

**CausalChaos! Comprehensive Causal Action  
QA Over Longer Causal Chains Grounded in  
Dynamic Visual Scenes---*Supplementary***

NeurIPS 2024 Paper ID: 241

# Contents

- Dataset examples
- More Dataset Stats
- MCQA – Qualitative results
- MCQA – CausalConfusion set
- Open-ended answer generation
- CausalChaos! Bringing improvements on real-world dataset (NextQA)-  
Qualitative results

# CausalChaos! Dataset samples



**Q: Why did Jerry put Tom's tail in the sandwich?**

**A: Jerry wanted Tom to bite his own tail.**

**E: Jerry wanted to inflict pain on Tom.**



**Q: Why did Jerry whistle?**

**A: Jerry was whistling for Spike.**

**E: Jerry was calling for Spike to save Jerry from Tom who was about to hit Jerry.**



**Q: Why was Jerry laughing on the ground?**

**A: Jerry was amused.**

**E: Jerry was amused at Tom whose pants had dropped.**



**Q: Why did Tom put a spoon of hot soup in front of Jerry?**

**A: Tom wanted Jerry to blow at the hot soup.**

**E: Tom thought the soup was too hot and wanted Jerry to help to cool the soup down.**

# More Dataset Stats

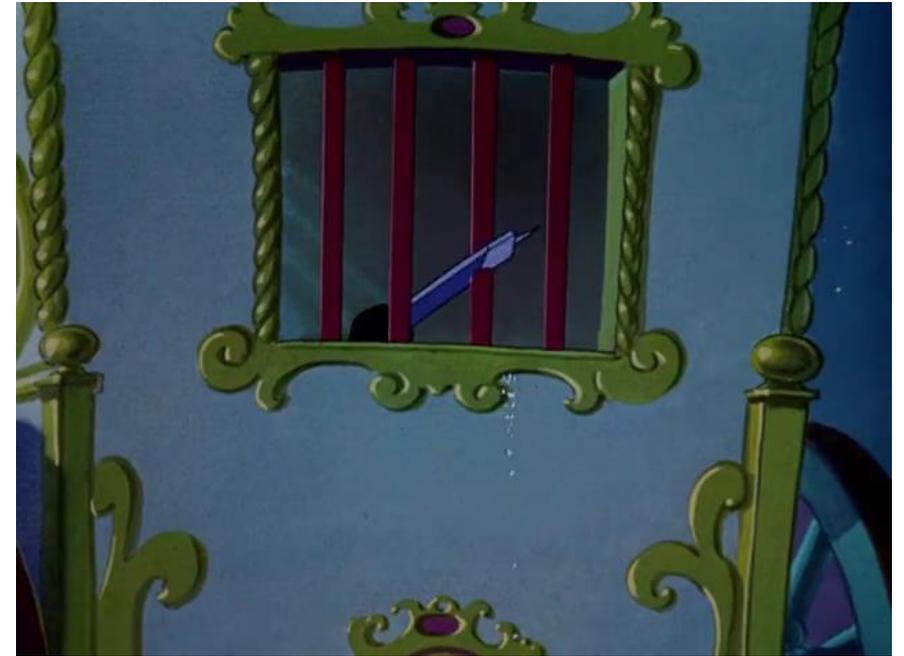
- Average clip frames : 357.95
- Longest of clips(frames) : 2315.0
- Average length of question : 6.54
- Longest length question : 17
- Longest question : why did jerry and tuffy put the wire into the water and turn on the freeze mode
- Average length of Answer : 7.76
- Longest length Answer : 26
- Longest 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
- Longest length Explanation: 30
- 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

# MCQA – Qualitative results

Question: why did the seal cut the rod?

**Answer GT:** because seal wanted to make a gap that is big enough for the seal to jump out

HGA Pred: because the seal wanted to swing with the wire off the roof



HGA and HME which incorporate motion information explicitly got it wrong because they took shortcut like matching motion in action snippet and verb.



Question: why did the seal cut the rod?

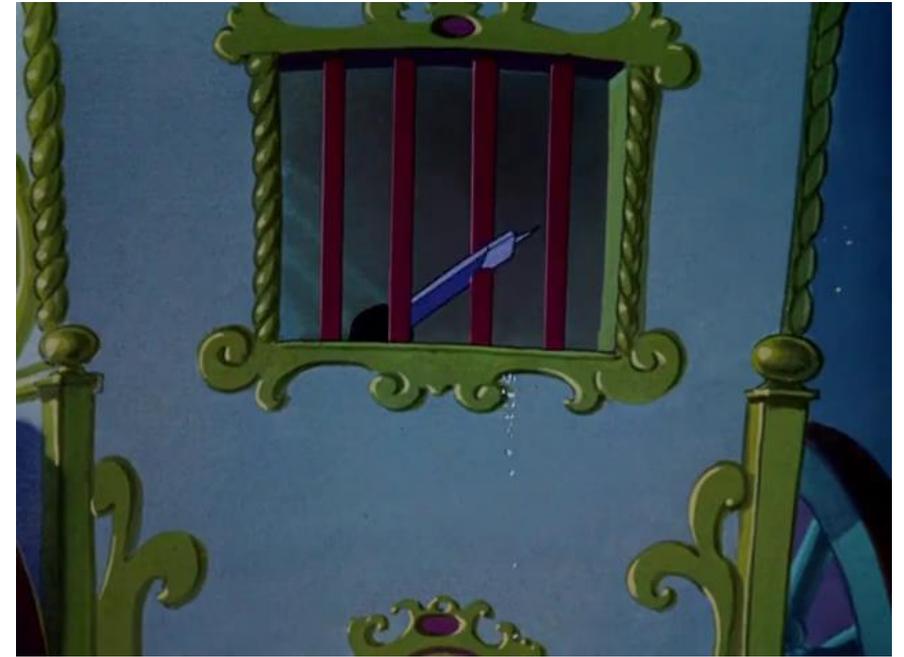
**Answer GT:** because seal wanted to make a gap that is big enough for the seal to jump out

MIST Pred: The seal wanted to get to the roof.

**Explanation GT:** because the seal wanted to jump out of the carriage to escape from the circus

MIST Pred: because the seal wanted to run away from tom

MIST predicted wrong answer because it probably gave too much importance to matching between pose in action snippet and the phrase/word in language space. MIST also potentially only considered partial evidence and based on that selected the answer.



Question: Why did the lion catch Jerry ?

**Answer:**The lion wanted to hush Jerry.

**MIST:**The lion was hungry.

**Explanation:**The lion was afraid that Jerry's scream would attract the attention of the ones who were searching for the lion that the lion wanted Jerry to stay quiet.

**MIST:**The lion wanted Jerry to help to get some food to the lion as the lion was hungry.



MIST predicted wrongly because it might gave visual information less consideration, and probably frequency-based bias.

Question: Why was Butch running ?

Answer: Butch wanted to slide out of the house on the bassinet.

MIST: Jerry escaped from Butch

HGA: Butch wanted to slide out of the house on the bassinet.

MIST wrongly predicted because it potentially selected answer where more nouns in language and video spaces matched (Jerry + Butch).

HGA got it correct because the explicit motion information it leverages helped it interpret the scene correctly.

Explicit motion information and denser sampling can help improve performance of models like MIST.



Question: Why did Tom slam the **keys** ?

Answer: Tom wanted to catch Jerry.

MIST: Tom wanted to make sure Jerry and Nibbles could not get the **key**.

HGA: Tom wanted to make sure Jerry and Nibbles could not get the **key**.

HCRN: Tom wanted to make sure that the room stays locked.

HME: Tom wanted to lock the door and make sure that no one can open the **door**.

CoMem: Tom wanted to lock the door and make sure that no one can open the **door**.

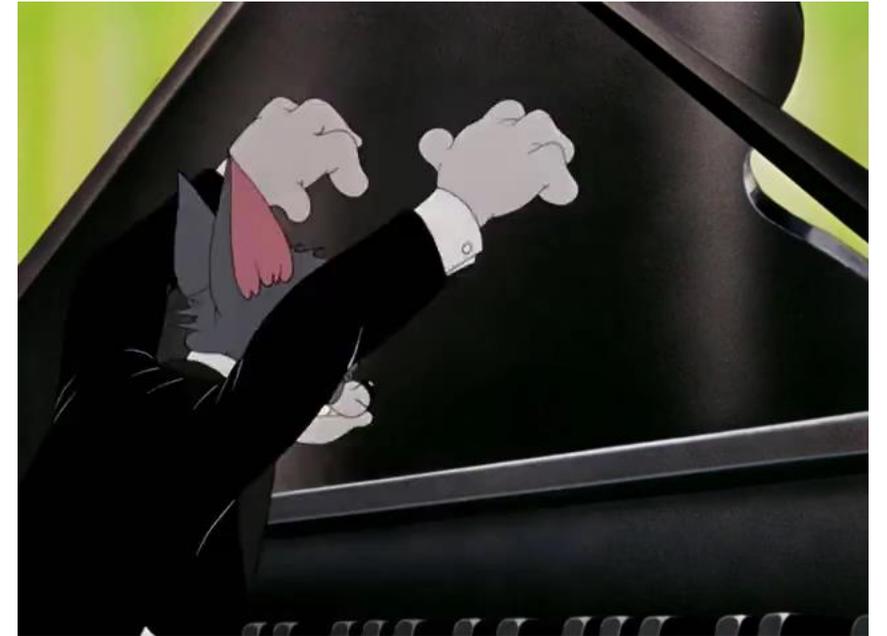
EVQA: Tom wanted to make sure that the room stays locked.

Some models (MIST, HGA) might be selecting an answer based on matching words in

Answer and question, rather than understanding video, Q, A and

Modeling causal relationship.

Other models (HME, CoMem) probably selected answer based on the closest distance in the embedding space (Key and Door); but here key referred to piano keys, not key to a door lock



Question: Why did Jerry put Tom's tail in the sandwich ?  
Correct Answer: Jerry wanted Tom to bite his own tail.  
HGA: Jerry wanted to stop himself falling.

Explicit motion information leveraging model, HGA, misinterpreted Jerry bending over as him falling. Based on this partial, incorrect interpretation, it prematurely jumped to wrong conclusion .



# Qualitative results –MCQA Causal Confusion

- Where CausalConfusion can confuse MIST
- Q: Why did Tom hold his head ?
- Correct Answer: Tom's head was in pain.
- MIST original prediction: Tom's head was in pain.
- MIST in the presence of CausalConfusion negative: Tom's head was **not** in pain.

# Qualitative results - MCQA

- Where CausalConfusion can confuse MIST
- Q: Why did Jerry bandage up the baby duck's wing ?
- Correct Answer: The baby duck had told Jerry the story of how it got hurt.
- MIST original prediction: The baby duck had told Jerry the story of how it got hurt.
- MIST in the presence of CausalConfusion negative: Jerry had told the baby duck the story of how it got hurt. (Actors swapped)

# Open-Ended Answer Generation- Qualitative results

# VisualGPT-2 Successes



- Question: Why did Jerry bandage up the baby duck's wing?
- Answer: Jerry was showing concern for the baby duck and wanted to heal the hurt wing.
- Pred: Jerry was showing concern for the baby duck and wanted to bandage the baby duck

- Question: Why did Jerry blow a kiss?
- Answer: Jerry was in love with Lilli.
- Pred: Jerry was in love with the female mouse.

- Question: Why did Jerry push a rollerblade?
- Answer: Jerry wanted Tom to trip on the rollerblade.
- Pred: Jerry wanted to use the rollerblade to block Tom

# VisualGPT-2 Failures



- Question: Why did the lion catch Jerry ?
- Answer: The lion wanted to hush Jerry.
- Pred: The lion wanted to eat Jerry.

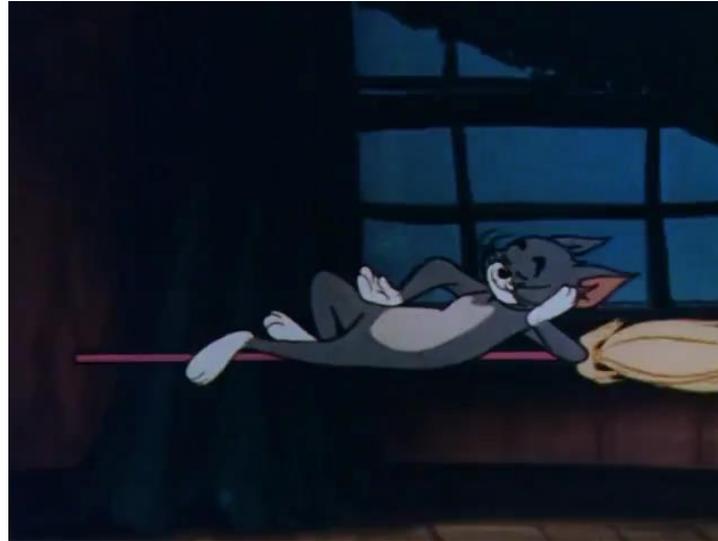


- Question: Why did Tom grab an axe?
- Answer: Tom wanted to get inside the house and catch the dog and Jerry.
- Pred: Tom wanted to use the axe to kill and get rid of Jerry.

# VideoChat2 Successes



- Question: Why did the goldfish wake up?
- Answer: Jerry knocked on the fishbowl to wake up the goldfish.
- Pred: The goldfish woke up due to the sound of the mouse.



- Question: Why did Tom slam the seat on Jerry?
- Answer: Tom was annoyed that Jerry was laughing at Tom when Tom got hurt on screen.
- Pred: Tom was upset and frustrated with Jerry, so he slammed the seat on him.

# VideoChat2 failures



- Question: Why did Quacker sit in the basket?
- Answer: Quacker wanted to float up in the air together with the air balloon.
- Pred: Quacker was sitting in the basket because he was being inflated.



- Question: Why was Tom struggling?
- Answer: Tom was stuck in a sack.
- Pred: Tom was struggling due to the heavy rain.

# Qualitative Results

- Open-ended answer generation
- Cases BlindGPT-2 generates wrong answer, but VisionGPT-2 is able to generate correct answers



- Q: why did the witch touch her head?
- GT A: The witch realised that her head was empty.
- BlindGPT-2: the witch wanted to lift up her head to
- VisionGPT-2: the witch realised her hat was missing

# Qualitative Results

- Open-ended answer generation
- Cases BlindGPT-2 generates wrong answer, but VisionGPT-2 is able to generate correct answers
- Q: Why did Jerry run into the metal can?
- GT A: Jerry did not see the metal can in front of him.
- BlindGPT-2: Jerry wanted to get out of the metal can and get out
- VisionGPT-2: Jerry did not see the metal can in front of Jerry



# Qualitative Results

- Open-ended answer generation
- We observed that sometimes both fail, but VisionGPT-2 gets the context right, for example, as shown below.
- Q: Why did Muscles blow the barrel of the gun?
- GT A: Muscles wanted to blow the bullets out through the other end.
- BlindGPT-2: Muscles wanted to let the the the wine to flow out
- VisionGPT-2: Muscles wanted to get rid of the barrel of the gun



# CausalChaos! Bringing improvements on real-world dataset (NextQA)-Qualitative results

*Red color: wrong prediction by model Green color: correct prediction (same as groundtruth)*



why does the man pull the baby sitting on the sled by the rope ?  
Just NextQA: **for boy to pull sled**  
+ CausalChaos: **help baby sledge**

*Enhances focus on finding better cause & effect*



**what** did the boy do with the toy scooper ?  
Just NextQA: **collect crabs**  
+ CausalChaos: **scoop water**

*Enhances visual grounding*



why did the adult squat down and opened his arm at the end of the video?  
Just NextQA: **stretching his arms**  
+ CausalChaos: **want to hug child**

*Encourages comprehensive analysis of the situation, rather than jumping to premature conclusion*

# Domain Gap testing

- We ask several models to describe the videos off-the-shelf, without any finetuning
- The idea is that if the models are able to describe the videos, then they are able to perceive the scene
- This would also indicate that the domain gap is potentially not the reason for the poor performance of models on our dataset, but it is more a problem of causal reasoning

# Domain Gap testing: Sample 1



**GPT-4o:** The frames from the video depict a sequence from a classic animated cartoon featuring a cat and a mouse. Here is a summary of the action:

1. The scene starts with a view of a fence and a garden.
2. The mouse is seen running and jumping with a joyful expression.
3. The cat appears, chasing the mouse with a playful and determined look.
4. The cat continues to chase the mouse, leaping and bounding energetically.
5. The cat eventually stops, looking surprised and somewhat dazed.
6. The mouse is then seen interacting with another mouse, who has a bow on her head, suggesting a friendly or romantic encounter.

The sequence captures the typical playful chase and humorous interactions between the cat and the mouse, characteristic of many classic animated cartoons.

**VideoChat2:** The video shows a cartoon of a cat chasing a mouse. The cat is shown running and jumping to catch the mouse.

**VideoLLAMA:** First, we see a cartoon cat in a red dress running on a green grassy lawn. Next, the camera pans over to a different view of the same cat running on the grass. Then, the camera zooms in on the cat as it runs. We see a bird flying over the same lawn as well. Finally, the camera pans over to a close-up of the cat jumping through a small hole in the fence. Overall, the video shows a cat running around on a green lawn, and there is a bird flying over it. We also see a small hole in the fence that the cat is jumping through.

# Domain Gap testing: Sample 2

**GPT-4o:** The frames depict a sequence from a classic animated cartoon featuring a cat and a mouse. Here's a summary of the action:

1. The cat is holding its head in frustration while standing near a small blue object that is emitting sparks.
2. The mouse emerges from the blue object, which appears to be some kind of firework or explosive device.
3. The mouse lifts the blue object and runs towards a hole in the wall.
4. The mouse throws the blue object into the hole, and it continues to emit sparks.
5. The cat looks on in surprise and concern as the mouse completes its action.
6. The cat jumps in shock, likely due to the noise or the unexpected outcome.
7. The blue object continues to emit sparks and rolls towards a doghouse.
8. The object enters the doghouse, and the scene ends with the cat running away in fear.

The sequence is typical of the comedic and chaotic interactions between the cat and mouse in this animated series.

**VideoChat2:** The video shows a cat and a mouse playing together in a garden. The cat is seen chasing the mouse around the garden while the mouse tries to escape. The cat eventually catches the mouse and starts playing with it.

**VideoLLAMA:** First, there is a cartoon cat playing with a tennis ball in a grassy area. Next, the cat is playing with a red tennis ball in the same area. Then, the cat is in a garden with a tennis ball. The cat is also in a garden with a red tennis ball in front of a red tennis ball in the grass. Then, there is a cartoon cat in a garden with a red tennis ball on the ground. The cat is also in a garden with a tennis ball on the grass. Finally, the cat is in a garden with a tennis ball in front of a red tennis ball on the ground. Throughout the video, there are different colored tennis balls being played with and in different locations. The cat is also seen playing with a tennis ball on the grass in different frames. There are red chairs on the porch and a large bush in the background of the garden. The video is described with high-quality descriptions.



# Domain Gap testing: Sample 3

**GPT-4o:** The frames depict a sequence from an animated cartoon featuring a cat and a mouse. Here's a summary of the action:

1. The mouse, painted blue, is seen running away from a door as a blue lion peeks out.
2. The cat, wearing a green and blue hat, chases the blue mouse.
3. The mouse escapes through the door, and the cat follows closely behind.
4. The cat reaches the door and starts to close it.
5. The cat, now standing by the closed door, appears to be satisfied and adjusts its hat.
6. The cat continues to stand by the door, looking smug and confident.
7. The cat starts to laugh, seemingly pleased with itself.
8. The cat strikes a triumphant pose, still laughing.
9. The cat begins to open the door again, still looking confident.
10. The cat peeks through the door, still smiling.
11. The cat's hand is seen closing the door.
12. The cat's hand is still on the door handle, closing it further.
13. The cat's hand is now holding a key.
14. The cat's hand inserts the key into the lock.
15. The cat's hand turns the key to lock the door.
16. The cat's hand is seen removing the key from the lock.

The sequence shows the cat chasing the mouse, the mouse escaping, and the cat locking the door, presumably to prevent the mouse from returning.

**VideoChat2:** The video features a cartoon cat who is dancing in a room. The cat is wearing a hat and appears to be having a good time.

**VideoLLAMA:** In the beginning of the video, there is a cartoon cat standing at a door with its mouth open. The cat is wearing a blue hat and looks like it is waiting for something. Next, the cartoon cat appears again at the door, and it seems to be in a different location. There is a blue and white wall and a pink picture in the background, as well as a blue chair and a white and brown chair. The cartoon cat is still wearing its blue hat. In the next scene, the cartoon cat is at a door in a room with a blue and white wall, and the cat seems to be looking out at something. The cat is wearing a brown hat and has a white shirt on. In the following scene, there is a cartoon cat at a door, and the cat seems to be looking into the room. There is a blue and white wall, a pink picture frame, and a blue and white blanket. The cat is wearing a brown hat and has a white shirt on. Then, there is another cartoon cat at a door, and the cat seems to be looking into the room. There is a blue and white wall, a pink picture frame, and a blue and white blanket. The cat is wearing a brown hat and has a white shirt on. Finally, the video ends with a cartoon cat standing at a door, and the cat seems to be looking out



# Domain Gap testing: Sample 4



**GPT-4o:** The frames from the video depict a sequence involving a small blue mouse and a menacing black cat in a dark alley. Here's a summary of the events:

1. The blue mouse is seen walking in a dark alley with a figure lying on the ground.
2. The mouse notices multiple pairs of glowing eyes appearing in the darkness, indicating the presence of many cats.
3. The mouse becomes alarmed and starts to run away.
4. The mouse hides behind a trash can, but a large black cat emerges from the can, looking threatening.
5. The cat tries to catch the mouse, but the mouse manages to escape.
6. The mouse runs out of the alley and onto a city street, continuing to flee.
7. The mouse hides behind another set of trash cans, trying to evade the pursuing cat.

The sequence portrays a classic chase scene with a small, clever mouse trying to escape from a larger, predatory cat in a dark, urban setting.

**VideoChat2:** The video shows a cartoon of a cat and a bear. The cat is seen running and the bear is chasing it. The scene is set in a dark alley with a trash can in the background.

**VideoLLAMA:** First, there is a cat in the street, and there is a dark street. Next, the cat is looking around, and there is a dog in the street. Then, there is a cat standing in a garbage dumpster in the night. After that, there is a cat in a dark street, and the cat is in a dumpster at night. The cat is in a street again, and there is a cat in a dark street. Then, there is a cat in a dumpster at night, and the cat is in a dumpster. Finally, there is a cat in a dumpster at night, and the cat is standing in a dark street. Throughout the video, there are various objects and animals present, including a dog, a cat, a street, a dark street, a dumpster, a garbage dumpster, a car, a pile of trash, a street, and a dark alleyway.

# Domain Gap testing: Sample 5



**GPT-4o:** The frames from the video depict a sequence of events in a cartoon setting, likely involving a comedic storyline. Here's a summary based on the frames:

1. A character, possibly a cat dressed as a cowboy, is seen near a building labeled "Mutt Dillin Sheriff." The character is interacting with a barrel marked with "XXX."  
2. The character appears to be struggling with the barrel, which eventually explodes, causing significant damage to the building.  
3. The character is left charred and disheveled, standing amidst the wreckage of the sheriff's office.  
4. The character examines a tattered piece of paper with the text "HAVE TRAP WILL TRAP."  
5. Another character, possibly a dog, approaches and hands over a piece of paper.  
6. The cat character reacts angrily, pointing a gun at the dog character.  
7. The scene shifts to a "Dry Gulp General Store," with a sign in the window advertising "SWISS" cheese.  
8. The final frame shows the store with some smoke or dust, indicating some form of disturbance or action.

The sequence suggests a humorous and chaotic interaction between the characters, typical of classic animated cartoons.

**VideoChat2:** The video features a cartoon character who is a cowboy. The cowboy is seen walking around in the desert, and he is accompanied by a dog. The video also shows a store that sells dry goods.

**VideoLLAMA:** The video starts with an animated cartoon character wearing a cowboy hat and holding a gun. The character is standing in front of a house that is in the middle of a river. The house is made of wood and there is a man standing on the roof. In the following seconds, the animated character is seen walking on the river and splashing water. The character then comes to a small wooden boat and jumps onto it, shooting the water below. The next scene shows the character in the water, shooting a bottle of water and throwing it into the air. In the following seconds, the character is seen riding a horse in the water and shooting a black bird. The character is then seen riding a horse and a black dog in the water, splashing water and riding on the water. Finally, the video shows the character on a small boat in the water, holding a white and black dog and riding on the water.