Invariant Grounding for Video Question Answering

Yicong Li[1], Xiang Wang[2]*, Junbin Xiao[1], Wei Ji[1], Tat-Seng Chua[1]
[1]National University of Singapore, [2]University of Science and Technology of China,
liyicong@u.nus.edu, xiangwang1223@gmail.com
junbin@comp.nus.edu.sg, jiwei@nus.edu.sg, dcscts@nus.edu.sg

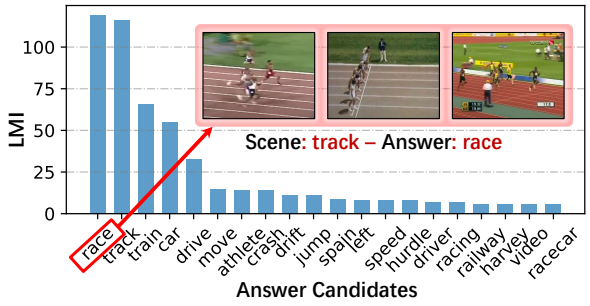
Abstract

Video Question Answering (VideoQA) is the task of answering questions about a video. At its core is understanding the alignments between visual scenes in video and linguistic semantics in question to yield the answer. In leading VideoQA models, the typical learning objective, empirical risk minimization (ERM), latches on superficial correlations between video-question pairs and answers as the alignments. However, ERM can be problematic, because it tends to over-exploit the spurious correlations between question-irrelevant scenes and answers, instead of inspecting the causal effect of question-critical scenes. As a result, the VideoQA models suffer from unreliable reasoning.

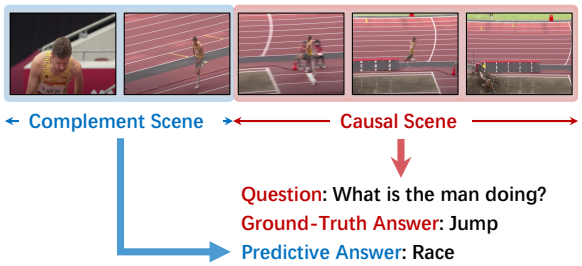
In this work, we first take a causal look at VideoQA and argue that invariant grounding is the key to ruling out the spurious correlations. Towards this end, we propose a new learning framework, Invariant Grounding for VideoQA (IGV), to ground the question-critical scene, whose causal relations with answers are invariant across different interventions on the complement. With IGV, the VideoQA models are forced to shield the answering process from the negative influence of spurious correlations, which significantly improves the reasoning ability. Experiments on three benchmark datasets validate the superiority of IGV in terms of accuracy, visual explainability, and generalization ability over the leading baselines. Our code is available at https://github.com/yl3800/IGV.

1. Introduction

Video Question Answering (VideoQA) [9] is growing in popularity and importance to interactive AI, such as vision-language navigation for in-home robots and personal assistants [2, 35]. It is the task of multi-modal reasoning, which answers the natural language question about the content of a given video. Clearly, inferring a reliable answer requires a deep understanding of visual scenes, linguistic semantics, and more importantly, the visual-linguistic alignments.



(a) Local mutual information (LMI) between the "track" scene and answers.



(b) Running example of how the "track" complement deviates the answering.

Figure 1. Running example. (a) Superficial correlations between visual scenes and answers; (b) Suffering from the spurious correlations, VideoQA model fails to answer the question.

Towards this end, a number of VideoQA models have emerged [8–10, 18, 38]. Scrutinizing these models, we summarize their common paradigm as a combination of two modules: (1) video-question encoder, which encapsulates the visual scenes of video and the linguistic semantics of question as representations; and (2) answer decoder, which exploits these representations to model the visual-linguistic alignment and yield an answer. Consequently, the criterion of empirical risk minimization (ERM) is widely adopted as the learning objective to optimize these modules — that is, minimizing the loss between the predictive answer and the ground-truth answer.

However, the ERM criterion is prone to over-exploiting the superficial correlations between video-question pairs and answers. Specifically, we use the metric of local mutual information (LMI) [27] to quantify the correlations between

*Corresponding author. This work is supported by the Sea-NExT Joint Lab

the "track" scene and answers. As Figure 1a shows, most videos with "track" scene are associated with the "race" answer. Instead of inspecting the visual-linguistic alignments (*i.e*. which scene is critical to answer the question), ERM blindly captures all statistical relations. As Figure 1b shows, it makes VideoQA model naively link the "track"-relevant videos with the strongly-correlated "race" answer, instead of the gold "jump" answer. Taking a causal look [23, 24] at VideoQA (see Section 3), we partition the visual scenes into two parts: (1) causal scene, which holds the question-critical information, and (2) its complement, which is irrelevant to the answer. We scrutinize that the complement is spuriously correlated with the answer, thus ERM hardly differentiates the effects of causal and complement scenes on the answer. Worse still, the unsatisfactory reasoning obstacles the VideoQA model to own the intriguing properties:

- **Visual-explainability** to exhibit "Which visual scene are the right reasons for the right answering?" [6, 26]. Taking Figure 1b as an example to answer "What is the man doing?", the model should attend the "jump" event present in the last three clips, rather than referring to the "track" complement in the first two clips. One straightforward solution is "learning to attend" [31, 36, 39] to ground some scenes via the attentive mechanism. Nonetheless, guided by ERM, such attentive grounding still suffers from the spurious correlations, thus making the highly-correlated complement grounded.
- **Introspective learning** to double-check "How would the predictive answer change if the causal scenes were absent?". On top of attentive grounding, the model needs to introspect whether the learned knowledge (*i.e*. attended scene) reliably and faithfully reflects the logic behind the answering. Briefly put, it should fail to answer the question if the causal scenes were removed.
- **Generalization ability** to enquire "How would the predictive answer response to the change of spurious correlations?". As spurious correlations poorly generalize to open-world scenarios, the model should instead latch on the causal visual-linguistic relations that are stable across different environments.

Inspired by recent invariant learning [3, 16, 33], we conjecture that invariant grounding is the key to distinguishing causal scenes from the complements and overcoming these limitations. By "invariant", we mean that the relations between question-critical scenes and answers are invariant regardless of changes in complements. Towards this end, we propose a new learning framework, Invariant Grounding for VideoQA (**IGV**). Concretely, it integrates two additional modules with into the VideoQA backbone model: a grounding indicator, a scene intervener. Specifically, the grounding indicator learns to attend the causal scenes for a given question and leaves the rest as the complement. Then, we collect visual clips from other training videos to compose a memory bank of complement stratification. For the causal part of interest, the scene intervener conducts the causal interventions [23, 24] on its complement — that is, replace it with the stratification sampled from the memory bank and compose the "intervened videos". After pairing the casual, complement, and intervened scenes with the question, we feed them into the backbone model to obtain the corresponding predictions: (1) causal prediction, which approaches the gold answer, so as to achieve visual explainability; (2) complement prediction, which contains no critical clues to the ground-truth answer, thus enforces the backbone model to perform introspective reasoning; and (3) intervened prediction, which is consistent with the causal prediction across different intervened complements. Jointly learning these predictions enables the backbone model to alleviate the negative influence of multi-modal data bias. It is worthwhile emphasizing that IGV is a model-agnostic strategy, which trains the VideoQA backbones in a plug-and-play fashion.

Our contributions are summarized as follows:

- We highlight the importance of grounding causal scenes from the complements to visual-explainability, generalization, and introspective learning of VideoQA models.
- We propose a new model-agnostic training scheme, IGV, which incorporates invariant grounding into the VideoQA models, to mitigate the negative influence of multi-modal data bias and enhance the multi-modal reasoning ability.
- On three benchmark datasets (*i.e*. MSRVTT-QA [38], MSVD-QA [38], NExT-QA [37]), we conduct extensive experiments to justify the superiority of IGV in training the VideoQA backbones. In particular, IGV significantly outperforms the state-of-the-art models.

2. Preliminaries

In this section, we summarize the common paradigm of VideoQA models. Throughout the paper, we denote the random variables and their deterministic values by upper-cased (*e.g*. V) and lower-cased (*e.g*. v) letters, respectively.

Modeling. Given the video-question pair (V, Q), the primer task of VideoQA is to generate an answer $\hat{A}$ as:

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

where $f_{\hat{A}}$ is the VideoQA model, which is typically composed of two modules: video-question encoder, and answer decoder. Specifically, the encoder includes two components: (1) a video encoder, which encodes visual scenes of the target video as a visual representation, such as motion-appearance memory design [9, 10], structural graph representation [12, 15, 20, 34], hierarchical architecture [8, 17]; and (2) a question encoder, which encapsulates linguistic semantics of the question into a linguistic represen-

tation, such as global/local representation of textual content [14, 32], graph representation of grammatical dependencies [20]. On top of these representations, the decoder learns the visual-linguistic alignments to generate the answer. In particular, the alignments are modeled via cross-modal interaction like graph alignment [20], cross-attention [14, 15, 18, 40] and co-memory [10], *etc*.

Learning. To optimize these modules, most of the leading VideoQA models [9, 10, 14, 15, 17] cast the multi-modal reasoning problem as a supervised learning task and adopt the learning objective of empirical risk minimization (ERM) as:

$$\min_{h} \mathcal{L}_{\text{ERM}}(\hat{A}, A), \tag{2}$$

where $\mathcal{L}_{\text{ERM}}$ is the risk function to measure the loss between the predictive answer $\hat{A}$ and ground-truth answer A, which is usually set as cross-entropy loss [10, 17] or hinge loss [9, 15, 37]. In essence, ERM encourages these VideoQA modules to capture the statistical correlations between the video-question pairs and answers.

3. Causal Look at VideoQA

From the perspective of causal theory [23, 24], we revisit the VideoQA scenario to show superficial correlations between video-question pairs and answers. We then analyze ERM's suffering from the spurious correlations.

3.1. Causal Graph of VideoQA

In general, multiple visual scenes are present in a video. But only part of the scenes are critical to answering the question of interest, while the rest hardly offers information relevant to the question. Moreover, the linguistic variations in different questions should activate different scenes of a video. These facts inspire us to split the video into the causal and complement parts in terms of the question. Here we use a causal graph [23, 24] to exhibit the relationships among five variables: input video V, input question Q, causal scene C, complement scene T, ground-truth answer A. Figure 2 illustrates the causal graph, where each link is a cause-and-effect relationship between two variables:

- $C \leftarrow V \rightarrow T$. The input video V consists of C and T. For example, the video in Figure 1b is the combination of the first two clips (*i.e*. C) and the last three clips (*i.e*. T).
- $V \rightarrow C \leftarrow Q$. The causal scene C is conditional upon the video-question pair (V, Q), which distills Q-relevant information from V. For a given V, the variations in Q result in different C.
- $Q \rightarrow A \leftarrow C$. The answer A is determined by the question Q and causal scene C, reflecting the visual-linguistic alignments. Considering the example in Figure 1b again, C is the oracle scene that perfectly explains why "jump" is labeled as the ground truth to answer the question.

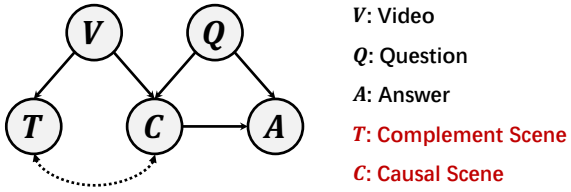


Figure 2. Causal graph of VideoQA

- $T \dashleftarrow\dashrightarrow C$. The dashed arrow summarizes the additional probabilistic dependencies [21, 22] between C and T. Such dependencies are usually caused by the selection bias or inductive bias during the process of data collection or annotation [5, 30]. For example, one mostly collects the videos with the "jump" events on the "track". Here we list three typical scenarios: (1) C is independent of T (*i.e*. $T \perp C$); (2) C is the direct cause of T (*i.e*. $C \rightarrow T$), or vise versa (*i.e*. $C \leftarrow T$); (3) C and T have a common cause E (*i.e*. $C \leftarrow E \rightarrow T$). See Appendix A for details.

3.2. Spurious Correlations

Taking a closer look at the causal graph, we find that the complement scene T and the ground-truth answer A can be spuriously correlated. Specifically, as the confounder [22–24] between T and A, Q and V open the backdoor paths: $T \leftarrow V \rightarrow C \rightarrow A$ and $T \leftarrow V \rightarrow C \leftarrow Q \rightarrow A$, which make T and A spuriously correlated even though there is no direct causal path from T to A. Worse still, $T \dashleftarrow\dashrightarrow C$ can amplify this issue. Assuming $C \rightarrow T$, C becomes an additional confounder to yield another backdoor path $T \leftarrow C \rightarrow A$. Such spurious correlations can be summarized as the probabilistic dependence: $A \not\perp T$.

As ERM naively captures the statistical correlations between video-question pairs and answers, it fails to distinguish the causal scene C and complement scene T, thus failing to mitigate the negative influence of spurious correlations. As a result, it limits the reasoning ability of VideoQA models, especially in the following aspects: (1) visual-explainability to reason about "Which visual scenes are the supporting evidence to answer the question?"; (2) introspective learning to answer "How would the answer change if the causal scenes were absent?"; and (3) generalization ability to enquire "How would the answer response to the change of spurious correlations?".

4. Methodology

We get inspiration from invariant learning [3, 16, 33] and argue that invariant grounding of causal scenes is the key to reducing the spurious correlations and overcoming the foregoing limitations. We then present a new learning framework, Invariant Grounding for VideoQA (**IGV**).

4.1. Invariant Grounding for VideoQA

Upon closer inspection on the causal graph, we notice that the ground-truth answer A is independent of the visual

complement T, only when conditioned on the question Q and the causal scene C, more formally:

$$A \perp T \mid C, Q. \tag{3}$$

This probabilistic independence indicates the invariance — that is, the relations between the (C, Q) pair and A are invariant regardless of changes in T. The causal relationship $Q \rightarrow A \leftarrow C$ is invariant across different T. Taking Figure 1b as an example, if the question and the causal scene (*i.e.* the last three clips) remain unchanged, the answer should arrive at "jump", no matter how the complement varies[1] (*e.g.* substitute the "track" clips by the "cloud"- or "sea"-relevant ones). This highlights that the (C, Q) pair is the key to shielding A from the influence of T.

Modeling. However, only the (V, Q) pair and A are available in the training set, while neither C nor the grounding function towards C is known. This motivates us to incorporate visual grounding into the VideoQA modeling, where the grounded scene $\hat{C}$ aims to estimate the oracle C and guide the prediction of answer $\hat{A}$. More formally, instead of the conventional modeling (*cf*. Equation (1)), we systematize the modeling process as:

$$\hat{C} = f_{\hat{C}}(V, Q), \quad \hat{A} = f_{\hat{A}}(\hat{C}, Q), \tag{4}$$

where $f_{\hat{C}}$ is the grounding model, and $f_{\hat{A}}$ is the VideoQA model that relies on the $(\hat{C}, Q)$ pair instead. See Section 4.2 for our implementations of $f_{\hat{C}}$ and $f_{\hat{A}}$.

Learning. Nonetheless, simply integrating visual grounding with the VideoQA model falls into the "learning to attend" paradigm, which still suffers from the spurious correlations and erroneously attends to the complement scenes as $\hat{C}$. To this end, we exploit the invariance property of C (*cf*. Equation (3)) and reformulate the learning objective of invariant grounding as:

$$\min_{f_{\hat{A}}, f_{\hat{C}}} \mathcal{L}_{\text{IGV}}(\hat{A}, A), \quad \text{s.t. } A \perp \hat{T} \mid \hat{C}, Q, \tag{5}$$

where $\mathcal{L}_{\text{IGV}}$ is the loss function to our IGV; $\hat{T} = V \setminus \hat{C}$ is the complement of $\hat{C}$. In the next section, we will elaborate how to implement $\mathcal{L}_{\text{IGV}}$ and achieve invariant grounding.

4.2. IGV Framework

Figure 3 displays our IGV framework, which involves two additional modules, the grounding indicator and scene intervener, beyond the VideoQA backbone model $f_{\hat{A}}$.

4.2.1 Grounding Indicator

For a video-question pair instance (v, q), at the core of the grounding indicator is to split the video instance v into two parts, $\hat{c}$ and $\hat{t}$, according to the question q. Towards this end, it first employs two independent LSTMs [11] to encode the visual and linguistic characteristics of v and q, respectively:

$$\mathbf{v}_g, \mathbf{v}_l = \text{LSTM}_1(v), \quad \mathbf{q}_g, \mathbf{q}_l = \text{LSTM}_2(q), \tag{6}$$

where the features of v are K fixed visual clips, while q is associated with L language tokens; LSTM_1 outputs $\mathbf{v}_l \in \mathbb{R}^{K \times d}$ as the local representations of clips, and yields the last hidden state $\mathbf{v}_g \in \mathbb{R}^d$ as the global representations of the holistic video. Analogously, LSTM_2 generates $\mathbf{q}_l \in \mathbb{R}^{L \times d}$ as the local representations of tokens, and makes the last hidden state $\mathbf{q}_g \in \mathbb{R}^d$ represent the question holistically, here d is the hidden dimension.

Upon these representations, the attention scores are constructed to indicate the importance of each visual clip. Here we devise $\mathbf{p}_{\hat{c}} \in \mathbb{R}^K$ to exhibit the probability of each clip belonging to the causal scene $\hat{c}$, while $\mathbf{p}_{\hat{t}} \in \mathbb{R}^K$ is in contrast to $\mathbf{p}_{\hat{c}}$ to show how likely each clip composes the complement $\hat{t}$. The formulations are as follows:

$$\mathbf{p}_{\hat{c}} = \text{Softmax}(\text{MLP}_1(\mathbf{v}_l) \cdot \text{MLP}_2(\mathbf{q}_g)^\top), \tag{7}$$

$$\mathbf{p}_{\hat{t}} = \text{Softmax}(\text{MLP}_3(\mathbf{v}_l) \cdot \text{MLP}_4(\mathbf{q}_g)^\top), \tag{8}$$

where four multilayer perceptrons (MLPs) are employed to distill useful information: $\text{MLP}_1(\mathbf{v}_l)$, $\text{MLP}_3(\mathbf{v}_l) \in \mathbb{R}^{K \times d'}$, $\text{MLP}_2(\mathbf{q}_g)$, $\text{MLP}_4(\mathbf{q}_g) \in \mathbb{R}^{d'}$; d' is the feature dimension. However, as the soft masks make $\hat{c}$ and $\hat{t}$ overlapped, the attentive mechanism cannot shield the answering from the influence of the complement. Hence, the grounding indicator produces discrete selections instead to make $\hat{c}$ and $\hat{t}$ disjoint. Nonetheless, simple sampling or selection is not differentiable. To achieve differentiable discrete selection, we apply Gumbel-Softmax [13]:

$$\mathbf{I} = \text{Gumbel-Softmax}([\mathbf{p}_{\hat{c}}, \mathbf{p}_{\hat{t}}]), \tag{9}$$

where Gumbel-Softmax is built upon the concatenation of $\mathbf{p}_{\hat{c}}$ and $\mathbf{p}_{\hat{t}}$ (*i.e.* $[\mathbf{p}_{\hat{c}}, \mathbf{p}_{\hat{t}}] \in \mathbb{R}^{K \times 2}$), and outputs the indicator vector $\mathbf{I} \in \mathbb{R}^{K \times 2}$ whose first and second column indexes $\hat{c}$ and $\hat{t}$ over k clips, respectively. As such, we can devise $\hat{c}$ and $\hat{t}$ as follows:

$$\hat{c} = \{I_{k0} \cdot v_k | I_{k0} = 1\}, \quad \hat{t} = \{I_{k1} \cdot v_k | I_{k1} = 1\}, \tag{10}$$

where I_{0k} and I_{1k} suggests that the k-th clip belongs to the causal and complement scenes, respectively.

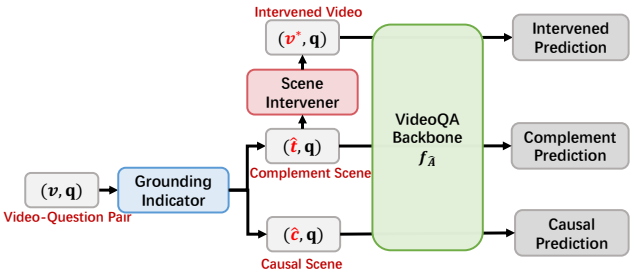


Figure 3. Overview of our IGV framework.

[1]Note that the complement substitutes will not involve the question-relevant scenes, in order to avoid creating additional paths from T to A.

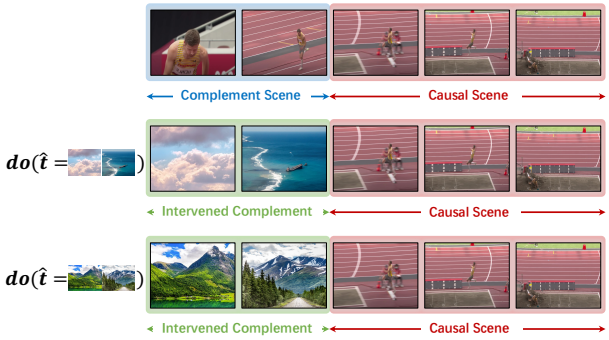


Figure 4. Illustration of interventional distribution.

4.2.2 Scene Intervener

It is challenging to learn the grounding indicator, owing to the lack of supervisory signals of clip-level importance. To remedy this issue, we propose the scene intervener, which preserves the estimated causal scene $\hat{c}$ but intervenes the estimated complement $\hat{t}$ to create the "intervened videos", as Figure 4 shows.

Specifically, for the observed video-question pairs during training, the scene intervener first collects visual clips from other training videos as a memory bank of complement stratification, $\hat{\mathcal{T}} = \{\hat{t}\}$. Then, for the video of interest $v = \hat{c} \cup \hat{t}$, the intervener conducts causal interventions [23, 24] on its $\hat{t}$ — that is, random sample a complement stratification $\hat{t}^* \in \hat{\mathcal{T}}$ to replace $\hat{t}$ and combine it with $\hat{c}$ at hand as a new video $v^* = \hat{c} \cup \hat{t}^*$.

It is worthwhile mentioning that, distinct from the current invariant learning studies [3, 16, 33] that only partition the training set into different environments, our scene intervener exploits the interventional distributions [29] instead. The interventional distribution (*i.e.*, the videos with the same interventions) can be viewed as one environment.

4.2.3 VideoQA Backbone Model

Inspired by [15], we design a simple yet effective architecture as our backbone predictor, where the video encoder is shared with the grounding indicator. It embodies convolutional graph networks (GCN) to propagate clip-level visual messages, then integrates cross-modal fused local and global representations via BLOCK fusion [4]. See Appendix B for the detailed architecture.

4.2.4 Joint Training

For a video-question pair instance (v, q), we have established the causal scene $\hat{c}$, complement scene $\hat{t}$, and intervened video v^* via the grounding indicator and scene intervener. Pairing them with q synthesizes three new instances: $(\hat{c}, q)$, $(\hat{t}, q)$, (v^*, q). We next feed these instances into the backbone VideoQA model $f_{\hat{A}}$ to obtain three predictions:

- **Causal prediction**. As the causal scene $\hat{c}$ is expected to be sufficient and necessary to answer the question q, we leverage its predictive answer $f_{\hat{A}}(\hat{c}, q)$ to approach the ground-truth answer a solely:

$$\mathcal{L}_{\hat{c}} = \text{XE}(f_{\hat{A}}(\hat{c}, q), a), \tag{11}$$

  where XE denotes the cross-entropy loss.
- **Complement prediction**. As no critical clues should exist in the complement scene $\hat{t}$ to answer the question q, we encourage its predictive answer $f_{\hat{A}}(\hat{t}, q)$ to evenly predict all answers. This uniform loss is formulated as:

$$\mathcal{L}_{\hat{t}} = \text{KL}(f_{\hat{A}}(\hat{t}, q), u), \tag{12}$$

  where KL denotes KL-divergence, and u is the uniform distribution over all answer candidates.
- **Intervened prediction**. According to the invariant constraint (*cf*. Equation (3)), the causal relationship between the causal scene and the answer is stable across different complements. To parameterize this constraint, we enforce all v's intervened versions to hold the consistent predictions:

$$\mathcal{L}_{v^*} = \mathbb{E}_{\hat{t}^* \in \mathcal{T}}(\text{KL}(f_{\hat{A}}(v^*, q), f_{\hat{A}}(\hat{c}, q))). \tag{13}$$

Aggregating the foregoing risks, we attain the learning objective of IGV:

$$\mathcal{L}_{\text{IGV}} = \mathbb{E}_{(v,q,a)\in\mathcal{O}^+} \mathcal{L}_{\hat{c}} + \lambda_1 \mathcal{L}_{\hat{t}} + \lambda_2 \mathcal{L}_{v^*}, \tag{14}$$

where $\mathcal{O}^+$ is the training set of the video-question pair (v, q) and the ground-truth answer a; λ_1 and λ_2 are the hyper-parameters to control the strengths of invariant learning. Jointly learning these predictions enables the VideoQA backbone model to uncover the question-critical scene, so as to mitigate the negative influence of spurious correlations between the question-irrelevant complement scene and answer. In the inference phase, we use the causal prediction $f_{\hat{A}}(\hat{c}, q)$ to answer the question.

5. Experiments

We conduct extensive experiments to answer the following research questions:

- **RQ1:** How effect is IGV in training VideoQA backbones as compared with the State-of-the-Art (SoTA) models?
- **RQ2:** How do the loss component and feature setting affect the performance?
- **RQ3:** What are the learning patterns and insights of IGV training?

Settings: We compare IGV with seven baselines from families of Memory, GNN and Hierarchy (Appendix C) on three VideoQA datasets: **NExT-QA** [37] which features causal

Table 1. Comparison of accuracy on NExT-QA test set. The **best** and second-best results are highlighted.

Models	Causal	Temp	Descrip	All
Co-Mem [10]	45.85	50.02	54.38	48.54
HCRN [17]	47.07	49.27	54.02	48.82
HME [9]	46.76	48.89	57.37	49.16
HGA [15]	48.13	49.08	57.79	50.01
IGV(Ours)	**48.56**	**51.67**	**59.64**	**51.34**
Abs. Improve	+0.43	+1.65	+1.85	+1.33

and temporal action interactions among multiple objects. It contains about 47.7K manually annotated questions for multi-choice QA collected from 5.4K videos with an average length of 44s. **MSVD-QA** [38] and **MSRVTT-QA** [38] are two prevailing datasets that focus on the description of video elements. They respectively contain 50K and 243K QA pairs with open answer space over 1.6K and 6K. For all three datasets, we follow their official data splits for experiments and report accuracy as evaluation metric.

Implementation Details: For the visual feature, we follow previous works [15, 17, 37] and extract video feature as a combination of motion and appearance representations by using the pre-trained 3D ResNeXt-101 and ResNet-101, respectively. Specifically, each video is uniformly sampled into K=16 clips, where each clip is represented by a combined feature vector $v_k^{d_v}$, where d_v equals 4096. Similar to [37], we obtain the contextualized word representation from the finetuned BERT model, and the feature dim d_q is 768. For our model, the dimension of the hidden states are set to d = 512, and the number of graph layers in IGV backbone predictor is 2. During training, IGV is optimized by Adam optimizer with the initial learning rate of 1e-4, which will be halved if no validation improvements in 5 epochs. We set the batch size to 256 and a maximum of 60 epochs. (See Appendix D for more details and complexity analysis).

5.1. Main Results (RQ1)

5.1.1 Comparisons with SoTA Methods

As shown in Table 1 and Table 2, our method outperforms SoTAs with questions of all sub-types surpassing their competitors. Specifically, we have two major observations:

First, on NExT-QA, IGV gains remarkable improvement on *temporal* type (+1.65%), the underlying explanation are: 1) *temporal* question generally corresponds to video content with a longer time span, which requires more introspective grounding of the causal scene. Fortunately, IGV's design philosophy comfort such demand by wiping out the trivial scenes, which takes up a huge proportion in *temporal* type, thus making the predicting faithful. 2) *temporal* questions tend to include a temporal indicative phase (*e.g.* "*at the end of the video*") that serves as a strong signal for grounding indicator to locate the target window.

Table 2. Comparison of accuracy on MSVD-QA and MSRVTT-QA test set. "†" indicates the result is re-implementation with the publicly available code

Models		MSVD-QA	MSRVTT-QA
Memory	AMU [38]	32.0	32.0
	HME [9]	33.7	33.0
	Co-Mem† [10]	34.6	35.3
GNN	HGA† [15]	35.4	36.1
	B2A [20]	37.2	36.9
Hierarchy	HCRN [17]	36.1	35.6
	HOSTR [8]	39.4	35.9
Causal view	IGV (Ours)	**40.8**	**38.3**
	Abs. Improve	+1.4	+1.4

Second, along with *descriptive* questions on NExT-QA, the result on MSRVTT-QA and MSVD-QA (both emphases on question of *descriptive* type) demonstrate the superiority in *descriptive* question across all three datasets (+1.85% on NExT-QA, +1.4% on MSRVTT-QA and MSVD-QA). Such improvement is underpinned by logic that answering *descriptive* questions requires scrutiny on the scene of interest, instead of a holistic view of the entire sequence. Accordingly, targeted prediction inducted by IGV concentrates reasoning on keyframes, thus achieving better performance. As a consequence, such improvement strongly validates that IGV generalizes better over various environments.

5.1.2 Backbone Agnostic

By nature, our IGV principle is orthogonal to backbone design, thus helping to boost any off-the-shelf SoTAs without compromising the underlying architecture. We therefore experimentally testify the generality and effectiveness of our learning strategy by marrying the IVG principle with methods from two different categories: Co-Mem [10] from memory-based architecture and HGA [15] from Graph-based method. Table 3 shows the results on three backbone predictors (including ours). Our findings are:

1. Better improvement for severe bias. We notice that the improvement on MSVD-QA (+3.1%∼4.7%) is considerably larger than that on MSRVTT-QA (+1.4%∼2%). Such expected discrepancy is caused by the fact that, although identical in question type, MSRVTT-QA is almost 5 times larger than MSVD-QA (#QA pairs 243K *vs* 50K). As a result, the baseline model trained on MSRVTT-QA is gifted with better generalization ability, whereas the model on MSVD-QA still suffers from severe shortcut correlation. For the same reason, the IGV framework achieves much better improvement in the severe-shortcut situation (*e.g.* MSVD-QA). Such discrepancy validates our motivation of eliminating statistic dependency.

2. Constant improvement for each method. Through

Table 3. IGV strategy is applied to different SoTAs methods. "+IGV" denoted our strategy is incorporated.

Models	MSVD-QA		MSRVTT-QA	
	Baseline	+IGV	Baseline	+IGV
Co-Mem [10]	34.6	**37.7**	35.3	**37.3**
HGA [15]	35.4	**38.8**	36.1	**37.5**
Our Backbone	36.1	**40.8**	36.3	**38.3**

Table 4. Study of IGV loss components

Variants	MSVD-QA		MSRVTT-QA	
	Our Backbone	Co-Mem [10]	Our Backbone	Co-Mem [10]
Baseline	36.1	34.6	36.3	35.3
$\mathcal{L}_{\hat{c}}$	36.0	33.3	36.7	36.0
$\mathcal{L}_{\hat{c}} + \mathcal{L}_{\hat{t}}$	37.4	36.1	37.8	36.8
$\mathcal{L}_{\hat{c}} + \mathcal{L}_{v^*}$	38.2	36.3	37.4	36.2
$\mathcal{L}_{\hat{c}} + \mathcal{L}_{\hat{t}} + \mathcal{L}_{v^*}$	**40.8**	**37.7**	**38.3**	**37.3**

row-wise inspection, we notice that for each benchmark, IGV can bring considerable improvement across different backbone models (+3.1%∼4.7% for MSVD-QA, +1.4%∼2% for MSRVTT-QA). Such stable enhancement strongly verifies our modal-agnostic statement.

5.2. In-Depth Study (RQ2)

5.2.1 Contributions of Different Loss Components

An in-depth comprehension of IGV framework requires careful scrutiny on its components. Alone this line, we exhaust the combination of IGV loss components and design three variants: $\mathcal{L}_{\hat{c}}$, $\mathcal{L}_{\hat{c}} + \mathcal{L}_{\hat{t}}$ and $\mathcal{L}_{\hat{c}} + \mathcal{L}_{v^*}$. Table 4 shows the result of the above variants on two benchmarks across two backbone predictors. Our observations are as follow:

- Using $\mathcal{L}_{\hat{c}}$ solely, which can be viewed as a special case of ERM-guided attention, hardly outperforms the baseline, because grounding indicators can not identify the causal scene without clip-level supervision. Such an expected result reflects our motivation in interventional design.
- $\mathcal{L}_{\hat{c}} + \mathcal{L}_{\hat{t}}$ and $\mathcal{L}_{\hat{c}} + \mathcal{L}_{v^*}$ matched equally in accuracy that consistently surpass baseline and $\mathcal{L}_{\hat{c}}$ in all cases. Such progress shows the effectiveness of intervention strategy and introspective regularization imposed on complement.
- In all cases, $\mathcal{L}_{\hat{c}} + \mathcal{L}_{\hat{t}} + \mathcal{L}_{v^*}$ further boosts the performance significantly, which shows $\mathcal{L}_{\hat{t}}$ and $\mathcal{L}_{v^*}$ contribute in different aspects and their benefits are mutually reinforcing.

5.2.2 Study of Feature

By convention, we study the effect of the input condition by ablation on the visual feature. Particularly, we denote **APP** for tests that adopt only appearance feature as input and **MOT** for tests that utilize motion feature alone. Figure 6a delivers results on two benchmarks, where we observe:

First, IGV can improve the performance significantly for all input conditions, which generalizes the effectiveness of our framework. Similar to Table 3, the improvement on MSVD-QA is larger than that on MSRVTT-QA, which solidifies our finding in Section 5.1.2.

Second, compared to motion feature, IGV brings distinctively larger improvements using appearance feature. Considering the causal nature of IGV, we conclude that static correlation tends to bias more in appearance feature.

5.2.3 Study of Hyper-parameter

To validate the sensitivity of IGV against the hyper-parameters, we conduct experiments with variations of λ_1 and λ_2 on two datasets. Without loss of generalization, we tune λ_1 (λ_2) as sample of $\{1.3^i \mid -10 \leq i \leq 10,\ i \in \mathbb{Z}\}$, while keeping the λ_2 (λ_1) as 1. According to Figure 6b, we have follow observations:

For MSVD-QA, we observe consistent peaks around 0.8 for both λ_1 and λ_2. Comparatively, fluctuation on MSRVTT-QA is more moderate, where tuning on λ only causes a 1.5% difference in their accuracy. It's noteworthy that IGV outperforms the baseline by a large margin (+3%) under all tests, which indicates IGV's robustness against variation of hyper-parameters. Additionally, comparing to λ_2, IGV is more sensitive to λ_1. Typically, the performance suffers a drastic degradation for λ_1 larger than 5 on both datasets. Whereas λ_2 maintain above 39% (MSVD-QA) and 37.5% (MSRVTT-QA) for all tests.

5.3. Qualitative analysis (RQ3)

As mentioned in Section 1 , IGV is empowered with visual-explainability, and is apt to account for the right scene for its prediction. Following this essence, we grasp the learning insight of IGV by inspecting some correct examples from the NExT-QA dataset and show the visualization in Figure 5. Concretely, each video comes with two questions that emphasize different parts of the video. We notice that, even for the same video, our grounding window is question-sensitive to enclose the explainable content with correct prediction. Nonetheless, we also observe results of *insufficient-grounding* on the third row Q2, where the girl starts to bend down before the last two frames, even though the most informative last two frames are encompassed.

6. Related works

Video Question Answering (VideoQA). Aiming to answer the question in a video scenario, VideoQA is defined as an escalation of imageQA, because the temporal nature of the input has enriched its reasoning process as well as the answer space. Previous efforts towards VideoQA establish their contribution on either a better multi-modal interaction or stronger video representation. Specifically, early studies tend to impose sophisticated cross-modal fusion via attention [15, 18, 40] or dynamic memory [9, 10, 38], while more recent approaches perform relation reasoning

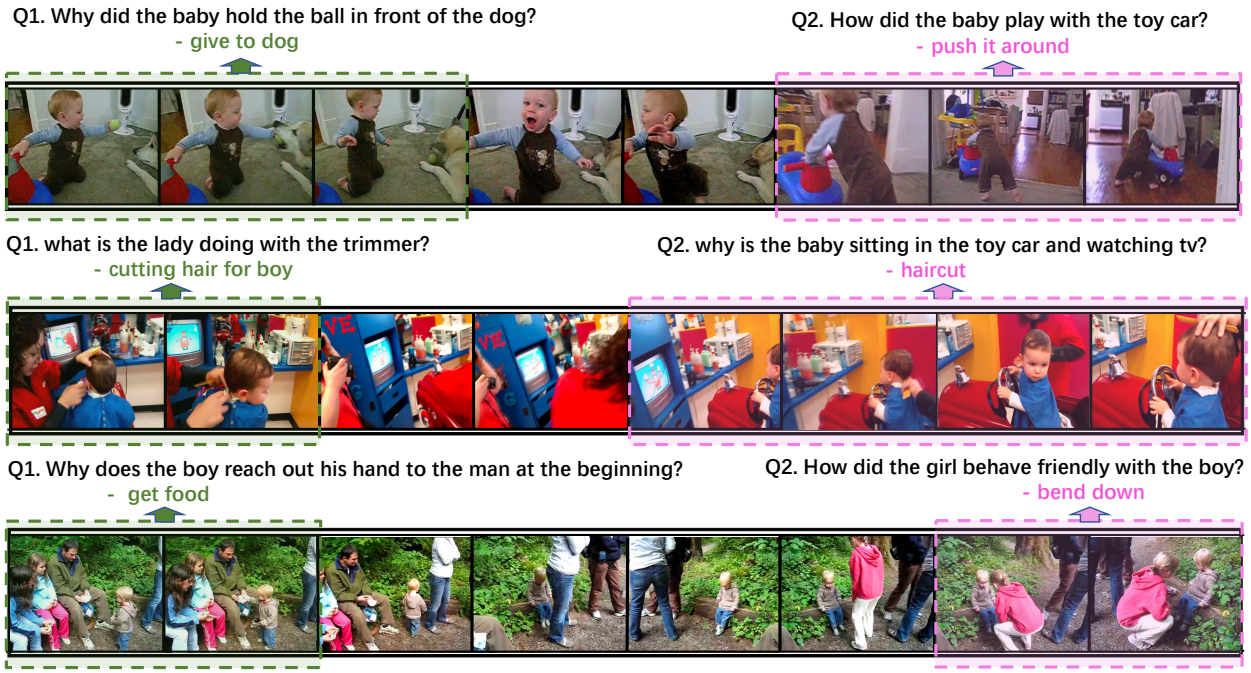
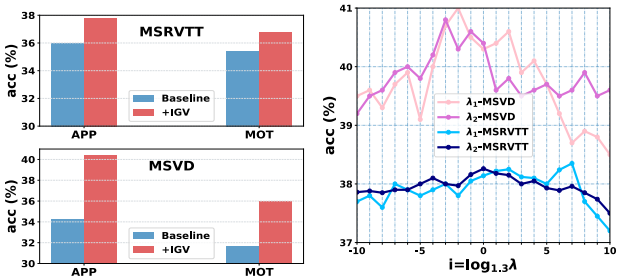


Figure 5. Visualization of grounding result on the correct prediction cases from NExT-QA. Each video comes with two questions that demand causal scene of different time span. The green and pink windows indicate the causal scenes for the corresponding questions.



(a) (b)

Figure 6. (a) Study of feature setting.; (b) Study of λ_1 and λ_2.

through visual or textual graph [12, 15, 20]. In addition, current efforts that model video as a hierarchical structure also intrigue wide interest. Among them, HCRN [17] stack conditional relation blocks in different feature granularity, whereas HOSTR [8] employs a spatio-temporal graph for multilevel reasoning. Despite their effectiveness, their visual-explainability still dwells on ERM-guided attention weights, which only reflect the intensity of feature-prediction correlation.

Invariant Learning. Multi-modal datasets tend to display inherent bias in some forms [1, 19, 25, 28]. In contrast to overarching reality, the collection process [5, 30] degrades its generalization ability by introducing undesirable correlations between the inputs and the ground truth annotations.

To overcome such correlation, invariant learning is developed to discover causal relations from the causal factors to the response variable, which remains constant across distributions. As the most prevailing formulation, IRM [3] promotes this philosophy from feature level to representation level by finding a data representation Υ, from which the optimal predictor φ can yield the prediction $\Upsilon \circ \varphi$ that is stable across all environments. In terms of environment acquisition, previous studies either manually partition the training set by prior knowledge [2], or generates data partition iteratively via adversarial environment inference [7, 33]. Our method, instead of partitioning the training, assumes no prophets about environments but performs causal intervention to perturb the original distribution. To the best of our knowledge, IGV is the first work that introduces invariant learning as a model-agnostic framework to VideoQA.

7. Conclusions

In this paper, we pinpoint that the spurious visual-linguistic correlations in VideoQA are triggered by question-irrelevant scenes. We propose a novel invariant grounding framework, IGV, to distinguish the causal scene and emphasize its causal effect on the answer. With the grounding indicator and scene intervener, IGV captures the causal patterns that remain stable across complements. Extensive experiments verify the effectiveness of IGV on different backbone VideoQA models.

Our future work includes two aspects: 1) the spurious correlations can nest in entities, object-level invariant learning is promising to alleviate this issue; 2) as the current intervention strategy might threaten the causal prediction by introducing complement with new shortcuts, we will explore new intervention methods.

References

[1] Ankesh Anand, Eugene Belilovsky, Kyle Kastner, Hugo Larochelle, and Aaron C. Courville. Blindfold baselines for embodied QA. *CoRR*, 2018. 8

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, pages 3674–3683, 2018. 1, 8

[3] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019. 2, 3, 5, 8

[4] Hedi Ben-younes, Rémi Cadène, Nicolas Thome, and Matthieu Cord. BLOCK: bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *AAAI*, pages 8102–8109. AAAI Press, 2019. 5, 2

[5] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Being negative but constructively: Lessons learnt from creating better visual question answering datasets. In *NAACL-HLT*, pages 431–441. Association for Computational Linguistics, 2018. 3, 8

[6] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, pages 10797–10806, 2020. 2

[7] Elliot Creager, Jörn-Henrik Jacobsen, and Richard S. Zemel. Environment inference for invariant learning. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 2189–2200. PMLR, 2021. 8

[8] Long Hoang Dang, Thao Minh Le, Vuong Le, and Truyen Tran. Hierarchical object-oriented spatio-temporal reasoning for video question answering, 2021. 1, 2, 6, 8, 3

[9] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, pages 1999–2007. Computer Vision Foundation / IEEE, 2019. 1, 2, 3, 6, 7

[10] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *CVPR*, 2018. 1, 2, 3, 6, 7

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997. 4

[12] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering, 2020. 2, 8

[13] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR (Poster)*, 2017. 4

[14] Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. pages 11101–11108, 2020. 3

[15] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI*, pages 11109–11116. AAAI Press, 2020. 2, 3, 5, 6, 7, 8

[16] David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron C. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, volume 139, pages 5815–5826, 2021. 2, 3, 5

[17] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. pages 9969–9978, 2020. 2, 3, 6, 8

[18] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *AAAI*, pages 8658–8665. AAAI Press, 2019. 1, 3, 7

[19] Yicong Li, Xun Yang, Xindi Shang, and Tat-Seng Chua. Interventional video relation detection. In *MM '21: ACM*, pages 4091–4099. ACM, 2021. 8

[20] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge to answer: Structure-aware graph interaction network for video question answering, 2021. 2, 3, 6, 8

[21] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. 3

[22] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009. 3

[23] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000. 2, 3, 5

[24] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 2, 3, 5

[25] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, pages 4035–4045. Association for Computational Linguistics, 2018. 8

[26] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *IJCAI*, pages 2662–2670, 2017. 2

[27] Tal Schuster, Darsh J. Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models, 2019. 1

[28] Jesse Thomason, Daniel Gordon, and Yonatan Bisk. Shifting the baseline: Single modality performance on visual navigation & QA. In *NAACL-HLT*, pages 1977–1983. Association for Computational Linguistics, 2019. 8

[29] Jin Tian, Changsung Kang, and Judea Pearl. A characterization of interventional distributions in semi-markovian causal models. In *AAAI*, pages 1239–1244, 2006. 5

[30] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 1521–1528. IEEE Computer Society, 2011. 3, 8

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 2

[32] Jianyu Wang, Bing-Kun Bao, and Changsheng Xu. Dualvgr: A dual-visual graph reasoning unit for video question answering. *CoRR*, abs/2107.04768, 2021. 3

[33] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition, 2021. 2, 3, 5, 8

[34] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, volume 11209, pages 413–431. Springer, 2018. 2

[35] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*, pages 6629–6638, 2019. 1

[36] Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. Visual relation grounding in videos. In *European Conference on Computer Vision*, pages 447–464. Springer, 2020. 2

[37] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786. Computer Vision Foundation / IEEE, 2021. 2, 3, 5, 6

[38] 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 *ACM MM*, pages 1645–1653, 2017. 1, 2, 6, 7

[39] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 37, pages 2048–2057, 2015. 2

[40] 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, 2017. 3, 7