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

## Supplementary Material

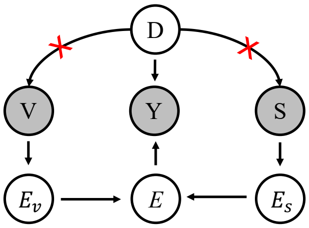


Figure 1. Causal formulation. Variables include data domain D, video V, sentence S, label Y, visual event features $E_v$, textual event features $E_s$, and alignment event features $E$. They are connected by directed edges which represent causal directions.

## Abstract

*In section A of this supplementary, we provide Backdoor Judgement Derivation; In section B, we show more qualitative results in Causal-VidQA [2].*

## A. Backdoor Judgement Derivation

In this section, we show the derivation of Eq.(1) in the main paper according to the backdoor judgment in Structural Causal Model(SCM) [4]. More theoretical derivations can be found in [4]. First, refer to Fig.1, we introduce three elemental "junctions" in SCM:

$\boldsymbol{V} \rightarrow \boldsymbol{E_v} \rightarrow \boldsymbol{E}$ is a chain junction. It provides a front-door path from $V$ to $E$. The effect of $V$ to $E$ is equal to the effect of $E_v$ to $E$.

$\boldsymbol{V} \leftarrow \boldsymbol{D} \rightarrow \boldsymbol{Y}$ is a confounding junction. It provides a back-door path from $V$ to $Y$ and introduces confounding associations between them. If conditioning on variable $D$, that is, given a specific value of $D$, the confounding path between $V$ and $Y$ is blocked.

$\boldsymbol{E_v} \rightarrow \boldsymbol{E} \leftarrow \boldsymbol{E_s}$ is a collider junction. It introduces a positive effect of $E_v$ and $E_s$ for $E$. If conditioning on variable $E$, the blocked path between $E_v$ and $E_s$ is open which creates a correlation between them. Conversely, $E_v$ and $E_s$ are independent when $E$ is not conditioned.

In Video Question Answering (VideoQA) task, we prefer to eliminate the spurious association and find the real association. Therefore, we can intervene in the chained junction and confounding junction to cut off spurious associations. We should not intervene in the collider junction, which will introduce new confounders. A preferred approach is the backdoor judgment. That is, with enough data to block all backdoor paths, we can establish real causality.

In Fig. 1, we block $\boldsymbol{V} \leftarrow \boldsymbol{D} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{S} \leftarrow \boldsymbol{D} \rightarrow \boldsymbol{Y}$ by intervening $V$ and $S$, represented by $P\left(Y|\, do\left(V,\ S\right)\right)$. Since there is no confounder between visual reasoning and textual reasoning, and the data are independent of each other, therefore:

$$P\left(Y|\, do\left(V,\ S\right)\right) = P\left(Y|\, do\left(V\right)\right) P\left(Y|\, do\left(S\right)\right). \quad (1)$$

$P\left(Y|\, do\left(V\right)\right)$ and $P\left(Y|\, do\left(S\right)\right)$ have the same structures in SCM. Here we take $P\left(Y|\, do\left(V\right)\right)$ in vision as an example to illustrate the backdoor judgment derivation. We block this backdoor path by changing the original training data $D$ into new data $\mathcal{T}_v$. We use lower case $v$, $y$, and $\tau_v$ to represent specific sample of $V$, $Y$, and $\mathcal{T}_v$. For interventional distribution, we use two rules:

**Rule 1** *Insertion/deletion of actions. When intervene $V$, the marginal distribution of $D$ is invariant, that is, $P(\tau_v|do(V)) = P(\tau_v)$.*

**Rule 2** *Action/observation exchange. When intervene $V$, The conditional probability of $Y$ in terms of $\mathcal{T}$ and $V$ is invariant, that is, $P(y|\tau_v, do(V)) = P(y|\tau_v, V)$.*

We can derive the desired interventional distribution by:

$$\begin{aligned} & P\left(Y|\, do\left(V\right)\right) \\ & = \sum_{\tau_v} P\left(Y|\, do\left(V\right), \mathcal{T}_v = \tau_v\right) P\left(\mathcal{T}_v = \tau_v|\, do\left(V\right)\right) && \text{(2a)} \\ & = \sum_{\tau_v} P\left(Y|\, do\left(V\right), \mathcal{T}_v = \tau_v\right) P\left(\mathcal{T}_v = \tau_v\right) && \text{(2b)} \\ & = \sum_{\tau_v} P\left(Y|V, \mathcal{T}_v = \tau_v\right) P\left(\mathcal{T}_v = \tau_v\right), && \text{(2c)} \end{aligned}$$

where Eq. 2a is based on the law of total probability. Eq. 2b follows Rule 1 and Eq. 2c follows Rule 2. The above equations can be summarized as:

$$P\left(Y|\, do\left(V\right)\right) = \sum_{\tau_v \in \mathcal{T}_v} P\left(Y|\, V, \tau_v\right) P\left(\tau_v\right). \quad (3)$$

The textual data $S$ is the same. Therefore, we can get Eq.1 in the main paper:

$$\begin{aligned} & P\left(Y|\, do\left(V,\ S\right)\right) \\ & = \sum_{\tau_v \in \mathcal{T}_v} P\left(Y|\, V, \tau_v\right) P\left(\tau_v\right) + \sum_{\tau_s \in \mathcal{T}_s} P\left(Y|S, \tau_s\right) P\left(\tau_s\right) \\ & && \text{(4)} \end{aligned}$$

**Algorithm 1:** Training process of MCR and Backbone Module

```
Input: Dataset: Video($V$), Question($Q$),
       Answer($A$).
Output: Trained parameters of MCR and Backbone
        Module (BM).
Initialize the parameters of MCR and BM;
for sample in paired V-Q-A do
    Detect the global object representations by Eq.3;
    Embed the question feature by Eq.5;
    Embed the answer feature by Eq.8;
    Capture the causal feature $a^c$ and irrelevant
      feature $a^{\bar{c}}$ in textual data by Eq.9;
    Optimize the textual part of MCR with $a^c, a^{\bar{c}}$ by
      Eq.12;
end
Intervene Answer $A$ to new data $\hat{A}$;
while Not done do
    for sample in paired V-Q-A($\hat{A}$) do
        Detect the global object representations by
          Eq.3;
        Capture the interaction feature by Eq.4;
        Embed the question feature by Eq.5;
        while $K_1$ epochs do
            Capture the causal feature $\mathcal{V}^c$ and
              irrelevant feature $\mathcal{V}^{\bar{c}}$ in visual data by
              Eq.6;
            Optimize the visual part of MCR with
              $\mathcal{V}^{\bar{c}}, \hat{\mathcal{V}}$ by Eq.11;
            Intervene Video $\mathcal{V}$ to new data $\hat{\mathcal{V}}$ by
              Eq.7;
        end
        while $K_2$ epochs do
            Capture the causal feature $a^c$ and
              irrelevant feature $a^{\bar{c}}$ in textual data by
              Eq.9;
            Optimize the textual part of MCR with
              $a^c, a^{\bar{c}}$ by Eq.12;
            Intervene Answer $A$ to new data $\hat{A}$;
        end
        Optimize BM with $A, \mathcal{V}, (\hat{A}, \hat{\mathcal{V}})$ by Eq.13;
    end
end
```

## B. Algorithm and More Results

**Algorithm.** To facilitate the reproduction of our work, we show the joint training process of the backbone model and the MCR model in Algorithm 1. $K_1$ and $K_2$ denote that we execute K2 epochs textual interventions after executing K1 epochs visual interventions for the visual causal module and the textual causal module.

Table 1. Ablation study on Causal-VidQA about visual causal loss weight ($\lambda_1$), textual causal loss weight ($\lambda_2$) and their elements.

| $\lambda_1, \lambda_2$ | 1, 3 | 1, 5 | 3, 1 | 5, 1 | 1, 1 |
|---|---|---|---|---|---|
| Acc | 49.75 | 49.45 | 47.98 | 46.32 | **50.96** |
| | w/o $\mathcal{L}_{c}^{v}$ | w/o $\mathcal{L}_{\bar{c}}^{v}$ | w/o $\mathcal{L}_{\bar{c}}^{s}$ | w/o $\mathcal{L}_{c}^{s}$ | All |
| Acc | 47.43 | 46.55 | 50.54 | 47.92 | **50.96** |

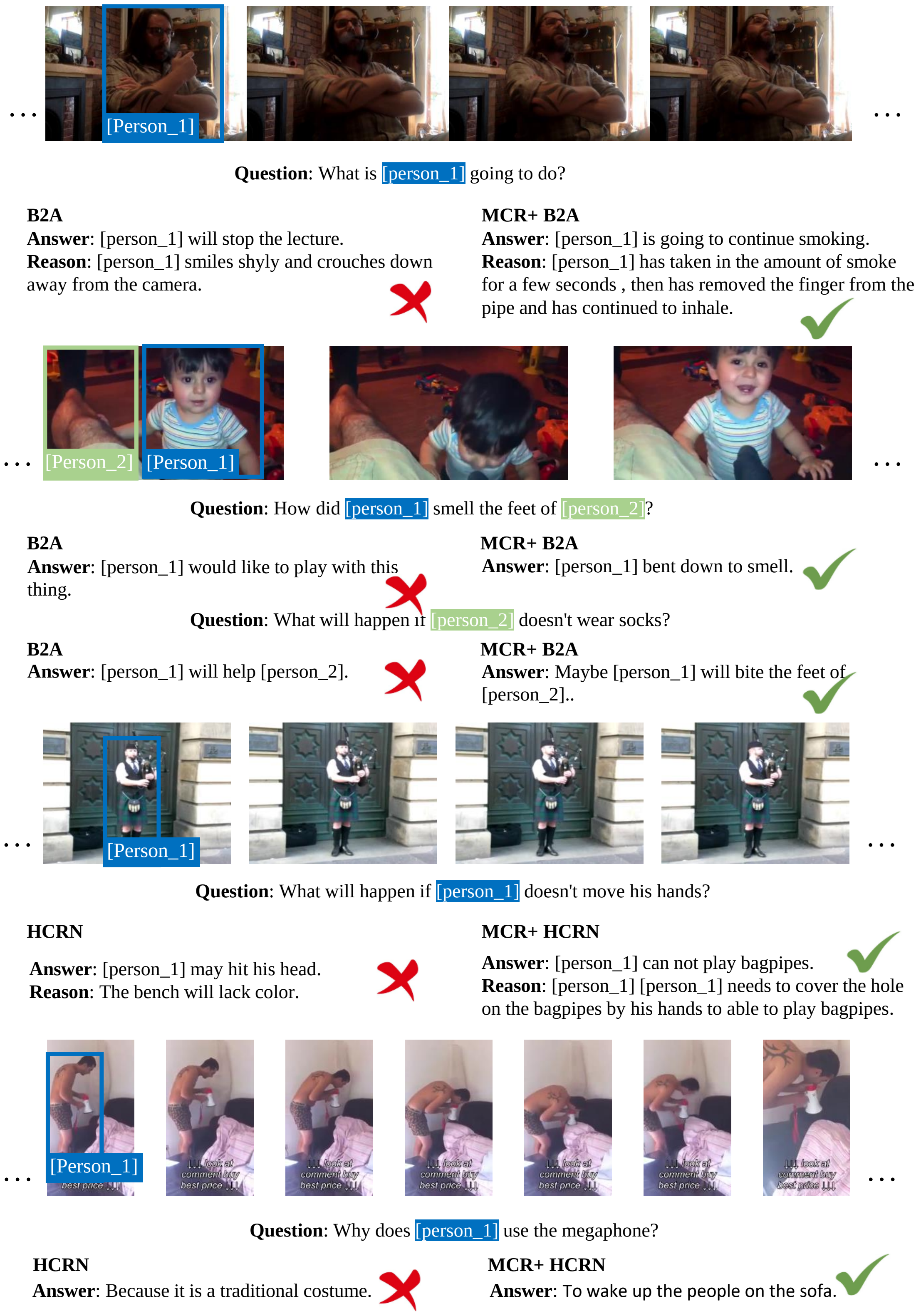
Table 2. Comparison of accuracy, number of parameters, and training time with state-of-the-art methods on Causal-VidQA dataset. **Acc** means accuracy. **Params** mean parameters of the architecture. **Time** indicates the time required for each backpropagation of the network. **FLOPs** means floating point operations per second in the training stage.

| | Acc | Params(M) | Time(s) | FLOPs(G) |
|---|---|---|---|---|
| HCRN [1] | 48.05 | 18.2 | 2.835 | 0.48 |
| MCR+HCRN | 50.86 | 22.6 | 14.91 | 6.45 |
| B2A [3] | 49.11 | 14.6 | 1.875 | 0.59 |
| MCR+B2A | 51.06 | 17.8 | 2.805 | 4.39 |

**Additional ablation study.** In Tab. 1, we conduct ablation studies on the setting of $\lambda_1$ and $\lambda_2$. When we increase the weight ($\lambda_2$) of the intervention visual data for training the model, the performance drops obviously. This may be because the intervention visual data has a greater impact on the model than the original data, making the model less generalizable on the test set which has a similar distribution to the original data. While the effect on intervention textual data ($\lambda_2$) is not obvious. Besides, we also show the effectiveness of loss functions for training MCR whose intervention effect affects the VideoQA performance of the backbone module. All loss functions are helpful for the selection of causal features, and help the model to focus on cause-and-effect related features that improve model generalization.

**Complexities.** In Tab. 2, we show model accuracy, the number of parameters, training time with (without) MCR, and floating point operations per second (FLOPs). We can see that compared to the backbone, i.e. HCRN and B2A, using MCR adds a small number of parameters and effectively improves accuracy. In terms of training time, since MCR requires additional 5 executions of the backbone inference model to obtain the results of causally related and irrelevant data on vision and text, it brings obvious time and FLOPs consumption. HCRN spends much time on feature inference, which also leads to a lot of time consumption in MCR. The time and FLOPs consumption on the B2A model is lower. Besides, it is worth mentioning that these costs are only incurred during training, and are consistent with the backbone during the inference phase.

**Qualitative evaluation.** In Fig. 2, we can observe spu-



**Question**: What is [person_1] going to do?

**B2A**
**Answer**: [person_1] will stop the lecture.
**Reason**: [person_1] smiles shyly and crouches down away from the camera. ✗

**MCR+ B2A**
**Answer**: [person_1] is going to continue smoking.
**Reason**: [person_1] has taken in the amount of smoke for a few seconds , then has removed the finger from the pipe and has continued to inhale. ✓

**Question**: How did [person_1] smell the feet of [person_2]?

**B2A**
**Answer**: [person_1] would like to play with this thing. ✗

**MCR+ B2A**
**Answer**: [person_1] bent down to smell. ✓

**Question**: What will happen if [person_2] doesn't wear socks?

**B2A**
**Answer**: [person_1] will help [person_2]. ✗

**MCR+ B2A**
**Answer**: Maybe [person_1] will bite the feet of [person_2].. ✓

**Question**: What will happen if [person_1] doesn't move his hands?

**HCRN**
**Answer**: [person_1] may hit his head.
**Reason**: The bench will lack color. ✗

**MCR+ HCRN**
**Answer**: [person_1] can not play bagpipes. ✓
**Reason**: [person_1] [person_1] needs to cover the hole on the bagpipes by his hands to able to play bagpipes.

**Question**: Why does [person_1] use the megaphone?

**HCRN**
**Answer**: Because it is a traditional costume. ✗

**MCR+ HCRN**
**Answer**: To wake up the people on the sofa. ✓

Figure 2. More examples with various question types demonstrate that our MCR helps various models find real associations and select the correct answers.

rious associations between videos and text. In textual association, as shown in the first example for answering the reason of the question, the local semantics of the sentence that is closely related to the video misleads the choice of the model. B2A chooses ”smiles shyly” and ”crouches down away” as the reason which matches the action in the video. While it over relies on the alignment relationship between words and video and ignores the requirements of the question. A reasonable approach is to infer a causally related answer based on the complete semantics of the question and answer.

In visual association, as shown in the last example, the ”megaphone” is not a usual object for waking up people. It appears in the ”traditional costume” more frequently. Focusing on the statistical relation instead of the human motion and interaction occurring in the video introduces untrustworthy ways of reasoning. Our MCR can change statistical spurious associations and remove data bias, allowing models to focus on appearance features and motion features in videos.

## References

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

[2] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21273–21282, 2022. 1

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

[4] Judea Pearl. *Causality*. Cambridge university press, 2009. 1