

CAVIAR: CRITIC-AUGMENTED VIDEO AGENTIC REASONING SUPPLEMENTAL MATERIAL

Anonymous authors

Paper under double-blind review

1 MODULE API

The full module API can be seen in 1. The modules included vary by the strategy being sampled.

```
@dataclasses.dataclass
class VideoSegment:
    """Class containing a segment of a video, defined by start and end
    times as MM:SS strings."""
    start: str
    end: str

def think(thought: str) -> None:
    """Tool to perform intermediate reasoning that doesn't directly look
    at the video.

    Args:
        thought (str): the thought to print.

    Returns:
        None
    """

def get_segment(start: str, end: str) -> VideoSegment:
    """Clips the segment of the input video between the times indicated
    by 'start' and 'end'.

    Args:
        start (str): the start time for the segment as an MM:SS string.
        end (str): the end time for the segment as an MM:SS string.

    Returns:
        A VideoSegment made from the input video from 'start' to 'end'.
    """

def find_when(query: str, video_segment: VideoSegment | None) -> str:
    """Tool to determine timing of a query such as an event or action.
    Tries to find time ranges in the video that may correspond to the
    query. Works best for simple queries.
    The reasoning may be faulty, so you must carefully consider each
    time range and determine if it is relevant to best answering the
    question.

    Args:
        query (str): the event or action to localize.
        video_segment (VideoSegment): if specified, only looks at the
        given segment of the input video.

    Returns:
        A string of time ranges identified as potentially relevant and the
        justification for each time range.
    """
```

```

054 def asr_understanding(question: str, answer_options: list[str] | None)
055 :
056     """Tool to understand the ASR transcript of the video. Considers the
057     ASR transcript of the full video and tries to answer the
058     question.
059     If the question cannot be answered from the transcript, returns any
060     potentially relevant information or timestamps based on the
061     transcript.
062     Args:
063         question (str): the question to answer.
064         answer_options (list[str]): Optional. If the set of possible
065         answers is known, they should be specified here as a list.
066     Returns:
067         A string containing the predicted answer to the question.
068     """
069
070 def retrieval_qa(question: str, answer_options: list[str] | None,
071                 video_segment: VideoSegment | None) -> str:
072     """Tool to answer a question from frames selected independently from
073     each minute of the video.
074     Looks at each minute of the video and tries to identify relevant
075     frames.
076     Then tries to answer the question using the retrieved frames.
077
078     Args:
079         question (str): the question to answer.
080         answer_options (list[str]): Optional. If the set of possible
081         answers is known, they should be specified here as a list.
082         video_segment (VideoSegment): if specified, only looks at the
083         given segment of the input video.
084
085     Returns:
086         A string containing the predicted answer to the question.
087     """
088
089 def finish(final_answer: str) -> str:
090     """Print final answer and exit.
091     After explaining your reasoning, output the final answer in the
092     format "Final Answer: (X)" where X is the correct digit choice.
093     Never say "unknown" or "unsure", or "None", instead provide your
094     most likely guess.
095
096     Args:
097         final_answer (str): reasoning followed by final answer.
098
099     Returns:
100         The parsed final answer.
101     """

```

Listing 1: Module API.

2 MODULE IMPLEMENTATION DETAILS

VideoSegment. This is a convenience class for identifying segments of the input video by their start and end time.

get_segment. This module selects the appropriate segment from the input video given start and end times, returning a **VideoSegment** corresponding to that range.

find_when. This corresponds to coarse temporal localization with a sliding window. Using a window of 100 frames, the model is instructed to identify any time ranges that could be relevant for the given query, ultimately providing all of these ranges along with justifications for each of them as a string.

asr_understanding. This module uses the ASR transcript of all the speech in the video with timestamps aligned for the words said at each second. If the query is a question that can be answered from the transcript, it tries to provide an answer. If not, it aims to provide any relevant information along with timestamps from the ASR given.

retrieval_qa. This module tries to use the base model to retrieve frames that appear visually relevant to a query, then tries to answer the query from them. It uses a sliding window of 64 frames, returning the indices of frames with high visual relevance from each window. Those frames are then retrieved and provided back to the base model, along with 56 uniform frames from the rest of the video for additional context, to produce a response to the query.

finish. This module indicates the agent has reached a final answer and ends the agentic inference procedure, returning the final answer.

```
You are a video understanding agent. Your task is to answer the
question provided using the available tools. These tools are
already imported and accessible to you. You will receive the
question and answer choices. You should use the tools iteratively
, deciding which tool to use next based on the results of the
previous tool call. Each turn, you should write a call to one tool
, enclosed in triple backticks. Continue this process until you
can confidently answer the question, selecting one of the provided
answer choices. Do not import any tools; they are already
available. Your final response should be one of the provided
answer choices.
```

Step by Step Instructions

1. ****Analyze the question:**** Carefully read the provided ‘question’ to understand the specific information required. Identify key entities, actions, and relationships mentioned in the question.
2. ****Select an initial tool:**** Based on your understanding of the question from step 1, choose the most appropriate tool from the available set of video interaction functions. Consider which tool will provide the most relevant information to answer the question efficiently.
3. ****Execute the selected tool:**** Call the chosen tool, providing any necessary parameters derived from the question analysis in step 1.
4. ****Analyze the tool’s output:**** Examine the results returned by the tool. This output might be textual descriptions, numerical data, or other relevant information about the video.
5. ****Evaluate progress:**** Based on the tool’s output from step 4, assess whether you have enough information to confidently answer the question. If yes, proceed to step 6. If not, go back to step 2 and select a different tool or use the same tool with different parameters based on the insights gained from the previous tool call. Consider how the previous tool’s output informs your choice of the next tool.
6. ****Formulate the final answer:**** Using the information gathered from all tool calls, select one of the provided answer choices that best answers the ‘question’. Clearly state your chosen answer choice.

Tools:

Listing 2: The preamble used for the reasoning agent.

3 STRATEGY SAMPLING DETAILS

For each dataset, 3 strategies are selected, each corresponding to a different subset of modules. These correspond to the strategies that can feasibly produce an answer for the given task, as well as the information allowed for the dataset (not including the base ‘think’ and ‘finish’ tools). For LVBench (and EgoSchema), the strategies are `get_segment` and `retrieval_qa`; `retrieval_qa` directly; and `get_segment`, `retrieval_qa`, and `find_when` (that is, all modules in use). For Neptune, as audio information is permitted, the same strategies are used with the addition of the `asr_understanding` module for the multi-module strategies. For the RTL task, the strategies are `get_segment` and `find_when`; `find_when` only; and `get_segment`, `retrieval_qa`, and `find_when` (again, all modules in use).

4 CRITIC PROMPTS

The critic is given the following preamble.

```
You are an expert in assessing and critiquing reasoning about videos.
Given a question and possible strategies from video reasoning
agents for solving it, assess which strategy seems the most likely
to result in a correct answer based on the provided examples.
If multiple strategies produce the same final answer, you may pick
multiple; you may not pick multiple answers with conflicting final
answers.
Observe that the reliability of the tools depends on their input; you
must consider the examples to determine when a given tool is
likely to hallucinate or give an accurate output, as evidenced by
the ultimate result.
Produce a critique per the examples followed by "Winning Strategies:"
followed by a comma separated list of the letters corresponding to
those strategies.
```

Listing 3: The preamble used for the critic.

The critic is given 4 in-context examples per prompt. These examples are written by the authors with the aim of being in-distribution for the datasets shown.

```
Input:
You will be given a question about a video and four possible answer
options. Question: At 12:10, what color suit is the host wearing
?Possible answer choices:
(1) Navy
(2) Tan
(3) Black
(4) White
Video length: 40:50

Strategy A (retrieval_qa, get_segment):
'''
input_video = get_segment(start='00:00', end='40:50')
'''

'''
segment = get_segment(start='12:10', end='12:20')
answer = retrieval_qa(question='What color suit is the host wearing?'
', video_segment=segment)
'''
The host is wearing a navy suit.

'''
final_answer = finish(f"The retrieval_qa tool determined the host
was wearing a navy suit at 12:10. Final Answer: (1)")
'''
```

```

216 The retrieval_qa tool determined the host was wearing a navy suit at
217 12:10. Final Answer: (1)
218
219 Strategy B (direct retrieval_qa):
220 Here's how to solve this:
221
222 1. **Locate the timestamp:** Go to the 12:10 mark in the video.
223
224 2. **Observe the host's attire:** At this point, the host (Jimmy
225 Kimmel) is wearing a dark suit and a dark-colored tie.
226
227 3. **Compare to options:** The suit appears to be black.
228
229 Therefore, the correct answer is (3) Black.
230 Final Answer: (3)
231
232 Strategy C (retrieval_qa, get_segment, find_when):
233 '''
234 input_video = get_segment(start='00:00', end='40:50')
235 '''
236
237 '''
238 segment = get_segment(start='12:10', end='12:20')
239 '''
240 ['12:10', '12:20']
241
242 '''
243 answer = retrieval_qa(question='At 12:10, what color suit is the
244 host wearing?', answer_options=['Navy', 'Tan', 'Black', 'White',
245 ], video_segment=segment)
246 '''
247 The image at 12:10 shows the host wearing a navy suit.
248
249 Final Answer: (1)
250
251 '''
252 final_answer = finish(final_answer='The image at 12:10 shows the
253 host wearing a navy suit. Final Answer: (1)')
254 '''
255 The image at 12:10 shows the host wearing a navy suit. Final Answer:
256 (1)
257
258 Critique:
259 A: 1. B: 3. C: 1. Strategies A and C succeed while Strategy B fails
260 because Strategies A and C successfully use the get_segment tool
261 to isolate the relevant part of the video.
262 Strategy B fails as the retrieval_qa tool alone cannot actually
263 locate a given timestamp in a video, resulting in a hallucinated
264 answer.
265
266 Winning Strategies:
267 A, C
268
269 Input:
270 You will be given a question about a video and four possible answer
271 options. Question: What does the man in charcoal suit do in the
272 next scene after the scene of a photo of six people end?Possible
273 answer choices:
274 (1) He shows his scars
275 (2) He takes off his shirt
276 (3) He shows his jewelry
277 (4) He shows his physique
278 Video length: 39:59

```

```

270 Strategy A (retrieval_qa, get_segment):
271 '''
272 input_video = get_segment(start='00:00', end='39:59')
273 '''
274
275 '''
276 think("I need to identify the scene where a photo of six people is
277       shown. Then I need to find the scene immediately following it
278       to see what the man in the charcoal suit does.")
279 '''
280 I need to identify the scene where a photo of six people is shown.
281 Then I need to find the scene immediately following it to see
282 what the man in the charcoal suit does.
283
284 '''
285 segment_with_photo = get_segment(start='00:00', end='01:00')
286 answer = retrieval_qa(question="What does the man in charcoal suit
287 do in the scene immediately after the scene with a photo of six
288 people?", video_segment=segment_with_photo)
289 '''
290 After the scene with the photo of six people, the man in the
291 charcoal suit walks down a hallway and is greeted by a group of
292 people who clap and cheer for him.
293
294 '''
295 think("The provided text doesn't describe any of the options (1) He
296 shows his scars, (2) He takes off his shirt, (3) He shows his
297 jewelry, (4) He shows his physique. The man is celebrated. I
298 will guess the closest option, which is none of them, but I will
299 choose the least violent option.")
300 '''
301 The provided text doesn't describe any of the options (1) He shows
302 his scars, (2) He takes off his shirt, (3) He shows his jewelry,
303 (4) He shows his physique. The man is celebrated. I will
304 guess the closest option, which is none of them, but I will
305 choose the least violent option.
306
307 '''
308 final_answer = finish("Final Answer: (1)")
309 '''
310 Final Answer: (1)
311
312 Strategy B (direct retrieval_qa):
313 Here's how to solve this step-by-step:
314
315 1. **Identify the relevant scene:** The question refers to the scene
316 after a photo of six people. Looking through the timestamps,
317 this appears to be around 20:49.
318
319 2. **Observe the next scene:** The next scene shows the man in the
320 charcoal suit (Harry Styles) pointing to and showing jewelry on
321 his arm.
322
323 3. **Match to the answer choices:** This action directly corresponds
324 to answer choice (3).
325
326 Therefore, the correct answer is **(3) He shows his jewelry**.
327 Final Answer: (3)
328
329 Strategy C (retrieval_qa, get_segment, find_when):
330 '''
331 input_video = get_segment(start='00:00', end='39:59')
332 '''
333

```

```

'''
think("First, I need to find when the scene with the photo of six
people ends. Then I can examine the next scene to see what the
man in the charcoal suit does.")
'''
First, I need to find when the scene with the photo of six people
ends. Then I can examine the next scene to see what the man in
the charcoal suit does.

'''
photo_end_time = find_when(query="scene with a photo of six people",
video_segment=None)
'''
["00:05", "00:24"]: The query asks to find the scene with a photo of
six people. From the provided frames, a photo of six people is
visible on the wall in the background starting at 00:05 and
remaining visible until 00:24. The photo is partially obscured
at times by other people and objects, but it is consistently
present in the background during this timeframe.
["10:39", "10:43"]: The query asks for a scene with a photo of six
people. At 10:39 through 10:43 a large group photo of more than
six people is shown.
["17:47", "17:53"]: The query asks for a scene with a photo of six
people. At 17:47-17:53, a photo of approximately six people is
visible behind a curtain as Harry Styles walks out onto the
stage. The photo is partially obscured and the exact number of
people is difficult to ascertain with certainty, but it appears
to show at least five.
["18:47", "19:59"]: The query asks for the time range of a scene
with a photo of six people. A photo of what appears to be six
people is visible on the wall behind the main set in frames
starting at 18:47 and continuing through 19:59. The photo is
partially obscured at times by other elements in the scene, but
it is consistently visible enough to be identified.
["20:03", "20:04"]: The query asks for the time range of a scene
with a photo of six people. At 20:03, a wide shot of the studio
audience is shown, and a photo of six people is visible on a
screen in the background. This photo remains visible until
approximately 20:04.
["23:15", "23:25"]: The query asks for the timestamp of a scene with
a photo of six people. In the provided video frames, a photo
of six people is visible on the wall in the background starting
at 23:15 and remaining visible until 23:25. The photo is
partially obscured at times by other elements in the frame, but
it is consistently present during this time range.
["35:17", "35:20"]: The provided video frames show a wide shot of a
studio audience at the 35:17 mark. The audience appears to
consist of more than six people. While the exact number is
difficult to determine from the provided frames, it is clearly a
large group. Therefore, this time range satisfies the query.
["39:03", "39:18"]: The provided video frames show a live
performance with an audience. At the end of the performance,
from 39:03 to 39:18, the camera focuses on the audience, and
there are enough people visible to potentially include a group
of five. The resolution isn't high enough to definitively
confirm the presence of a specific photo of six people, but the
scene depicts a large group of people, making it a plausible
match for the query.

'''
think("The most likely candidate for the 'scene with a photo of six
people' is the longest continuous segment: 18:47-19:59. Let's
use that as the end point for the previous scene.")
'''

```

```

378 The most likely candidate for the 'scene with a photo of six people'
379 is the longest continuous segment: 18:47-19:59. Let's use that
380 as the end point for the previous scene.
381
382 '''
383 next_scene_segment = get_segment(start='19:59', end='20:05')
384 '''
385 ['19:59', '20:05']
386
387 '''
388 answer = retrieval_qa(question="What does the man in the charcoal
389 suit do in the scene from 19:59 to 20:05?", answer_options=['He
390 shows his scars', 'He takes off his shirt', 'He shows his
391 jewelry', 'He shows his physique'], video_segment=
392 next_scene_segment)
393 '''
394 The provided video frames show a talk show segment featuring three
395 men. From 19:59 to 20:02, the man in the gray suit (Harry
396 Styles) sits on a couch and gestures with his hands. At 20:03,
397 the camera angle changes to show the audience, and the three men
398 are still on the couch. There is no indication that the man in
399 the gray suit shows his scars, takes off his shirt, shows a
400 tattoo, or shows his muscles. He is simply participating in a
401 conversation.
402
403 Final Answer: (4)
404
405 '''
406 final_answer = finish("Based on the analysis of the video segment
407 from 19:59 to 20:05, the man in the charcoal suit (assuming this
408 refers to one of the men in the talk show segment, as there's
409 no charcoal suit explicitly mentioned) does not perform any of
410 the actions listed in the options. However, since an answer must
411 be chosen, and the retrieval_qa tool did not provide a
412 definitive answer, I will select the option that is most likely
413 to be present in a talk show setting, which is showing his
414 physique. This is a weak inference, but given the constraints,
415 it's the best guess. Final Answer: (4)")
416 '''
417 Based on the analysis of the video segment from 19:59 to 20:05, the
418 man in the charcoal suit (assuming this refers to one of the men
419 in the talk show segment, as there's no charcoal suit
420 explicitly mentioned) does not perform any of the actions listed
421 in the options. However, since an answer must be chosen, and
422 the retrieval_qa tool did not provide a definitive answer, I
423 will select the option that is most likely to be present in a
424 talk show setting, which is showing his physique. This is a
425 weak inference, but given the constraints, it's the best guess.
426 Final Answer: (4)
427
428 Critique:
429 A:1. B: 3. C: 4. Strategy B succeeds while Strategies A fails
430 because Strategy A incorrectly applies the get_segment tool to
431 the first minute of the video and only considers that, resulting
432 in it not being able to find a correct answer and ultimately
433 guessing. Strategy C applies the find_when tool, then trims a
434 segment and does not find the relevant information, then
435 guessing as well.
436
437 Winning Strategies:
438 B
439
440 Input:
441

```

```

432 You will be given a question about a video and four possible answer
433 options. Question: What is the main reason the vlogger goes out
434 again, where she also buys makeup?Possible answer choices:
435 (1) To go exercise
436 (2) To eat chicken
437 (3) To buy coffee mugs
438 (4) To go to the park
439 Video length: 30:25
440
441 Strategy A (retrieval_qa, get_segment):
442 '''
443 input_video = get_segment(start='00:00', end='30:25')
444 '''
445
446 think("I need to find a segment of the video where the vlogger buys
447 makeup. I'll try to identify this segment using retrieval_qa,
448 and then I can use that information to answer the question about
449 why she went out.")
450 '''
451 I need to find a segment of the video where the vlogger buys makeup.
452 I'll try to identify this segment using retrieval_qa, and then
453 I can use that information to answer the question about why she
454 went out.
455 '''
456 answer = retrieval_qa(question="When does the vlogger buy makeup?",
457 answer_options=None, video_segment=input_video)
458 '''
459 The vlogger buys makeup at 16:10-16:11 and 17:20-17:25.
460 '''
461 think("The vlogger buys makeup in two segments: 16:10-16:11 and
462 17:20-17:25. I'll use retrieval_qa again, focusing on these
463 segments, to determine the reason for her outing.")
464 '''
465 The vlogger buys makeup in two segments: 16:10-16:11 and
466 17:20-17:25. I'll use retrieval_qa again, focusing on these
467 segments, to determine the reason for her outing.
468 '''
469 segment1 = get_segment(start='16:10', end='16:11')
470 segment2 = get_segment(start='17:20', end='17:25')
471 reason = retrieval_qa(question="What is the main reason the vlogger
472 goes out again, where she also buys makeup?", answer_options=['
473 To go exercise', 'To eat chicken', 'To buy coffee mugs', 'To go
474 to the park'], video_segment=segment1)
475 if reason == "": #check if the first segment didn't provide enough
476 information
477 reason = retrieval_qa(question="What is the main reason the
478 vlogger goes out again, where she also buys makeup?",
479 answer_options=['To go exercise', 'To eat chicken', 'To buy
480 coffee mugs', 'To go to the park'], video_segment=segment2)
481 '''
482 None
483 '''
484 final_answer = finish("Based on the analysis of the video segments
485 where the vlogger buys makeup, I was unable to definitively
determine the main reason for her outing from the provided
options. However, given the context of beauty product purchases,
it's less likely she's going to the gym, eating salmon, or
going to the park. Therefore, I will eliminate those options.
Final Answer: (3)")

```

```

486     '''
487     Based on the analysis of the video segments where the vlogger buys
488     makeup, I was unable to definitively determine the main reason
489     for her outing from the provided options. However, given the
490     context of beauty product purchases, it's less likely she's
491     going to the gym, eating salmon, or going to the park. Therefore
492     , I will eliminate those options.  Final Answer: (3)
493
494     Strategy B (direct retrieval_qa):
495     Here's how to solve this step-by-step:
496
497     1. **Identify the relevant section:** The question asks about the
498     vlogger's second outing, where she buys makeup. This occurs
499     after the gym visit.
500
501     2. **Review the frames:** The frames show the vlogger going to the
502     park after the gym. She then goes shopping and buys makeup at
503     Kicks.
504
505     3. **Select the answer:** The main reason for her second outing is
506     to go to the park.
507
508     Therefore, the correct answer is (4) To go to the park.
509     Final Answer: (4)
510
511     Strategy C (retrieval_qa, get_segment, find_when):
512     '''
513     input_video = get_segment(start='00:00', end='30:25')
514     '''
515
516     '''
517     think("I need to find when the vlogger goes out and buys makeup. I'
518     ll use 'find_when' to locate that event, then use 'retrieval_qa'
519     to determine the reason.")
520     '''
521     I need to find when the vlogger goes out and buys makeup. I'll use
522     'find_when' to locate that event, then use 'retrieval_qa' to
523     determine the reason.
524
525     '''
526     when_out_buying_makeup = find_when(query="vlogger goes out and buys
527     makeup", video_segment=input_video)
528     '''
529
530     ["00:01", "01:59"]: The video shows a vlogger waking up in her room,
531     doing some laundry, and then seemingly getting ready to go out.
532     While there's no explicit depiction of buying makeup, the
533     context strongly suggests this is part of her preparation. The
534     time range encompasses her actions from waking up to preparing
535     laundry, which could be interpreted as part of her routine
536     before going out to buy makeup.
537
538     ["02:19", "02:25"]: The video shows the vlogger in her bathroom at
539     02:15, where she picks up a toothbrush and looks at her tablet.
540     From 02:19-02:25, she is shown selecting items from a bathroom
541     cabinet. This is consistent with the query.
542
543     ["05:24", "05:31"]: The vlogger is seen leaving her apartment at
544     04:02 and returning at 04:48. At 05:24-05:31, she is seen
545     putting away what appears to be a package of makeup on a shelf
546     in her bathroom. This is supported by the context of the query
547     and the visual of a package being placed on a shelf.
548
549     ["08:51", "09:16"]: The video shows a vlogger leaving their home at
550     08:30, walking outside, and then entering a grocery store at
551     08:51. The vlogger is seen selecting various grocery items,
552     including celery, grapes, blueberries, carrots, broccoli, and
553     lettuce, between 08:51 and 09:16. Therefore, the time range

```

```

540     reflects the vlogger's grocery shopping trip, which aligns with
541     the query's description of buying products.
542     ["14:00", "14:02"]: The query asks about the vlogger buying makeup.
543     The provided video frames show the vlogger in their home,
544     interacting with items that could be interpreted as makeup (
545     frames 14:00-14:02) and later organizing clothes in their closet
546     (frames 14:39-15:25). There is no visual evidence of the
547     vlogger going out to buy anything. Therefore, only the initial
548     interaction with what appears to be makeup is considered
549     relevant.
550     ["16:09", "16:18"], ["17:17", "17:25"]: The video shows the vlogger
551     going to a store, Kicks, and buying makeup. The vlogger is seen
552     in the store from 16:09 to 16:18. At 17:17-17:25, the vlogger
553     unpacks the Kicks bag, revealing makeup.
554     ["18:12", "18:17"]: The query asks for the time range when the
555     vlogger goes out and buys makeup. The provided video frames
556     show the vlogger in their room, at their vanity. At 18:12, there
557     are shopping bags from the store Kicks on the vanity. This
558     strongly suggests that the vlogger has recently returned from a
559     shopping trip where they purchased makeup from Kicks. There is
560     no footage of the actual shopping trip itself within this video
561     chunk.
562     ["20:25", "20:51"]: The video shows a person preparing coffee. From
563     20:25 to 20:51, the person is shown unpacking two ceramic brown
564     mugs. The contents of the mugs are visible and appear to be
565     coffee. Therefore, this time range is identified as the time
566     when the vlogger is shown preparing coffee.
567     ["29:20", "29:41"]: The video shows the vlogger in her bathroom at
568     28:00-28:18 applying makeup and then brushing her teeth. At
569     29:20-29:41, she is shown painting her nails. The presence of
570     various makeup in both locations supports the query.
571
572     '''
573     reason = retrieval_qa(question="What is the main reason the vlogger
574     goes out again, where she also buys makeup?", answer_options=['
575     To go exercise', 'To eat chicken', 'To buy coffee mugs', 'To go
576     to the park'], video_segment=get_segment(start='16:09', end='
577     17:25'))
578     '''
579     The vlogger goes to a store and buys coffee mugs. Then she goes to
580     another store and buys makeup. Therefore, the main reason she
581     goes out again is to buy coffee mugs.
582
583     Final Answer: (3)
584
585     '''
586     final_answer = finish("The provided text from 'retrieval_qa'
587     indicates the vlogger buys makeup at a store called Kicks.
588     However, it also mentions buying coffee mugs. The question asks
589     for the *main* reason for going out again. While makeup is
590     mentioned, the provided text strongly suggests the primary
591     reason for the second outing is to buy coffee mugs. Therefore,
592     the answer is (3). Final Answer: (3)")
593     '''
594     The provided text from 'retrieval_qa' indicates the vlogger buys
595     makeup at a store called Kicks. However, it also mentions
596     buying coffee mugs. The question asks for the *main* reason for
597     going out again. While makeup is mentioned, the provided text
598     strongly suggests the primary reason for the second outing is to
599     buy coffee mugs. Therefore, the answer is (3). Final Answer:
600     (3)
601
602     Critique:
603     A: 3. B: 4. C: 3.

```

```

594 Strategy B succeeds while Strategies A fails because Strategy A
595 looks at particular time ranges of the video and concludes it is
596 unable to find a correct answer, thus guessing.
597 While the approach is generally sound, Strategy A gives up upon not
598 finding an answer initially. Strategy C looks at a narrow range
599 and follows an implication to guess an answer. Thus Strategy B
600 is preferred.
601
602 Winning Strategy:
603 B

```

Listing 4: LVBench critic examples.

```

605 Input:
606 You will be given a question about a video and five possible answer
607 options. Question: What are the main points regarding proper form
608 and breathing during the barbell squat? Possible answer choices:
609 (1) Keep your back straight, core engaged, and chest up throughout
610 the movement. Inhale as you descend and exhale forcefully as you
611 drive up.
612 (2) Maintain a slight arch in your back, push your hips back first,
613 and keep your knees aligned with your toes. Breathe in during the
614 eccentric phase (lowering) and breathe out during the concentric
615 phase (rising).
616 (3) Focus on pushing through your heels, keep your head down, and
617 hold your breath at the bottom of the squat for stability. Exhale
618 as you return to the starting position.
619 (4) Prioritize depth over weight, allowing your hips to descend below
620 your knees. Inhale deeply at the top, hold your breath throughout
621 the squat, and exhale at the top.
622 (5) Use a wide stance, keep your toes pointed outwards, and focus on
623 squeezing your glutes at the top of the movement. Breathe
624 naturally throughout the exercise.
625 Video length: 01:35
626
627 Strategy A (get_segment, retrieval_qa, asr_understanding):
628 '''
629 input_video = get_segment(start='00:00', end='01:35')
630 '''
631
632 think("I need to understand the main points about form and breathing
633 during barbell squats. I'll use the ASR understanding tool to
634 analyze the video transcript. If that doesn't work, I'll try '
635 retrieval_qa' on the video frames.")
636 '''
637
638 I need to understand the main points about form and breathing during
639 barbell squats. I'll use the ASR understanding tool to analyze
640 the video transcript. If that doesn't work, I'll try 'retrieval_qa
641 ' on the video frames.
642 '''
643
644 asr_understanding(question='What are the main points regarding proper
645 form and breathing during the barbell squat?', answer_options=['
646 (1) Keep your back straight, core engaged, and chest up
647 throughout the movement. Inhale as you descend and exhale
648 forcefully as you drive up.', '(2) Maintain a slight arch in your
649 back, push your hips back first, and keep your knees aligned with
650 your toes. Breathe in during the eccentric phase (lowering) and
651 breathe out during the concentric phase (rising).', '(3) Focus on
652 pushing through your heels, keep your head down, and hold your
653 breath at the bottom of the squat for stability. Exhale as you
654 return to the starting position.', '(4) Prioritize depth over

```

```

648     weight, allowing your hips to descend below your knees. Inhale
649     deeply at the top, hold your breath throughout the squat, and
650     exhale at the top.', '(5) Use a wide stance, keep your toes
651     pointed outwards, and focus on squeezing your glutes at the top of
652     the movement. Breathe naturally throughout the exercise.'])
653     '''
654     The video emphasizes maintaining a slight arch in the back
655     (00:23-00:27), pushing the hips back first (00:28-00:31), and
656     ensuring knees track with the toes (00:32-00:35). For breathing,
657     it advises inhaling during the descent and exhaling during the
658     ascent (00:55-01:00).
659     '''
660     think("The ASR transcript supports answer choice (2). The video
661     highlights maintaining a slight arch in your back, initiating the
662     movement with the hips, aligning knees with toes, and coordinating
663     breathing with the phases of the squat.")
664     '''
665     The ASR transcript supports answer choice (2). The video highlights
666     maintaining a slight arch in your back, initiating the movement
667     with the hips, aligning knees with toes, and coordinating
668     breathing with the phases of the squat.
669     '''
670     final_answer = finish("Final Answer: (2)")
671     '''
672     Final Answer: (2)
673     Strategy B (direct retrieval_qa):
674     The video primarily shows a demonstration of the barbell squat and
675     highlights key points on form. It doesn't explicitly detail the
676     breathing technique. However, general weightlifting principles
677     suggest inhaling during the lowering phase and exhaling during the
678     lifting phase. Additionally, proper form includes keeping the
679     back straight, core engaged, and chest up.
680     Final Answer: (1)
681     Strategy C (get_segment, retrieval_qa find_when, asr_understanding):
682     '''
683     input_video = get_segment(start='00:00', end='01:35')
684     '''
685     '''
686     think("I need to find the main points regarding form and breathing in
687     the barbell squat. I'll use 'asr_understanding' to see if the
688     transcript mentions these details.")
689     '''
690     I need to find the main points regarding form and breathing in the
691     barbell squat. I'll use 'asr_understanding' to see if the
692     transcript mentions these details.
693     '''
694     asr_understanding_result = asr_understanding(question="What are the
695     main points regarding proper form and breathing during the barbell
696     squat?", answer_options=[
697         "Keep your back straight, core engaged, and chest up throughout
698         the movement. Inhale as you descend and exhale forcefully as
699         you drive up.",
700         "Maintain a slight arch in your back, push your hips back first,
701         and keep your knees aligned with your toes. Breathe in during
702         the eccentric phase (lowering) and breathe out during the
703         concentric phase (rising).",

```

```

702     "Focus on pushing through your heels, keep your head down, and
703     hold your breath at the bottom of the squat for stability.
704     Exhale as you return to the starting position.",
705     "Prioritize depth over weight, allowing your hips to descend below
706     your knees. Inhale deeply at the top, hold your breath
707     throughout the squat, and exhale at the top.",
708     "Use a wide stance, keep your toes pointed outwards, and focus on
709     squeezing your glutes at the top of the movement. Breathe
710     naturally throughout the exercise."
711 ]
712 """
713 The video focuses on maintaining a slight arch in your back
714 (0:23-0:27), pushing your hips back first (0:28-0:31), and keeping
715 your knees aligned with your toes (0:32-0:35). For breathing, it
716 advises to inhale during the descent (eccentric phase) and exhale
717 during the ascent (concentric phase) (0:55-1:00).
718 """
719 think("The transcript suggests answer (2), highlighting back arch, hip
720 movement, knee alignment, and breathing coordinated with the
721 squat phases. Let's use 'retrieval_qa' to cross-check this.")
722 """
723 The transcript suggests answer (2), highlighting back arch, hip
724 movement, knee alignment, and breathing coordinated with the squat
725 phases. Let's use 'retrieval_qa' to cross-check this.
726 """
727 retrieval_qa_result = retrieval_qa(question="What are the main points
728 regarding proper form and breathing during the barbell squat?",
729 answer_options=[
730     "Keep your back straight, core engaged, and chest up throughout
731     the movement. Inhale as you descend and exhale forcefully as
732     you drive up.",
733     "Maintain a slight arch in your back, push your hips back first,
734     and keep your knees aligned with your toes. Breathe in during
735     the eccentric phase (lowering) and breathe out during the
736     concentric phase (rising).",
737     "Focus on pushing through your heels, keep your head down, and
738     hold your breath at the bottom of the squat for stability.
739     Exhale as you return to the starting position.",
740     "Prioritize depth over weight, allowing your hips to descend below
741     your knees. Inhale deeply at the top, hold your breath
742     throughout the squat, and exhale at the top.",
743     "Use a wide stance, keep your toes pointed outwards, and focus on
744     squeezing your glutes at the top of the movement. Breathe
745     naturally throughout the exercise."
746 ], video_segment=input_video)
747 """
748 The video demonstrates proper form for barbell squats, emphasizing
749 keeping the back straight, core engaged, and chest up. While
750 breathing is not explicitly addressed, general knowledge suggests
751 inhaling on the descent and exhaling forcefully on the ascent.
752 """
753 Final Answer: (1)
754 """
755 final_answer = finish("The ASR transcript suggests (2), focusing on
756 back arch and hip movement. However, 'retrieval_qa' prioritizes a
757 straight back and doesn't fully confirm the breathing technique.
758 Considering general weightlifting knowledge and the potential for
759 ASR misinterpretation, (1) is the more reliable answer. Final
760 Answer: (1)")
761 """
762 The ASR transcript suggests (2), focusing on back arch and hip
763 movement. However, 'retrieval_qa' prioritizes a straight back and

```

756 doesn't fully confirm the breathing technique. Considering general
 757 weightlifting knowledge and the potential for ASR
 758 misinterpretation, (1) is the more reliable answer. Final Answer:
 759 (1)
 760
 Critique:
 761 A: 2. B: 1. C: 1. Strategies B and C consider the visual content, but
 762 deny the information can be found in the video and rely on general
 763 knowledge.
 764
 Winning Strategies:
 765 A
 766
 Input:
 767 You will be given a question about a video and five possible answer
 768 options. Question: To prepare this dish, what are the main steps
 769 involved? Possible answer choices:
 770 (1) The main steps include chopping vegetables, boiling them in water,
 771 adding spices and herbs, pouring in coconut milk, simmering the
 772 mixture, and finally garnishing with cilantro.
 773 (2) The main steps include chopping vegetables, sauteing them in a
 774 skillet, adding spices and dry fruits, pouring in beef broth,
 775 baking the mixture in the oven, and finally garnishing with green
 776 and red bell peppers.
 777 (3) The main steps include chopping vegetables, sauteing them in a wok
 778 , adding milk and herbs, pouring in vegetable broth, simmering the
 779 mixture, and finally garnishing with parsley.
 780 (4) The main steps include chopping vegetables, sauteing them in a pan
 781 , adding lentils, pouring in water, adding spices and cilantro,
 782 simmering the lentils, and finally garnishing with chopped
 783 tomatoes.
 784 (5) The main steps include chopping fruits, sauteing them in a pan,
 785 adding spices and herbs, pouring in chicken broth, simmering the
 786 mixture, and finally garnishing with cilantro.
 787 Video length: 09:43
 Strategy A (get_segment, retrieval_qa, asr_understanding):
 788 '''
 789 input_video = get_segment(start='00:00', end='09:43')
 790 '''
 791
 792 '''
 793 answer = asr_understanding(question='To prepare this dish, what are
 794 the main steps involved?', answer_options=['The main steps include
 795 chopping vegetables, boiling them in water, adding spices and
 796 herbs, pouring in coconut milk, simmering the mixture, and finally
 797 garnishing with cilantro.', 'The main steps include chopping
 798 vegetables, sauteing them in a skillet, adding spices and dry
 799 fruits, pouring in beef broth, baking the mixture in the oven, and
 800 finally garnishing with green and red bell peppers.', 'The main
 801 steps include chopping vegetables, sauteing them in a wok, adding
 802 milk and herbs, pouring in vegetable broth, simmering the mixture,
 803 and finally garnishing with parsley.', 'The main steps include
 804 chopping vegetables, sauteing them in a pan, adding lentils,
 805 pouring in water, adding spices and cilantro, simmering the
 806 lentils, and finally garnishing with chopped tomatoes.', 'The main
 807 steps include chopping fruits, sauteing them in a pan, adding
 808 spices and herbs, pouring in chicken broth, simmering the mixture,
 809 and finally garnishing with cilantro.'])
 810
 811 '''
 812 The main steps are chopping onions and carrots (00:32-01:15), sauteing
 813 them in olive oil with garlic and ginger (01:23-02:58), adding
 814 lentils and water (03:12-04:05), adding turmeric and cumin
 815 (04:30-05:20), simmering until the lentils are tender

```

810         (05:35-07:10), and finally garnishing with chopped cilantro and
811         tomatoes (07:22-08:01).
812
813     """
814     think("None of the provided options exactly match the steps in the ASR
815           transcript. Option (1) is closest since it mentions chopping
816           vegetables, adding spices, and simmering. However, it omits key
817           ingredients and steps like sauteing.")
818     """
819     None of the provided options exactly match the steps in the ASR
820     transcript. Option (1) is closest since it mentions chopping
821     vegetables, adding spices, and simmering. However, it omits key
822     ingredients and steps like sauteing.
823
824     """
825     final_answer = finish("Final Answer: (1)")
826     """
827     Final Answer: (1)
828
829     Strategy B (direct retrieval_qa):
830     The video demonstrates a recipe for lentil soup. The chef chops
831     vegetables like onions, carrots, and garlic. They then saute these
832     vegetables in a pan with oil. Lentils and water are added,
833     followed by spices like cumin and turmeric. The mixture simmers
834     until the lentils are cooked through. Finally, it's garnished with
835     chopped tomatoes and cilantro.
836
837     The closest answer is (4).
838
839     Final Answer: (4)
840     Strategy C (get_segment, retrieval_qa find_when, asr_understanding):
841     """
842     input_video = get_segment(start='00:00', end='09:43')
843     """
844
845     """
846     answer = asr_understanding(question='To prepare this dish, what are
847           the main steps involved?', answer_options=['The main steps include
848           chopping vegetables, boiling them in water, adding spices and
849           herbs, pouring in coconut milk, simmering the mixture, and finally
850           garnishing with cilantro.', 'The main steps include chopping
851           vegetables, sauteing them in a skillet, adding spices and dry
852           fruits, pouring in beef broth, baking the mixture in the oven, and
853           finally garnishing with green and red bell peppers.', 'The main
854           steps include chopping vegetables, sauteing them in a wok, adding
855           milk and herbs, pouring in vegetable broth, simmering the mixture,
856           and finally garnishing with parsley.', 'The main steps include
857           chopping vegetables, sauteing them in a pan, adding lentils,
858           pouring in water, adding spices and cilantro, simmering the
859           lentils, and finally garnishing with chopped tomatoes.', 'The main
860           steps include chopping fruits, sauteing them in a pan, adding
861           spices and herbs, pouring in chicken broth, simmering the mixture,
862           and finally garnishing with cilantro.'])
863     """
864     The main steps include chopping onions and carrots (0:35-1:10),
865     sauteing them in olive oil (1:15-2:05), adding spices like cumin
866     and turmeric (2:10-2:45), then adding lentils and water
867     (2:50-3:30). The mixture simmers for some time (3:45-5:00) and is
868     finally garnished with cilantro and chopped tomato (5:10-5:45).
869
870     """
871     think("None of the options perfectly align with the ASR transcript.
872           Option (1) is somewhat similar as it mentions chopping vegetables,

```

```

864         spices, and simmering. However, it misses crucial details like
865         sauteing and the specific ingredients mentioned.")
866     """
867     None of the options perfectly align with the ASR transcript. Option
868     (1) is somewhat similar as it mentions chopping vegetables, spices
869     , and simmering. However, it misses crucial details like sauteing
870     and the specific ingredients mentioned.
871
872     """
873     final_answer = finish("Final Answer: (1)")
874     """
875     Final Answer: (1)
876
877     Critique:
878     A: 1. B: 4. C: 1. A and C consider the ASR, but give up after finding
879     it doesn't directly result in an answer.
880
881     Winning Strategies:
882     B
883
884     Input:
885     You will be given a question about a video and five possible answer
886     options. Question: In what order does the video discuss the
887     following topics: introduction, stock taillight operation,
888     Morimoto XB LED taillight replacement, taillight features, and
889     installation and warranty? Possible answer choices:
890     (1) Introduction, stock taillight operation, taillight features,
891     Morimoto XB LED taillight replacement, installation and warranty.
892     (2) Morimoto XB LED taillight replacement, introduction, taillight
893     features, stock taillight operation, installation and warranty.
894     (3) Introduction, Stock taillight operation, Morimoto XB LED taillight
895     replacement, taillight features, Installation and warranty.
896     (4) Introduction, Morimoto XB LED taillight replacement, installation
897     and warranty, taillight features, stock taillight operation.
898     (5) Stock taillight operation, introduction, Morimoto XB LED taillight
899     replacement, taillight features, installation and warranty.
900     Video length: 01:15
901
902     Strategy A (get_segment, retrieval_qa, asr_understanding):
903     """
904     input_video = get_segment(start='00:00', end='01:15')
905     """
906
907     """
908     answer = asr_understanding(question='In what order does the video
909     discuss the following topics: introduction, stock taillight
910     operation, Morimoto XB LED taillight replacement, taillight
911     features, and installation and warranty?', answer_options=[
912     'Introduction, stock taillight operation, taillight features,
913     Morimoto XB LED taillight replacement, installation and warranty.'
914     , 'Morimoto XB LED taillight replacement, introduction, taillight
915     features, stock taillight operation, installation and warranty.',
916     'Introduction, Stock taillight operation, Morimoto XB LED
917     taillight replacement, taillight features, Installation and
918     warranty.', 'Introduction, Morimoto XB LED taillight replacement,
919     installation and warranty, taillight features, stock taillight
920     operation.', 'Stock taillight operation, introduction, Morimoto XB
921     LED taillight replacement, taillight features, installation and
922     warranty.'])
923     """
924     The video first provides an introduction (00:02-00:10). Next, it
925     discusses the stock taillight operation (00:11-00:25). Then, it
926     details the Morimoto XB LED taillight replacement (00:26-00:38).
927     Following this, the video describes the taillight features

```

```

918         (00:39-00:55). Finally, it covers installation and warranty
919         (00:56-01:08). Therefore, the correct order is (3).
920
921     """
922     final_answer = finish("Final Answer: (3)")
923     """
924     Final Answer: (3)
925
926     Strategy B (direct retrieval_qa):
927     The video begins with an introduction of Morimoto and then shows the
928     stock taillight operation. After that, it shows the Morimoto XB
929     LED taillight replacement and then discusses its features. Finally
930     , it shows the installation process and discusses its warranty.
931     Therefore, the correct order is introduction, stock taillight
932     operation, Morimoto XB LED taillight replacement, taillight
933     features, and installation and warranty.
934
935     Final Answer: (1)
936     Strategy C (get_segment, retrieval_qa find_when, asr_understanding):
937     """
938     input_video = get_segment(start='00:00', end='01:15')
939     """
940
941     """
942     answer = asr_understanding(question='In what order does the video
943     discuss the following topics: introduction, stock taillight
944     operation, Morimoto XB LED taillight replacement, taillight
945     features, and installation and warranty?', answer_options=['
946     Introduction, stock taillight operation, taillight features,
947     Morimoto XB LED taillight replacement, installation and warranty.'
948     , 'Morimoto XB LED taillight replacement, introduction, taillight
949     features, stock taillight operation, installation and warranty.',
950     'Introduction, Stock taillight operation, Morimoto XB LED
951     taillight replacement, taillight features, Installation and
952     warranty.', 'Introduction, Morimoto XB LED taillight replacement,
953     installation and warranty, taillight features, stock taillight
954     operation.', 'Stock taillight operation, introduction, Morimoto XB
955     LED taillight replacement, taillight features, installation and
956     warranty.'])
957     """
958     The video first provides an introduction (00:02-00:10). Next, it
959     discusses the stock taillight operation (00:11-00:25). Then, it
960     details the Morimoto XB LED taillight replacement (00:26-00:38).
961     Following this, the video describes the taillight features
962     (00:39-00:55). Finally, it covers installation and warranty
963     (00:56-01:08). Therefore, the correct order is (3).
964
965     """
966     final_answer = finish("Final Answer: (3)")
967     """
968     Final Answer: (3)
969
970     Critique:
971     A: 3. B: 1. C: 3. The question asks about what is discussed in the
972     video, implying speech, but B is not able to consider the ASR and
973     fails.
974
975     Winning Strategies:
976     A, C
977
978     Input:
979     You will be given a question about a video and five possible answer
980     options. Question: Why does Jamie decide to stay home instead of

```

```

972     going to the library at 11:00 a.m. on a Saturday? Possible answer
973     choices:
974     (1) Jamie gets a call from a friend and decides to meet them for lunch
975         instead.
976     (2) Jamie realizes the library is closed on Saturdays.
977     (3) Jamie's car breaks down, preventing them from driving to the
978         library.
979     (4) Jamie starts a new book at home and is too engrossed to leave.
980     (5) Jamie remembers they have a prior engagement, a family gathering,
981         preventing them from going to the library.
982     Video length: 01:59
983
984     Strategy A (get_segment, retrieval_qa, asr_understanding):
985     """
986
987     input_video = get_segment(start='00:00', end='01:59')
988     """
989
990     answer = asr_understanding(question="Why does Jamie decide to stay
991         home instead of going to the library at 11:00 a.m. on a Saturday
992         ?", answer_options=['Jamie gets a call from a friend and decides
993         to meet them for lunch instead.', 'Jamie realizes the library is
994         closed on Saturdays.', 'Jamie\'s car breaks down, preventing them
995         from driving to the library.', 'Jamie starts a new book at home
996         and is too engrossed to leave.', 'Jamie remembers they have a
997         prior engagement, a family gathering, preventing them from going
998         to the library.'])
999     """
1000
1001     Jamie decides to stay home instead of going to the library because
1002     they remember they have a prior family gathering
1003     [01:25:00-01:32:00].
1004
1005     """
1006
1007     final_answer = finish("Based on the ASR transcript, Jamie decides to
1008         stay home instead of going to the library because they remember
1009         they have a prior family gathering. Final Answer: (5)")
1010
1011     """
1012
1013     Based on the ASR transcript, Jamie decides to stay home instead of
1014     going to the library because they remember they have a prior
1015     family gathering. Final Answer: (5)
1016
1017
1018     Strategy B (direct retrieval_qa):
1019     The video shows Jamie checking the time on their watch, which reads
1020     11:00 AM. They then pack a bag with books and prepare to leave the
1021     house, even putting on their shoes. However, they receive a text
1022     message and after reading it, unpack their bag and take off their
1023     shoes. They then proceed to sit on the couch and relax. The video
1024     doesn't explicitly say why the plans changed, but the text message
1025     is likely the reason. The video ends with a short scene of a
1026     group of people, suggesting a possible social gathering.
1027
1028
1029     Final Answer: (1)
1030
1031
1032     Strategy C (get_segment, retrieval_qa find_when, asr_understanding):
1033     """
1034
1035     input_video = get_segment(start='00:00', end='01:59')
1036     """
1037
1038     """
1039
1040     found_times = find_when(query="Jamie's plans are interrupted",
1041         video_segment=input_video)
1042     """

```

```

1026 ["00:35", "00:45"]: The query asks for the time range when Jamie's
1027 plans are interrupted. Based on the video and transcript, Jamie's
1028 plans to go to the library are interrupted when they trip over
1029 their shoelaces after putting on their shoes. This occurs between
1030 00:35 and 00:42. The video shows them putting on their shoes,
1031 tripping, and then sitting down.
1032
1033 '''
1034 segment_of_interest = get_segment(start='00:35', end='00:45')
1035 '''
1036 ['00:35', '00:45']
1037
1038 '''
1039 answer = retrieval_qa(question="Why does Jamie decide to stay home
1040 instead of going to the library at 11:00 a.m. on a Saturday?",
1041 answer_options=['Jamie gets a call from a friend and decides to
1042 meet them for lunch instead.', 'Jamie realizes the library is
1043 closed on Saturdays.', 'Jamie\'s car breaks down, preventing them
1044 from driving to the library.', 'Jamie starts a new book at home
1045 and is too engrossed to leave.', 'Jamie remembers they have a
1046 prior engagement, a family gathering, preventing them from going
1047 to the library.'], video_segment=segment_of_interest)
1048 '''
1049 The video shows Jamie putting on their shoes and then tripping over
1050 their shoelaces. There is no indication they got a call, realized
1051 the library was closed, had car trouble, or started a new book.
1052 The video also does not show a family gathering. Therefore, none
1053 of the options 1-5 are supported by the video.
1054
1055 However, if we must choose the most likely option, option (3) is the
1056 most plausible. Jamie tripping could be interpreted as an
1057 indication of car trouble, albeit a very indirect one.
1058
1059 Final Answer: (3)
1060
1061 '''
1062 final_answer = finish(final_answer="The video shows Jamie putting on
1063 their shoes and then tripping over their shoelaces. There is no
1064 indication they got a call, realized the library was closed, had
1065 car trouble, or started a new book. The video also does not show a
1066 family gathering. Therefore, none of the options 1-5 are
1067 supported by the video. However, if we must choose the most likely
1068 option, option (3) is the most plausible. Jamie tripping could be
1069 interpreted as an indication of car trouble, albeit a very
1070 indirect one. Final Answer: (3)")
1071 '''
1072 The video shows Jamie putting on their shoes and then tripping over
1073 their shoelaces. There is no indication they got a call, realized
1074 the library was closed, had car trouble, or started a new book.
1075 The video also does not show a family gathering. Therefore, none
1076 of the options 1-5 are supported by the video. However, if we must
1077 choose the most likely option, option (3) is the most plausible.
1078 Jamie tripping could be interpreted as an indication of car
1079 trouble, albeit a very indirect one. Final Answer: (3)
1080
1081 Critique:
1082 A: 5. B: 1. C: 3. B makes a guess after not finding an explicit answer
1083 in the visuals. C selects an overly narrow range to consider, as
1084 the reason for the plans being interrupted may not be exactly when
1085 they are interrupted.
1086
1087 Winning Strategies:
1088 A
1089

```

Listing 5: Neptune critic examples.

```

1080
1081 Input:
1082 Question: When is the trampoline first used in the video?
1083 Video length: 04:23
1084
1085 Strategy A (get_segment, find_when):
1086 '''
1087 input_video = get_segment(start='00:00', end='04:23')
1088 '''
1089
1090 time_ranges = find_when(query='the trampoline gets its first use',
1091 video_segment=input_video)
1092 '''
1093 ["00:45", "01:15"]: The video shows a person assembling a trampoline.
1094 The trampoline is shown fully assembled at 00:45 and the assembly
1095 process is shown up to 01:15.
1096 ["02:30", "02:55"]: The video shows a person jumping on the trampoline
1097 . This is the first time the trampoline is shown being used, from
1098 02:30 to 02:55.
1099 '''
1100 think("The first time the trampoline is used is arguably at 00:45,
1101 when it is assembled. However, the question asks when it gets its
1102 *first use*, which implies actual use by a person. Therefore, the
1103 relevant time range is 02:30-02:55.")
1104 '''
1105 The first time the trampoline is used is arguably at 00:45, when it is
1106 assembled. However, the question asks when it gets its *first
1107 use*, which implies actual use by a person. Therefore, the
1108 relevant time range is 02:30-02:55.
1109 '''
1110 final_answer = finish("The question asks when the trampoline gets its
1111 first use, implying actual use by a person. While the trampoline
1112 is assembled at 00:45-01:15, its first actual use begins at 02:30
1113 when a person jumps on it. Therefore, the relevant time range is
1114 02:30-02:55. Final Answer: [02:30, 02:55]")
1115 '''
1116 The question asks when the trampoline gets its first use, implying
1117 actual use by a person. While the trampoline is assembled at
1118 00:45-01:15, its first actual use begins at 02:30 when a person
1119 jumps on it. Therefore, the relevant time range is 02:30-02:55.
1120 Final Answer: [02:30, 02:55]
1121 '''
1122
1123 Strategy B (get_segment, find_when, retrieval_qa):
1124 '''
1125 input_video = get_segment(start='00:00', end='04:23')
1126 '''
1127
1128 first_use = find_when(query='first use of the trampoline',
1129 video_segment=input_video)
1130 '''
1131 ["00:45", "01:50"]: The video shows a person assembling a trampoline.
1132 The first appearance of the trampoline occurs at 00:45, when it's
1133 partially assembled. The trampoline is shown in several stages of
1134 assembly until 01:50.
1135 ["03:10", "03:40"]: The video shows the finished trampoline being
1136 showcased in a backyard. The trampoline is fully assembled and
1137 ready for use at 03:10, and is shown from various angles until
1138 03:40.

```

```

1134 '''
1135 think("The first time range [00:45, 01:50] seems to describe the
1136     trampoline being assembled, but the second range [03:10, 03:40]
1137     shows the trampoline in its final assembled state, ready for use.
1138     Therefore, the answer is the second range.")
1139 '''
1140 The first time range [00:45, 01:50] seems to describe the trampoline
1141 being assembled, but the second range [03:10, 03:40] shows the
1142 trampoline in its final assembled state, ready for use. Therefore
1143 , the answer is the second range.
1144 '''
1145 final_answer = finish("The first time range [00:45, 01:50] seems to
1146     describe the trampoline being assembled, but the second range
1147     [03:10, 03:40] shows the trampoline in its final assembled state,
1148     ready for use. Therefore, the answer is the second range. Final
1149     Answer: [03:10, 03:40]")
1150 '''
1151 The first time range [00:45, 01:50] seems to describe the trampoline
1152 being assembled, but the second range [03:10, 03:40] shows the
1153 trampoline in its final assembled state, ready for use. Therefore
1154 , the answer is the second range. Final Answer: [03:10, 03:40]
1155 '''
1156 Strategy C (find_when):
1157 '''
1158 input_video = get_segment(start='00:00', end='03:53')
1159 '''
1160 '''
1161 query = "When is the trampoline first used in the video?"
1162 result = find_when(query, None)
1163 '''
1164 ["00:45", "01:15"]: The video shows a person assembling a trampoline.
1165     The trampoline is shown fully assembled at 00:45 and the assembly
1166     process is shown up to 01:15.
1167 ["02:30", "02:55"]: The video shows a person jumping on the trampoline
1168 . This is the first time the trampoline is shown being used, from
1169 02:30 to 02:55.
1170 '''
1171 think("The first result mentions the trampoline being assembled from
1172     00:45 to 01:15, but this is assembly, not use. The second result
1173     shows the trampoline in use from 02:30 to 02:55. I will choose
1174     the second range as the answer.")
1175 '''
1176 The first result mentions the trampoline being assembled from 00:36 to
1177     01:59, but this is assembly, not use. The second result shows the
1178     trampoline in use from 02:30 to 02:55. I will choose the second
1179     range as the answer.
1180 '''
1181 final_answer = "The trampoline's first use is shown from 02:30 to
1182     02:55. Final Answer: [02:30, 02:55]"
1183 finish(final_answer)
1184 '''
1185 The trampoline's first use is shown from 02:30 to 02:55. Final Answer:
1186     [02:30, 02:55]
1187 '''
1188 Critique:
1189 Strategies A and C actually perceive a person using the trampoline and
1190 pick the associated range. Strategy B only sees the finished
1191 trampoline and guesses based on that, but does not directly see
1192 the part relevant to the query.

```

```

1188
1189 Winning Strategies:
1190 A, C
1191
1192 Question: When is the woman showing a harder variation of the exercise
1193 ?
1194 Video length: 01:18
1195
1196 Strategy A (get_segment, find_when):
1197 """
1198 input_video = get_segment(start='00:00', end='01:18')
1199 """
1200
1201 """
1202 harder_exercise_times = find_when(query='woman showing a harder
1203 variation of the exercise', video_segment=input_video)
1204 """
1205 ["00:03", "00:27"], ["00:30", "01:15"]]: The video shows a woman
1206 performing a modified push-up against a wall from 0:03 to 0:27.
1207 From 0:30 to 1:15, she performs standard push-ups on the floor.
1208 The query asks for a harder version of the exercise. The standard
1209 push-up on the floor is considered more challenging than the
1210 modified version against the wall.
1211 """
1212 think("The tool 'find_when' returned two time ranges: ['00:03', '00:27'
1213 ] and ['00:30', '01:15']. The first range shows a modified push-
1214 up against a wall, and the second shows standard push-ups. The
1215 standard push-up is described as the harder variation. Therefore,
1216 the final answer should be the second range.")
1217 """
1218 The tool 'find_when' returned two time ranges: ['00:03', '00:27'] and
1219 ['00:30', '01:15']. The first range shows a modified push-up
1220 against a wall, and the second shows standard push-ups. The
1221 standard push-up is described as the harder variation. Therefore,
1222 the final answer should be the second range.
1223 """
1224 final_answer = finish("The tool 'find_when' returned two time ranges:
1225 ['00:03', '00:27'] and ['00:30', '01:15']. The first range shows a
1226 modified push-up against a wall, and the second shows standard
1227 push-ups. The standard push-up is described as the harder
1228 variation. Therefore, the final answer should be the second range.
1229 Final Answer: [00:30, 01:15]")
1230 """
1231 The tool 'find_when' returned two time ranges: ['00:03', '00:27'] and
1232 ['00:30', '01:15']. The first range shows a modified push-up
1233 against a wall, and the second shows standard push-ups. The
1234 standard push-up is described as the harder variation. Therefore,
1235 the final answer should be the second range. Final Answer: [00:30,
1236 01:15]
1237
1238 Strategy B (get_segment, find_when, retrieval_qa):
1239 """
1240 input_video = get_segment(start='00:00', end='01:18')
1241 """
1242
1243 """
1244 times = find_when(query='woman showing a harder variation of the
1245 exercise', video_segment=input_video)
1246 """
1247 ["00:03", "00:27"], ["00:30", "01:15"]]: The video shows a woman
1248 performing a modified push-up against a wall from 0:03 to 0:27.

```

```

1242     From 0:30 to 1:15, she performs standard push-ups on the floor.
1243     The query asks for a harder version of the exercise. The standard
1244     push-up on the floor is considered more challenging than the
1245     modified version against the wall.
1246
1247     """
1248     think("The tool 'find_when' returned two time ranges: 00:03-00:27 and
1249           00:30-01:15. The description indicates that 00:30-01:15 shows a
1250           harder variation of the exercise. I will use this time range as
1251           the final answer.")
1252     """
1253     The tool 'find_when' returned two time ranges: 00:03-00:27 and
1254     00:30-01:15. The description indicates that 00:30-01:15 shows a
1255     harder variation of the exercise. I will use this time range as
1256     the final answer.
1257
1258     """
1259     final_answer = finish("The tool 'find_when' returned two time ranges:
1260                           00:03-00:27 and 00:30-01:15. The description indicates that
1261                           00:30-01:15 shows a harder variation of the exercise. I will use
1262                           this time range as the final answer. Final Answer: [00:30, 01:15]"
1263                           )
1264     """
1265     The tool 'find_when' returned two time ranges: 00:03-00:27 and
1266     00:30-01:15. The description indicates that 00:30-01:15 shows a
1267     harder variation of the exercise. I will use this time range as
1268     the final answer. Final Answer: [00:30, 01:15]
1269
1270     Strategy C (find_when):
1271     """
1272     input_video = get_segment(start='00:00', end='01:18')
1273     """
1274
1275     """
1276     think("I need to find when the woman is showing a harder variation of
1277           the exercise. I'll first try to find any mentions of 'harder' or
1278           similar terms related to the exercise difficulty.")
1279     """
1280     I need to find when the woman is showing a harder variation of the
1281     exercise. I'll first try to find any mentions of 'harder' or
1282     similar terms related to the exercise difficulty.
1283
1284     """
1285     query_result = find_when(query="harder exercise", video_segment=None)
1286     """
1287     ["00:03", "00:27"]: The query "harder exercise" is subjective. However
1288     , based on the video content, the frames from 00:03 to 00:27 show
1289     a person performing a modified exercise, which could be considered
1290     an "easier exercise." The exercise is shown for a sustained
1291     period. Since a harder exercise was requested, it is likely shown
1292     at a different time in the video.
1293
1294     """
1295     final_answer = finish("Based on the tool's analysis, the woman
1296                           performs a harder version of the exercise between 00:03 and 00:27.
1297                           The tool identified this segment, but it may have mistaken it for
1298                           an easier variation. Final Answer: [00:03, 00:27]")
1299     """
1300     Based on the tool's analysis, the woman performs a harder version of
1301     the exercise between 00:03 and 00:27. The tool identified this
1302     segment, but it may have mistaken it for an easier variation.
1303     Final Answer: [00:03, 00:27]
1304
1305     Critique:

```

```

1296 Strategies A and B find ranges corresponding to both versions of the
1297 exercise and compare them to find the harder variation.
1298 Strategy C only identified one range, for a potentially easier version
1299 , then did not recognize the alternative to compare to.
1300
1301 Winning Strategies:
1302 A, B
1303
1304 Question: When are we given a zoomed in view of the guitar strings?
1305 Video length: 01:53
1306
1307 Strategy A (get_segment, find_when):
1308 '''
1309
1310
1311 times = find_when(query='zoomed in view of the guitar strings',
1312 video_segment=input_video)
1313 '''
1314 ["00:00", "01:52"]: The video shows a person playing a guitar
1315 throughout its duration. The query asks for a zoomed in view of
1316 the guitar strings. Therefore, the entire video shows the event.
1317
1318 '''
1319 final_answer = finish("The query asks for when a zoomed in view of the
1320 guitar strings is given. The response from 'find_when' indicates
1321 that a guitar is played throughout the entire video. Therefore,
1322 the entire video is relevant. Final Answer: [00:00, 01:53]")
1323
1324 The query asks for when a zoomed in view of the guitar strings is
1325 given. The response from 'find_when' indicates that a guitar is
1326 played throughout the entire video. Therefore, the entire video is
1327 relevant. Final Answer: [00:00, 01:53]
1328
1329 Strategy B (get_segment, find_when, retrieval_qa):
1330 '''
1331
1332
1333 times = find_when(query='zoomed in view of the guitar strings',
1334 video_segment=input_video)
1335 '''
1336 ["00:00", "01:52"]: The video shows a person playing a guitar
1337 throughout its duration. The query asks for a zoomed in view of
1338 the guitar strings. Therefore, the entire video shows the event.
1339
1340 '''
1341 final_answer = finish(final_answer="The video shows a person playing a
1342 guitar throughout its duration. The query asks for a zoomed in
1343 view of the guitar strings. Therefore, the entire video shows the
1344 event. Final Answer: [00:00, 01:53]")
1345
1346 The video shows a person playing a guitar throughout its duration. The
1347 query asks for a zoomed in view of the guitar strings. Therefore,
1348 the entire video shows the event. Final Answer: [00:00, 01:53]
1349
1350 Strategy C (find_when):
1351 '''
1352
1353 input_video = get_segment(start='00:00', end='01:53')
1354 '''

```

```

1350 '''
1351 think("I need to find when the video shows a close-up of the guitar
1352 strings. I'll use the 'find_when' tool to search for this.")
1353 '''
1354 I need to find when the video shows a close-up of the guitar strings.
1355 I'll use the 'find_when' tool to search for this.
1356
1357 '''
1358 find_when(query="close-up of guitar strings", video_segment=None)
1359 '''
1360 ["01:10", "01:25"]: The video shows a person playing guitar. From
1361 [01:10] to [01:25], the camera zooms in to show a close-up view of
1362 the guitar strings as the person plays them.
1363
1364 '''
1365 think("The 'find_when' tool returned a single time range: 01:10-01:25.
1366 This seems to directly answer the question. I will use this as
1367 my final answer.")
1368 '''
1369 The 'find_when' tool returned a single time range: 01:10-01:25. This
1370 seems to directly answer the question. I will use this as my final
1371 answer.
1372
1373 '''
1374 finish("The find_when tool identified the time range 01:10-01:25 as
1375 showing a close-up of the guitar strings. Final Answer: [01:10,
1376 01:25]")
1377 '''
1378 The find_when tool identified the time range 01:10-01:25 as showing a
1379 close-up of the guitar strings. Final Answer: [01:10, 01:25]
1380
1381 Critique:
1382 Strategies A and B choose the whole video, despite the query asking
1383 for a zoomed in view, implying a relative change in the video.
1384 Strategy C instead identifies a particular range.
1385
1386 Winning Strategies:
1387 C
1388
1389 Input:
1390 Question: When is the person least energetic in the video?
1391 Video length: 00:49
1392
1393 Strategy A (get_segment, find_when):
1394 '''
1395 input_video = get_segment(start='00:00', end='00:49')
1396 '''
1397
1398 '''
1399 activity_times = find_when(query='person least energetic',
1400 video_segment=input_video)
1401 '''
1402 ["00:11", "00:18"], ["00:19", "00:24"], ["00:33", "00:41"]: The query
1403 asks to identify the time ranges where the person is least
energetic. Based on the provided video frames, the person appears
to be sleeping during the periods from 00:11 to 00:18, from 00:19
to 00:24, and from 00:33 to 00:41. These segments represent
periods of minimal movement and apparent rest.
'''
think("The tool returned multiple time ranges: ['00:11', '00:18'], ['
00:19', '00:24'], ['00:33', '00:41']. Since the question asks for

```

```

1404         the time the person is LEAST energetic, and the first two ranges
1405         are contiguous, I will combine them into a single range.")
1406     """
1407     The tool returned multiple time ranges: ['00:11', '00:18'], ['00:19',
1408         '00:24'], ['00:33', '00:41']. Since the question asks for the
1409         time the person is LEAST energetic, and the first two ranges are
1410         contiguous, I will combine them into a single range.
1411     """
1412     final_answer = "The person is least energetic from 00:11 to 00:24, and
1413         from 00:33 to 00:41. Final Answer: [00:11, 00:24], [00:33, 00:41]"
1414     "
1415     finish(final_answer)
1416     """
1417     The person is least energetic from 00:11 to 00:24, and from 00:33 to
1418         00:41. Final Answer: [00:11, 00:24], [00:33, 00:41]
1419
1420     Strategy B (get_segment, find_when, retrieval_qa):
1421     """
1422     input_video = get_segment(start='00:00', end='00:49')
1423     """
1424
1425     """
1426     activity_times = find_when(query='person is least energetic',
1427         video_segment=input_video)
1428     """
1429     ["00:11", "00:18"]: The query asks to identify when a person is least
1430         energetic in the video. The video shows a person who appears to be
1431         sleeping between 00:11 and 00:18. This is the period of least
1432         visible movement and likely the time of lowest energy.
1433     """
1434     think("The find_when tool suggests the person is least energetic
1435         between 00:11 and 00:18. This seems reasonable based on the
1436         description. I will use this as my final answer.")
1437     """
1438     The find_when tool suggests the person is least energetic between
1439         00:11 and 00:18. This seems reasonable based on the description.
1440         I will use this as my final answer.
1441     """
1442     final_answer = finish("Final Answer: [00:11, 00:18]")
1443     """
1444     Final Answer: [00:11, 00:18]
1445
1446     Strategy C (find_when):
1447     """
1448     input_video = get_segment(start='00:00', end='00:49')
1449     """
1450
1451     """
1452     think("I need to find when the person is least energetic. I'll use
1453         find_when to search for inactivity.")
1454     """
1455     I need to find when the person is least energetic. I'll use find_when
1456         to search for inactivity.
1457     """
1458     activity_times = find_when(query="person is least energetic",
1459         video_segment=None)
1460     """
1461     ["00:11", "00:18"]: The query asks to identify when a person is least
1462         energetic in the video. The person appears to be sleeping between

```

```

1458     00:11 and 00:18, indicating low energy. There may be other periods
1459     of low activity, but this period of apparent sleep is the
1460     clearest indication of low energy.
1461
1462     '''
1463     think("The find_when tool suggests the person is least energetic
1464           between 00:11 and 00:18. This seems like the most obvious period
1465           of inactivity (sleeping). I will use this as my final answer.")
1466     '''
1467     The find_when tool suggests the person is least energetic between
1468     00:11 and 00:18. This seems like the most obvious period of
1469     inactivity (sleeping). I will use this as my final answer.
1470
1471     '''
1472     final_answer = finish("The find_when tool suggests the person is least
1473                           energetic between 00:11 and 00:18. This seems like the most
1474                           obvious period of inactivity (sleeping). Therefore, I'm choosing
1475                           this as the final answer. Final Answer: [00:11, 00:18]")
1476     '''
1477     The find_when tool suggests the person is least energetic between
1478     00:11 and 00:18. This seems like the most obvious period of
1479     inactivity (sleeping). Therefore, I'm choosing this as the final
1480     answer. Final Answer: [00:11, 00:18]
1481
1482     Winning Strategies:
1483     A

```

Listing 6: ActivityNet-RTL critic examples.

```

1482     Input:
1483     You will be given a question about a video and five possible answer
1484     options, where C refers to the person wearing the camera. Question
1485     : What is the main action being performed by C in the video, and
1486     provide specific moments that support your observation. Possible
1487     answer choices:
1488     (1) The main action performed by C is applying makeup. This is evident
1489     as C picks up a brush, looks in a mirror, and applies something
1490     to their face.
1491     (2) C's main action in the video is eating a meal. This is clearly
1492     evident because C lifts a fork to their mouth multiple times and
1493     chews.
1494     (3) The main action performed by C in the video is reading a book.
1495     This is evident in the fact that C holds a book open, turns pages,
1496     and moves their eyes across the page.
1497     (4) In the video, the main action performed by C is talking on the
1498     phone. This is evident because C holds a phone to their ear and
1499     speaks into it several times.
1500     (5) The main action performed by C in the video is typing on a
1501     computer. This is clearly seen as C sits at a desk, places their
1502     hands on a keyboard, and moves their fingers.
1503     Video length: 01:48
1504
1505     Strategy A (retrieval_qa, get_segment):
1506     '''
1507     input_video = get_segment(start='00:00', end='01:48')
1508     '''
1509
1510     '''
1511     answer = retrieval_qa(question='What is the main action being
1512                           performed by C in the video, and provide specific moments that
1513                           support your observation.', answer_options=['The main action
1514                           performed by C is applying makeup. This is evident as C picks up a
1515                           brush, looks in a mirror, and applies something to their face.',
1516                           'C\'s main action in the video is eating a meal. This is clearly

```

```

1512     evident because C lifts a fork to their mouth multiple times and
1513     chews.', 'The main action performed by C in the video is reading a
1514     book. This is evident in the fact that C holds a book open, turns
1515     pages, and moves their eyes across the page.', 'In the video, the
1516     main action performed by C is talking on the phone. This is
1517     evident because C holds a phone to their ear and speaks into it
1518     several times.', 'The main action performed by C in the video is
1519     typing on a computer. This is clearly seen as C sits at a desk,
1520     places their hands on a keyboard, and moves their fingers.'],
1521     video_segment=input_video)
1522     """
1523     C is clearly typing on a computer throughout the video. They are
1524     sitting at a desk with their hands on a keyboard and their fingers
1525     moving. While there may be other actions present, the consistent
1526     and primary action is typing.
1527
1528     Final Answer: (5)
1529
1530     """
1531     final_answer = finish(final_answer='C is clearly typing on a computer
1532     throughout the video. They are sitting at a desk with their hands
1533     on a keyboard and their fingers moving. While there may be other
1534     actions present, the consistent and primary action is typing.\n\
1535     nFinal Answer: (5)')
1536     """
1537     C is clearly typing on a computer throughout the video. They are
1538     sitting at a desk with their hands on a keyboard and their fingers
1539     moving. While there may be other actions present, the consistent
1540     and primary action is typing.
1541
1542     Final Answer: (5)
1543
1544     Strategy B (direct retrieval_qa):
1545     C is clearly typing on a computer throughout the video. They are
1546     sitting at a desk with their hands on a keyboard and their fingers
1547     are moving. This is the primary and most consistent action shown
1548     .
1549
1550     Final Answer: (5)
1551
1552     Strategy C (retrieval_qa, get_segment, find_when):
1553     """
1554     input_video = get_segment(start='00:00', end='01:48')
1555     """
1556
1557     """
1558     think("The question asks about the main action C performs. I'll use '
1559     find_when' to check for specific actions mentioned in the options,
1560     like 'applying makeup', 'eating', 'reading', 'talking on the
1561     phone', and 'typing'. This will help identify the main action.")
1562     """
1563     The question asks about the main action C performs. I'll use '
1564     find_when' to check for specific actions mentioned in the options,
1565     like 'applying makeup', 'eating', 'reading', 'talking on the
1566     phone', and 'typing'. This will help identify the main action.
1567
1568     """
1569     makeup_times = find_when(query='C applies makeup', video_segment=
1570     input_video)
1571     eating_times = find_when(query='C eats a meal', video_segment=
1572     input_video)
1573     reading_times = find_when(query='C reads a book', video_segment=
1574     input_video)

```

```

1566 phone_times = find_when(query='C talks on the phone', video_segment=
1567     input_video)
1568 typing_times = find_when(query='C types on a computer', video_segment=
1569     input_video)
1570 '''
1571 ["00:12", "00:18"]: At around 00:12, C appears to pick up a small
1572     object, which is potentially a phone, and brings it to their ear.
1573     The interaction with the object continues until about 00:18.
1574     Although it is not definitively clear whether C is talking, their
1575     actions suggest they may be on a phone call.
1576 '''
1577 think("'find_when' only returned a possible phone call. The limited
1578     information isn't conclusive. Let's use 'retrieval_qa' to check if
1579     C is focused on work, which could suggest typing.")
1580 '''
1581 'find_when' only returned a possible phone call. The limited
1582     information isn't conclusive. Let's use 'retrieval_qa' to check if
1583     C is focused on work, which could suggest typing.
1584 '''
1585 work_focused = retrieval_qa(question="Is C focused on work in the
1586     video?", answer_options=["Yes", "No"], video_segment=input_video)
1587 '''
1588 The video shows C sitting at a desk and using a computer. Although it's
1589     not shown exactly what they are doing, their posture and actions
1590     suggest they may be working.
1591 '''
1592 Final Answer: (2)
1593 '''
1594 think("The response suggests work at a desk and computer use. That
1595     supports typing. However, the short phone interaction is also a
1596     possibility. Since 'eating' seems unlikely in a work context, I'll
1597     guess between phone and typing, leaning towards phone due to the
1598     'find_when' result.")
1599 '''
1600 The response suggests work at a desk and computer use. That supports
1601     typing. However, the short phone interaction is also a
1602     possibility. Since 'eating' seems unlikely in a work context, I'll
1603     guess between phone and typing, leaning towards phone due to the
1604     'find_when' result.
1605 '''
1606 final_answer = finish(final_answer="The 'find_when' result hinted at
1607     phone use. 'retrieval_qa' suggested work at a desk with a computer
1608     , which supports 'typing.' Lacking definitive proof, I'll guess '
1609     talking on the phone' based on the slight 'find_when' evidence.
1610     Final Answer: (4)")
1611 '''
1612 The 'find_when' result hinted at phone use. 'retrieval_qa' suggested
1613     work at a desk with a computer, which supports 'typing.' Lacking
1614     definitive proof, I'll guess 'talking on the phone' based on the
1615     slight 'find_when' evidence. Final Answer: (4)
1616 '''
1617 Critique:
1618 A: 5. B: 5. C: 4. Strategy C attempts an overly complex strategy with
1619     the find_when tool and ends up guessing due to limited information
1620     .
1621 '''
1622 Winning Strategies:
1623 A, B
1624 '''
1625 Input:

```

```

1620 You will be given a question about a video and five possible answer
1621 options, where C refers to the person wearing the camera. Question
1622 : Describe how C prepared the ingredients and assembled the
1623 sandwich in the video, noting key steps and transitions. Possible
1624 answer choices:
1625 (1) C prepared the ingredients by ordering them online and having them
1626 delivered. They then assembled the sandwich by layering the
1627 ingredients between two slices of bread. They added condiments
1628 and cut the sandwich in half.
1629 (2) C prepared the ingredients by taking pre-sliced ingredients from
1630 the refrigerator. They then assembled the sandwich by layering the
1631 ingredients between two slices of bread. They added condiments
1632 and cut the sandwich in half.
1633 (3) C prepared the ingredients by chopping vegetables and slicing meat
1634 . They then assembled the sandwich by layering the ingredients
1635 between two slices of bread. They added condiments and cut the
1636 sandwich in half.
1637 (4) C prepared the ingredients by microwaving pre-packaged meals. They
1638 then assembled the sandwich by layering the ingredients between
1639 two slices of bread. They added condiments and cut the sandwich
1640 in half.
1641 (5) C prepared the ingredients by going to a restaurant and getting
1642 takeout. They then assembled the sandwich by layering the
1643 ingredients between two slices of bread. They added condiments
1644 and cut the sandwich in half.
1645 Video length: 02:45
1646
1647 Strategy A (retrieval_qa, get_segment):
1648 '''
1649 input_video = get_segment(start='00:00', end='02:45')
1650 '''
1651
1652 answer = retrieval_qa(question='Describe how C prepared the
1653 ingredients and assembled the sandwich in the video, noting key
1654 steps and transitions. ', answer_options=['C prepared the
1655 ingredients by ordering them online and having them delivered.
1656 They then assembled the sandwich by layering the ingredients
1657 between two slices of bread. They added condiments and cut the
1658 sandwich in half.', 'C prepared the ingredients by taking pre-
1659 sliced ingredients from the refrigerator. They then assembled the
1660 sandwich by layering the ingredients between two slices of bread.
1661 They added condiments and cut the sandwich in half.', 'C prepared
1662 the ingredients by chopping vegetables and slicing meat. They
1663 then assembled the sandwich by layering the ingredients between
1664 two slices of bread. They added condiments and cut the sandwich
1665 in half.', 'C prepared the ingredients by microwaving pre-packaged
1666 meals. They then assembled the sandwich by layering the
1667 ingredients between two slices of bread. They added condiments
1668 and cut the sandwich in half.', 'C prepared the ingredients by
1669 going to a restaurant and getting takeout. They then assembled the
1670 sandwich by layering the ingredients between two slices of bread.
1671 They added condiments and cut the sandwich in half.'],
1672 video_segment=input_video)
1673 '''
1674
1675 The video shows C assembling a sandwich. They take various ingredients
1676 , like sliced meat and vegetables, from the refrigerator. They
1677 layer these ingredients between two slices of bread, add
1678 condiments, and then cut the sandwich in half. There is no
1679 indication of ordering food online, chopping vegetables,
1680 microwaving meals, or getting takeout. The ingredients appear to
1681 be pre-sliced.
1682
1683 Final Answer: (2)

```

```

1674
1675 """
1676 final_answer = finish("The video clearly shows C using pre-sliced
1677 ingredients from the refrigerator. Therefore, the most accurate
1678 description is option (2). Final Answer: (2)")
1679 """
1680 The video clearly shows C using pre-sliced ingredients from the
1681 refrigerator. Therefore, the most accurate description is option
1682 (2). Final Answer: (2)
1683
1684 Strategy B (direct retrieval_qa):
1685 The video shows C making a sandwich. They retrieve pre-sliced
1686 ingredients, like lettuce, tomato, and meat, from the refrigerator
1687 . They then assemble the sandwich, add condiments, and cut it in
1688 half. The ingredients were prepared by being pre-sliced and stored
1689 in the refrigerator.
1690
1691 Final Answer: (3)
1692
1693 Strategy C (retrieval_qa, get_segment, find_when):
1694 """
1695 input_video = get_segment(start='00:00', end='02:45')
1696 """
1697
1698 """
1699 think("To understand how the ingredients were prepared, I'll use '
1700 find_when' to look for actions like chopping or retrieving from
1701 the refrigerator, and then use 'retrieval_qa' to select the best
1702 answer.")
1703 """
1704 To understand how the ingredients were prepared, I'll use 'find_when'
1705 to look for actions like chopping or retrieving from the
1706 refrigerator, and then use 'retrieval_qa' to select the best
1707 answer.
1708
1709 """
1710 preparation_times = find_when("C preparing the sandwich ingredients",
1711 input_video)
1712 """
1713 ["00:05", "00:40"]: C retrieves pre-sliced sandwich ingredients from
1714 the refrigerator.
1715
1716 """
1717 answer = retrieval_qa(question="Describe how C prepared the
1718 ingredients and assembled the sandwich in the video, noting key
1719 steps and transitions.", answer_options=["C prepared the
1720 ingredients by ordering them online and having them delivered.
1721 They then assembled the sandwich by layering the ingredients
1722 between two slices of bread. They added condiments and cut the
1723 sandwich in half.", "C prepared the ingredients by taking pre-
1724 sliced ingredients from the refrigerator. They then assembled the
1725 sandwich by layering the ingredients between two slices of bread.
1726 They added condiments and cut the sandwich in half.", "C prepared
1727 the ingredients by chopping vegetables and slicing meat. They
1728 then assembled the sandwich by layering the ingredients between
1729 two slices of bread. They added condiments and cut the sandwich
1730 in half.", "C prepared the ingredients by microwaving pre-packaged
1731 meals. They then assembled the sandwich by layering the
1732 ingredients between two slices of bread. They added condiments
1733 and cut the sandwich in half.", "C prepared the ingredients by
1734 going to a restaurant and getting takeout. They then assembled the
1735 sandwich by layering the ingredients between two slices of bread.
1736 They added condiments and cut the sandwich in half."],
1737 video_segment=input_video)

```

```

1728 '''
1729 The video shows C making a sandwich. They retrieve pre-sliced
1730 ingredients, like lettuce, tomato, and meat, from the refrigerator
1731 . They then assemble the sandwich, add condiments, and cut it in
1732 half. The ingredients were prepared by being pre-sliced and stored
1733 in the refrigerator.
1734
1735 Final Answer: (3)
1736
1737 '''
1738 final_answer = finish("Since C takes pre-sliced ingredients from the
1739 refrigerator, the correct answer is (3). Final Answer: (3)")
1740 '''
1741 Since C takes pre-sliced ingredients from the refrigerator, the
1742 correct answer is (3). Final Answer: (3)
1743
1744 Critique:
1745 A: 2. B: 3. C: 3. Strategy A correctly identifies that the ingredients
1746 are pre-sliced from the refrigerator but fails to choose the
1747 matching answer.
1748
1749 Winning Strategies:
1750 B, C
1751
1752 Input:
1753 You will be given a question about a video and five possible answer
1754 options. Question: Compare the lighting conditions in the room at
1755 the start of the video to those near the end. How has the lighting
1756 evolved over the course of the video? Possible answer choices:
1757 (1) The lighting in the earlier segments of the video is primarily
1758 provided by a single overhead lamp, while the lighting in the
1759 latter segments is primarily provided by multiple spotlights.
1760 (2) The lighting in the earlier segments of the video is primarily
1761 natural light from a window, while the lighting in the latter
1762 segments is primarily provided by an overhead lamp.
1763 (3) The lighting in the earlier segments of the video is primarily
1764 provided by multiple spotlights, while the lighting in the latter
1765 segments is primarily natural light from a window.
1766 (4) The lighting in the earlier segments of the video is primarily
1767 natural light from a window, while the lighting in the latter
1768 segments is dimmer, perhaps due to cloud cover or sunset.
1769 (5) The lighting in the earlier segments of the video is dimmer,
1770 perhaps due to cloud cover or sunset, while the lighting in the
1771 latter segments is primarily natural light from a window.
1772 Video length: 03:01
1773
1774 Strategy A (retrieval_qa, get_segment):
1775 '''
1776 input_video = get_segment(start='00:00', end='03:01')
1777 '''
1778
1779 '''
1780 think("To compare the lighting in the earlier vs. latter segments, I'
1781 ll first use retrieval_qa on the first minute and the last minute.
1782 This should give a general idea of the differences. Then I can
1783 refine my approach if needed.")
1784 '''
1785 To compare the lighting in the earlier vs. latter segments, I'll first
1786 use retrieval_qa on the first minute and the last minute. This
1787 should give a general idea of the differences. Then I can refine
1788 my approach if needed.
1789
1790 '''
1791 first_minute_segment = get_segment(start='00:00', end='01:00')

```

```

1782 first_minute_answer = retrieval_qa(question="Describe the lighting in
1783 this segment.", video_segment=first_minute_segment)
1784
1785 last_minute_segment = get_segment(start='02:00', end='03:00')
1786 last_minute_answer = retrieval_qa(question="Describe the lighting in
1787 this segment.", video_segment=last_minute_segment)
1788
1789 think(f"First minute: {{first_minute_answer}}. Last minute: {{
1790 last_minute_answer}}. Now I need to compare these and select the
1791 best matching answer choice.")
1792
1793 First minute: The lighting appears to be dim, likely from a weak
1794 overhead light source.. Last minute: The room is brightly lit by
1795 natural light coming in from a large window.. Now I need to
1796 compare these and select the best matching answer choice.
1797
1798
1799 think("The first minute shows dim lighting, which could be interpreted
1800 as cloud cover or sunset. The last minute shows bright natural
1801 light. Option (4) seems close, mentioning dimmer lighting
1802 initially and then a change, but it doesn't specifically mention
1803 the natural light. However, it's the closest option.")
1804
1805 The first minute shows dim lighting, which could be interpreted as
1806 cloud cover or sunset. The last minute shows bright natural light.
1807 Option (4) seems close, mentioning dimmer lighting initially and
1808 then a change, but it doesn't specifically mention the natural
1809 light. However, it's the closest option.
1810
1811
1812 final_answer = finish("Reasoning: The initial dim lighting suggests
1813 cloud cover or sunset as per option (4), even though the later
1814 segment has bright natural light, not just a generic change. Final
1815 Answer: (4)")
1816
1817 Reasoning: The initial dim lighting suggests cloud cover or sunset as
1818 per option (4), even though the later segment has bright natural
1819 light, not just a generic change. Final Answer: (4)
1820
1821
1822 Strategy B (direct retrieval_qa):
1823 The video starts with dim lighting, likely due to cloud cover or the
1824 time of day. As the video progresses, the lighting becomes
1825 brighter and it appears more natural light is entering the room,
1826 likely from a window. This suggests the sun might have come out
1827 from behind the clouds or the video was filmed as the day
1828 progressed.
1829
1830
1831 Final Answer: (5)
1832
1833
1834 Strategy C (retrieval_qa, get_segment, find_when):
1835
1836 input_video = get_segment(start='00:00', end='03:01')
1837
1838
1839 think("To compare the lighting, I'll analyze the first and last
1840 minutes using find_when and then use retrieval_qa for a more
1841 detailed comparison.")
1842
1843

```

```

1836 To compare the lighting, I'll analyze the first and last minutes using
1837 find_when and then use retrieval_qa for a more detailed
1838 comparison.
1839
1840 '''
1841 early_segment = get_segment(start='00:00', end='01:00')
1842 '''
1843 ['00:00', '01:00']
1844
1845 '''
1846 early_lighting = find_when(query="Describe the lighting conditions.",
1847                             video_segment=early_segment)
1848 '''
1849 ["00:00", "00:59"]: The lighting conditions appear dim, likely due to
1850 a weak overhead light source or cloud cover outside.
1851
1852 '''
1853 late_segment = get_segment(start='02:00', end='03:01')
1854 '''
1855 ['02:00', '03:01']
1856
1857 '''
1858 late_lighting = find_when(query="Describe the lighting conditions.",
1859                             video_segment=late_segment)
1860 '''
1861 ["02:00", "03:00"]: The room is brightly lit by natural light
1862 streaming in from a large window.
1863
1864 '''
1865 comparison = retrieval_qa(question="Compare the lighting conditions in
1866 the room at the start of the video to those near the end. How has
1867 the lighting evolved over the course of the video?",
1868                             answer_options=[
1869     "The lighting in the earlier segments of the video is primarily
1870     provided by a single overhead lamp, while the lighting in the
1871     latter segments is primarily provided by multiple spotlights.",
1872     '
1873     "The lighting in the earlier segments of the video is primarily
1874     natural light from a window, while the lighting in the latter
1875     segments is primarily provided by an overhead lamp.",
1876     "The lighting in the earlier segments of the video is primarily
1877     provided by multiple spotlights, while the lighting in the
1878     latter segments is primarily natural light from a window.",
1879     "The lighting in the earlier segments of the video is primarily
1880     natural light from a window, while the lighting in the latter
1881     segments is dimmer, perhaps due to cloud cover or sunset.",
1882     "The lighting in the earlier segments of the video is dimmer,
1883     perhaps due to cloud cover or sunset, while the lighting in
1884     the latter segments is primarily natural light from a window."
1885 ], video_segment=input_video)
1886 '''
1887 The video starts with dim lighting, likely due to cloud cover or the
1888 time of day. As the video progresses, the lighting becomes
1889 brighter and it appears more natural light is entering the room,
1890 likely from a window. This suggests the sun might have come out
1891 from behind the clouds or the video was filmed as the day
1892 progressed.
1893
1894 Final Answer: (5)
1895
1896 '''
1897 final_answer = finish("Based on the analysis, the early segment has
1898 dim lighting (potentially due to cloud cover), while the late
1899

```

```

1890     segment is lit by natural light. This matches option (5). Final
1891     Answer: (5)"""
1892     """
1893     Based on the analysis, the early segment has dim lighting (potentially
1894     due to cloud cover), while the late segment is lit by natural
1895     light. This matches option (5). Final Answer: (5)
1896
1897     Critique: A: 4. B: 5. C: 5. Strategy A initially uses the valid
1898     strategy of using segments at the beginning and end, but
1899     ultimately does not logically connect the information to the
1900     appropriate answer choice.
1901
1902     Winning Strategies: B, C
1903
1904     Input:
1905     You will be given a question about a video and five possible answer
1906     options, where C refers to the person wearing the camera. Question
1907     : What is the primary action performed with the backpack at the
1908     beginning and the main purpose of using a backpack in the context
1909     of this video? please highlight the significance of this action in
1910     the overall travel process. Possible answer choices:
1911     (1) The primary action performed with the backpack at the beginning is
1912     to open it. the main purpose of using a backpack in the context
1913     of this video is to store belongings. opening the backpack allows
1914     for easy access to items needed for the trip.
1915     (2) The primary action performed with the backpack at the beginning is
1916     to close it. the main purpose of using a backpack in the context
1917     of this video is to carry belongings. closing the backpack ensures
1918     that items are secure and do not fall out during travel.
1919     (3) The primary action performed with the backpack at the beginning is
1920     to put it on. the main purpose of using a backpack in the context
1921     of this video is to carry belongings. putting the backpack on
1922     allows for hands-free travel and even weight distribution.
1923     (4) The primary action performed with the backpack at the beginning is
1924     to pack it. the main purpose of using a backpack in the context
1925     of this video is to carry belongings. packing the backpack ensures
1926     that all necessary items are brought on the trip.
1927     (5) The primary action performed with the backpack at the beginning is
1928     to take it off. the main purpose of using a backpack in the
1929     context of this video is to carry belongings. taking the backpack
1930     off allows for a rest and reduces strain on the back.
1931     Video length: 02:48
1932
1933     Strategy A (retrieval_qa, get_segment):
1934     """
1935     input_video = get_segment(start='00:00', end='02:48')
1936     """
1937
1938     """
1939     answer = retrieval_qa(question='What is the primary action performed
1940     with the backpack at the beginning and the main purpose of using a
1941     backpack in the context of this video? please highlight the
1942     significance of this action in the overall travel process.',
1943     answer_options=['The primary action performed with the backpack at
    the beginning is to open it. the main purpose of using a backpack
    in the context of this video is to store belongings. opening the
    backpack allows for easy access to items needed for the trip.', '
    The primary action performed with the backpack at the beginning is
    to close it. the main purpose of using a backpack in the context
    of this video is to carry belongings. closing the backpack ensures
    that items are secure and do not fall out during travel.', 'The
    primary action performed with the backpack at the beginning is to
    put it on. the main purpose of using a backpack in the context of

```

```

1944     this video is to carry belongings. putting the backpack on allows
1945     for hands-free travel and even weight distribution.', 'The primary
1946     action performed with the backpack at the beginning is to pack it
1947     . the main purpose of using a backpack in the context of this
1948     video is to carry belongings. packing the backpack ensures that
1949     all necessary items are brought on the trip.', 'The primary action
1950     performed with the backpack at the beginning is to take it off.
1951     the main purpose of using a backpack in the context of this video
1952     is to carry belongings. taking the backpack off allows for a rest
1953     and reduces strain on the back.'], video_segment=input_video)
1954     """
1955     The video shows a person packing a backpack at the beginning. They
1956     carefully organize and place items inside. This is essential for
1957     travel as it ensures all needed items are brought along. The main
1958     purpose of the backpack in the video is to carry belongings.
1959     Final Answer: (4)
1960     """
1961     think("The response aligns with option 4, showing the backpack is
1962     initially packed, crucial for travel to ensure having necessary
1963     items. The main purpose is indeed carrying belongings.")
1964     """
1965     The response aligns with option 4, showing the backpack is initially
1966     packed, crucial for travel to ensure having necessary items. The
1967     main purpose is indeed carrying belongings.
1968     """
1969     final_answer = finish("Final Answer: (4)")
1970     """
1971     Final Answer: (4)
1972     Strategy B (direct retrieval_qa):
1973     The video shows a person packing a backpack at the beginning. They
1974     carefully organize and place items inside. This is essential for
1975     travel as it ensures all needed items are brought along. The main
1976     purpose of the backpack in the video is to carry belongings.
1977     Final Answer: (4)
1978     Strategy C (retrieval_qa, get_segment, find_when):
1979     """
1980     input_video = get_segment(start='00:00', end='02:48')
1981     """
1982
1983     """
1984     think("I'll use 'find_when' to identify the initial backpack action
1985     and 'retrieval_qa' to determine its purpose in the video, then
1986     analyze the options.")
1987     """
1988     I'll use 'find_when' to identify the initial backpack action and '
1989     retrieval_qa' to determine its purpose in the video, then analyze
1990     the options.
1991     """
1992     initial_action_times = find_when("someone interacts with a backpack",
1993     input_video)
1994     """
1995     ["00:05", "00:25"]: The video shows a person actively packing a
1996     backpack. They are placing clothes, toiletries, and other items
1997     inside the backpack.

```

```

1998 ["01:50", "01:55"]: The person is seen wearing the backpack and
1999 walking through an airport terminal.
2000
2001 '''
2002 purpose = retrieval_qa(question="What is the main purpose of using a
2003 backpack in the context of this video?", answer_options=[
2004     "The main purpose of using a backpack in the context of this video
2005     is to store documents.",
2006     "The main purpose of using a backpack in the context of this video
2007     is to carry belongings.",
2008     "The main purpose of using a backpack in the context of this video
2009     is to organize toiletries.",
2010     "The main purpose of using a backpack in the context of this video
2011     is to conceal valuables.",
2012     "The main purpose of using a backpack in the context of this video
2013     is to transport food."
2014 ], video_segment=input_video)
2015 '''
2016 The video shows a person packing and then carrying a backpack filled
2017 with personal belongings during their travel. Therefore, the main
2018 purpose of the backpack is to carry belongings.
2019
2020 Final Answer: (2)
2021
2022 '''
2023 final_answer = finish("'find_when' shows the backpack is first packed,
2024 and 'retrieval_qa' confirms it's used to carry belongings.
2025 Knowing packing is essential for bringing necessary items, option
2026 (3), putting on the backpack for hands-free carrying, aligns with
2027 the main purpose and initial action. Final Answer: (3)")
2028 '''
2029 'find_when' shows the backpack is first packed, and 'retrieval_qa'
2030 confirms it's used to carry belongings. Knowing packing is
2031 essential for bringing necessary items, option (3), putting on the
2032 backpack for hands-free carrying, aligns with the main purpose
2033 and initial action. Final Answer: (3)
2034
2035 Critique:
2036 A: 4. B: 4. C: 2. Strategy C correctly finds that the backpack is
2037 first packed and then used to carry belongings, and reasons that
2038 the main purpose is to carry belongings, but ultimately
2039 illogically concludes the purpose is to put the backpack on
2040 despite this.
2041
2042 Winning Strategies:
2043 A, B
2044 "").strip()
2045
2046 else:
2047 CRITIC_EXAMPLES = textwrap.dedent(f"""
2048 Input:
2049 You will be given a question about a video and four possible answer
2050 options. Question: At 12:10, what color suit is the host wearing
2051 ?Possible answer choices:
2052 (1) Navy
2053 (2) Tan
2054 (3) Black
2055 (4) White
2056 Video length: 40:50
2057
2058 Strategy A (retrieval_qa, get_segment):
2059 '''
2060 input_video = get_segment(start='00:00', end='40:50')
2061 '''

```

```

2052
2053
2054     '''
2055     segment = get_segment(start='12:10', end='12:20')
2056     answer = retrieval_qa(question='What color suit is the host wearing?
2057     ', video_segment=segment)
2058     '''
2059     The host is wearing a navy suit.
2060
2061     '''
2062     final_answer = finish(f"The retrieval_qa tool determined the host
2063     was wearing a navy suit at 12:10. Final Answer: (1)")
2064     '''
2065     The retrieval_qa tool determined the host was wearing a navy suit at
2066     12:10. Final Answer: (1)
2067
2068     Strategy B (direct retrieval_qa):
2069     Here's how to solve this:
2070
2071     1. **Locate the timestamp:** Go to the 12:10 mark in the video.
2072
2073     2. **Observe the host's attire:** At this point, the host (Jimmy
2074     Kimmel) is wearing a dark suit and a dark-colored tie.
2075
2076     3. **Compare to options:** The suit appears to be black.
2077
2078     Therefore, the correct answer is (3) Black.
2079     Final Answer: (3)
2080
2081     Strategy C (retrieval_qa, get_segment, find_when):
2082     '''
2083     input_video = get_segment(start='00:00', end='40:50')
2084     '''
2085
2086     '''
2087     segment = get_segment(start='12:10', end='12:20')
2088     '''
2089     ['12:10', '12:20']
2090
2091     '''
2092     answer = retrieval_qa(question='At 12:10, what color suit is the
2093     host wearing?', answer_options=['Navy', 'Tan', 'Black', 'White'
2094     ], video_segment=segment)
2095     '''
2096     The image at 12:10 shows the host wearing a navy suit.
2097
2098     Final Answer: (1)
2099
2100     '''
2101     final_answer = finish(final_answer='The image at 12:10 shows the
2102     host wearing a navy suit. Final Answer: (1)')
2103     '''
2104     The image at 12:10 shows the host wearing a navy suit. Final Answer:
2105     (1)
2106
2107     Critique:
2108     A: 1. B: 3. C: 1. Strategies A and C succeed while Strategy B fails
2109     because Strategies A and C successfully use the get_segment tool
2110     to isolate the relevant part of the video.
2111     Strategy B fails as the retrieval_qa tool alone cannot actually
2112     locate a given timestamp in a video, resulting in a hallucinated
2113     answer.
2114
2115     Winning Strategies:

```

```

2106 A, C
2107
2108 Input:
2109 You will be given a question about a video and four possible answer
2110 options. Question: What does the man in charcoal suit do in the
2111 next scene after the scene of a photo of six people end?Possible
2112 answer choices:
2113 (1) He shows his scars
2114 (2) He takes off his shirt
2115 (3) He shows his jewelry
2116 (4) He shows his physique
2117 Video length: 39:59
2118
2119 Strategy A (retrieval_qa, get_segment):
2120 '''
2121
2122 input_video = get_segment(start='00:00', end='39:59')
2123 '''
2124
2125 think("I need to identify the scene where a photo of six people is
2126 shown. Then I need to find the scene immediately following it
2127 to see what the man in the charcoal suit does.")
2128 '''
2129
2130 I need to identify the scene where a photo of six people is shown.
2131 Then I need to find the scene immediately following it to see
2132 what the man in the charcoal suit does.
2133 '''
2134
2135 segment_with_photo = get_segment(start='00:00', end='01:00')
2136 answer = retrieval_qa(question="What does the man in charcoal suit
2137 do in the scene immediately after the scene with a photo of six
2138 people?", video_segment=segment_with_photo)
2139 '''
2140
2141 After the scene with the photo of six people, the man in the
2142 charcoal suit walks down a hallway and is greeted by a group of
2143 people who clap and cheer for him.
2144 '''
2145
2146 think("The provided text doesn't describe any of the options (1) He
2147 shows his scars, (2) He takes off his shirt, (3) He shows his
2148 jewelry, (4) He shows his physique. The man is celebrated. I
2149 will guess the closest option, which is none of them, but I will
2150 choose the least violent option.")
2151 '''
2152
2153 The provided text doesn't describe any of the options (1) He shows
2154 his scars, (2) He takes off his shirt, (3) He shows his jewelry,
2155 (4) He shows his physique. The man is celebrated. I will
2156 guess the closest option, which is none of them, but I will
2157 choose the least violent option.
2158 '''
2159
2160 final_answer = finish("Final Answer: (1)")
2161 '''
2162
2163 Final Answer: (1)
2164
2165 Strategy B (direct retrieval_qa):
2166 Here's how to solve this step-by-step:
2167
2168 1. **Identify the relevant scene:** The question refers to the scene
2169 after a photo of six people. Looking through the timestamps,
2170 this appears to be around 20:49.

```

```

2160 2. **Observe the next scene:** The next scene shows the man in the
2161 charcoal suit (Harry Styles) pointing to and showing jewelry on
2162 his arm.
2163
2164 3. **Match to the answer choices:** This action directly corresponds
2165 to answer choice (3).
2166
2167 Therefore, the correct answer is **(3) He shows his jewelry**.
2168 Final Answer: (3)
2169
2170 Strategy C (retrieval_qa, get_segment, find_when):
2171 '''
2172
2173
2174 '''
2175 think("First, I need to find when the scene with the photo of six
2176 people ends. Then I can examine the next scene to see what the
2177 man in the charcoal suit does.")
2178
2179 First, I need to find when the scene with the photo of six people
2180 ends. Then I can examine the next scene to see what the man in
2181 the charcoal suit does.
2182
2183 '''
2184 photo_end_time = find_when(query="scene with a photo of six people",
2185 video_segment=None)
2186
2187 '''
2188 ["00:05", "00:24"]: The query asks to find the scene with a photo of
2189 six people. From the provided frames, a photo of six people is
2190 visible on the wall in the background starting at 00:05 and
2191 remaining visible until 00:24. The photo is partially obscured
2192 at times by other people and objects, but it is consistently
2193 present in the background during this timeframe.
2194 ["10:39", "10:43"]: The query asks for a scene with a photo of six
2195 people. At 10:39 through 10:43 a large group photo of more than
2196 six people is shown.
2197 ["17:47", "17:53"]: The query asks for a scene with a photo of six
2198 people. At 17:47-17:53, a photo of approximately six people is
2199 visible behind a curtain as Harry Styles walks out onto the
2200 stage. The photo is partially obscured and the exact number of
2201 people is difficult to ascertain with certainty, but it appears
2202 to show at least five.
2203 ["18:47", "19:59"]: The query asks for the time range of a scene
2204 with a photo of six people. A photo of what appears to be six
2205 people is visible on the wall behind the main set in frames
2206 starting at 18:47 and continuing through 19:59. The photo is
2207 partially obscured at times by other elements in the scene, but
2208 it is consistently visible enough to be identified.
2209 ["20:03", "20:04"]: The query asks for the time range of a scene
2210 with a photo of six people. At 20:03, a wide shot of the studio
2211 audience is shown, and a photo of six people is visible on a
2212 screen in the background. This photo remains visible until
2213 approximately 20:04.
2214 ["23:15", "23:25"]: The query asks for the timestamp of a scene with
2215 a photo of six people. In the provided video frames, a photo
2216 of six people is visible on the wall in the background starting
2217 at 23:15 and remaining visible until 23:25. The photo is
2218 partially obscured at times by other elements in the frame, but
2219 it is consistently present during this time range.
2220 ["35:17", "35:20"]: The provided video frames show a wide shot of a
2221 studio audience at the 35:17 mark. The audience appears to
2222 consist of more than six people. While the exact number is

```

```

2214         difficult to determine from the provided frames, it is clearly a
2215         large group. Therefore, this time range satisfies the query.
2216     ["39:03", "39:18"]: The provided video frames show a live
2217         performance with an audience. At the end of the performance,
2218         from 39:03 to 39:18, the camera focuses on the audience, and
2219         there are enough people visible to potentially include a group
2220         of five. The resolution isn't high enough to definitively
2221         confirm the presence of a specific photo of six people, but the
2222         scene depicts a large group of people, making it a plausible
2223         match for the query.
2224
2225     """
2226     think("The most likely candidate for the 'scene with a photo of six
2227         people' is the longest continuous segment: 18:47-19:59. Let's
2228         use that as the end point for the previous scene.")
2229     """
2230     The most likely candidate for the 'scene with a photo of six people'
2231         is the longest continuous segment: 18:47-19:59. Let's use that
2232         as the end point for the previous scene.
2233
2234     """
2235     next_scene_segment = get_segment(start='19:59', end='20:05')
2236     """
2237     ['19:59', '20:05']
2238
2239     """
2240     answer = retrieval_qa(question="What does the man in the charcoal
2241         suit do in the scene from 19:59 to 20:05?", answer_options=['He
2242         shows his scars', 'He takes off his shirt', 'He shows his
2243         jewelry', 'He shows his physique'], video_segment=
2244         next_scene_segment)
2245     """
2246     The provided video frames show a talk show segment featuring three
2247         men. From 19:59 to 20:02, the man in the gray suit (Harry
2248         Styles) sits on a couch and gestures with his hands. At 20:03,
2249         the camera angle changes to show the audience, and the three men
2250         are still on the couch. There is no indication that the man in
2251         the gray suit shows his scars, takes off his shirt, shows a
2252         tattoo, or shows his muscles. He is simply participating in a
2253         conversation.
2254
2255     Final Answer: (4)
2256
2257     """
2258     final_answer = finish("Based on the analysis of the video segment
2259         from 19:59 to 20:05, the man in the charcoal suit (assuming this
2260         refers to one of the men in the talk show segment, as there's
2261         no charcoal suit explicitly mentioned) does not perform any of
2262         the actions listed in the options. However, since an answer must
2263         be chosen, and the retrieval_qa tool did not provide a
2264         definitive answer, I will select the option that is most likely
2265         to be present in a talk show setting, which is showing his
2266         physique. This is a weak inference, but given the constraints,
2267         it's the best guess. Final Answer: (4)")
2268     """
2269     Based on the analysis of the video segment from 19:59 to 20:05, the
2270         man in the charcoal suit (assuming this refers to one of the men
2271         in the talk show segment, as there's no charcoal suit
2272         explicitly mentioned) does not perform any of the actions listed
2273         in the options. However, since an answer must be chosen, and
2274         the retrieval_qa tool did not provide a definitive answer, I
2275         will select the option that is most likely to be present in a
2276         talk show setting, which is showing his physique. This is a
2277         weak inference, but given the constraints, it's the best guess.
2278         Final Answer: (4)

```

```

2268
2269 Critique:
2270 A:1. B: 3. C: 4. Strategy B succeeds while Strategies A fails
2271 because Strategy A incorrectly applies the get_segment tool to
2272 the first minute of the video and only considers that, resulting
2273 in it not being able to find a correct answer and ultimately
2274 guessing. Strategy C applies the find_when tool, then trims a
2275 segment and does not find the relevant information, then
2276 guessing as well.
2277
2278 Winning Strategies:
2279 B
2280
2281 Input:
2282 You will be given a question about a video and four possible answer
2283 options. Question: What is the main reason the vlogger goes out
2284 again, where she also buys makeup?Possible answer choices:
2285 (1) To go exercise
2286 (2) To eat chicken
2287 (3) To buy coffee mugs
2288 (4) To go to the park
2289 Video length: 30:25
2290
2291 Strategy A (retrieval_qa, get_segment):
2292 '''
2293 input_video = get_segment(start='00:00', end='30:25')
2294 '''
2295
2296 '''
2297 think("I need to find a segment of the video where the vlogger buys
2298 makeup. I'll try to identify this segment using retrieval_qa,
2299 and then I can use that information to answer the question about
2300 why she went out.")
2301 '''
2302 I need to find a segment of the video where the vlogger buys makeup.
2303 I'll try to identify this segment using retrieval_qa, and then
2304 I can use that information to answer the question about why she
2305 went out.
2306
2307 '''
2308 answer = retrieval_qa(question="When does the vlogger buy makeup?",
2309 answer_options=None, video_segment=input_video)
2310 '''
2311 The vlogger buys makeup at 16:10-16:11 and 17:20-17:25.
2312
2313 '''
2314 think("The vlogger buys makeup in two segments: 16:10-16:11 and
2315 17:20-17:25. I'll use retrieval_qa again, focusing on these
2316 segments, to determine the reason for her outing.")
2317 '''
2318 The vlogger buys makeup in two segments: 16:10-16:11 and
2319 17:20-17:25. I'll use retrieval_qa again, focusing on these
2320 segments, to determine the reason for her outing.
2321
2322 '''
2323 segment1 = get_segment(start='16:10', end='16:11')
2324 segment2 = get_segment(start='17:20', end='17:25')
2325 reason = retrieval_qa(question="What is the main reason the vlogger
2326 goes out again, where she also buys makeup?", answer_options=['
2327 To go exercise', 'To eat chicken', 'To buy coffee mugs', 'To go
2328 to the park'], video_segment=segment1)
2329 if reason == "": #check if the first segment didn't provide enough
2330 information

```

```

2322         reason = retrieval_qa(question="What is the main reason the
2323             vlogger goes out again, where she also buys makeup?",
2324             answer_options=['To go exercise', 'To eat chicken', 'To buy
2325             coffee mugs', 'To go to the park'], video_segment=segment2)
2326     '''
2327     None
2328     '''
2329     final_answer = finish("Based on the analysis of the video segments
2330         where the vlogger buys makeup, I was unable to definitively
2331         determine the main reason for her outing from the provided
2332         options. However, given the context of beauty product purchases,
2333         it's less likely she's going to the gym, eating salmon, or
2334         going to the park. Therefore, I will eliminate those options.
2335         Final Answer: (3)")
2336     '''
2337     Based on the analysis of the video segments where the vlogger buys
2338     makeup, I was unable to definitively determine the main reason
2339     for her outing from the provided options. However, given the
2340     context of beauty product purchases, it's less likely she's
2341     going to the gym, eating salmon, or going to the park. Therefore
2342     , I will eliminate those options. Final Answer: (3)
2343
2344     Strategy B (direct retrieval_qa):
2345     Here's how to solve this step-by-step:
2346
2347     1. **Identify the relevant section:** The question asks about the
2348     vlogger's second outing, where she buys makeup. This occurs
2349     after the gym visit.
2350
2351     2. **Review the frames:** The frames show the vlogger going to the
2352     park after the gym. She then goes shopping and buys makeup at
2353     Kicks.
2354
2355     3. **Select the answer:** The main reason for her second outing is
2356     to go to the park.
2357
2358     Therefore, the correct answer is (4) To go to the park.
2359     Final Answer: (4)
2360
2361     Strategy C (retrieval_qa, get_segment, find_when):
2362     '''
2363     input_video = get_segment(start='00:00', end='30:25')
2364     '''
2365
2366     '''
2367     think("I need to find when the vlogger goes out and buys makeup. I'
2368     ll use 'find_when' to locate that event, then use 'retrieval_qa'
2369     to determine the reason.")
2370     '''
2371
2372     I need to find when the vlogger goes out and buys makeup. I'll use
2373     'find_when' to locate that event, then use 'retrieval_qa' to
2374     determine the reason.
2375
2376     '''
2377     when_out_buying_makeup = find_when(query="vlogger goes out and buys
2378     makeup", video_segment=input_video)
2379     '''
2380     ["00:01", "01:59"]: The video shows a vlogger waking up in her room,
2381     doing some laundry, and then seemingly getting ready to go out.
2382     While there's no explicit depiction of buying makeup, the
2383     context strongly suggests this is part of her preparation. The
2384     time range encompasses her actions from waking up to preparing

```

laundry, which could be interpreted as part of her routine before going out to buy makeup.

["02:19", "02:25"]: The video shows the vlogger in her bathroom at 02:15, where she picks up a toothbrush and looks at her tablet. From 02:19-02:25, she is shown selecting items from a bathroom cabinet. This is consistent with the query.

["05:24", "05:31"]: The vlogger is seen leaving her apartment at 04:02 and returning at 04:48. At 05:24-05:31, she is seen putting away what appears to be a package of makeup on a shelf in her bathroom. This is supported by the context of the query and the visual of a package being placed on a shelf.

["08:51", "09:16"]: The video shows a vlogger leaving their home at 08:30, walking outside, and then entering a grocery store at 08:51. The vlogger is seen selecting various grocery items, including celery, grapes, blueberries, carrots, broccoli, and lettuce, between 08:51 and 09:16. Therefore, the time range reflects the vlogger's grocery shopping trip, which aligns with the query's description of buying products.

["14:00", "14:02"]: The query asks about the vlogger buying makeup. The provided video frames show the vlogger in their home, interacting with items that could be interpreted as makeup (frames 14:00-14:02) and later organizing clothes in their closet (frames 14:39-15:25). There is no visual evidence of the vlogger going out to buy anything. Therefore, only the initial interaction with what appears to be makeup is considered relevant.

["16:09", "16:18"], ["17:17", "17:25"]: The video shows the vlogger going to a store, Kicks, and buying makeup. The vlogger is seen in the store from 16:09 to 16:18. At 17:17-17:25, the vlogger unpacks the Kicks bag, revealing makeup.

["18:12", "18:17"]: The query asks for the time range when the vlogger goes out and buys makeup. The provided video frames show the vlogger in their room, at their vanity. At 18:12, there are shopping bags from the store Kicks on the vanity. This strongly suggests that the vlogger has recently returned from a shopping trip where they purchased makeup from Kicks. There is no footage of the actual shopping trip itself within this video chunk.

["20:25", "20:51"]: The video shows a person preparing coffee. From 20:25 to 20:51, the person is shown unpacking two ceramic brown mugs. The contents of the mugs are visible and appear to be coffee. Therefore, this time range is identified as the time when the vlogger is shown preparing coffee.

["29:20", "29:41"]: The video shows the vlogger in her bathroom at 28:00-28:18 applying makeup and then brushing her teeth. At 29:20-29:41, she is shown painting her nails. The presence of various makeup in both locations supports the query.

'''

```
reason = retrieval_qa(question="What is the main reason the vlogger
goes out again, where she also buys makeup?", answer_options=['
To go exercise', 'To eat chicken', 'To buy coffee mugs', 'To go
to the park'], video_segment=get_segment(start='16:09', end='
17:25'))
```

'''

The vlogger goes to a store and buys coffee mugs. Then she goes to another store and buys makeup. Therefore, the main reason she goes out again is to buy coffee mugs.

Final Answer: (3)

'''

```
final_answer = finish("The provided text from 'retrieval_qa'
indicates the vlogger buys makeup at a store called Kicks.
However, it also mentions buying coffee mugs. The question asks
```

Table 1: **EgoSchema Results.** We report accuracy on the evaluation subset of 500. Direct inference results use Gemini 1.5 Flash. Other results reported from respective method papers.

	Accuracy (%) $\uparrow$
SeViLA Yu et al. (2023)	25.7
LLoVi Zhang et al. (2023)	57.6
VideoAgent Wang et al. (2024a)	60.2
VideoTree Wang et al. (2024b)	66.2
Direct Inference	70.6
CAViAR	72.2

```

for the *main* reason for going out again. While makeup is
mentioned, the provided text strongly suggests the primary
reason for the second outing is to buy coffee mugs. Therefore,
the answer is (3). Final Answer: (3)"
'''
The provided text from 'retrieval_qa' indicates the vlogger buys
makeup at a store called Kicks. However, it also mentions
buying coffee mugs. The question asks for the *main* reason for
going out again. While makeup is mentioned, the provided text
strongly suggests the primary reason for the second outing is to
buy coffee mugs. Therefore, the answer is (3). Final Answer:
(3)

Critique:
A: 3. B: 4. C: 3.
Strategy B succeeds while Strategies A fails because Strategy A
looks at particular time ranges of the video and concludes it is
unable to find a correct answer, thus guessing.
While the approach is generally sound, Strategy A gives up upon not
finding an answer initially. Strategy C looks at a narrow range
and follows an implication to guess an answer. Thus Strategy B
is preferred.

Winning Strategy:
B

```

Listing 7: EgoSchema critic examples.

5 ADDITIONAL RESULTS

5.1 EGOschema

We also report results on the EgoSchema dataset Mangalam et al. (2023) for comparison with other approaches.

5.2 GPT-4O-MINI

To show the capabilities of our method with other base models, we obtained preliminary results with the GPT-4o-mini model. Due to cost, lack of credit availability, and query limits, we perform this proof of concept experiment on 25% of LVBench. We see a substantial gain in performance.

REFERENCES

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Table 2: **LVBench GPT-4o-mini Results.** We report accuracy on 25% of the evaluation set. Direct inference results use GPT-4o-mini. Other results reported from respective method papers.

	Accuracy (%) $\uparrow$
Direct Inference	43.2
CAViAR	49.0

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024a.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. *arXiv preprint arXiv:2405.19209*, 2024b.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023.