



CLiViS:

Unleashing Cognitive Map through Linguistic-Visual Synergy for Embodied Visual Reasoning

Supplementary Material

In this supplementary material, we provide additional details and analyses to complement the main paper. Sec. 1 elaborates on the implementation details, covering experimental parameters, relation graph construction, and compute resources. Sec. 2 presents extended quantitative results with a fine-grained evaluation on the EgoTempo benchmark. Sec. 3 provides a comparative analysis with frontier multimodal models regarding accuracy. This is followed by an analysis of reasoning dynamics in Sec. 4. Sec. 5 visualizes the model’s decision-making process through comprehensive qualitative examples. All the prompts used in our framework are placed in the `prompts.txt` folder in the supplementary material.

1. More Implementation Details

Additional Experimental Parameters. During evaluation, we set the temperature to 0.5 and $\text{top-p} = 0.9$ for LLM components, while VLMs use a temperature of 0.3 and $\text{top-p} = 0.9$. Considering hardware limitations and efficiency, we preprocess videos by sampling frames at 0.5 FPS on OpenEQA and EgoSchema, while limiting frame count to 32 for EgoTempo. We also pre-scale all videos from the EgoTempo and OpenEQA datasets to a resolution of 480p prior to experimentation.

Implementation Details of Relation Graph. The relational graph is implemented using Neo4j, a graph database system. Neo4j is an open source NoSQL graph database designed to efficiently store and process highly interconnected data, supporting complex graph traversals and relationship queries. It is widely used in applications such as knowledge graphs. As input to the LLM during inference, the subgraph extracted from the full relational graph includes the following components: (1) all designated key entities (nodes); (2) all paths between key nodes, along with intermediate nodes and relationships on those paths; (3) nodes and relationships directly connected to each key node; (4) all associated nodes and relationships linked to activity-type nodes. In addition, we limit the maximum path length to 10 steps when retrieving paths between key nodes to ensure the query efficiency.

Experiments Compute Resources. All VLMs used in our experiments, including Qwen2-VL (7B), Qwen2.5-VL (7B), InternVL2.5 (8B), InternVL3 (8B), and VideoL-LaMA3 (7B), are deployed and executed locally. The hard-

ware setup for experiments consists of 2 A100 80G GPUs and 4 A6000 48G GPUs. LLMs such as Qwen2.5-Max and DeepSeek-V3 are accessed via APIs provided by a cloud computing platform.

2. Supplementary Results

To further understand how different methods handle diverse instructions, we conduct a fine-grained evaluation on the EgoTempo dataset. The questions in EgoTempo are categorized into 10 distinct types, including Action-Specific Objects (AO), Locating Objects (LO), Spatial Relations (SR), Object and Action Sequences (OS, AS), Object-Specific Actions (OA), Temporal Events (TE), Future Prediction (FU), Action and Object Counting (AC, OC).

As shown in Table 1, CLiViS consistently outperforms all compared methods across most categories. Notably, compared to others, CLiViS delivers substantial improvements in reasoning-intensive categories such as Spatial Relations (SR), Action Sequences (AS), and Temporal Event (TE), demonstrating its superior capability for multi-step, long-horizon visual reasoning.

3. Comparison with Frontier Multimodal Models

We have conducted additional experiments using GPT-4.1 and Gemini-2.5-flash on the EgoTempo benchmark to evaluate performance. The results are presented in Table 2.

As shown in the table, CLiViS delivers superior accuracy compared to standalone models (36.8% vs 34.2% for GPT-based systems, 30.8% vs 27.4% for Gemini-based systems). This performance gap highlights CLiViS’s effectiveness in complex embodied reasoning tasks, where its iterative LLM-VLM interaction and dynamic cognitive mapping enable deeper analysis compared to standard end-to-end approaches.

4. Reasoning Dynamics

Figure 1 illustrates a multivariate bubble chart, where the horizontal axis represents video duration, the vertical axis indicates the number of inference rounds performed by CLiViS, bubble color represents the accuracy (darker is better) and bubble size reflects the number of samples. From the visualization, we have the following observations: (1) While most videos can be resolved within 2–3 reasoning

Table 1. **Performance Comparison on Various Question Categories of the EgoTempo benchmark.** Best results are marked in **bold**, and the second-best is underlined.

Models	AO	LO	SR	OS	AS	OA	TE	FU	AC	OC	Avg.
Socratic-based Models											
Qwen2.5-VL + Qwen2.5-Max	8.0	10.0	0.0	4.0	0.0	6.0	6.0	18.0	0.0	6.0	5.8
Qwen2.5-VL + DeepSeek-V3	10.0	4.0	4.0	2.0	4.0	12.0	4.0	10.0	0.0	2.0	5.2
InternVL3 + Qwen2.5-Max	6.0	2.0	2.0	4.0	0.0	4.0	2.0	4.0	0.0	0.0	2.4
InternVL3 + DeepSeek-V3	10.0	2.0	0.0	0.0	0.0	4.0	4.0	8.0	0.0	2.0	3.0
VideoLLaMA3 + Qwen2.5-Max	10.0	4.0	2.0	2.0	2.0	6.0	10.0	10.0	0.0	2.0	4.8
VideoLLaMA3 + DeepSeek-V3	14.0	2.0	0.0	2.0	4.0	8.0	10.0	18.0	4.0	2.0	6.4
End-to-End VLM Models											
Qwen2-VL	18.0	8.0	<u>28.0</u>	4.0	0.0	6.0	10.0	2.0	18.0	10.0	9.6
Qwen2.5-VL	<u>28.0</u>	18.0	24.0	10.0	4.0	6.0	12.0	26.0	16.0	18.0	16.2
InternVL2.5	20.0	14.0	14.0	4.0	2.0	8.0	10.0	4.0	12.0	24.0	10.2
InternVL3	22.0	22.0	22.0	6.0	2.0	6.0	20.0	8.0	20.0	42.0	17.0
VideoLLaMA3	<u>28.0</u>	<u>26.0</u>	24.0	4.0	2.0	24.0	<u>22.0</u>	12.0	14.0	42.0	19.8
Video Reasoning Models											
VideoAgent	20.0	10.0	0.0	6.0	2.0	6.0	12.0	14.0	0.0	6.0	7.6
Video-R1	18.0	12.0	22.0	4.0	0.0	12.0	14.0	18.0	28.0	<u>36.0</u>	16.4
CLiViS (Qwen2.5-VL)	20.0	18.0	32.0	12.0	<u>12.0</u>	18.0	16.0	<u>22.0</u>	<u>26.0</u>	16.0	19.6
CLiViS (InternVL3)	36.0	30.0	22.0	28.0	16.0	12.0	28.0	16.0	8.0	32.0	<u>23.0</u>
CLiViS (VideoLLaMA3)	36.0	22.0	32.0	<u>16.0</u>	16.0	<u>22.0</u>	20.0	<u>22.0</u>	18.0	30.0	23.4

Table 2. **Comparison with Frontier Multimodal Models on EgoTempo.**

Models	Accuracy (%)
GPT-4.1	34.2
Gemini-2.5-flash	27.4
CLiViS (w/ GPT-4.1)	36.8
CLiViS (w/ Gemini-2.5-flash)	30.8

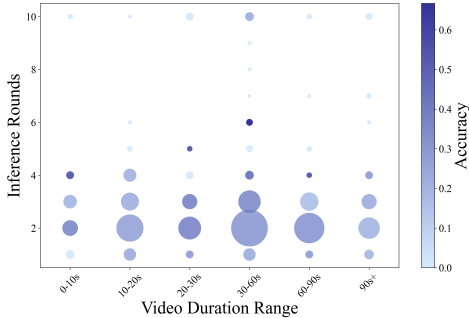


Figure 1. **Correlation between reasoning rounds and video duration.**

rounds, the distribution of reasoning steps gradually shifts toward 3–5 rounds as video duration increases, indicating that CLiViS adaptively allocates deeper reasoning for complex inputs. (2) Within each duration range, a positive correlation emerges between the reasoning rounds and the accuracy, revealing that deeper reasoning may lead to higher-quality responses. These observations highlight CLiViS’s ability to dynamically modulate reasoning rounds based on temporal complexity, achieving a favorable balance be-

tween efficiency and performance.

5. More Qualitative Results

To illustrate CLiViS’s reasoning process and interpretability, we present a qualitative case from the EgoSchema dataset in Figure 2 to Figure 4, that traces every step from cognitive initialization to sub-instruction generation and execution and cognitive update. We highlight relevant information throughout the reasoning trace, marking correct details in green and erroneous ones in red. Overall, this example validates four key advantages of CLiViS:

- **Comprehensive scene parsing.** Segmenting the video into clips and generating per-segment descriptions enable the VLM to better capture fine-grained details, resulting in a more comprehensive and precise cognitive graph.
- **Clear task decomposition.** At each reasoning iteration, the LLM allocates the next perception task based on accumulated cognition, ensuring that every subtask yields fresh and relevant insights.
- **Self-correcting ability.** It demonstrates the ability to proactively detect and resolve errors or inconsistencies in the VLM-generated descriptions, significantly improving the overall robustness and logical consistency of the reasoning.
- **Strong interpretability.** Every subtask, VLM response, and cognitive map update is fully traceable, providing transparent decision paths.

These results demonstrate that CLiViS not only maintains high accuracy in embodied visual reasoning tasks involving long-range dependencies and complex semantic instructions, but also delivers an auditable modular inference chain, validating both its effectiveness and interpretability.



Question

Considering the main actions in the video, how would you summarize the video's central activity, and what role does the interaction with the girl play in this context?

- option 0: While c is meticulously cleaning a shelf, the brief interaction with the girl serves as a minor interruption, but ultimately does not affect the central activity being performed.
- option 1: Currently, c is constructing a shelf diligently. the fleeting interaction with the girl serves as a brief interruption, yet it does not impact or influence the central activity significantly.
- option 2: C is painting a shelf. the interaction with the girl is a brief distraction, but does not affect the central activity.
- option 3: C is repairing a shelf. the interaction with the girl is a brief interruption, but does not affect the central activity.
- option 4: Currently, c is diligently decorating a shelf. the brief, friendly interaction with the girl momentarily interrupts, but ultimately does not affect the central activity at hand.

Ground Truth: option 2

Segmented Description (VLM)

00:00-00:30

Action: A man is **painting a shelf** with white paint. He dips his **brush** into the paint and then paints the shelf.\nDetails: The video shows a person diligently painting a shelf. They are mixing paint in a **bucket** and applying it to the shelf with a brush. The **workspace** is cluttered, indicating ongoing work or repairs.

00:30-01:00

Action: A man is **painting a shelf** with white paint. He dips his brush into the paint and then applies it to the shelf.\nDetails: The scene is set in what appears to be a workspace or garage, with tools and materials scattered around. The person is wearing a light-colored sweatshirt and seems focused on their work. There are additional paint containers nearby, indicating ongoing or multiple painting projects.

01:00-01:30

Action: A man is **painting a shelf** white. He **dips his paintbrush into the paint bucket** and then paints the shelf.\nDetails: The scene appears to be in a workshop or garage setting, with tools and other materials visible around them.

01:30-02:00

Action: A man is **painting a wall** with white paint. He dips his brush into the paint bucket and then paints the wall.\nDetails: The setting appears to be a basement or utility area, with various tools and materials scattered around, including another paint bucket and some boxes.

02:00-02:30

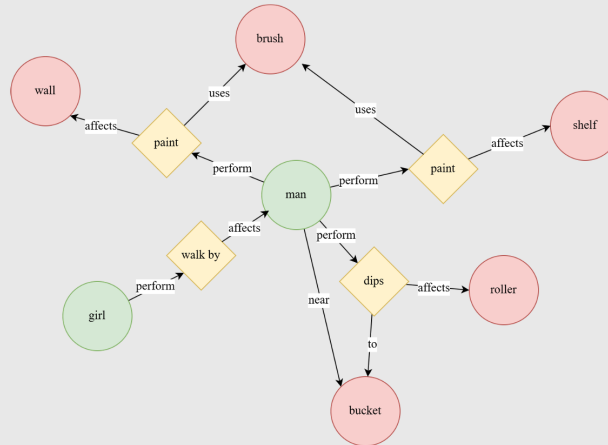
Action: A man is **painting a wall** with a roller. He dips the roller into a bucket of paint and then applies it to the wall.\nDetails: The video shows a person diligently painting a wall. They are using a paint roller and a bucket of white paint, moving back and forth to apply the paint evenly. The scene appears to be in a room under renovation or decoration, with tools and materials visible around them.

02:30-03:00

Action: A man is **painting a shelf** with white paint. He dips his brush into the paint can and applies it to the shelf. **A young girl walks by and watches him paint.**\nDetails: The video shows a person painting a shelf in a room. They dip a brush into a paint can and apply white paint to the shelf's surface. The scene includes another person observing, and there are various tools and materials around, indicating a DIY or renovation setting.\nLast frame: The video ends with the person continuing to paint.

Figure 2. Qualitative results of Segmented Description.

Cognitive Map & Memory Initialization (LLM)



Key Entities: "man", "girl", "shelf", "brush", "bucket"

Global Analysis (VLM)

To summarize the central activity in the video, it appears that the person is engaged in painting a shelf. The individual is seen dipping a paintbrush into a bucket of white paint and applying it to the surface of the shelf. This repetitive action suggests that the primary focus is on painting or refurbishing the shelf.

Regarding the interaction with the girl, there is no visible interaction between the person and any girl in the provided frames. Therefore, any mention of an interaction with a girl does not align with the visual content of the video. The options mentioning interactions with a girl are likely incorrect based on the available evidence.

Sub-Instruction Generation

Round 1:

LLM:

Analysis: The current information indicates that the central activity in the video is painting, specifically a shelf or wall. However, there is conflicting information about the interaction with the girl. The preliminary conclusion states there is no visible interaction with the girl, but the description for the time period 00:02:30-00:03:00 mentions a young girl walking by and watching the man paint. This suggests that the interaction with the girl might indeed occur, but it is brief and does not affect the main activity. To confirm this and ensure accuracy, we need to clarify whether the girl's presence and the interaction are depicted in the video.

```
{
  "period": "00:02:30-00:03:00",
  "instruction": "Does the video show a girl walking by and watching the man paint?"
}
```

VLM:

(00:02:30-00:03:00) Yes.

Rationale:

```
{'rationale': 'The video shows a girl walking by and watching the man paint', 'related_area': 'workspace'}
```

Figure 3. Qualitative results of Cognitive Map and Sub-Instruction Generation.

Round 2:**LLM:**

Analysis: The video content clearly shows that the central activity is painting, specifically a shelf or wall. During the period 00:02:30-00:03:00, a young girl is observed walking by and watching the man paint. This confirms that there is indeed an interaction, albeit brief, with the girl. Given this information, the best fitting option would be one that acknowledges the central activity of painting and recognizes the girl's presence as a minor distraction that does not impact the main task.

Instruction:

```
```json
{
 "period": "full video",
 "instruction": "Is the central activity consistently painting throughout the video?"
}
...

```

**VLM:**

(full video) Yes, the man paints the shelf throughout the video.

**Rationale:**

{'rationale': 'The central activity in the video is consistently painting, as confirmed by observing the full video.', 'related\_area': 'workspace'}

---

**Round 3:****LLM:**

Analysis: The video content consistently shows a man engaged in painting activities throughout all periods, primarily focusing on painting a shelf or wall. The central activity is confirmed to be painting, with no significant deviation from this task. During the period 00:02:30-00:03:00, a young girl is observed walking by and watching the man paint. This indicates a brief interaction, though minimal and non-intrusive, aligning with the description that it serves as a minor distraction but does not affect the central activity. Based on this analysis, the best fitting option would be one that acknowledges the central activity of painting and recognizes the girl's presence as a brief distraction.

```
```json
{
  "reason": "The video consistently shows the man painting a shelf, with a brief interaction involving a girl walking by and observing during 00:02:30-00:03:00. This aligns with Option 2, where the central activity is painting and the interaction is a minor distraction.",
  "answer": "option 2"
}
...

```

Correct

Figure 4. Qualitative results of Iterative Reasoning and Answer Generation.