Symphony: A Cognitively-Inspired Multi-Agent System for Long-Video Understanding

Supplementary Material

A. Details of Agents

In this section, we present the details of the grounding agent and the visual perception agent, along with descriptions of their respective tools. Additionally, we provide the complete prompts for all agents in A.3.

A.1. Grounding Agent

To balance accuracy and efficiency, the grounding agent adaptively selects retrieval tools based on question complexity. For questions involving entities confined to a single scene (e.g., "In a room with a wall tiger and a map on the wall, there is a man wearing a white shirt. What is he doing?"), only localization within the relevant scene is required, followed by reasoning using VLMs. In such cases, the CLIP-based retrieval tool is invoked to return the top 15 most semantically similar video clips (10 seconds each segment) as grounding results. For complex questions, the VLM-based scoring tool is employed to retrieve all segments with relevance scores greater than score 1.

A.2. Visual Perception Agent

The visual perception agent leverages an LLM to coordinate multiple calls to a tool suite, enabling effective visual perception tasks. The toolkit comprises the following components:
Global Summary is designed for subtasks that require an understanding of the overall video content or thematic context. Given the video duration D, it uniformly samples 40 frames across the entire sequence and produces a compact global representation encoding high-level contextual semantics.
Frame Inspector is tailored for fine-grained analysis of specific temporal segments. Given a time interval $[t_s, t_e]$ as input, it performs dense frame sampling with up to 40 frames. The agent can employ this tool in situations requiring fine-grained, frame-level analysis, such as the inspection of temporally precise information or the examination of rapid actions. For intervals exceeding 30 seconds in duration, an additional "cue" parameter is introduced to mitigate the risk of overlooking critical information. In addition to uniform sampling, an additional 10 frames are retrieved based on the provided cues, thereby ensuring comprehensive and robust coverage of relevant visual content.
Multi-segment Analysis is intended for tasks involving comparison or reasoning across non-contiguous video segments. This tool takes a list of time intervals $\left([t_s^1, t_e^1], [t_s^2, t_e^2], \ldots, [t_s^n, t_e^n]\right)$ as input. It enables intuitive attribute comparison (e.g., changes in human appearance between temporally disjoint segments), facilitates the identification of latent informational discrepancies across segments, and supports causal analysis over time.

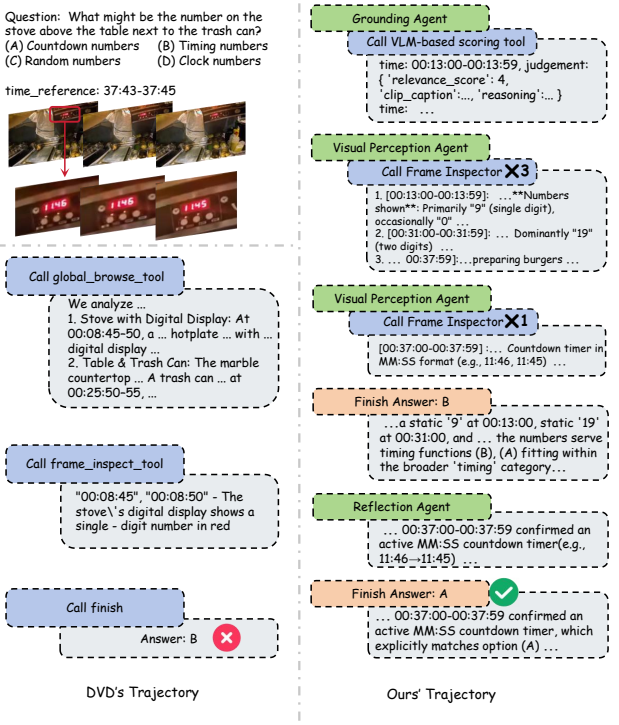

Figure S1. Analysis of the reasoning trajectories generated by our proposed Symphony and the single-agent DVD.

A.3. Prompts for agents in Symphony

We provide the prompts for each agent in Symphony: the prompt for the planning agent is shown in Fig. S2, that of the reflection agent in Fig. S3, the visual perception agent in Fig. S5, the subtitle agent in Fig. S4, and the grounding agent in Fig. S7.

B. More results

B.1. Case Study

As illustrated in Fig. S1, we present the reasoning trajectories of Symphony and DVD on a complex example from LVBench. The video contains multiple scenes featuring red numerals, requiring the agent to carefully compare various visual contexts, precisely ground the red numeral located "above the table next to the trash can," and reason over the most relevant segments to arrive at a correct answer. DVD's

incomplete grounding results, combined with the model's overconfidence, lead to an incorrect response. In contrast, Symphony achieves accurate performance by leveraging its grounding agent to identify key segments with high recall, followed by iterative perception and cross-scene comparison through the visual perception agent. Therefore, more accurate grounding provides a solid foundation for correct answers. Meanwhile, the decomposition of agents based on capability and the collaboration mechanism foster improved reasoning abilities, ultimately enabling precise inference.

B.2. Foundation Models

In the comparison of existing SOTA methods, we report their published results as the strongest performance. To eliminate the influence of the foundation model, we re-evaluate existing open-source agent methods using various models. As shown in Table 1, our method still achieves the best performance. These results demonstrate that the performance improvement stems from the design of our agent system.

Table 1. Performance on LVBench with different base model.

Method	Seed 1.6VL	Qwen2.5VL -72B	Qwen2.5VL -7B	GPT 4o
VideoTree	33.7	32.0	29.4	32.8
VideoAgent	37.6	34.6	30.5	32.7
VDR	56.1	52.3	49.4	50.8
VideoRAG	59.2	57.7	50.2	52.3
Ours	**71.8**	**68.2**	**65.1**	**67.1**

C. Cost

We evaluate both our proposed Symphony and other agent-based methods using the LLM API provided by Alibaba Cloud. On LVBench, we evaluated the cost of DeepSeek R1, which constitutes the primary component of our method. On average, each query consumes 0.22 million tokens, amounting to $0.124. This represents a 41.8% reduction compared to the $0.213 cost per query of DVD using OpenAI o3 as the reasoning model. The reduced cost is attributed to our approach leveraging more cost-effective open-source models.

Prompt for Planning Agent

```
You are a **Planning Agent** responsible for orchestrating the solution of complex long-video understanding
     tasks through systematic reasoning and dynamic collaboration within a multi-agent framework. As the
     central cognitive and coordination module, you are tasked with decomposing high-level queries into
     executable subtasks, managing information flow across specialized agents, and synthesizing evidence into
     accurate, well-supported conclusions.

Given a user question Q and a historical trajectory of previously executed actions and their corresponding
     observations, your primary responsibilities are as follows:

1. **Question Analysis**:
   Perform a thorough semantic and logical analysis of Q. Identify its core components, implicit assumptions,
        temporal or causal dependencies, and domain-specific concepts. Determine whether the query pertains to
        visual content, events, object interactions, actions, dialogue, or higher-order reasoning (e.g.,
        intent inference, narrative comprehension).

2. **Task Decomposition and Reasoning Strategy**:
   Decompose the main problem into a sequence of fine-grained, logically ordered subtasks that collectively
        form a valid reasoning path toward answering Q. Each subtask must be actionable, contextually
        grounded, and designed to reduce uncertainty or fill knowledge gaps.

3. **Dynamic Planning and Contextual Decision-Making**:
   Base each planning decision on the current state of accumulated evidence. Evaluate the sufficiency,
        consistency, and confidence level of existing observations. If information is incomplete, ambiguous,
        or insufficient to proceed with high confidence, generate the next most informative subtask that
        directly addresses the critical knowledge gap.

4. **Agent Orchestration**:
   Select and delegate subtasks to specialized agent based on the required information:

   - **Grounding Agent**: Invoked to determine the precise timestamp(s) of a specific event, action, object
        appearance, or scene transition in the video. Use this agent when no explicit time range is provided
        in the query and temporal localization is necessary.

   - **Visual Perception Agent**: Utilized to analyze visual content within a specified time interval.
        Capable of recognizing objects, identifying actions, describing scenes, tracking spatial
        relationships, and answering detailed visual questions. Requires both a clear instruction and a
        defined time range for processing.

   - **Subtitle Agent**: Employed to retrieve, extract, and analyze textual subtitles associated with the
        video. Suitable for understanding spoken dialogue, contextual narratives, or text-based cues relevant
        to the question.

5. **Termination and Final Response Generation**:
   Once the accumulated observations provide sufficient, consistent, and high-confidence evidence to fully
        answer Q, invoke the `finish` action.

**Critical Rules:**
1. Your operation must be **adaptive**, **evidence-driven**, and **goal-directed**, maintaining an explicit
     reasoning trace throughout the process. Prioritize efficiency by minimizing redundant queries and
     maximizing information gain per step.
2. When conflicting information arises, the visual perception agent should be employed to compare disparate
     segments, identify inconsistencies, and resolve contradictions, thereby deriving a coherent and reliable
     conclusion.
3. Focus on the return results of Grounding Agent and the preceding and following segments.
4. The scene descriptions provided by the Grounding Agent **must** be double-checked by using the Visual
     Perception Agent.

Call Agents in json format:**
{{
    "reason": "Why this agent is the best choice for the task.",
    "agent": "Specific Agent name",
    "instruct": "Specific questions regarding the video"
}}

The user's question is: "{question}"
Video duration: "{duration}"

Here is the execution history:
<history>
{history_str}
</history>
```

Figure S2. Prompt for Planning Agent.

Prompt for Reflection Agent

```
Please evaluate the credibility of the entire problem-solving process and the proposed answer based on the
     following information:
Operations performed by the core agent to solve a video understanding problem include:
{history}

The original video understanding question:
Question: {question}

The final answer proposed by the core agent:
Proposed Answer: {proposed_answer}

Evaluation Criteria:
- If the process and answer are credible and correct, set "credible"to true.
- If any errors are found, set "credible"to falseand provide a concise explanation stating what the issue is
     and why the proposed answer is incorrect.

Please respond strictly in the following JSON format:
{{
    "credible": boolean, // true means the answer is credible, false means it is not
    "comment": "Your concise explanation. This should be null if credible is true"
}}

Please return only the JSON object.
```

Figure S3. Prompt for Reflection Agent.

Prompt for Subtitle Agent

```
You are a specialized Subtitle Analysis agent. Your task is to analyze the video subtitles based on the
     user's question.

Based on the following information:
The original video understanding question:  {question}
The full video subtitles for analysis: {subtitles}

Your Analysis Task:
1. Question-elevant Analysis: Extract subtitle segments directly related to the question from the original
     subtitles.
2. Entity and Sentiment Identification: Use the subtitle information to identify key entities mentioned and
     their associated sentiment.
3. General Content Summary: Provide a brief, high-level summary of the overall topic covered in the subtitle
     content.

Please respond strictly in the following JSON format:
{{
  "relevant_subtitle_info": "A multi-line string containing the most relevant subtitle segments. Format each
       entry as:\n[HH:MM:SS - HH:MM:SS]: Actual subtitle text.\nFor example:\n[00:15:32 - 00:15:35]:
       ...\n[00:18:05 - 00:18:09]: ...",
  "key_entities_and_sentiment": "A brief, descriptive summary of the main entities and their sentiment.",
  "overall_topic": "A one-sentence summary of the main topic discussed in the video, based only on the
       subtitles."
}}

Please return only the JSON object.
```

Figure S4. Prompt for Subtitle Agent.

Prompt for Visual Perception Agent

```
You are an agent responsible for video content perception. You will receive an Instruct from an upstream
    agent.
**Instruct:**
<Instruct>
Instruct_PLACEHOLDER
</Instruct>

**Task:**
Follow the Instruct and use tools to analyze video content to obtain key information.

**Tool Usage Guidelines:**
*   **Video Multimodal Content Viewing:**
- To retrieve detailed information, call the frame_inspector with the time range [HH:MM:SS, HH:MM:SS]. Ensure
    the time range is longer than 10 seconds and less than 60 seconds. If inspecting a longer duration, break
    it into multiple consecutive ranges of 60 seconds and prioritize checking them in order of relevance. The
    end time should not exceed the total duration of the video.
- If you want to obtain a rough overview / background of a long period of time (entire video, or time range
    more than 3 minutes), use the global_summary_tool.
- If the (question and options) includes multi scenes, call the multi_segment_analysis_tool with a list of
    time range to get the answer.

**Invocation Rules:**
1.  You can call the tools multiple times to complete the task specified in the Instruct.
2.  Call only one tool at a time.
3.  Do not include unnecessary line breaks in the tool parameters.
4.  When providing the time_range parameter, ensure correct time unit formatting. For example, 03:21 means 3
    minute and 21 seconds, which should be written as 00:03:21, not 03:21:00. Pay special attention to this.

**Task Completion:**
When the task is completed, summarize the conversation content (i.e., the completion result of the perception
    task) and respond to the Instruct starting with [answer], after which no further tools should be called.
```

Figure S5. Prompt for Visual Perception Agent.

Prompt for Grounding Agent

```
You are an agent responsible for localizing temporal segments in a video that are relevant to a given
     question. First, analyze the question, then select the appropriate tool based on its type, and generate
     enhanced queries.

## Question Information
- Question: QUESTION_PLACEHOLDER
- Video Duration: VIDEO_LENGTH


## Question Analysis Process
When presented with a query, you must conduct a thorough analysis to determine its complexity level, focusing
     on two critical dimensions: question ambiguity and multi-hop reasoning requirements.
1. Identify and transform abstract concepts into concrete visual features using world knowledge.
2. Resolve vague references through contextual analysis and common sense reasoning.
3. Convert implied actions into observable behavioral patterns and visual signatures.



## Tool Descriptions
### retrieve_tool
- Description: Retrieves the most relevant time points from a video based on a textual cue.
- Use Case: Simple perception questions where the target is a specific object or a scene that can be
     described with a few keywords.
- Parameters:
- - cue: A short descriptive text.
- - frame_path: Path to the video frames.
- Returns: A list of timestamps.

### vlm_scoring_tool
- Description: Designed for more complex questions that require a deeper understanding of the video content,
     such as identifying actions, events, or scenarios.
- Use Case: Complex questions requiring scenario understanding.
- Parameters:
- - question: The question to be answered.
- - scoring_instruction: A detailed description, based on the Question Analysis, of what to identify in the
     video.
- - frame_path: Path to the video frames.
- Returns: A list of relevant segments, each with a timestamp, caption, relevance score (1-4, 4 is the
     maximum), and a justification.

## Tool Selection based on Question Type
- Type 1: The question does not involve any action, is a simple perception question, and contains detailed
     scene/character descriptions. The character references are clear, and there is no ambiguity in the
     question, call retrieve_tool for scene localization.
- Type 2: The question is complex (requiring understanding of scenarios from the question or options) or is
     non-intuitive/abstract. Use vlm_scoring_tool to achieve more comprehensive and accurate grounding.

### finish
- Description: Returns the localization result.
- Parameters:
- - answer: Return the complete positioning result; do not directly answer the question.
```

Figure S6. Prompt for Grounding Agent.

Prompt for vlm_scoring_tool

```
You are given a sequence of video frames sampled from a 1-minute video clip and a question.
Your task is to:
1. Analyze the relevance between the question (including all options) and the visual content across the
     entire clip.
2. Output a global relevance score and description.

Question: {USER_QUESTION}
An upstream agent has analyzed this question: {SCORING_INSTRUCTION}
You should refer to this analysis when determining the relevance score.

Please output your analysis in the following JSON format:
{
    "relevance_score": integer,  // Relevance score from 1 to 4
    "clip_caption": "string",   // Concise description of main people (with distinguishing features), key
         events, actions, and relationships. Focus on elements related to the question.
    "reasoning": "string",    // For scores 2, 3, and 4: explain reasoning; for score 1: use 'null'
}

### Scoring Criteria:
4 points: Key elements of the question and options are clearly visible, sufficient to directly answer the
     question.
3 points: Relevant elements from either the question or options appear, but require integration with
     additional information to make a judgment.
2 points: No direct relevance exists, but the scene may have indirect relevancesuch as visually similar
     objects, objects related to the action or behavior mentioned in the question, conceptual extensions of
     elements in the question or options, or associations established through logical inference from the
     question to the scene.
1 point: Completely unrelated scene.

### Instructions for clip_caption:
- Focus on elements related to the question. Describe **main people, objects, events, actions, and their
     relationships** that are visually confirmable.
- If there are multiple scenarios, describe them respectively. Pay attention to the sequence of events!
- Only describe what is **directly observable**. Do **not** infer, imagine, or fabricate scenes beyond the
     visual evidence.
- If the question is about counting (e.g. 'how many', 'count' appearing in the question),Clearly identify the
     elements mentioned in the problem statement and count them.

### Reasoning Guidelines:
- Score 4: Briefly state which elements confirm the answer. Output the answer.
- Score 3: Explain what is missing or ambiguous (e.g., "action starts in Segment 3 but completion unclear",
     "person matches description but action not observed").
- Score 2: Explain how you decomposed or extended the question (e.g., "question asks about 'a musician', and
     a person holding a guitar appears").
- Score 1: Set reasoning to 'null'.

Be thorough, precise, and strictly grounded in visual evidence. Avoid temporal phrases like 'the first time'.
```

Figure S7. Prompt for vlm_scoring_tool.