ILVR-Agent: Decomposing Instance-Level Reasoning in Long-Form Videos via Chain-of-Agent Thinking

Anonymous ACL submission

Abstract

Recently, agentic video reasoning methods have demonstrated significant potential by incentivizing tool-thinking capabilities through Reinforcement Learning (RL). However, existing agentic approaches struggle with Instance-level Long-form Video Reasoning (ILVR), which demands extensive cross-frame evidence aggregation, due to the scaling of reasoning chains and tool-thinking trajectories. To address these challenges, we introduce **ILVR-Agent**, a multi-agent framework powered by **Chain-of-Agent Thinking (CoAT)**, which modularizes complex reasoning chains and facilitates modular tool-thinking with specialized agents. Specifically, we systematically develop ILVR-Agent across three perspectives: dataset, method, and benchmark. First, we design an end-to-end multi-agent engine to meticulously curate **ILVR-Instruction**, a large-scale, high-quality instruction dataset tailored for ILVR. Additionally, the ILVR-Agent method orchestrates a collaborative reasoning pipeline by modularizing intricate reasoning chains into: retrieval, planning and execution, subsequently invoking specialized agents with task-specific tool-thinking. Furthermore, to enhance tool-thinking efficiency, we propose PA-GRPO, an RL framework that incorporates process-aware supervision via LLM-as-Judge, explicitly validating each tool invocation throughout the reasoning trajectory. Finally, we establish **ILVR-Bench**, a comprehensive benchmark for evaluating the ILVR capabilities of Video-LLMs. Extensive experiments and analyses demonstrate that our ILVR-Agent method achieves promising performance on both instance-level and general long-form video reasoning.

1 Introduction

Instance-level Long-form Video Reasoning (ILVR) refers to instance-centric reasoning within long-form video content, which is foundational to autonomous robotic navigation (Pore et al., 2023), human-computer interaction (Kosch et al., 2023), and autonomous driving (Chib and Singh, 2023). However, ILVR remains a challenging task, as it requires not only fine-grained spatio-temporal perception but also the capability to localize and aggregate cross-frame evidence for complex instance interactions.

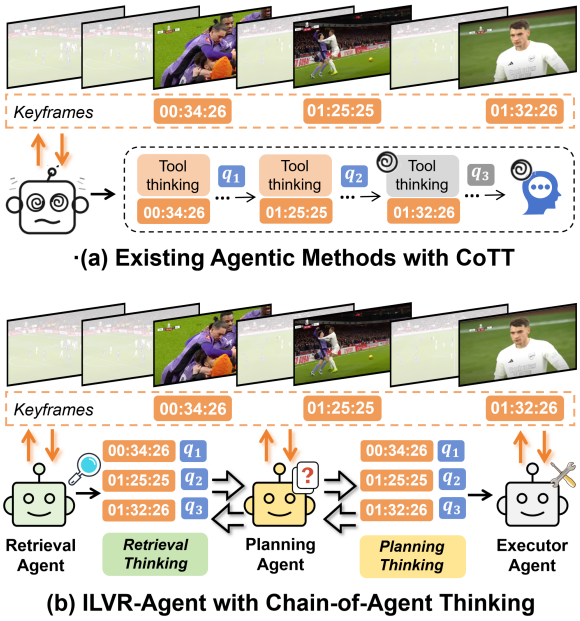


Figure 1: Comparison between existing agentic video reasoning methods and ILVR-Agent. (a) Extremely long Chain-of-Tool-Thought (CoTT) in existing agentic methods undermines long-range logical consistency and leads to reasoning drift. (b) ILVR-Agent facilitates Chain-of-Agent Thinking (CoAT) by modularizing complex reasoning chains and enabling specialized tool-thinking with task-specific agents.

To address these challenges, recent advancements in Video-LLMs (Wei and Chen, 2025; Liu et al., 2025) and Reinforcement Learning (RL) strategies (Shao et al., 2024) have substantially expanded models' context windows and enhanced their reasoning capabilities, respectively. However, despite extended context capacities, comprehending the information density of hour-level videos and

facilitating long-chain reasoning within a single inference pass remains intractable. Motivated by the human cognitive process for long-form video reasoning, recent approaches (Fan et al., 2024; Kugo et al., 2025) introduce agent-based frameworks that integrate the reasoning and decision-making capabilities of Large Language Models (LLMs) with the visual perception abilities of Vision-Language Models (VLMs), thereby enabling interactive and iterative video reasoning through multi-turn inference. However, such video agent systems primarily rely on predefined reasoning chains, limiting their adaptability to open-ended reasoning tasks in real-world scenarios.

To explore adaptive reasoning in video agents, recent works (Dong et al., 2025; Fei et al., 2024) introduce Agentic RL, endowing agents with **tool-thinking** capabilities. Such agentic methods facilitate adaptive tool invocation based on interactive environmental feedback, iteratively refining the evidence aggregation process. E.g., VideoExplorer (Yuan et al., 2025) designs a training pipeline comprising supervised trajectory initialization with trajectory-level preference optimization, thereby reinforcing step-by-step tool invocations based on real-time feedback. However, existing agentic approaches struggle with ILVR due to the scaling of reasoning chains and tool-thinking trajectories. As shown in Fig. 1(a), when video duration scales to hour-level, the elongated reasoning chains involving multi-turn interactions inevitably accumulate overwhelming contextual noise. This interference disrupts the ongoing tool-thinking process, ultimately undermining long-range logical consistency and inducing reasoning drift.

To address these challenges, we introduce **ILVR-Agent**, a multi-agent framework powered by **Chain-of-Agent Thinking (CoAT)**, as depicted in Fig. 1(b). Unlike existing agentic video reasoning approaches that rely on a single agent to centralize diverse tool-thinking paradigms, ILVR-Agent modularizes complex reasoning chains and facilitates specialized tool-thinking through task-specific agents: a *Planning Agent* for task decomposition and evidence sufficiency assessment, a *Retrieval Agent* for temporal grounding and information extraction, and an *Executor Agent* for sub-task resolution. Specifically, we systematically develop the ILVR-Agent ecosystem across three perspectives: dataset, method, and benchmark.

Dataset. To mitigate the scarcity of high-quality ILVR data, we design an end-to-end multi-agent data curation pipeline, by synergizing four specialized modules: *Multi-granularity Captioner*, *QA Generator*, *Instance Annotator*, and *Multimodal Reviewer*. We curate **ILVR-Instruction**, the first large-scale instruction-following dataset dedicated to instance-level reasoning, which comprises 76K samples for Supervised Fine-Tuning (SFT) and 12K samples for Reinforcement Learning (RL).

Method. ILVR-Agent orchestrates a collaborative reasoning pipeline through CoAT, which structures reasoning as an iterative agent coordination loop. The Planning Agent dynamically decomposes queries and dispatches sub-tasks to specialized agents, while the Retrieval Agent performs temporal grounding via retrieval tool-thinking and the Executor Agent handles inference-oriented sub-tasks via executor tool-thinking. Furthermore, to enhance tool-thinking efficiency, we propose **PA-GRPO** (Process-Aware GRPO), an RL framework that incorporates process-level supervision via *LLM-as-Judge*, explicitly validating each tool invocation throughout the reasoning trajectory.

Benchmark. To facilitate comprehensive evaluation of ILVR capabilities, we introduce **ILVR-Bench**, a benchmark encompassing two complementary evaluation protocols: *ILVR-Bench-QA* for multiple-choice QA, and *ILVR-Bench-QAR* that additionally requires explicit rationale generation. The benchmark comprises 1,000 high-quality samples spanning 10 task types across four temporal granularities (1-, 10-, 30-, and 100-minute).

Our contributions are summarized as follows:

- We design an end-to-end multi-agent data curation pipeline and construct **ILVR-Instruction**, the first large-scale instruction-tuning dataset dedicated to ILVR.
- We propose **ILVR-Agent**, a multi-agent framework that introduces **Chain-of-Agent Thinking (CoAT)**, effectively modularizing complex reasoning chains through specialized Planning, Retrieval, and Executor agents.
- We introduce **PA-GRPO**, a process-aware reinforcement learning framework leveraging LLM-as-Judge to provide step-wise supervision, significantly improving the efficiency of tool-thinking in long-range video reasoning.
- We establish **ILVR-Bench**, featuring two evaluation protocols across 10 task types that rigorously assess capabilities spanning from fine-grained perception to temporal reasoning.

2 Related Work

2.1 Agent-based Video Reasoning

Inspired by the human cognitive process for long-form video reasoning, recent methods (Fan et al., 2025; Yang et al., 2025c; Shi et al., 2025) introduce agent-based frameworks. Specifically, these frameworks decouple the reasoning task into a cognitive-perceptual pipeline: where LLMs function as the "brain" for high-level planning and decision-making, whereas VLMs serve as the "eyes" for visual perception. Early explorations in this domain, such as ViperGPT (Surís et al., 2023), pioneered modular orchestration by leveraging LLMs as high-level planners to synthesize executable programs that invoke specialized vision modules. Building upon this modularity, VideoAgent (Wang et al., 2024) further refined the iterative reasoning workflow. Unlike traditional single-pass architectures, VideoAgent emphasizes a human-like evidence compilation process, employing an LLM-based controller to dynamically localize task-relevant temporal segments and progressively aggregate information to resolve complex, long-horizon queries. However, such video agent systems primarily rely on predefined reasoning chains, which limits the adaptability to open-ended reasoning tasks in real-world scenarios.

2.2 Agentic Video Reasoning

To explore adaptive reasoning in video agents, recent works (Li et al., 2025; Xu et al., 2025; Rutherford et al., 2024; Zhang et al., 2023) introduce Agentic RL, a post-training strategy that reinforces iterative tool-thinking and the refinement of reasoning trajectories based on environmental feedback. The core motivation is to enable agents to autonomously learn optimal tool-invocation strategies rather than relying solely on zero-shot prompting. E.g., VideoExplorer (Yuan et al., 2025) advocates for the "thinking with video" principle, which seamlessly intertwines planning, temporal grounding, and scalable perception into a coherent reasoning process. It achieves task-oriented video understanding by iteratively formulating sub-questions and locating relevant moments within the video. Furthermore, Ego-R1 (Tian et al., 2025) establishes a structured Chain-of-Tool-Thought (CoTT) process by leveraging a DeepSeek-R1-style (Guo et al., 2025) training paradigm (i.e., cold-start and GRPO phases), enabling step-by-step and interactive tool invocations based on real-time feedback. Despite these advances, existing agentic approaches struggle with Instance-level Long-form Video Reasoning (ILVR), which demands extensive cross-frame evidence aggregation. This difficulty stems from the complexity bottleneck when scaling reasoning chains and tool-thinking trajectories within a single agent.

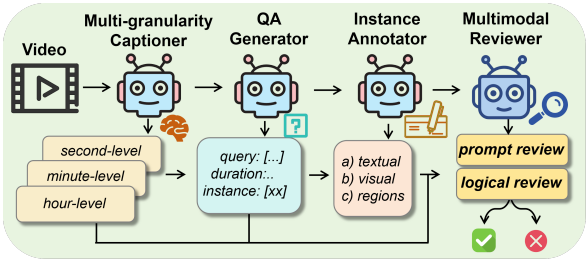


Figure 2: A multi-agent data engine for the construction of ILVR Dataset.

3 Method

3.1 Preliminary

ILVR Task Formulation. ILVR entails the capability to parse and respond to instance-centric queries involving multiple target instances and their complex interactions. Specifically, given a query such as *'How do* <instance$_1$> *and* <instance$_2$> *interact with each other throughout the video?"*, where <instance$_i$> denotes an instance prompt, the model must reason over intricate instance relationships and long-horizon temporal dependencies. To this end, we present a unified formulation for ILVR. Formally, let V denote the input video and $\mathbf{P} = \{P_1, P_2, \ldots, P_n\}$ be a set of n user-specified instance prompts. Each prompt P_i can be instantiated through one or more modalities: textual descriptions, visual exemplars, or spatial regions (i.e., bounding boxes). The learning objective is to maximize the conditional log-likelihood of generating the ground-truth response y, given the video V, instance prompts $\mathbf{P}$, and query q:

$$\mathcal{L} = \sum_{(V,\mathbf{P},q,y)} \log P(y \mid V, P_1, \ldots, P_n, q). \quad (1)$$

3.2 Multi-Agent Engine for ILVR Dataset

To mitigate the scarcity of high-quality ILVR data, we design an end-to-end multi-agent data curation pipeline to construct ILVR-Instruction, the first large-scale instruction-following dataset tailored for instance-level reasoning. The resulting dataset comprises 76K samples for supervised fine-tuning (SFT) and 12K samples for reinforcement learning (RL). As illustrated in Fig. 2, our pipeline is

structured into four modular components, each orchestrated by specialized expert models.

Multi-granularity Captioner. To capture the complex dynamics of long-form videos, we employ *Qwen3-VL-235B* (Bai et al., 2025a) to establish a multi-granularity captioning framework spanning three temporal scales (second-, minute-, and hour-levels), enabling the synthesis of instruction data with diverse temporal distributions.

QA Generator. Building upon multi-granularity captions, the QA Generator (*GPT-5*) synthesizes instruction-following pairs across four temporal horizons (1-, 10-, 30-, and 100-minute). Reflecting the inherent complexities of ILVR, the generated QA data is structured along two dimensions: *Perception* and *Reasoning*, further decomposed into four fine-grained sub-types: *Object*, *Action*, *Spatial*, and *Temporal*.

Instance Annotator. To enable precise instance-level prompting, the Instance Annotator (*Qwen3-VL-235B*) synthesizes multimodal prompts for each target instance, comprising: (i) textual descriptions capturing semantic details; (ii) visual exemplars providing appearance cues; and (iii) spatial regions establishing spatial grounding. Such multimodal prompts effectively mitigate semantic ambiguity inherent in purely linguistic references.

Multimodal Reviewer. We implement a two-stage verification process to ensure data integrity. First, *Gemini-2.5-Pro* (Comanici et al., 2025) audits the spatio-temporal alignment and semantic accuracy of instance prompts. Subsequently, *DeepSeek-V3* (Liu et al., 2024) performs logical verification by generating step-by-step reasoning trajectories, where visual evidence is encapsulated within `<caption>` tags and analytical reasoning within `<think>` tags.

3.3 ILVR-Agent with PA-GRPO

Unlike existing agentic video reasoning approaches (Tian et al., 2025; Yuan et al., 2025) that rely on a single agent to centralize diverse tool-thinking for different tasks, we introduce ILVR-Agent, a multi-agent framework powered by chain-of-agent thinking, modularizing complex reasoning chains and facilitates modular tool-thinking with specialized agents. We first describe the overall architecture and the Chain-of-Agent Thinking paradigm (§3.3.1), followed by the Retrieval Agent (§3.3.3), the Planning Agent (§3.3.2), the Executor Agent (§3.3.4) and the PA-GRPO (§3.3.5).

3.3.1 Overall Architecture

As depicted in Fig. 3, ILVR-Agent comprises three specialized agents: a *Planning Agent* $\mathcal{A}_{\text{plan}}$ for task decomposition and evidence sufficiency assessment, a *Retrieval Agent* $\mathcal{A}_{\text{ret}}$ for temporal grounding and information extraction, and an *Executor Agent* $\mathcal{A}_{\text{exe}}$ for sub-task resolution. This modular design decouples the monolithic reasoning chain prevalent in single-agent systems into discrete, manageable components.

Chain-of-Agent Thinking (CoAT). Central to our framework is the Chain-of-Agent Thinking paradigm, which addresses contextual noise accumulation in long-form video reasoning. Given a query q and instance prompts $\mathbf{P}$ over video V, CoAT structures reasoning as an iterative agent coordination loop:

$$s_t = \mathcal{A}_{\text{plan}}(q, \mathbf{P}, \mathcal{H}_{<t}), \tag{2}$$

$$r_t = \mathcal{A}_*(s_t),\ \mathcal{A}_* \in \{\mathcal{A}_{\text{ret}}, \mathcal{A}_{\text{exe}}\}, \tag{3}$$

$$\mathcal{H}_t = \mathcal{H}_{<t} \cup \{(s_t, r_t)\}, \tag{4}$$

where s_t denotes the sub-task dispatched at step t, r_t is the returned result from specialized agent $\mathcal{A}_*$, and $\mathcal{H}_t$ denotes the interaction history. The planner $\mathcal{A}_{\text{plan}}$ dynamically orchestrates sub-task decomposition conditioned on prior feedback until convergence:

$$y = \mathcal{A}_{\text{plan}}(q, \mathbf{P}, \mathcal{H}_T) \text{ s.t. } \textsc{Sufficient} = \texttt{true}. \tag{5}$$

Unlike single-agent approaches with ever-growing context, CoAT isolates task-specific reasoning within specialized agents and enables adaptive re-planning, thereby preserving long-range logical consistency.

3.3.2 Planning Agent

The Planning Agent $\mathcal{A}_{\text{plan}}$ serves as the central coordinator, orchestrating multi-agent collaboration through two core functions: *hierarchical task decomposition* and *evidence sufficiency assessment* using planning tool-thinking.

Task Decomposition. Given query q, instance prompts $\mathbf{P}$, and interaction history $\mathcal{H}_{<t}$, $\mathcal{A}_{\text{plan}}$ decomposes the problem into atomic sub-tasks:

$$s_t = \textsc{Decompose}(q, \mathbf{P}, \mathcal{H}_{<t}). \tag{6}$$

Each sub-task s_t is classified and dispatched accordingly:

$$\textsc{Dispatch}(s_t) \rightarrow \begin{cases} \mathcal{A}_{\text{ret}}, & \text{if } s_t \in \mathcal{S}_{\text{ret}} \\ \mathcal{A}_{\text{exe}}, & \text{if } s_t \in \mathcal{S}_{\text{exe}} \end{cases}, \tag{7}$$

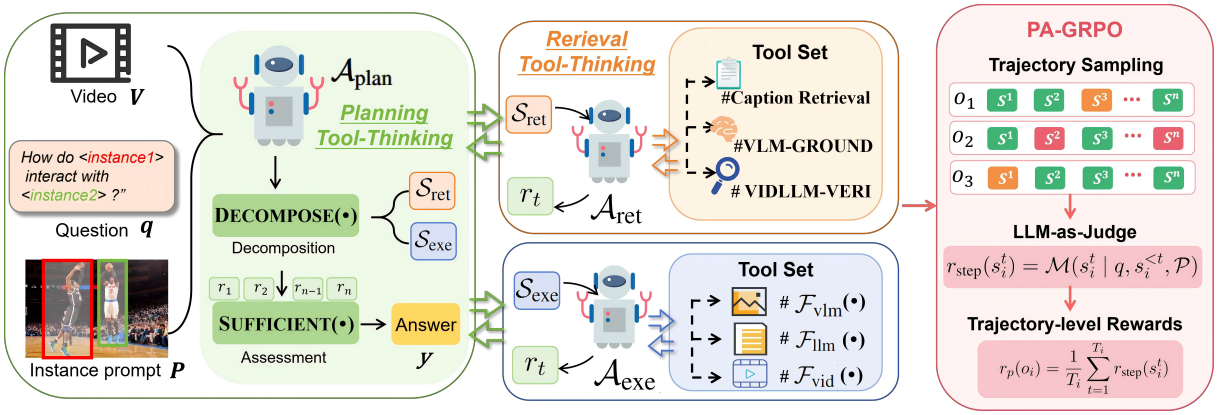


Figure 3: The overall architecture of ILVR-Agent, containing three specialized agents: *Planning Agent* $\mathcal{A}_{\text{plan}}$, *Retrieval Agent* $\mathcal{A}_{\text{ret}}$ and *Executor Agent* $\mathcal{A}_{\text{exe}}$. The framework performs chain-of-agent thinking, modularizing complex reasoning chains and facilitates modular tool-thinking with specialized agents.

where $\mathcal{S}_{\text{ret}}$ and $\mathcal{S}_{\text{exe}}$ denote retrieval and executable task spaces, respectively.

Evidence Sufficiency Assessment. Upon receiving result r_t from specialized agents, $\mathcal{A}_{\text{plan}}$ evaluates whether $\mathcal{H}_t$ provides sufficient evidence:

$$\textsc{Sufficient}(\mathcal{H}_t, q) \rightarrow \{\texttt{true}, \texttt{false}\}. \quad (8)$$

If satisfied, the agent synthesizes the final response via $y = \textsc{Aggregate}(\mathcal{H}_T, q)$; otherwise, it initiates another decomposition-dispatch cycle. This iterative mechanism enables adaptive reasoning depth based on query complexity.

3.3.3 Retrieval Agent

The Retrieval Agent $\mathcal{A}_{\text{ret}}$ localizes relevant temporal segments and extracts associated information from long-form videos via retrieval tool-thinking. This mechanism enables $\mathcal{A}_{\text{ret}}$ to iteratively invoke specialized tools, progressively narrowing down to compact yet informative video intervals. To interpret multimodal instance prompts comprehensively, we equip $\mathcal{A}_{\text{ret}}$ with three complementary tools:

Caption-based Retrieval. We construct a hierarchical caption index at two temporal granularities: fine-grained ($\mathcal{C}_1$, 1-min) and coarse-grained ($\mathcal{C}_{10}$, 10-min). Given textual cues in s_t, we employ an LLM to perform semantic retrieval:

$$\mathcal{T}_{\text{cand}} = \text{LLM-}\textsc{Ret}(s_t, \mathcal{C}_1 \cup \mathcal{C}_{10}), \quad (9)$$

where $\mathcal{T}_{\text{cand}}$ contains candidate temporal segments. This coarse-to-fine strategy enables efficient navigation of hour-level content.

Visual Grounding. For visual cues in s_t (e.g.,, visual exemplars or bounding boxes), $\mathcal{A}_{\text{ret}}$ leverages a VLM to ground these references:

$$c_v = \text{VLM-}\textsc{Ground}(s_t), \quad (10)$$

where c_v denotes the detected visual instance clues.

Video-based Verification. To refine $\mathcal{T}_{\text{cand}}$ and mitigate false positives, we introduce a verification step using a Video-LLM:

$$\mathcal{T}', \mathcal{K} = \textsc{VidLLM-Veri}(\mathcal{T}_{\text{cand}}, s_t, c_v), \quad (11)$$

which outputs refined temporal boundaries $\mathcal{T}'$ and verified knowledge $\mathcal{K}$. The final result is then formulated as $r_t = (\mathcal{T}', \mathcal{K})$.

3.3.4 Executor Agent

The Executor Agent $\mathcal{A}_{\text{exe}}$ handles inference-oriented sub-tasks via executor tool-thinking, dynamically selecting appropriate reasoning tools. Given $s_t \in \mathcal{S}_{\text{exe}}$, $\mathcal{A}_{\text{exe}}$ routes it based on input modality:

$$r_t = \mathcal{F}_*(s_t, \mathcal{H}_{<t}),\ \mathcal{F}_* \in \{\mathcal{F}_{\text{llm}}, \mathcal{F}_{\text{vlm}}, \mathcal{F}_{\text{vid}}\}, \quad (12)$$

where $\mathcal{F}_{\text{llm}}$ handles text-based reasoning, $\mathcal{F}_{\text{vlm}}$ addresses image-level understanding (e.g.,, object recognition, spatial reasoning), and $\mathcal{F}_{\text{vid}}$ performs temporal reasoning over video segments. The result r_t is subsequently aggregated into $\mathcal{H}_t$ for evidence accumulation.

3.3.5 PA-GRPO

While GRPO (Guo et al., 2025) has demonstrated effectiveness in text-based reasoning, it fundamentally relies on outcome-level rewards. In tool-thinking scenarios, where reasoning trajectories encompass multiple intermediate tool invocations, outcome-level supervision fails to capture the validity of individual reasoning steps. To address this limitation, we propose **Process-Aware Group Relative Policy Optimization (PA-GRPO)**, an RL framework that incorporates process-level supervision to explicitly validate tool invocations throughout the reasoning trajectory.

The key insight of PA-GRPO is to leverage an *LLM-as-Judge* paradigm to evaluate individual reasoning steps within tool-thinking trajectories, subsequently aggregating these step-wise assessments into trajectory-level process rewards. In contrast to training a dedicated Process Reward Model (Khalifa et al., 2025) requiring costly step-level annotations, our approach harnesses a capable LLM with structured evaluation prompts to assess whether each tool invocation is *well-formatted* and *contextually appropriate*.

Step-wise Process Reward. Given query q, we sample a group of G reasoning trajectories $\{o_i\}_{i=1}^{G}$ from policy π_θ. Each trajectory o_i comprises T_i reasoning steps: $o_i = (s_i^1, \ldots, s_i^{T_i})$, where each s_i^t encapsulates a tool invocation with its accompanying rationale. For each step, we employ an LLM judge $\mathcal{M}$ guided by evaluation rubric $\mathcal{P}$:

$$r_{\text{step}}(s_i^t) = \mathcal{M}(s_i^t \mid q, s_i^{<t}, \mathcal{P}), \tag{13}$$

where $s_i^{<t}$ denotes preceding context. The rubric $\mathcal{P}$ directs the judge to score: (1) *tool invocation appropriateness*, and (2) *format compliance*. The trajectory-level process reward is obtained via mean pooling:

$$r_p(o_i) = \frac{1}{T_i} \sum_{t=1}^{T_i} r_{\text{step}}(s_i^t). \tag{14}$$

Reward Formulation. The composite reward R_i integrates outcome-level accuracy and process-level quality:

$$R_i = r_a(o_i) + \lambda \cdot r_p(o_i), \tag{15}$$

where $r_a(o_i) \in \{0, 1\}$ denotes the binary correctness indicator, and λ is a balancing coefficient.

Optimization Objective. Following the group relative paradigm, we compute advantages via intra-group normalization:

$$A_i = \frac{R_i - \mu(\{R_j\}_{j=1}^{G})}{\sigma(\{R_j\}_{j=1}^{G})}, \tag{16}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote mean and standard deviation, respectively. The policy optimization objective is:

$$\mathcal{J}(\theta) = \mathbb{E}_{q,\{o_i\}} \left[\frac{1}{G} \sum_{i=1}^{G} \mathcal{L}_i(\theta) \right], \tag{17}$$

$$\mathcal{L}_i(\theta) = \min \left(\rho_i A_i,\ \text{clip}(\rho_i, 1 \pm \epsilon) A_i \right) - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \tag{18}$$

where $\rho_i = \pi_\theta(o_i \mid q) / \pi_{\theta_{\text{old}}}(o_i \mid q)$ is the importance ratio, ϵ controls the clipping range, and β governs the KL penalty against reference policy π_{ref}.

3.4 ILVR-Bench

To comprehensively evaluate model capabilities in ILVR, we introduce ILVR-Bench, a benchmark encompassing two complementary evaluation protocols: (1) **ILVR-Bench-QA**, a multiple-choice question-answering task, and (2) **ILVR-Bench-QAR**, an extension that requires models to provide explicit rationales alongside predictions. Visual examples and detailed statistics are provided in the Appendix.

3.4.1 ILVR-Bench-QA

We curate a dataset of 42 hour-level videos spanning four domains: *Film & Television*, *Life Record*, *Knowledge*, and *Documentary*, sourced and re-curated from established public benchmarks (Song et al., 2024; Yang et al., 2025a,b; He et al., 2024). Through a multi-agent pre-annotation pipeline with rigorous human verification, we construct 1,000 high-quality multiple-choice questions covering 10 task types, including object recognition, spatial perception, and temporal reasoning. To enable a comprehensive assessment of long-horizon video reasoning, the temporal spans of QA pairs are stratified into four granularity levels (1-, 10-, 30-, and 100-minute intervals), facilitating systematic assessment across varying temporal horizons.

3.4.2 ILVR-Bench-QAR

To further probe the reasoning capabilities of agentic frameworks, ILVR-Bench-QAR mandates structured rationale generation alongside answer prediction. Each sample in ILVR-Bench-QAR comprises an additional `<caption>` module providing task-relevant visual evidence and `<think>` module articulating the step-by-step reasoning chain. The incorporation of reasoning process mitigates shortcut-based guessing and enables fine-grained evaluation of reasoning interpretability in ILVR systems.

4 Experiments

4.1 Experiment Setup

Training pipeline. Inspired by the DeepSeek-R1 RL paradigm, we train ILVR-Agent (comprising Planning and Retrieval Agents) via a two-stage strategy. In the first stage, we perform supervised

Method	Size	GLVR		ILVR	
		MME	LVB	QA	QAR
MLLMs					
Qwen2.5-VL (Bai et al., 2025b)	72B	62.2	47.3	42.5	36.1
GPT-4o (Hurst et al., 2024)	-	65.3	48.9	45.1	38.5
Gemini-1.5-Pro (Team et al., 2024)	-	**67.4**	33.1	46.4	39.1
Video Agent					
VideoAgent (Wang et al., 2024)	7B	50.8	36.1	34.7	25.2
LLaVA-OV+T^* (Ye et al., 2025)	7B	46.3	37.2	35.1	26.5
Agentic Methods					
VideoExplorer (Yuan et al., 2025)	3B	62.1	43.1	39.5	32.4
Ego-R1 (Tian et al., 2025)	3B	64.9	44.3	42.6	34.2
ILVR-Agent	3B	**67.4**	**49.5**	50.6	44.2
ILVR-Agent*	3B	**67.4**	**49.5**	**56.2**	**49.1**

Table 1: Performance comparison on GLVR (MME: VideoMME-l, LVB: LVBench) and ILVR (QA: ILVR-Bench-QA, QAR: ILVR-Bench-QAR) benchmarks. To ensure a fair comparison with existing baselines that exclusively utilize textual prompts, **ILVR-Agent** is evaluated using text-based instance prompts, whereas **ILVR-Agent*** incorporates multimodal prompts.

fine-tuning (SFT) on a pretrained language model using our synthetic tool-thinking dataset. In the second stage, we apply PA-GRPO to the SFT-initialized models to further enhance the multi-turn tool-thinking capabilities.

Benchmarks. We evaluate the performance of the ILVR-Agent on two General Long-form Video Reasoning (GLVR) benchmarks: *Video-MME-l* (Fu et al., 2025) (long-version without subtitles) and *LVBench* (Wang et al., 2025). To further evaluate the capability of ILVR, we introduce ILVR-Bench-QA and ILVR-Bench-QAR. Additional details about experiment setup are provided in Appendix. All results are reported as the mean of 3 independent runs with different random seeds.

4.2 Main Results

Results on GLVR. As shown in Table 1, ILVR-Agent achieves 67.4% on VideoMME-l and 49.5% on LVBench, establishing SOTA performance on General Long-form Video Reasoning (GLVR). Compared to Video Agent methods, ILVR-Agent outperforms VideoAgent by +16.6% (32.7% relative improvement) on VideoMME-l and +13.4% (37.1%) on LVBench. Against Agentic Methods with comparable model capacity, ILVR-Agent surpasses the strongest baseline Ego-R1 by +2.5% and +5.2% on the two benchmarks, respectively. Notably, with only 3B parameters, ILVR-Agent matches Gemini-1.5-Pro on VideoMME-l while exceeding it by +16.4% on LVBench. We attribute these gains to the CoAT paradigm, which delegates sub-tasks to specialized agents equipped with task-specific tool-thinking capabilities. This modular design enhances both tool invocation accuracy and reasoning efficiency, yielding more coherent reasoning chains than single-agent counterparts that suffer from cross-task interference.

Strategy	VideoMME-l	ILVR-QAR
CoTT	64.9	34.2
CoAT(Ours)	**66.5**	**45.8**

Table 2: Ablation study of Chain-of-Agent Thinking (CoAT).

Results on ILVR. The advantages of ILVR-Agent become more salient on instance-level reasoning tasks, which demand longer reasoning chains and finer-grained temporal grounding. On ILVR-Bench-QA, ILVR-Agent achieves 50.6%, surpassing VideoAgent and Ego-R1 by +15.9% and +8.0%, respectively. The performance gap widens considerably on ILVR-Bench-QAR, where explicit rationale generation is required: ILVR-Agent attains 44.2%, outperforming VideoAgent by +19.0% (75.4% relative gain) and Ego-R1 by +10.0%. The amplified improvements on QAR over QA (75.4% vs. 45.8% against VideoAgent; 29.2% vs. 18.8% against Ego-R1) underscore that ILVR-Agent produces substantially more faithful reasoning traces. We ascribe this to the task-based decoupling mechanism in CoAT, which effectively mitigates contextual noise accumulation, a critical bottleneck when reasoning chains scale with instance-level complexity. By compartmentalizing retrieval, planning, and execution within dedicated agents, ILVR-Agent preserves logical consistency across extended multi-turn interactions, whereas monolithic single-agent architectures exhibit pronounced reasoning drift under comparable settings. Additionally, ILVR-Agent*, which leverages multimodal prompts, demonstrates substantial improvements over its text-only counterpart (e.g., 56.2 vs. 50.6 on QA), highlighting the effectiveness of multimodal prompt understanding.

Effectiveness of CoAT. Table 2 compares Chain-

RL Strategy	VideoMME-l	ILVR-QAR
GRPO	66.5	45.8
PA-GRPO (Ours)	**67.4**	**49.1**

Table 3: Ablation study of reinforcement learning (RL) strategies.

Method	Video-MME-l		ILVR-QAR	
	Acc. ↑	# Call ↓	Acc. ↑	# Call ↓
Ego-R1 (Tian et al., 2025)	64.9	7.4	34.2	10.2
ILVR-Agent	**67.4**	**6.4**	**49.1**	**8.6**

Table 4: Efficiency and accuracy comparison. ILVR-Agent achieves higher accuracy with fewer tool calls across benchmarks.

of-Agent Thinking (CoAT) with the single-agent Chain-of-Tool-Thought (CoTT) (Tian et al., 2025). While both achieve comparable performance on VideoMME-l (66.5% vs.64.9%), CoAT outperforms CoTT by +11.6% on ILVR-QAR. This disparity stems from the inherent complexity of ILVR tasks, which require extensive cross-frame evidence aggregation and prolonged reasoning trajectories. CoTT suffers from contextual noise accumulation within a single agent, whereas CoAT mitigates this bottleneck through task-based decoupling and specialized agent collaboration. These results validate that modularizing tool-thinking across expert agents yields superior reasoning consistency for long-chain video understanding.

Effectiveness of PA-GRPO. Table 3 presents ablations comparing PA-GRPO against vanilla GRPO. PA-GRPO achieves consistent improvements: +1.8% on Video-MME-l and +4.0% on ILVR-Bench-QAR. The amplified gains on QAR can be attributed to its longer reasoning trajectories with more frequent tool invocations, where outcome-level rewards in vanilla GRPO provide sparse supervision that fails to discriminate between valid and spurious intermediate steps. In contrast, PA-GRPO introduces step-wise process rewards that explicitly evaluate tool invocation appropriateness and format compliance, enabling more precise credit assignment for intermediate actions. The substantial improvement gap (8.9% vs.2.7%) empirically validates that process-level supervision becomes increasingly critical as reasoning complexity scales. Notably, by leveraging an LLM-as-Judge, PA-GRPO achieves these gains without the annotation overhead typically associated with training dedicated process reward models.

Statistics of Tool Calling Counts. Table 4 compares reasoning efficiency between ILVR-Agent and Ego-R1. ILVR-Agent consistently achieves higher accuracy with fewer tool calls: on Video-MME, +2.5% with 13.5% fewer calls (7.4 → 6.4); on ILVR-QAR, +14.9% with 15.7% fewer calls (10.2 → 8.6). This dual improvement arises from fundamental architectural differences. Ego-R1's monolithic CoTT paradigm centralizes all operations within a single agent, incurring redundant tool calls to compensate for accumulated contextual noise as reasoning complexity scales. Conversely, ILVR-Agent's CoAT paradigm decomposes reasoning into task-specific workflows, where each specialized agent maintains a focused context window and enables precise tool selection. The amplified gains on ILVR-QAR (43.6% vs.3.9% relative improvement) demonstrate that modular multi-agent architectures scale more favorably than single-agent counterparts for complex reasoning chains.

5 Conclusion

In this paper, we systematically develop ILVR-Agent across: dataset, method, and benchmark. First, we design an end-to-end multi-agent engine to curate ILVR-Instruction, a large-scale, high-quality instruction dataset tailored for ILVR. Additionally, we introduce ILVR-Agent, a multi-agent framework with Chain-of-Agent Thinking (CoAT) by modularizing intricate reasoning chains, subsequently invoking specialized agents with task-specific tool-thinking. Furthermore, we propose PA-GRPO that incorporates process-aware supervision via LLM-as-Judge, explicitly validating tool invocations throughout the reasoning trajectory. Finally, we establish ILVR-Bench, a comprehensive benchmark for evaluating ILVR. Experiments demonstrate that ILVR-Agent achieves promising performance on both GLVR and ILVR.

6 Limitations

While our multi-agent pipeline ensures high-quality annotations, the source videos predominantly represent mainstream categories, which may not fully capture the tail distributions of complex visual scenes. The robustness of ILVR-Agent against significant domain shifts requires more comprehensive empirical investigation. Addressing this gap through large-scale, open-domain data scaling is a priority for our subsequent research.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Pranav Singh Chib and Pravendra Singh. 2023. Recent advancements in end-to-end autonomous driving using deep learning: A survey. *IEEE Transactions on Intelligent Vehicles*, 9(1):103–118.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, and 1 others. 2025. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*.

Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. 2025. Embodied videoagent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6342–6352.

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. In *European Conference on Computer Vision*, pages 75–92. Springer.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yichen He, Yuan Lin, Jianchao Wu, Hanchong Zhang, Yuchen Zhang, and Ruicheng Le. 2024. Storyteller: Improving long video description through global audio-visual character identification. *arXiv preprint arXiv:2411.07076*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. 2025. Process reward models that think. *arXiv preprint arXiv:2504.16828*.

Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W Woźniak. 2023. A survey on measuring cognitive workload in human-computer interaction. *ACM Computing Surveys*, 55(13s):1–39.

Noriyuki Kugo, Xiang Li, Zixin Li, Ashish Gupta, Arpandeep Khatua, Nidhish Jain, Chaitanya Patel, Yuta Kyuragi, Yasunori Ishii, Masamoto Tanabiki, and 1 others. 2025. Videomultiagents: A multi-agent framework for video question answering. *arXiv preprint arXiv:2504.20091*.

Weizhen Li, Jianbo Lin, Zhuosong Jiang, Jingyi Cao, Xinpeng Liu, Jiayu Zhang, Zhenqiang Huang, Qianben Chen, Weichen Sun, Qiexiang Wang, and 1 others. 2025. Chain-of-agents: End-to-end agent foundation models via multi-agent distillation and agentic rl. *arXiv preprint arXiv:2508.13167*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ruyang Liu, Shangkun Sun, Haoran Tang, Wei Gao, and Ge Li. 2025. Flow4agent: Long-form video understanding via motion prior from optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23817–23827.

Ameya Pore, Zhen Li, Diego Dall'Alba, Albert Hernansanz, Elena De Momi, Arianna Menciassi, Alicia Casals Gelpi, Jenny Dankelman, Paolo Fiorini, and Emmanuel Vander Poorten. 2023. Autonomous navigation for robot-assisted intraluminal and endovascular procedures: A systematic review. *IEEE Transactions on Robotics*, 39(4):2529–2548.

Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Garðar Ingvarsson Juto, Timon Willi, Ravi Hammond, Akbir Khan, Christian Schroeder de Witt, and 1 others. 2024. Jaxmarl: Multi-agent rl environments and algorithms in jax. *Advances in Neural Information Processing Systems*, 37:50925–50951.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Haoyuan Shi, Yunxin Li, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. 2025. Animaker: Multi-agent animated storytelling with mcts-driven clip generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11888–11898.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. 2025. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.

Hongchen Wei and Zhenzhong Chen. 2025. Visual context window extension: A new perspective for long video understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4281–4289.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. 2025. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, and 1 others. 2025a. Egolife: Towards egocentric life assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28885–28900.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Ying-Cong Chen. 2025b. Seed-story: Multimodal long story generation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1850–1860.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. 2025c. Vca: Video curious agent for long video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20168–20179.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, and 1 others. 2025. Re-thinking temporal search for long-form video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8579–8591.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. 2025. Videoexplorer: Think with videos for agentic long-video understanding. *arXiv preprint arXiv:2506.10821*.

Kaiqing Zhang, Sham M Kakade, Tamer Basar, and Lin F Yang. 2023. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *Journal of Machine Learning Research*, 24(175):1–53.

A Implementation Details

A.1 Training Data.

Our training corpus is constructed through the multi-agent data engine described in Sec. 3.2, comprising 76K SFT samples and 12K RL samples. We detail the data formulation for each agent below. And all tools within agents are powered by the Qwen3-LM and Qwen3-VL models.

Retrieval Agent. For SFT, we construct 42K retrieval tool-thinking trajectories that decompose the temporal localization process into interpretable reasoning steps. Specifically, each trajectory sequentially invokes three specialized tools: *Caption-based Retrieval* for coarse-to-fine semantic matching, *Visual Grounding* for visual instance clue extraction, and *Video-based Verification* for candidate refinement (in Sec. 3.3.3). This structured decomposition enables the agent to learn systematic retrieval strategies rather than end-to-end shortcuts. For RL, we collect 6K samples, each consisting of a video paired with multimodal retrieval clues (textual descriptions, visual exemplars and patial regions) as input, with the ground-truth temporal segments serving as the reward signal.

Planning Agent. For SFT, we curate 34K planning tool-thinking trajectories that capture the hierarchical task decomposition process and iterative agent coordination loop. Each trajectory demonstrates how to parse complex queries into atomic sub-tasks and appropriately dispatch them to either the Retrieval Agent $\mathcal{A}_{\text{ret}}$ or the Executor Agent $\mathcal{A}_{\text{exe}}$, thereby teaching the model effective orchestration strategies. For RL, we leverage 6K Instance-level Long-form Video Reasoning (ILVR) multiple-choice questions, where the agent receives reward signals based on final answer correctness, encouraging holistic optimization of the planning policy.

A.2 Two-stage Training

Both the Retrieval Agent and Planning Agent are optimized through a unified two-stage training paradigm: SFT on specialized tool-thinking trajectories, followed by reinforcement learning via PA-GRPO (Sec. 3.3.5). While the training framework remains consistent, the reward formulations are tailored to each agent's task-specific objectives.

Retrieval Agent. In the RL stage, we train $\mathcal{A}_{\text{ret}}$ using samples with ground-truth temporal segments. Given predicted temporal interval $\hat{\mathcal{T}} = [\hat{t}_s, \hat{t}_e]$ and ground-truth interval $\mathcal{T}^* = [t_s^*, t_e^*]$, the outcome-level accuracy is computed via temporal Intersection-over-Union (tIoU):

$$\text{tIoU}(\hat{\mathcal{T}}, \mathcal{T}^*) = \frac{\max(0, \min(\hat{t}_e, t_e^*) - \max(\hat{t}_s, t_s^*))}{\max(\hat{t}_e, t_e^*) - \min(\hat{t}_s, t_s^*)}. \tag{19}$$

The composite reward integrates localization accuracy with process-level quality:

$$R_i^{\text{ret}} = \mathbb{I}[\text{tIoU}(\hat{\mathcal{T}}_i, \mathcal{T}^*) \geq \tau] + \lambda \cdot r_p(o_i), \tag{20}$$

where τ is the IoU threshold, $\mathbb{I}[\cdot]$ denotes the indicator function, and $r_p(o_i)$ represents the process reward aggregated from step-wise evaluations.

Planning Agent. For $\mathcal{A}_{\text{plan}}$, we leverage ILVR multiple-choice questions during RL training. Given predicted answer $\hat{a}$ and ground-truth answer a^*, the outcome-level accuracy is defined as exact match:

$$\text{EM}(\hat{a}, a^*) = \mathbb{I}[\hat{a} = a^*]. \tag{21}$$

The composite reward formulation follows:

$$R_i^{\text{plan}} = \text{EM}(\hat{a}_i, a^*) + \lambda \cdot r_p(o_i), \tag{22}$$

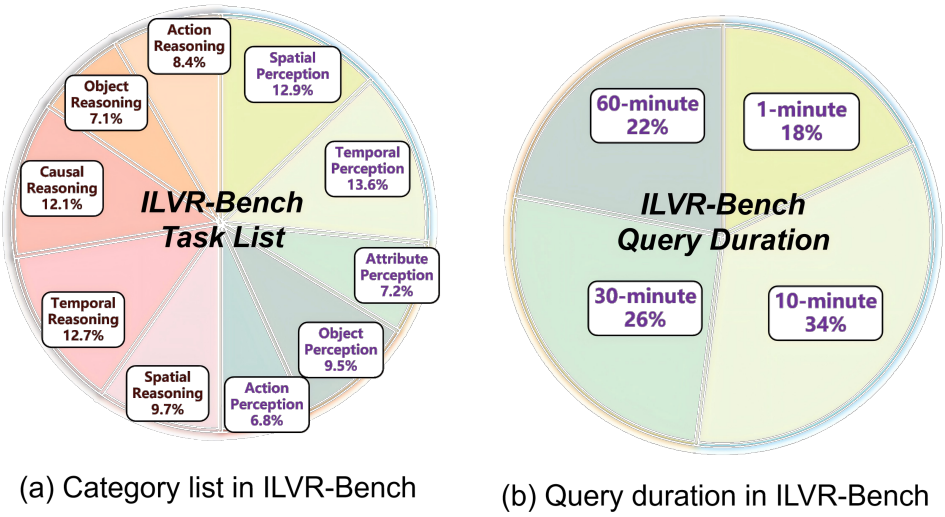
where the process reward $r_p(o_i)$ assesses the validity of task decomposition and agent dispatching decisions throughout the planning trajectory. λ is set to 1.

B ILVR-Bench

B.0.1 Statistics and Characteristics

Task Distribution. As illustrated in Figure 4(a), ILVR-Bench encompasses a diverse array of 10 task categories, which can be systematically grouped into two complementary dimensions: *perception* and *reasoning*. The perception-oriented tasks comprise Temporal Perception (13.6%), Spatial Perception (12.9%), Object Perception (9.5%), Attribute Perception (7.2%), and Action Perception (6.8%), collectively accounting for approximately 50% of the benchmark. Correspondingly, the reasoning-oriented tasks include Temporal Reasoning (12.7%), Causal Reasoning (12.1%), Spatial Reasoning (9.7%), Action Reasoning (8.4%), and Object Reasoning (7.1%). This balanced distribution ensures comprehensive coverage of both low-level perceptual understanding and high-level cognitive reasoning capabilities, thereby enabling a holistic evaluation of ILVR systems.

Temporal Span Distribution. Figure 4(b) presents the distribution of query durations across four temporal granularity levels. Notably, the 10-minute interval constitutes the largest proportion (34%),



(a) Category list in ILVR-Bench (b) Query duration in ILVR-Bench

Figure 4: Data characteristics of ILVR-Bench.

followed by 30-minute (26%), 60-minute (22%), and 1-minute (18%) spans. This stratified design deliberately emphasizes medium- to long-range temporal dependencies, with over 80% of queries requiring reasoning across intervals exceeding one minute. Such a distribution reflects real-world scenarios where understanding hour-level videos necessitates the integration of contextual information dispersed across extended temporal horizons, rather than relying solely on localized frame-level perception.

B.1 Evaluation Metric: QAR-Score

To rigorously quantify model performance on ILVR-Bench-QAR, we introduce the **QAR-Score**, a composite metric designed to assess the alignment between predicted answers and their underlying rationales. Formally, the score S_i for the i-th sample is defined as a gated objective:

$$S_i = \mathbb{I}(\hat{y}_i = y_i) \cdot [\alpha \cdot \mathcal{G}(\mathcal{C}_i) + \beta \cdot \mathcal{G}(\mathcal{T}_i)], \quad (23)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function that returns 1 if and only if the predicted answer $\hat{y}_i$ matches the ground-truth label y_i, and 0 otherwise. This gating mechanism enforces a necessary condition: reasoning quality is evaluated exclusively when the prediction is correct, thereby penalizing models that generate plausible-sounding but ultimately erroneous rationales.

For correctly answered instances, $\mathcal{G}(\cdot) \in [0, 1]$ represents an LLM-based automated evaluator that performs multi-dimensional assessment along two axes: (1) the *perceptual fidelity* of the visual evidence $\mathcal{C}_i$, measuring whether the extracted visual cues are accurate and task-relevant; and (2) the *inferential soundness* of the reasoning chain $\mathcal{T}_i$, evaluating logical coherence and evidence utilization. The weighting coefficients are empirically set to $\alpha = 0.4$ and $\beta = 0.6$, prioritizing the reasoning component to reflect the inherent complexity of multi-step logical deduction in long-horizon video understanding.

B.2 Visual Exemplar

Figure 5 illustrates representative samples from ILVR-Bench-QA and ILVR-Bench-QAR. In ILVR-Bench-QA, each instance comprises a multiple-choice question paired with an *instance prompt*, which provides essential contextual information to disambiguate entity references within the query. Specifically, the instance prompt integrates visual exemplars with spatially localized regions (*e.g.*, bounding boxes) and corresponding textual descriptions, thereby grounding abstract entity mentions to concrete visual referents.

Building upon this foundation, ILVR-Bench-QAR augments each sample with two structured annotation modules. The `<caption>` module furnishes temporally-anchored visual evidence, explicitly specifying the relevant time intervals and describing the observed scene content that substantiates the answer. The `<think>` module decomposes the reasoning process into a sequence of logically coherent steps, each articulating intermediate inferences that bridge perceptual observations to the final conclusion. This hierarchical annotation schema not only facilitates answer verification but also enables systematic evaluation of multi-step

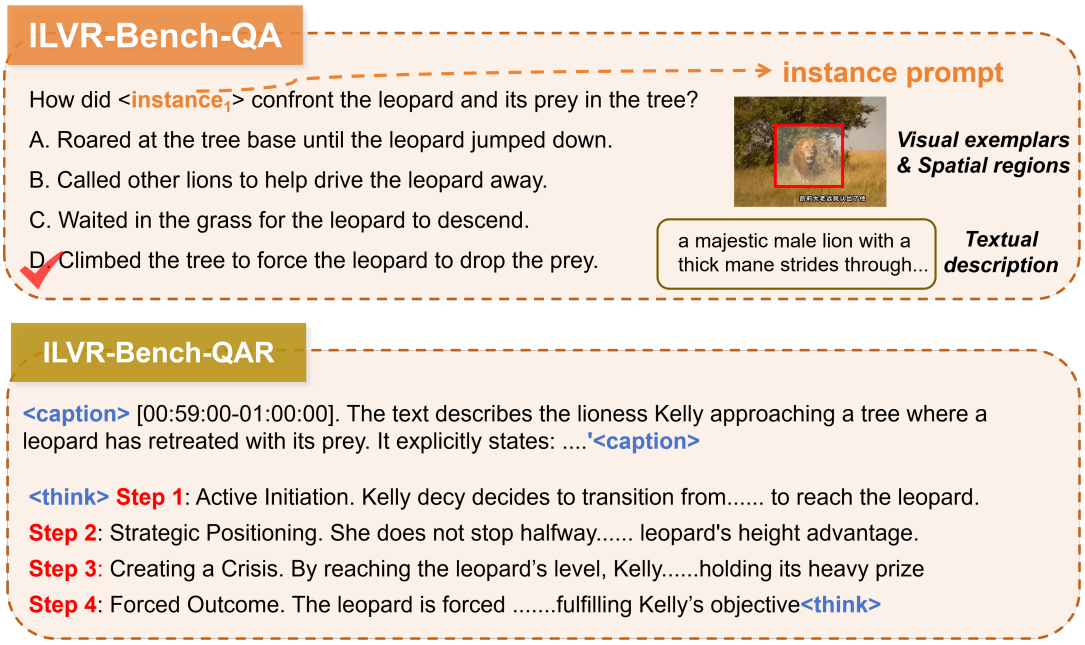


Figure 5: Visual Exemplar of ILVR-Bench.

reasoning capabilities in long-form video understanding.

C Qualitative Results

C.1 CoTT vs. CoAT

We present a representative example to illustrate the distinctions between Chain-of-Tool Thinking (CoTT) and our Chain-of-Agent Thinking (CoAT), as shown in Fig. 6. The task requires identifying that the character is *dreaming* of a body journey based on visual cues including a yellow submarine, fleshy tunnels, and X-ray imagery.

CoTT Limitations. As depicted in Fig. 6(a), conventional CoTT employs a single agent that sequentially invokes retrieval and Video-LLM tools. Despite multiple iterations, this approach fails to detect the critical "dreaming" context due to: (i) *evidence fragmentation*—processing retrieved segments in isolation without semantic integration, and (ii) *contextual noise accumulation*—irrelevant intermediate outputs diluting the reasoning focus, ultimately leading to an incorrect answer.

CoAT Advantages. In contrast, our ILVR-Agent with CoAT (Fig. 6(b)) decomposes the task across specialized agents. The Planning Agent formulates retrieval strategies explicitly targeting "George sleeping or dreaming". The Retrieval Agent then identifies additional segments (20:00-23:00) containing dream sequence evidence overlooked by CoTT. Finally, the Executor Agent verifies the "fantastical, dreamlike" characteristics, enabling correct answer derivation. This modular design effectively mitigates reasoning chain saturation while enhancing evidence discovery through task-specific specialization.

C.2 Case Study of ILVR-Agent

To illustrate the effectiveness of CoAT in ILVR-Agent, we present two representative cases in Fig. 7.

In the first case, the Planning Agent initially formulates abstract emotional labels (e.g., "Sad", "Funny") as retrieval clues. Upon receiving null results from the Retrieval Agent, the Planning Agent adaptively refines its strategy by replacing vague semantic descriptors with concrete perceptual cues (e.g., "beautiful weather"), thereby enabling successful evidence retrieval. This demonstrates the **self-reflective refinement** capability of our framework.

The second case further highlights the **cross-agent collaborative reasoning** mechanism. When the Executor Agent fails to verify the initial hypothesis regarding the necklace's function within the specified temporal segment, the Planning Agent re-evaluates the evidence sufficiency and re-invokes the Retrieval Agent with reformulated queries. This iterative loop ensures comprehensive evidence collection before final answer derivation.

These cases collectively validate two core advantages of ILVR-Agent: (1) *adaptive query reformu-*

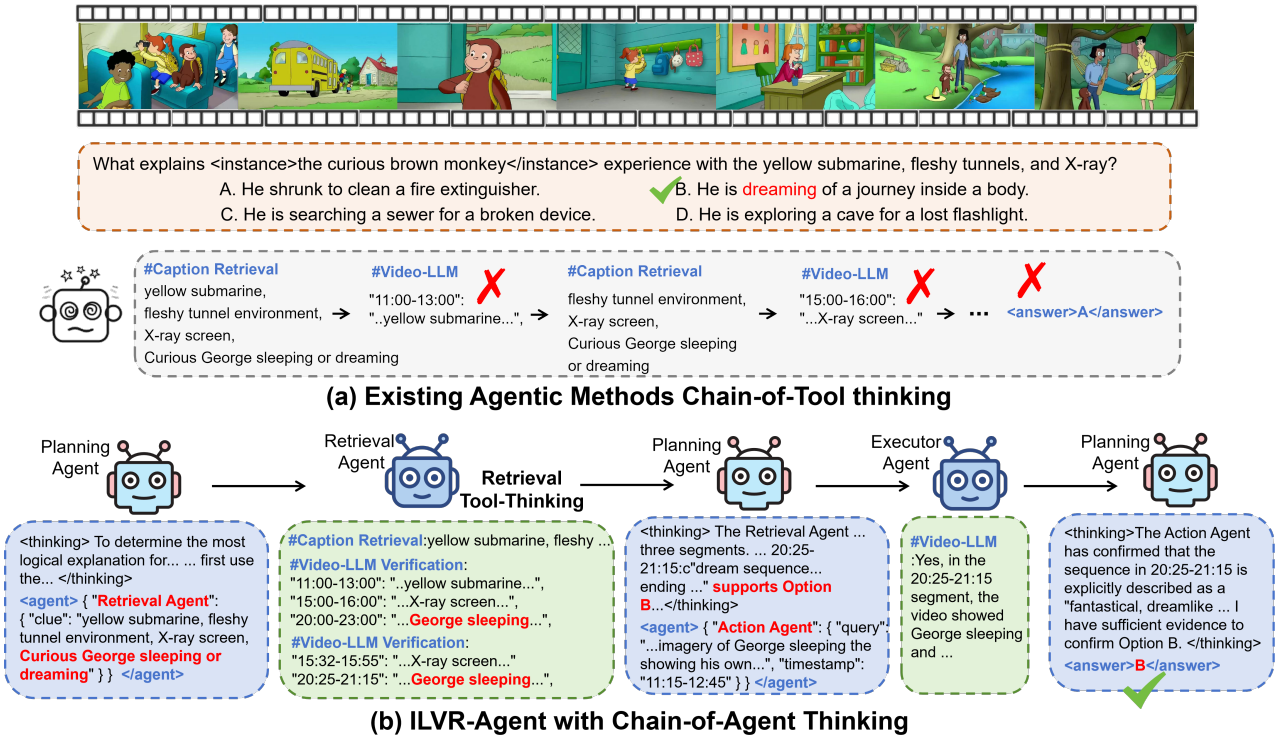


Figure 6: Visual comparison between **Chain-of-Tool Thinking (CoTT)** and **Chain-of-Agent Thinking (CoAT)** on an ILVR example. CoTT (top) operates as a single-agent pipeline, where tool invocations (caption retrieval and Video-LLM) are applied sequentially but often fail to preserve cross-frame consistency, leading to incorrect reasoning. CoAT (bottom) decomposes the reasoning task into specialized agents—Retrieval, Action, and Verification—which collaboratively ground temporal evidence (e.g., "George sleeping or dreaming") and converge on the correct answer (B).

lation through inter-agent feedback propagation, which mitigates retrieval failures caused by semantic ambiguity; and (2) *dynamic evidence accumulation* via multi-round agent collaboration, which addresses the inherent uncertainty in long-form video understanding. Such mechanisms substantially enhance the robustness and interpretability of the reasoning process.

D Verification Details

To ensure the reliability of the 1,000 multiple-choice questions generated by our multi-agent pipeline, we conducted rigorous human verification. Annotators were provided with a standardized interface displaying the video clip, system-generated question, candidate options, and multimodal instance prompt. Each instance was evaluated according to the following criteria:

- **Video-Grounded Answerability:** The question must be answerable solely based on visual and temporal information in the video, without requiring external knowledge that contradicts the depicted content.
- **Answer Uniqueness:** Exactly one unambiguous correct answer must exist among the four options. Ambiguous instances were flagged for revision.
- **Task Type Alignment:** Each question must correctly correspond to one of the 10 pre-defined task types (*e.g.*, object recognition, temporal reasoning).
- **Temporal Granularity Consistency:** The temporal distribution of questions must align with the four stratified granularity levels (1-, 10-, 30-, and 100-minute intervals).
- **Multimodal Prompt Validity:** Accurate temporal localization and coherent alignment between visual content and textual queries were verified.
- **Linguistic Clarity:** Questions and options were examined for grammatical correctness and freedom from ambiguity.

Ethical Considerations. All datasets used in this work are publicly available for research purposes. The human verification was conducted internally by the authors to ensure annotation quality and

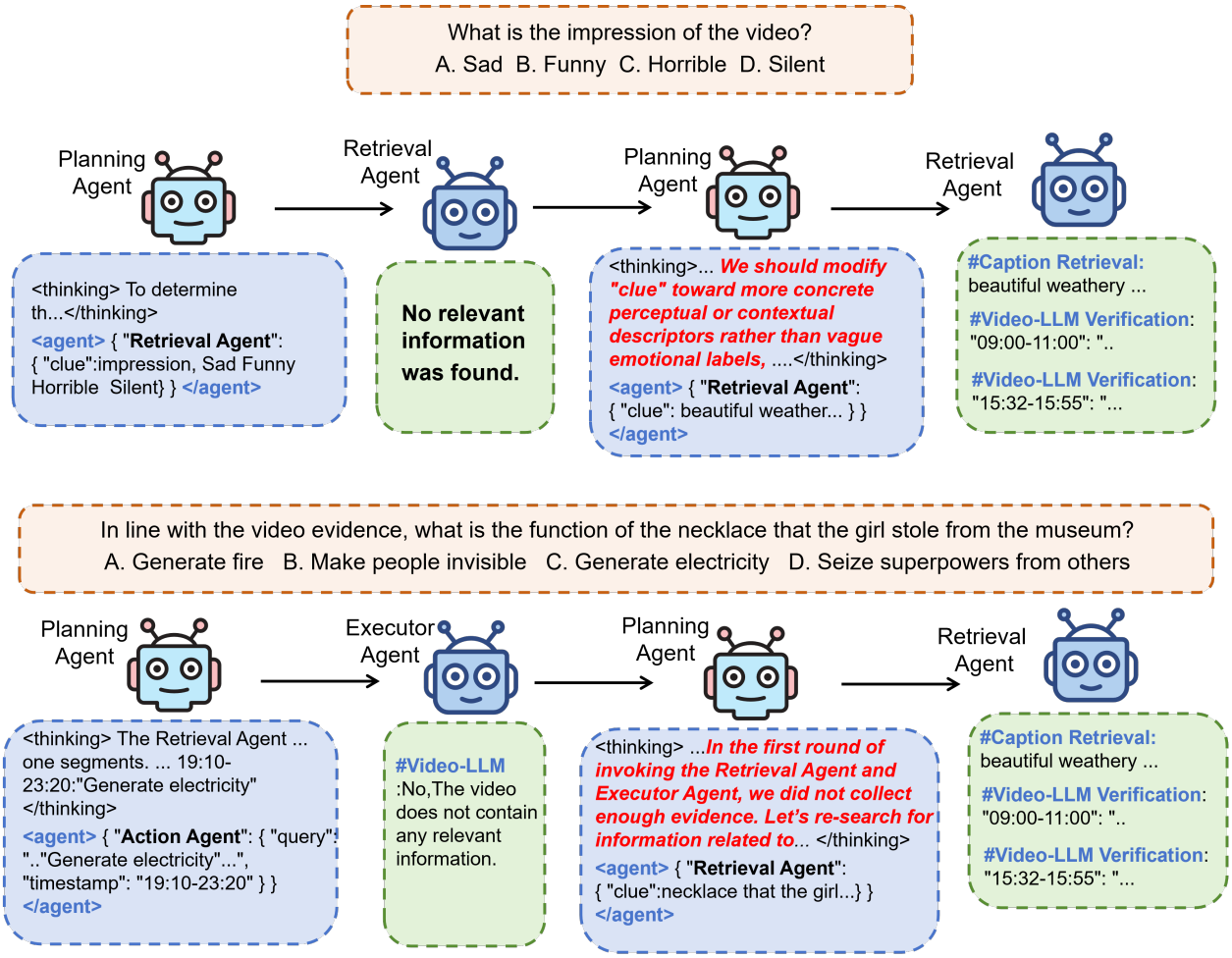


Figure 7: Case Study of ILVR-Agent.

consistency. No personally identifiable information was collected during the verification process.