Examples

task_types	question	answer
Temporal Understanding; Sequential; Counting	How many "teaspoons of sugar" does Ferny add into the small metal pitcher, after placing the moka pot onto the stove?	After placing the moka pot onto the stove, Ferny adds "6 teaspoons of sugar," into the small metal pitcher
Temporal Understanding; Sequential; Needle	What prompts the announcer to get excited and say "look out, some right-handed thunder from Grant Nelson"?	Grant Nelson breaks away from the defense and dunks a ball with one hand.
Subscene; Needle	What was the speaker referring to when he said, "we found these down here last night"?	He was referring to a jar of gummy bears and a jar of malted milk balls.
Tacking Spurious Correlations	In a scene in the input clip, a group of people is having a meeting with a table in the middle, with a man saying \"If you can't bring down the charging bull, then don't wave the red cape at it.\" Who are they referring to?	They are referring to Superman as he is shown holographically in the middle of the table.
General Holistic Reasoning; Inference	Based on the audio and visual components in this "Inception" trailer, why is Dom Cobb feeling "guilt"?	Based on the audio and visual components in this "Inception" trailer, Dom Cobb is feeling "guilt" because his wife Mal began to question her own reality, an idea that came from him and one he "planted in her mind," which ultimately led to her demise.
General Holistic Reasoning; Inference	What is Casey's relationship to Candice?	The two are married and have a child.

Inference	Why does the young woman yell "Oh, Oh, Oh," after lifting up the barbecue grill's hood handle?	The young woman yells "Oh Oh Oh," after lifting up the barbecue grill's hood handle because it falls down on the ground.
General Holistic Reasoning; Inference	How did Jose Ramirez explain his goal as a singer during his interview, and what facial expressions or gestures accompanied his response?	He said he wanted to sing blues and soulful roots of the bluesy genre. He was making eye contact with the camera with little hand gestures.
Temporal Understanding; Sequential; Inference	After Michele opens up the second door using her key, what is the long beeping sound that instantly goes off?	After Michele opens up the second door using her key, the long beeping sound that instantly goes off is the security alarm which she turns off.
Temporal Understanding; Sequential	What is the order of these events? (A) Dolphins are filmed under the water. (B) The woman explains why she is not doing the turtle release event. (C) The text "El Sultan" appears on screen with a fruit bowl. (D) The fish tacos appear on screen.	(D)(A)(C)(B)
Temporal Understanding; Sequential; Inference	What is the man doing after he says, "and as always, stay tuned," and why?	The man is shown playing a melody on his guitar. Since this takes place at the end of the video, it is used as an outro for the video.
Sequential; Object Interaction Reasoning	After the man at the beginning of the video says "Wha'd you do", how does the purple character disappear?	The purple man falls backwards through a blue portal, leaving an axe where he once stood.

Below are more examples of questions for different task types. In the section Sub Task types, the finegrained subcategories are mentioned, and each of the task type can have questions from these subcategories.

Task Types

1. **Temporal Understanding**
    a. **General Examples**
        i. What happens before the woman in the blonde hair says to Captain America, \"This is going to work Steve\"?
        ii. What happens after Casey introduces Ollie?
        iii. What does the girl with the pink goggles say to her friend wearing a one-piece bathing suit, right before she's about to jump into the swimming pool?

2. **Sequential**
    a. **General Examples**
        i. What is the order of events in this video? \nA) Both girls are doing the splits on the beach\nB) One of the girls says, \"a bee was chasing us.\" \nC) One of the girls says, \"we're in a hot tub guys.\"\nD) Both girls are seen munching on snacks at the beach
        ii. What does the man in the video do after saying \"I really really really really wanted to, I really wanted to see the sunrise before I left
        iii. Which happens first in the video: the interviewer asking Collin Murray-Boyles about the defense or the Raptors scoring 66 points?

3. **Subscene –** Subscene refers to “captioning” a segment of the video based on the segment conditioning in the question.
    a. **General Examples**
        i. Describe what happens in the match when the score is 118-2 in favor of UGA?
        ii. In the third quarter, with 9:49 on the clock, how does the Laker's bench react when Lakers' Player 10 dunks the ball? (Answer: The players on the bench stood up and applauded, with some of them pointing one finger up.)
        iii. When the fight scenes showing the inside of a warehouse, and a car burning on the freeway appear on screen, what are the two words the narrator uses to refer to himself? (Answer: He referred to himself as \"the shadows\", and \"vengeance\")
4. **General Holistic Reasoning**
    a. **General Examples**
        i. What was the point of the video breaking up the game into clips?
        ii. Describe how the colored bars falling from the top of the screen relate to the song being played
        iii. What happens to the majority of the cars at the end of each of the clips and how do most people react in the clips?
5. **Inference – Understanding intentions, purposes and causal relationships**

a. **General Examples**
    i. Why do the teams switch after the scoreboards and after the announcer says, \"She just added some emphasis at the end\"?
    ii. Why did the wicketkeeper scream in excitement, and the bowler belly bump his teammate when the score is 145-2?
    iii. Based on the audio and visual components in this \"Inception\" trailer, why is Dom Cobb feeling \"guilt\"?

6. **Context**
    a. **General Examples**
        i. When Amari Williams discusses his biggest takeaway from the last two weeks, what does the black and white billboard say at the top left of the screen?
        ii. What visual elements are present in the background and foreground when the man says \"That's how we're gonna win, Mason get back\"
        iii. What visual elements appear as the man starts to create music using the bottles?
7. **Needle**
    a. **General Examples**
        i. Describe the graphic that pops up when the man says \"Definitely check it out in the masterclass that is in the link below
        ii. During the first quarter, who is the player that hits his \"first three of the night\" when a minute and ten seconds are left on the digital clock.
        iii. Who is the player that walks up to receive the \"MBA 2K26 Game MVP trophy,\" from Eric Watson, the VP of Sales Operations at Kia America?
8. **Referential Grounding**
    a. **General Examples**
        i. Who are the two people visually present when the man says \"What is the protocol with Emily and Morgan\"
        ii. What creates the distinct humming and buzzing sound in the video and what causes it to make that noise?
        iii. What does the commentator say when the score is 51-2?
9. **Counting**
    a. **General Examples**
        i. Throughout the video, how many counts does the dance instructor use to choreograph the structure for the line dancing?
        ii. How many times did the Nuggets score from the start of the game till when Dalton Knecht was mentioned by an announcer?
10. **Comparative**
    a. **General Examples**
        i. What is the difference between the man's shirt before and after he says, \"I'll really appreciate it\"?
        ii. What are the two biggest distinctions in Dave Herrero's appearance when he's on-stage performing a song compared with when he's giving on camera interviews?

iii. What is the primary difference with the man before he tells us to keep watching and after?

11. Object Interaction Reasoning

a. General Examples

i. How does the clay change after the man talks about the difference between a \"pour over and a coffee dripper\"?

ii. What is the visual effect that is created when the man places the second prism in the path of the spectrum?

iii. What causes the audio to get distorted during the swimming with dolphins segment?

12. Audio-Visual Stitching

a. General Examples

i. When Steven first refers to the inventor of AeroPress, is the inventor of AreoPress in the same room as Steven or is it a separate clip spliced in?

ii. How do the spliced musical clips in the interview pace the interview?

iii. What is the purpose of the segment in the beginning that says \"Super street Talent\" with the musical jingle?

13. Tacking Spurious Correlations

a. General Examples

i. In a scene in the input clip, a group of people is having a meeting with a table in the middle, with a man saying \"If you can't bring down the charging bull, then don't wave the red cape at it.\" Who are they referring to? (Answer: They are referring to Superman as he is shown holographically in the middle of the table.)

ii. Describe the unique event that occurs in what the man refers to as \"the first demonstration\" with the cone shaped cylinder? (Answer: A cone shaped cylinder is able to roll up against an incline plane.)

iii. In a scene in the input clip, a person walks towards a human gathering in a mask, with a man shouting in the background "What do you believe in?" -- to which the mad responds saying: (Answer: "I believe in, whatever doesn't kill you, simply makes you stranger.”)

Sub Task Types

1. Human Behavior Understanding

“What did the person do first — wave at the camera or pick up the cup?”

“Which happened after the woman entered the room — did she sit down or start talking?”

“In what sequence did the man smile, raise his hand, and walk away?”

2. Scene Recognition

“Which location appears first — the kitchen or the park?”

“Did the video transition to the night scene before or after the fireworks began?”

“In what order do the indoor and outdoor scenes occur?”

3. OCR Recognition

“Which text is displayed first — the title ‘Welcome’ or the subtitle ‘Part 1’?”

“Does the number ‘2025’ appear before or after the words ‘Final Countdown’?”

“In what order do the signs ‘Stop’ and ‘Caution’ appear on screen?”

4. Causal Reasoning

“Did the glass fall before or after it shattered?”

“Which happened first — the man tripped or the box slid across the floor?”

“Did the alarm sound before the people began running?”

5. Intent Understanding

“Did she glance at the exit before or after standing up?”

“Which action occurred first — the person checking their watch or heading toward the door?”

“In what order did the man look at the ball, then kick it?”

7. Hallucination

“Did the fireworks start after the crowd looked up?”

“Was there any point in the video where a car horn sounded before a car appeared on screen?”

“Did the text ‘Game Over’ appear after the character fell, or did that never happen?”

8. Multi-Detail Understanding

“In what order did the boy run, jump, and wave to his friends?”

“Which sequence occurred: (a) dog barked, (b) door opened, (c) person entered?”

“List the order in which the camera zoomed in, the lights dimmed, and the announcer spoke.”