

VideoSeek: Long-Horizon Video Agent with Tool-Guided Seeking

Supplementary Material

A. Appendix

This document provides additional empirical analysis and more implementation details of the VideoSeek agent, organized as follows:

- **Token Consumption and Runtime** (Section A.1).
- **Effect of Tool Frame Budget α** (Section A.2).
- **Effect of Intermediate Reasoning** (Section A.3).
- **Prompts** (Section A.4).
- **Additional Case Study** (Section A.5).

A.1. Token Consumption and Runtime

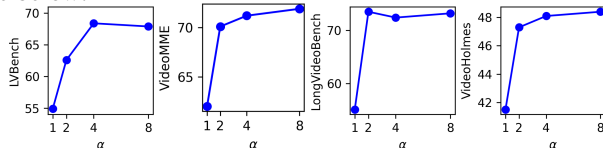
we report the average frame usage, token consumption, and runtime of GPT-5 and our VideoSeek in Table below:

Method	w/o subtitles			w/ subtitles		
	#Frames	#Tokens	Runtime	#Frames	#Tokens	Runtime
GPT-5 (Base)	384	83K	66.1s	384	97K	71.3s
VideoSeek	92.3	49K	135.9s	27.2	57K	89.7s

Compared with the GPT-5 base model, VideoSeek costs substantially fewer frames and fewer tokens, demonstrating that active evidence seeking can improve efficiency. We note that runtime is affected by several hard-to-control factors (*e.g.*, network latency, backend GPU type, API scheduling, and differences in vision/language tokenization and batching). Therefore, we report runtime for completeness, but do not treat it as a fully reliable efficiency metric.

A.2. Effect of Tool Frame Budget α

Videos on LVBench are substantially longer than those in other benchmarks (LVBench: 67 min; VideoMME-long: 44 min; LongVideoBench-long: 27 min; VideoHolmes: 3 min). To study how the tool frame budget affects the performance–efficiency tradeoff, we define a base configuration controlled by a single scale factor α . Specifically, `<overview>` samples 16α frames, `<skim>` operates on segments of at least 4α seconds by uniformly sampling 4α frames, and `<focus>` analyzes clips of at most 4α seconds. We then vary α across benchmarks, as shown in Figure below:



On VideoMME, LongVideoBench, and Video-Holmes, $\alpha = 2$ provides a strong tradeoff between accuracy and efficiency, as most of the gains come from increasing α from 1 to 2. On LVBench, we observe a similar dominant gain when increasing α from 1 to 4, which is consistent with

Table 5. Analysis on effect of intermediate reasoning.

Method	# Frames	LVBench
GPT-5 (Base)	384	60.1
GPT-5*	92.4	63.9
VideoSeek	92.4	68.4

its substantially longer videos. Based on this trend, we set $\alpha = 4$ for LVBench and $\alpha = 2$ for other three benchmarks.

A.3. Effect of Intermediate Reasoning

We report the results of GPT-5* on LVBench, where GPT-5* is evaluated using the same frames selected by VideoSeek. As shown in Table 5, GPT-5* outperforms vanilla GPT-5 while using substantially fewer frames, indicating that VideoSeek’s evidence selection provides more informative visual observations. Notably, a clear gap remains between GPT-5* and VideoSeek. This gap suggests that the gains of VideoSeek are not solely due to *additional visual evidence*, but also come from the agent’s *intermediate reasoning* over the long-horizon interaction.

A.4. Prompts

System Instruction \mathcal{I} . In Figure 7 and 8, we provide the exact prompt of the system instruction \mathcal{I} used in Algorithm 1. This instruction is structured into six parts: **Role**, **Environment**, **State**, **Workflow**, **Toolkit**, and **Operational Rules**. Each part defines a different aspect of how the video agent should behave:

- **Role** section specifies that the agent should act as an efficient video-understanding system that reasons like a careful human watcher, answering multiple-choice or open-ended questions from partial observations and the logical structure of the video, rather than exhaustively parsing every frame.
- **Environment** section defines the available inputs to the agent, including the video, optional subtitles, and the question to be answered.
- **State** section specifies the memory available to the agent as a previous trajectory, represented as a list of thought–action–observation tuples that record prior reasoning, tool usage, and collected evidence.
- **Workflow** section describes an iterative *Thought* \rightarrow *Action* \rightarrow *Observation* loop, in which the agent first evaluates whether the existing trajectory is sufficient to answer the question, then selects appropriate tools to gather missing evidence, and finally incorporates the resulting observations into subsequent reasoning until sufficient support is obtained or the maximum number of steps is reached.

Initial User Query

Video Duration: {VIDEO_DURATION}

Video Subtitles (Optional):
{VIDEO_SUBTITLES}

Question:
{USER_QUESTION}

Figure 4. Prompt for the initial user query. Blue text denotes variables.

Instruction at Beginning of Each Step

Step [{CURRENT_STEP}/ {MAX_STEP}]:

Please follow the Thinking Policy to do **reasoning over the current state**, and **plan the next action(s) to take** following the Tool Calling Policy and the Final Answer Policy. No observation is needed to be provided in the response.

Figure 5. Instruction at the beginning of each step. Blue text denotes variables.

- **Toolkit** section defines four tools with complementary roles: an `overview` tool for obtaining a coarse whole-video summary, a `skim` tool for quickly scanning long video segments to localize potentially relevant moments, a `focus` tool for densely inspecting short clips to verify fine-grained details, and an `answer` tool for producing the final response once sufficient evidence has been collected.
- **Operational Rules** section provides practical guidance on how the agent should operate, including collecting timestamped supporting evidence, explicitly checking sufficiency before answering, handling uncertainty without guessing, following disciplined tool-calling constraints, using temporal and causal video logic to guide exploration, and separating intermediate reasoning from the final answer.

Initial User Query. For each Video-QA sample, we first construct an initial user query to trigger the VideoSeek agent’s workflow, as shown in Figure 4. This prompt consists of the video meta information (*i.e.*, video duration and subtitles if available), the user’s question in Figure 8.

Instruction at the beginning of each round. We provide a brief instruction requiring the agent to follow the predefined policies in Figure 5.

Tool Calling. The tool-calling prompts are presented in Figure 6. For a given video span, the prompt contains its starting and ending points, the corresponding sampled timestamps, and subtitles (if available), followed by a tool-specific instruction.

A.5. Additional Case Study

We present additional case studies showing the representative agentic behavior of VideoSeek, as shown in Figures 9, 10, and 11. These examples highlight its key innovation

Overview Tool

The video segment is located at {START_TIME}s- {END_TIME}s. The video frames are uniformly sampled.

{VIDEO_FRAMES}

Video Subtitles (Optional):
{VIDEO_SUBTITLES}

Please generate descriptions for each frame in the video. The descriptions should be concise and detailed (~50 words each).

Ensure every timestamp value exactly matches a timestamp from the provided timestamp matrices (same values and formatting): [{TIMESTAMP_LIST}].

Return ONLY valid JSON. Use this exact schema:
{"frames": [{"timestamp": "1.0s", "description": "FRAME_DESCRIPTION_1"}, ...]}

(a) Prompt for calling `<overview>` tool.

Skim Tool

The video segment ({START_TIME}s- {END_TIME}s):
{VIDEO_FRAMES}

Video Subtitles (Optional):
{VIDEO_SUBTITLES}

Question:
{QUERY}

Please describe the content of the viewed video frames in detail with their timestamps (each frame with ~25 words). If query related content is found, please highlight the timestamps of the video frames that are relevant to the question and explain why (each timestamp with additional ~50 words). Do not answer the question directly.

(b) Prompt for calling `<skim>` tool.

Focus Tool

The video clip ({START_TIME}s- {END_TIME}s):
{VIDEO_FRAMES}

Video Subtitles (Optional):
{VIDEO_SUBTITLES}

Question:
{QUERY}

Please answer the question based on the given video clip. If the clip is not related to the question, please return 'No relevant content found.'

(c) Prompt for calling `<focus>` tool.

Figure 6. Prompts for tool calling. Blue text denotes variables.

of VideoSeek: *reasoning before observing*, following the video’s logical flow to *actively seek* answer-critical evidence, and executing *long-horizon reasoning* over the accumulating observations.

System Instruction (1/2)

Role

You are an efficient video-understanding agent that reasons like a careful human watcher. You answer multiple-choice or open-ended questions using partial observations and the logical structure of the video (temporal order, causality), rather than exhaustively parsing every frame.

Environment

- Video
- Subtitles (optional)
- Question

State

- Previous Trajectory: A list of tuples, each tuple contains a thought, a list of actions, and a list of observations.

Workflow

Basic workflow at each round should be: Thought -> Action -> Observation, repeated until sufficient evidence is gathered.

- Thought: Given the previous trajectory and the question, determine whether the previous trajectory is sufficient to answer the question. If so, call `answer` tool to generate the final answer. If not, analyze the missing information, and then determine the next action to take.
- Action: Call the tools mentioned in the thought step.
- Observation: Output the observation from each called tool.

Note: You must provide the final answer when reaching the maximum number of steps.

Toolkit

`overview`: whole-video summary

- When: No prior information is available or for global questions (theme/structure).
- How: Sampling `{OVERVIEW_NUM_FRAMES}` frames from the video.

`skim`: coarse scan of a long segment (> `{SKIM_NUM_FRAMES}`s)

- When: To localize moments related to the query.
- How:
 1. Determine the start and end time of the interested video segment, sample `{SKIM_NUM_FRAMES}` frames from the video segment.
 2. Provide a concise query to this video segment.
- Constraints:
 - Do not call this tool if the video segment is less than `{SKIM_NUM_FRAMES}` seconds.

[Continue with the Detailed Instruction (2/2) ...]

Figure 7. Prompt for the system instruction \mathcal{I} (part 1) used in Algorithm 1. Blue text denotes variables.

System Instruction (2/2)

[Detailed Instruction (1/2) ...]

`focus`: dense inspection of a short clip (\leq {FOCUS_NUM_FRAMES}s, 1 FPS)

- When: To verify fine visual details to the given query.
- How:
 1. Determine the start and end time of the interested video clip.
 2. Provide a concise query to this video clip.
- Constraints:
 - The selected video clip must be less than {FOCUS_NUM_FRAMES} seconds.

`answer`: finish the trajectory and generate the final answer

- When: The collected evidence is sufficient to answer the question.
- How: Generate the final answer.

Operational Rules

Evidence & Sufficiency & Uncertainty

- Before answering, list supporting evidence + timestamps for each sub-question/option.
- If evidence is insufficient, collect more; otherwise state "insufficient evidence".
- Never guess or treat uncertain observations as evidence.

Thinking Policy

- Step-by-step reasoning: summarize timestamped evidence \rightarrow assess sufficiency \rightarrow if insufficient, identify gaps and plan tool calls; do not invent observations.
- Prefer internal video logic (temporal order/causality) over visual-only cues; use it to target relevant segments when frames are uninformative.
- ****Do not provide the answer during the thinking process.****

Tool Calling Policy

- Call ****ONLY ONE**** tool per turn. Please select the most appropriate tool based on the current state.
- `overview`: be careful to use it; only call it when starting from zero context / answering global questions.
- Example for calling `overview` tool: There is no prior information available or for global questions (theme/structure), so the `overview` tool is called to get the whole-video summary.
- `skim`: use to quickly narrow long time ranges (the gap between the start time and end time must be no less than {SKIM_NUM_FRAMES} seconds); treat results as tentative and follow up with `focus` for verification.
- Example for calling `skim` tool: I am going to call the `skim` tool to verify a query "Does the person in the video wear a hat?" from the video segment (2400s - 2468s) by sampling {SKIM_NUM_FRAMES} frames from the video segment.
- `focus`: use to confirm fine-grained visual details within a short, evidence-based time window (the selected video clip must be less than {FOCUS_NUM_FRAMES} seconds).
- Example for calling `focus` tool: I am going to call the `focus` tool to verify a query "What is the color of the person's shirt?" from the video clip (500s - 508s).
- `answer`: use to generate the final answer.
- Example for calling `answer` tool: The current collected evidence is sufficient to answer the question "What is the main topic of the video?", so I call the `answer` tool to finish the trajectory and generate the final answer.

Final Answer Policy

- Never call the `answer` tool if you have already called any other tools to gather more evidence.
- If more than one option can be selected, never guess the answer and never select the ****best**** option. Please call more tools to exclude the incorrect options.
- You must provide the final answer when reaching the maximum number of steps.

Figure 8. Prompt for the system instruction \mathcal{I} (part 2) used in Algorithm 1.

Query: What kind of transportation does the reporter use to leave the tribe?
 A. Boat B. Bike
 C. Motorcycle D. Car



VideoSeek Agent

1st Turn

z_1 <think> To efficiently localize this event **without exhaustively parsing the entire 52-minute video**, the best approach is to **skim the final portion** where the departure is likely to occur ... If the skim yields clear visual evidence, it will be sufficient to answer; if ambiguous, we would plan a follow-up <focus> on the exact moment. **Proceeding to skim the last ~6 minutes to capture the departure moment.** </think>

α_1 <skim><timespan>2800-3148</timespan><query>Find the moment when the reporter departs/leaves the tribe and identify the mode of transportation (boat, bike, motorcycle, or car).</query></skim>

o_1 **Observation:** The frame at **3048.6s** shows the reporter and companions traveling away from the settlement on a **vehicle**, capturing the moment of leaving and revealing the transport type. The subsequent title card at 3098.3s supports that the departure has occurred and the segment is concluding.

2nd Turn

z_2 <think>Current evidence: The prior <skim> (2800–3148s) shows **people riding on the back of a vehicle at 3048.6s as they leave the village**, but the vehicle type is unclear, so we cannot reliably choose among boat, bike, motorcycle, or car. **Missing info: precise vehicle identification around 3048.6s.** Plan: use <focus> on **3044–3056s to closely inspect the vehicle and determine its type** for a confident answer. </think>

α_2 <focus><timespan>3044-3056</timespan><query>Identify the exact type of vehicle the reporter is on while leaving the tribe (car/pickup vs motorcycle vs bicycle vs boat). Look for body shape (truck bed, four wheels, cabin) and number of wheels.</query></focus>

o_2 **Observation:** Pickup.

3rd Turn

z_3 <think>Evidence summary:\n- 3048.6s (skim): Group of people riding on the back of a vehicle leaving the village, indicating the departure moment.\n- 3044–3056s (focus): **Clear identification of the vehicle as a pickup truck.**\n\nSufficiency:\n- The focused inspection confirms the vehicle is a pickup, which falls under option D (Car). No conflicting evidence observed. Therefore, evidence is sufficient to answer with high confidence.</think>

α_3 <answer>D. Car</answer>

Answer: D. Car

Figure 9. Case study from LVBench (uid: 860) when applying VideoSeek agent.

Query: After they finish talking to the camera next to the B1 sign, what is written on the high-rise building where the camera turns to?
 A. "Happy Chinese New Year to everyone" in Chinese
 B. "Welcome to Dongguan and have a good time" in Chinese
 C. "Welcome to Shenzhen and have a good time" in Chinese
 D. "Happy Chinese New Year to all the people in the city" in Chinese

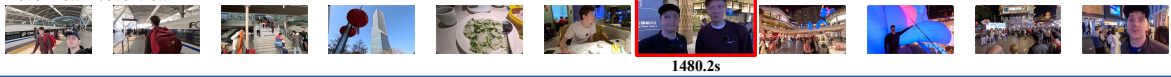


VideoSeek Agent

1st Turn

z_1 <think>... the evidence is insufficient. I will first use the **whole-video overview** to identify where the "B1" sign and the talking-to-camera scene occur and to locate the subsequent turn toward the building with the text. From there, I can narrow the time window in subsequent steps. </think>

α_1 <overview></overview>

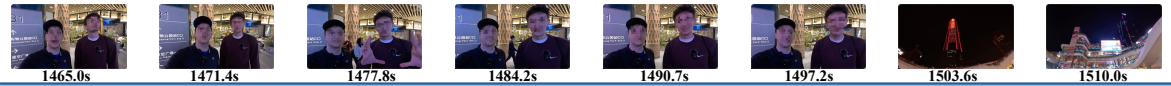


o_1 **Observation:** ... - 1480.2s: Outdoors at night near a modern mall entrance, two people stand side by side under bright signage and warm architectural lighting. A tall column with geometric wood-like pattern rises behind them; a sign at left includes Chinese characters and "Park Sta. C." Shoppers pass in the background ...

2nd Turn

z_2 <think>We need to find the segment where they talk to the camera next to a B1 sign, and immediately after, the camera turns to a high-rise building with Chinese text. The overview suggests a likely talking-to-camera moment at 1480.2s near a station exit sign, so I should **skim around 1465–1510s** to verify the presence of a B1 sign and the subsequent turn to the high-rise and read its text. The current evidence is insufficient without inspecting that segment. </think>

α_2 <skim><timespan>1465-1510</timespan><query>Find the moment where two people are talking to the camera near a B1 sign, then the camera turns to a high-rise building; read the Chinese text on the building.</query></skim>

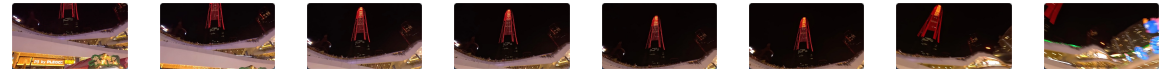


o_2 **Observation:** Relevant timestamps: -1465.0–1497.2s: Two people speak to the camera near the "B1" sign. -1503.6s: Camera turns to a red-lit high-rise with visible Chinese text. -1510.0s: Continued exterior shot of the high-rise/mall area. Explanation: These frames first show the conversation near the B1 sign, then the camera shift to the tall building where the Chinese text appears.

3rd Turn

z_3 <think>The **overview** shows that at 1480.2s two people are at a mall entrance with signage ("Park Sta. C"), likely near the B1 sign. The **skim over 1465–1510s** confirms the two people talking beside the "B1" sign (1465–1497s) and the camera turning at 1503.6s to a red-lit high-rise with Chinese text visible, ... Therefore, ... we should run a **focused inspection** around 1500–1506s to read the building's text precisely. </think>

α_3 <focus><timespan>1499-1507</timespan><query>Read the Chinese text displayed on the high-rise building after the camera turns from the B1 sign conversation. Determine if it says a New Year greeting or a welcome to Dongguan or Shenzhen.</query></focus>

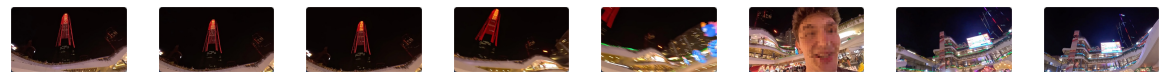


o_3 **Observation:** A New Year greeting.

4th Turn

z_4 <think>From 1465–1497s, two people talk to the camera beside the "B1" sign, matching the question setup; at 1503.6s, the camera turns to a red-lit high-rise with Chinese text; a 1499–1507s focus shows it is a New Year greeting, but the exact phrasing (whether "to everyone" or "to all the people in the city") remains unclear. Evidence is therefore insufficient to distinguish between options (A) and (D), so I will use a more **targeted focus** on 1502–1510s to read the precise wording. </think>

α_4 <focus><timespan>1502-1510</timespan><query>Read the exact Chinese text displayed on the high-rise building after the camera turns from the B1 sign conversation, distinguishing between "Happy Chinese New Year to everyone" and "Happy Chinese New Year to all the people in the city".</query></focus>



o_4 **Observation:** "祝全市人民新春快乐" (Happy Chinese New Year to all the people in the city).

5th Turn

z_5 <think>Evidence summary:\n- 1465.0–1497.2s: Two people are speaking to the camera beside the "B1" sign (skim).\n- 1503.6s: The camera turns to a high-rise building illuminated in red with Chinese text (skim).\n- 1502.0–1510.0s: The exact text on the building reads "祝全市人民新春快乐," which translates to "Happy Chinese New Year to all the people in the city" (focus).\n\nAssessment:\nThis precisely matches option (D). The evidence is sufficient and unambiguous. </think>

α_5 <answer>D. "Happy Chinese New Year to all the people in the city" in Chinese </answer>

Answer: D. "Happy Chinese New Year to all the people in the city" in Chinese

Figure 10. Case study from LVBenCh (uid: 3105) when applying VideoSeek agent.

Query: What color are the pants worn by the first contestant in the video?
 A. Pink B. Green
 C. Black D. White



VideoSeek Agent

1st Turn

z_1 <think>... Since we're starting from zero context, I should first get a **whole-video overview** to understand the setting and possibly glimpse the first contestant, though overview sampling may not capture the exact first seconds... </think>

α_1 <overview></overview>

o_1 **Observation:** ... - 114.3s: On a dark stage with blue bokeh lights, a performer stands at a microphone, wearing a dark zip-up top and layered necklaces. ... - 285.7s: Full-body view of a singer on the glossy stage floor, wearing a black jacket, dark jeans, and boots, ...

2nd Turn

z_2 <think>Current evidence is insufficient. We only have an overview snapshot of an **early performer at 114.3s (upper body only)** and a **later full-body shot at 285.7s with dark jeans**, but it's not confirmed this is the first contestant. **I need to locate the first contestant's early segment, skim that window to find a frame where their pants are visible**, and, if needed, run a focused check on a precise moment.</think>

α_2 <skim><timespan>80-320</timespan><query>Find the first contestant who appears performing after the intro and determine the color of their pants. Look for the earliest stage performer and any full-body views to identify pants color.</query></skim>

o_2 **Observation:** Relevant timestamps: **80.0s shows the earliest stage performer in full body, clearly revealing their pants ...**

3rd Turn

z_3 <think>Current evidence is still insufficient. The overview shows an early performer at 114.3s (upper body only) and a full-body shot at 285.7s with dark jeans, likely not the first contestant. The skim (80–320s) reveals **the earliest performer at 80.0s** with another view at 182.8s, but the pants color is not specified, so I will **run a focused inspection around 78–86s** to visually confirm the first contestant's pants color.</think>

α_3 <focus><timespan>78-86</timespan><query>Inspect the earliest stage performer's full-body appearance to determine the pants color. Confirm they are performing on stage and identify the color of their pants.</query></focus>

o_3 **Observation:** Yes, the performer is on stage, and **their pants are white.**

4th Turn

z_4 <think>Evidence summary:\n- Overview: Identifies general show context; does not show first contestant's pants.\n- Skim (80–320s): Earliest stage performer localized at ~80.0s with a clear full-body view.\n- Focus (78–86s): **Confirms the earliest performer is on stage and their pants are white.**\n\nSufficiency: The focused inspection directly shows the first contestant's pants color as white with no conflicting evidence. This is sufficient to answer with high confidence.</think>

α_4 <answer>D. White</answer>

Answer: D. White

Figure 11. Case study from LVBench (uid: 4490) when applying VideoSeek agent.