

VideoThinker: Building Agentic VideoLLMs with LLM-Guided Tool Reasoning

Supplementary Material

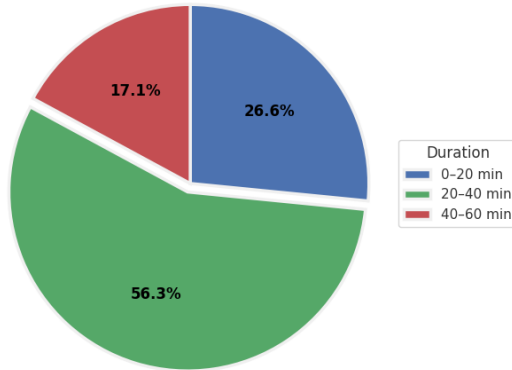


Figure 1. Distribution of Video Durations in CoTs.

001 A. Supplementary Materials

002 A.1. Prompt Details

003 As shown in Figure 2, we provide the detailed Tool reason-
004 ing prompt used in our framework. This prompt specifies
005 the expected output format for tool calls, defines all avail-
006 able tools, and outlines the instructions for each task. For
007 comparison, the Direct reasoning prompt is defined as fol-
008 lows:

009 Question: {question}, output a single letter. Best
010 option:

011 A.2. Tools

012 As shown in Table 1, our framework provides a set of tools
013 for long-form video reasoning at multiple levels of granu-
014 larity. The toolkit combines retrieval modules, such as Clip
015 Retrieval and Subtitle Retrieval, with zoom-level tools, in-
016 cluding Frame Zoom, Subtitle Zoom, and Caption Zoom, to
017 support both coarse- and fine-grained analysis. The Subtitle
018 Summary module further offers high-level textual abstrac-
019 tion. Together, these tools enable effective and interpretable
020 reasoning over long-duration videos.

021 A.3. Training Data

022 We analyze the distribution of video lengths in the CG-
023 bench synthetic dataset, as illustrated in Figure 1. More than
024 half of the videos have durations between 20 and 40 min-
025 utes, while those lasting 40- 60 minutes account for 17.1%.
026 Most data samples involve approximately 3-5 tool calls.

```

You are in an environment with access to video analysis tools. \nTool call format:
<tool_call>
  {"name": "tool_name", "arguments": {"param": value}}
</tool_call>
Available tools:
# Tools
You may call one or more functions to assist with the user query.
<tools>
<tool_call>
  <name>frame_zoom</name>
  <function>
    { "name": "frame_zoom",
      "description": "Time range in seconds as [start_time, end_time]. The smaller the time
interval, the more detailed the analysis will be. This tool is useful when zoom in the temporal
interval to get the visual content of interval. ",
      "parameters": {"interval": {"type": "array", "items": {"type": "number"}}},
      "required": ["interval"] }
    </function>
  </tool_call>
<tool_call>
  <name>subtitle_zoom</name>
  <function>
    { "name": "subtitle_zoom",
      "description": "Locate the interval where subtitle appears. This tool is useful when the
question mentions subtitle.",
      "parameters": {"interval": {"type": "array", "items": {"type": "number"}}},
      "required": ["interval"] }
    </function>
  </tool_call> .....
---
## Task Instructions: Use the available tools as needed, following the XML-style format guidelines,
to answer the question.

```

Figure 2. The prompt is designed to enable VideoLLM to think with videos using the available tools.

027 A.4. Implementation Details

028 We present the training script implemented with Swift, as
029 shown in Figure 3.

030 A.5. More results

031 Table 2 presents the comparison results on LVBench across
032 six tasks, including Entity Recognition, Event Understanding,
033 Key Information Retrieval, Temporal Grounding, Reasoning,
034 and Summarization. Our method VideoThinker achieves the best overall performance with an average
035 accuracy of 48.9%, comparable to GPT-4o (48.9%) and
036 clearly outperforming previous LLM-based agents such as
037 VideoTree (28.8%) and VideoAgent (29.3%). ReZoom-
038 V also shows strong results on KIR (58.0%) and ER
039 (49.6%), demonstrating its effectiveness in retrieval tasks.
040 We present additional LVBench cases in Figure 4-9.
041

Tool	Parameters	Description
Clip_Retrieval	video path, query, topk	Retrieves the top- k video clips most relevant to the given query based on semantic similarity.
Subtitle_Retrieval	video path, query, topk	Retrieves the top- k subtitle segments relevant to the query from Whisper-transcribed subtitles based on semantic similarity.
Subtitle_Summary	video path, query	Summarizes subtitles using an LLM to provide concise, query-focused contextual understanding for reasoning.
Frame_Zoom	video path, interval	Extracts video frames from the specified time interval.
Subtitle_Zoom	video path, interval	Extracts subtitles from the specified time interval.
Caption_Zoom	video path, interval	Extracts visual captions from the specified time interval.

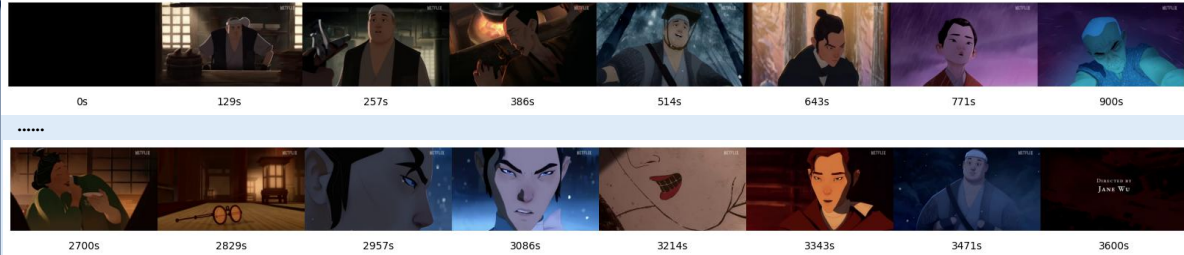
Table 1. Tool suites for long-form video reasoning, integrating retrieval-based and zoom-level analysis capabilities.

Methods	ER (%)	EU (%)	KIR (%)	TG (%)	Rea (%)	Sum (%)	Overall (%)
<i>VideoLLMs</i>							
Gemini-1.5-Pro [4]	32.1	30.9	39.3	31.8	27.0	32.8	33.1
GPT-4o [2]	48.9	49.5	48.1	40.9	50.3	50.0	48.9
Qwen2.5-VL-72B [1]	–	–	–	–	–	–	47.4
VideoChat-Flash [3]	51.1	46.0	49.0	38.9	48.5	34.5	48.2
<i>Agentic LLMs</i>							
VideoTree [6]	30.3	25.1	26.5	27.7	31.9	25.5	28.8
VideoAgent [5]	28.0	30.3	28.0	29.3	28.0	36.4	29.3
VideoThinker (Ours)	49.6	48.5	58.0	43.6	45.5	35.2	48.9

Table 2. Comparison on LVBench across various tasks: Entity Recognition (ER), Event Understanding (EU), Key Information Retrieval (KIR), Temporal Grounding (TG), Reasoning (Rea), Summarization (Sum), and Overall performance.

```
swift sft \  
  --model Qwen2.5-VL-7B-Instruct \  
  --train_type lora \  
  --dataset vcbench/train.jsonl \  
  --load_from_cache_file true \  
  --torch_dtype bfloat16 \  
  --num_train_epochs 3 \  
  --per_device_train_batch_size 1 \  
  --per_device_eval_batch_size 1 \  
  --learning_rate 1e-4 \  
  --lora_rank 64 \  
  --lora_alpha 16 \  
  --freeze_vit True \  
  --target_modules all-linear \  
  --gradient_accumulation_steps 2 \  
  --eval_steps 50 \  
  --save_steps 500 \  
  --save_total_limit 5 \  
  --logging_steps 5 \  
  --output_dir save_vcbench_3b \  
  --warmup_ratio 0.05 \  
  --dataloader_num_workers 8 \  
  --use_chat_template True \  
  --max_length 200000
```

Figure 3. The training script with Swift.



Tool_Prompt: You are in an environment with access to video analysis tools.

What year appears in the opening caption of the video?

(A) 1636 (B) 1366 (C) 1363 (D) 1633



[Begin of Agentic Thinking] To determine the year that appears in the opening caption of the video, I need to locate the very beginning of the video and examine the visual content where the caption is likely displayed.

Step 1: Start by analyzing the earliest part of the video, focusing on the opening scene. I will use the 'Frame_Zoom' tool to narrow down to the first few seconds where opening captions typically appear.

```
<tool_call>
{"name": "Frame_Zoom", "arguments": {"interval": [0, 10]}}
</tool_call>
```



The zoom into the interval [0, 10] revealed the start of the video with "A NETFLIX SERIES" and some animated fire effects, but no visible year in the caption. However, opening captions sometimes appear slightly after the initial logo or animation.

Step 2: Expand the analysis slightly to cover a broader opening segment, just after the Netflix branding, to ensure we don't miss any text that might appear next. I will now zoom into the interval [10, 20] to check for any subsequent captions.

```
<tool_call>
{"name": "Frame_Zoom", "arguments": {"interval": [10, 20]}}
</tool_call>
```



The zoom into the interval [10, 20] revealed the text: **"In 1633, Japan closed its borders to the outside world."** This clearly indicates that the year appearing in the opening caption is **"1633"**. **[End of Agentic Thinking]**

Final Answer:
D 1633.

Figure 4. VideoThinker's agentic tool reasoning on LVBench, testing its ability to retrieve key information from a 61-minute video (Cm73ma6lbc).



0s 129s 257s 386s 514s 643s 771s 900s

.....

Tool_Prompt: You are in an environment with access to video analysis tools.

Why does the protagonist tie iron to her limbs?

(A) Because she wants to strengthen her body (B) Because irons are decorations (C) Because she has superpower and iron surpress it

(D) Because tying iron is a culture tradition

[Begin of Agentic Thinking] To determine why the protagonist ties iron to her limbs, I need to analyze the context in which this action occurs. Since the question involves a narrative or symbolic reason (not purely visual), I should first look for any relevant subtitles or dialogue that might explain her motivation. I will start by gathering hints from the subtitles that could provide context about the protagonist's actions.

```
<tool_call>
  {"name": "Subtitle_Summary", "arguments": {}}
</tool_call>
<tool_use_result>
Subtitle_Summary
Summary: The subtitles do not contain any information about the protagonist tying iron to her limbs, nor do they explain the reason for such an action. None of the provided dialogue or scenes reference iron being tied to limbs for strength, decoration, suppression of powers, or cultural tradition. Therefore, based on the given subtitles, there is no clear evidence to support or explain any of the answer choices.
</tool_use_result>
```

Since the subtitles do not provide a clear explanation, I will now attempt to locate specific scenes where the protagonist is seen tying iron to her limbs. This visual context may help infer the reason based on surrounding actions or symbolism. I will search for a scene that matches the description of "the protagonist tying iron to her limbs."

```
<tool_call>
  {"name": "Clip_Retrieval", "arguments": {"text": "protagonist tying iron to her limbs", "top_k": 3}}
</tool_call>
<tool_use_result>
Clip_Retrieval
The interval: [(2140, 2160), (2450, 2460)]
</tool_use_result>
```

I have identified two time intervals where the protagonist is seen tying iron to her limbs: [2140, 2160] and [2450, 2460]. To analyze this behavior in detail, I will first zoom into the earlier and longer interval [2140, 2160] to observe the context and actions more clearly.

```
<tool_call>
  {"name": "Frame_Zoom", "arguments": {"interval": [2140, 2160]}}
</tool_call>
```



The zoomed-in analysis of the interval [2140, 2160] suggests that the protagonist is using iron as part of a training exercise, specifically in connection with a forge and physical conditioning. The context implies that the iron is being used to strengthen her body through resistance, not for decoration, cultural tradition, or suppression of powers. To ensure this interpretation is consistent, I will now examine the second interval [2450, 2460] where the same action occurs, to confirm the pattern and rule out alternative explanations.

```
<tool_call>
  {"name": "Frame_Zoom", "arguments": {"interval": [2450, 2460]}}
</tool_call>
```



The analysis of the second interval [2450, 2460] reinforces the conclusion from the first interval: the protagonist ties iron to her limbs as part of physical preparation, likely to strengthen her body for an upcoming challenge or battle. The deliberate and focused manner in which she applies the iron, combined with the absence of any symbolic or decorative context, rules out options (B) and (D). There is also no indication that she possesses superpowers or that the iron suppresses them, eliminating option (C). Instead, both visual intervals consistently support the idea that the iron serves a functional, physical training purpose. **[End of Agentic Thinking]**

Final Answer:
A. Because she wants to strengthen her body

Figure 5. VideoThinker's agentic tool reasoning on LVBench, testing its reasoning ability from a 61-minute video (Cm73ma6Ibcs).



.....

Tool_Prompt: You are in an environment with access to video analysis tools.

What does the woman in red do after she collides with the man?

- (A) She picks up things on the ground on her own and throws them away (B) She picks up things on the ground with the man
(C) She picks up things on the ground on her own (D) She picks up things on the ground with a passerby



[Begin of Agentic Thinking] To determine what the woman in red does after she collides with the man, I need to first locate the moment of collision between the woman in red and the man. I will start by searching for a scene that matches this description.

```
<tool_call>
{"name": "Clip_Retrieval", "arguments": {"text": "woman in red collides with man", "top_k": 1}}
</tool_call>
```

```
<tool_use_result>
Clip_Retrieval
The interval: [(850, 860)]
</tool_use_result>
```

The collision between the woman in red and the man occurs in three possible intervals: [850, 860], [940, 950], and [2160, 2170]. To determine which one is most relevant and to analyze the aftermath of the collision, I will first zoom into the earliest interval [850, 860] to examine the visual details.

```
<tool_call>
{"name": "zoom_interval", "arguments": {"interval": [850, 860]}}
</tool_call>
```



Since both visual interval have not yielded a clear answer, I will now attempt to locate more precise moments where actions like "picking up things" occur, especially in the context of a collision.

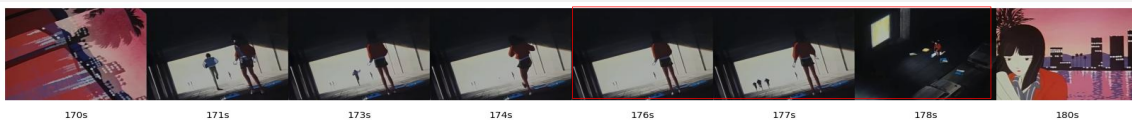
I will use 'Clip_Retrieval' with a more focused description: "woman picking up things after collision".

```
<tool_call>
{"name": "Clip_Retrieval", "arguments": {"text": "woman picking up things after collision", "top_k": 3}}
</tool_call>
```

```
<tool_use_result>
Clip_Retrieval
The interval: [(170, 180), (940, 950), (2960, 2970)]
</tool_use_result>
```

The new intervals where a "woman picking up things after collision" may have occurred are [170, 180], [940, 950], and [2960, 2970]. Let's now zoom into [170, 180], which could be an earlier event in the video.

```
<tool_call>
{"name": "Frame_Zoom", "arguments": {"interval": [170, 180]}}
</tool_call>
```



The zoomed-in analysis of the interval [170, 180] clearly shows the woman in red picking up items from the ground in a dimly lit room, with no interaction with the man or any passerby. She performs the action independently. **[End of Agentic Thinking]**

Final Answer:

C. She picks up things on the ground on her own

Figure 6. VideoThinker's agentic tool reasoning on LVBench, testing its event-understanding ability using a 55-minute video (q01CUy_gwdU).



.....

Tool_Prompt: You are in an environment with access to video analysis tools.

What color is the first plane?

(A) Yellow (B) Red (C) White (D) Blue



[Begin of Agentic Thinking] To determine the color of the first plane, I need to locate when the plane appears in the video. Since the question refers to the "first" plane, I should identify the earliest occurrence of a plane in the scene.

I will start by searching for scenes that match the description of a "plane" to get the relevant time intervals.

```
<tool_call>
  {"name": "Clip_Retrieval", "arguments": {"text": "plane", "top_k": 3}}
</tool_call>
```

```
<tool_use_result>
```

```
Clip_Retrieval
```

```
The interval: [(570, 580), (590, 610)]
```

```
</tool_use_result>
```

The intervals where the plane appears are (570, 580) and (590, 610). Since we are looking for the "first" plane, the earliest occurrence is in the interval (570, 580). I will now zoom into this interval to get a more detailed view of the scene and observe the color of the plane.

```
<tool_call>
  {"name": "Frame_Zoom", "arguments": {"interval": [570, 580]}}
</tool_call>
```

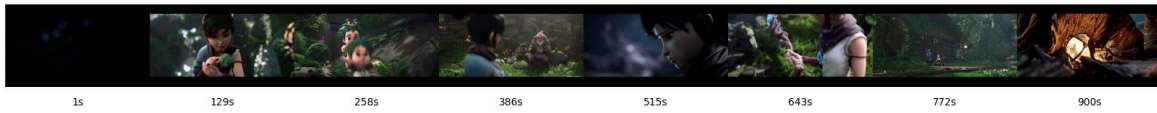


The zoomed-in analysis of the interval [570, 580] clearly indicates that the first plane is yellow. The description mentions that the plane's body and wings are painted in a bright yellow color, and there are no red or blue planes in this sequence. **[End of Agentic Thinking]**

Final Answer:

A. Yellow

Figure 7. VideoThinker's agentic tool reasoning on LVBench, testing its event-recognition ability using a 55-minute video (q01CUy_gwdU).



.....

Tool_Prompt: You are in an environment with access to video analysis tools.

What color are the runes on the bottom of the girl's feet at 38:18 in the video?

(A) Golden (B) White (C) Red (D) Black



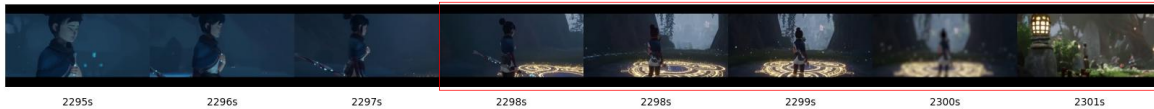
[Begin of Agentic Thinking] To determine the color of the runes on the bottom of the girl's feet at 38:18, I need to analyze the visual content around that timestamp. Since the question refers to a very specific visual detail (runes on the feet) at an exact time (38:18), I should first zoom into a narrow interval around that moment to inspect the frame details.

I'll start by zooming into an interval centered around 38:18, say from 38:15 to 38:21 (2295 to 2301 seconds), to get a close look at the scene.

<tool_call>

```
{"name": "Frame_Zoom", "arguments": {"interval": [2295, 2301]}}
```

</tool_call>



The zoomed-in analysis of the interval around 38:18 (2295 to 2301 seconds) clearly indicates that the runes on the bottom of the girl's feet are glowing with a warm, golden light. The visual description confirms that the runes are golden in color. **[End of Agentic Thinking]**

Final Answer:

A. Golden

Figure 8. VideoThinker's agentic tool reasoning on LVBench, testing its temporal grounding ability using a 64-minute video (28C1eC8cZks).



..... 1s 129s 258s 386s 515s 643s 772s 900s

Tool_Prompt: You are in an environment with access to video analysis tools.

What happens between 22:17-42:05?

(A) Reporter spends three months with the Penan people, understanding their psychological characteristics and documenting the lifestyle of some Penan people living in the city

(B) Reporter spends three months with the Penan people, learning about their history, lifestyle, how to hunt, how to use darts, and also documented the lifestyle of some Penan people living in the city

(C) Reporter spends three months with the Penan people, learning about their history, lifestyle, how to hunt, how to use darts, and also recorded the lifestyle of some Penan people living in the village

(D) Reporter spends three months with the Penan people, understanding their psychological characteristics and recording the lifestyle of some of the Penan people living in the countryside

[Begin of Agentic Thinking] The question asks about what happens between 22:17 and 42:05 in the video. This is a 19-minute and 48-second interval, which is quite long, so we need to analyze the content during this period carefully.

First, let's convert the timestamps to seconds:

- 22:17 = (22 × 60) + 17 = 1337 seconds
- 42:05 = (42 × 60) + 5 = 2525 seconds

So, we are analyzing the interval [1337, 2525].

Since the options involve themes like "learning history, lifestyle, hunting, darts," and "documenting lifestyle in city/village/countryside," we should determine whether these topics are discussed or shown in the video during this interval.

A good starting point is to get subtitle hints from this interval to understand the main themes discussed.

```
<tool_call>
{"name": "Subtitle Summary ", "arguments": {"text": "22:17 to 42:05"}}
</tool_call>
```

```
<tool_use_result>
Subtitle Summary
The hints: Between 22:17 and 42:05, the reporter documents his time living with the Penan people, learning their survival skills, history, and traditional practices such as hunting with blowpipes and using poison darts. He stays with semi-nomadic families in jungle camps, observes their daily life, and records their methods of foraging, building shelters, and using medicinal plants. The focus is on their traditional lifestyle in the forest, not in the city. Key activities include harvesting sago, making bark cloth, and learning about their ethnobotanical knowledge..
</tool_use_result>
```

The subtitle hints indicate that during the interval [22:17–42:05], the reporter:

- Lives with the Penan people
- Learns their history and traditional practices
- Learns hunting with blowpipes and using poison darts
- Observes daily life in jungle camps (i.e., village/forest setting)
- Records foraging, shelter-building, medicinal plant use, sago harvesting, and bark cloth making
- Focus is on traditional lifestyle in the forest, ****not in the city****

Now, let's compare this with the options:

(A) Mentions "psychological characteristics" and "living in the city" — not supported by the hints

(B) Mentions learning history, lifestyle, hunting, darts — supported; but says "living in the city" — contradicted

(C) Mentions learning history, lifestyle, hunting, darts — supported; and "living in the village" — consistent with jungle camps/forest

(D) Mentions "psychological characteristics" — not mentioned; "countryside" is vague, but "understanding psychological characteristics" is not supported

So, (C) appears to be the best fit. But to confirm, let's verify if the term "village" or "jungle camp" appears in the scene description during this interval.

.....

```
(tool_use)
The subtitle at [1611–1620] seconds (~26:51–27:00) states:
**"The villagers have a hornbill as a pet."**
Followed by: **"Hornbills are the emblem of Sarawak."**
This confirms that the term "village" is used in context to describe the Penan or local community living in a traditional setting (jungle/village), not in an urban environment. The mention of a hornbill as a pet also aligns with traditional forest life.
Combined with earlier evidence:
- The reporter learns history, lifestyle, hunting, and dart use
- Focus is on jungle camps and forest living
- "City" is mentioned once but without contextual support about documenting urban Penan lifestyle
- "Village" is mentioned in context of traditional life
Thus, option (C) is fully supported:
> *(C) Reporter spends three months with the Penan people, learning about their history, lifestyle, how to hunt, how to use darts, and also recorded the lifestyle of some Penan people living in the village**
This matches the video content both visually (jungle camp scenes) and through subtitles (village, traditional practices). [End of Agentic Thinking]
```

Final Answer:
C

Figure 9. VideoThinker's agentic tool reasoning on LVBench, testing its summarization ability using a 53-minute video (TiQBTesZUIJQ).

042 **References**

- 043 [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin
044 Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun
045 Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint*
046 *arXiv:2502.13923*, 2025. 2
- 047 [2] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman,
048 Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda,
049 Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv*
050 *preprint arXiv:2410.21276*, 2024. 2
- 051 [3] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan
052 Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yanan He,
053 Chenting Wang, et al. Videochat-flash: Hierarchical com-
054 pression for long-context video modeling. *arXiv preprint*
055 *arXiv:2501.00574*, 2024. 2
- 056 [4] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell,
057 Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent,
058 Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking mul-
059 timodal understanding across millions of tokens of context.
060 *arXiv preprint arXiv:2403.05530*, 2024. 2
- 061 [5] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-
062 Levy. Videoagent: Long-form video understanding with large
063 language model as agent. In *European Conference on Com-*
064 *puter Vision*, pages 58–76. Springer, 2024. 2
- 065 [6] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong
066 Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal.
067 Videotree: Adaptive tree-based video representation for llm
068 reasoning on long videos. In *Proceedings of the Computer Vi-*
069 *sion and Pattern Recognition Conference*, pages 3272–3283,
070 2025. 2