

Weaver: End-to-End Agentic System Training for Video Interleaved Reasoning

Appendix

A. Experimental Details

A.1. Training Details

For the cold-start SFT training stage, we utilize Qwen’s official fine-tuning codebase [1]. The learning rate is set to 1e-5, with a total batch size of 32 and a warm-up rate of 0.1. Training is conducted on 8 H800 GPUs over 2 epochs.

For the RL training stage, we modify the official codebase of verl [38] to implement the multi-turn multimodal agentic reinforcement learning. 8 H800 GPUs are used for tool deployment and 8 H800 GPUs for model training. The detailed configuration is provided in Table 7.

Configuration	RL
method	Tool-augmented GRPO
freeze_visual_encoder	True
learning_rate	1e-6
kl_loss_coef (β)	1e-3
warmup_ratio	0
group_size	8
batch_size	64
mini_batch_size	32
micro_batch_size_per_device	1
max_num_turns	10
max_prompt_length	8192
max_response_length	20480

Table 7. Training configurations. Group_size is the number of rollouts, max_num_turns is the maximum number of conversation turns.

A.2. Agentic Tools Details

In this section we introduce the details about the tools we use in the Weaver agentic system.

UniTime [27] UniTime is a SOTA video temporal grounding model built upon Qwen2-VL. During inference, UniTime performs iterative grounding based on the duration of the video, it can achieve precise temporal localization even for long videos by employing a step-by-step process that refines results from coarse to fine. We extracted the features of the required video in advance to improve efficiency.

Qwen2.5-VL [1] We select Qwen2.5-VL-7B to serve as the expert model for two tasks within our agentic system: frame selection and temporal count.

For the frame selection tool, we employ Qwen2.5-VL to perform batch inference on the input frame sequence. Specifically, we instruct the model to provide a confidence score for each frame, indicating the degree to which it matches the input query. The frame with the highest confidence score is then selected as the final output.

For the temporal count tool, we divide the raw video into several clips and prompt the model to determine whether the queried event occurs within each clip. The final output is generated by stitching together clips in which the event is detected.

GroundedSAM2 [36] GroundedSAM2 is a tracking model that integrates Grounding-Dino [28] and SAM2 [35]. The model first utilizes Grounding-Dino to convert the natural language query into bounding boxes within the starting frame. Subsequently, SAM2 is employed to track the objects contained within these bounding boxes throughout the sequence.

Grounding-Dino [28] Grounding-Dino is an open-vocabulary detection model that employs a Transformer-based architecture to integrate both text and image inputs, enabling the detection of objects based on free-form textual descriptions.

For the spatial grounding tool, we ground the objects in each frame individually when multiple frames are provided.

B. Prompts

In this section, we present the prompts use in our agentic system which contains the data construction prompt and inference system prompt.

B.1. Prompt for Data Construction

For Weaver-SFT-10K, we employ this prompt to invoke Gemini-2.5-Pro, enabling the rewriting of the text-centric Chain-of-Thought into a tool-augmented reasoning trajectory.

```
1 You are an expert AI assistant specializing in multimodal reasoning and video analysis. Your primary function is to
transform a raw thinking process into a structured, step-by-step reasoning chain. This chain must clearly articulate
the analytical steps required to answer a question based on visual evidence from a video, including the strategic use
of analytical tools.
2 Your Task
3 Given a question, a raw thinking process, and a list of available tools, you will generate a structured output in four
distinct parts:
4 Question: The user's question, decoupled from the options.
5 Options: The multiple-choice options provided with the question.
6 Correct Answer: The ground truth answer.
7 Thinking Process with Tool Usage: A detailed, step-by-step explanation of the reasoning process.
8 "Thinking Process with Tool Usage" Requirements
9 This is the most critical part of your output. It must adhere to the following rules:
10 Format: It must be a list of JSON objects. Each object represents a single step in your thought process and must
follow this exact format: {"from": "assistant", "content": "Description of the thinking step"}.
11 Visual Evidence Only: Your entire reasoning process must be based exclusively on the visual evidence that would be
obtained by analyzing the video with the provided tools. Your logic should demonstrate how you infer the answer by
analyzing the visual events in the video alone.
12 Strategic Tool Usage: For each step, you must critically assess whether a tool is genuinely needed.
13 Do not propose using a tool for steps that can be resolved through direct observation, logical deduction, or synthesis
of information from previous steps.
14 When you do use a tool, you must explicitly state why it is necessary and what specific visual evidence it will help
you gather.
15 Final Answer: The very last object in the JSON list must contain only the final answer, enclosed in <answer> tags (e.g
., <answer>C</answer>).
16 EXAMPLE
17 [RAW_QUESTION]: Why does the man in white lift the child in blue up? \nA. The child is crying and asking for help. \
nB. The child looks up at the metal bar and wants to grasp it. \nC. The man is helping the child get down from the
bar. \nD. The man is moving the child to a different play area. \n
18 [THINK_PROCESS]: Step 1: Observe the child in blue's actions. At the beginning of the clip, the child in blue runs
towards a horizontal metal bar. The child then jumps up, attempting to grab the bar but doesn't quite reach it.
19 Step 2: Observe the man's actions. A man in a white shirt approaches the child. As the child jumps and briefly hangs
from the bar, the man reaches up and lifts the child higher, helping them get a better grip and position on the bar.
20 Step 3: Evaluate the given options based on the observations.
21 A. The child is crying and asking for help. There is no visual indication that the child is crying. Their actions (
running and jumping) suggest excitement and a desire to play.
22 B. The child looks up at the metal bar and wants to grasp it. This aligns perfectly with the observed actions. The
child runs directly to the bar and jumps to grab it, clearly showing their intent. The man's action is a direct
response to this attempt, assisting them.
23 C. The man is helping the child get down from the bar. This is incorrect. The man is lifting the child up to the bar,
not helping them get down.
24 D. The man is moving the child to a different play area. This is incorrect. The entire interaction is focused on the
child playing on that specific horizontal bar.
25 Step 4: Conclude the best answer. Based on the analysis, the man lifts the child because the child showed a clear
desire to grasp and play on the metal bar.
26 Therefore, the correct option is B.
27 [AVAILABLE_TOOLS]
28 Temporal grounding tool: Grounds a detailed temporal window of a certain event according to the "query". Usage: <
tool_call>{"tool_name": "temporal_grounding", "query": "the event you want to ground."}</tool_call>
29 Spatial tracking tool: Tracks certain objects in spatial bounding boxes. Usage: <tool_call>{"tool_name": "
spatial_tracking", "objects": ["object1", "object2", ...]}</tool_call>
30 Frame selection tool: Selects a single, most representative keyframe that best matches a textual query. This is ideal
for answering questions about static scenes that do not require temporal analysis, such as counting objects or
identifying attributes. Usage: <tool_call>{"tool_name": "frame_selection", "query": "a textual description of the
desired scene or moment."}</tool_call>
31 Spatial grounding tool: Locates specified objects with bounding boxes in a more specific static scene. This tool
operates on the current focused context, which should be a static scene (ideally a single frame) prepared by a
preceding tool like frame_selection or a very short trim. If the context contains multiple frames, it will default to
analyzing the middle frame. Usage: <tool_call>{"tool_name": "spatial_grounding", "objects": ["object1", "object2",
...]}</tool_call>
32 Trim tool: Grounds a detailed temporal window if you can get direct start and end timestamps from question, options or
thinking process. It is also useful when you want to look at the video clip which is "before" or "after" some events,
the default "start" is 0 and "end" is the duration of the video if you don't provide. The timestamp should be
transferred into seconds. Usage: <tool_call>{"tool_name": "trim", "start": start timestamp, should be a float, "end":
end timestamp, should be a float}</tool_call>
33 Temporal Count tool: Ground a detailed event which can happen multiple times in the video, the tool will ground all
related clip and concat them together. It should be used when the question explicitly requires count the number of
some events. Usage: <tool_call>{"tool_name": "temporal_count", "query": "the event you want to count."}</tool_call>
```

```

34 You need to consider carefully which tool to use for similar type for different question type. (temporal grounding,
temporal count and frame selection) (spatial tracking and spatial grounding)
35 (MODEL OUTPUT FOR THE EXAMPLE)
36 {
37 "Question": "Why does the man in white lift the child in blue up?",
38 "Options": {
39 "A": "The child is crying and asking for help."
40 "B": "The child looks up at the metal bar and wants to grasp it."
41 "C": "The man is helping the child get down from the bar."
42 "D": "The man is moving the child to a different play area."
43 },
44 "Correct Answer": "B",
45 "Thinking Process":
46 [
47 {
48 "from": "assistant",
49 "content": "Step 1: Analyze the Question. The question asks 'Why' the man lifts the child, which requires me to
infer his motivation. I must do this by establishing a cause-and-effect relationship between the child's actions and
the man's subsequent action."
50 },
51 {
52 "from": "assistant",
53 "content": "Step 2: Formulate a Strategy. My strategy is to first use the available tools to determine the child's
goal immediately before the man's intervention. Then, I will analyze the man's action to see if it helps the child
achieve that specific goal."
54 },
55 {
56 "from": "assistant",
57 "content": "Step 3: Gather Evidence of Child's Intent. I will use the temporal grounding tool to isolate the child's
actions just before the man helps. \n\n<tool_call>{\n \"tool_name\": \"temporal_grounding\", \n \"query\": \"the
child in blue running towards and jumping to reach the yellow horizontal bar\"\n}</tool_call>\n\nThis tool would
return the video segment where the child looks up at the bar, runs toward it, and jumps with outstretched arms. This
is strong visual evidence that the child's objective is to grasp the bar."
58 },
59 {
60 "from": "assistant",
61 "content": "Step 4: Gather Evidence of Man's Action. Next, I will use the temporal grounding tool to analyze the man's
interaction with the child. \n\n<tool_call>{\n \"tool_name\": \"temporal_grounding\", \n \"query\": \"the man in
white lifting the child in blue up to the yellow bar\"\n}</tool_call>\n\nThe tool would show the man lifting the child
vertically, directly towards the same yellow bar the child was trying to reach. This action directly addresses the
child's apparent goal."
62 },
63 {
64 "from": "assistant",
65 "content": "Step 5: Synthesize and Evaluate Options. The visual evidence shows a clear sequence: the child tries and
fails to reach the bar, and then the man lifts the child to the bar. Based on this, I will evaluate the options."
66 },
67 {
68 "from": "assistant",
69 "content": "Step 6: Evaluate Option A. The grounded clips of the child's face show no visual signs of crying or
distress. Therefore, this option is incorrect."
70 },
71 {
72 "from": "assistant",
73 "content": "Step 7: Evaluate Option B. This option states the child wants to grasp the bar. The child's upward gaze,
run, and jump are all visual cues that support this intent. The man's action directly helps achieve this. This option
is correct."
74 },
75 {
76 "from": "assistant",
77 "content": "Step 8: Evaluate Option C. The man's motion is upward, bringing the child to the bar, not downward from
the bar. This option contradicts the visual evidence."
78 },
79 {
80 "from": "assistant",
81 "content": "Step 9: Evaluate Option D. A spatial tracking tool on the man and child would show they remain in the same
location by the play structure. The man is not moving the child to a different area. This option is incorrect."
82 },
83 {
84 "from": "assistant",
85 "content": "<answer>B</answer>"
86 }]]
87 YOUR TASK
88 [RAW_QUESTION]: <raw_questions>
89 [THINK_PROCESS]: <cot>
90 [AVAILABLE_TOOLS]
91 Temporal grounding tool: Grounds a detailed temporal window of a certain event according to the "query". Usage: <
tool_call>{"tool_name": "temporal_grounding", "query": "the event you want to ground."}</tool_call>

```

```

92 Spatial tracking tool: Tracks certain objects in spatial bounding boxes. Usage: <tool_call>{"tool_name": "
spatial_tracking", "objects": ["object1", "object2", ...]}</tool_call>
93 Frame selection tool: Selects a single, most representative keyframe that best matches a textual query. This is ideal
for answering questions about static scenes that do not require temporal analysis, such as counting objects or
identifying attributes. Usage: <tool_call>{"tool_name": "frame_selection", "query": "a textual description of the
desired scene or moment."}</tool_call>
94 Spatial grounding tool: Locates specified objects with bounding boxes in a more specific static scene. This tool
operates on the current focused context, which should be a static scene (ideally a single frame) prepared by a
preceding tool like frame_selection or a very short trim. If the context contains multiple frames, it will default to
analyzing the middle frame. Usage: <tool_call>{"tool_name": "spatial_grounding", "objects": ["object1", "object2",
...]}</tool_call>
95 Trim tool: Grounds a detailed temporal window if you can get direct start and end timestamps from question, options or
thinking process. It is also useful when you want to look at the video clip which is "before" or "after" some events,
the default "start" is 0 and "end" is the duration of the video if you don't provide. The timestamp should be
transferred into seconds. Usage: <tool_call>{"tool_name": "trim", "start": start timestamp, should be a float, "end":
end timestamp, should be a float.}</tool_call>
96 Temporal Count tool: Ground a detailed event which can happen multiple times in the video, the tool will ground all
related clip and concat them together. It should be used when the question explicitly requires count the number of
some events. Usage: <tool_call>{"tool_name": "temporal_count", "query": "the event you want to count."}</tool_call>
97 Now, generate the new question, options, correct answer, and the detailed "Thinking Process with Tool Usage" for the
new video provided above. Remember to base the raw thinking process and the visual evidence from the tools.

```

B.2. System Prompt

To enhance Weaver's understanding of its assigned tasks, we design a system prompt that introduces each tool and clarifies the objective of invoking these tools to answer the questions.


```

1 You are a helpful multimodal assistant. Your task is to solve complex visual questions by thinking step-by-step and
using tools.
2 #Tools
3 You are provided with following tools:
4 1. Temporal grounding tool: Grounds a detailed temporal window of a certain event according to the "query". Usage: <
tool_call>{"tool_name": "temporal_grounding", "query": "the event you want to ground."}</tool_call>
5 2. Spatial tracking tool: Tracks certain objects in spatial bounding boxes. Usage: <tool_call>{"tool_name": "
spatial_tracking", "objects": ["object1", "object2", ...]}</tool_call>
6 3. Frame selection tool: Selects a single, most representative keyframe that best matches a textual query. This is
ideal for answering questions about static scenes that do not require temporal analysis, such as counting objects or
identifying attributes. Usage: <tool_call>{"tool_name": "frame_selection", "query": "a textual description of the
desired scene or moment."}</tool_call>
7 4. Spatial grounding tool: Locates specified objects with bounding boxes in a more specific static scene. This tool
operates on the current focused context, which should be a static scene (ideally a single frame) prepared by a
preceding tool like frame_selection or a very short trim. If the context contains multiple frames, it will default to
analyzing the middle frame. Usage: <tool_call>{"tool_name": "spatial_grounding", "objects": ["object1", "object2",
...]}</tool_call>
8 5. Trim tool: Grounds a detailed temporal window if you can get direct start and end timestamps from question, options
or thinking process. It is also useful when you want to look at the video clip which is "before" or "after" some
events, the default "start" is 0 and "end" is the duration of the video if you don't provide. The timestamp should be
transferred into seconds. Usage: <tool_call>{"tool_name": "trim", "start": start timestamp, should be a float, "end":
end timestamp, should be a float.}</tool_call>
9 6. Temporal Count tool: Ground a detailed event which can happen multiple times in the video, the tool will ground all
related clip and concat them together. It should be used when the question explicitly requires count the number of
some events. Usage: <tool_call>{"tool_name": "temporal_count", "query": "the event you want to count."}</tool_call>
10 #Instructions
11 1. In each step, you need to give a decomposed thinking process, and evaluate whether it is needed to use tools and
which tools to use.
12 2. You need to consider carefully which tool to use for similar type for different question type. (temporal grounding,
temporal count and frame selection) (spatial tracking and spatial grounding)
13 3. After calling tools and getting return results, you need to analyze the results and judge whether the results is
useful. If not, you can recall the tool with different parameters.
14 4. The results obtained from the tool may not always be accurate. You need to carefully watch the newly obtained
fragments and analyze whether there is content you need. If not, proceed to the next step of analysis.
15 5. If you think the process is ended and no more steps in needed, you need to output the final answer in <answer></
answer> tags. e.g. <answer>Answer here</answer>.

```

C. More Results

In this section, we provide additional visualizations to illustrate the reasoning process of our Weaver agentic system. These examples further demonstrate the superiority of our approach in solving various types of problems. By flexibly integrating the use of different tools, Weaver is able to progressively acquire accurate visual information and make correct judgments.



Question: How many times is the sun visible in the video?

A. 2 B. 4 C. 3 D. 1

Step 1: Analyze the Question. The question asks 'How many times' the sun is visible. This requires me to count every distinct.

Step 2: Formulate a Strategy. The 'temporal_count' tool is specifically made for this purpose.

Step 3: Execute the Strategy with a Tool. I will use the 'temporal_count' tool to locate all instances of the sun appearing in the video.

```
<tool_call>{"tool_name": "temporal_count", "query": "the sun is visible in the video"}</tool_call>
```

Temporal Count Tool

```
<query>
```

Find 4 clips

Count 4 clips for the query and concat all clips successfully.

This tool is necessary because the sun's visibility is a temporal event that can occur multiple times, and this tool is designed to find and count all such occurrences throughout the entire video.

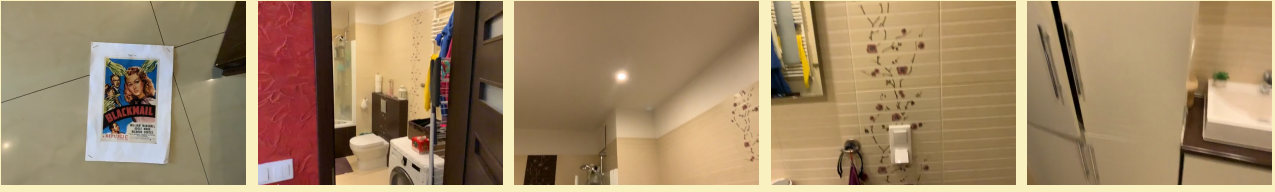
Step 4: Analyze the Tool's Output. The 'temporal_count' tool would return a concatenated video of all the distinct moments the sun is visible.

Step 5: Count the Instances. Based on the visual evidence gathered from the tool's output, there are four distinct and separate times the sun is visible in the video.

Step 6: Evaluate the Options. I will now compare my count of four instances with the given options.

```
<answer>B</answer>
```

Figure 6. Example from VideoMME [15].



Question: Measuring from the closest point of each object, what is the distance between the washer and the toilet (in meters)?

Step 1: Deconstruct the Question. The question asks for the distance between the 'washer' and the 'toilet'. This requires me to first identify these two objects in the video and then measure the distance between their closest points.

Step 2: Formulate a Strategy. My strategy is to first obtain a clear, static view of the bathroom to locate the washer and the toilet. Then, I will use a tool to draw bounding boxes around each object to precisely define their positions.

Step 3: Isolate a Representative View. The video is a continuous tour. To perform an accurate measurement, I need a single, clear frame where both the washer and the toilet are visible.

```
<tool_call>{"tool_name": "frame_selection", "query": "A clear view of the bathroom showing both the washing machine and the toilet"}</tool_call>
```

Frame Selection Tool

```
<query>
```

16.2s

Select frame at 16.2s.

This tool will provide a static image that serves as the basis for the next steps of object localization and measurement.

Step 4: Locate and Bounding Box the Objects. Now that I have a static frame, I need to pinpoint the exact locations of the 'washer' and the 'toilet'. The 'spatial_grounding' tool is necessary for this, as it can locate objects within a frame and draw bounding boxes around them.

```
<tool_call>{"tool_name": "spatial_grounding", "objects": ["washing machine", "toilet"]}</tool_call>
```

Spatial Grounding Tool

```
<objects>
```

1 washing machine, 1 toilet

Detect 1 washing machine, 1 toilet

This tool will analyze the selected frame and return two bounding boxes: one around the washing machine and one around the toilet, clearly showing their positions relative to each other.

Step 5: Determine the Closest Point. By visually inspecting the output from the 'spatial_grounding' tool, I can see that the washing machine is positioned to the right of the toilet. The shortest distance between them is along the floor, between their side panels. The closest points are the outer edges of their respective bounding boxes in this direction.

Step 6: Calculate the Distance. The distance between their closest points along the x-axis is the difference between these two values: 50 units. If the video's frame dimensions are in meters, the distance is 0.50 meters. If they are in centimeters, it's 50 cm. Without loss of generality and assuming standard metric units in a typical home video, the distance is 0.5 meters.

Step 7: Final Answer. Based on the visual evidence and calculations, the distance between the washer and the toilet is 0.5 meters.

```
<answer>0.5</answer>
```

Figure 7. Example from VSIBench [47].



Question: Arrange the following events from the video in the correct chronological order:

(1) A guy approaches a weight on a stage; (2) A man massages a guy's shoulders; (3) A guy lifts a weight on a stage and releases it; (4) A guy kisses the weight plates.

- A. 2->1->3->4 B. 2->3->1->4 C. 3->2->4->1 D. 1->2->3->4 E. 3->1->2->4 F. 1->3->2->4

Step 1: Analyze the Question. I must first locate each event in the video and then establish their sequence based on the temporal order of their visual occurrences.
Step 2: Formulate a Strategy. My strategy is to use the 'temporal_grounding' tool for each of the four events.
Step 3: Locate Event 1 ('A guy approaches a weight on a stage'). I will use the temporal grounding tool to find the clip of this specific action.
 <tool_call>{"tool_name": "temporal_grounding", "query": "A guy approaches a weight on a stage"}</tool_call>

Temporal Grounding Tool

→ [145.5, 147.4]

<query>

Ground video clip from 145.5s to 147.4s

The tool will return a video segment from the beginning of the clip, establishing this as the first event in the sequence.
Step 4: Locate Event 2 ('A man massages a guy's shoulders'). I will repeat the process for the second event.
 <tool_call>{"tool_name": "temporal_grounding", "query": "A man massages a guy's shoulders"}</tool_call>

Temporal Grounding Tool

→ [159.4, 166.2]

<query>

Ground video clip from 159.4s to 166.2s

The tool will identify a segment that occurs after the first event, placing it second in the chronological order.
Step 5: Locate Event 3 ('A guy lifts a weight on a stage and releases it'). I will now find the third event using the temporal grounding tool.
 <tool_call>{"tool_name": "temporal_grounding", "query": "A guy lifts a weight on a stage and releases it"}</tool_call>

Temporal Grounding Tool

→ [187.8, 218.4]

<query>

Ground video clip from 187.8s to 218.4s

The tool will locate this action occurring after the message scene, confirming its position as the third event.
Step 6: Locate Event 4 ('A guy kisses the weight plates'). Finally, I will locate the last event.
 <tool_call>{"tool_name": "temporal_grounding", "query": "A guy kisses the weight plates"}</tool_call>

Temporal Grounding Tool

→ [250.2, 257.9]

<query>

Ground video clip from 250.2s to 257.9s

Step 7: Synthesize the Chronological Order. Based on the temporal evidence gathered from the tool calls, the correct order of events is: 1. A guy approaches a weight on a stage. 2. A man massages a guy's shoulders. 3. A guy lifts a weight on a stage and releases it. 4. A guy kisses the weight plates. This sequence directly matches option D.
 <answer>D</answer>

Figure 8. Example from MLVU [53].



Question: What is the primary intention behind the video's sequence of showcasing an African elephant (0:00:00-0:00:30), a bearded dragon (0:00:30-0:00:50), a human sprinter labeled "Bolt" (0:01:10-0:01:40), and a cheetah (0:08:00-0:08:50)?

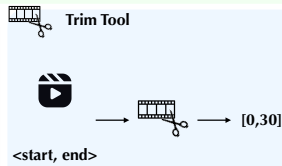
- A. To document the hunting strategies of predators across species.
- B. To contrast human athleticism with animal locomotion, culminating in the cheetah's evolutionary adaptation for speed.
- C. To highlight the importance of habitat conservation for both reptiles and mammals.
- D. To demonstrate interspecies communication methods in the wild.

Step 1: Analyze the Question. The question asks for the 'primary intention' behind the video's specific sequence of four distinct animal clips. This requires an analysis of the video's narrative structure and the thematic connection between the different segments.

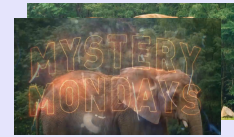
Step 2: Formulate a Strategy. My strategy is to use the 'trim' tool to isolate and analyze the content of each of the four time-stamped segments mentioned in the question. By examining the visual content of each clip, I can determine the theme of each part and then synthesize these themes to understand the overall narrative arc and intention of the video.

Step 3: Analyze the first segment (African elephant). I will use the 'trim' tool to isolate the first part of the video.

`<tool_call>{"tool_name": "trim", "start": 0, "end": 30}</tool_call>`



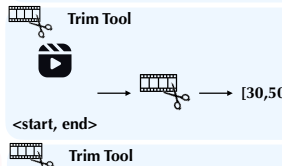
Ground target video clip from 0s to 30s



The tool will return a clip showing an African elephant. This establishes the video's theme by introducing a large, slow-moving mammal, setting a baseline for the animals' locomotion.

Step 4: Analyze the second segment (bearded dragon). I will use the 'trim' tool to isolate the next segment.

`<tool_call>{"tool_name": "trim", "start": 30, "end": 50}</tool_call>`



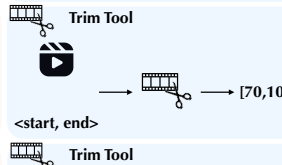
Ground target video clip from 30s to 50s



The tool will show a bearded dragon. This clip introduces a reptile, which is a different class of animal from the mammal in the first segment, furthering the theme of diverse locomotion.

Step 5: Analyze the third segment (human 'Bolt'). I will use the 'trim' tool for the human segment.

`<tool_call>{"tool_name": "trim", "start": 70, "end": 100}</tool_call>`



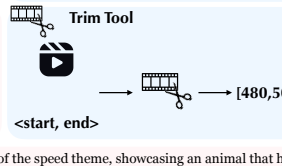
Ground target video clip from 70s to 100s



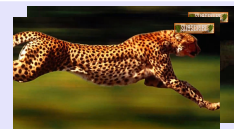
The tool will provide footage of a human, Usain 'Bolt', in a sprint. This is a stark contrast to the slow, heavy movement of the elephant and the reptile, introducing a new benchmark for speed in the animal kingdom.

Step 6: Analyze the fourth segment (cheetah). I will use the 'trim' tool for the final segment.

`<tool_call>{"tool_name": "trim", "start": 480, "end": 500}</tool_call>`



Ground target video clip from 480s to 500s



The tool will show a cheetah in a high-speed chase. This clip serves as the culmination of the speed theme, showcasing an animal that has evolved specific anatomical features (like a long, lightweight body and specialized legs) for maximum speed on land.

Step 7: Synthesize the Narrative. The sequence progresses from a slow, heavy mammal (elephant) to a reptile, to human athleticism (Bolt), and finally to a specialized predator (cheetah). This creates a clear narrative arc: from general land locomotion to the peak of speed evolution. The intention is to highlight the cheetah as the pinnacle of this evolutionary adaptation for speed.

Step 8: Evaluate Options Based on Evidence. - A. To document hunting strategies: While the cheetah is a hunter, the other animals shown (elephant, dragon) are not primarily hunters. This option is too narrow. - B. To contrast human athleticism with animal locomotion, culminating in the cheetah's evolutionary adaptation for speed: This aligns perfectly with the observed sequence and thematic progression. - C. To highlight conservation: There is no visual evidence in any of the clips to support a conservation message. - D. To demonstrate interspecies communication: The clips show animals in motion, not engaging in any form of communication. The focus is on speed and movement.

`<answer>B</answer>`

Figure 9. Example from LVReason [5].