LongVideo-R1: Smart Navigation for Low-cost Long Video Understanding

Supplementary Material

The supplementary document provides (1) details of our hierarchical video definition in A;(2)comprehensive implementation details including the prompts we used for data generation and the experiment setup in B;(3)more qualitative examples and ultra long video examples in C;(4)failure examples and analysis in D.

A. Hierarchical Video Definition

To enable efficient localization of relevant segments within long videos without requiring any preprocessing, we represent each video using a hierarchical tree structure. Given a tree depth D and width W, the root node corresponds to the entire video. The root is evenly divided into W child segments, each of which is recursively divided into another W sub-segments. Repeating this process yields a hierarchical video tree with D levels.

In practice, we set $D = 3$, and choose W adaptively according to video length. The leaf-level segment length is fixed at 16 seconds. Let *Duration* denote the total video length, then the number of leaf segments is $\mathit{Duration}/16$. We determine the width as:

$$W = \left(\frac{\mathit{Duration}}{16}\right)^{1/D},$$

which typically lies within 4 to 8 across all datasets.

To ensure that deeper nodes receive fine-grained visual signals, we adjust both frame sampling rate and spatial resolution according to the hierarchy. For the caption model, we use {256, 128, 64, 32} frames from level 0 to level 3, respectively. The corresponding image resolutions are set to

$$\frac{512}{2\sqrt{2}}, \frac{512}{2}, \frac{512}{\sqrt{2}}, 512,$$

ensuring that the overall number of visual tokens remains approximately constant across levels. This design guarantees consistent computation cost per caption call while enabling finer temporal and spatial detail as the model traverses downward in the hierarchy.

B. Implementation Details

B.1. Environment Setup

During both data generation and model evaluation, we use Qwen-2.5-VL-72B as our *video_cap* model and Qwen-2.5-VL-32B as our *video_qa* model. These two components can be replaced by other models; for instance, Qwen-2.5-VL-32B is also capable of serving as an effective captioning model.

For CoTWT data generation, we employ GPT-5 as the central reasoning model. All SFT and RL training is conducted on a cluster with 8 NVIDIA A100 GPUs (80GB). Mixed-precision training and FSDP sharding are used to maximize training throughput.

Configuration	SFT	RL
Model init	Qwen3-8B	LongVideo-R1-SFT
LLM sequence length	32768	32768
Learning rate	1×10^{-5}	1×10^{-6}
Learning rate schedule	cosine decay	constant
Training epochs	3	2
Global batch size	32	12
Training steps	384	696
Rollout numbers	-	16

Table 8. Training hyper-parameters.

B.2. Data Generation

We use a proprietary GPT-5 model to generate the CoTWT supervision signals. Since CGBench provides timestamp annotations for each question, we use CGBench as the primary source for CoTWT construction. CGBench contains approximately 1200 videos and 12,000 QA pairs. We use 800 videos and around 8000 QA pairs to construct the SFT data; after filtering, we obtain 5600 high-quality CoTWT trajectories.

The remaining 400 videos and approximately 4200 QA pairs are reserved for RL training.

The prompts used for data generation and caption extraction are listed in Table 9 and Table 10 of the appendix. Initially, we provided only the root-level caption as GPT-5's starting context, but this resulted in unstable behavior and low accuracy (around 30%). We found that providing the W child captions from the highest-level nodes as initial information substantially improves stability and accuracy.

For all datasets (CGBench, LVBench, VideoMME-Long), the tree width W is set between 4 and 8. For EgoSchema, due to its large number of short 2-minute videos, we set W between 3 and 8.

B.3. Training hyper-parameters

We adopt Qwen3-8B as the central reasoning model for both SFT and RL training. The model receives the W highest-level captions as its initial observation and interacts with the video tree through a sequence of tool calls.

Training consists of two phases:

Table 9. System prompt for data generation.

[BEGIN OF GOAL]
You are a reasoning assistant designed to answer questions about a long video through hierarchical captions. The video is organized into three levels of temporal granularity:
1. High-level: The video is divided into **width** major segments.
2. Medium-level: Each High-level segment is further divided into **width** sub-segments.
3. Low-level: Each Medium-level segment is further divided into **width** finer sub-segments.
You will be asked a question about the video.
At the beginning, you are given **only the High-level captions**.
Your goal is to answer the question as accurately as possible.
[END OF GOAL]

[BEGIN OF REASONING AND TOOL USAGE INSTRUCTIONS]
1. Reason first:
Before taking any action, carefully analyze whether the current information (captions you already have) is sufficient to answer the question.
2. If sufficient:
Directly provide your final answer inside ⟨answer⟩⟨/answer⟩ tags.
3. If insufficient:
Identify which part(s) of the video might contain the needed information. Then use one of the following tools:
- To obtain finer captions:
⟨tool⟩get_caption((high_segment_id, medium_segment_id, low_segment_id))⟨/tool⟩
- Each of the three IDs is an integer from 1 to **width**.
- To request a Medium-level caption, provide (high_segment_id, medium_segment_id) only.
- To request a Low-level caption, provide the full triplet (high_segment_id, medium_segment_id, low_segment_id).
- To query visual information from the actual video segment:
⟨tool⟩video_qa((high_segment_id, medium_segment_id, low_segment_id), query)⟨/tool⟩
- This tool sends the corresponding Low-level video segment to a specialized video QA module.
- The query should specify what exact information you need (e.g., "what color is the person's shirt?", "what object is on the table?").
- You may only use video_qa after you have already retrieved the corresponding Low-level caption for that segment.
4. Restriction:
In each reasoning round, you may only call one tool (either 'get_caption' or 'video_qa') once to obtain new information.
[END OF REASONING AND TOOL USAGE INSTRUCTIONS]

[BEGIN OF FORMAT INSTRUCTIONS]
Your reasoning and actions must follow this structure exactly: ⟨think⟩Your internal reasoning process here. Analyze what information you have, what is missing, and which part might be relevant.⟨/think⟩ ⟨tool⟩(get_caption or video_qa call here, if needed)⟨/tool⟩ or ⟨think⟩ . . . ⟨/think⟩ ⟨answer⟩Your final answer here (only when you are confident the information is sufficient).⟨/answer⟩
[END OF FORMAT INSTRUCTIONS]

SFT. We train the model to imitate CoTWT trajectories by predicting reasoning process, search actions and answers. This stage helps the model acquire hierarchical search behavior and structured video reasoning skills.

RL. During RL, we pre-extract all hierarchical captions to accelerate training, while the *video_qa* tool is invoked in real time. Qwen-2.5-VL-32B is deployed on two GPUs to serve as the *video_qa* module, while the remaining six GPUs are dedicated to RL training.

The detailed hyper-parameters are listed in Table 8.

B.4. Time consumption calculation

The total inference time of LongVideo-R1 consists of three components: (1) the forward-pass time of the reasoning model T_1, (2) the captioning time required for processing a

Table 10. Video Caption Model Prompt.

You are a video understanding expert. Please create a detailed description with timestamp information for the current video clip (which contains multiple frames arranged in chronological order).
You are given **num_frame** uniformly sampled frames from the video, along with the timestamp (in seconds) of each frame in the entire video.
Description Guidelines:
-Dialogue Description Guidelines:
1)In addition to video frames, subtitle information for this video segment is also provided.
2)The output description must faithfully include the given subtitle content. Do not invent or add dialogues that are not provided. Avoid redundant repetition, maintain the original order of the lines, and ensure the sentences flow smoothly.
3)Your output should be around **num_words** words.
-Whenever reasonable, include the provided timestamps in your description.
1)For multiple frames with short intervals that depict the same continuous action, you may merge them into a single description.
2)For example:This video begins at 0.0s with a scene featuring two individuals seated outdoors, engaging in a conversation. The subtitles indicate they are discussing the impact of the pandemic on their ability to shoot videos at a bar. By 14.0s, the dialogue shifts to their newfound regular appearances on a show called Scam Nation. At 28.0s, the conversation turns to the promotion of a product named Kraken, encouraging viewers to visit a website for...
Output Format:
Your response should be in the following format, wrapped with ⟨caption⟩⟨/caption⟩ tags: "⟨caption⟩This clip (video) XXX⟨/caption⟩".

video segment T_2, and (3) the time required for the video_qa model to answer a query T_3. Let the average number of calls to LongVideo-R1, the video_cap model, and the video_qa model be C_1, C_2, and C_3, respectively. Then the expected time cost for answering one question is:

$$T = C_1T_1 + C_2T_2 + C_3T_3.$$

Using VideoMME-Long as an example, the model requires on average 10.5 reasoning rounds per question. Therefore, $C_1 = 10.5$. The video_qa model is invoked infrequently, with an average of $C_3 = 0.36$ calls per question.

Since the average tree width for VideoMME-Long is $W = 5$, the initial step requires obtaining the W highest-level captions. During the subsequent reasoning process, every reasoning step except the one that triggers a video_qa call requires an additional caption. Thus the expected number of caption calls is:

$$C_2 = W + C_1 - 1 - C_3 = 5 + 10.5 - 1 - 0.36 = 14.14.$$

Assuming Qwen-2.5-VL-32B is used for both video_cap and video_qa, the empirical average runtimes are:

$$T_1 \approx 2.5\text{s}, \quad T_2 \approx 7.0\text{s}, \quad T_3 \approx 2.7\text{s}.$$

Therefore, the expected end-to-end time required to answer a single question on VideoMME-Long is:

$$T = 10.5 \times 2.5 + 14.14 \times 7.0 + 0.36 \times 2.7 \approx 135\text{s}.$$

This result reflects the full hierarchical search procedure, including both caption retrieval and occasional fine-grained Video_QA queries.

C. More Qualitative Examples

We provide additional qualitative results (Figure6, Figure 7) and ultra-long video examples (Figure 5) in this section. These examples illustrate LongVideo-R1's ability to perform hierarchical search, disambiguate similar scenes across hours-long content, and jointly use both high-level and fine-grained information.

The examples include cases from TV series such as *Downton Abbey*, where the model successfully navigates multi-hour narratives, repeatedly locating the correct characters, objects, or events despite substantial visual similarity across episodes.

D. Failure Examples

Although LongVideo-R1 performs well across various long-video benchmarks, failure cases still occur (Figure 8, Figure 9). When a visually similar but irrelevant object appears in the video, the model sometimes commits to the wrong branch and fails to return to the correct segment.

We also find that simple textual hints can often guide the model back to the correct segment and enable it to produce the correct answer.

After Miss Mary introduced Mr. Pamuk to her father, Earl Robert, *what questions did Robert ask Pamuk enthusiastically*?

A 15-episode TV series with a duration of over 10 hours

<think>…This is a *specific dialogue moment*, *likely in a scene involving Lady Mary Crawley and Kemal Pamuk*, which is mentioned in High-level Caption 3. To answer accurately, I need to look at the medium or low level captions around the…</think>
<tool>video_cap(3,6)</tool>

<think>…The medium-level caption describes scenes of a fox hunt and a church visit, but it does not mention the introduction of Kemal Pamuk to Robert Crawley. The introduction scene is likely in another sub-segment of this high-level segment…</think>
<tool>video_cap(3,6,1)</tool>

⋮

<think>…caption now *contains the specific lines of dialogue*: Robert Crawley says "Heavens, you have been in the wars," then asks Pamuk "How do you do, My Lord? Did you have a good day?" This *matches the question asked after the introduction*….</think>
<tool>video_cap(3,7,2)</tool>

<think>The introduction scene is fully described: Robert greets Kemal Pamuk and asks, "How do you do, My Lord? Did you have a good day?" This corresponds to the question, asking about his day's experience. To ensure accuracy, I can confirm if any other question was asked. The low-level captions show no other question, so the answer is clear.</think>
<answer>Robert asked Pamuk's experiences of the day.<answer>

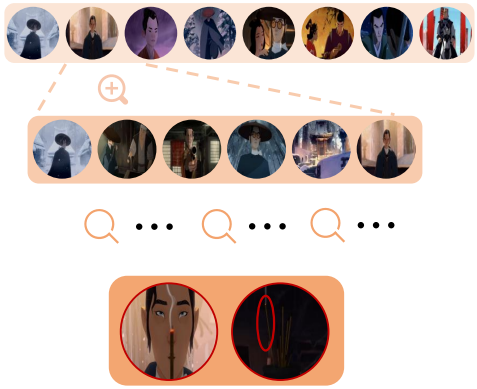
Question: How many *sticks* does the protagonist *put in the incense burner*?

<think>The question asks about the number of sticks the protagonist puts in the incense burner. This detail is likely in a scene involving lighting incense, which is mentioned in High-level Caption 2 between 635.0s and 654.2s … <think>
<tool> VideoCap(2,1) </tool>

<think>The medium-level caption (2,1) confirms that the incense lighting happens between 624.8s and 650.8s, but it still does not specify the number of incense sticks used. To find that detail, I need the low-level caption for the sub-segment… <think>
<tool> VideoCap(2,1,2) </tool>

⋮

<think> …does not mention the number of sticks either. It seems none of the retrieved captions explicitly state the number of incense sticks.The detail is likely in the initial lighting or placement scene, but captions have omitted it…<think>
<tool> VideoQA((2,1,3), How many incense sticks the man place in the burner?)</tool>
<answer>The protagonist puts *one incense stick* into the incense burner. <answer>



Question: During *Putin's first speech* on stage in the video, *how many women* appeared *when the camera panned offstage*?

<think>The question asks about Putin's first speech on stage and how many women appeared when the camera panned offstage. This is a very specific visual detail likely to occur in a segment showing Putin speaking. From the high-level captions… <think>
<tool> VideoCap(1,1) </tool>

<think>he medium-level caption for (1,1) *describes Putin speaking at the podium* but does *not mention the number of women appearing when the camera panned offstage*. To find this detail, I need to look at the low-level captions for the part of his …<think>
<tool> VideoCap(1,1,6) </tool>

⋮

<think>The low-level caption (1,1,5) *mentions a transition to the audience between 63.5s and 65.9s, which could be when the camera pans offstage*, but it does not specify the number of women. Since the exact detail is missing, and we already have …<think>
<tool> VideoQA((2,1,3), how many women appear when the camera pans offstage?)</tool>
<answer> *one woman appears* among the seated audience. <answer>

Figure 7. More qualitative examples.

Question: What kind of *transportation* do the *people carry on the boat*?

<think>The question asks about what kind of transportation people carry on the boat. The high-level captions mention boats in High-level Caption 6, but they don't specify what transportation is carried on the boat. To find this detail… <think>
<tool> VideoCap(6,6) </tool>

<think>The medium-level caption for ***(6,6) mentions a boat scene*** but does not specify what transportation is carried on the boat. To find that detail, I need to look at the low-level captions for the boat scene, which occurs between 3074.6s and possibly…<think>
<tool> VideoCap(6,6,1) </tool>

⋮

<think>The low-level caption (6,6,2) does not mention the boat or transportation carried on it. The boat scene seems to be in (6,6,1), but that caption didn't specify what's on the boat. Since captions are not giving the needed detail, I should use video_qa on (6,6,1) …<think>
<tool> VideoQA((6,6,1), What transportation do the people carry on the boat?)</tool>
< observation >no other form of transportation being carried on the boat. < /observation >

<think>The video_qa result indicates that no vehicle or transportation is carried on the boat, Given the information, the correct answer is not present.<think>
<answer> No transportation was carried on the boat </answer>

Your answer is incorrect, I will give you some hints: The segment that contains information *relevant to this question is 4*.

<think> The user indicates that the relevant segment is 4, but *I have been focusing on segment 6*. This suggests that the boat scene with carried transportation is in High-level segment 4, not 6. *I need to reorient*. High-level 4 mentions a. … <think>
<tool> VideoCap(4,1) </tool>

⋮

<think> The low-level caption *(4,4,6) clearly mentions that the group is unloading a motorcycle from the boat*. This directly answers the question about what kind of transportation the people carry on the boat. The correct answer is "Motocycle." <think>
<answer> People carried a motocycle on the boat </answer>

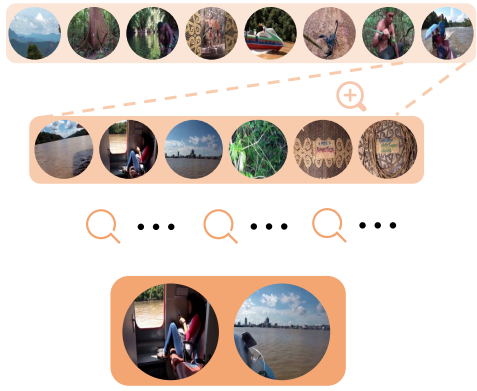
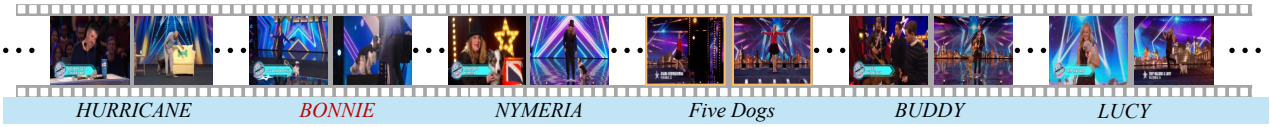


Figure 8. LongVideo-R1 may sometimes be disturbed by similar information, but people can guide the model back on track with just a few hints.



Question: What is the *name of the dog* that *plays the piano*?

A 102 minutes video with about 20 performers

<think>The question asks for the name of the dog that plays the piano. The high-level captions provided do not mention any dog playing piano; they describe various dog acts, tricks, and performances, but not the specific detail of a dog playing piano … <think>
<tool> VideoCap(6,3) </tool>

<think>…within High-level 6, possibly (6,2) or (6,4). ***Since High-level 6 mentions a woman performing with a dog accompanied by a pianist***, I should check the medium-level caption for (6,2) to see if it contains the piano act and possibly the dog's …<think>
<tool> VideoCap(6,2) </tool>

⋮

<think>The videoqa result for (6,2,5) *indicates the dog's name is "Mouse"* but clarifies that *the dog isn't shown playing the piano in these frames*. However, the question specifically asks for the dog that plays the piano, so "Mouse" might not be the correct answer…<think>
<tool> VideoQA((6,2,2), "What is the name of the dog that plays the piano?")</tool>
< observation >The name of the dog that plays the piano is not provided . < /observation >

<think>After multiple attempts, none of the captions or videoqa results have revealed the dog's name. I cannot deduce it from the given information …<think>
<answer> I cannot deduce the dog's name. **</answer>**

Your answer is incorrect, I will give you some hints: The segment that contains information *relevant to this question is 1*.

<think> The user indicates that the relevant segment is 1, but the high-level caption for segment 1 describes a dog performing tricks, not playing piano. I should get the medium-level caption for (1,3) or (1,4) to see if the dog's name is given… <think>
<tool> VideoCap(1,3) </tool>

⋮

<think> The *caption (1,3,4) explicitly mentions the dog's name as Bonnie*, *confirming that Bonnie is the dog in the act with the piano*. Since the question asks for the name of the dog that plays the piano, the correct answer is Bonnie.<think>
<answer> The name of the dog that plays the piano is *Bonnie*. **</answer>**

Figure 9. LongVideo-R1 may sometimes be disturbed by similar information, but people can guide the model back on track with just a few hints.