LongVideoAgent: Multi-Agent Reasoning with Long Videos

Anonymous ACL submission

Abstract

Recent advances in multimodal LLMs and systems that *use tools* for long-video QA point to the promise of reasoning over hour-long episodes. However, many methods still compress content into lossy summaries or rely on limited toolsets, weakening temporal grounding and missing fine-grained cues. We propose a multi-agent framework in which a master LLM coordinates a grounding agent to localize question-relevant segments and a vision agent to extract targeted textual observations. The master agent plans with a step limit, and is trained with reinforcement learning to encourage concise, correct, and efficient multi-agent cooperation. This design helps the master agent focus on relevant clips via grounding, complements subtitles with visual detail, and yields interpretable trajectories. On our proposed *LongTVQA* and *LongTVQA+* which are episode-level datasets aggregated from TVQA/TVQA+, our multi-agent system significantly outperforms strong non-agent baselines. Experiments also show reinforcement learning further strengthens reasoning and planning for the trained agent.

1 Introduction

Multimodal large language models (MLLMs) extend LLMs beyond text to perceive and reason over multimodal signals, such as visual frames, audio, and subtitles. A key emerging challenge is robust *long video* understanding, where information is sparsely distributed across hours of content and multiple modalities (e.g., frames, and dialogue cues). Early instruction-tuned systems such as Video-LLaMA (Zhang et al., 2023a; Lin et al., 2024a) demonstrated that LLMs can be adapted to jointly process sampled video frames, marking an initial step toward multimodal video reasoning. However, current models remain limited to short clips or coarse summaries and struggle with fine-grained, temporally extended queries. Crucially, most prior systems are *non-agentic* models: they process a static, pre-encoded or down-sampled video. Converting the full visual stream into compressed representations in the LLM's textual space shifts the burden of temporal reasoning to this early stage—often lossy and irreversible, making it difficult to recover fine-grained evidence. These limitations motivate an *agentic*, tool-augmented paradigm that can actively decide what to observe next, when to query external visual or other tools, and when enough grounded evidence has been gathered to respond. Despite recent advances, the field still lacks a solution that jointly achieves efficiency, multimodal completeness, and fine-grained temporal reasoning in long videos.

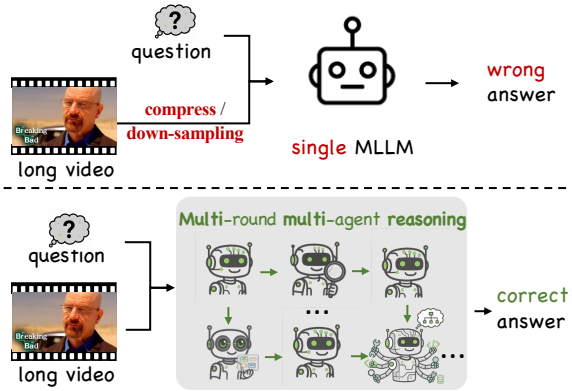

Figure 1: Traditional single-pass MLLMs that ingest entire long videos in one context—typically (may through heavy downsampling and compression) often miss crucial evidence and produce wrong answers, whereas LongVideoAgent conducts *multi-agent*, *multi-round*, and *multimodal* reasoning to extract sparse, task-relevant cues and answer correctly.

Recent works have begun to frame long video understanding as an *agent-driven process*, rather than a passive encoding task. Notably, VideoAgent (Fan et al., 2024; Wang et al., 2024b) introduced an agent-based framework where a central LLM actively conduct video analysis. In this paradigm, the LLM agent iteratively queries external vision mod-

els (tools) to retrieve and interpret video frames, progressively compiling the information needed to answer a given query. This interactive strategy mirrors human cognitive behavior and has demonstrated promising effectiveness. These findings highlight the potential of *tool-augmented LLM agents* in achieving both efficiency and accuracy. However, the initial incarnation of VideoAgent relies on a less powerful toolset, primarily generic vision-language foundation models for captioning and image retrieval. Such tools are often insufficient for capturing fine-grained semantics, precise object references, or subtle temporal cues. This restricts the agent's ability to understand complex scenes and reason over long temporal spans. Moreover, current frameworks underutilize the LLM's inherent reasoning abilities and lack mechanisms for multi-step decision making or reinforcement-based planning.

In this paper, shown as Figure 1 we address these challenges by proposing a new *multi-agent-based framework* for long video understanding that strategically incorporates agents. Our system adopts a multi-agent architecture, where a central MASTERAGENT is responsible for reasoning and answering, while coordinating with other specialized agents. Specifically, a GROUNDINGAGENT locates video segments relevant to the question, and a VISIONAGENT extracts detailed visual information from the selected clips (e.g., objects, faces, actions). The master agent gathers these outputs to iteratively reason over the accumulated evidence. To guide the reasoning process, we design a reward-driven training strategy that encourages the master agent to conduct structured, multi-step reasoning. In each iteration, the master agent generates sub-queries, invokes either the grounding or vision agent as needed, and integrates the returned information before deciding on the next step. When it determines that enough evidence has been collected, it produces a final answer. By designing a reward function that penalizes irrelevant tool use and incoherent reasoning, we guide the agent to "think" in a proper format, effectively learning when to explore the video with tools and when it has gathered sufficient evidence to answer the question. Furthermore, to evaluate long-form video reasoning in a realistic setting, we construct a new benchmark dataset *LongTVQA* and *LongTVQA+*. This dataset extends the well-known TVQA video question answering task to much longer video durations, providing a rigorous testbed for our agent.

Our *Agent-with-Tools* approach demonstrates superior performance on the LongTVQA benchmark, outperforming all existing baselines by a significant margin. Through ablation studies, we show that both the multi-agent architecture and the reward-guided training contribute critically to the agent's gains. Our system not only achieves higher accuracy, but also exhibits interpretable decision-making, coordinating sub-agents to select relevant video segments and extract fine-grained visual information essential for reasoning. These results underscore the benefit of an agentic framework for long video understanding.

Our contributions are threefold:

- a modular *multi-agent* architecture in which a master LLM coordinates grounding and vision specialists.
- a *reward-driven* agentic reinforcement learning training scheme that promotes concise, step-wise reasoning.
- episode-level long video datasets LongTVQA and LongTVQA+ are proposed under which our system achieves state-of-the-art results.

2 Related Work

2.1 Video Question Answering

Early work focused on memory and attention mechanisms over appearance–motion features (Gao et al., 2018). This evolved into multimodal transformers designed for efficient frame sampling (Lei et al., 2021). Recent trends emphasize retrieval-aware reasoning and efficient tokenization for long videos, as well as integrating LLM-based reasoning with video encoders (Zhang et al., 2023a) and employing agentic planners that iteratively gather evidence (Wang et al., 2024b). Long-form systems further explore sparse memory and temporal grounding techniques to handle hour-scale inputs (Song et al., 2024). Meanwhile, some approaches focus on identifying or generating sparse key evidence to support long-video QA (Song et al., 2025). These developments motivate long-form VideoQA systems that selectively retrieve segments under a limited context budget.

2.2 LLM Agents

LLM agents couple chain-of-thought with *actions*: planning, tool calls, and iterative evidence gathering. Foundational agent ideas include ReAct

(interleaving reasoning and actions), Self-Ask (decomposition with search), and WebGPT (browser QA with citations) (Yao et al., 2023b; Press et al., 2022; Nakano et al., 2021). Toolformer shows self-supervised API-calling, while orchestration frameworks (HuggingGPT/Gorilla-style) route subtasks to expert models (Schick et al., 2023; Shen et al., 2023). In multimodal settings, MM-ReAct wires LLMs to vision experts via prompting, and program-of-thought systems like ViperGPT compose perception modules through executable code for transparent, verifiable reasoning (Yang et al., 2023; Surís et al., 2023). For long videos, agentic designs such as VideoAgent/VideoAgent-style frameworks use memory, targeted retrieval, and temporal grounding to operate under strict context budgets while improving faithfulness (Wang et al., 2024b). Beyond planning, video-RAG pipelines extract ASR/OCR/objects and retrieve evidence to augment LVLMs for factual responses (Luo et al., 2024). In addition, long-horizon multimodal agents with persistent memory and structural planning further enhance reliability for extended videos, e.g., Long-Seeing, VideoTree, and Koala (Long et al., 2025; Wang et al.; Tan et al., 2024); and general reasoning paradigms such as Chain-of-Thought, Least-to-Most, Tree-of-Thoughts, and Generative Agents provide foundations for decomposition and memory (Wei et al., 2023; Zhou et al., 2023; Yao et al., 2023a; Park et al., 2023). Retrieval-first paradigms like Retrieving-to-Answer complement agent pipelines with a retrieve-then-reason template (Pan et al., 2023). (We also include the alternative ReAct entry for key consistency (Yao et al., 2023c).)

2.3 Multi-Modal LLMs

Modern MLLMs combine strong vision encoders with instruction-tuned LLMs. CLIP pretraining provides broad visual–text transfer (Radford et al., 2021). Flamingo introduces a perceiver-style resampler for few-shot multimodal learning (Alayrac et al., 2022); BLIP-2/InstructBLIP bridge frozen encoders and LLMs (Li et al., 2023a; Dai et al., 2023). LLaVA scales visual instruction tuning with open ingredients, and LLaVA-OneVision unifies high-resolution perception and efficient tokens for images/video (Liu et al., 2023, 2024). Recent video-tuned variants (e.g., Video-LLaVA) and training-free token schedulers (e.g., SlowFast-LLaVA) further improve temporal coverage and efficiency (Lin et al., 2024b; Xu et al., 2024b). Proprietary MLLMs (GPT-4/4o; Gemini 1.5) show long-context multimodal reasoning (Achiam et al., 2023; Gemini Team, 2024), while open models (Qwen2-VL, InternVL) narrow the gap via dynamic resolution, OCR, and video pipelines (Wang et al., 2024a; Chen et al., 2024). Complementary advances focus on unifying image–video tokens with few, informative representations (e.g., MiniGPT4-Video, Video-ChatGPT, Video-LaVIT, LLaMA-VID, LongVU, PLLaVA, LLaVA-Video, Chat-UniVi) (Ataallah et al., 2024; Maaz et al., 2024; Jin et al., 2024b; Li et al., 2023b; Shen et al., 2024; Xu et al., 2024a; Zhang et al., 2025; Jin et al., 2024a), and on long-context optimization or adaptive input selection (e.g., InternVideo2.5, LongVLM, Long Context training, self-adaptive sampling, simple-but-effective alignment, and question-instructed tuning) (Wang et al., 2025; Weng et al., 2024; Zhang et al., 2024b; Han et al., 2024; Zhang et al., 2024a; Romero and Solorio, 2024). Comprehensive analyses of video understanding in large multimodal models (e.g., Apollo) situate these models within broader capabilities and evaluation protocols (Zohar et al.). For key harmonization with the bibliography, we also include the alternate Video-LLaMA entry (Zhang et al., 2023b). However, most models still face long-video constraints (context length, retrieval). This motivates combining video-native encoders, instruction tuning, retrieval, and tool use for scalable long-form VideoQA.

3 Method

As shown in Figure 2, we cast long-video QA as *multi-agent reasoning*, where a master agent LLM coordinates a grounding agent to temporally localize question-relevant segments and a vision agent to extract targeted observations from those segments. The system proceeds iteratively, maintaining a running context that accumulates subtitles, relevant segment tags, and vision observations, and it produces an answer once the master agent judges that sufficient evidence has been gathered. For open-source LLMs serving as the master agent, we apply reinforcement learning to encourage accurate, concise, and cooperation-efficient behavior while keeping the other agents frozen. At inference, the process yields clear, step-by-step traces aimed at solving the question at hand.

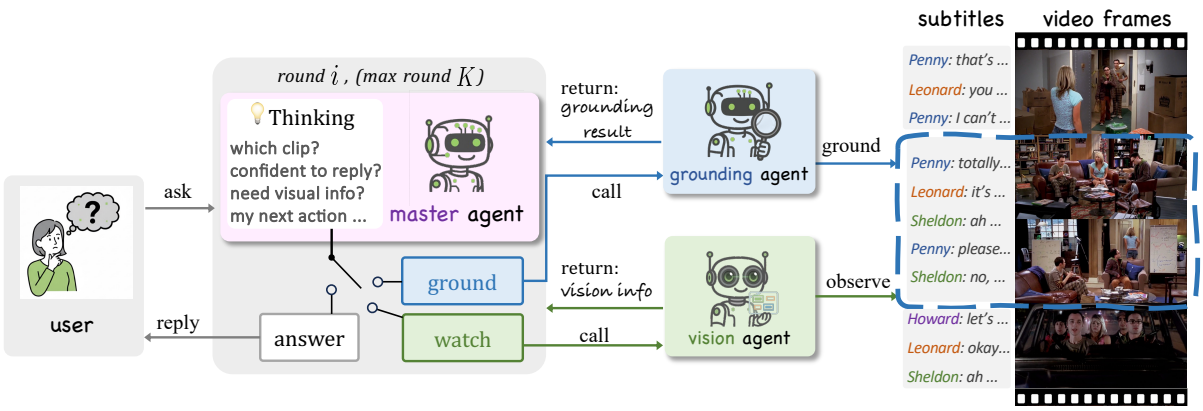

Figure 2: Architecture of LongVideoAgent. A MasterAgent runs for up to K rounds, collaborating with a GroundingAgent to localize relevant clips from videos and a VisionAgent to read fine-grained cues from the localized frames. Evidence accumulates until the MasterAgent feels confident to answer the user.

Table 1: System prompt for LongVideoAgent.

System Prompt — LongVideoAgent
You are an agent that answers questions about a long video episode. You may use two tools: a *grounding agent* to localize relevant segments and a *vision agent* to extract visual facts from the localized segment. Produce *concise, direct* answers.
Context you may receive. All subtitles and the user question q. When a segment has been localized, you will also have a tag `<clipX>` (e.g., `<clip2>`). When the vision agent has been called, you will see its textual response.
Available actions (choose exactly one per turn). **A — Visual query:** If current visual information is insufficient, or you need visual details conditioned on the subtitles for the current `<clipX>`, call the vision engine with `<visual_query>query </visual_query>`. **B — (Re)Grounding:** If the current text/visual evidence conflicts with the question, or the current location cannot support a confident answer, call the grounding agent with `<request_grounding>`. **C — Answer:** If evidence is sufficient, return the final answer with `<answer>... </answer>`. The answer must be concise and direct.
Guidelines. (1) Be conservative with tool calls; answer when sufficient. (2) Do not hallucinate visual details; only use the vision agent for facts not inferable from subtitles. (3) Each turn targets the current `<clipX>` (if any); if none exists, prefer (re)grounding before visual query.

3.1 Multi-agent System Framework

Master agent behavior and training. Specifically, the master agent follows the instruction schema in the *System Prompt* (Table 1) and the multi-turn policy in Algorithm 1that coordinates two other specialist agents: a grounding agent and a vision agent. Given an episode with its full subtitles and a question, the master runs a bounded loop (at most K steps). At each turn it emits exactly one structured action token, `<watch>` for a visual read, `<request_grounding>` for (re)localization, or `<answer>` to terminate. After the corresponding agent is invoked, its textual output is appended to the context of the master agent. For open-source masters, we optimize the policy with GRPO while keeping the grounding and vision agents fixed. the rollouts terminated by action tokens in Algorithm 1 provide the trajectories for training and evaluation.

Grounding agent. Given the question and subtitles, the grounding agent proposes a temporal segment and returns a symbolic tag `<clip_X>` marking the relevant portion of the episode. By default the window context is 1; when larger, the agent outputs a short run of consecutive tags. The master may re-query grounding to refine or validate the segment as reasoning progresses.

Vision agent. Conditioned on `<clip_X>` and an on-demand prompt that specifies the current visual need, the vision agent extracts textual observations from frames within the localized segment (e.g., objects/entities, attributes, actions, OCR/on-screen text, scene cues). These observations are appended to the context and guide the next decision; the loop terminates when the master judges the accumulated visual evidence sufficient to answer.

3.2 Reinforcement Learning for LongVideoAgent

For open-source LLMs serving as the master agent, we fine-tune the master with GRPO while keeping the grounding and vision agents frozen. Long-video QA is cast as a finite-horizon decision pro-

Algorithm 1 LONGVIDEOAGENT with Multi-Turn Reasoning

Require: Subtitles $\mathcal{S}$; question q; video V; MASTERAGENT parameters π_θ; maximum steps K; GROUNDINGAGENT; VISIONAGENT.
Ensure: Final answer $\hat{y}$.

```
Initialize rollout sequence y ← ∅
Initialize step count t ← 0
while t < K do
    Initialize current action LLM rollout sequence y_t ← ∅
    while True do
        Generate response token y_i ~ π_θ(· | S, q, V, y + y_t)
        Append y_i to rollout sequence y_t ← y_t + y_i
        if y_i in [</visual_query>, </request_grounding>, </answer>, <eos>] then
            break
        end if
    end while
    y ← y + y_t
    if <visual_query> detected in y_t then
        Extract visual query q_vis ← PARSE(y_t, <visual_query>, </visual_query>)
        Retrieve vision results d = VISIONAGENT(q_vis, V)
        Insert clip tag with subtitles into rollout y ← y + d
    else if <request_grounding> detected in y_t then
        Retrieve grounding results clipTag = GROUNDINGAGENT(q, S)
        Insert clip tag into rollout y ← y + clipTag + S(clipTag)
    else if <answer> detected in y_t then
        Extract predicted answer ŷ ← PARSE(y_t, <answer>, </answer>); Normalize ŷ (trim spaces/punctuation)
        Insert final answer into rollout y ← y + "The answer is: " + ŷ
        return final answer ŷ
    else
        Ask for rethink y ← y + "The action is not correct. Only <visual_query>, <request_grounding>, or <answer>."
    end if
    Increment step count t ← t + 1
end while
return final generated response y for q
```

cess: at each action step after reasoning the policy emits exactly one structured action token (`<watch>`, `<request_grounding>`, or `<answer>`).

Trajectory. A full response terminates upon emitting `<answer>...</answer>` or reaching K steps. We index decision steps by $t \in \{0, 1, \ldots, T\}$ with $T \leq K$. At step t, the policy π_θ first plans and then emits a contiguous action string a_t ending with exactly one closing tag from {`</watch>`, `</request_grounding>`, `</answer>`}. If not terminating, the system appends feedback from the invoked agent o_t (e.g., a vision observation or a clip tag) to the context for the next step.

Rewards. We use two simple, rule-based rewards as supervision for reinforcement learning: (i) *Structural validity* $r_t^{\text{fmt}} \in \{0, 1\}$ grants 1 if the action string contains exactly one top-level tag with proper closure and no extraneous text; otherwise 0. (ii) *Answer correctness* $r^{\text{ans}} \in [0, 1]$ is awarded at termination via exact match on the multiple-choice answer; if no valid `<answer>` appears, $r^{\text{ans}} = 0$.

Objective and optimization. We seek a policy that produces well-formed actions at every step and a correct final answer. To balance these goals, the trajectory reward return is $R(\tau) = \alpha \sum_{t=0}^{T} r_t^{\text{fmt}} + r^{\text{ans}}$ where $\alpha > 0$ weights the per-step structural shaping and r^{ans} supplies the terminal task reward. r_t^{fmt} encourages the master to emit exactly one correct action tag at each decision, while r^{ans} evaluates only the final `<answer>`. If no valid and correct answer is produced, $r^{\text{ans}} = 0$.

We optimize the master agent with GRPO on sampled rollouts: for each episode, the policy generates an action sequence, receives structural rewards at action boundaries and a terminal answer reward, and we compute sequence-level advantages with a learned value baseline. Policy updates follow the GRPO objective with standard clipping and entropy regularization, while the grounding and vision agents remain frozen. This minimal, two-signal objective provides sufficient guidance to learn structured, multi-turn coordination without additional dense rewards.

4 Experiments

4.1 Datasets

We build *LongTVQA* and *LongTVQA+* on top of TVQA and TVQA+. TVQA spans six TV shows with 152.5K multiple-choice QAs over 21.8K clips (60–90s) with subtitles and moment annotations;

Table 2: Performance on *LongTVQA* and *LongTVQA+*. The left block lists model attributes (*Agentic*, *Input*, *RL fine-tune*); the right block reports validation accuracy (%). GPT-4o and Gemini-2.5 Pro are *multimodal* baselines that process and accept the full long video directly. Methods labeled `Agentic` indicate the model operates as the MasterAgent; methods labeled `AgenticRL` additionally denote RL fine-tuning. Parenthesized green numbers denote absolute gains over the immediately preceding (non-agentic or non-RL) setting. We observe that: (i) our multi-agent framework,LongVideoAgent, consistently outperform their non-agentic counterparts; (ii) agentic RL yields additional gains, especially for smaller open-source models; (iii) using frames provides visual evidence beyond subtitles, and generally outperforms subtitle-only inputs; (iv) closed-source models remain strong, but the gap narrows much when open-source models adopt agentic designs and agentic RL.

Method	**Multi-agent**	**Input**	**RL Finetune**	**Accuracy (%)**	
				LongTVQA	LongTVQA+
Closed-source (M)LLMs					
GPT-4o	✗	Subtitle+Frame	✗	70.78	78.32
Gemini-2.5 Pro	✗	Subtitle+Frame	✗	78.90	81.28
GPT5-mini	✗	Subtitle	✗	62.40	66.70
Agentic-GPT5-mini	✓	Subtitle+Frame	✗	71.11(+8.71)	78.90(+12.20)
Grok	✗	Subtitle	✗	76.90	81.80
Agentic-Grok	✓	Subtitle+Frame	✗	**82.65**(+5.75)	**85.60**(+3.80)
Open-source LLMs					
DeepSeek671B	✗	Subtitle	✗	68.99	75.04
Agentic-DeepSeek671B	✓	Subtitle+Frame	✗	70.30(+1.31)	79.70(+4.66)
Agentic-Qwen3B	✓	Subtitle+Frame	✗	23.50	27.70
AgenticRL-Qwen3B	✓	Subtitle+Frame	✓	47.40(+23.90)	50.10(+22.40)
Agentic-Qwen7B	✓	Subtitle+Frame	✗	46.10	60.30
AgenticRL-Qwen7B	✓	Subtitle+Frame	✓	**60.20**(+14.10)	**70.80**(+10.50)

questions require joint dialogue–visual reasoning (Lei et al., 2018). TVQA+ refines a subset with spatio-temporal grounding—adding precise timestamps and 310.8K frame-level boxes for referenced entities (29.4K QAs from 4,198 clips, mainly TBBT)—supporting joint QA and temporal/spatial localization (Lei et al., 2020).

To obtain *LongTVQA* and *LongTVQA+*, we aggregate all clips from the same TV episode into a single *episode-level* (hour-scale) sequence. For each episode, we merge the visual stream, subtitles, and all associated questions; clip timestamps are re-indexed into the episode timeline, and TVQA+ bounding boxes are preserved at their corresponding frames. Unless otherwise noted, we report results on the original validation splits after this episode-level aggregation.

4.2 Baselines

We include both open-source and closed-source models (see Table 2), including representative open-source LLMs such as *DeepSeek-R1* (DeepSeek-AI et al., 2025) and *Qwen2.5-3B/7B* (Qwen et al., 2025) , and closed-source models such as *Grok*, *GPT5-mini* (OpenAI, 2025),GPT-4o (OpenAI) and *Gemini 2.5 Pro* (Comanici and so on, 2025). We adopt unified experimental settings to ensure comparability across backbones. We evaluate *base LLMs in a non-agent mode* and *our agent system*, with the base LLM serving as the master agent in both cases. In the non-agent mode, the model consumes the full subtitles and does not invoke grounding or vision agents. We then compare our agent system against the corresponding non-agent runs on the *same* backbones, so that observed gains can be attributed to agentic behavior rather than backbone differences. For open-source backbones, we also report results *with vs. without* reinforcement learning under the identical evaluation protocol described in §4.3. Closed-source models are evaluated as released, without additional training.

4.3 Experimental Setup

By default we use Grok-4-fast-reasoning for temporal localization and GPT-4o as the vision agent. The window context is set to 1, meaning the agent conditions on a single localized clip (no adjacent clips), and the maximum execution steps are $K=5$. All methods read the full episode subtitles. In the non-agent setting, no external modules are invoked. In the agent setting, the master agent receives symbolic temporal tag(s) `<clip_X>`, which is produced by the grounding agent, that marks the grounded clip(s) on the episode timeline, and it may request on-demand *textual* observations from the vision agent via prompting for selected frames within

Table 3: A case of LongVideoAgent showing the collaboration among multiple agents.

Question & Choices
Question: Where is Sheldon sitting when he is accompanied by a man? Choices: a0: The Mall a1: A Movie Theatre a2: A Park **a3: A Bus Stop** ✓ a4: The Store
LongVideoAgent Execution
Master Agent: <think> Subtitles alone do not specify the location; I will first localize with grounding, then request a visual read. </think> <request_grounding> ⇒ Grounding Agent
Grounding Agent: *Grounding information.* Subtitles: *"Sheldon: … apparently we're just two peas in a pod. … Mrs Cooper: Here. Thank you."* Clip tag: <s05e06_seg02_clip_15>
Master Agent: <think> The grounded subtitles still don't reveal the setting; I'll request a visual description of the localized clip. </think> <visual_query> visual description of scene in *The Big Bang Theory* S05E06 segment 02 clip 15 where Sheldon is with a man </visual_query> ⇒ Vision Agent
Vision Agent: *Visual description.* Sheldon is sitting on a bench at night in an urban sidewalk setting; a waste container and building windows are visible nearby.
Master Agent: <think> Bench + sidewalk + trash can + windows strongly indicate a bus stop rather than a mall, theatre, store, or park. No further tools needed. </think> <answer> a3: A Bus Stop ✓ </answer> ⇒ User

that segment (e.g., objects and attributes, OCR/text, brief scene cues). The master agent only consumes text (subtitles, the <clip_X> tag, and optional visual observations); no raw images are passed to the master agent. We report results on the *validation* splits of LongTVQA and LongTVQA+, using *Answer Accuracy (Acc)* as the primary metric (the questions are multiple choice) and additionally *Grounding Accuracy* for experiments that involve clip grounding. For reinforcement learning, we use GRPO with a learning rate of 5×10^{-6}, up to 2,000 optimization steps, a KL coefficient of 10^{-3}, batch size 4, rollout count N=4, and temperature 1.0. Training Qwen2.5-7B took 12 hours on 4× NVIDIA H800 GPUs, while the 3B variant took 6 hours under the same setup.

4.4 Performance

Table 2 presents overall validation accuracy. Moving from the non-agent setting to our agent with temporal grounding yields an absolute gain of *4.7* points; enabling vision after grounding brings a further *5.8*-point improvement, establishing the best accuracy in the table. This provides direct evidence for the effectiveness of a *multi-agentic* pipeline

Table 4: Ablations and analysis of LongVideoAgent.

(a) **Comparison of non-agent vs. multi-agent performance.** Agentic components progressively improve performance: adding grounding outperforms the non-agent baseline, and adding vision agent yields the best results.

Setting	**Accuracy (%)**
Non-agent (Text-only)	64.3
Multi-Agent (Grounding)	69.0
Multi-Agent (Grounding + Vision)	**74.8**

(b) **Effect of max steps K.** Increasing the MasterAgent step budget generally raises both grounding and overall accuracy until reaching a saturation position.

K	**Grounding Accuracy (%)**	**Accuracy (%)**
2	67.00	68.30
5	71.00	73.67
10	72.00	73.67

(c) **Effect of evidence window size.** Larger temporal windows supply richer context for grounding and vision.

Window	**Grounding Accuracy (%)**	**Accuracy (%)**
1	71.67	70.33
2	78.67	75.00
3	**81.94**	**77.26**

(d) **Vision model ablation.** Stronger VisionAgent shows higher overall accuracy, reflecting higher quality extraction of visual information from frames.

Vision model	**Grounding Acc. (%)**	**Acc. (%)**
Qwen3-VL-235B	71.00	73.67
GPT-4o	73.30	**78.00**

that can localize the relevant clips and performs targeted visual inspection. In addition, for several open-source LLMs(as master agent), reinforcement learning consistently improves over their inference-only counterparts under identical prompts and evaluation; notably, the Qwen2.5 7B model with RL attains accuracy comparable to *GPT-5-mini* (closed-source) on our protocol. Illustrative examples in Table 3 and Table 5 effectively demonstrate the efficacy of our approach, with additional cases provided in the supplementary materials.

4.5 Ablation Studies and Analysis

Execution step limit K. Table 4b varies the *upper bound* on agent actions per question. Increasing K from 2 to 5 raises temporal localization accuracy from 67.0 to 71.0 (+4.0) and answer accuracy from 68.30 to 73.67 (+5.37). Raising the limit to K=10 yields no additional gain in answer accuracy (73.67) while localization nudges to 72.0, indicating diminishing returns. We therefore adopt K=5 as the default.

Table 5: Example showing LongVideoAgent reasoning and repeatedly invoking the Grounding and Vision agents to gather evidence and answer correctly.

Question & Choices
Question: What side of the bed is Sheldon when he is closer to the window? Choices: **a0: the left side** ✓ a1: the right side a2: the middle a3: not on the bed a4: the chair
LongVideoAgent Execution
Master Agent: <think> Subtitles alone cannot disambiguate left vs. right; I will first ground the segment, then request visual evidence. </think> <request_grounding> localize the referenced segment </request_grounding> ⇒ Grounding Agent
Grounding Agent: *Grounding information.* Subtitles (localized): <s10e04_seg02_clip_09> *"Sheldon: Okay. Now, on this side, … Bernadette: Raj knows,"*
Master Agent: <think> Grounded subtitles still don't reveal which side is nearer the window; I'll request a visual read. </think> <visual_query> Big Bang Theory season 10 episode 4 — bedroom scene layout (Sheldon & Amy); identify bed/window configuration </visual_query> ⇒ Vision Agent
Vision Agent: *Visual description (1).* Bedroom scene; bed centered against the wall; Sheldon sits on the bed; general layout and decor described, but the side nearest the window is not specified.
Master Agent: <think> The first visual read lacks the window–bed side relation; I need a more precise visual query focused on the window's position relative to Sheldon. </think> <visual_query> The Big Bang Theory S10E04 seg02 clip 09 — which side of the bed is next to the window, and where is Sheldon relative to it? </visual_query> ⇒ Vision Agent
Vision Agent: *Visual description (2).* A window with patterned curtains is on the *left* side of the bed; Sheldon is seated against the headboard near that window.
Master Agent: <think> Window is on the left side of the bed and Sheldon is closer to it; therefore he is on the left side. No further tools needed. </think> <answer> a0: the left side ✓ </answer> ⇒ User

Window context. With GPT-4o as the vision agent, expanding the evidence window from 1 to 2 *adjacent clips* improves localization from 71.67 to 78.67 (+7.00) and answer accuracy from 70.33 to 75.00 (+4.67); using 3 clips yields further gains to 81.94 and 77.26 (additional +3.27 and +2.26), as shown in Table 4c. The adjacent context helps disambiguate references across shots, but larger windows require more visual queries and latency while delivering smaller accuracy gains; hence we keep *Window*= 1 for the default and study larger windows in ablations.

Vision model. Table 4d compares perception backbones. GPT-4o attains 73.30 localization and 78.00 answer accuracy, outperforming Qwen3-VL-235B-a22b at 71.00 and 73.67 by +2.30 and +4.33, respectively. The gap indicates that stronger visual recognition (small objects, OCR, fine attributes) translates into better end-task accuracy in long-form QA, so we adopt GPT-4o as the default vision agent.

Contribution of agentic components. Table 4a decomposes the gains when moving *from a single LLM* to a *multi-agent, multimodal* system. Adding temporal grounding to the same backbone increases answer accuracy from 64.3 to 69.0 (+4.7), showing that identifying the relevant clip filters distractors and focuses reasoning. Enabling vision after grounding further lifts accuracy to 74.8 (+5.8 over grounding; +10.5 overall): targeted visual inspection complements subtitles with concrete object/text cues and can validate or refine grounding through repeated calls when uncertain. Because backbones and prompts are held fixed, these improvements are attributable to the agentic procedure. We suggest grounding narrows the context length for reasoning and guide the master agent's attention, while vision supplies the missing fine-grained evidence.

5 Conclusion

We presented a multi-agent framework, LongVideoAgent, for long-form video question answering in which a Master agent coordinates a GroundingAgent for temporal localization and a VisionAgent for targeted perception. The framework is model-agnostic: we evaluate it with both closed- and open-source LLMs; for open-source masters, we fine-tune with GRPO to encourage accurate, concise, and cooperation-efficient behavior while keeping the other agents frozen. Equipped with a unified context and GRPO training that combines structural and answer rewards, the system where open-source LLMs act as the master agent yields transparent, step-by-step traces and achieves strong gains on *LongTVQA* / *LongTVQA+* over non-agent baselines. Ablations show that grounding+vision is essential, modest step limits suffice, adjacent-window context helps, and stronger perception yields higher accuracy, validating the effectiveness of the framework. Future work includes richer modalities(like audio track and knowledge background), finer grounding and larger-scale RL training.

Limitations

Our work has several practical limitations. First, based on TVQA and TVQA+ we rely on provided subtitles as the primary textual channel and do not process raw audio; in future work we plan to integrate an audio-to-subtitles (ASR) module to capture raw speech. Second, the vision and grounding modules are kept fixed during RL. Jointly optimizing them could further improve robustness and accuracy. Lastly, the reward is intentionally simple (format + answer correctness), which may still have room for improvements.

References

Josh Achiam and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac and 1 others. 2022. Flamingo: A visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. 2024. MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding with Interleaved Visual-Textual Tokens. *arXiv preprint*. ArXiv:2404.03413 [cs].

Zheng Chen and 1 others. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*. InternVL 1.5/2 report.

Gheorghe Comanici and so on. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Wenliang Dai, Junnan Li, Dongxu Li, and 1 others. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. VideoAgent: A memory-augmented multimodal agent for video understanding. *arXiv preprint arXiv:2403.11481*.

Jingjing Gao, Runzhou Ge, Kai Chen, and Ram Nevatia. 2018. Motion-appearance co-memory networks for video question answering. In *CVPR*. Dual attention over appearance and motion.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Wei Han, Hui Chen, Min-Yen Kan, and Soujanya Poria. 2024. Self-Adaptive Sampling for Efficient Video Question-Answering on Image–Text Models. *arXiv preprint*. ArXiv:2307.04192 [cs].

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024a. Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. *arXiv preprint*. ArXiv:2311.08046 [cs].

Yang Jin, Zhicheng Sun, Kun Xu, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, Kun Gai, and Yadong Mu. 2024b. Video-LaVIT: Unified Video-Language Pre-training with Decoupled Visual-Motional Tokenization. *arXiv preprint*. ArXiv:2402.03161 [cs].

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*. End-to-end video–language with sparse frames.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. Tvqa: Localized, compositional video question answering. In *EMNLP*.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023b. LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models. *arXiv preprint*. ArXiv:2311.17043 [cs].

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024a. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024b. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint*. ArXiv:2311.10122 [cs].

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*. LLaVA.

Haotian Liu and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. 2025. Seeing, Listening, Remembering, and Reasoning: A Multimodal Agent with Long-Term Memory. *arXiv preprint*. ArXiv:2508.09736 [cs].

Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. 2024. Video-RAG: Visually-aligned Retrieval-Augmented Long Video Comprehension. *arXiv preprint*. ArXiv:2411.13093 [cs].

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv preprint*. ArXiv:2306.05424 [cs].

Reiichiro Nakano and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. In *NeurIPS*.

OpenAI. Hello GPT-4o.

OpenAI. 2025. Introducing GPT-5.

Junting Pan, Ziyi Lin, Yuying Ge, Xiatian Zhu, Renrui Zhang, Yi Wang, Yu Qiao, and Hongsheng Li. 2023. Retrieving-to-Answer: Zero-Shot Video Question Answering with Frozen Large Language Models. *arXiv preprint*. ArXiv:2306.11732 [cs].

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *arXiv preprint*. ArXiv:2304.03442 [cs].

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Jacob Andreas. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*. Introduces Self-Ask and search integration.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, and 1 others. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

David Romero and Thamar Solorio. 2024. Question-Instructed Visual Descriptions for Zero-Shot Video Question Answering. *arXiv preprint*. ArXiv:2402.10698 [cs].

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. 2024. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding. *arXiv preprint*. ArXiv:2410.17434 [cs].

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, and 1 others. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. In *NeurIPS*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. MovieChat: From Dense Token to Sparse Memory for Long Video Understanding. *arXiv preprint*. ArXiv:2307.16449 [cs].

Zhende Song, Chenchen Wang, Jiamu Sheng, Chi Zhang, Shengji Tang, Jiayuan Fan, and Tao Chen. 2025. DreamFrame: Enhancing Video Understanding via Automatically Generated QA and Style-Consistent Keyframes. *arXiv preprint*. ArXiv:2403.01422 [cs].

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *ICCV*.

Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A. Plummer, Bryan Russell, and Kate Saenko. 2024. Koala: Key frame-conditioned long video-LLM. *arXiv preprint*. ArXiv:2404.04346 [cs].

Peng Wang, Shuai Bai, Sinan Tan, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024b. VideoAgent: Long-form video understanding with large language model as agent. *arXiv preprint arXiv:2403.10517*.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025. InternVideo2.5: Empowering Video MLLMs with Long and Rich Context Modeling. *arXiv preprint*. ArXiv:2501.12386 [cs].

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting

Elicits Reasoning in Large Language Models. *arXiv preprint*. ArXiv:2201.11903 [cs].

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. 2024. LongVLM: Efficient Long Video Understanding via Large Language Models. *arXiv preprint*. ArXiv:2404.03384 [cs].

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024a. PLLaVA : Parameter-free LLaVA Extension from Images to Videos for Video Dense Captioning. *arXiv preprint*. ArXiv:2404.16994 [cs].

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. 2024b. SlowFast-LLaVA: A Strong Training-Free Baseline for Video Large Language Models. *arXiv preprint*. ArXiv:2407.15841 [cs].

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, and 1 others. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv preprint*. ArXiv:2305.10601 [cs].

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *ICLR*. Interleaving tool use with chain-of-thought.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023c. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint*. ArXiv:2210.03629 [cs].

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024a. A Simple LLM Framework for Long-Range Video Question-Answering. *arXiv preprint*. ArXiv:2312.17235 [cs].

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint*. ArXiv:2306.02858 [cs].

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024b. Long Context Transfer from Language to Vision. *arXiv preprint*. ArXiv:2406.16852 [cs].

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-Video: Video Instruction Tuning With Synthetic Data. *arXiv preprint*. ArXiv:2410.02713 [cs].

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *arXiv preprint*. ArXiv:2205.10625 [cs].

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, Serena Yeung-Levy, and Xide Xia. Apollo: An Exploration of Video Understanding in Large Multimodal Models.

A Appendix

A.1 LLM Usage Disclosure

We used OpenAI ChatGPT only for minor language edits (grammar, spelling, clarity). It did not contribute to the ideas, methods, experiments, analyses, or technical content, and all edits were reviewed and verified by the authors.

A.2 More cases of LongVideoAgent

Table 6: The case of LongVideoAgent demonstrates the effective collaboration among multiple agents, specifically highlighting the Master Agent's invocation of the Vision Agent, resulting in accurate outcomes.

Question & Choices
Question: What color shirt was Sheldon wearing when he was knocking on Penny 's door ? Choices: a0: Blue a1: Black **a2: Yellow** ✓ a3: Red a4: Purple
LongVideoAgent Execution
Master Agent: <think> ... To determine the color of Sheldon's shirt, visual information from this specific clip is required. I lack that knowledge, so I need to search for visual details of the scene. </think> <visual_query> visual description of scene in *The Big Bang Theory* S05E06 segment 02 clip 15 where Sheldon is with a man </visual_query> ⇒ Vision Agent - **Vision Agent**: *Visual description.* He's wearing a yellow short-sleeve T-shirt with a repeating multicolored (red/green/yellow) Rubik's-cube print, layered over a light gray long-sleeve shirt. - **Master Agent**: <think> The Vision Agent described Sheldon's shirt in The Big Bang Theory S05E06, segment 02, clip 15 as a yellow T-shirt with a Rubik's cube print. The question asks for the shirt color when Sheldon knocks on Penny's door, and option a2 (Yellow) is correct. Thus, the answer is a2. </think> <answer> a2 ✓ </answer> ⇒ User

Table 7: A case of LongVideoAgent showing the collaboration among multiple agents.

Question & Choices
Question: What was on the counter in front of Leonard when he was saying he had some friends who were not genius ? Choices: a0: A bowl of salad . a1: A pot roast . a2: A plate of steaks . **a3: Several bottles .** ✓ a4: A computer .
LongVideoAgent Execution
Master Agent: <think> The provided subtitles describe the dialogue but contain no visual details about the counter in front of Leonard during the key line ("some of my best friends aren't geniuses"). To determine what was on the counter, visual information from the clip is necessary, as the text alone is insufficient. </think> <visual_query> Big Bang Theory S03E06 seg02 clip03 visual description kitchen counter Leonard Penny friends conversation </visual_query> ⇒ Vision Agent - **Vision Agent**: *Visual description.* When Leonard delivers that line the kitchen/living-room counters and table are cluttered. Key visible items: - A cluster of brown beer bottles (several) on the coffee/table area in front of the couch. - Additional clear glass bottles and smaller glass containers mixed in with the beer bottles. - An aluminum soda/beer can (blue) on the counter near the bottles. - A plastic water bottle (clear) on the counter. - A yellow plastic bowl (being held/placed on the table). - A pizza box on the dining counter. - Snack/chip bags near the kitchen area. - Blue disposable plates and a red plastic cup on the table. - Various condiment/olive-oil-type bottles and small jars near the kitchen prep area. - Refrigerator with photos/magnets in the background. Overall it's a party/after-party spread: multiple empty/half bottles, a pizza box, snack bags, disposable plates/cups and a yellow serving bowl. - **Master Agent**: <think> The subtitles pinpoint the key line to the end of the clip, during the conversation in the kitchen area. The visual description confirms the setting is the cluttered kitchen counter with party remnants, including multiple beer bottles, clear glass bottles, a water bottle, a can, plates, bowls, pizza box, etc. The question asks what was on the counter in front of Leonard, and option a3 (Several bottles .) matches the description of several bottles. Thus, the answer is a3. </think> <answer> a3 ✓ </answer> ⇒ User

Table 8: The case of LONGVIDEOAGENT exemplifies the seamless coordination among multiple agents, specifically showcasing the Master Agent's effective integration with the Grounding Agent and Vision Agent, ultimately deriving accurate image-based information and reasoning to achieve the correct solution.

Question & Choices
Question: What is behind Penny when she asks how the potato clock works ? Choices: a0: Curtains a1: Window a2: Door **a3: Books** ✓ a4: Table
LONGVIDEOAGENT Execution
MASTER AGENT: <think> The provided subtitles from clip discuss Bernadette, Howard, Sheldon, and Penny waiting for someone (likely Professor Proton), but make no mention of a potato clock or Penny asking about it. There are significant discrepancies between the provided information and the question, and no visual details are given. **I should request grounding to relocate to the correct clip containing the potato clock scene with Penny**. </think> <grounding_query> What is behind Penny when she asks how the potato clock works in The Big Bang Theory? </grounding_query> ⇒ GROUNDING AGENT
GROUNDING AGENT: *New clip.* <clip_12>UNKNAME : I 'm sorry to hear about your troubles . Uh , thanks . <eos> Penny : But if you do n't mind me asking , uh , the potato clock , how does it work ? (...) <eos> UNKNAME : No , I get that .</clip_12>
MASTER AGENT: <think> The initial clip was irrelevant, and this is the correct relocated scene. However, to determine what's behind Penny (e.g., curtains, window, door, books, table), **I need visual details of the scene, which are not available in the text**. </think> <visual_query> What is behind Penny in The Big Bang Theory season 6 episode 22 scene where she asks Professor Proton about the potato clock? </visual_query> ⇒ VISION AGENT
VISION AGENT: *Visual description.* A bookshelf filled with books.
MASTER AGENT: <think> The search for visual details returns a description of a bookshelf filled with books behind her, which aligns with option a3 (Books). This is sufficient to answer without discrepancies or need for further actions. </think> <answer> a3 ✓ </answer> ⇒ USER