

OVBench: How Far is Your Video-LLMs from Real-World Online Video Understanding?

Supplementary Material

6. More Details of Evaluation

6.1. Evaluation for Online Models on Forward Active Responding

As no existing online models can satisfy the demand imposed by our original designs, we choose not to cover this part in our main paper. We introduce an effective evaluation metric tailored for each task consisting of two different dimensions.

Guidance for evaluation metrics design.

- **Accuracy-Based.** The model’s responses should, first of all, be correct without misleading information. We judge the effectiveness of the answer given by the model, and simply average all of them to give the accuracy.
- **Score-Based.** Based on the accuracy-based evaluation, we encourage the response to be both accurate and timely and therefore devise a scoring metric.

Details of evaluation metrics. Given the user’s queries Q_{t_i} at time t_i , the referred events E_j (such as a specific step of a tutorial procedure) with the time interval from t_j to t'_j , the appropriate response A_m at time t_m , the model’s responses $R_{m'}$ at time t'_m , the evaluation function $F(R_{m'}, A_m)$, which directly compare the models’ responses against the right ones, the evaluation metrics of different models are formally given as follows.

1. **[REC]** In this task, the query is only made at a certain time before one complete repetition event happens. In our benchmark, the query is made at the start of the video, i.e. only Q_0 is made.

• Accuracy-Based.

$$Acc = \frac{\sum_{i=1}^N F(R_{m'}, A_m)}{N}$$

• Score-Based.

$$Score = \sum_{i=1}^N e^{i \cdot p_1} \cdot F(R_{m'}, A_m) \cdot 2^{-(m'-m) \cdot p_2}$$

where $F(R_{m'}, A_m) = [A_m == R_{m'}]$, which gives 1 if the model’s response is the same as the answer, and gives 0 otherwise. p_1 and p_2 are parameters to balance the weight. In our evaluation, they are set to 0.2 and 0.05 respectively.

2. **[SSR]** In this task, a query like *Illustrate me on how to make a sandwich according to the video* is made before the start of the procedure. Akin to [REC], the query is only made at the start of the video, i.e. only Q_0 is made.

• Accuracy-Based.

$$Acc = \frac{\sum_{i=1}^N F(R_{m'}, A_m)}{N}$$

• Score-Based.

$$Score = \sum_{i=1}^N F(R_{m'}, A_m) \cdot 2^{-(m'-m) \cdot p}$$

where we leverage GPT-4o to give $F(R_{m'}, A_m)$, measuring the effectiveness of $R_{m'}$ given the reference answer A_m and relevant visual content. p is set to 0.5 to balance weight in our evaluation.

3. **[CRR]** In this setting, queries are made before every A_m , i.e. $range(i) == range(m)$.

• Accuracy-Based.

$$Acc = \frac{\sum_{i=1}^N F(R_{m'}, A_m)}{N}$$

• Score-Based.

$$Score = \sum_{i=1}^N F(R_{m'}, A_m) \cdot 2^{-(m'-m) \cdot p}$$

where we leverage GPT-4o to give $F(R_{m'}, A_m)$, measuring the effectiveness of $R_{m'}$ given the reference answer A_m and relevant visual content. p is set to 0.5 to balance weight in our evaluation.

Prompt Design. To adapt to the online scenarios, we constructed streaming mode prompts with accurate timestamps and also deleted the complicated instructional statement compared to 6.2. Prompts and examples of models’ responses are shown in 8.

6.2. Prompt Design for Offline Models on Forward Active Responding

The *Forward Active Responding* task is intrinsically inappropriate for offline models, as these models only support queries about existing video contents and can not receive additional visual frames after the query is made. However, considering the superiority of offline models against existing online models, we design a multiple-triggering evaluation pipeline and prompt offline models to decide whether the current time is appropriate for answering the user’s query. Formally, given the user’s query Q_{t_0} at t_0 , we leverage offline models to decide at $t_i, i \geq 1; t_i > t_0$ whether

6.1. Online Models on Forward Active Responding		
Prompt Used & Response Examples		
[REC]	[SSR]	[CRR]
Repetition Event Count	Sequential Steps Recognition	Clues Reveal Responding
<p>[Prompt] In the video, the man/woman is [Action] repetitively. Remind me every time when he/she finishes one.</p> <p>[Examples] [Action] Showing something to the camera [Complete Query] In the video, the man/woman is showing something to the camera repetitively. Remind me every time when he/she finishes one. [Query Time] 0:00/Start of the video [GT Times] (0:00-0:07) - (0:09-0:19) - (0:21-0:25) - (0:27-0:34)</p> <p>[Response] [videollm-online] ① You look around.</p>	<p>[Prompt] Illustrate me on the steps of [Procedure] according to the video.</p> <p>[Examples] [Procedure] Make Sugar Coated Haws [Complete Query] In the video, the man/woman is showing something to the camera repetitively. Remind me every time when he/she finishes one. [Query Time] 0:00/Start of the video [Steps] - String the fruit together - Melt the sugar - Soak sugar gourd in sugar [GT Times] (0:44-1:07) - (1:22-1:19) - (1:51-2:11)</p> <p>[Response] [videollm-online] ② You hold a haw in your left hand ③ You hold a haw in your left hand ... ④ You hold a haw in your left hand</p>	<p>[Prompt] [Question]</p> <p>[Examples] [Question/Complete Query] The woman in the black coat walks towards the direction of the man in yellow, what action does she do with the man? [Clues Reveal Time] 5:17 [Answer] She walks past him. [Query Time] 5:10</p> <p>[Response] [videollm-online] ⑤ It seems like the woman in the black coat is walking towards the man in the yellow coat. She is likely to interact with him.</p>
6.2. Offline Models on Forward Active Responding		
[REC]	[SSR]	[CRR]
Repetition Event Count	Sequential Steps Recognition	Clues Reveal Responding
<p>[Prompt] You're a helpful assistant proficient in video question-answering. You're watching a video in which people may perform a certain type of action repetitively. The person performing are referred to as 'they' in the following statement. Your task is to count how many times did different people in the video perform this kind of action in total. Now, answer the following question: [Question] Your response type should be INT, for example, 0/1/2/3...</p> <p>[Examples] [Question] How many times did they showing something to the camera? [Ground Truth] 1 - 2 - 3 - 4 [GT Times] (0:00-0:07) - (0:09-0:19) - (0:21-0:25) - (0:27-0:34) [Query Times] (0:07) - (0:19) - (0:25) - (0:34)</p> <p>[Response] [Gemini-1.5 Pro] 1 - 1 - 1 - 6 [GPT-4o] 0 - 0 - 1 - 3 [Qwen-VL-72B] 1 - 4 - 1 - 4 [Qwen-VL-7B] 1 - 4 - 1 - 4 [LLaVA-NeXT-Video-7B] 2 - 2 - 2 - 2 [LLaVA-OneVision-7B] 2 - 0 - 2 - 1 [LongVU-7B] 3 - 3 - 3 - 3 [Flash-VStream-7B] 2 - 2 - 3 - 3</p>	<p>[Prompt] You're a helpful assistant proficient in video question-answering. You're watching a tutorial video which contain a sequential of steps. The following is one step from the whole procedures:</p> <p>[Query Step] Your task is to decide: Is the man/woman in the video currently carrying out this step? Return "Yes" if the man/woman in the video is currently performing this step; Return "No" if not</p> <p>[Examples] [Procedure] Make Sugar Coated Haws [Query Step] melt the sugar [Ground Truth] N - N - Y - Y [Step Intervals] 1:22 - 1:46 [Query Times] (1:17) - (1:20) - (1:24) - (1:27) - (1:46)</p> <p>[Response] [Gemini-1.5 Pro] N - N - N - N - Y [GPT-4o] N - N - N - Y - Y [Qwen-VL-72B] N - N - N - N - N [Qwen-VL-7B] N - N - N - N - Y [LLaVA-NeXT-Video-7B] N - N - N - Y - Y [LLaVA-OneVision-7B] N - N - N - N - Y [LongVU-7B] N - N - N - N - Y [Flash-Vstream-7B] Y - N - Y - Y - Y</p>	<p>[Prompt] You're a helpful assistant proficient in video question-answering. You're responsible of answering questions based on the video content. The following question are relevant to the latest frames, i.e. the end of the video. [Question] Decide whether existing visual content, especially latest frames, i.e. frames that near the end of the video, provide enough information for answering the question. Return "Yes" if existing visual content has provided enough information; Return "No" otherwise.</p> <p>[Examples] [Question] The woman in the black coat walks towards the direction of the man in yellow, what action does she do with the man? [Ground Truth] N - N - Y - Y - Y [Clues Reveal Time] 5:17 [Query Times] (5:10) - (5:13) - (5:19) - (5:27) - (5:47)</p> <p>[Response] [Gemini-1.5 Pro] N - N - N - N - N [GPT-4o] N - N - N - Y - N [Qwen-VL-72B] N - N - N - N - N [Qwen-VL-7B] N - N - N - N - N [LLaVA-NeXT-Video-7B] N - Y - N - N - Y [LLaVA-OneVision-7B] Y - Y - Y - Y - Y [LongVU-7B] N - N - N - Y - N</p>

Figure 8. **Prompts used for Online(up) and Offline(down) Models on Forward Active Responding and Response Examples.** Despite our vision for online models, existing online models, like videollm-online, are still far from satisfactory, showing limited adaptation ability, and would easily encounter collapse when processing complicated or out-of-training-domain video and queries. Offline models are inclined to perform random guessing when the queries contain words like "is/currently/ongoing".

video contents from t_0 to t_i offer sufficient clues. Specifically, for each of the tasks under the *Forward Active Responding* setting, instructional prompts and examples of models' [6][59] responses are shown in Fig. 8.

7. More Details of Benchmark Construction

7.1. Human-annotated QA Generation

We leverage meticulous human labor for part of the QA generation.

Real-Time Visual Perception. For tasks, including [STU], [OJR], and [ATR], we invite volunteers to propose candidate questions in supplement to our Video-LLMs-based automatic generation pipeline. This procedure is designed to alleviate possible bias and increase diversity. Specifically, we provide our volunteers with the following guidelines:

- Watch the video and decide whether this candidate is appropriate for constructing questions that can be classified into the above three types.

911
912
913
914

915 6.3. Prompt Design for Models on Backward Tracing and Real-Time Visual Perception

916
917 We use the clip from the beginning to the query time to
918 query models. Prompts and examples of models' responses
919 are shown in Fig. 9.

920
921
922
923
924
925
926
927
928
929
930
931
932
933

6.3-1 Online Models on Backward Tracing and Real-Time Visual Perception		
Prompt Used & Response Examples		
[ACR] Action Recognition	[OCR] Optical Character Recognition	[ASI] Action Sequence Identification
<p>[Prompt] Question: [Question] Options: [Options] Answer with the option's letter from the given choices directly.</p> <p>[Examples] [Question] What is he doing? [Prompt] Question: What is he doing? Options: A: He is wiping something with a rag. B: He is hitting objects with a tool. C: He is inspecting an object closely. D: He is connecting the pipe to the interface. Answer with the option's letter from the given choices directly. [Query Time] 4:47 [GT] B(Time 3:21)->D(Time 1:43)->A(Time 4:47)</p> <p>[Response] [videollm-online] (4:47) You are inspecting the object closely.</p>	<p>[Prompt] Question: [Question] Options: [Options] Answer with the option's letter from the given choices directly.</p> <p>[Examples] [Question] What name and number are visible on the back of this person's jacket? [Prompt] Question: What name and number are visible on the back of this person's jacket? Options: A: LAFFONT, 00. B: TOM, 21. C: MOTEEA, 18. D: GUS, 83. Answer with the option's letter from the given choices directly. [Query Time] 7:46 [GT] D(Time 1:43)->C(Time 5:09)->A(Time 7:46)</p> <p>[Response] [videollm-online] (7:46) According to the information provided, the name and number on the back of the person's jacket are "LAFFONT, 00".</p>	<p>[Prompt] Question: [Question] Options: [Options] Answer with the option's letter from the given choices directly.</p> <p>[Examples] [Question] Where did I put the shoe? [Prompt] Question: Where did I put the shoe? Options: A: Under the table. B: On the shelf. C: Shoes organizer at the back of the door. D: Unable to answer. Answer with the option's letter from the given choices directly. [Query Time] 7:00 [GT] D(Time 7:00) [Clues Time] 8:00</p> <p>[Response] [videollm-online] (7:00) You put the shoe on the shelf.</p>
6.3-2 Offline Models on Backward Tracing and Real-Time Visual Perception		
Prompt Used & Response Examples		
[ACR] Action Recognition	[OCR] Optical Character Recognition	[ASI] Action Sequence Identification
<p>[Prompt] Question: [Question] Options: [Options] Answer with the option's letter from the given choices directly.</p> <p>[Examples] [Question] What name and number are visible on the back of this person's jacket? [Prompt] Question: What name and number are visible on the back of this person's jacket? Options: A: LAFFONT, 00. B: TOM, 21. C: MOTEEA, 18. D: GUS, 83. Answer with the option's letter from the given choices directly. [Query Time] 4:47 [GT] B(Time 3:21)->D(Time 1:43)->A(Time 4:47)</p> <p>[Response] [Gemini-1.5 Pro] A [GPT-4o] A [Owen-VL-72B] D [Owen-VL-7B] B [LLaVA-NeXT-Video-7B] D [LLaVA-OneVision-7B] D [LongVU-7B] A [Flash-VStream-7B] D</p>	<p>[Prompt] Question: [Question] Options: [Options] Answer with the option's letter from the given choices directly.</p> <p>[Examples] [Question] Where did I put the shoe? [Prompt] Question: Where did I put the shoe? Options: A: Under the table. B: On the shelf. C: Shoes organizer at the back of the door. D: Unable to answer. Answer with the option's letter from the given choices directly. [Query Time] 7:00 [GT] D(Time 7:00) [Clues Time] 8:00</p> <p>[Response] [Gemini-1.5 Pro] A [GPT-4o] D [Owen-VL-72B] D [Owen-VL-7B] D [LLaVA-NeXT-Video-7B] A [LLaVA-OneVision-7B] C [LongVU-7B] A [Flash-VStream-7B] A</p>	<p>[Prompt] Question: [Question] Options: [Options] Answer with the option's letter from the given choices directly.</p> <p>[Examples] [Question] Where did I put the shoe? [Prompt] Question: Where did I put the shoe? Options: A: Under the table. B: On the shelf. C: Shoes organizer at the back of the door. D: Unable to answer. Answer with the option's letter from the given choices directly. [Query Time] 7:00 [GT] D(Time 7:00) [Clues Time] 8:00</p> <p>[Response] [Gemini-1.5 Pro] D [GPT-4o] D [Owen-VL-72B] D [Owen-VL-7B] D [LLaVA-NeXT-Video-7B] A [LLaVA-OneVision-7B] C [LongVU-7B] A [Flash-VStream-7B] A</p>

Figure 9. **Prompts used for Online(up) and Offline(down) Models on Real-Time Visual Perception and Response Examples.** Three tasks including [ACR], [OCR], and [ASI] are included as demonstrations. Our benchmarks involve a large ratio of questions, whose answers shift over time, which means that models can hardly figure out the answer by randomly selecting frames from original videos.

- 934 • Selected appropriate moments for problem construction.
935 Consider whether the moment contains: 1. Obvious spa-
936 tial relationships between several objects; 2. Interested
937 objects, such as something that appears in the moment,
938 and so on; 3. Objects with unusual attributes, e.g. green
939 fire, smooth woods.
- 940 • Construct options for the questions. Ensure that 1. Op-
941 tions should be relevant to the visual content; 2. Incorrect
942 options should bring misleading information from the vi-
943 sual content; 3. Options should be as close in length as
944 possible.

945 **Clue Reveal Responding.** For our novel [CRR] task, we
946 find it difficult to construct satisfactory question proposals
947 by straightly prompting Video-LLMs with original video
948 content as reference or LLMs with the provided scripts and
949 subtitles as reference. So we recruit volunteers to propose

queries and corresponding answers. Our guidelines for vol-
950 untees are as follows:

- Find scenes with apparent discontinuity. For example, character A performs a certain action at query time Q_i . However, the action's complete process or outcome is not immediately shown during query time.
- Continue watching the video, find clues for your query, and annotate the clues revealing time as A_i .
- Try to provide concise timestamps, let A_i be the time when enough visual information has just been revealed.

8. Additional Dataset Analysis

8.1. Task and Sample Distribution

Fig. 10 illustrates the distribution of questions and videos in OVBench across the twelve tasks listed in Fig. 2.

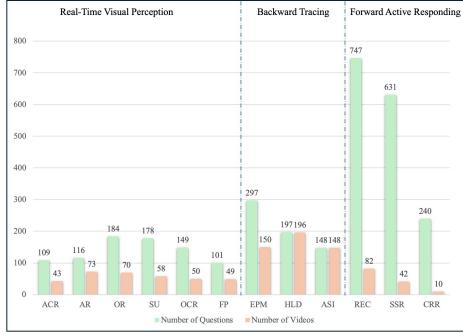


Figure 10. Distribution of questions and video in OVBench.

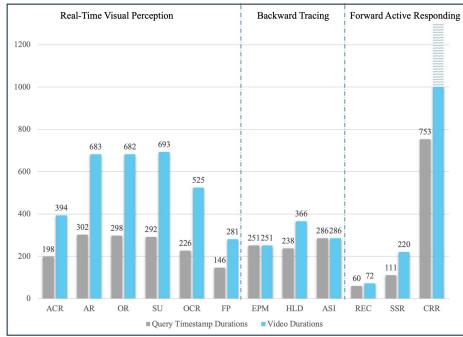


Figure 11. Distribution of averaged query timestamps and video duration (in seconds) in OVBench. Specifically, the averaged video duration in CRR is 6,857 seconds.

964 8.2. Query Timestamps and Video Duration

965 Fig. 11 illustrates the distribution of averaged query times-
966 tamps and video duration in OVBench across the twelve
967 tasks listed in Fig. 2.

968 9. Limitations

969 While we have tried hard to cover a wide range of rea-
970 sonable video domains and QA generation methods, the
971 scarcity of existing datasets with annotations that fit require-
972 ments, the unsatisfactory results of automatic QA genera-
973 tion, and the high human annotation cost, hinder diversity
974 and can cause potential bias.

975 **Offline Models for Online Video Understanding.** As
976 implied in our analysis 8, offline models usually perform
977 random guesses in the forward active responding scenarios,
978 making our evaluation unfair. For example, a model that
979 always outputs "Yes" can still achieve a score above zero
980 in our evaluation. Moreover, the absence of online mod-
981 els with satisfactory performance, makes our benchmarks
982 more suitable for future advancements. We hope our inten-
983 sive work and intuitive ideas can guide the development of
984 video understanding models toward real-world online video
985 understanding.

10. Licenses

The annotations of our OVBench are provided to the com-
987 munity under CC BY-NC-SA 4.0 license. By download-
988 ing our dataset from our website or other sources, the user
989 agrees to adhere to the terms of CC BY-NC-SA 4.0 and li-
990 censes of the source datasets. Download links are provided
991 for our self-crawled YouTube videos. Licenses of the source
992 datasets are listed in 2

Dataset	License
QAEGo4D [2]	N/A
OpenEQA [1]	MIT License
STAR [51]	Apache License 2.0
HiREST [57]	MIT License
YouCook2 [64]	MIT License
CrossTask [65]	BSD 3-Clause License
COIN [45]	Research Purpose Only
Ego4D [15]	MIT License
THUMOS'14 [21]	Research Purpose Only
THUMOS'15 [14]	Research Purpose Only
Perception Test [39]	CC BY 4.0
MovieNet [17]	N/A
E.T.Bench [32]	CC BY 4.0

Table 2. License of source datasets in OVBench.

11. Data Examples

We provide more examples extracted from our benchmark.
995 We try to cover different video categories in every task to
996 offer a holistic overview of OVBench.
997

[EPM] Episodic Memory

998

QA-Ego4D

Clue Time: 5:22
Question: Where was the kitchen paper towel before I picked it?
Options: A. top of the kitchen sink; B. under the sink;
C. on the counter; D. on the stove

Clue Time: 3:44
Question: What did I pick from the fridge?
Options: A. bread; B. milk; C. vegetable; D. water

Clue Time: 5:25
Question: How many rolls of paper towel did I cut?
Options: A. two; B. one; C. three; D. four

Clue Time: 6:36
Question: Did I leave the drawer open?
Options: A. yes; B. no

Query At the End

Clue Time: 1:59
Question: Where was the paper towel before I picked it?
Options: A. on the counter; B. on the cupboard;
C. on the table; D. in the dish washer

Clue Time: 3:05
Question: What drawer did I pull?
Options: A. dish washer; B. microwave;
C. cupboard; D. fridge

Clue Time: 4:09
Question: What did I pour in the container?
Options: A. flour; B. bread crumbs; C. salt; D. pepper

Clue Time: 3:13
Question: Where was the plate before I picked it?
Options: A. on the counter; B. on the cupboard
C. on the table; D. in the dish washer

Query At the End

Open-EQA

Clue Time: 0:43
Question: What is the gold object on the nightstand?
Options: A. A painting; B. A mirror;
C. A nightlamp; D. A vase

Clue Time: 0:50
Question: Is this home on the first floor?
Options: A. Yes, it's on the first floor;
B. No, it's on the second floor;

Clue Time: 1:10
Question: Where can I sit and eat if I don't want to use the dining table?
Options: A. Use the kitchen bar counter; B. Use the floor in the hallway;
C. Use the bed in the bedroom; D. Use the couch in the living room

Clue Time: 1:10
Question: What color is the smoke detector?
Options: A. White; B. Yellow or off-white;
C. Black; D. Blue

Query At the End

[HLD] Hallucination Detection

QA-Ego4D

Query Time: 6:11
Question: what did I put in the black dustbin?
Options: A. empty water bottles; **B. Unable to answer;**
C. old newspapers; D. food scraps

Query Time: 6:44
Question: Where did I put the vacuum cleaner head?
Options: A. closet; **B. Unable to answer;**
C. bathroom; D. kitchen

999



Clue Time: 7:10

Clue Time: 7:30

Query Time: 6:48
Question: Where were game boards?
Options: **A. Unable to answer;** B. in the shelves;
C. in the fridge; D. in the bags



Clue Time: 7:55

Open-EQA

Query Time: 0:35
Question: what color is the flower in the bottom floor?
Options: A. Red; **B. Unable to answer;**
C. Blue; D. White



Clue Time: 1:15

[ASI] Action Sequential Identification

1000

STAR

Clue Time: 0:02

Reference Time: 0:10



Query Time: 0:27/End of the video

Question: What happened before the person took the pot?

- Options: A. close the refrigerator.; B. Unable to answer.;
C. Took the box.; D. Throw the broom.; E. Open the book.

Reference Time: 0:10

Clue Time: 0:02



Query Time: 11:46/End of the video

Question: What does the person do after cover up and cook for 6 to 8 minutes?

- Options: A. chop up the tomatoes.; B. chop up oleander.;
C. add more garam masala powder.; D. add cloves to the pot

YouCook2

Clue Time: 0:02

Reference Time: 0:10



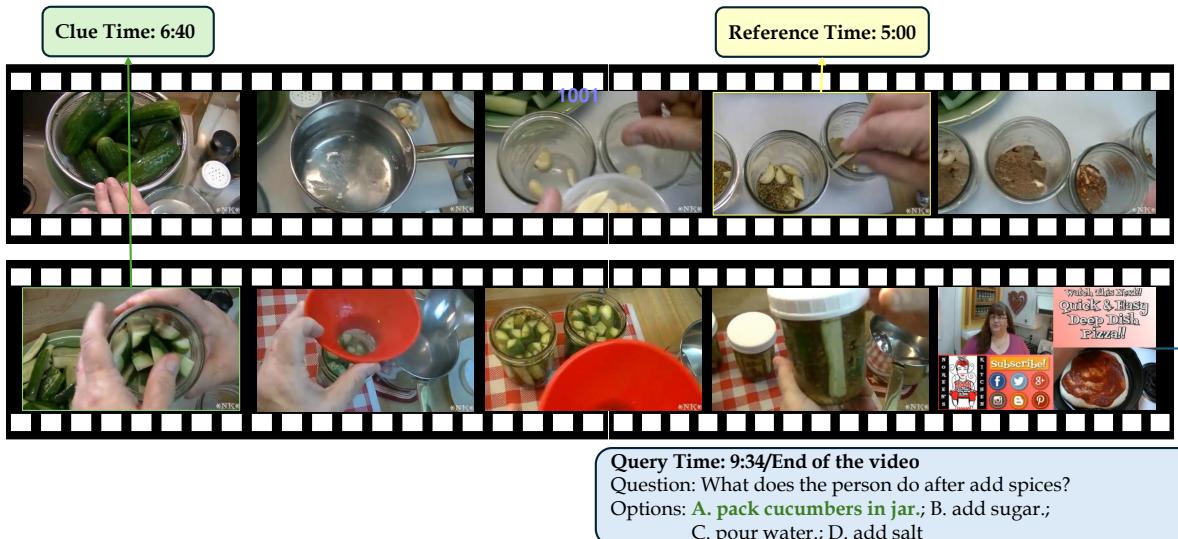
Query Time: 9:03/End of the video

Question: What does the person do before roughly chop garlic and peppers and add them to the bowl?

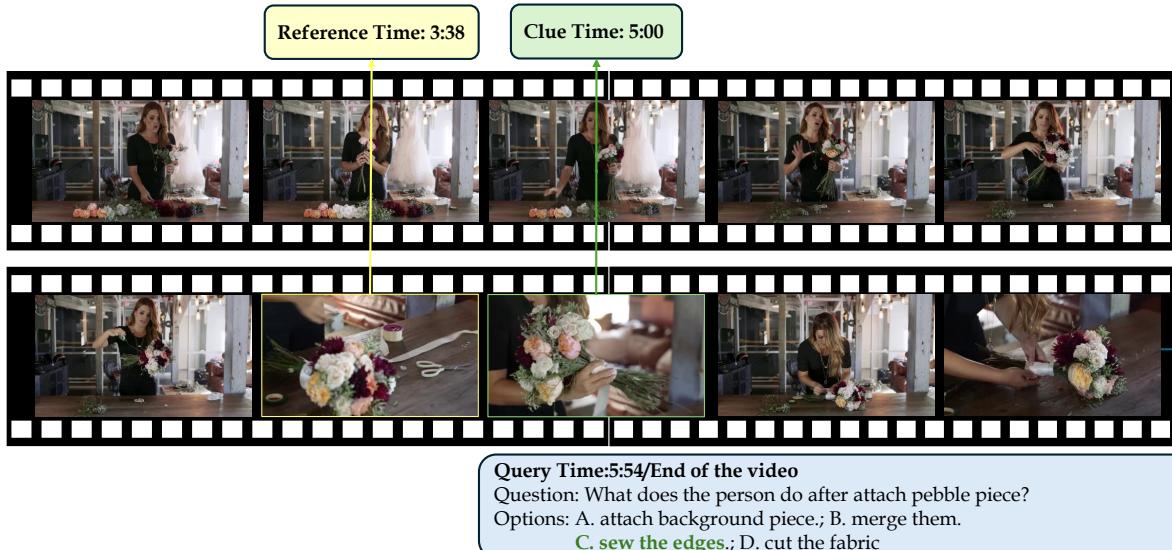
- Options: A. pour some olive oil on the kabob.; B. pour olive oil on shrimp.;
C. cut up some onions and peppers into squares.; D. skewer the vegetables and shrimps

[ASI] Action Sequential Identification

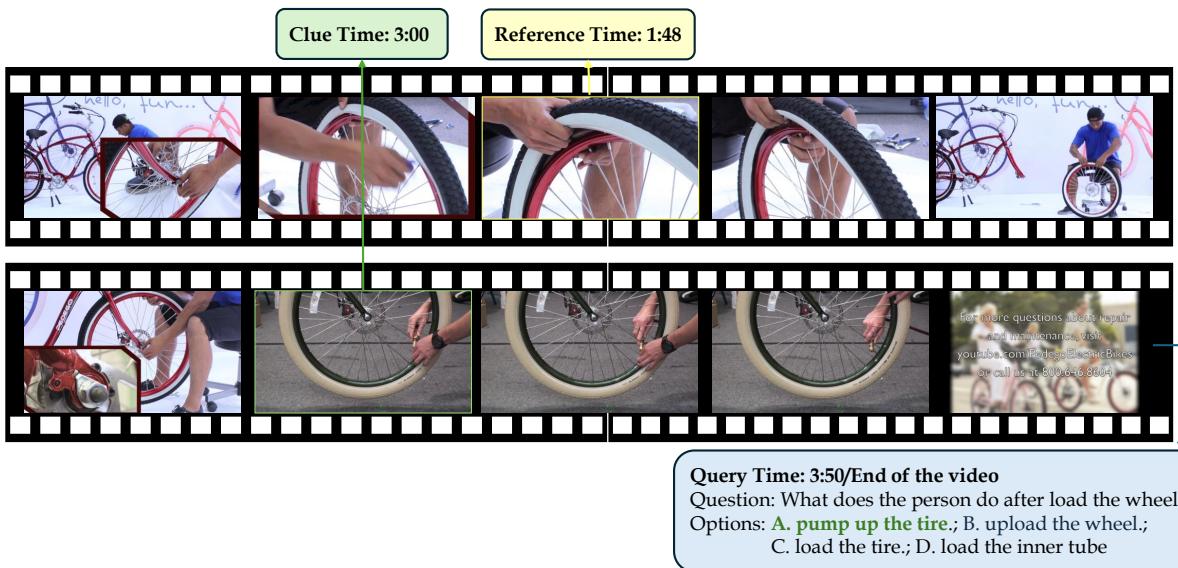
CrossTask



HiREST



COIN



[STU] Spatial Understanding

1002

Ego4D

Query Time/Clue Time: 1:27

- Question: What is the relative position of the person to the car ?
Options: A. The person is standing at the front of the car.;
B. The person is on the co-pilot side of the car;,
C. The person is standing beside the driver's side of the car ;
D. The person is behind the trunk of the car.



Query Time/Clue Time: 1:47

- Question: Which container is located closer to the top left corner of the table?
Options: A. The blue container;
B. The white container;
C. The red container;
D. The green container.

YouTube



Query Time/Clue Time: 0:12

- Question: Which road did I take?
Options: **A. The road on the right;**
B. The road on the left;
C. Unable to answer.

YouTube-Human Annotated



[OJR] Object Recognition

Ego4D

Query Time/Clue Time: 2:08

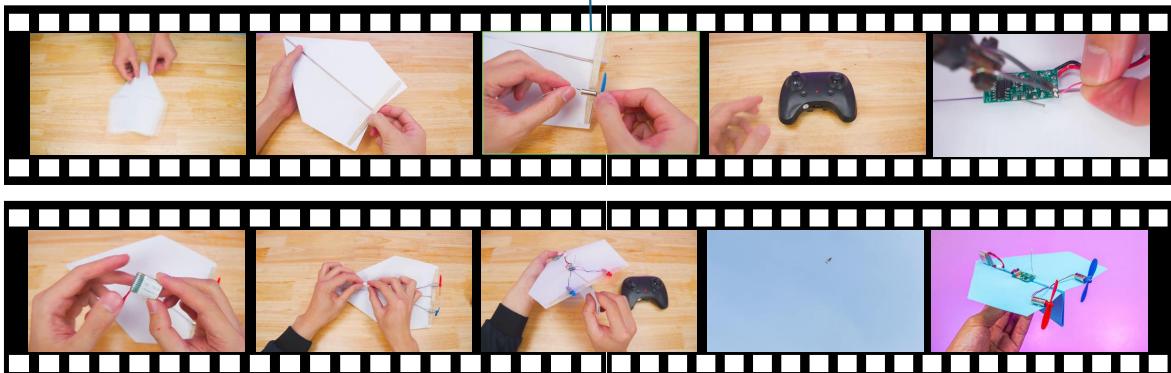
Question: What object is being used to construct the paper aircraft wing?

Options: A. A plastic frame; ;

B. A wooden stick..; 1003

C. A paper sheet;

D. A metal rod.



YouTube

Query Time/Clue Time: 1:27

Question: What object is being attached to the back of it?

Options: A. A bag; ;

B. A saddlebag; ;

C. A basket;

D. A pannier.



YouTube-Human Annotated

Query Time/Clue Time: 4:28

Question: What weapon do I have in my hand?

Options: **A. Bow;** B. Sword;

C. Unable to answer.

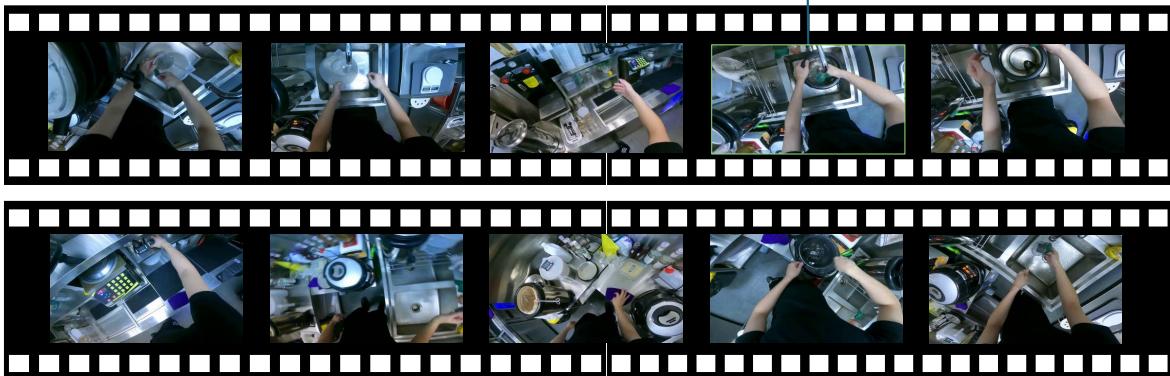
D. Axe. E. Spear



[ATR] Attribute Recognition

1004

Ego4D



YouTube



YouTube-Human Annotated



[ACR] Action Recognition

Ego4D



Query Time/Clue Time: 4:18

Question: What action is he performing with the blue checkered cloth?

- Options:
- A. He is wiping the tabletop with the cloth.;
 - B. He is folding the cloth.;**
 - C. He is covering a basket with the cloth;
 - D. He is tying the cloth around his neck.

Query Time/Clue Time: 1:03

Question: What is he doing?

- Options:
- A. He is playing a musical instrument.;
 - B. He is taking off his clothes.;
 - C. He is taking a shower;
 - D. He is answering the phone.**



YouTube

Query Time/Clue Time: 2:37

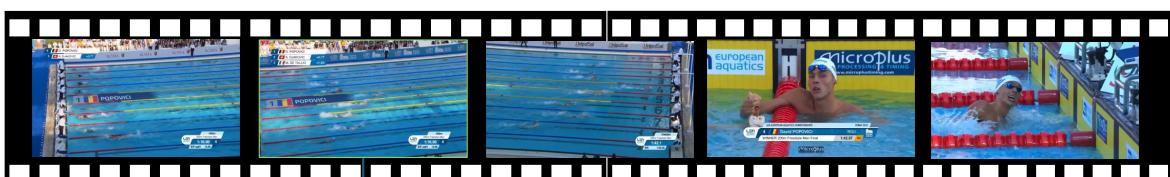
Question: What is she doing with the chicken in her hands?

- Options:
- A. She is placing the chicken into a pot of water.;
 - B. She is putting the chicken down.;**
 - C. She is holding the chicken while preparing to cook it.;
 - D. She is cutting the chicken into smaller pieces..



[OCR] Optical Character Recognition

1006



Query Time/Clue Time: 2:16

Question: What is the leading player's time at the 150m mark?
Options: A. 1:16.96 B. 1:15.89;
C. 1:14.32. D. 1:19.01

Query Time/Clue Time: 0:23

Question: What is the text on the package?
Options: A. ANDOUILLE.;
B. THICK CUT HAM.;
C. THICK CUT PORK;
D. THICK CUT BACON.

Query Time/Clue Time: 1:35

Question: What is the text on the package?
Options: A. ANDOUILLE.;
B. THICK CUT HAM.;
C. THICK CUT PORK;
D. THICK CUT BACON.

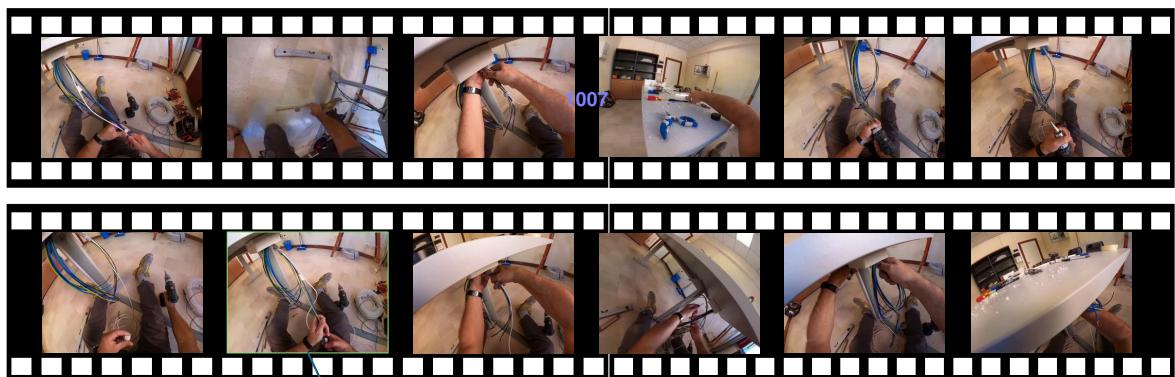


Query Time/Clue Time: 10:12

Question: What text is displayed now?
Options: A. Milk Bucket.;
B. Butter;
C. Peanut Butter Cup.;
D. Peppermint Swirl.

YouTube

[FPD] Future Prediction



Query Time/Clue Time: 1:45

Question: What is the person preparing to manipulate?

- Options:
- A. The person is about to handle or manipulate the wire;**
 - B. The person is about to handle or manipulate the plastic tubing;
 - C. The person is about to handle or manipulate the circuit board;
 - D. The person is about to handle or manipulate the metal sheet

Query Time/Clue Time: 0:59

Question: What action is this person preparing to take ?

- Options:
- A. The person is about to turn on the faucet;**
 - B. The person is preparing to start the car engine;
 - C. The person is about to open the refrigerator door;
 - D. The person is reaching for the light switch.



Query Time/Clue Time: 0:59

Question: What is this person about to do?

- Options:
- A. The person is about to press a button.;
 - B. The person is about to grab a book;
 - C. The person is about to open the drawer;**
 - D. The person is about to adjust the lamp.



Ego4D

[REC] Repetition Event Count

1008

Thumos 14/15

Query Time: 0.26/1.53/End of the video
Question: How many times did they clean and jerk?
Answers: 1/2/2



Clue Time: 0:01-0.26

Clue Time: 0:27-1.53

Clue Time: 0:00-0.06

Clue Time: 0:07-0.14



Clue Time: 0:17-0.24

Query Time: 0.06/0.14/0.24/0.34/End of the video
Question: How many times did they showing something to the camera?
Answers: 1/2/3/4/4

Clue Time: 0:25-0.34

Perception Test

Clue Time: 0:11-0.13

Clue Time: 0:13-0.14



Clue Time: 0:14-0.15

Clue Time: 0:15-0.17

Query Time: 0.06/0.13/0.14/0.15/0.17/0.27/End of the video
Question: How many times did they take something out of something?
Answers: 0/1/2/3/4/4/4

[SSR] Sequential Steps Recognition



[CRR] Clues Reveal Responding

MovieNet

Query Time: 2:25

Question: Women came out of doors of different colors and went into the center door. What is the purpose of doing so?

Clue Time: 3:25

Answer: To listen to the older woman talking



1010



Query Time: 2:25

Question: The man picked up several books from the ground. What does the man do with the books he picked up?

Clue Time: 3:25

Answer: He handed these books to the woman

Query Time: 4:02

Question: A policeman is driving the car away, what is his destination

Clue Time: 5:15

Answer: A residential house with a woman inside.



Query Time: 10:19

Question: The policeman is stepping into a wooden house, what does the police see?

Clue Time: 10:26

Answer: A black man sitting on a bench.

Query Time: 3:23

Question: The woman with the gray dress is stepping into the room, what does she do in the room?

Clue Time: 3:27

Answer: She talks to the man



Query Time: 8:07

Question: The couple are sitting on their trunks, who do they meet then

Clue Time: 8:28

Answer: A group of friends