Supplementary Material for Knowledge Proxy Intervention for Deconfounded Video Question Answering

Jiangtong Li[1], Li Niu[1*], Liqing Zhang[1*]
[1] Department of Computer Science and Engineering, MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
{keep_moving-Lee,ustcnewly,lqzhang}@sjtu.edu.cn

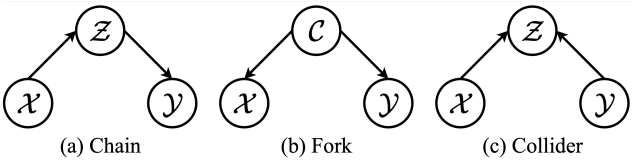
In this document, we provide additional materials to supplement our main submission. In Section 1, we introduce more background on the causal graph and causal intervention. In Section 2, we derive some formulations in detail, which are omitted in the main paper due to space limitation. In Section 3, we elaborate more details about the dataset. In Section 4, we provide more details about the implementation of our KPI framework, including the choice of backbone model, hyper-parameters and training strategy. In Section 5, we provide more details about the implementation of knowledge space. In Section 6, we provide more details about the implementation of video embeddings. In Section 7, we perform significance test to reveal the stability of KPI framework. In Section 8, we provide more qualitative examples to testify the effectiveness of KPI framework.

1. Causal Graph and Causal Interventation

1.1. Causal Graph

Causal graph [13] is a high-level road-map indicating the causal relationship among different variables. Besides, causal graph is a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where nodes $\mathcal{N}$ represent variables and arrows $\mathcal{E}$ represent the causal relationship between two nodes. For example, in Figure 1 (a), $\mathcal{X} \rightarrow \mathcal{Z}$ indicates that $\mathcal{X}$ cause $\mathcal{Z}$, while $\mathcal{X} \nrightarrow \mathcal{Y}$ indicates there is no direct causal effect from $\mathcal{X}$ to $\mathcal{Y}$.

Besides, if a variable $\mathcal{C}$ is the common cause of two variables $\mathcal{X}, \mathcal{Y}$ (in Figure 1 (b)), $\mathcal{C}$ is called the confounder, which will induce spurious correlation between $\mathcal{X}$ and $\mathcal{Y}$ to disturb the recognition of the causal effect between them. In particular, such spurious correlation is brought by the backdoor path created by the confounder. Formally, a backdoor path between $\mathcal{X}$ and $\mathcal{Y}$ is defined as any path from $\mathcal{X}$ to $\mathcal{Y}$ that starts with an arrow pointing into $\mathcal{X}$. For example, in Figure 1 (b), we use $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{C}$ to represent the *"sales of ice-cream "*, *"death from drowning"*, and *"season"*, respectively. From the causal point of view, $P(\mathcal{Y} = y|\mathcal{X} = x) - P(\mathcal{Y} = y)$ should be zero, since there is no causal relationship between *"sales of ice-cream "* and *"death from drowning"*. However, due to the existence of $\mathcal{C}$, $P(\mathcal{Y} = y|\mathcal{X} = x) - P(\mathcal{Y} = y)$ is not zero, and spurious correlation is built by the confounder, *"season"*. Therefore, if we want to deconfound two variables for the true causal effect, we should cut the backdoor path between them. In the abovementioned case, cutting backdoor path can be formulated as $P(\mathcal{Y} = y|\mathcal{X} = x, \mathcal{C} = c) - P(\mathcal{Y} = y|\mathcal{C} = c) = 0$



(a) Chain (b) Fork (c) Collider

Figure 1. Three elemental structures in causal graph.

1.2. Elemental Structures

As shown in Figure 1, there are three elemental structures in causal graph, which construct the whole graph. In this section, we will introduce some basic rules about them.
Chain, $\mathcal{X} \rightarrow \mathcal{Z} \rightarrow \mathcal{Y}$. In chain structure, $\mathcal{Z}$ is the intermediate variable to correlate $\mathcal{X}$ and $\mathcal{Y}$, which indicates that if we know the value of $\mathcal{Z}$, $\mathcal{X}$ will not give any extra information to predict $\mathcal{Y}$. Therefore, if we directly intervene $\mathcal{Z}$ as a specific value, we block this chain structure.
Fork, $\mathcal{X} \leftarrow \mathcal{C} \rightarrow \mathcal{Y}$. In fork structure, $\mathcal{C}$ is the confounder to produce a backdoor path between $\mathcal{X}$ and $\mathcal{Y}$, which would induce spurious correlation between $\mathcal{X}$ and $\mathcal{Y}$. Therefore, if we directly intervene $\mathcal{C}$ as a specific value, the true causal relationship between $\mathcal{X}$ and $\mathcal{Y}$ would be revealed.
Collider, $\mathcal{X} \rightarrow \mathcal{Z} \leftarrow \mathcal{Y}$. In collider structure, variable $\mathcal{X}$ and $\mathcal{Y}$ are independent by nature; however, if we know the value of $\mathcal{Z}$, $\mathcal{X}$ and $\mathcal{Y}$ will be correlated at once. Therefore, if we want to study the causal relationship between $\mathcal{X}$ and $\mathcal{Y}$, we should leave the collider structure alone.

These three structures are the elemental structure for a causal graph and provide us with the basic tool of blocking

*Corresponding author.

paths between two variables. To sum up, if we want to block two variables within a chain or fork, we should intervene in the intermediate node, whereas, if we want to block two variables within a collider, we should do nothing.

1.3. do-calculus

do-calculus $P(\mathcal{Y}|do(\mathcal{X}=x))$ is first introduce by [12], which can be used to represent the causal effect of $\mathcal{X}$ on $\mathcal{Y}$. In this section, we will introduce how the **do-calculus** works. Note that we will use $do(\mathcal{X})$ to represent $do(\mathcal{X}=x)$ in the following sections.

As mentioned in the main submission, *do-calculus* is a type of causal intervention, which means that we actively assign a value to the variable instead of passively observing it. For example, in Figure 1 (b), $do(\mathcal{X})$ indicates that we set the variable $\mathcal{X}$ as value x and ignore the cause from its parent node, *i.e.*, $\mathcal{Z} \nrightarrow \mathcal{X}$. In brief, while applying **do-calculus** to a variable, we cut off all the arrows ending at this variable. Therefore, when we calculate $P(\mathcal{Y}|do(\mathcal{X}))$, no confounder will cause $\mathcal{X}$ and $\mathcal{Y}$, which ensures that the probability reflects the causal effect.

To derive the probability formula with **do-calculus**, we will introduce three rules in [12].

Rule 1. If a variable $\mathcal{X}$ is irrelevant to $\mathcal{Y}$, then the probability distribution of $\mathcal{Y}$ will not change:

$$P(\mathcal{Y}|z,\mathcal{X}) = P(\mathcal{Y}|z). \tag{1}$$

Rule 2. If a set $\mathcal{Z}$ of variable blocks all backdoor paths from $\mathcal{X}$ to $\mathcal{Y}$, then conditional on $\mathcal{Z}$, $do(\mathcal{X})$ is equivalent to observe $\mathcal{X}$:

$$P(\mathcal{Y}|z,do(\mathcal{X})) = P(\mathcal{Y}|z,X). \tag{2}$$

Rule 3. If there are no causal paths from $\mathcal{X}$ to $\mathcal{Y}$, $do(\mathcal{X})$ can be directly removed from $P(\mathcal{Y}|do(\mathcal{X}))$:

$$P(\mathcal{Y}|do(\mathcal{X})) = P(\mathcal{Y}). \tag{3}$$

2. Formula Derivations

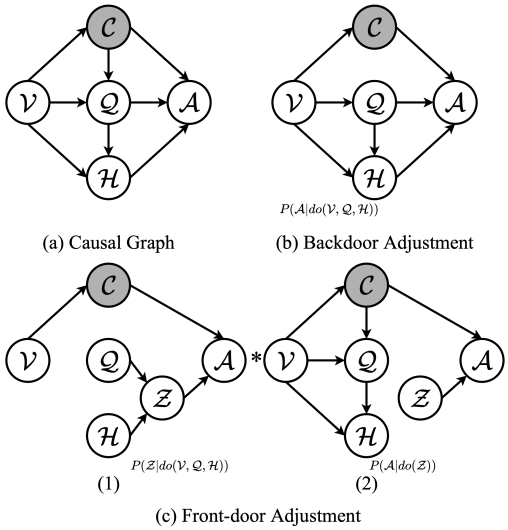
For convenience, we draw our cause graph and the corresponding causal intervention in Figure 2 and write down the equation for front-door adjustment again, and add some intermediate steps for the derivation. As shown in Figure 2 (c), the causal intervention for front-door adjustment is split into two parts,

2.1. Proof of Equation (4)

Equation (4) of the main submission is formulated as

$$P(\mathbf{z}|do(\mathcal{V},\mathcal{Q},\mathcal{H})) = P(\mathbf{z}|\mathcal{V},\mathcal{Q},\mathcal{H}) = P(\mathbf{z}|\mathcal{Q},\mathcal{H}).$$

The first step is because $\mathcal{V},\mathcal{Q},\mathcal{H}$ do not have a backdoor path to $\mathbf{z}$ and **Rule 2**. The section step is due to **Rule 3** and the fact that $\mathcal{V}$ does not have the causal path to $\mathbf{z}$.



(a) Causal Graph (b) Backdoor Adjustment

(c) Front-door Adjustment

Figure 2. The causal graph and causal intervention for VideoQA.

2.2. Proof of Equation (5)

Equation (5) of the main submission is formulated as

$$\begin{aligned}
&P(\mathcal{A}|do(\mathbf{z})) \\
&= \sum_{\mathbf{v}}\sum_{\mathbf{q}}\sum_{\mathbf{h}} P(\mathcal{A}|do(\mathbf{z}),\mathbf{q},\mathbf{h},\mathbf{v})P(\mathbf{q},\mathbf{h},\mathbf{v}|do(\mathbf{z})) \\
&= \sum_{\mathbf{v}}\sum_{\mathbf{q}}\sum_{\mathbf{h}} P(\mathcal{A}|\mathbf{z},\mathbf{q},\mathbf{h},\mathbf{v})P(\mathbf{q},\mathbf{h},\mathbf{v}|do(\mathbf{z})) \\
&= \sum_{\mathbf{v}}\sum_{\mathbf{q}}\sum_{\mathbf{h}} P(\mathcal{A}|\mathbf{z},\mathbf{q},\mathbf{h},\mathbf{v})P(\mathbf{q},\mathbf{h},\mathbf{v}) \\
&= \sum_{\mathbf{v}}\sum_{\mathbf{q}}\sum_{\mathbf{h}} P(\mathcal{A}|\mathbf{z},\mathbf{q},\mathbf{h},\mathbf{v})P(\mathbf{v})P(\mathbf{q}|\mathbf{v})P(\mathbf{h}|\mathbf{q},\mathbf{v}).
\end{aligned}$$

The first step is according to the Bayes rules. The second step is because $\mathbf{z}$ does not have backdoor path to $\mathcal{A}$ and **Rule 2**. The third step is because $\mathbf{z}$ does not have causal path to $\mathbf{v},\mathbf{q},\mathbf{h}$ and **Rule 3**. The last step is due to the chain rule of conditional probability.

2.3. Proof of Formula (7)

Let's first start with a simple case:

$$\begin{aligned}
P(\mathcal{Y}|\mathcal{X}) &= \sum_{x}(P(\mathbf{x})P(\mathcal{Y}|\mathbf{x})) \\
&= \mathbb{E}_x[P(\mathcal{Y}|\mathbf{x})] = \mathbb{E}_x[\text{Softmax}(g(\mathbf{x}))],
\end{aligned}$$

where $P(\mathcal{Y}|\mathbf{x})$ is estimated by $\text{Softmax}(g(\mathbf{x}))$. By [18], we can use Weighted Geometric Mean (WGM) to approximate the expectation $\mathbb{E}_x[\text{Softmax}(g(\mathbf{x}))]$, *i.e.*,

$$P(\mathcal{Y}|\mathcal{X}) = \mathbb{E}_x[\text{Softmax}(g(\mathbf{x}))] \approx \text{WGM}[\text{Softmax}(g(\mathbf{x}))].$$

Then let's show how to apply Weighted Geometric Mean (WGM) [24] to move the outer expectation into the feature

level. Given the $\text{Softmax}(g(\mathbf{x})) \propto \exp(g(\mathbf{x}))$, the weighted geometric mean (WGM) of $P(\mathcal{Y}|\mathbf{x})$ is:

$$\text{WGM}[\text{Softmax}(g(\mathbf{x}))] = \prod_{\mathbf{x}} \exp(g(\mathbf{x}))^{P(\mathbf{x})}$$
$$= \prod_{\mathbf{x}} \exp(g(\mathbf{x})P(\mathbf{x})) = \exp[\sum_{\mathbf{x}} g(\mathbf{x})P(\mathbf{x})] = \exp[\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})]]$$

Furthermore, to guarantee the sum of $P(\mathcal{Y}|\mathcal{X})$ to be 1, we normalize the $\text{WGM}[\text{Softmax}(g(\mathbf{x}))]$ as

$$\begin{aligned}\text{NWGM}[\text{Softmax}(g(\mathbf{x}))] &= \frac{\exp[\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})]]}{\sum_{\mathbf{x}} \exp[\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})]]} \\ &= \text{Softmax}(\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})]).\end{aligned}$$

Therefore, the NWGM can be used to approximate the expectation $\mathbb{E}_x[\text{Softmax}(g(\mathbf{x}))]$ and move the outer expectation into the feature level.

In Formula (7) of the main submission, we apply the same approximation and derivation, arriving at

$$\begin{aligned}&P(\mathcal{A}|do(\mathcal{V}, \mathcal{Q}, \mathcal{H})) \\ &= \mathbb{E}_{\mathbf{v}}\mathbb{E}_{[\mathbf{q}|\mathbf{v}]}\mathbb{E}_{[\mathbf{h}|\mathbf{q},\mathbf{v}]}\mathbb{E}_{[\mathbf{z}|\mathcal{Q},\mathcal{H}]}[P(\mathcal{A}|\mathbf{v}, \mathbf{q}, \mathbf{h}, \mathbf{z})] \\ &= \mathbb{E}_{\mathbf{v}}\mathbb{E}_{[\mathbf{q}|\mathbf{v}]}\mathbb{E}_{[\mathbf{h}|\mathbf{q},\mathbf{v}]}\mathbb{E}_{[\mathbf{z}|\mathcal{Q},\mathcal{H}]}[\text{Softmax}[g(\mathbf{v}, \mathbf{q}, \mathbf{h}, \mathbf{z})]] \\ &\approx \text{Softmax}[g(\mathbb{E}_{\mathbf{v}}[\mathbf{v}], \mathbb{E}_{[\mathbf{q}|\mathbf{v}]}[\mathbf{q}], \mathbb{E}_{[\mathbf{h}|\mathbf{q},\mathbf{v}]}[\mathbf{h}], \mathbb{E}_{[\mathbf{z}|\mathcal{Q},\mathcal{H}]}[\mathbf{z}])].\end{aligned}$$

3. Datasets

3.1. MSVD-QA and MSRVTT-QA

MSVD-QA [22] and MSRVTT-QA [22] are the extensions of two video description datasets, *i.e.*, MSVD [2] and MSRVTT [23]. Specifically, the question-answer pairs from these two datasets are generated from video descriptions through an automatic method [22]. For MSVD-QA and MSRVTT-QA, there are 50K and 243K question-answer pairs, respectively. Both of these two datasets consist of five different types of questions, including what, who, how, when, and where. The task is open-ended and aims to identify the answer from a pre-defined answer set. Note that since the question-answer pair is generated automatically, there are some questions that correspond to multiple different answers.

3.2. TGIF-QA

TGIF-QA [8] is a large-scale benchmark dataset for VideoQA [8], which collect 72K animated GIFs along with 165K question-answer pairs to form four sub-tasks: Count, Action, Transition, and FrameQA. (1) Repetition count (Count.) aims to count the number of repetitions of objects or actions in a video, and the answer is a number; (2) Repeating action (Action.) aims to identify a repetitive action in a video, and the answer is chosen from 5 candidates; (3) State transition (Transition.) is also a multiple-choice task which aims to identify the temporal transition of two states, *i.e.*, actions or activities; (4) Frame QA (FrameQA) aims to answer a question from a single video frame. This task is formulated as a classification problem aiming to indicate the correct answer from a pre-defined set. Considering that the Count problem is not directly affected by dataset bias, we mainly focus on the Action, Transition, and FrameQA sub-tasks in this work.

3.3. NeXT-QA

NeXT-QA [19] is a large-scale human-annotated dataset for VideoQA [19], which collects 5440 videos from VidOR [16] along with 100k question-answer pairs. Different from previous datasets, NeXT-QA aims to advance video understanding from describing to explaining temporal actions. To this end, NeXT-QA formulates two sub-task, *i.e.*, multi-choice QA and Generative QA, and for each sub-task, there are three types of questions, including Causal (Why and How), Temporal ($Previous$, $Present$, and $Next$), and Description ($Binary$, $Location$, $Count$, and $Other$). In this work, we focus on the multi-choice QA sub-task.

3.4. Causal-VidQA

Causal-VidQA [11] is a large-scale dataset for reasoning VideoQA, which includes 27k videos from Kinetics-700 [10]. For each video, there are four types human-annotated questions, *i.e.*, Description, Explanation, Prediction, and Counterfactual. These four question focus on simple description (Description), evidence reasoning (Explanation), and commonsense reasoning (Prediction and Counterfactual). To verify the reasoning ability of existing methods, Causal-VidQA requires the methods to provide a correct answer and to offer a proper reason justifying why that answer is correct. Considering that a question may correspond to more than one rational answers and reasons, Causal-VidQA is also formulated as multi-choice QA.

4. Implementation Details

For video representation, we extract the motion and appearance feature using the I3D-ResNeXt-101 [21, 5] and ResNet-152 [6] with 8 clips per video and 16 frames per clip. For HQGA [20], which requires the object features [15], we follow the setting in HQGA [20] by detecting 20 objects in NExT-QA and 5 objects in other datasets. For language representation, we follow NExT-QA [19] and obtain the contextualized word representation using the fine-tuned BERT model. Note that the initial results in CoMem [4] and HGA [9] are conducted based on GloVe [14] feature. Therefore, we reproduce them with the official code and the contextualized word representation. For the size of feature space, we set k_V, k_Q, k_H as 500, and the actual number of causal concepts decides k_Z. For MSVD-QA, MSRVTT-QA, TGIF-Action,

TGIF-Transition, TGIF-FrameQA, NExT-QA, and Causal-VidQA, k_Z are 10,249, 21,612, 18,974, 19,836, 18,856, 24,799, and 25,645, respectively. We set $d = 512$ for video-question alignment and train each model 25 epochs with an initial learning rate of 5e-5. During training, the KPI framework is optimized by Adam optimizer, and the learning rate is decayed when validation stops improving for 5 epochs. For other hyper-parameters, we follow the setting in corresponding works.

5. Implementation of Knowledge Space

5.1. Key Words and Phrases

In Sec.4.1 of the main submission (step 1), we have mentioned to "*extract the key words and phrases from question and answer*". For the key words and phrase extraction, we first explore the off-the-shelf NLTK [1] to label the part-of-speech (POS) tag of each word in questions and answers, and then select the verb and noun as the key words. Furthermore, based on the POS-tag of each word, we extract the noun phrase and the verb phrase with pre-defined grammar. For more details, please refer to NLTK [1].

5.2. Knowledge Graph and Knowledge Semantic Embedding

In our KPI framework, we explore two knowledge graphs (ConceptNet [17] and Atomic [7]) to filter out the correlations and select causal relations. ConceptNet [17] focuses more on physical-entity relations, which is helpful to discover entity relations in descriptive questions. Besides, Atomic [7] pays more attention on event-centered and social-interaction relations, which could contribute more on evidence and commonsense reasoning.

For ConceptNet [17], the nodes are mainly single words. Therefore, we directly match the head and tail of each correlated concept pair with the node in ConceptNet following the strategy in Sec.4.1 of main submission (step 4). For Atomic [7], the nodes include both single words, phrase and short sentence. In order to use the event-centered and social-interaction relations within Atomic, we first extract the key words and phrases (see Section 5.1) from node n_i as node concepts N_i. During the matching process, if the $h_i \in N_p$ and $t_i \in N_q$, where $h_i - t_i$ is $i-$th a correlated concept pair and node n_p and n_q are adjacent nodes in knowledge graph. To ensure the causal concepts are representative, we only use the correlated concept pairs, which appears more than 20 times in the training dataset, during the matching. Since the causal concepts cannot directly be used by our method, we 1) concatenate the head, tail, and relation of each causal concept together, 2) use a pre-trained BERT [3] to get the contextualized word representation, 3) average the contextualized word representation along word dimension to get the knowledge semantic embeddings.

6. Implementation of Video Embeddings

In Sections 4.2 and 4.3 of the main submission, we exploit the self-attention layer along with average-pooling layer to get the video embeddings. In this section, we will provide more details about the implementation.

Given a video, we extract the motion features $\mathbf{V}_m \in \mathbb{R}^{n_{v_m} \times d_v}$, appearance features $\mathbf{V}_a \in \mathbb{R}^{n_{v_a} \times d_v}$, and object features $\mathbf{V}_o \in \mathbb{R}^{n_{v_o} \times d_v}$ using the I3D-ResNeXt-101 [21, 5], ResNet-152 [6], and Faster-RCNN with ResNet-152 [15], respectively, where $d_v = 2048$ and $n_{v_m}, n_{v_a}, n_{v_o}$ are decided by datasets. The motion features $\mathbf{V}_m$, appearance features $\mathbf{V}_a$, and object features $\mathbf{V}_a$ are gathered together as the video features $\mathcal{V}$. For each video sub-features $\mathbf{V}_* \in [\mathbf{V}_m, \mathbf{V}_a, \mathbf{V}_o]$ from video features $\mathcal{V}$, the self-attention layer along with average-pooling layer is formulated as

$$\begin{aligned}
\mathbf{V}_*^{sa} &= \mathrm{Softmax}(\frac{\mathbf{V}_* \boldsymbol{W}_1 (\mathbf{V}_* \boldsymbol{W}_2)^T}{\sqrt{d_v}}) \mathbf{V}_* \boldsymbol{W}_3, \\
\mathbf{V}_*^{out} &= \mathrm{LayerNorm}(\mathbf{V}_*^{sa} + \mathbf{V}_*) \boldsymbol{W}_{out} \\
\mathbf{v}_* &= \mathrm{AveragePool}(V_*^{out}),
\end{aligned}$$

where $\mathbf{v}_* \in \mathbb{R}^{d_v}$ represent the video sub-embeddings, and $\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3, \boldsymbol{W}_{out}$ are trainable parameters.

7. Significance Test

We perform the significance test towards the improvement between our KPI framework and HQGA [20] on MSVD-QA [22] and NeXT-QA [19]. On MSVD-QA and NeXT-QA datasets, we run HQGA equipped with KPI framework ten times with random seeds ranging from 1 to 10. The improvement beyond original HQGA are $2.1 \pm 4 \times 10^{-1}$ and $3.2 \pm 4 \times 10^{-1}$ in terms of accuracy. At the significance level of 0.05, we perform a significance test to verify that our KPI framework can stably boost the original HQGA. The p-values of the improvements are 7.37×10^{-8} and 1.81×10^{-9}, which is far below 0.05, demonstrating the superiority of our KPI framework is statistically significant.

8. More Qualitative Examples

In Figure 3, we inspect the predictive answer of three different video instances along with two different questions, and the top attended causal concepts based on each video and question. We observe that, for the same video and different questions, our model is capable of retrieving different causal concepts, which reveals that our framework is capable of achieving the two aforementioned requirements, *i.e.*, (1) emphasize the knowledge from the question and the aligned features; (2) collect the causal concepts for answer prediction. However, as shown in the last video example, our model may also fail to capture the useful causal concepts from the knowledge space. We suspect that this fail-

Q: What is the dog doing?

A: on the right
swimming
lick baby's hand
playing with sticks
play with ball

Top Z: <dog-CapableOf-swim>
<water-ObjectUse-feed dog>
<swim-RelatedTo-water>
<water-ObjectUse-wash dog>
<person-xWant-play with dog>

Q: Why does the black dog have a blue vest around it s body?

A: attack black dog
keep warm
snatched the bone
stay afloat
want to join in

Top Z: <vest-ObjectUse-keep safe>
<swim-RelatedTo-water>
<lifevest-ObjectUse-float in water>
<dog-CapableOf-swim>
<vest-ObjectUse-protect body>

Q: What did the cat do after jumped into the sink with water?

A: sit up then roll back
play with the hand
watch the water flow
follow the string
walk past

Top Z: <tap-ObjectUse-get water>
<sink-isFilledBy-water>
<cat-AtLocation-sink>
<sink-ObjectUse-drain water>
<eye-RelatedTo-watch>

Q: Why is the cat shaking its head at the end of the video?

A: drink water
distracted by something
follow the toy
sleepy
dry face

Top Z: <head-MannerOf-shake>
<water-Cause-wet>
<shake-RelatedTo-water>
<shake head-xIntend-remove water>
<cat-AtLocation-sink>

Q: Where could this be happening?

A: park
room
museum
farm
workplace

Top Z: <airplane-AtLocation-hanger>
<airplane-AtLocation-base>
<airplane-AtLocation-sky>
<warehouse-RelatedTo-roof>
<airplane-ObjectUse-land>

Q: Why is the big silver aeroplane at an indoor space?

A: **work**
moving his body
forgot to be closed
presents wrappers
display for exhibition

Top Z: <airplane-ObjectUse-land>
<work-RelatedTo-indoors>
<airplane-ObjectUse-travel>
<worker-CapableOf-repair>
<repair-xNeed-follow instruction>

Figure 3. The visualization of three VideoQA cases from NExT-QA [19]. Top Z indicates the causal concepts from $\mathbf{Z}$ with the top-5 highest attention weight. Predicted answers are highlighted in boldface. Correct (*resp.* Wrong) answers are highlighted in green (*resp.* red)

ure case is due to these two reasons: (1) the causal concepts within the knowledge space cannot link the airplane with the exhibition, and (2) the environment in the video does not have obvious characters about museum and exhibition, which make the causal concepts searching and answer prediction extremely hard.

References

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 4

[2] David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL 2011*, pages 190–200, 2011. 3

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*, pages 4171–4186, 2019. 4

[4] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *CVPR 2018*, pages 6576–6585, 2018. 3

[5] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR 2018*, pages 6546–6555, 2018. 3, 4

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR 2016*, pages 770–778, 2016. 3, 4

[7] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI Conference on Artificial Intelligence*, 2020. 4

[8] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *CVPR 2017*, pages 1359–1367, 2017. 3

[9] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI 2020*, pages 11109–11116, 2020. 3

[10] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv*, 2017. 3

[11] Jiangtong Li, Li Niu, and Liqing Zhang. From Representation to Reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *CVPR 2022*, pages 21241–21250, 2022. 3

[12] Judea Pearl. Causality: models, reasoning and inference. *Springer*, 2000. 2

[13] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. Causal inference in statistics: A primer. *John Wiley & Sons.*, 2019. 1

[14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *EMNLP 2014*, pages 1532–1543, 2014. 3

[15] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS 2015*, pages 91–99, 2015. 3, 4

[16] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *ICMR 2019*, pages 279–287, 2019. 3

[17] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI 2017*, pages 4444–4451, 2017. 4

[18] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014. 2

[19] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *CVPR 2021*, pages 9777–9786, 2021. 3, 4, 5

[20] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *AAAI 2022*, pages 2804–2812, 2022. 3, 4

[21] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR 2017*, pages 5987–5995, 2017. 3, 4

[22] 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 2017*, pages 1645–1653, 2017. 3, 4

[23] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR 2016*, pages 5288–5296, 2016. 3

[24] 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 2015*, pages 2048–2057, 2015. 2