

ViTED : Video Temporal Evidence Distillation

Supplementary Material

This section contains supplementary material to support the main paper. The contents include:

- (A7) Preliminary study on how often evidence chains are actually needed for VideoQA.
- (A8) Evidence pool generation details and hyper-parameters, and an illustration of our hierarchical evidence pool (Table A7).
- (A9) Details of full prompts used throughout Sec. 3.
- (A10) Baseline details and training setting details to supplement Sec. 4.
- (A11) Dataset details including training and evaluation set composition and splits.
- (A12) Evaluation protocol details for the human study of evidence quality to supplement Sec. 4.
- (A13) Additional ablation experiments to supplement Sec. 4.
- (A14) Expanded sets of qualitative results to add to those presented already in Figures 3 and 5.
- (A15) A discussion on the scope and limitations of our approach.

A supplementary html page with video versions of paper figures is also attached.

A7. How often are evidence chains required?

In Sec. 3 we highlighted the need for evidence chain reasoning. We investigate what percentage of the questions in current VideoQA datasets actually require evidence chains to reason about the final answer. We manually annotate 50 samples and found that 54% (27/50) of videos in NExTQA-val do indeed require evidence. The remaining samples can be answered by single frame or a textual caption from the video itself. We further investigate how visual evidence is distributed across the video by computing an *entailment score* — a score for how likely the description of different parts of the video entail the answer of the question. We calculate this by prompting an LLM and validate that high-scoring video segments do contain essential evidence to help question answering.

A8. Evidence Pool Generation Details

We present additional details and examples of our evidence pool generation strategy to supplement Sec. 3.1.

A8.1. Hyper-parameters

Table A6 summarizes the key hyper-parameters across the three main phases: Evidence Pool Generation (Sec. 3.1), Evidence Search (Sec. 3.2), and Model Training (Sec. 3.3).

Evidence Pool Generation involves hierarchical segmentation with $N = 5$ levels ensures evidence extraction across temporal granularities using LLaMA-3.2-Vision-Instruct-11B and a tailored prompt.

Evidence Search employs LLaMA-3.1-8B-Instruct with a beam width $W = 4$, threshold $T = 0.7$, and a maximum of 3 iterations for evidence refinement.

Model Training involves two stages: instruction tuning and temporal evidence distillation, with distinct configurations for TimeChat and LLaVA-Video. Key components, including Q-Formers and adapters, are optimized with LoRA ranks of 32 and 128, ensuring reproducibility of the pipeline.

A8.2. Hierarchical Visual Evidence Chunk

The hierarchical narrated evidence approach divides a video into multiple levels of granularity, ranging from a global context to fine-grained, localized details. This structured segmentation allows for capturing evidence across different temporal scales. The five hierarchical levels are as follows:

- **Hier1 (Global)**: Represents the entire video as one single segment, providing a broad context for understanding the overall scene. This level covers the full video duration with $L = 1$ and $S = 1$, capturing general, high-level information about the content.
- **Hier2 (Sectional)**: Divides the video into two large segments, each covering half of the video. This level is used to capture larger shifts or sections of the video, such as changes in scenes, major actions, or transitions. Here, $L = 1/2$ and $S = 1/4$.
- **Hier3 (Detailed)**: Breaks the video into smaller chunks (e.g., one-quarter of the video duration), allowing for a more detailed view of specific events and interactions. This level captures important actions or events that may not be apparent at a larger granularity. In this case, $L = 1/4$ and $S = 1/8$.
- **Hier4 (Fine-Grained)**: Divides the video into even smaller segments, focusing on precise details such as body language, gestures, or smaller interactions. This level allows for a closer examination of fine-grained events. The segment length is $L = 1/8$ with a stride of $S = 1/16$.
- **Hier5 (Atomic)**: The smallest possible temporal chunk, capturing the finest details of the video. At this level, the model focuses on very localized moments such as rapid actions or fleeting events. Here, $L = 1/16$ and $S = 1/16$.

Parameter	Value/Details
Evidence Pool Generation	
Hierarchical Levels (N)	5
Segment Length (L) and Stride (S)	$\{(1/16, 1/16), (1/8, 1/16), (1/4, 1/8), (1/2, 1/4), (1, 1)\}$
Evidence Model	LLaMA-3.2-Vision-Instruct-11B
Evidence Search	
Refinement Model	LLaMA-3.1-8B-Instruct
Beam Width (W)	4
Probability Threshold (T)	0.7
Maximum Iterations	3
Model Training	
Base Models	TimeChat, LLaVA-Video
Training Stages	Stage-1: Instruction Tuning (Answer tokens), Stage-2: Temporal Evidence Distillation (Evidence + Answer tokens)
Evidence Chain Filtering Criterion	Correct answer likelihood ($f(C_i^*) = 1$)
Epochs per Stage	TimeChat: 10 epochs; LLaVA-Video: 1 epoch
Model Parameters to Train	TimeChat: Image Q-Former, Video Q-Former, Linear Layer (96 input frames). LLaVA-Video: Adapter, LLM Backbone.
LoRA Rank	TimeChat: 32; LLaVA-Video: 128

Table A6. **Summary of Parameters and Settings.** Key parameters used in Evidence Pool Generation, Evidence Search, and Model Training. Metrics are detailed to ensure reproducibility.

This hierarchical segmentation method ensures that evidence from a wide range of temporal scales is captured, from broad contextual understanding to very specific and rapid events. We showcase one example of our hierarchical evidence pool in Table A7.

Frame 1	Frame 2	Frame 3	Frame 4	Frame 5	Frame 6
					
Question: Why are there people surrounding the two people fencing? Options: A. guide. B. performing. C. watching them. D. play for fun. E. to maintain the distance.					
Hierarchy 1: 0. [0.0-0.062seconds] There are no people surrounding the two people fencing. The image only shows two individuals engaging in a fencing match, with no additional individuals present. 1. [0.062-0.125seconds] There are no people surrounding the two people fencing. The image shows a person fencing with another person in a yard, with no other people visible. ... 15. [0.938-1.0seconds] The video shows a scene where two individuals are engaging in a fencing match, with a group of people standing around them. The people surrounding the fencers appear to be watching the match, possibly as spectators or participants in the activity. The presence of these people suggests that the fencing match is part of a larger event or gathering, such as a tournament, training session, or social gathering. The fact that the people are standing around the fencers, rather than actively participating in the match, implies that they are observing the activity rather than engaging in it themselves.					
Hierarchy 2: 0. [0.0-0.125seconds] The video shows two people fencing in a backyard, with a fence surrounding the area. The presence of the fence suggests that the two people are fencing in a designated area, possibly to ensure safety and prevent damage to surrounding objects. 1. [0.062-0.188seconds] The image shows two people wearing fencing masks and holding fencing weapons, with a third person standing nearby. The third person is likely a coach or instructor, providing guidance and support to the two fencers. ... 13. [0.812-0.938seconds] The image depicts two individuals engaged in a fencing match, with a third person observing from a distance. The presence of the third person suggests that they may be a coach, instructor, or spectator, providing guidance and support to the fencers.					
Hierarchy 3: 0. [0.0-0.25seconds] There are no people surrounding the two people fencing. The image shows two people fencing in a backyard, with a house and trees in the background. The two people are wearing fencing gear and holding swords, and appear to be engaged in a fencing match. There is no evidence of people surrounding them. 1. [0.125-0.375seconds] There are no people surrounding the two people fencing. The image only shows two people fencing, and there is no indication of any other individuals present. ... 6. [0.75-1.0seconds] The image depicts two individuals engaging in a fencing match, with a third person standing nearby, likely serving as a referee or instructor. The presence of the third person is likely due to the fact that the two fencers are children, and the third person is there to ensure their safety and provide guidance on proper fencing techniques.					
Hierarchy 4: 0. [0.0-0.5seconds] The video shows two people engaging in a fencing match, with one person holding a foil and the other holding an epee. The scene is set in a backyard, with trees and bushes visible in the background. There are no people surrounding the two fencers. 1. [0.25-0.75seconds] The image shows two people fencing in a backyard, with one person wearing a white shirt and black pants, and the other wearing a white shirt and white pants. The person on the left is holding a fencing sword, while the person on the right is holding a fencing mask. There are no people surrounding them in the image. 2. [0.5-1.0seconds] The individuals surrounding the two people fencing are likely wearing protective gear, including masks, gloves, and fencing jackets, which are characteristic of the sport of fencing. This protective gear is worn to prevent injury during the match. The presence of these individuals, likely coaches or referees, suggests that they are involved in the fencing competition, possibly officiating or providing guidance to the participants.					
Hierarchy 5: 0. [0.0-1.0seconds] The image shows two people fencing, with one person holding a sword and the other person holding a foil. The person on the left is wearing a white shirt and black pants, while the person on the right is wearing a white shirt and white pants. The person on the right is holding a sword in their right hand, while the person on the left is holding a foil in their left hand. 1. [0.0-1.0seconds] There are no people surrounding the two people fencing in the image. The background of the image shows a grassy area with trees and bushes, suggesting that the fencing is taking place outdoors. There is no indication of any people or objects surrounding the two fencers.					

Table A7. **Example of our Hierarchical Evidence Pool.** All five levels are shown with generated descriptions. Note: [0.0-1.0 seconds] is normalized to video length.

A9. Detailed Prompts for Evidence Generation

We provide detailed prompts corresponding to the high-level prompts in the main paper, specifically, textual evidence narration from video chunks in Sec. 3.1, converting evidence to chain-of-thought and evidence candidate refinement in Sec. 3.2.

Textual evidence narration from video chunks

[Video] Sampled Video Frames

[Instruction] Please provide short and concise evidence from the video that can help answer the question. The format should be as follows:

Evidence: your_evidence_here

[Output] Evidence

The evidence for each chunk is then paired with its temporal window between $[0, 1]$, normalized by the duration of the video (e.g., `[start-end] evidence`).

Evidence candidate refinement

[Video] Sampled Video Frames

[Instruction] Use the following video transcript to gather a list of evidence to help answer the question “question”.

Options: options

Transcript: transcript

Provide the evidence in the following json format that will help reach the answer in a step by step manner. Format:

```
{
  "evidence_chain": [
    {
      "start_time": float,
      "end_time": float,
      "evidence": str
    },
    ...
  ]
}
```

Limit your evidence chain to at most `beam_width` steps. Respond directly with the json. Please return the evidence as a valid JSON object with proper formatting. Ensure all strings are enclosed in double quotes (") and no invalid syntax is used.

[Output] Evidence

Converting evidence to chain-of-thought

[Video] Sampled Video Frames

[Instruction] You're the assistant to seek the visual evidence chain from the video to answer the question “question”

Visual Evidence Observed from Video: transcript

The total duration of the video is `vid_duration`. Each evidence is the narrated question-relevant information within the `[t1-t2seconds]` interval of the video.

Please utilize both the timestamps of the evidence and the temporal hint in the question, and also focus on the objects/events in the evidence that strongly indicate the moment described in the question, and then think step-by-step using the most relevant evidence to derive your answer. Please rewrite relevant evidence and its temporal span into a chain-of-thought reasoning based on the video. Such as, as the question ask about “what does the man do after he enters the room in the end of the video?”, we find that both `[t1-t2seconds]` and `[t3-t4seconds]` intervals show the man entering the room, since the question is asking end of the video, we look at the latter interval and find that he is picking up a cup after entering the room, thus the answer is xxx. Please provide your step-by-step reasoning `full_chain_of_thought` and keep the `[t1-t2seconds]` when you describe the visual evidence. You can merge `[t1-t2seconds]` and `[t3-t4seconds]` as `[t1-t4seconds]` when they're the same evidence information. Based on your step-by-step reasoning, select the most appropriate option letter as your `final_answer`. Please try to only include the evidence that is relevant and necessary for answering the question. Format:

```
{
  "full_chain_of_thought": str,
  "final_answer": str
}
```

Respond directly with the JSON.

[Output] Evidence

A10. Additional Implementation Details

We provide more details for two of our baselines in Sec. 4 below.

Direct Multi-Evidence Grounding This method tests whether LLaMA-3.2-Vision-Instruct-11B can replace a hierarchical evidence pool and multi-hop search by directly generating temporally-aware, multi-hop evidence in one pass. We prompt the model to output evidence by referencing specific video timestamps, formatted as `[start_time-end_time]`, with accompanying descriptions. This approach assesses if the model can synthesize detailed, multi-hop evidence without the need for additional structuring or search processes. Prompt: “[Question] [Option] Please

provide detail sequence of information of each part of the video that help answering the question. The format should be in the form of: [start_time2-end_time2] This clip 1 shows that xxx which indicate xxx. [start_time2-end_time2] This clip 2 shows that xxx which indicate xxx...: ...”

GT-Guided Evidence Sampling This strategy uses GT-based filtering to iteratively refine evidence. The model initially generates an evidence chain in response to a question and potential answers, then reviews the evidence against the video data to identify the answer most aligned with the ground truth. If the generated evidence chain does not yield the correct answer, we adjust the model’s response by varying temperature settings or re-prompting, allowing up to three iterations to improve the evidence quality. This method evaluates whether the model can consistently generate valid, temporally-grounded evidence chains using direct GT guidance. Prompt: “[Question] [Option] Please provide your evidence chain in order in the video that help answering the question.”

A11. Training and Evaluation Dataset Details

We provide more details about the training and evaluation datasets used in Sec. 4.

Training Data For Stage-1 training, we use the training splits of PerceptionTest (7.4K), NExT-QA (34.1K), STAR (45.7K) and a collection of 127.1K instances from public long video QA datasets. For Stage-2, we train with additional instances of our temporal evidence data synthesized by LLaMA-3.2-Vision-Instruct-11B and LLaMA-3.1-8B-Instruct following Sec. 3.1 and 3.2, bringing the total to 291K. Specifically, PerceptionTest (12.3K), NExT-QA (58.3K), STAR (75.7K) and long video QA (145.1K).

Evaluation Data We evaluate our models on the validation splits of CinePile [32], PerceptionTest [30], NExT-QA [48], STAR [47], MVBench [17], NExT-GQA [49].

A12. Evidence Quality Evaluation Protocol

As mentioned in Sec. 4.1, we collect human annotations to verify the quality of evidence chains. We provide detailed instructions and evaluation protocol for the study. We recruit two graduate students with expertise in video understanding, and ask each of them annotate 50 examples across five aspects. The instructions for the annotation task are provided below.

Evidence quality annotation instructions

Objective: The purpose of this evaluation is to assess the quality of evidence chains generated for answering video-based questions. Your task is to review the provided evidence chains in the context of the video content and score their quality across five distinct aspects. Your feedback will help refine and improve the performance of the evidence generation system.

Workflow: You will be shown a video and a question related to the video. A generated evidence chain, which includes textual descriptions of events and reasoning, will be presented. Carefully review the video and identify the key events occurring between them. Read the question and ensure you understand what is being asked. Compare the evidence chain to the video frames and question, evaluating it based on the five aspects described below.

- **Temporal:** Does the temporal window match the evidence text?
- **Faithfulness:** Is the evidence faithful to the video content?
- **Logical:** Is the reasoning logical across the evidence?
- **Relevance:** How relevant is the evidence chain to the video/question?
- **Completeness:** Does the evidence chain capture all required information in the video to answer the question?

Important Notes:

- Be as objective as possible. Focus on the content and structure of the evidence chain.
- If you are unsure about any aspect, provide comments for clarification.
- Your evaluation will directly impact the improvement of the evidence generation system.

The scoring rubric is presented in Table A8 and Table A9 shows some examples of the collected human annotations.

A13. Additional Ablation Experiments

We present additional ablation experiments to supplement Sec. 4.4.

Evidence Generator and CoT Narrator In Table A10, we test various evidence generator models and the Chain-of-Thought (CoT) narrator models. Larger models like LLaVA-NEXT-Video-32B achieved the highest average performance, particularly excelling in causal and descriptive questions. While models with a smaller CoT narrator generally performed worse (3B versus 8B in full model), indicating a strong L-CoT model is useful for complex multi-step tasks.

	Temporal	Faithfulness	Logical	Relevance	Completeness
3	The evidence correctly identifies the time sequence of events.	The evidence is fully consistent with the video content.	The evidence forms a coherent and logical reasoning chain.	The evidence is directly relevant to the question and frames.	The evidence includes all critical information needed to answer the question.
2	The temporal sequence is somewhat accurate but contains minor errors.	The evidence is mostly accurate but includes minor inconsistencies.	The reasoning is partially logical but has gaps or weak links.	The evidence is somewhat relevant but includes unnecessary information.	The evidence captures most key details but omits some minor elements.
1	The evidence significantly misrepresents the time sequence.	The evidence is misleading or contains major inaccuracies.	The reasoning is illogical or lacks coherence.	The evidence is irrelevant or off-topic.	The evidence is incomplete and misses significant details.

Table A8. **Scoring rubric for evidence quality.** Annotators score each evidence chain on a three-point scale, across five aspects.






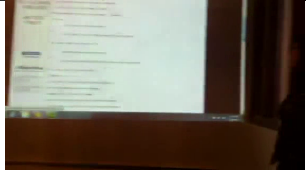
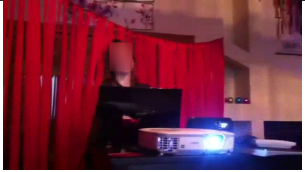

Frame 1	Frame 2	Frame 3	Frame 4
			
<p>Question: How many animals are involved in the video? Options: A. five. B. four. C. one. D. three. E. two.</p> <p>Evidence: As the video starts, we look at the interval from 0.0 to 11.252 seconds in the video, and observe that there is only one animal, a dog. However, this observation does not necessarily indicate the number of animals in the video throughout its duration. We then examine the subsequent intervals to determine the number of animals involved. In the interval from 5.581 to 16.924 seconds in the video, we see a dog and a cat, which suggests that there are at least two animals in the video. Furthermore, in the intervals from 16.924 to 22.505 seconds in the video, and from 22.505 to 28.086 seconds in the video, we continue to observe a dog and a cat, confirming that there are indeed two animals present in the video.</p> <p>Answer: Therefore, based on the visual evidence, it is evident that the number of animals involved in the video is two.</p> <p>Scores: Temporal: 3, Faithfulness: 3, Logical: 3, Relevance: 3, Completeness: 2</p>			
Frame 1	Frame 2	Frame 3	Frame 4
			
<p>Question: Why is there a projector on the table? Options: A. displaying lyrics of the song. B. listening to speaker. C. to feed baby. D. eat. E. drink waters.</p> <p>Evidence: From the given visual evidence, we can observe that the projector on the table is being used for presentation or display purposes from 0.0 to 4.621 seconds and 9.242 to 11.535 seconds. It is also mentioned that the projector is connected to a laptop and is being used to display a presentation or slideshow from 0.0 to 18.485 seconds. However, from 18.485 to 36.97 seconds, the image depicts a man playing a guitar on stage, with a projector positioned on the table in front of him. The projector is likely being used to display visual elements or lyrics for the performance, enhancing the overall experience for the audience.</p> <p>Answer: This suggests that the projector is being used to display lyrics of the song during the performance.</p> <p>Scores: Temporal: 3, Faithfulness: 2, Logical: 3, Relevance: 2, Completeness: 3</p>			

Table A9. **Human Evaluation of Evidence Quality.** Two examples are provided above. Temporal evidence evidence segments are highlighted for visualization only.

A14. Additional Qualitative Evaluations

A14.1. Evidence Examples

Figure A6 illustrates examples of evidence chains generated for video question answering. For more examples, please refer to our ‘vited_evidence_data.html’.

A14.2. Video Question Answering

Figure A7 demonstrates the effectiveness of our approach in answering complex video-based questions. For more examples, please refer to our ‘vited_prediction.html’.

A15. Scope and Limitations of VITED

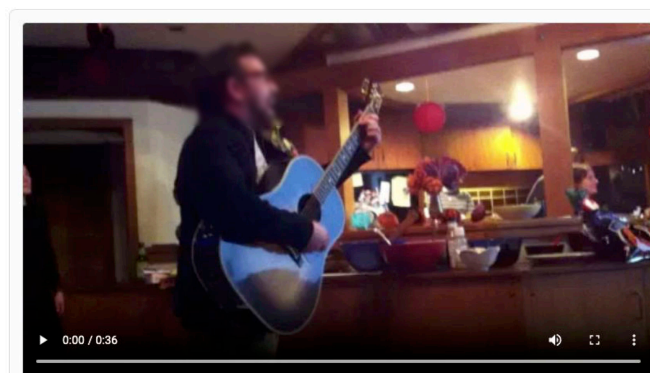
VITED represents a significant step forward in enabling evidence-based temporal reasoning for complex video question answering. However, its scope is inherently influenced

Model	NExT-QA			Avg
	Temporal	Causal	Descriptive	
Evidence Generator				
LLaVA-OneVision-0.5B [16]	72.27	75.17	81.47	75.22
LLaMA-3.2-Vision-Instruct-11B [16]	73.46	76.34	81.03	76.14
LLaVA-OneVision-7B [16]	74.26	76.91	85.33	77.36
LLaVA-NEXT-Video-32B [57]	74.38	78.52	83.01	77.88
CoT Narrator				
LLaMA-3.2-Instruct-3B [4]	70.78	74.68	81.72	74.52

Table A10. **Ablation on Evidence Generator** 1) Evidence Generator: model for evidence pool generation, and 2) CoT Narrator: LLM for evidence refinement and searching.

by the nature of current video datasets and limitations of large language models (LLMs). The approach excels at scenarios with relatively simple question structures requiring 1-3 reasoning hops, as observed in benchmarks like NExT-QA and STAR. These tasks align well with the hierarchical evidence pool and chain-of-thought generation strategies utilized. However, questions demanding deeper reasoning or involving rare, highly nuanced interactions remain challenging, given the constrained model’s ability to synthesize and interpret uncommon event sequences.

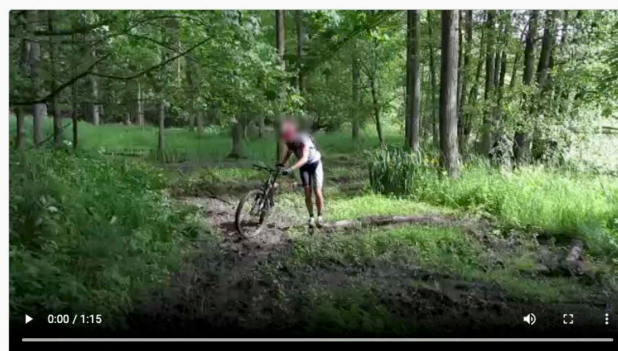
Additionally, the reliance on automated evidence generation introduces imperfections. Evidence chains, while effective in many cases, may propagate errors from initial noisy predictions, such as hallucinated or vague descriptions of video segments. These issues can compound, especially in long videos with sparse critical evidence. Furthermore, while the hierarchical evidence framework attempts to capture granular and global contexts, it is not infallible in identifying or relating temporally distant yet causally connected events. Future work could enhance the model’s robustness by refining temporal evidence representation and incorporating richer, multi-modal cues to address these limitations.



Question: Why is there a projector on the table? options: a. displaying lyrics of the song. b. listening to speaker. c. to feed baby. d. eat. e. drink waters.

Answer: A

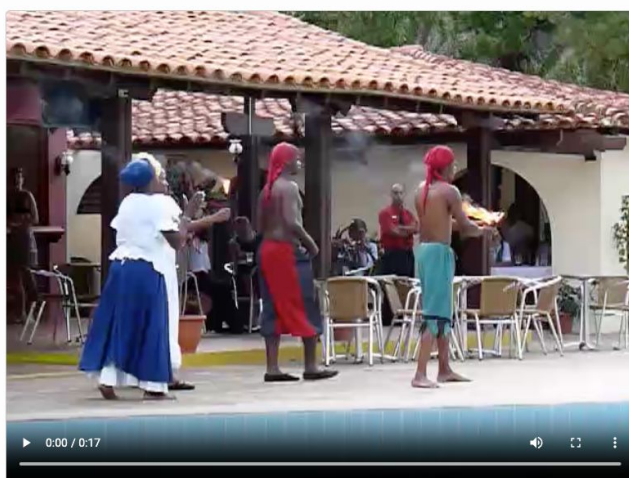
Temporal Evidence Data: From the given visual evidence, we can observe that the projector on the table is being used for presentation or display purposes from [0.0-4.621] seconds and [9.242-11.535] seconds. It is also mentioned that the projector is connected to a laptop and is being used to display a presentation or slideshow from [0.0-18.485] seconds. However, from [18.485-36.97] seconds, the image depicts a man playing a guitar on stage, with a projector positioned on the table in front of him. The projector is likely being used to display visual elements or lyrics for the performance, enhancing the overall experience for the audience. This suggests that the projector is being used to display lyrics of the song during the performance.



Question: Why are the men walking instead of cycling? options: a. fire show costume. b. can not cycle on that ground. c. give spotlight to boy in yellow. d. help him walk. e. enjoy the walk.

Answer: B

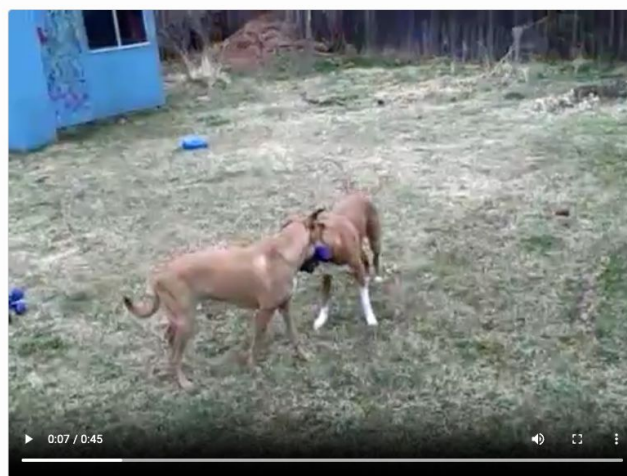
Temporal Evidence Data: The question asks why the men are walking instead of cycling. From 0.0 to 4.689 seconds in the video, the men are exerting effort to move the bike forward, which is consistent with walking. This suggests that they are not riding their bikes at this point. From 4.689 to 9.454 seconds in the video, the muddy trail and uneven terrain make it difficult for the men to ride their bikes, so they are instead opting to walk it. This is a crucial piece of evidence that directly addresses the question. From 18.907 to 23.597 seconds in the video, the path they are walking on is muddy and overgrown with grass and weeds, suggesting that it may be difficult to ride a bike on this terrain. This further supports the idea that the terrain is not suitable for cycling. Based on this evidence, the most likely reason the men are walking instead of cycling is that the terrain is too difficult for them to ride their bikes.



Question: How did the man in red shorts show support for the man in green at the start? options: a. walk. b. blow. c. hold onto handle for support. d. project his voice. e. clap.

Answer: E

Temporal Evidence Data: To answer the question, we need to find the moment when the man in red shorts shows support for the man in green at the start. We can start by looking at the relevant evidence. From 0.0 to 1.088 seconds in the video, the man in red shorts initially stands with his arms at his sides, facing away from the man in green. However, this does not indicate support. From 1.088 to 2.194 seconds in the video, the man in red shorts shows support for the man in green at the start by holding his hand up in the air, with his palm facing the man in green. This is a clear indication of support. From 3.299 to 4.388 seconds in the video, the man in red shorts is seen with his right arm raised, as if he is about to clap or make a gesture of support. This also suggests support, but it is not as direct as the previous evidence. Therefore, the most relevant evidence is that the man in red shorts holds his hand up in the air to show support for the man in green at the start.

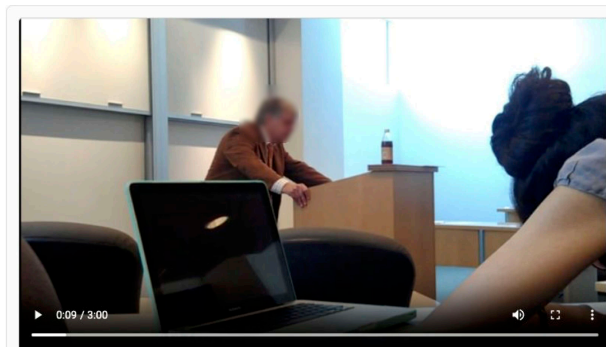


Question: What did the dark coloured dog do after it won the bottle? options: a. lick the other dog. b. run around with it. c. swim after it. d. follow black dog. e. lighter coloured.

Answer: B

Temporal Evidence Data: From 0.0 to 5.7 seconds in the video, The dark-colored dog is seen playing with the bottle on the ground after winning it. However, this is not directly related to the question, so we move on. From 2.827 to 5.7 seconds in the video, The dark-colored dog appears to have won the bottle, as it is seen holding the bottle in its mouth. This indicates that the dark-colored dog has won the bottle. From 5.7 to 8.573 seconds in the video, The dark-colored dog is seen carrying the bottle away from the light-colored dog and moving towards the blue house in the background. This shows that the dark-colored dog is moving away from the light-colored dog with the bottle. From 8.573 to 11.4 seconds in the video, The dark-colored dog began to run away with the bottle in its mouth, indicating that it has claimed victory and is now in possession of the prize. This is the most relevant evidence to the question, as it shows the dark-colored dog running away with the bottle.

Figure A6. Evidence Examples



Question: Why is the lady in purple bending over the table? a. stabilize food. b. resting. c. blow the candles. d. resting. e. writing. identify specific time segments in the video that help answer the question. share the evidence you find in these relevant intervals to support the correct option.

Answer: E

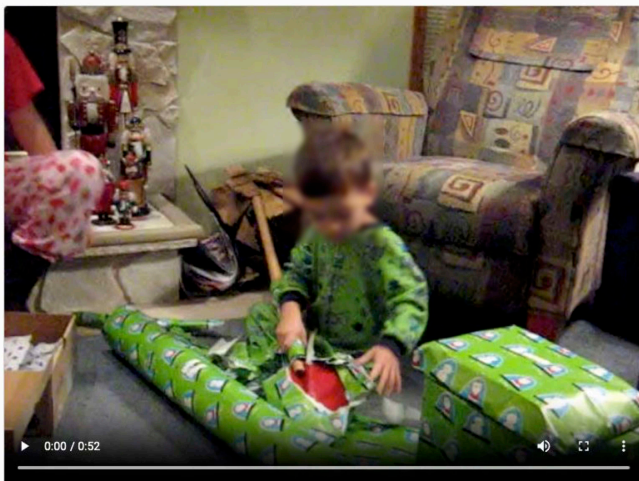
Prediction: As we start observing the video from 0.0 to 22.512 seconds in the video, we see the lady in purple bending over the table, which suggests that she is likely reading or writing something on the table. This observation is further supported by the fact that she is holding a pen or pencil in her hand, indicating that she is engaged in writing or drawing. As we continue to observe the video from 22.512 to 33.775 seconds in the video, we see that the lady in purple is still bending over the table, and her hand is visible, holding a pen or pencil, which is consistent with our previous observation. This behavior is also observed in the interval 33.775 to 45.037 seconds in the video, where the lady in purple is bending over the table, holding a pen or pencil, and appears to be writing or drawing on the table. Finally, in the interval 45.037 to 67.65 seconds in the video, we see that the lady in purple is bending over the table, holding a pen or pencil, and appears to be writing or drawing on the table. Based on these observations, it is clear that the lady in purple is writing or drawing on the table. Therefore, we can conclude that E is the right choice.



Question: How does the woman respond when the priest lets go of her head? a. listen to priest talk. b. sit on grass. c. touch her feet. d. put her hands under child. e. smile at camera. answer with the option letter.

Answer: A

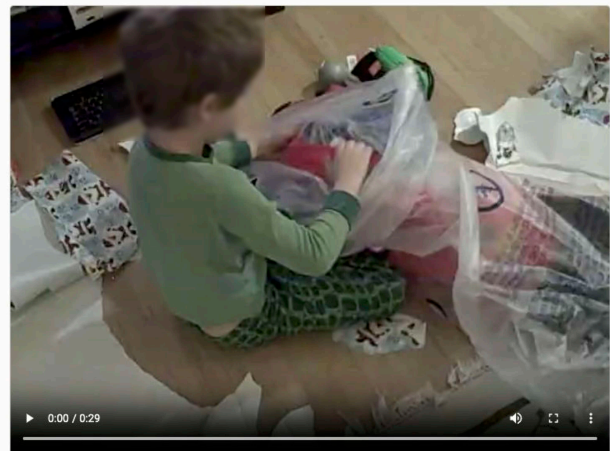
Prediction: The correct answer is: A. The question asks how the woman responds when the priest lets go of her head. From 0.0 to 4.95 seconds in the video, the woman appears to be startled or surprised by the priest letting go of her head. This suggests that she is not expecting this action. From 19.8 to 24.75 seconds in the video, the woman responds with a slight smile and a nod of her head when the priest lets go of her head. This indicates that she is accepting of the priest's action and is not experiencing any negative emotions. Therefore, the most likely response of the woman when the priest lets go of her head is to smile and nod her head.



Question: What did the boy do with the wrappers after he tore them from the spear? a. pick it up. b. moves the fork. c. stabilise himself. d. catch it. e. throw it aside. answer with the option letter.

Answer: E

Prediction: The correct answer is: E. The question asks what the boy did with the wrappers after he tore them from the spear. From 0.0 to 6.579 seconds in the video, the boy is seen tearing the wrappers off the spear and then throwing them on the floor. This indicates that he threw the wrappers aside after tearing them from the spear. Therefore, the most appropriate option is E. throw it aside.



Question: What does the boy do after walking around the toy at the end? a. put hand inside the toy. b. get off the dest. c. sit down. d. comes back. e. blow to the people sitting around him. answer with the option letter.

Answer: C

Prediction: The correct answer is: A. The question asks what the boy does after walking around the toy at the end. From 0.0 to 14.57 seconds in the video, the boy walks around the toy and then sits down on the floor. This suggests that he may be taking a break or resting after playing with the toy. However, this is not the end of the video. From 14.57 to 29.14 seconds in the video, the boy walks around the toy and then sits down on the floor. This is the most relevant evidence for answering the question, as it shows the boy's action after walking around the toy at the end of the video.

Figure A7. Video Question Answering Showcases