representation [10, 14, 17, 46]; (2) the question encoder embeds the contextual information into linguistic representation through multi-scale semantic integration [17, 36, 42] or grammatical dependencies parsing [26].
- Answer decoder, which abridges the encoded visual-linguistic information via cross-modal interaction methods like graph alignment [26] and progressive attention [30, 36], then generates the prediction accordingly.

Learning. To optimize the video-question encoder and answer decoder, current VideoQA models usually adopt the scheme of empirical risk minimization (ERM) [8, 17, 19, 30], which measures and minimizes the risk between the ground-truth answer A and predictive answer \hat{A} :

$$\min \mathcal{L}_{\text{ERM}}(\hat{A}, A). \quad (2)$$

In essence, ERM recklessly takes the video content as a whole and enforces the risk deduction over compassion of question and every video frame, which hardly discovers a reliable interpretation to exhibit the visual-linguistic alignment.

3 VIDEOQA REFORMULATION

Here we argue that disclosing “Which part of the video is critical to answering the question?” is the key to presenting the visual-linguistic alignment explicitly. To this end, we take a causal [27] look at the reasoning process of VideoQA, then formalize it as a Structure Causal Model (SCM) [29] by investigating the causal relationships among five variables: input video V , question Q , causal scene C , environment scene E , ground-truth answer A .

3.1 Causal Graph of VideoQA

Figure 2 illustrates the causal graph, where each link depicts the cause-effect relationship between two variables:

- $Q \rightarrow C, E \leftarrow V$. Given the question of interest Q , the video V can be partitioned into two parts: (1) the causal scene C , which retains the question-critical information and naturally serves as the rationale for answering, (2) the environment scene E , which gathers the cues irrelevant to the question-answering. For example, to answering “What is the girl doing?” in Figure 1, C should be the first two clips describing the “girl-riding on-horse” scene, while E should be the last clip about the “meadow” scene. Moreover, the varying semantics of different questions will emphasize different C .
- $Q \rightarrow A \leftarrow C$. The visual knowledge in the causal scene C and the linguistic semantics in the question Q collaborate together to determine the answer A . Furthermore, this path, which presents the visual-linguistic alignment, internally interprets the reasoning.
- $E \dashrightarrow C$. The dashed arrow sketches additional probabilistic dependencies [28] between C and E , which typically arise from selection bias [39]. For example, the “meadow” scene is frequently collected as a common environment for the “horse-riding” scene.

3.2 Beyond ERM

With inspections on prior VideoQA studies, we investigate their inability to distinguish the causal and non-causal effects of scenes. Specifically, in conventional VideoQA models, video and question

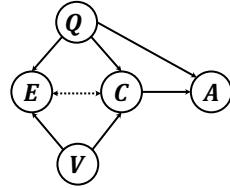


Figure 2: Causal Graph of VideoQA

are directly paired together to model their interaction and approach the golden answer, consequently. Inevitably, taking video as a whole leaves the contributions of scenes untouched, thus failing to differentiate C from E and forgoing their function divergence towards the answer. Worse still, ERM enforces these models to blindly capture the statistical correlations between the video-question pairs and answers. As such, the visual-linguistic alignment hinges easily on the spurious correlations between E and A , owing to the backdoor paths [29], which hinders the generalization of models [25, 47]. Therefore, identifying the causal scene C is the critical to addressing these limitations.

4 METHODOLOGY

To ground the causal scene C in the video V , we take a closer look at the VideoQA SCM (*i.e.*, Figure 2a), and emphasize the essential differences between C and E . Specifically, given the causal scene C and question Q , the answer A is determined, regardless of the variations in the environment scene E :

$$A \perp E \mid C, Q, \quad (3)$$

where \perp denotes the probabilistic independence.

Rationalization. During training, the oracle grounding rationale C is out of reach, while only the input (V, Q) pair and training target A are observed. Such an absence motivates VideoQA to embrace video grounding in its modeling. Specifically, in light of question Q , the estimated causal scene \hat{C} is grounded from the massive V to approach the oracle C and then generate prediction \hat{A} via $Q \rightarrow A \leftarrow C$. To systematize this relation, the causal-equivariance principle introduces an equivariant transformation T_e to each of the parent variables (*i.e.*, C and Q), and expects a proportionate change in the response variable (*i.e.*, A). On top of SCM, we formally present such notions as:

$$T_e(\hat{A}) = f_{\hat{A}}(T_e(\hat{C}), T_e(Q)). \quad (4)$$

Meanwhile, environment-invariant principle formulated Equation (3) in the sense that imposing an invariant-transformation T_i on the estimated environment \hat{E} should not trigger variation of answer A :

$$\hat{A} = f_{\hat{A}}(T_i(\hat{E}), Q), \quad (5)$$

To this end, we parameterize our learning framework, EIGV, as a combination of equivariant and invariant principles, which comprises three additional modules on top of the ERM-guided backbone: grounding indicator, intervener, and disruptor. In a nutshell, we display our EIGV framework in Figure 3.

Data representation. Following previous efforts [10, 17], we encode video instance v as a sequence of K fixed visual clips, while question instance q is encoded into a similar form with a fixed length of language tokens L . Then, visual and linguistic features

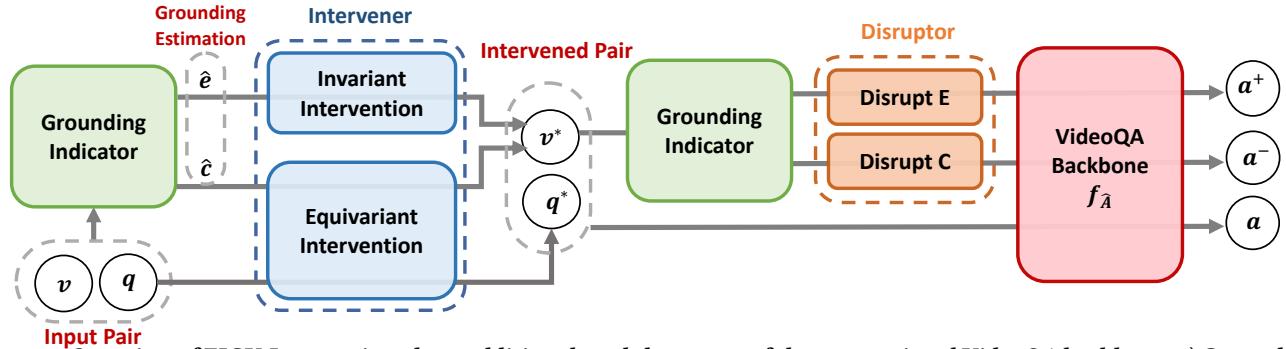


Figure 3: Overview of EIGV. It comprises three additional modules on top of the conventional VideoQA backbone: 1) Grounding indicator, 2) Intervener, and 3) Disrupter. First, the grounding indicator learns the estimation of causal scene \hat{c} and environment \hat{e} . Next, two interventions are imposed on the causal and non-causal factors to compose the intervened pair (v^*, q^*) . Finally, based on the re-grounded result, the disruptor creates contrastive samples, which are further feed into the VideoQA backbone.

are applied with a linear layer and an LSTM [13], respectively, to align their dimension. As a result, we acquire the output of linear layer $v \in \mathbb{R}^{k \times d}$ as the final video representation and the last hidden state of LSTM $q \in \mathbb{R}^d$ as the holistic question representation.

4.1 Grounding Indicator

Scene partition is fundamental to the rationale discovery, whose core is to estimate the value of C and E via a hard split on video V . Given an input sample (v, q) , the grounding indicator aims to access the causal scene and environment scene via their estimated value \hat{c} and \hat{e} according to question Q . Concretely, we first construct two cross-modal attention modules to indicate the probability of each visual clip of being causal scene ($p_{\hat{c}} \in \mathbb{R}^K$) and environment scene ($p_{\hat{e}} \in \mathbb{R}^K$):

$$p_{\hat{c}} = \text{Softmax}(\text{FC}_1(v) \cdot \text{FC}_2(q)^T), \quad (6)$$

$$p_{\hat{e}} = \text{Softmax}(\text{FC}_3(v) \cdot \text{FC}_4(q)^T), \quad (7)$$

where $\text{FC}_1, \text{FC}_2, \text{FC}_3, \text{FC}_4$ are fully connected layers that align cross-modal representations. However, gathering messages via a soft mask still makes the visual information on different clips overlap. As discussed in Section 3.2, guided by ERM, the conventional attention mechanism is unable to block the influence of \hat{e} , thus undermining the veracity of \hat{c} . As a correction, the grounding indicator makes a discrete selection over the clip-wise attention result to generate a disjoint group of the causal scene. We leverage Gumbel-Softmax [16] to manage a differentiable selection on attentive probabilities and compute the indicator vector $I \in \mathbb{R}^{K \times 2}$ on the two attention scores over each clip (i.e., $p_{\hat{c},i}, p_{\hat{e},i}; i \in K$). Formally, I is derived as:

$$I = \text{Gumbel-Softmax}([p_{\hat{c}}; p_{\hat{e}}]), \quad (8)$$

where $[;]$ denote concatenation. The first and second column of I (i.e., I_0 and I_1) index the attribution of \hat{c} and \hat{e} over K clips, respectively. To this end, we estimate \hat{c} and \hat{e} as follows:

$$\hat{c} = I_0 \cdot v, \quad \hat{e} = I_1 \cdot v, \quad s.t. v = \hat{c} + \hat{e}. \quad (9)$$

4.2 Intervener

In absence of clip-level supervision, learning grounding indicators requires dedicated exploitation of the equivariance-invariance principle. On this demand, we propose the intervener, which prompts

the estimated rationale to the oracle by intervening \hat{c} and \hat{e} . Figure 4 describes the functionality of $do(\cdot)$ – the intervention operator that successively manipulated SCM over E and C . Concretely, two intervention operations are configured to realize the equivariant and invariant transformation defined in Equations (4) and (5).

To fulfill the causal-equivariant principle, we design the E-intervention on the causal scene \hat{c} , which applies a linear interpolation between two data points on their causal factors – C, Q and A . By casting the same mixing ratio $\lambda_0 \sim \text{Beta}(\alpha, \alpha)$ on all causal factors, the equivariant intervener learns to capture the causal connection of $C, Q \rightarrow A$. In particular, we attain the intervened causal scene $c^* \in \mathbb{R}^{K \times d}$, question $q^* \in \mathbb{R}^d$ and answer $a^* \in \mathbb{R}$ as follow:

$$c^* = \lambda_0 \cdot \hat{c} + (1 - \lambda_0) \cdot \hat{c}', \quad (10)$$

$$q^* = \lambda_0 \cdot q + (1 - \lambda_0) \cdot q', \quad (11)$$

$$a^* = \lambda_0 \cdot a + (1 - \lambda_0) \cdot a', \quad (12)$$

where \hat{c}', q' and a' are causal factors from a second sample.

To achieve the environment-invariant principle, we devise the I-intervention that adopts a similar mixing strategy to the environment scene \hat{e} . Notably, by drawing the mixing ratios λ_1 from a second distribution that is distinct from the equivariant one (i.e., $\lambda_1 \sim U(0, 1)$), the invariant intervener learns to rule out the influence of environment scene on the answer, which essentially refines the ERM-guided scheme at our will. Formally, we arrive the intervened environment scene e^* by:

$$e^* = \lambda_1 \cdot \hat{e} + (1 - \lambda_1) \cdot \hat{e}', \quad (13)$$

where \hat{e}' is the estimated environment scene of a second sample.

In practice, the equivariant and invariant intervention operations are performed in parallel on different parts of v , and the intervened video $v^* \in \mathbb{R}^{K \times d}$ is composed of $do(C = c^*)$ and $do(E = e^*)$:

$$v^* = c^* + e^*. \quad (14)$$

4.3 Disruptor

To fully exploit the privilege of the proposed principles, we employ contrastive learning as an auxiliary objective to establish a good representation that maintains the desired properties of \hat{c} and \hat{e} . Specifically, we first compose a memory bank π as a collection of visual clips from other training videos. Then, we apply the grounding

indicator a second time on top of the intervened variables, where the re-grounded causal and environment scene are manipulated to set up the contrastive twins as follows:

- In light of the environment-invariance principle, positive video is developed in the sense that changing the environment scene will not provoke disagreement in answer semantics. Thus, the disruptor synthesizes a positive video v^+ by disrupting the v^* on its environment part – that is, replacing the environment scene with a random stratification sampled from the memory bank.¹
- Built upon the causal-equivariance principle, the negative counterpart v^- is created by a similar disruption but on the causal scene of v^* , where substitution on the question-critical causal part should raise inconsistency in answer space. Apart from the visual negatives, the disruptor also creates linguistic alternatives to enhance the distinctiveness of the vision-language alignment. Specifically, it disrupts the combination of the intervened input (v^*, q^*) and pairs the video with random sample question q_r to create a second view of negative samples (v^*, q_r) .

To this end, we attain the answer representation of anchor a and its contrastive counterparts a^+, a^- by feeding the paired positive and negative samples to backbone VideoQA model $f_{\hat{A}}$:

$$a = f_{\hat{A}}(v^*, q^*), \quad (15)$$

$$a^+ = f_{\hat{A}}(v^+, q^*), \quad (16)$$

$$a^- = f_{\hat{A}}([(v^-, q^*); (v^*, q_r)]), \quad (17)$$

where $[;]$ denotes concatenation.

Notably, EIGV is designed to be model-agnostic, which aims to promote any VideoQA backbone built on frame-level visual inputs.

4.4 Optimization

By far, we set up the intervened vision-language instance (v^*, q^*) for a pair of input (v, q) , and further constitute its contrastive counterparts based on the estimated grounding result. To steer the learning process away from the conventional ERM pitfall, we establish two learning objectives on top of their output a, a^+, a^- :

- **Contrastive loss.** To reflect the invariant property of the environment scene while maintaining the distinctiveness of the causal scene, we borrow the definition of InfoNCE [40], and construct the contrastive objective as follows:

$$\mathcal{L}_{CL} = -\log\left(\frac{\exp(a^\top a^+)}{\exp(a^\top a^+) + \sum_n^N \exp(a^\top a_n^-)}\right), \quad (18)$$

where N is the number of negative samples, a_n^- denotes negative answer generated by one of negative samples.

- **ERM loss.** Estimating the rationale requires a robust causal connection from V, Q to A . Thus, we imposed an entropy-based risk function $XE(\cdot)$ on (v^*, q^*) to approach the intervened answer a^* :

$$\mathcal{L}_{ERM} = XE(f_{\hat{A}}(v^*, q^*), a^*), \quad (19)$$

As a result, the overall training objective of EIGV is the aggregation of the foregoing risks:

$$\mathcal{L}_{EIGV} = \mathbb{E}_{(v, q, a) \in O} \mathcal{L}_{ERM} + \beta \mathcal{L}_{CL}, \quad (20)$$

¹Note that the environment substitutes will not involve the question-relevant scenes, to avoid creating additional paths from E to A .

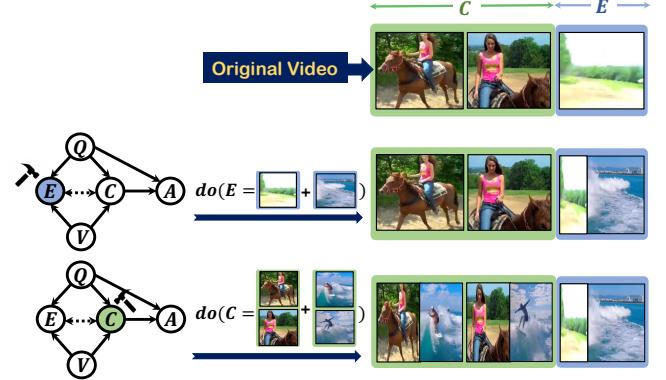


Figure 4: We illustrate the invariant and equivariant interventions in the second and third rows, respectively. The effects on Q and A are omitted for illustration purposes.

where O is the set of training instances (v, q) alongside their ground-truth answer a ; β is the hyper-parameter that balances the strength of contrastive learning. The joint optimization disentangles the mischief of environment scene, thus fulfilling the desired interpretation by locating the causal pattern. During inference, EIGV generates the predication \hat{a} without the intervener and disruptor involved, and gives the interpretation \hat{c} as the partition result of the grounding indicator.

5 EXPERIMENT

In this section, we show the experimental results to answer the following research questions.

- **RQ1** How effective is EIGV in discovering the causal pattern and improving the model generalization across different settings?
- **RQ2** How does the sub-module and feature setting contribute to the performance?
- **RQ3** What pattern does EIGV capture in rationale discovery?

5.1 Settings

Datasets. We conduct experiments on three benchmark datasets that challenge the model’s reasoning capacity from different aspects: **MSVD-QA** [50] and **MSRVTT-QA** [50] mainly emphasize the recognition ability by asking the descriptive questions, where 50K and 243K question-answer pairs are automatically generated from the human-labeled video captions, respectively. **NExT-QA** [48] pinpoints the causal and temporal relations among objects in the video. It contains 47.7K questions with answers in the form of multi-choice, which are manually annotated from 5.4K videos.

Baseline. We validate the effectiveness of EIGV across backbone VideoQA models of three kinds: 1) **Memory-based** methods that foster a storage of input sequence via auxiliary memory design, such as AMU [50], HME [7] and Co-Mem [8]. 2) **Graph-based** methods that leverage the expressiveness of graph network to model the interaction between visual and language elements, which involves methods like L-GCN [14], B2A [26] and HGA [17]. 3) **Hierarchy-based** methods include HCRN [19], PGAT [30], HOSTR [6], MSPAN [10] and HQGA [49]. In common, they exploit the multi-granularity

Table 1: Comparison with SoTAs. Our results are highlighted.

Model	MSVD-QA	MSRVTT-QA	NExT-QA
SoTAs	AMU [50]	32.0	-
	HME [7]	33.7	49.2
	B2A [26]	37.2	-
	L-GCN [14]	34.3	49.5
	HCRN [19]	36.1	48.9
	PGAT [30]	39.0	-
	HOSTR [6]	39.4	-
	HQGA [49]	<u>41.2</u>	<u>51.8</u>
	IGV [21]	40.8	51.3
	Co-Mem [8] + EIGV	34.6 39.8 +5.2	48.5 37.2 +1.9
Ours	HGA [17]	36.6	50.0
	+ EIGV	40.8 +4.2	38.5 +1.8
	MSPAN [10]	40.3	50.7
	+ EIGV	42.6 +2.3	39.3 +1.3

nature of visual elements and realize the hierarchical reasoning via bottom-up architecture. In Specific, we test the generalization of EIGV by marrying our learning principles to three backbones of different categories: memory-based Co-Mem [8], graph-based HGA [17] and hierarchy-based MSPAN [10].

Implementation Detail. For input representation, we encode the video instance as a sequence of $K=16$ clips, where each clip is represented as a combination of appearance and motion features extracted from the pre-trained ResNet-152 and 3D ResNeXt-101. For the linguistic feature, we follow [48] and obtain the contextualized word representation using the fine-tuned BERT model. In the hyper-parameters setting, we set $d = 512$ for cross-modal alignment, then train the model for 80 epochs with an initial learning rate of 5e-5. During optimization, EIGV is trained with Adam optimizer and we decay the learning rate when validation stops improving for 5 epochs. The balance ratio β is set to 0.75.

5.2 Main Result (RQ1)

Table 1 shows the overall result of our method and the SoTAs on three benchmark datasets: MSVD-QA, MSRVTT-QA and NExT-QA. Our observations are summarized as follows:

- Across all three benchmark datasets, the proposed EIGV outperforms SoTA by a distinct margin (+1.2%~2.3%). Such prevailing performance indicates the overall effectiveness of our design, which underpins the theoretical soundness of the equivariant and invariant principles.
- Narrowing the inspection to each of the three backbones, EIGV brings each backbone model a sharp gain across all benchmark datasets (+1.3%~5.2%), which evidences its model-agnostic property. Nevertheless, we notice that the improvements fluctuate across the backbones. As a comparison, on MSVD-QA and MSRVTT-QA benchmarks, EIGV acquires more favorable gains with backbone Co-Mem and HGA than it does with MSPAN. This is because the multi-granularity hierarchy empowers the MSPAN

Table 2: Evaluation on the effectiveness of sub-modules

Ablation	MSVD-QA		NExT-QA	
	MSPAN	HGA	MSPAN	HGA
SoTA Backbone	40.3	36.6	50.7	50.0
+ Mixup [53]	41.0	38.3	52.0	51.8
+ Intervener	41.5	39.6	52.5	52.7
+ Disruptor	40.9	37.6	51.0	51.1
+ Disrupt-Q	40.6	37.0	50.8	51.0
+ Disrupt-V	40.7	37.3	51.0	50.8
EIGV	42.6	40.8	52.9	53.7

with robustness, especially to questions of the descriptive type. Therefore, it achieves stronger backbone performances on benchmarks that focus on the descriptive question (*i.e.*, MSVD-QA and MSRVTT-QA), which, in turn, account for the contribution of EIGV to some extent, thus makes improvement of MSPAN less remarkable. In contrast, when it comes to the causal and temporal question (*i.e.*, NExT-QA) where the inherit advantage of MSPAN backbone vanishes, EIGV shows equivalent improvements on all three backbones (+2.2%~3.7%).

- Comparing the average improvement across different benchmarks, we notice that EIGV achieves the best improvement on MSVD-QA (+2.3%~5.2%) while relatively moderate gains on MSRVTT-QA (+1.3%~1.9%) and NExT-QA (+2.2%~3.7%). The reason for such discrepancy is that MSVD-QA is relatively small in size, which constrains the reasoning ability of the backbone models by limiting their exposure to training instances. As a comparison, MSVD-QA is five-time smaller than MSRVTT-QA in terms of QA pairs (43K vs 243K), and three-time smaller than NExT-QA in terms of video instances (1970 vs 5440). However, such deficiency caters to the focal point of EIGV that develops better in a less generalized situation, thus leading to more preferable growth on MSVD-QA.

5.3 In-Depth Study (RQ2)

5.3.1 What are the effect of EIGV’s components? To comprehensively understand the reasoning mechanism of EIGV, we poke its structure with careful scrutiny. Specifically, we explore the effectiveness of the proposed intervener and disruptor by analyzing their performance with different backbones on two benchmarks. We report the corresponding performances in Table 2 and summarize our findings as follows:

- **Effectiveness of Intervener.** We first testify the substantial efficacy of the intervener by comparing its permanence (“+Intervener”) to the backbone. This brings constant gains across different settings (+1.2%~3%), which demonstrates the stability of our design. Then, we compare the result of the intervener with the conventional mixup augmentation [53], which can be considered as a simplified case of the intervener that only applies the equivariant intervention to the entire training video. The result shows that our design outperforms the conventional mixup in all cases. This manifests that the benefit of invariant

Table 3: Study of feature setting. “APP” and “MOT” denotes using appearance and motion feature individually.

Method	MSVD-QA		NExT-QA		
	MSPAN	HGA	MSPAN	HGA	
APP	SoTA Backbone	40.1	35.0	49.7	48.3
	+ EIGV	41.0	39.5	52.0	52.4
MOT	SoTA Backbone	37.8	33.6	49.4	47.5
	+ EIGV	39.3	38.5	51.1	51.7

intervention is fundamental, and the functionality of invariance and equivariance principle are mutually reinforced.

- **Effectiveness of Disruptor.** We validate the disruptor design by investigating its components — the substitution on video (“+Disrupt-V”) and the permutation on question (“+Disrupt-Q”), respectively. Albeit moderate, improvement on (“+Disrupt-V”) shows that stressing causal scene can benefit visual robustness. A similar trend also applies to “+Disrupt-Q” as well, the constant improvement in all settings shows that acknowledging artificial corrosion in (v, q) matching can strengthen vision-language alignment, which is in line with the current finding in the cross-model pre-train literature [31]. Furthermore, the overall result on “+Disrupt” shows that the advancement of “+Disrupt-V” and “+Disrupt-Q” can be amplified by further integration.

5.3.2 What are the effects of different feature settings? To answer this question, we perform uni-feature tests for the visual representation. Concretely, instead of combining the appearance and motion features together and then manipulating them as a whole, we run ablative experiments on each of them solely. As shown in Table 3, under all circumstances, EIGV can improve models trained with appearance and motion features in equivalent magnitude, even though the appearance feature is demonstrated to be more visually informative in backbone comparison. This verifies that our improvement is ascribed to both feature modes rather than accessing only one of them.

5.3.3 What are the effects of hyper-parameters? Justifying a reliable design requires a sensitivity test on its hyper-parameters. As shown in Figure 5, we probe the potency of EIGV by investigating the distribution of the equivariance mixing ratio and the number of negative samples. Our observations are as follows:

- Figure 5a shows how EIGV performs compared to the SoTA backbone and the conventional mixup augmentation. Specifically, we adjust α to vary the distribution that the equivariant mixing ratio λ_0 draws from, and cross-check the performance of EIGV (“+EIGV”) against its counterparts (“SoTA Backbone” and “+HGA”) on two backbone models. Mixup, despite some improvement, its generalization is limited by the choice of the backbone. For MSPAN backbone, even the heavily tuned α fails to make a reasonable improvement. In contrast, EIGV successively outperforms mixup augmentation and backbone methods in every α setting, which recalls our finding in Section 5.3.1 and justifies the necessity of the environment-invariance principle.
- Figure 5b demonstrates how the performance of EIGV varies as the number of negative sample increase. We notice that the

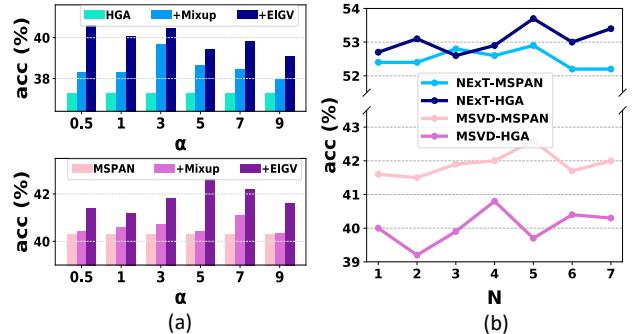


Figure 5: Hyperparameter analysis. (a) Study of α on MSVD-QA, which controls the equivariant mixing ratio by $\lambda_0 \sim \text{Beta}(\alpha, \alpha)$. Performance of two EIGV enhanced models – HGA (top) and MSPAN (bottom) are reported, alongside the SoTA backbone and mixup augmented performances. (b) Study on the impact of the negative sample number N , where EIGV with two backbones (i.e., MSPAN and HGA) on two benchmark datasets (NExT-QA and MSVD-QA) are reported.

predictive curves keep rising until N reaches around five, which indicates that EIGV learns distinctive grounding rationale as more negative samples are considered. This is in line with the finding in the contrastive learning community that additional negative pairs bring about more desirable outcome [11].

5.4 Quantitative Study (RQ3)

By nature, EIGV is equipped with intrinsic visual interpretability. To capture the learning insight of EIGV, we inspect the predictive answer of some video instances along with their grounded interpretations and show the visualization in Figure 6, where each row provides a video instance and two questions that emphasize the visual content in different temporal spans. Notably, even for the same video instance, EIGV is able to accredit different scenes for the different questions. Nonetheless, we also observe the insufficient grounding result in Row 3 Q1, where the grounding result partially covers the dog swimming scene, while the whole video is answerable to the question.

6 RELATED WORK

Video Question Answering. Established to answer the question in dynamic visual content, VideoQA is bred through the task of ImageQA but has broadened its definition by assembling a temporal dimension. To make the task intriguing, the VideoQA benchmark has gone beyond the problem of description [50] and built several datasets to challenge temporal reasoning and even causal reflection [48]. As a result, VideoQA has experienced an aggressive expansion in the architecture design. Chronologically, early efforts tend to enact alignment through cross-modal attention [20, 52] or enhanced memory design [7, 8, 50], while more recent works leverage the expressiveness of the graph neural network and perform the relation reasoning as node propagation [17, 30] or graph alignment [26]. In addition, current designs modify the representation of video and manipulate the temporal sequence from a hierarchical angle. Following this line, HCRN [19] first came out with the conditional

Q1. how does the man in white prevent the white and black dog from running away?

1. fenced up give food
2. use collar strap
3. running behind
4. put dog in cage



Q2. why is the man in white bent down in the middle?

1. carry the brown dog
2. look at the ants
3. pick up the white item
4. protect against the chemical
5. caress the dog

Q1. how did the boy stand up in the middle of the video?

1. lady pull him up
2. push himself up sideways
3. use rope
4. in crutches
5. hinge on the man's feet



Q2. what did the girl in white do after she was at the top of the slide?

1. switch on it
2. prepares to slide
3. push boy on swing
4. jump off seat and spread arms
5. happy

Q1. what is the dog doing?

1. on the right
2. swimming
3. lick baby's hand
4. playing with sticks
5. play with ball



Q2. what did the lady in blue do when she reached the stairs?

1. put head on table
2. sit on stairs
3. goes to get food
4. move hands around
5. jump downstairs

Figure 6: Visualization of discovered grounding rationale. Each row comes with a video instance and two questions that target at different scene. The green and pink windows indicate the rationales for the corresponding questions.

relation module as building blocks that operate through different video intervals, whereas HOSTR and PGAT made their advancement by incorporating visual content from different granularity. MSPAN, however, established cross-scale feature interaction on top of the hierarchy. Despite effective, their intrinsic rationale has long been overlooked. To the best of our knowledge, EIGV is the first work that probes intrinsic interpretation.

Invariant Learning. Given a encoder $f(\cdot)$ and input x , a representation $f(x)$ is invariant to operation T , if $\forall x : f(G(x)) = f(x)$. In practice, this invariant property has a long history in presenting visual content (e.g., HOG [15]), which has recently been renovated by deep learning in form of risk minimization. As its most prevailing form, IRM [2] fosters this philosophy by posing an environment invariant prior and discovering the underlying causal pattern by reducing cross-environment variance. Different from previous studies that create environment via inductive re-grouping [1] or adversarial inference [5, 43–45], our method conducts causal intervention that perturbs the original sample distribution to form a new one. The most relevant work is [21], where an invariant framework is introduced as a model-agnostic framework. However, EIGV gains better generalization ability by marrying equivariance as complementary learning principle.

Visual Interpretability Machine interpretability can be achieved in various methods, such as clustering [24] or disentanglement

[37]. Our design can be vested in the category of attribution discovery, which investigates the contribution of different input elements toward the prediction. Based on whether the prediction and interpretation are yielded simultaneously, two categories are further defined: 1) post-hoc methods that generated the interpretation after prediction, such as backpropagation methods (e.g., grad-CAM [35]). 2) self-interpretable method that cast prediction and interpretation at the same stage. Unlike the post-hoc method that traces the interpretative clue from the output of the black-box, the self-interpretable model builds a transparency model via methods such as prototype generation [3] or structural delineation [47]. In fact, previous works tend to focus on static image. EIGV, however, approaches the video interpretation in a multi-modal situation.

7 CONCLUSION

In this paper, we present EIGV — a model-agnostic explainer, that empowers the SOTA VideoQA model with intrinsic interpretability. In the light of the causality, we formulate our learning principles — causal-equivariance and environment-invariance by incorporating three constituents, the grounding indicator, the intervener, and the disruptor, which manage a robust rationale discovery. Experiments across three benchmarks validate EIGV’s fulfillment in both interpretation and accuracy.

REFERENCES

- [1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In *CVPR*. 3674–3683.
- [2] Martin Arjovsky, Léon Bottou, Ishaaan Gulrajani, and David Lopez-Paz. 2019. Invariant Risk Minimization. *CoRR* abs/1907.02893 (2019).
- [3] Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, and Cynthia Rudin. 2018. This looks like that: deep learning for interpretable image recognition. *CoRR* (2018), 8928–8939.
- [4] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yuetong Zhuang. 2020. Counterfactual Samples Synthesizing for Robust Visual Question Answering. In *CVPR*. 10797–10806.
- [5] Elliot Creager, Jörn-Henrik Jacobsen, and Richard S. Zemel. 2021. Environment Inference for Invariant Learning. In *ICML*. 2189–2200.
- [6] Long Hoang Dang, Thao Minh Le, Vuong Le, and Truyen Tran. 2021. Hierarchical Object-oriented Spatio-Temporal Reasoning for Video Question Answering. In *IJCAI*. 636–642.
- [7] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering. In *CVPR*. 1999–2007.
- [8] Jiayang Gao, Runzhou Ge, Kai Chen, and Ram Nevatia. 2018. Motion-Appearance Co-Memory Networks for Video Question Answering. In *CVPR*. 6576–6585.
- [9] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *AAAI*. 3681–3688.
- [10] Zhicheng Guo, Jiaxuan Zhao, Licheng Jiao, Xu Liu, and Lingling Li. 2021. Multi-Scale Progressive Attention Network for Video Question Answering. In *ACL*. 973–978.
- [11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*. 9726–9735.
- [12] Juhyeon Heo, Sunghwan Joo, and Taesup Moon. 2019. Fooling Neural Network Interpretations via Adversarial Model Manipulation. In *NeurIPS*. 2921–2932.
- [13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* (1997), 1735–1780.
- [14] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-Aware Graph Convolutional Networks for Video Question Answering. In *AAAI*. 11021–11028.
- [15] Shih-Shinh Huang, Hsin-Ming Tsai, Pei-Yung Hsiao, Meng-Qui Tu, and Er-Liang Jian. 2011. Combining Histograms of Oriented Gradients with Global Feature for Human Detection. In *MMM*. 208–218.
- [16] Eric Jiang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. In *ICLR*.
- [17] Pin Jiang and Yahong Han. 2020. Reasoning with Heterogeneous Graph Alignment for Video Question Answering. In *AAAI*. 11109–11116.
- [18] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2019. The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations. In *IJCAI*. 2801–2807.
- [19] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2021. Hierarchical Conditional Relation Networks for Multimodal Video Question Answering. *Int. J. Comput. Vis.* 129 (2021), 3027–3050.
- [20] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019. Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering. In *AAAI*. 8658–8665.
- [21] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant Grounding for Video Question Answering. In *CVPR*. 2928–2937.
- [22] Yicong Li, Xun Yang, Xindi Shang, and Tat-Seng Chua. 2021. Interventional video relation detection. In *ACM MM*. 4091–4099.
- [23] Fei Liu, Jing Liu, Weineng Wang, and Hanqing Lu. 2021. HAIR: Hierarchical Visual-Semantic Relational Reasoning for Video Question Answering. In *ICCV*. 1678–1687.
- [24] Tom Monnier, Thibault Groueix, and Mathieu Aubry. 2020. Deep Transformation-Invariant Clustering. In *NeurIPS*.
- [25] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual VQA: A Cause-Effect Look at Language Bias. In *CVPR*. 12700–12710.
- [26] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. 2021. Bridge To Answer: Structure-Aware Graph Interaction Network for Video Question Answering. In *CVPR*. 15526–15535.
- [27] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys* (2009), 96–146.
- [28] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.
- [29] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*.
- [30] Liang Peng, Shuangji Yang, Yi Bin, and Guoqing Wang. 2021. Progressive Graph Attention Network for Video Question Answering. In *ACM MM*. 2871–2879.
- [31] Alex Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*. 8748–8763.
- [32] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *KDD*. 1135–1144.
- [33] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *IJCAI*. 2662–2670.
- [34] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* (2019), 206–215.
- [35] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *ICCV*. 618–626.
- [36] Ahjeong Seo, Gi-Cheon Kang, Joonhan Park, and Byoung-Tak Zhang. 2021. Attend What You Need: Motion-Appearance Synergistic Networks for Video Question Answering. In *ACL*. 6167–6177.
- [37] Yujun Shen, Ceyuan Yang, Xiaooou Tang, and Bolei Zhou. 2020. InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs. *TPAMI* (2020), 2004–2018.
- [38] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *AIES*. 180–186.
- [39] Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *CVPR*. IEEE Computer Society, 1521–1528.
- [40] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR* abs/1807.03748 (2018).
- [41] Hui Wang, Dan Guo, Xian-Sheng Hua, and Meng Wang. 2021. Pairwise VLAD Interaction Network for Video Question Answering. In *ACM MM*. 5119–5127.
- [42] Jianyu Wang, Bing-Kun Bao, and Changsheng Xu. 2021. DualVGR: A Dual-Visual Graph Reasoning Unit for Video Question Answering. *CoRR* abs/2107.04768 (2021).
- [43] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. 2021. Causal Attention for Unbiased Visual Recognition. In *ICCV*. 3071–3080.
- [44] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue. In *SIGIR*. 1288–1297.
- [45] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, Min Lin, and Tat-Seng Chua. 2022. Causal Representation Learning for Out-of-Distribution Recommendation. In *WWW*. 3562–3571.
- [46] Xiaolong Wang and Abhinav Gupta. 2018. Videos as Space-Time Region Graphs. In *ECCV*. 413–431.
- [47] Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-seng Chua. 2022. Discovering Invariant Rationales for Graph Neural Networks. In *ICLR*. 18446–18458.
- [48] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *CVPR*. 9777–9786.
- [49] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022. Video as Conditional Graph Hierarchy for Multi-Granular Question Answering. In *AAAI*. 2804–2812.
- [50] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yuetong Zhuang. 2017. Video Question Answering via Gradually Refined Attention over Appearance and Motion. In *ACM MM*. 1645–1653.
- [51] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. 2021. Causal Attention for Vision-Language Tasks. In *CVPR*. 9847–9857.
- [52] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2017. Leveraging Video Descriptions to Learn Video Question Answering. In *AAAI*. 4334–4340.
- [53] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.
- [54] Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video Question Answering: Datasets, Algorithms and Challenges. *arXiv preprint arXiv:2203.01225* (2022).