北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

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

THU-AM-243

Chuanqi Zang, Hanqing Wang, Mingtao Pei, Wei Liang

School of Computer Science and Technology, Beijing Institute of Technology
Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing

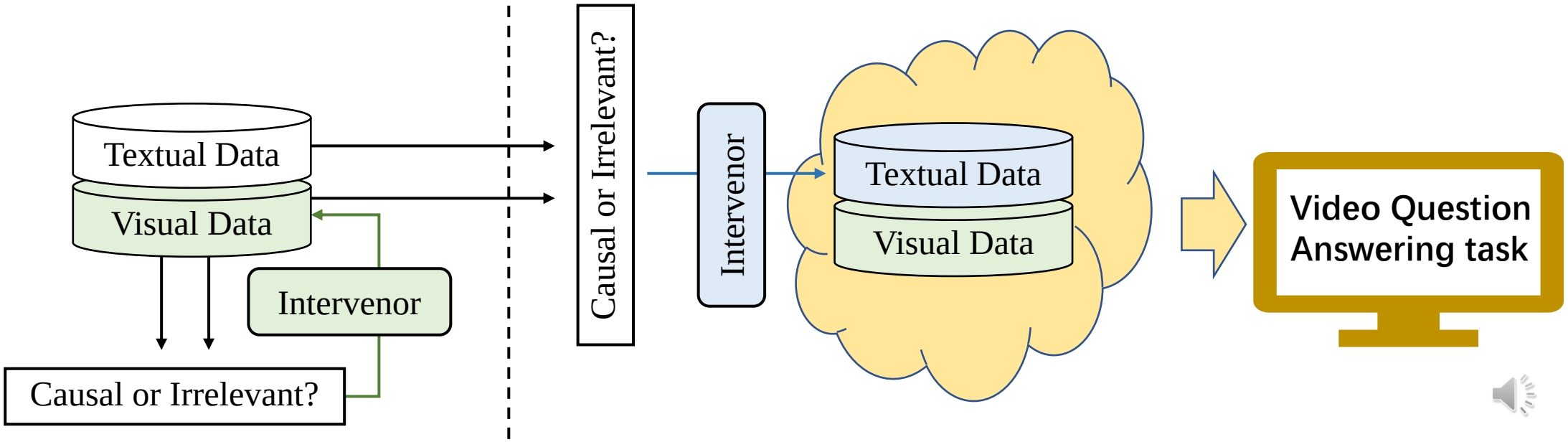
Preview

Visual Data

Objects, background, actions appear in the video, which one is a **clue** for the question?

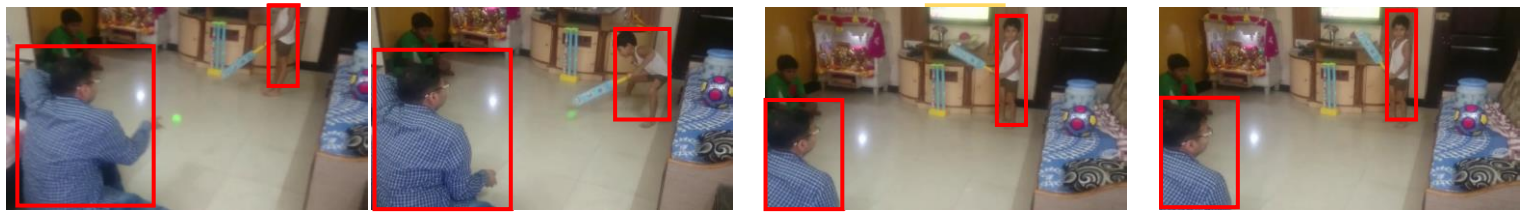
Textual Data

When the answer is a paragraph, can **key words** represent sentence semantics?



Background

Answer a series of questions based on the video.



question: "Where is [person_2]?“
answer: "[person_2] is sitting in a [person_2].",
"[person_2] is sitting in a car.",
"[person_2] is in a house.",
"[person_2] is at the station and standing near exit entrance."
"[person_2] is standing next to the slide."

question: "Why is [person_2] holding the stick? “
answer: "Because she cannot open it.",
"[person_1] wants to train to change its body.",
"[person_2] is having fun with [person_1].",
"Because [person_1] is smoking hookah.",
"To help push the process of the proposal."

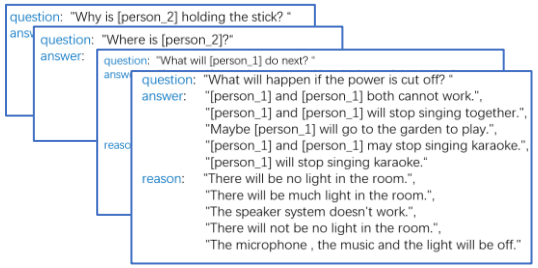
question: "What will [person_1] do next? “
answer: "[person_1] is ready to start making the base of the device.",
"[person_1] may want to pick up the ball next.",
"[person_1] will have a rest.",
"[person_1] is bound to keep playing piano.",
"It is predicted that [person_1] will give the ball for [person_2].“
reason: "It is heavy rainy.",
"[person_1] seems to enjoy it a lot.",
"[person_1] wears skates , holds a ice hockey stick and skates around the ice rink.",
"Because the bike got stuck.",
"[person_1] would like to take care of [person_2]."

question: "What will happen if the power is cut off? “
answer: "[person_1] and [person_1] both cannot work.",
"[person_1] and [person_1] will stop singing together.",
"Maybe [person_1] will go to the garden to play.",
"[person_1] and [person_1] may stop singing karaoke.",
"[person_1] will stop singing karaoke.“
reason: "There will be no light in the room.",
"There will be much light in the room.",
"The speaker system doesn't work.",
"There will not be no light in the room.",
"The microphone , the music and the light will be off."

Background

Existing work:

Zeng et al., 2017; Yang et al., 2020



Question; Answer

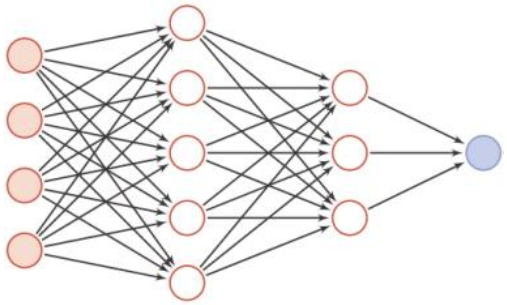
Textual encoder

Video frames

Visual encoder

Jang et al., 2017; Xu et al., 2017;
Zhao et al., 2017;

video-text alignment and fusion



Memory Networks: Tapaswi et al., 2016; Kim et al., 2019; Gao et al., 2018; Fan et al., 2019

Transformer: Li et al. ,2019; Lei et al., 2021; Yang et al., 2021; Zellers et al., 2021; Fu et al., 2021

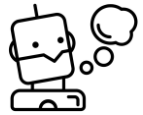
Graph Neural Networks: Huang et al., 2020; Jiang and Han, 2020; Park et al., 2021; [Wang et al., 2021

Background

Problems

Statistical bias exits in visual elements:

Question: What will happen if the power is cut off?



TV, indoor → singing karaoke

Answer:

A. [person_1] will stop singing karaoke. **Predict**

B. [person_1] and [person_2] both cannot work. **G.T.**

Background

Problems

Statistical bias exits in textual elements:

Question: What will happen if the girl sprains?

Answer: The girl will stop.

Someone, help → **Reason**

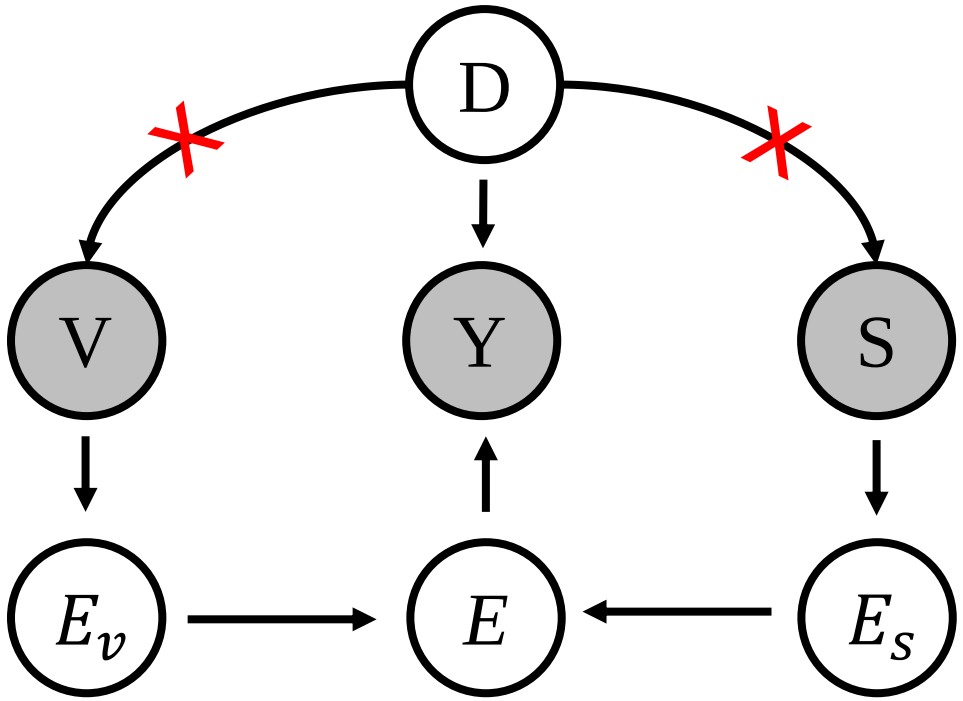
Reason:

A. There are a lot people here, and can find someone to help at any time. **Predict**

B. The girl can't exercise because of a sprain and needs to rest. **G.T.**

Model

Causal model



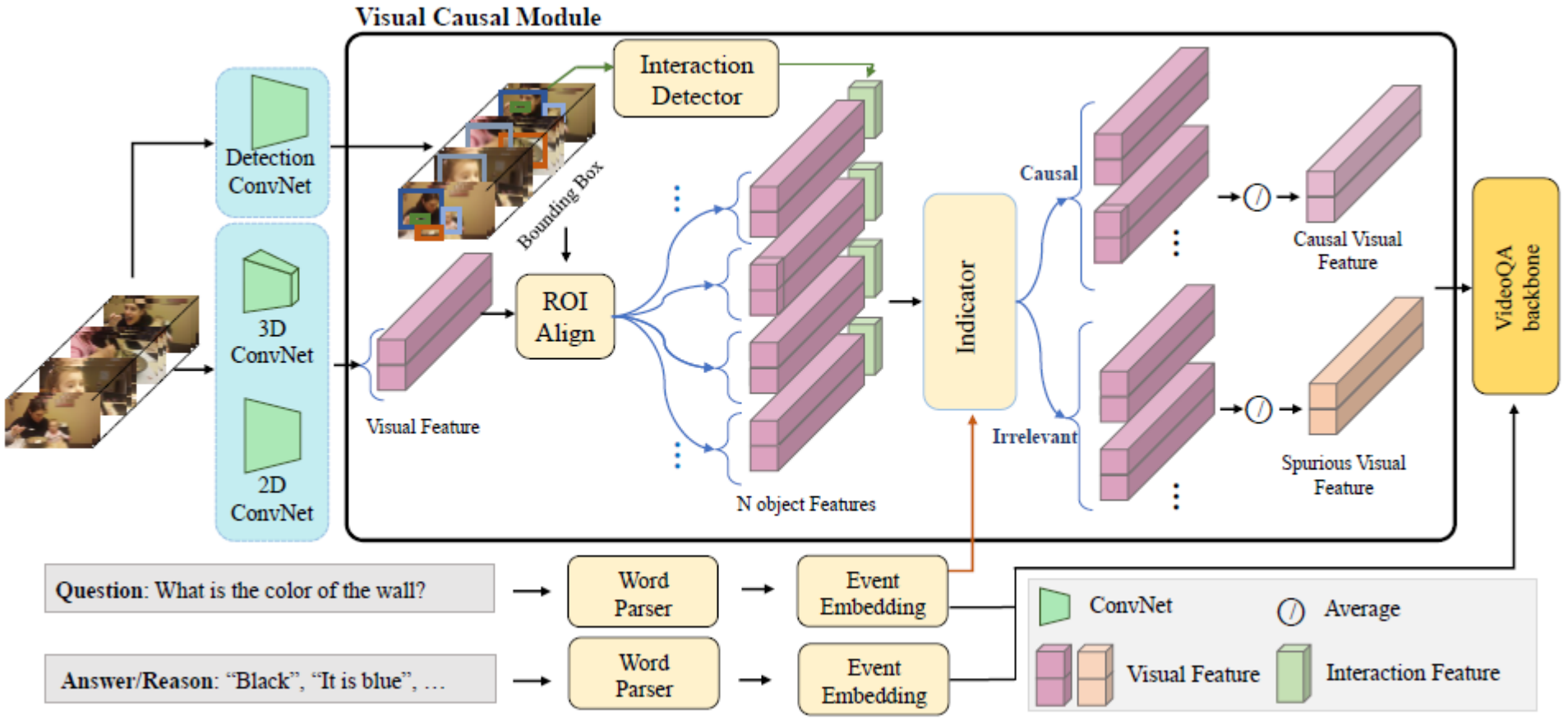
$$P\left(Y|\ do\left(V,\ S\right)\right) = \sum_{\tau \in \mathcal{T}} P(Y|V,S,\tau)P(\tau)$$

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

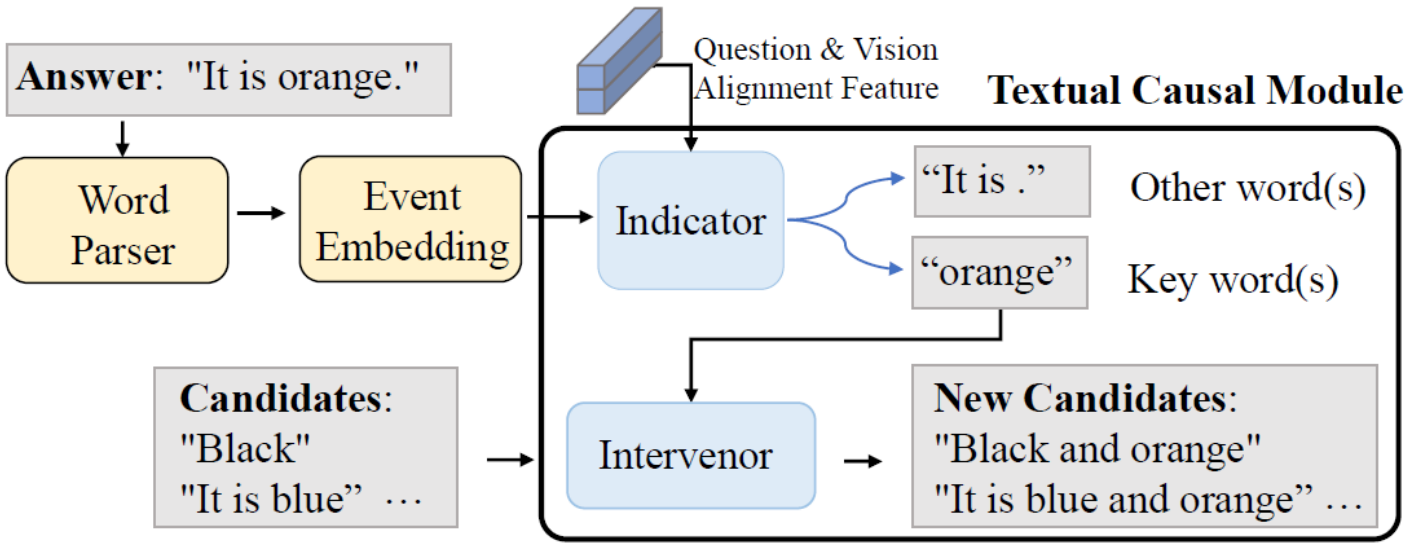
$$+ \sum_{\tau_s \in \mathcal{T}_s} P\left(Y|S,\tau_s\right) P\left(\tau_s\right)$$

Judea Pearl. *Causality*. Cambridge university press, 2009

Visual Causal Module



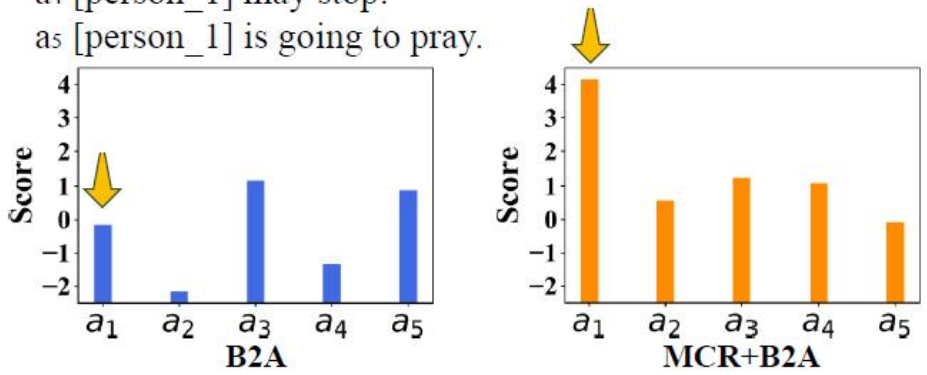
Textual Causal Module



Visualization

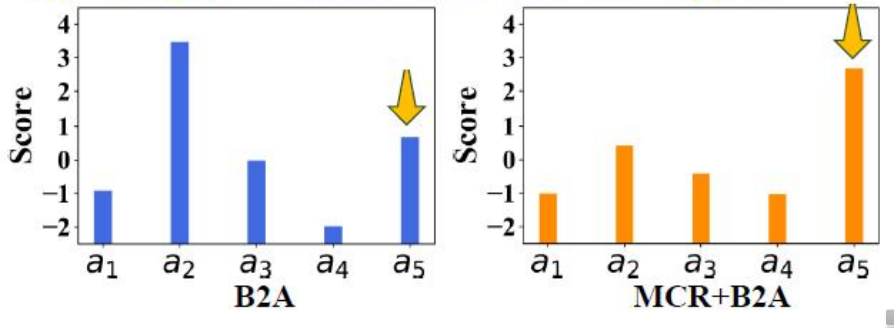
Question: What is [person_1] going to do?

✓ a1 **[person_1] is going to continue smoking.**
a2 [person_1] is going to play the flute.
✗ a3 **[person_1] will stop the lecture.**
a4 [person_1] may stop.
a5 [person_1] is going to pray.



Question: What will [person_1] do next?

a1 [person_1] will continue to perform after [person_1].
✗ a2 **[person_1] will lift the third one.**
a3 [person_1] will separate the two legs.
a4 [person_1] will keep on dancing.
✓ a5 **[person_1] will further introduce this stopper.**



Summary

Main contributions:

- Discover two new types of causal challenges for both visual data and textual data.
- Propose an object-level causal relationship extraction strategy to establish the real association between objects and language semantics
- Propose a keyword broadcasting strategy to cut off the spurious influence of local textual information.

Thanks for your watching!