CausalChaos! Dataset for Comprehensive Causal Action Question Answering Over Longer Causal Chains Grounded in Dynamic Visual Scenes
Appendix/Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

A Appendix / supplemental material

We have provided the Appendix (**a PDF file + PowerPoint presentation**). Alternatively, Appendix can be obtained from here: `https://drive.google.com/drive/folders/1TytAcmDw11fbSCtKDlSWFoLeZyC6fgOX?usp=sharing`. This is with the permission of Chairs.

Contents

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

A.1 Extended Related Work

A.1.1 Datasets/Benchmarks

To drive the progress in Video QA, researchers have developed various datasets with distinct focuses and contributions. In the following, we will delve into the landscape of Video QA datasets, examining their characteristics, limitations, and the specific areas they address.

Early Video QA Datasets. During the early stages of Video QA research, several datasets relied on video captions or descriptions to automatically generate questions and corresponding answers. Examples of these datasets include *MovieFIB* [15], *MSVD-QA* [20], *MSRVTT-QA* [20], *YouTube2Text* [6], *open-ended QA*, *Zeng et al.*, and *Video Context-QA*[26]. These datasets played a crucial role in the initial exploration of Video QA but were primarily limited to object and action recognition. They lacked the ability to go beyond these basic visual cues, which posed limitations in understanding complex interactions and causal relationships within videos.

TGIF-QA [7, 8] focuses on short videos and relies on captions to generate questions and answers, but it is limited in its coverage of object interactions and causal reasoning. On the other hand, *ActivityNet-QA* [23] annotates longer web videos, offering a broader range of content, but it also lacks in capturing complex reasoning. Both datasets fall short in capturing the depth and complexity required for comprehensive video question answering; they did not fully explore questions and answers involving object interactions and causal relationships.

The *Social-IQ* [24] dataset is designed to address questions related to human social behavior in videos, relying on multimodal cues for answering. This dataset emphasizes the importance of understanding social interactions and dynamics within video content. By focusing on human behavior, Social-IQ offers a unique perspective in video question answering. However, it should be noted that this dataset primarily relies on multimodal cues, meaning that the answers to the questions heavily depend on the combination of visual and other sensory information, and as such cannot be used as a visual reasoning dataset. While it provides valuable insights into social aspects of video content, the dataset may not fully capture the broader context and reasoning required for comprehensive video question answering.

CLEVRER [22] dataset covers temporal and causal relationships using collision events between various objects. Being limited to simple, inanimate objects and collision events, it does not cover reasoning involving emotions and intentions; objects do not have any characteristics; actions do not have motivation or rationale; limited set of events; scene does not have an actually involved background; character-object interaction is lacking. Moreover, the reasoning is over a shorter temporal horizon than ours. Since it is a synthetic dataset with programmatically generated QA pairs, there is also a lack of diversity in natural language descriptions of the events and human judgments. *CLEVRER-Humans* [17] bridges this language gap, but, other shortcomings still persist.

AGQA [5] focuses on spatio-temporal scene understanding. For example, Did they <action1> or <action2>for longer? What did the person do after <action>? What were they <action> first/last? *STAR*, a situation reasoning dataset, additionally, covers prediction and feasibility questions. However, they do not cover explanatory “why” questions like ours.

We have discussed and compared with *NextQA* [19], *CausalVidQA* [11], *IntentQA* [13] in detail in the main paper. Here we provide some additional details on them. NextQA [19] contains descriptive (related to location, counting, yes/no), temporal (related to temporal ordering previous, next), and causal questions. CausalVidQA [11] contains descriptive, causal, predictive, and counterfactual. While they provide a rationale for predictive and counterfactual, they do not explore and provide multi-level answers and explanations for Why-questions, while our dataset does provide them. IntentQA [13], a concurrent dataset explores understanding motivations based on context. Their dataset is derived from NextQA causal and temporal questions, but they construct their dataset in a contrastive manner such that the same actions under different contexts lead to different underlying intents.

A.1.2 Models

In the following, we have discussed the state-of-the-art VideoQA models that we have benchmarked on VisCAQA dataset in the main paper. We have discussed their central concepts and unique design characteristics.

- *BlindQA* [1]. In this approach, no visual information is leveraged. Answers are chosen directly based on the questions. In a nutshell, this model learns a mapping from question to answer. Higher performance by method would suggest that the dataset contains questions that are not visually-grounded.
- *EVQA* [1]. This method extends BlindQA baseline by incorporating the visual stream modeled by an LSTM.
- *Spatio-Temporal Reasoning in Visual Question Answering (STVQA)* [7]. This work introduces three novel video QA tasks that demand spatio-temporal reasoning skills to answer questions accurately. In addition, a new TGIF-QA dataset has been created to facilitate research in this field. To address this issue, a dual-LSTM-based approach with both spatial and temporal attention mechanisms has been proposed as a baseline model.
- *Motion-Appearance Co-Memory Networks (CoMem)* [4]. A novel Video QA framework, combining Dynamic Memory Network (DMN) principles with motion and appearance features. This innovative approach leverages a co-memory attention mechanism to incorporate both motion and appearance cues. It employs a temporal conv-deconv network to create multi-level contextual information and utilizes a dynamic fact ensemble method for constructing dynamic temporal representations tailored to specific questions.
- *Heterogeneous Memory Enhanced Multimodal Attention Model (HME)* [2].This innovative end-to-end trainable Video QA framework begins by generating global context-aware visual and textual features. It achieves this by interacting the current inputs with memory contents. Subsequently, it integrates these multimodal features through attentional fusion to make accurate inferences for answering questions.
- *HCRN* [10]. This is a hierarchical framework with conditional relation networks as building blocks models input video at multiple scales (clip-, full video-level) in a cascaded manner. Visual features at each level are conditioned on the question features. The joint representation is fed into the classifier for answer prediction.
- *HGA* [9]. Leverages heterogeneous graph reasoning module and a co-attention unit to capture the local and global correlations between video clips, linguistic concepts and their cross-modal correspondences.
- *Multimodal Iterative Spatial-temporal Transformer (MIST)* [3]. MIST, designed for long-form Video Question Answering (VideoQA), revolutionizes conventional dense spatial-temporal self-attention. It accomplishes this by utilizing two critical modules: segment and region selection, which adaptively pick out frames and image regions tied to the questions. Following this, it processes diverse visual concepts effectively with an attention mechanism. This process occurs iteratively across multiple layers, empowering the model with multi-event reasoning capabilities.

A.2 Implementation details

We use the publicly available github code repository: `https://github.com/doc-doc/NExT-QA` for BlindQA, EVQA, CoMem, HME, HCRN, and HGA.

A.3 Details regarding MIST-CC

We design MIST-CC, a multitask version of MIST that learns to generate causal chains as an auxiliary task. With MIST-CC, our goal is not to generate perfect causal chains during testing, per se; but to focus on providing guiding signal to the model to improve the performance on question-answering task. Framework for MIST-CC is shown in Figure 1. We build upon the publicly available implementation of MIST. We implement the Causal Chain Generator in our MIST-CC framework using a single-layer gated recurrent unit (GRU) with a 1024-dimensional hidden state; and dropout rate of 0.2. The overall multitask (MTL) objective function to be minimized is the summation of: 1) multichoice question answering loss ($\mathcal{L}_{MCQA}$); 2) causal chain generation loss ($\mathcal{L}_{CCG}$) (Equation 1). We set α to 1. To obtain vanilla MIST, we set β to 0; while to obtain MIST-CC, we set β to 0.1. We use ADAM optimizer with an initial learning rate of 1e-4. We do not use learning rate schedulers. All the models are trained for 30 epochs, and the best version of a model is selected based on the performance on the validation set.

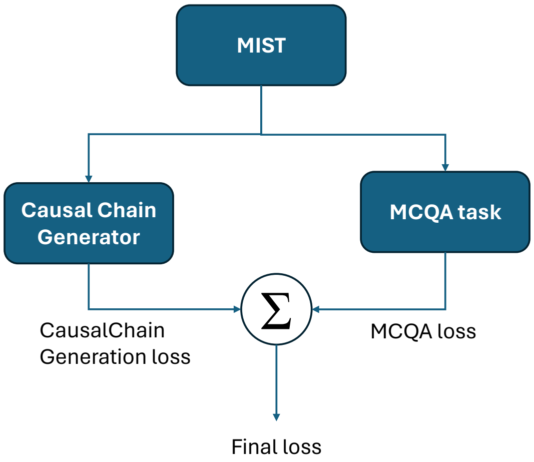


Figure 1: **MIST-CC framework.**

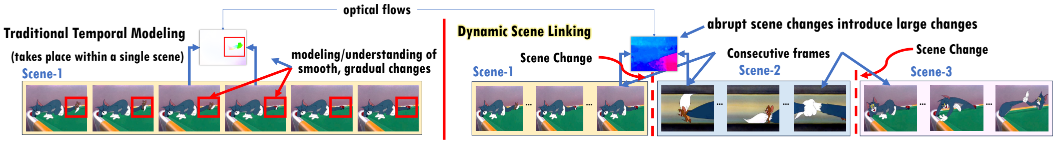

optical flows
Traditional Temporal Modeling
(takes place within a single scene)
Scene-1
modeling/understanding of smooth, gradual changes
Dynamic Scene Linking
abrupt scene changes introduce large changes
Scene Change
Scene-1
Consecutive frames
Scene-2
Scene Change
Scene-3

Figure 2: **Challenge of dynamic scene linking.** Here we have shown temporal modeling or understanding in a traditional sense; and compared it with dynamic scene linking. Our full causal chain for this particular example consisted of three scenes. In traditional temporal modeling, typically the dependencies are modeled over gradual transitions—typically, within a single scene. Notice the optical flow from traditional temporal modeling where Jerry is going into a pool table hole. On the other hand, notice the abrupt scene change, which causes disruption in visual flow, resulting in large amplitude and widespread optical flow. We have used optical flow as one way to illustrate the magnitude of change (but other measures may also be used).

$$\mathcal{L}_{MTL} = \alpha\mathcal{L}_{MCQA} + \beta\mathcal{L}_{CCG} \tag{1}$$

A.4 Annotators' Background

Five undergraduate students from computer science and electrical engineering disciplines were recruited as annotators. All the annotators listed at least Fluency as the English language skills level.

A.5 Dynamic Scene Linking

Our CausalChaos dataset involves more abrupt/frequent scene (event) changes than existing causal video QA datasets. Causal links or clues needed to solve causal relationships in QA pairs in our dataset are embedded in different scenes/events. Thus, video QA models must link these scenes/events together to understand the story. We term this problem Dynamic Scene Linking. In the following, we briefly discuss a related problem of temporal modelling and differentiate Dynamic Scene Linking from it. Although humans can seamlessly link such scenes, dynamic scene linking introduces a novel challenge for video understanding models in addition to temporal modelling.

Temporal modelling or understanding in video understanding generally refers to the process of analyzing and interpreting the temporal dynamics or changes within a sequence of frames in a video—typically within a single scene (refer to Figure 2). This involves capturing and understanding the patterns of motion, action, and context over time. Temporal modeling techniques aim to extract meaningful information from the temporal dimension of video data.

In the traditional sense, temporal modelling involves techniques that capture the gradual changes and transitions occurring within a video sequence. This includes methods like optical flow and 3D convolutional neural networks. These techniques are designed to capture the temporal dependencies and patterns of continuity or gradual evolution in videos.

On the other hand, abrupt scenes or shot changes, such as those found in cartoons like Tom and Jerry, represent sudden and significant shifts in the content or context of a video, however, these changes are causally linked. These changes can include shifts in location, characters, actions, or camera perspectives. Unlike gradual temporal changes, abrupt changes occur rapidly and may disrupt the continuity of the narrative or visual flow. While temporal understanding typically involves linking very nearby dependencies, dynamic scene linking involves linking across abrupt scene changes. For example, in Figure 2, in Scene-2, some of the things the model needs to be able to understand are: 1) the thing that Jerry is carrying is Tom's tail from its partial observation; 2) white furry hand is of Tom. Implicit causal reasoning plays a crucial role in establishing continuity between scenes, even when objects are seen partially or there are view changes. By relying on their understanding of cause-and-effect relationships within the narrative, humans can seamlessly integrate partial views and view changes into their mental model of the story. Similarly, models are required to do implicit causal reasoning for dynamic scene linking, and can benefit from incorporating capabilities such as forming a mental 'world model' of the story.

Temporal modeling techniques in the context of abrupt scene changes need to be able to detect and handle these sudden transitions effectively. While some traditional temporal modeling methods may capture gradual changes well, they might struggle to handle abrupt changes efficiently. Specialized algorithms or models may be required to identify and adapt to such abrupt scene changes.

A.6 Full-size Tables and Figures

Examples from CausalChaos! dataset. For easier viewing, we have provided dataset video examples from Figure 1 from the main paper in the accompanying PowerPoint presentation.

CausalChaos! vs NextQA dataset. For easier viewing, we have provided dataset video examples from Figure 2(c) from the main paper in the accompanying PowerPoint presentation.

A.7 Further Dataset Stats and Examples

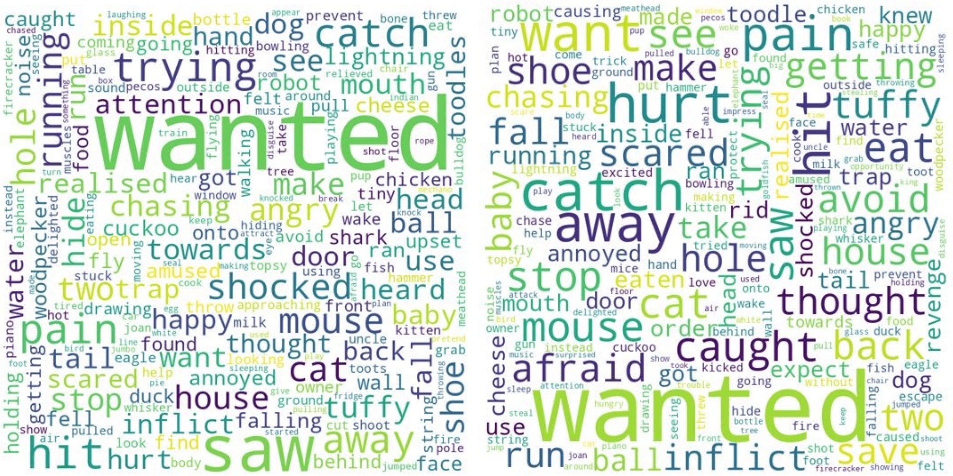
Dataset Stats:

- Average clip length : 357.95 frames
- Longest of clips : 2315.0 frames
- Average length of question : 6.54 words
- Longest length question : 17 words
- Longest question : Why did Jerry and Tuffy put the wire into the water and turn on the freeze mode?
- Average length of Answer : 7.76 words
- Longest length Answer : 26 words
- Longest Answer : Tom saw his tail and hind legs at the top of the pipe while Tom's head and front legs were at the bottom of the pipe.
- Average length of Explanation: 13.88 words
- Longest length Explanation: 30 words
- Longest Explanation: Tom thought Jerry would walk into the hole and into Tom's mouth but Jerry let the toy mouse go first and Tom ate the toy mouse thinking it was Jerry.

Dataset examples. For easier viewing of the videos, we have provided them in the accompanying PowerPoint presentation.

A.8 Wordclouds

Wordclouds are illustrated in Figure 3.



(a) (b)

Figure 3: **Wordclouds** corresponding to (a) **Answers**; (b) **Explanations**.

A.9 Value in Multi-level Answers to Why-Questions

We richly annotate our dataset with multi-level answers to "Why"-questions behind the actions of characters in the Tom and Jerry cartoon. In this section, we discuss why and how a significant value lies in providing multi-level answers to "why" questions (or exploring various layers of causality or explanation) regarding characters' actions (or, in general life, people's actions). "Why" questions do not always have simple, straightforward answers—they often involve multiple layers of explanation and understanding. *In our dataset, we consider cartoon characters and their actions, but here for more generality, we take human behavior as a case for discussion.* Human behavior is typically multifaceted, influenced by a variety of factors including personal experiences, emotions, social context, cultural background, cognitive processes, etc.

There can be multiple layers of rationale behind any action or decision. For example, someone might choose to volunteer at a homeless shelter. On the surface, the reason may seem obvious – to help those in need. But delving deeper, you might find additional motivations such as personal fulfillment, a desire to contribute to the community, religious beliefs, or even social pressure from peers.

Recognizing the complexity of human behavior and understanding that there can be multiple, intertwined reasons for why people act the way they do is essential for empathy, effective communication, and building strong interpersonal relationships. Some of the ways multi-level answers to 'Why'-questions help are as follows:

- Understanding Motivations: At the surface level, a person's actions may seem straightforward, but delving deeper can reveal the underlying motivations and intentions driving those actions. Multi-level answers can help uncover these motivations, providing a more comprehensive understanding of human behavior.
- Contextual Understanding: Human behavior is complex and influenced by a variety of factors, including personal experiences, societal norms, cultural background, and psychological factors. Providing multi-level answers allows for a more nuanced understanding of the context in which the behavior occurs, shedding light on the various influences at play.
- Predictive Insights: By understanding the multiple layers behind people's actions, it becomes easier to predict future behavior. Recognizing patterns in motivations and behaviors can help anticipate how individuals might act in different situations, enabling better decision-making and planning.
- Empathy and Compassion: Exploring the deeper reasons behind someone's actions fosters empathy and compassion. It allows us to see beyond the surface behavior and understand the person's perspective, experiences, and struggles, leading to more meaningful interactions and relationships. While this might be more of a human experience-related factor and might

be limited to very specialized machines like empathetic robots, but still shows potentially how multi-level answers help.

- Problem Solving and Conflict Resolution: In situations where conflicts arise or problems need to be addressed, understanding the multi-level reasons behind people's actions can facilitate more effective problem-solving and conflict-resolution strategies. It enables individuals to address underlying issues rather than just surface-level symptoms.

Overall, providing multi-level answers to “why" questions behind people's actions enhances our understanding of human behavior, promotes empathy and compassion, and facilitates better decision-making and problem-solving.

A.10 CausalChain Details and Examples

In the following, we have provided some examples from our dataset to illustrate causal chains of different lengths. Note that, answer and explanation were combined when computing the length of causal chains.

> Question: Why did Jerry slide?
> Answer+Explanation: Jerry slid down the clock because he saw Spike trying to catch Tom and was confident that Spike's attention was on Tom and not Jerry.

Length of causal chain: 2. In the given event involving Jerry, Spike, Tom, and a clock, we can identify several causal relationships that form a chain of events. Let's break it down:

1. Event A: Jerry sees Spike trying to catch Tom.
2. Event B: Jerry is confident that Spike's attention is on Tom.
3. Event C: Jerry slides down the clock.

Now, let's consider the causal relationships:

- Event A causes Jerry's perception of Spike's actions.
- Event B is influenced by Jerry's perception of Spike's actions.
- Event B causes Jerry's confidence in Spike's attention being on Tom.
- Event C is influenced by Jerry's confidence in Spike's attention.

So, we can identify at least two causal links or events in this scenario. Each event contributes to the next in a causal chain, leading to the final action of Jerry sliding down the clock.

> Question: Why did Jerry go below chicken?
> Answer+Explanation: Jerry went below the chicken sitting on her nest to hide and get protection from Tom.

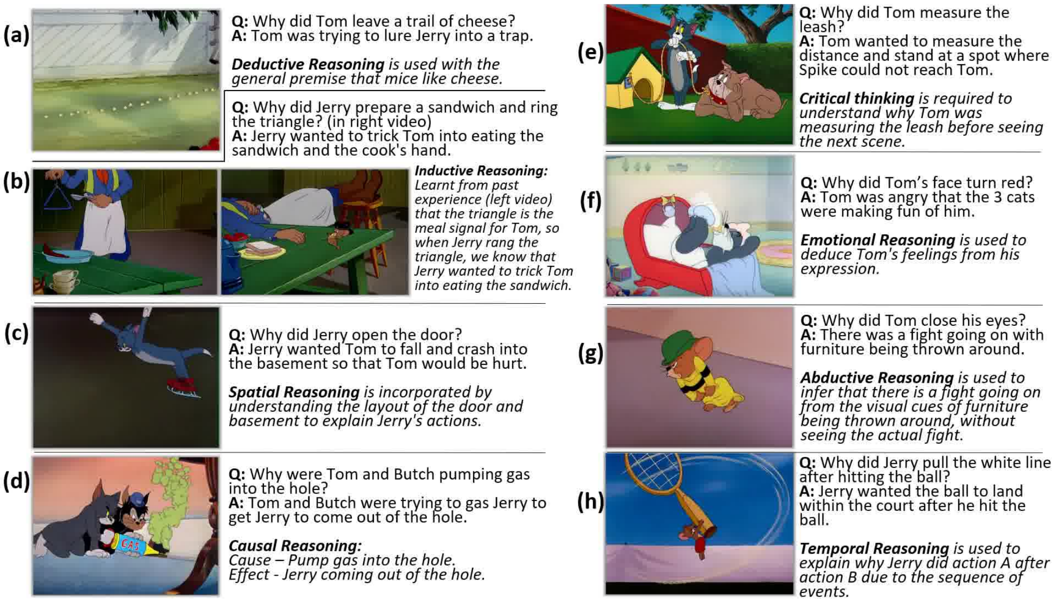
Breakdown of the events and causal links in this case is:

- Event A: Tom poses a threat to Jerry.
- Event B: Jerry seeks protection and safety.
- Event C: Jerry goes below the chicken sitting on her nest as a protective measure.

Here, we can identify that there are *two* causal links in the causal chain of the event.

A.11 Further details on Causal Chain Length Comparison Experiment

We leveraged GPT-4o [18] to compute the lengths of causal chains involved in the QA pairs from our and existing causal video QA datasets [19, 11, 13]. We randomly sampled 100 causal-Why QA pairs from all the datasets. Then, we used the following prompt: “What is the causal chain in the following question-answer pair? Please return the causal chain in the form of event_A->event_B->event_C...If



Q: Why did Tom leave a trail of cheese?
A: Tom was trying to lure Jerry into a trap.

Deductive Reasoning *is used with the general premise that mice like cheese.*

Q: Why did Jerry prepare a sandwich and ring the triangle? (in right video)
A: Jerry wanted to trick Tom into eating the sandwich and the cook's hand.

Inductive Reasoning: *Learnt from past experience (left video) that the triangle is the meal signal for Tom, so when Jerry rang the triangle, we know that Jerry wanted to trick Tom into eating the sandwich.*

Q: Why did Jerry open the door?
A: Jerry wanted Tom to fall and crash into the basement so that Tom would be hurt.

Spatial Reasoning *is incorporated by understanding the layout of the door and basement to explain Jerry's actions.*

Q: Why were Tom and Butch pumping gas into the hole?
A: Tom and Butch were trying to gas Jerry to get Jerry to come out of the hole.

Causal Reasoning:
Cause – Pump gas into the hole.
Effect - Jerry coming out of the hole.

Q: Why did Tom measure the leash?
A: Tom wanted to measure the distance and stand at a spot where Spike could not reach Tom.

Critical thinking *is required to understand why Tom was measuring the leash before seeing the next scene.*

Q: Why did Tom's face turn red?
A: Tom was angry that the 3 cats were making fun of him.

Emotional Reasoning *is used to deduce Tom's feelings from his expression.*

Q: Why did Tom close his eyes?
A: There was a fight going on with furniture being thrown around.

Abductive Reasoning *is used to infer that there is a fight going on from the visual cues of furniture being thrown around, without seeing the actual fight.*

Q: Why did Jerry pull the white line after hitting the ball?
A: Jerry wanted the ball to land within the court after he hit the ball.

Temporal Reasoning *is used to explain why Jerry did action A after action B due to the sequence of events.*

Figure 4: **Examples of various types of reasoning required by our dataset.** *Please zoom in & view in AdobeReader to play the embedded videos.*

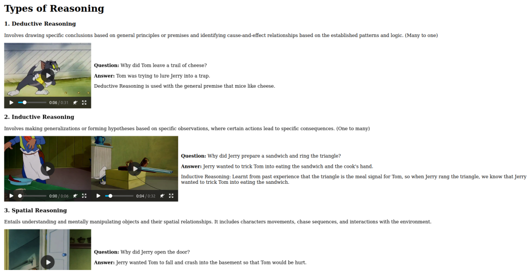
no cause-and-effect relationship is addressed, then output 0...Question: [question added here] Answer: [answer added here]." to ask GPT-4o to obtain the causal chain in each QA pair. We define the length of a causal chain as the number of links in that chain. For example, "event_A->event_B" has a length of 1, while "event_A->event_B->event_C" has a length of 2. Since our dataset contains multi-level answers, we combined answers and explanations using GPT-4 to get an overall answer to reflect the true length of the full causal chain involved. We use these overall answers when computing the causal chains for our dataset. Extracted causal chains from all datasets were manually verified. Human verifiers agreed 89% of the times with the causal chains. Once the lengths of causal chains for all the samples are computed, we average them to get the average causal chain length for a dataset. We repeat the process for all datasets and then compare them.

A.12 Details on Types of Reasoning

A.12.1 Definitions of Types of Reasoning

In the following, we have provided the definitions of various types of reasoning.

1. **Deductive Reasoning** involves drawing specific conclusions based on general principles or premises. Questions from our dataset can require video understanding models to answer based on established patterns or cause-and-effect relationships between characters' actions (*e.g.*, Figure 4(a)).
2. **Inductive Reasoning** involves making generalizations or forming hypotheses based on specific observations. Tom and Jerry episodes contain such episode-specific actions or features or nuances, *e.g.*, as in Figure 4(b). Answering causal questions related to such actions involves inductive reasoning.
3. **Spatial Reasoning** involves predicting and understanding spatial relationships or configurations. Our dataset requires Video-QA models to have an understanding of the physical space and how the characters navigate it, interactions with the environment, including concepts such as distance, direction, and obstacles (crashing in or avoiding it). For example, as shown in Figure 4(c).
4. **Causal Reasoning** involves understanding cause-and-effect relationships between actions and their consequences. In the process of answering questions in our dataset, Video QA models will be required to engage in causal reasoning by linking the characters' actions or



Types of Reasoning

1. Deductive Reasoning

Involves drawing specific conclusions based on general principles or premises and identifying cause-and-effect relationships based on the established patterns and logic. (Many to one)

Question: Why did Tom leave a trail of cheese?

Answer: Tom was trying to lure Jerry into a trap.

Deductive Reasoning is used with the general premise that mice like cheese.

2. Inductive Reasoning

Involves making generalizations or forming hypotheses based on specific observations, where certain actions lead to specific consequences. (One to many)

Question: Why did Jerry prepare a sandwich and ring the triangle?

Answer: Jerry wanted to trick Tom into eating the sandwich and the cook's hand.

Inductive Reasoning: Learnt from past experience that the triangle is the meal signal for Tom, so when Jerry rang the triangle, we know that Jerry wanted to trick Tom into eating the sandwich.

3. Spatial Reasoning

Entails understanding and mentally manipulating objects and their spatial relationships. It includes characters movements, chase sequences, and interactions with the environment.

Question: Why did Jerry open the door?

Answer: Jerry wanted Tom to fall and crash into the basement so that Tom would be hurt.

Figure 5: **Screenshot of a guide used as a part to explain types of reasoning to human subjects.**

sub-events within an episode to the resulting consequences and understanding the cause-and-effect chains in the cartoon as depicted in Figure 4(d).

5. **Critical Thinking** in our setting encompasses a range of cognitive processes, including analysis, and evaluation by analyzing the visual cues, and interpreting the characters' actions. A way to judge the complexity of critical thinking questions is by measuring the lengths of causal chains. To get the full picture, it is also important how difficult each link in this causal chain is to be inferred from the video (refer to Figure 4(e))

6. **Emotion Reasoning** involves recognition and understanding of emotions and how they can affect behavior and decision-making. Our dataset requires models to perform emotion/facial expression recognition and link them to characters' actions/behaviors. For example, as shown in Figure 4(f).

7. **Abductive reasoning** involves making an inference or hypothesis based on limited or incomplete information, in order to explain or interpret a situation or phenomenon. Our dataset contains questions that involve making inferences from partial information, *e.g.*, it is to be inferred that Tom was scared because there is a fight going on from the visual cues of furniture being thrown around, without seeing the actual fight as shown in Figure 4(f).

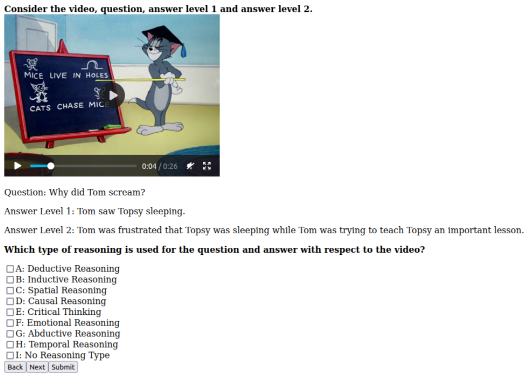
8. **Temporal Reasoning** refers to understanding and reasoning about the sequence/ordering of events over time—understanding the relationships between different actions, and identifying causal relationships amongst them (Why A is done before/after B) as in Figure 4(g).

A.12.2 Human Study Details

Human subjects in human studies had a background in the disciplines of computer science and electrical engineering; from undergraduate student level to postdoctoral level. Five human subjects participated in the study. For determining reasoning types, the subjects were first explained reasoning types and given brief training on identifying those. A screenshot is shown in Figure 5. Subjects were then shown the question-answer pairs (unseen during the briefing) and asked to choose the reasoning types. Screenshots are shown in Figure 6, Figure 7. In the interface, all the reasoning types were listed out, and the subjects had the freedom to select multiple reasoning types if they thought an instance contained more than one type of reasoning. Additionally, a "No Reasoning Type" option was available to subjects in case they deemed that no reasoning was involved.

A.13 Extended analysis of models' performance

In the main paper, we discussed two of the major limitations of VideoQA models. Other limitations include: 1) some models like MIST do not leverage explicit motion information using, *e.g.*, spatiotemporal convolutional neural networks. Due to this, they might inaccurately infer the scene based on a single static frame, instead of motion containing video clips. We believe these models can further improve their performance by incorporating explicit motion information. 2) Our CausalChaos dataset introduces the challenge of reasoning by linking scenes/shots. This is different from traditional temporal modeling, which is done within a scene. Scenes/shots involve an abrupt change in the scene. Traditional temporal modeling typically is geared toward smoother transitions; abrupt changes violate this condition. So while traditional temporal modeling does aid on our dataset, abrupt scene changes in our dataset poses a further challenge for VideoQA models which is not adequately addressed by traditional temporal models like 3DCNN feature extractors. We evaluated BLIP-2, VideoLLaMA and



Consider the video, question, answer level 1 and answer level 2.

Question: Why did Tom scream?

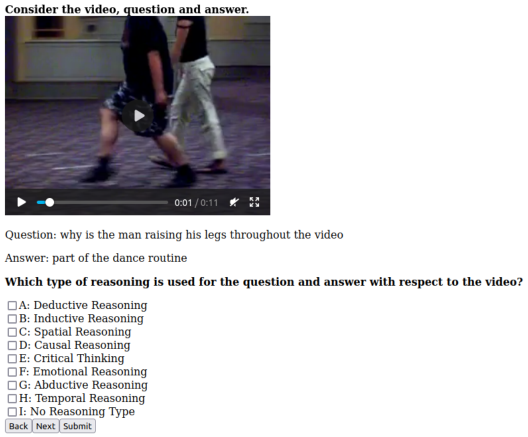
Answer Level 1: Tom saw Topsy sleeping.

Answer Level 2: Tom was frustrated that Topsy was sleeping while Tom was trying to teach Topsy an important lesson.

Which type of reasoning is used for the question and answer with respect to the video?

□ A: Deductive Reasoning
□ B: Inductive Reasoning
□ C: Spatial Reasoning
□ D: Causal Reasoning
□ E: Critical Thinking
□ F: Emotional Reasoning
□ G: Abductive Reasoning
□ H: Temporal Reasoning
□ I: No Reasoning Type

Figure 6: **Screenshot of user interface used for collecting responses from human subjects for CausalChaos! dataset.**



Consider the video, question and answer.

Question: why is the man raising his legs throughout the video

Answer: part of the dance routine

Which type of reasoning is used for the question and answer with respect to the video?

□ A: Deductive Reasoning
□ B: Inductive Reasoning
□ C: Spatial Reasoning
□ D: Causal Reasoning
□ E: Critical Thinking
□ F: Emotional Reasoning
□ G: Abductive Reasoning
□ H: Temporal Reasoning
□ I: No Reasoning Type

Figure 7: **Screenshot of user interface used for collecting responses from human subjects for NextQA dataset.**

VideoChat2 on MCQA and Open-Ended Answer Generation (OEAG) task. We noticed that BLIP-2 model performed better on MCQA, while VideoChat2 and VideoLLaMA performed better on OEAG task. Notably, OEAG requires better language modeling than that required in MCQA task.

Qualitative results. We have provided the qualitative analysis of models' failure cases in the accompanying PowerPoint presentation for the following tasks/cases:

1. **Multi-Choice Question Answering (MCQA)**
2. **Open-Ended Answer Generation (OEAG)**
3. **Incorporating our data into real-world Video QA**

Full results on OEAG are presented in Table 1.

A.14 Discussion on cartoon physics

Cartoon physics often operates within its own set of rules and logic, which may differ from real-world physics but still maintain consistency within the cartoon's universe. These rules might include exaggerated movements, gravity-defying actions, and other fantastical elements that wouldn't occur in reality but are accepted within the context of the cartoon world. Despite the departure from real-world physics, there is often an internal consistency to how these cartoon physics operate within their respective universes.

Model	BLEU-1		BLEU-2		BLEU-3		METEOR		ROUGE		SPICE		CIDEr		S-BERT		CapsMIX
	A	E	A	E	A	E	A	E	A	E	A	E	A	E	A	E	
BlindQA [1]	0.2412	0.1912	0.1039	0.0618	0.0415	0.0221	0.1243	0.0891	0.2732	0.1885	0.2956	0.0703	0.1034	0.0403	0.4987	0.4195	2.7646
UATT [21]	0.2661	0.2203	0.1257	0.0805	0.0575	0.0330	0.1409	0.1046	0.3171	0.2293	0.3547	0.2775	0.1197	0.0397	0.5495	0.4767	3.3928
HME [2]	0.2650	0.2072	0.1301	0.0713	0.0554	0.0219	0.1504	0.0914	0.3125	0.2273	0.3184	0.0826	0.0936	0.0639	0.5328	0.4737	3.0975
HGA [9]	0.2891	0.2263	0.1588	0.0897	0.0789	0.0323	0.1856	0.1130	0.3388	0.2399	0.3980	0.2426	0.2782	0.0573	0.6026	0.4561	3.7872
BlindGPT-2	0.3770	0.3165	0.2530	0.1749	0.1632	0.1091	0.2452	0.1761	0.3858	0.2966	0.4305	0.3529	1.2482	0.7690	0.6755	0.6271	6.6006
VisionGPT-2	0.3878	0.3095	0.2605	0.1727	0.1738	0.1091	0.2560	0.1756	0.3934	0.2941	0.4498	0.3385	1.3725	0.7539	0.6760	0.6350	6.7582
BLIP-2 [12]	0.1381	0.0815	0.0451	0.0256	0.0167	0.0059	0.0664	0.0480	0.1618	0.1312	0.0530	0.0422	0.2279	0.1046	0.3837	0.3614	1.8931
Video-LLaMA [25]	0.1241	0.1181	0.0419	0.0344	0.0115	0.0098	0.1477	0.1163	0.1719	0.1435	0.2055	0.1383	0.0836	0.0430	0.5734	0.4834	2.4464
VideoChat2 [14]	0.2353	0.2116	0.0823	0.0776	0.0250	0.0253	0.1769	0.1295	0.2667	0.2168	0.3264	0.2547	0.3980	0.2910	0.6445	0.5908	**3.9524**

(a)

Model	BLEU-1		BLEU-2		BLEU-3		METEOR		ROUGE		SPICE		CIDEr		S-BERT		CapsMIX
	A	E	A	E	A	E	A	E	A	E	A	E	A	E	A	E	
BlindQA [1]	0.2193	0.1795	0.0772	0.0468	0.0326	0.0149	0.1122	0.0813	0.2501	0.1767	0.2960	0.1147	0.1099	0.0434	0.5129	0.4260	2.6935
UATT [21]	0.2693	0.1947	0.1257	0.0581	0.0440	0.0146	0.1466	0.0847	0.3226	0.2107	0.3255	0.2012	0.1444	0.0646	0.5364	0.4829	**3.2260**
HME [2]	0.2475	0.1830	0.1031	0.0549	0.0363	0.0154	0.1417	0.0845	0.2820	0.2173	0.2933	0.2549	0.0831	0.0655	0.5165	0.4929	3.0719
HGA [9]	0.2586	0.1842	0.1085	0.0517	0.0365	0.0148	0.1433	0.0932	0.2909	0.1806	0.2908	0.2128	0.0877	0.0266	0.5231	0.4078	2.9111

(b)

Model	BLEU-1		BLEU-2		BLEU-3		METEOR		ROUGE		SPICE		CIDEr		S-BERT		CapsMIX
	A	E	A	E	A	E	A	E	A	E	A	E	A	E	A	E	
BlindQA [1]	0.2312	0.1847	0.0869	0.0550	0.0319	0.0198	0.1263	0.0930	0.2519	0.1806	0.2758	0.2447	0.0667	0.0265	0.4844	0.4153	2.7747
UATT [21]	0.2659	0.1936	0.1180	0.0645	0.0548	0.0223	0.1423	0.0868	0.3025	0.2200	0.3449	0.2595	0.0940	0.0713	0.5447	0.4750	3.2601
HME [2]	0.2640	0.1930	0.1238	0.0644	0.0556	0.0239	0.1457	0.0868	0.2998	0.2155	0.3115	0.1623	0.0944	0.0698	0.5341	0.4761	3.1207
HGA [9]	0.2674	0.2293	0.1337	0.0823	0.0620	0.0308	0.1703	0.1123	0.3186	0.2363	0.3734	0.2558	0.1972	0.0477	0.5765	0.4692	**3.5628**

(c)

Table 1: **OEAG Results** on our dataset. (a) **UD** split; (b) **PS** split; (c) **UN** split.

Despite the departure from real-world physics, humans/video understanding models can apply causal reasoning within the context of cartoon physics to predict the consequences of characters' actions. For example, if a character steps off a cliff, humans expect them to fall downwards due to gravity, even if the fall is exaggerated or prolonged for comedic effect. This consistency can allow video-understanding models to anticipate and understand the outcomes of actions within the cartoon world, facilitating their ability to follow the storyline and engage with the humor and narrative.

Furthermore, the consistency of cartoon physics enables humans to make logical connections between different events and understand the progression of the story. By recognizing patterns and understanding how actions lead to specific outcomes, video understanding models can engage in causal reasoning to predict future events and comprehend the logic of the cartoon universe.

A.15 CausalConfusion incorrect/negative answer generation

Samples from the dataset created using Vanilla Hard Negative mining:

> Q: Why did Tom dip his fingers in the ink?

Correct A: To draw a mouse hole on the wall.
Incorrect A(1): Tom wanted Jerry to mistake Tom's finger for a sausage.
Incorrect A(2): Tom was preparing to eat.
Incorrect A(3): Tom wanted to see if there was ink in the pen.
Incorrect A(4): Tom's hand was in pain.
Correct E: Tom was trying to trick Jerry by drawing a fake mouse hole on the wall.
Incorrect E(1): Tom's hand was in pain from hitting Jerry with the vase.
Incorrect E(2): Tom was excited to eat Jerry who was on Tom's plate.
Incorrect E(3): Tom thought there was no ink in the pen as the ink did not come out when Jerry pulled the pen.
Incorrect E(4): Tom wanted to trick Jerry to mistake Tom's finger for a sausage so that Tom could catch Jerry when Jerry tried to steal Tom's finger.

Q: Why did Tom climb onto the gate?
Correct A: The bull was charging towards Tom.
Incorrect A(1): Tom was trying to get away from Spike.
Incorrect A(2): Tom wanted to get to a higher point on the tree.
Incorrect A(3): because Tom heard barking sounds and was scared.
Incorrect A(4): Tom was trying to get away from Spike and Tyke.
Correct E: The bull was charging towards Tom so Tom climbed onto the gate to avoid getting hurt by the bull.
Incorrect E(1): because Jerry imitated Spike to bark at Tom to scare Tom into climbing up the tree.
Incorrect E(2): Tom was scared of Spike who was chasing Tom and climbed up the tree to get away from Spike.
Incorrect E(3): Tom was dressed as a bird and wanted to climb higher on a tree to take off.
Incorrect E(4): Tom saw Spike and saw Tyke barking and wanted to get away from them.

Examples of Vanilla Hard Negatives vs. *CausalConfusion* Negatives:

Q: Why did Tom dip his fingers in the ink?
Correct A: To draw a mouse hole on the wall.
Vanilla Hard Negatives
Incorrect A(1): Tom wanted Jerry to mistake Tom's finger for a sausage.
Incorrect A(2): Tom was preparing to eat.
Incorrect A(3): Tom wanted to see if there was ink in the pen.
Incorrect A(4): Tom's hand was in pain.
CausalConfusion version
Incorrect A(1): To not draw a mouse hole on the wall.
Incorrect A(2): Tom was preparing to eat.
Incorrect A(3): Tom wanted to see if there was ink in the pen.
Incorrect A(4): Tom's hand was in pain.

Q: Why did Tom climb onto the gate?
Correct A: The bull was charging towards Tom.
Vanilla Hard Negatives
Incorrect A(1): Tom was trying to get away from Spike.
Incorrect A(2): Tom wanted to get to a higher point on the tree.
Incorrect A(3): because Tom heard barking sounds and was scared.
Incorrect A(4): Tom was trying to get away from Spike and Tyke.
CausalConfusion version
Incorrect A(1): The bull was not charging towards Tom.
Incorrect A(2): Tom was charging towards the bull.
Incorrect A(3): because Tom heard barking sounds and was scared.
Incorrect A(4): Tom was trying to get away from Spike and Tyke.

A.16 CapsMIX extended details

However, we note that with such a wide range of metrics, it is difficult to get a comprehensive insight into models' performances & compare them. To address that, we introduce a comprehensive metric, termed Caps-MIX (Captioning Metrics Integration eXpert), which integrates all the previously mentioned scores after normalizing them to their theoretical best values. This **1)** makes it easier to compare models using a single number and **2)** combines the characteristics of individual metrics, each measuring performance from a unique perspective. We avoid using the WUPS score [16], as it is designed for single-word answers and is not suitable for our dataset's detailed responses.

A.17 Negative societal impact

While our dataset has a positive attribute of being synthetic in nature. And as such, we do not suggest deploying models trained on our dataset in real-world applications. Causal reasoning models trained on real-world data can potentially be used to find out or estimate why people carried out actions. This, in-turn, can be used to deduce further actionable insights into people's behavior. This might justifiably be seen as an intrusion of privacy, especially, without consent. Thus, such systems shall not be deployed/used without the consent of all the parties involved. We suggest that this space should be regularized by governing bodies, and consent from the end-users, and parties being monitored is inevitable.

A.18 Compute details

We used machine with following specifications: Intel(R) Xeon(R) W-2245 CPU@3.90GHz; 64GB RAM; 2x Nvidia A5000 24GB.

A.19 Link to Dataset

We have included the following dataset files in the supplementary.

1. File containing all the annotations
2. Vanilla hard negative sets
3. CausalConfusion set

The dataset files are also publicly available at: `https://github.com/LUNAProject22/CausalChaos`.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 3, 11

[2] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 3, 11

[3] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14773–14783, 2023. 3

[4] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585, 2018. 3

[5] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 2

[6] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2712–2719, 2013. 2

[7] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 2, 3

[8] Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Video question answering with spatio-temporal reasoning. *International Journal of Computer Vision*, 127: 1385–1412, 2019. 2

[9] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11109–11116, 2020. 3, 11

[10] 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. 3

[11] 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 (CVPR)*, 2022. 2, 7

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 11

[13] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11963–11974, 2023. 2, 7

[14] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. 2024. 11

[15] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893, 2017. 2

[16] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27, 2014. 13

[17] Jiayuan Mao, Xuelin Yang, Xikun Zhang, Noah Goodman, and Jiajun Wu. Clevrer-humans: Describing physical and causal events the human way. *Advances in Neural Information Processing Systems*, 35: 7755–7768, 2022. 2

[18] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. [Online; accessed 31-May-2024]. 7

[19] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 2, 7

[20] 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 *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 2

[21] Hongyang Xue, Zhou Zhao, and Deng Cai. Unifying the video and question attentions for open-ended video question answering. *IEEE Transactions on Image Processing*, 26(12):5656–5666, 2017. 11

[22] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 2

[23] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 2

[24] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019. 2

[25] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore, 2023. Association for Computational Linguistics. 11

[26] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124:409–421, 2017. 2