

# Towards Video Thinking Test: A Holistic Benchmark for Advanced Video Reasoning and Understanding

## Supplementary Material

### 1. Annotation Detail

We present the number of human hours at each stage in the *Data Curation Process* as follows. In total, the annotation process cost 8227.32 human hours.

Table 1. Time Estimation for Dataset Curation Process. Notes: \*2 indicates that two people are required for this stage; \*4 refers to four natural adversarial questions per video; \*5 covers all questions for human baseline annotation. Q stands for Question; A stands for Answer; R stands for Rationale.

Dataset Curation Stage	#Data	Hour/Data	Total (hour)
Trial Data Annotation	226	0.5	113
Trial Data Alignment	226	0.25	56.5
Complex Video Collection	2,977	0.16	496.17
Complex Video Alignment	2,977	0.05*2	297.7
Primary Q&A&R Annotation	2,338	0.5	1,169
Primary Q&A&R Alignment	2,338	0.3*2	1,402.8
Sampling Check	1,344	0.25*2	672
Adversarial Question Annotation	1,300*4	0.16	832
Adversarial Question Alignment	1,300*4	0.08*2	832
Human Baseline Annotation	1,300*5	0.16	1,040
Total			8227.32

### 2. Mathematical Definition of the Robustness Score

- $\mathcal{A}_{\text{primary\_correct}}$  be the set of videos where the primary open-ended question is answered correctly.
- $\mathcal{A}_{\text{paraphrased\_correct}}$  be the set of videos where the paraphrased open-ended question is answered correctly.
- $\mathcal{A}_{\text{correctly\_led\_correct}}$  be the set of videos where the correctly-led open-ended question is answered correctly.
- $\mathcal{A}_{\text{wrongly\_led\_correct}}$  be the set of videos where the wrongly-led open-ended question is answered correctly.
- $\mathcal{A}_{\text{multiple\_choice\_correct}}$  be the set of videos where the multiple-choice question is answered correctly.

The set of videos where all five questions are answered correctly, denoted as  $\mathcal{A}_{\text{full\_correct}}$ , is the intersection of all these sets:

$$\mathcal{A}_{\text{full\_correct}} = \mathcal{A}_{\text{primary\_correct}} \cap \mathcal{A}_{\text{paraphrased\_correct}} \cap \mathcal{A}_{\text{correctly\_led\_correct}} \cap \mathcal{A}_{\text{wrongly\_led\_correct}} \cap \mathcal{A}_{\text{multiple\_choice\_correct}}$$

Thus, the Robustness Score (RB) becomes:

$$R = \frac{|\mathcal{A}_{\text{full\_correct}}|}{|\mathcal{A}_{\text{primary\_correct}}|}$$

Where  $|\mathcal{A}|$  denotes the cardinality (size) of the set  $\mathcal{A}$ , representing the number of videos in that set.

### 3. Prompt for Evaluating Open-ended Answer

Table 2 shows the prompt for evaluating open-ended answers. A score of 3 or higher is considered correct, while scores below 3 are deemed incorrect. We refer to the prompt introduced in VideoChatGPT [? ].

#### System Message

You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here's how you can accomplish the task:

#### INSTRUCTIONS:

- Focus on the meaningful match between the predicted answer and the correct answer.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the answer.

Please evaluate the following video-based question-answer pair:

Question: question

Correct Answer: answer

Predicted Answer: pred

Provide your evaluation only as a yes/no and score where the score is an integer value between 0 and 5, with 5 indicating the highest meaningful match. " Please generate the response in the form of a Python dictionary string with keys 'pred' and 'score', where value of 'pred' is a string of 'yes' or 'no' and value of 'score' is in INTEGER, not STRING.

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string.

For example, your response should look like this: 'pred': 'yes', 'score': 4.

Table 2. System message for evaluating the open-ended answer.

### 4. Error Analysis

In this section, we give more analysis about the errors made by GPT-4o.

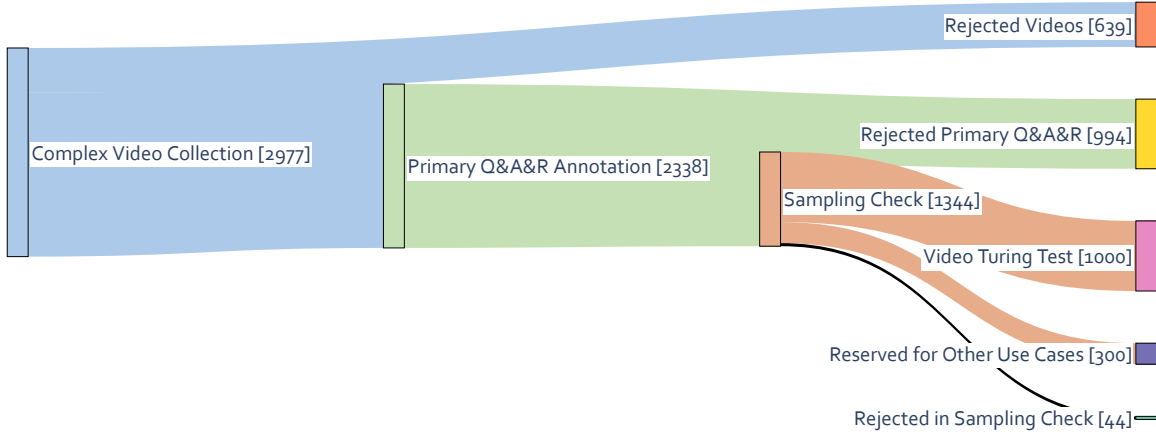


Figure 1. The data annotation flow of Video Turing Test. Q stands for Question; A stands for Answer; R stands for Rationale.








<b>Elements Attributes</b> → <i>Unusual</i>  <b>Q-1:</b> Which part of the person on the ladder precisely catches the hammer handed by the individual on the ground? <b>A:</b> The worker on the ladder clamps the hammer with his buttocks. The worker below adeptly tosses the hammer upwards with one hand, to the worker high on the ladder. <b>GPT-4o:</b> The hammer is received by the worker catching it mid-air with right hand.	<b>Event Attributes</b> → <i>Unclear</i>  <b>Q-3:</b> After solving the Rubik's Cube, which color is facing up? <b>A:</b> The color of the magic side facing up is white, which can be determined by looking at the top face of the cube in the image. The orange color is the face facing the kid. <b>GPT-4o:</b> The color facing up on the Rubik's Cube is yellow.	<b>Plot Attributes</b> → <i>World Knowledge - Metaphor</i>  <b>Imperfect Cognition</b> <b>Q-5:</b> What does this man want to imitate? <b>GPT-4o:</b> The man is imitating riding a motorcycle. <b>Required World Knowledge</b>  <i>*Battle in a trench</i>
<b>Elements Localization</b> → <i>Spatial Arrangement - Action Sender/Receiver</i>  <b>Q-2:</b> What are the characteristics of the person who puts the bamboo into the machine? <b>A:</b> The man placing the bamboo into the machine is wearing blue-black leather gloves, and black-and-white checked sleeves. The person in blue apron is the one moving bamboo away from the machine <b>GPT-4o:</b> The person putting the bamboo into the machine is wearing a blue apron over a camouflage-patterned shirt and red gloves.	<b>Event Counting</b> → <i>Spatial Arrangement - The boundary of an event</i>  <b>Q-4:</b> How many times did this person fold the lotus root stem in total? <b>A:</b> The person folds the lotus root stem a total of 15 times. Every time he folds the lotus stem, he bends it downward and switches between his left and right hands. <b>GPT-4o:</b> The person folded the lotus root stem a total of 20 times	<b>Plot Attributes</b> → <i>Complex Plot - In-context Reasoning</i>  <b>Prefect Recognition</b> <b>Q-6:</b> What does the two boy near the table do in the video? <b>GPT-4o:</b> They take turns flipping a partially filled water bottle, attempting to make it land upright on the table. <b>Imperfect Cognition</b> <b>Q-7:</b> Who is the winner of the game? <b>GPT-4o:</b> The winner of the game is the man in blue as he reached 60 successful bottle flips first. (A: The boy in gray shirt.)

Figure 2. **Error cases in typical question types.** We mark *rationale* answers with a grey background. Video links of each case : Q-1 Q-2, Q-3, Q-4, Q-5, Q-6 & Q-7.

#### 4.1. Recognition: Detecting objects and their attributes

In this subsection, we analyze errors in six question types focused on the “element” and “event” categories. These errors typically stem from visual complexity, challenging the recognition capabilities of the model.

**Element Attributes and Event Attributes.** In this category, 80% of errors involve unclear or unusual subjects in the questions, which relate to elements or events. These errors are linked to issues of unclear and the presence of unusual content in visual complexity. For instance, as depicted in Fig. 2-Q1, when confronted with unusual content, the model often defaults to the most common outcome rather than what is actually depicted in the video. For clarity is-

sues, as shown in Fig. 2-Q3, the model struggles to accurately identify the color of a small Rubik’s Cube in the video frames.

**Event Counting.** In this category, one specific errors arise from the model’s difficulty in accurately identifying the start and end points of repeated events, despite correctly classifying the event type (Fig. 2-Q4).

**Element Localization and Event Localization.** Errors in this category, which make up 79%, are related to spatio-temporal challenges. In spatial terms, a common error occurs when multiple individuals are present in a scene, and the model incorrectly assigns actions to the wrong person. This issue is particularly prevalent in interactions involving

two people, leading to confusion over who is performing and who is receiving the action (Fig. 2-Q2).

**Positional Relationship.** Understanding the relative positions of elements is a fundamental human skill. Yet, we observed that models struggle with this task. For instance, when asked whether element A is on the left or right side of B, the model typically responds “left” if A visually appears on the left side of the video frame. This response disregards their actual spatial relationship within the context of the video. Such findings indicate a significant limitation in the model’s ability to accurately interpret positional relationships.

**Displacement.** For a frame-based model, these questions challenge the model’s ability to track the development of the event across consecutive frames. For instance, considering the displacement of an object from the previous frame to the current one poses a significant challenge if the model’s vision encoder struggles with fine-grained spatial localization grounding [? ].

#### **4.2. Cognition:** *Reasoning the likely intents, goals, and social dynamics of people*

In this subsection, we analyze errors in question types associated with the “plot.” These errors are typically due to narrative complexity. When prompted, the model demonstrates recognition-level perception abilities; however, the narrative complexity challenges the model in addressing “cognition” level questions.

**Character Reaction and Character Motivation.** As discussed in Sec.??, world knowledge significantly contributes to narrative complexity. Fully understanding characters’ reactions and motivations requires applying this knowledge. Commonly, this involves grasping psychological activities, which are subjective by nature. To answer relevant questions effectively, the model must do more than just describe the video; it needs to link these descriptions to broader world knowledge.

**Plot Attributes and Objective Causality.** The typical errors in “plot attributes and objective causality” stem from a lack of world knowledge and in-context reasoning ability. An interesting aspect of necessary world knowledge is its multi-modal nature, essential for correct responses in this category. For example, as shown in Fig. 2-Q5, while the model can accurately describe a man’s actions in the video, understanding what these actions imply—such as imitating a battle scene in a trench—requires linking the video content with relevant world scenes. Moreover, the model’s limited in-context reasoning is evident as it struggles to integrate diverse perceptual inputs into a cohesive understanding of social dynamics, despite accurately answering recognition-level questions about actions observed in the video.