JOINT LATENT REASONING AND OBSERVATION FOR AGENTIC VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective video agents must learn not only how to answer, but also where to look next. Transferring this capability to smaller models is challenging because each observation serves two purposes: supporting an answer and guiding the search for further evidence. Textual descriptions may capture what is currently visible while omitting cues needed for subsequent acquisition. We introduce **Latent Agentic Reasoning for Video Analysis (LARVA)**, a framework that distills multi-round video agents through jointly learned latent reasoning and visual observations. Rather than communicating through captions, an observer passes compact continuous representations to a reasoner that selects the next acquisition action or produces the final answer. Crucially, action and answer supervision propagate through this interface, training the observer to communicate information useful for downstream decisions. Observation reconstruction and reasoning-distribution alignment provide auxiliary supervision from textual teacher trajectories. Across five video-understanding benchmarks, LARVA achieves strong performance with 4B and 7B backbones. On VideoHolmes, Video-MME, and LVBench, the 7B model outperforms a textual CoT student trained on the same teacher trajectories by 5.1, 2.5, and 3.7 percentage points, respectively. These results highlight the value of jointly learning how agents represent evidence and act on it when distilling video-understanding capabilities.

1 INTRODUCTION

A video agent succeeds through a sequence of interdependent decisions: observe part of a video, reason about the available evidence, decide where to look next, and incorporate the new observation before answering. Recent agentic video systems implement this loop with tools that selectively inspect temporal regions (Li et al., 2026; Lin et al., 2026; Liu et al., 2026a). Each observation therefore affects both the interpretation of current evidence and the acquisition of future evidence. A missed temporal cue can redirect the search or cause premature stopping, leaving later reasoning without the evidence it needs. Distilling this capability into a smaller model requires preserving the coupling between observation, reasoning, and action throughout the interaction.

The representations passed around this loop determine what information can influence each decision. Many video agents express intermediate reasoning as textual chain-of-thought (CoT) (Wei et al., 2022); modular systems also communicate observations through descriptions or query-specific responses (Lin et al., 2026; Liu et al., 2026a). This repeatedly requires selecting which visual and reasoning details to verbalize. A description sufficient to identify an event may omit the temporal context needed to choose the next interval, leaving subsequent reasoning without direct access to that cue. Passing selected frames directly to the reasoner avoids textualization but adds visual tokens to its accumulating context (Zhang et al., 2026; Liu et al., 2026b). We ask: *how can latent representations improve the complete observation–reasoning–action loop, from using current evidence to acquiring the evidence needed next?*

Continuous latent representations provide a promising foundation. DyLaR combines latent perception and reasoning for video question answering, MACF communicates compact visual representations among video agents, and LatentRAG jointly learns latent reasoning and retrieval for iterative search (Xia et al., 2026; Chen et al., 2026; Zheng & Worring, 2026). Our focus is jointly distilling latent reasoning and visual observations within a closed-loop video agent. Latent observations condition reasoning about the next acquisition; that acquisition determines the next visual input, whose

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

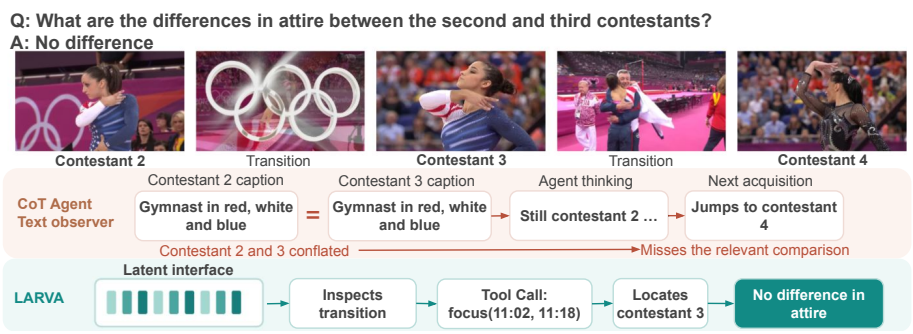


Figure 1: When asked to compare the contestants' attire, the CoT agent receives identical captions ("Gymnast in red, white, and blue"), confuses their identities, and skips ahead to the fourth contestant. LARVA samples around the transitions and correctly identifies no difference in attire.

latent representation feeds back into subsequent reasoning. Their value must therefore be assessed through both the decisions made along the trajectory and the final answer.

Learning this loop from textual teacher trajectories presents two challenges. First, the trajectories supervise reasoning traces, observations, and actions, but provide no direct targets for the latent states that replace the textual intermediates. Second, a representation useful in isolation may be ineffective in the loop: an observation latent may reconstruct a teacher description yet fail to support the reasoner's next decision. Our key design principle is to *jointly shape latent observations and reasoning by the downstream actions and answers they support*. The continuous interface makes it possible to train how evidence is encoded together with how it is used, while textual supervision anchors the learned states to the teacher trajectory.

We introduce **Latent Agentic Reasoning for Video Analysis (LARVA)**, which implements this principle through a jointly trained reasoner–observer architecture. At each round, the reasoner uses accumulated observations to compute latent reasoning states and select an explicit tool call or final answer. The observer encodes the requested frames into latent tokens that directly condition the next reasoning step. This closes the feedback loop between representing evidence and selecting further evidence, without decoding intermediate reasoning or observations into text. Each reasoning and observation block contains eight latent tokens in our main configuration, while raw visual processing remains in the observer.

To train LARVA, we replay teacher acquisition actions while replacing textual intermediates with generated latent states. Later action and answer losses propagate through the latent observation interface, so the observer receives supervision from how the reasoner uses its output. Observation reconstruction anchors visual content, while CoT-distribution alignment trains the combined latent context to induce next-reasoning predictions similar to those induced by teacher text in a frozen reference model. These objectives connect visual encoding, intermediate reasoning, and explicit decisions during training. The teacher fixes acquisitions during replay; at inference, the learned representations guide the student's own acquisition loop.

Our contributions advance both the design and learning of the complete video-agent loop:

- **A jointly learned latent observation–reasoning–action loop.** We introduce LARVA, which makes the representation of acquired visual evidence a trainable part of the acquisition policy. Latent observations feed directly into latent reasoning, which selects the next explicit action and thereby determines the evidence available in subsequent rounds. This formulation couples learning *how to represent evidence*, *how to reason over it*, and *where to look next* within a single agentic framework.
- **Learning visual communication from downstream decisions.** We develop a trajectory-distillation objective that trains what the observer communicates through the actions and answers its output supports. Action and answer supervision jointly shapes the observer and reasoner, while reconstruction and CoT-distribution alignment transfer textual supervision to their latent states. Joint training improves the latent/latent configuration by 5.3–7.2 percentage points across three diagnostic benchmarks, taking it from below the textual baseline to the strongest evaluated representation configuration (Table 2).

- **Stronger video understanding with acquisition gains across answerers.** LARVA improves over CoT students distilled from the same teacher trajectories on all five benchmarks with both 4B and 7B backbones, gaining 5.1, 2.5, and 3.7 points on VideoHolmes, Video-MME, and LVBench with the 7B model (Table 1). Its acquisition trajectories also improve accuracy with each of three fixed terminal answerers across all three diagnostic benchmarks; a fixed 27B visual answerer gains 1.7–4.2 points (Table 4). These results connect stronger final answers to improvements in the evidence-acquisition process itself.

2 Related Work

2.1 Agentic Video Understanding

Because only a small fraction of frames in a long video is typically relevant to a query, recent work replaces exhaustive encoding with agents that iteratively decide what evidence to acquire. VideoSeek follows a think–act–observe loop with multi-granularity observation tools (Lin et al., 2026), and EVA trains a single multimodal large language model (MLLM) with supervised and reinforcement learning to plan what to watch, acquire frames, and reflect on the resulting evidence (Zhang et al., 2026). These systems differ in how reasoning and perception are coupled. Unified agents such as EVA interpret the acquired frames directly within the reasoning model. Reasoner–observer designs such as LensWalk and VideoSeek instead let a language-model reasoner choose the temporal range and sampling strategy, while a vision-language observer returns evidence (Li et al., 2026; Lin et al., 2026). Other systems decompose the task further: LongVideoAgent coordinates grounding and vision agents (Liu et al., 2026a), VideoExplorer places a text-only reasoning model above a multimodal toolkit (Yuan et al., 2025), and VideoARM accumulates tool outputs in a hierarchical memory (Yin et al., 2026). Many of these systems are training-free.

Across these designs, reasoning and perception communicate through explicit interfaces. Intermediate reasoning is generated as text, and modular observers typically convert visual evidence into textual descriptions for the reasoner (Liu et al., 2026a; Li et al., 2026). Language is interpretable and convenient for combining independently developed models, but it requires sequential verbalization and projects high-dimensional visual evidence into a discrete representation before that evidence can influence subsequent decisions. LARVA retains the reasoner–observer decomposition but learns a continuous interface between the two roles.

2.2 Latent Reasoning and Communication

Latent reasoning replaces decoded reasoning steps with continuous hidden states. Coconut feeds the last hidden state back as the next input embedding and progressively replaces textual CoT steps through a curriculum (Hao et al., 2024), while CODI distills explicit CoT into a compact latent sequence by aligning teacher and student hidden states (Shen et al., 2025). In the visual domain, LaViT supervises latent visual thoughts with a teacher's visual semantics and attention patterns (Wu et al., 2026), and DyLaR combines separate perception and reasoning latents for video question answering over a fixed set of sampled frames (Xia et al., 2026). Continuous representations have also been used for communication and interaction. MACF trains local video agents to pass compact latents to a coordinator (Chen et al., 2026); ALAR learns latent reasoning from action supervision and switches adaptively between latent and explicit reasoning across tool-use turns (Jung et al., 2026); and LatentRAG jointly trains a large language model (LLM) and a retriever for iterative search with latent reasoning and queries (Zheng & Worring, 2026). LARVA differs from them by jointly distilling latent reasoning and latent visual observations within a closed-loop video agent, so that action and answer losses shape the observation interface and the reasoning states.

3 Method

LARVA learns a continuous interface between visual observation and sequential decision making. A reasoner selects which video evidence to acquire, and an observer encodes the selected frames into latent tokens that condition subsequent decisions. We distill this interaction from teacher trajectories using action and answer supervision, together with auxiliary observation reconstruction and CoT-distribution alignment. The central design is to train the observation interface jointly with the reasoner, so that its representations support both interpreting the evidence and decision-making.

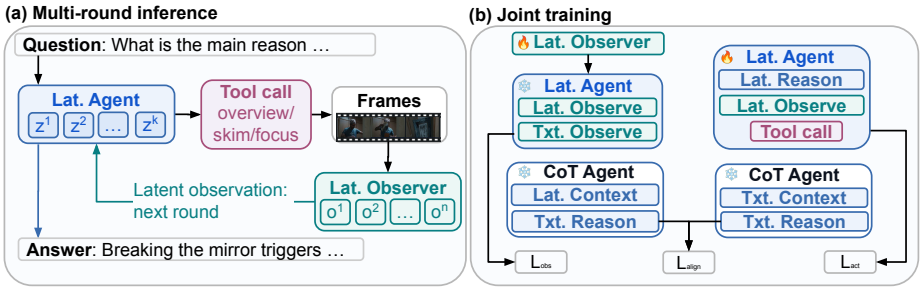


Figure 2: **Overview of LARVA. (a)** During multi-round inference, the latent agent selects video acquisitions and integrates compact latent observations to guide subsequent tool calls and produce an answer. **(b)** Joint training combines observation reconstruction $\mathcal{L}_{\text{obs}}$, CoT distribution alignment $\mathcal{L}_{\text{align}}$, and action supervision $\mathcal{L}_{\text{act}}$. The alignment loss uses a frozen CoT agent to match text- and latent-conditioned reasoning distributions, supervising both the latent agent and observer.

3.1 Sequential Visual Evidence Acquisition

Given a video V and a question Q, the agent alternates between reasoning, acquisition, and observation. Let $\mathcal{H}_{t-1}$ denote the retained interaction context before decision t. The reasoner first computes a latent reasoning block Z_t, then generates either a tool call A_t or a final answer. A nonterminal action specifies a tool type, a temporal interval when applicable, and an optional query. Executing it selects frames X_t from V; the observer returns a set of latent observation tokens O_t^i for each frame X_t^i, paired with the frame's timestamp, which enters the interaction context for subsequent decisions. The previous Z_t is also retained during the next round of decision making. A terminal answer triggers no further observation.

We use T to count nonterminal acquisition rounds and index the terminal answer decision by $T+1$. Thus, the final answer is conditioned on $\mathcal{H}_T$, which includes the observations returned by all completed acquisitions. Tool calls and answers remain explicit text, while intermediate reasoning and observation use continuous states.

Observation tools. We use the same VideoSeek (Lin et al., 2026) tools for controlled textual and latent students: **Overview** uniformly samples 64 frames across the video, while **Skim** and **Focus** each sample 16 frames from a specified interval for cue search and targeted inspection, respectively. Appendix A.1 details their roles.

3.2 Latent Reasoner–Observer Interface

LARVA separates two computational roles: a reasoner $\mathcal{R}_\theta$ for action selection and answering, and an observer $\mathcal{O}_\phi$ for visual processing. In our implementation, the roles share a backbone and use separate LoRA adapters (Hu et al., 2022). The role-specific parameters θ and ϕ include a projector after the backbone's last layer (Appendix A.2). Raw visual tokens are processed by the observer; the reasoner receives its continuous outputs rather than the frames or a decoded description.

Latent reasoning. Before each decision, the reasoner autoregressively computes a fixed-length block $Z_t = [z_t^1, \ldots, z_t^k]$, where $z_t^i \in \mathbb{R}^d$ and d is the latent embedding dimension. Writing f_θ for its continuous-state update, we have

$$z_t^i = f_\theta(Q, \mathcal{H}_{t-1}, z_t^{<i}), \qquad i = 1, \ldots, k. \tag{1}$$

Here $z_t^{<i}$ denotes preceding latents in the block. The model then decodes a textual tool call or answer conditioned on $(Q, \mathcal{H}_{t-1}, Z_t)$; Appendix A.2 gives the token-level factorization.

Latent observation. Let C_t^{obs} denote the observer input, containing the selected frames X_t, tool type c_t, and query q_t when present. The observer autoregressively computes n continuous states:

$$o_t^i = g_\phi(C_t^{\text{obs}}, o_t^{<i}), \qquad O_t = [o_t^1, \ldots, o_t^n] \in \mathbb{R}^{n \times d}. \tag{2}$$

Here g_ϕ is the observer's continuous-state update. The resulting O_t is inserted into the reasoner's context without decoding it into words. This makes the observer's output a learned interface to subsequent reasoning and acquisition. Each observation contributes a fixed number of latent positions

to the reasoner's context, although the context still grows across rounds and the observer must still process the selected visual input.

3.3 Trajectory Distillation and Joint Optimization

Textual supervision and initialization. A teacher reasoner and a teacher observer interact with the video to generate a trajectory

$$\widetilde{\mathcal{T}} = \left(\{(\widetilde{R}_t, \widetilde{A}_t, \widetilde{O}_t)\}_{t=1}^{T}, \widetilde{y} \right), \tag{3}$$

where $\widetilde{R}_t$ is textual reasoning, $\widetilde{A}_t$ is a tool call, $\widetilde{O}_t$ is the teacher observer's response to the selected frames, and $\widetilde{y}$ is the teacher's final answer. A terminal reasoning trace, when available, is denoted by $\widetilde{R}_{T+1}$. We first train a textual CoT student on the teacher reasoning, actions, and answers, and distill a textual observer from the teacher observations. These models initialize the latent reasoner and observer, respectively. A frozen copy of the trained CoT student, denoted by $p_{\bar{\psi}}$, provides the reference distributions for alignment.

Replaying acquisition trajectories. During training, we replay teacher acquisition actions while replacing textual intermediates with generated latent blocks. The reasoner predicts each teacher action, whose execution supplies frames for the latent observer and the next decision. At inference, the reasoner selects its own actions. Appendix A.3 details replay and teacher forcing.

Action and answer supervision. We supervise explicit decisions through the generated latent context:

$$\mathcal{L}_{\text{act}} = -\sum_{t=1}^{T} \sum_{j} \log p_\theta(\widetilde{a}_{t,j} \mid Q, \mathcal{H}_{t-1}, Z_t, \widetilde{a}_{t,<j}) - \sum_{j} \log p_\theta(\widetilde{y}_j \mid Q, \mathcal{H}_T, Z_{T+1}, \widetilde{y}_{<j}). \tag{4}$$

Here Z_{T+1} precedes the terminal answer. Each target token is scored using only its preceding target prefix; serialization and teacher-forcing details are in Appendix A.3.

The continuous observation interface allows downstream decision losses to update the observer as well as the reasoner. In particular, O_t can influence actions at steps $t+1$ and later, and the terminal answer; it cannot influence the already-selected action at step t. This supervision connects the representation of acquired evidence to its subsequent use. The teacher actions and frame-selection operations are fixed during replay, so this gradient path runs through latent computation rather than through the discrete acquisition operation.

Observation reconstruction. To anchor latent observations to teacher-described visual content, an auxiliary textual decoding branch reconstructs $\widetilde{O}_t$ from O_t. Denoting its decoding distribution by p_{rec}, we minimize

$$\mathcal{L}_{\text{obs}} = -\sum_{t=1}^{T} \sum_{j} \log p_{\text{rec}}(\widetilde{o}_{t,j} \mid O_t, \widetilde{o}_{t,<j}). \tag{5}$$

This objective encourages O_t to preserve information needed to recover the teacher observation. It complements decision supervision, which trains the interface for later actions and answering. Textual reconstruction is an auxiliary training branch; it is not part of the inference loop.

CoT-distribution alignment. To supervise how observations combine with prior reasoning, we compare next-CoT distributions under the frozen reference $p_{\bar{\psi}}$. Let $q_{t,j}$ and $p_{t,j}$ denote its predictions from textual and latent contexts, respectively, at token j after acquisition t. Both branches use the same question, teacher actions, and next-CoT prefix; only reasoning and observation representations differ. Appendix A.4 defines the contexts and token distributions explicitly.

Let $\mathcal{I}$ contain the acquisition steps with an available $\widetilde{R}_{t+1}$. We align predictions at corresponding token positions using

$$\mathcal{L}_{\text{align}} = \sum_{t \in \mathcal{I}} \sum_{j} D_{\text{KL}}(\text{sg}[q_{t,j}] \,\|\, p_{t,j}), \tag{6}$$

where sg denotes stop-gradient. The reference parameters remain frozen, while derivatives through the latent-context branch supervise the states supplied by the reasoner and observer. Alignment thus encourages latent contexts to support next-reasoning predictions similar to those induced by teacher text, without requiring direct targets for the latent vectors.

Overall objective. The three losses provide complementary supervision for explicit decisions, observation content, and the interpretation of latent contexts:

$$\mathcal{L} = \lambda_{\text{act}}\mathcal{L}_{\text{act}} + \lambda_{\text{obs}}\mathcal{L}_{\text{obs}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}. \tag{7}$$

The coefficients λ_{act}, λ_{obs}, and λ_{align} control their relative contributions. Joint optimization couples the observer's representations to the reasoner's use of those representations, while reconstruction and alignment supply supervision from the textual teacher trajectories.

3.4 Inference

At inference, the reasoner generates a latent reasoning block and then selects a tool call or produces the final answer. Each tool call acquires frames for the observer, whose latent output conditions the following decision. The loop ends when the reasoner answers or reaches the acquisition limit, after which it answers using the accumulated context. Our main configuration uses 8 latent tokens per reasoning and observation block and permits at most eight acquisition rounds. Neither textual CoT nor textual observations are decoded during inference; tool calls and final answers remain textual.

4 Experiments

Our experiments evaluate LARVA as a complete observation–reasoning–action loop. We first test whether it improves video understanding over students distilled from the same teacher trajectories. We then examine the contributions of latent reasoning, latent observations, and joint optimization, before evaluating inference cost and the usefulness of the acquired evidence. Answerer swaps, a human assessment of incremental evidence, and teacher-trajectory comparisons provide complementary views of acquisition behavior.

4.1 Experimental Setup

Training data and implementation. We generate 20,000 interaction trajectories from Video-Marathon (Lin et al., 2025), the VideoHolmes training set (Cheng et al., 2025), and LongVideo-Reason (Chen et al., 2025), filtering trajectories with incorrect final answers or invalid tool calls. DeepSeek-V4.1-Flash (DeepSeek-AI et al., 2026) and Qwen3.8-27B (Qwen Team, 2026) serve as the teacher reasoner and observer. We instantiate LARVA with Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-4B (Bai et al., 2025a), sharing a backbone with separate reasoner and observer LoRA adapters. The main configuration uses eight latent tokens per reasoning and observation block and at most eight acquisition rounds. Initialization, optimization settings, and acquisition tools are detailed in Appendix A.5.

Benchmarks and comparisons. We report accuracy on VideoHolmes (Cheng et al., 2025), Video-MME (Fu et al., 2025), LVBench (Wang et al., 2025), LongVideoBench (Wu et al., 2024), and MLVU (Zhou et al., 2024), excluding Video-MME captions from model input. CoT Policy learns the teacher's reasoning, tool calls, and answers; Direct Policy omits the reasoning traces. Both use the same distilled textual observer and source trajectories, controlling for the supervision data. EVA (Zhang et al., 2026), VideoMind (Liu et al., 2026b), DyLaR (Xia et al., 2026), and MACF (Chen et al., 2026) provide broader comparisons but differ in architecture and training. Unless otherwise stated, diagnostic experiments use Qwen2.5-VL-7B.

4.2 Main Results

Consistent gains across benchmarks and model scales. LARVA outperforms the CoT student on all five benchmarks with both backbones (Table 1). With Qwen2.5-VL-7B, it improves accuracy from 39.2 to 44.3 on VideoHolmes, 59.9 to 62.4 on Video-MME, and 41.3 to 45.0 on LVBench: gains of **5.1, 2.5, and 3.7 percentage points**, respectively. It also gains 1.3 points on LongVideoBench and 0.4 on MLVU. With Qwen3-VL-4B, the corresponding gains are 5.5, 1.9, 2.2,

Table 1: Main results. Accuracy (%) on five video-understanding benchmarks. Teacher models are listed for reference. CoT Policy and Direct Policy are controlled students distilled from the same teacher trajectories as LARVA.

Method	VideoHolmes (Cheng et al., 2025)	Video-MME (Fu et al., 2025)	LVBench (Wang et al., 2025)	LongVideoBench (Wu et al., 2024)	MLVU (Zhou et al., 2024)
Teacher models					
DeepSeek-V4.1-Flash (DeepSeek-AI et al., 2026)	60.4	80.1	63.3	71.9	77.2
Qwen3.8-27B (Qwen Team, 2026)	50.3	72.5	56.8	67.4	75.2
Qwen2.5-VL-7B (Bai et al., 2025b)					
EVA (Zhang et al., 2026)	37.2	<u>60.2</u>	<u>43.3</u>	55.0	68.3
VideoMind (Liu et al., 2026b)	<u>41.0</u>	58.2	40.8	56.3	64.4
DyLaR (Xia et al., 2026)	<u>41.0</u>	59.8	41.2	<u>57.4</u>	–
MACF (Chen et al., 2026)	–	58.2	37.7	52.5	47.6
CoT Policy	39.2	59.9	41.3	56.8	<u>68.4</u>
Direct Policy	34.6	53.4	37.6	52.0	61.8
LARVA (Ours)	**44.3**	**62.4**	**45.0**	**58.1**	**68.8**
Qwen3-VL-4B (Bai et al., 2025a)					
DyLaR (Xia et al., 2026)	<u>47.5</u>	57.6	38.5	<u>56.8</u>	–
CoT Policy	42.7	<u>60.2</u>	<u>39.1</u>	55.0	<u>64.9</u>
Direct Policy	33.2	55.8	36.7	52.3	59.5
LARVA (Ours)	**48.2**	**62.1**	**41.3**	**57.4**	**66.7**

Table 2: Contributions of latent reasoning and observation. Accuracy (%) of students with textual or latent reasoning and observation interfaces, with (✓) or without (✗) joint training of the reasoner and observer, on Qwen2.5-VL-7B (Bai et al., 2025b). All variants use the same training examples and tools.

Interface			Accuracy (%)		
Reasoning	Observation	Joint	VideoHolmes (Cheng et al., 2025)	Video-MME (Fu et al., 2025)	LVBench (Wang et al., 2025)
Textual	Textual	✗	39.2	59.9	41.3
Latent	Textual	✗	42.7	<u>61.9</u>	<u>43.6</u>
Textual	Latent	✗	38.6	59.3	40.7
Textual	Latent	✓	<u>43.5</u>	60.8	42.0
Latent	Latent	✗	37.4	55.2	39.7
Latent	Latent	✓	**44.3**	**62.4**	**45.0**

2.4, and 1.8 points. The unweighted mean improvement over the five benchmarks is 2.6 points for 7B and 2.8 points for 4B.

LARVA also achieves the highest reported accuracy within each backbone group in Table 1. For the 7B backbone, it exceeds the strongest listed alternative by 3.3 points on VideoHolmes, 2.2 on Video-MME, and 1.7 on LVBench. The Direct Policy falls below the CoT student on every benchmark with both backbones, whereas LARVA improves on CoT while replacing textual intermediates with learned latent states. Together, these comparisons support learning intermediate representations for the acquisition loop rather than simply omitting intermediate reasoning.

4.3 Representation and Training Ablations

Latent reasoning improves the textual-observation student. Replacing textual reasoning with latent reasoning while retaining textual observations raises accuracy by 3.5, 2.0, and 2.3 points on VideoHolmes, Video-MME, and LVBench (Table 2). Full LARVA further improves on this variant by 1.6, 0.5, and 1.4 points. The strongest configuration therefore combines latent reasoning with jointly learned latent observations, rather than obtaining its full benefit from latent reasoning alone.

Table 3: Inference cost per question, averaged over VideoHolmes, Video-MME, and LVBench. Reasoning tokens are generated by the reasoner, and observation tokens are received by it. Latency is measured on NVIDIA H200 GPUs. All local models use 7B backbones; EVA and the students use Qwen2.5-VL-7B (Bai et al., 2025b). †Relies on closed-source APIs, so latency includes API response time; excluded from Table 1.

Method	Turns	Tokens		Frames	Latency (s)
		Reasoning	Observation		
EVA (Zhang et al., 2026)	2.6	361.2	13984.4	71.4	<u>11.8</u>
VideoMind (Liu et al., 2026b)	3.0	16.6	19253.2	511.5	21.4
LongVideoAgent† (Liu et al., 2026a)	2.5	3196.9	168.8	21.7	30.8
CoT Policy	4.8	1752.5	1457.2	112.5	48.3
Direct Policy	4.4	0.0	1506.5	110.9	39.0
LARVA (Ours)	3.2	12.8	577.4	75.9	**8.2**

Joint optimization makes the latent interface effective. Latent observations are not sufficient by themselves: without joint training, the textual-reasoning/latent-observation variant scores 38.6, 59.3, and 40.7, below the fully textual student on all three benchmarks. Joint training raises these scores to 43.5, 60.8, and 42.0. The effect is larger when both interfaces are latent: joint optimization improves accuracy from 37.4 to 44.3, 55.2 to 62.4, and 39.7 to 45.0, gaining **6.9, 7.2, and 5.3-point**.

For the non-joint latent/latent variant, we first train the latent reasoner with textual observations, then freeze it and train the latent observer with the objectives in Section 3. Its weaker performance highlights the challenge of fitting a continuous observation interface to a fixed receiver trained on text. Joint training instead allows the reasoner and observer to adapt together, turning the latent/latent configuration from below the textual baseline into the strongest evaluated variant.

Latent-token capacity. Increasing observation or reasoning latents from one to eight generally improves accuracy, with smaller gains between six and eight. Very short observation sequences are particularly restrictive, supporting sufficient capacity for visual communication. Appendix A.7 provides the complete sweep and analysis (Figure 3).

4.4 Inference Cost

Table 3 summarizes average inference cost over VideoHolmes, Video-MME, and LVBench. LARVA records 8.2 seconds per question compared with 48.3 for CoT Policy and 39.0 for Direct Policy, corresponding to **5.9× and 4.8× speedups** in the reported setting. Relative to CoT, average acquisition turns decrease from 4.8 to 3.2 and processed frames from 112.5 to 75.9, reductions of 33.3% and 32.5%, respectively. Thus, its higher accuracy is accompanied by less visual processing on average.

LARVA also reduces the reported intermediate reasoning positions from 1752.5 to 12.8 and observation positions received by the reasoner from 1457.2 to 577.4. These counts describe the reasoner's interface; the observer still incurs the cost of encoding selected frames. LARVA has the lowest reported latency among the listed systems, but frame counts, token types, architectures, and API overhead differ across systems. We therefore interpret these measurements as the cost of each complete configuration rather than an isolated estimate of the savings from latent representations.

4.5 Evidence Acquisition Analysis

Acquired evidence benefits multiple terminal answerers. To assess whether gains extend beyond LARVA's own answer decoder, we replace the answerer when the acquisition policy stops or exhausts its action budget. Table 4 compares CoT and LARVA trajectories with each terminal answerer held fixed. LARVA trajectories improve accuracy in all nine answerer–benchmark comparisons. With the CoT answerer, the gains are 4.4, 1.2, and 0.2 points; with the latent answerer, they are 1.5, 2.4, and 3.5 points. A fixed 27B visual answerer gains **1.7, 4.2, and 2.6 points**, respectively.

The benefit to a separate visual answerer supports the usefulness of LARVA's selected evidence beyond compatibility with its own latent answerer. This complements the end-to-end results: improvements persist when terminal answering is held fixed. The swap experiment evaluates complete

Table 4: Evidence quality via answerer swaps. Trajectories are generated by the CoT or latent student; at the terminal decision, the acquired evidence is passed to the listed answerer.

Trajectory	Answerer	VideoHolmes	Video-MME	LVBench
CoT	CoT	39.2	59.9	41.3
	Latent	42.8	60.0	41.5
	27B Visual	49.9	63.7	49.6
Latent	CoT	43.6	61.1	41.5
	Latent	44.3	62.4	45.0
	27B Visual	**51.6**	**67.9**	**52.2**

Table 5: Human assessment of incremental evidence. Evidence-gain rate is adjudicated Yes divided by 100 questions. Uncertain cases remain in the denominator and are reported as counts. Both models receive the same 64-frame Overview and a 16-frame additional acquisition budget.

Model	Evidence-gain rate (%)	Uncertain (count)
Text reasoning + latent observation	11	7
Latent reasoning + textual observation	15	5
Full LARVA	35	8

acquisition trajectories, whose realized lengths and frame counts may differ; the human study below targets incremental evidence under a fixed additional-frame budget.

Human evaluation of incremental evidence. We conduct a paired comparison between full LARVA and the latent-reasoning/textual-observation ablation, both using Qwen2.5-VL-7B, on 100 held-out questions. Each model receives the same 64-frame Overview and makes one additional Skim/Focus call returning 16 frames. This fixes the initial visual coverage and the additional-frame budget, focusing the comparison on the usefulness of the next acquisition.

Two independent judges assess anonymized, randomly ordered acquisition sets for new answer-relevant clues absent from the Overview. Model identities, reasoning traces, and answers are hidden; repeated clues do not count. Judgments are Yes, No, or Uncertain, and disagreements are adjudicated. Appendix A.6 details the reference checklist and annotation procedure.
Full LARVA yields new answer-relevant evidence on **35%** of questions, compared with **15%** for the latent-reasoning/textual-observation and **11%** for text reasoning and latent observation. Full LARVA produces up to **24-percentage-point gain** under the same additional-frame budget (Table 5). Uncertain cases are retained in the denominator for both models. This result complements the answerer swaps by showing that the full model more often selects frames containing an additional useful clue, independently of its final-answer generation. This measures the first acquisition after a shared Overview; it does not measure evidence coverage across an entire trajectory.

Closer agreement with teacher acquisition trajectories. LARVA exceeds both textual students in tool-sequence similarity, temporal IoU, and optional-tool agreement across all three diagnostic benchmarks. Relative to CoT, its tool-sequence scores increase by 4.6, 0.8, and 1.3 points, and temporal IoU by 1.2, 2.3, and 2.0 points. These results support better transfer of teacher acquisition behavior, complementing the answerer swaps that assess the usefulness of acquired evidence. Appendix A.8 gives the metric definitions and complete results (Table 6).

5 Conclusion

We introduce LARVA, a framework that jointly learns latent reasoning and latent visual observations for agentic video understanding. By combining action supervision, observation reconstruction, and CoT distribution alignment, LARVA learns compact intermediate representations that support multi-round evidence acquisition and answer generation. Experiments across three benchmarks show that LARVA outperforms existing methods while reducing inference latency. Further analyses demonstrate stronger evidence-gathering capability than textual distilled students and closer agreement with the teacher's acquisition behavior. These results highlight joint latent training as an effective approach to building accurate and efficient video agents.

6 LIMITATIONS

LARVA has several limitations. First, it operates on sampled visual frames and does not process audio. It therefore cannot directly use dialogue, environmental sounds, or other acoustic cues, limiting its applicability to questions requiring audiovisual understanding.

Second, training relies on teacher-generated trajectories and a frozen CoT student for distribution alignment. These supervision sources may contain errors or suboptimal acquisition decisions that LARVA can inherit. Moreover, training on teacher trajectories may provide limited coverage of the histories encountered during student execution, particularly after an incorrect acquisition.

Finally, compact latent representations are less directly interpretable than textual reasoning and observations. Although tool calls and acquired frames remain inspectable, identifying whether a failure originates in visual compression, reasoning, or evidence selection is more difficult. Compression may also discard fine-grained details needed for some questions.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing, figure preparation, and discussions of research framing and experimental design. We also used teacher models to generate the training trajectories described in the paper. The authors reviewed AI-assisted content for accuracy and consistency and take responsibility for the manuscript, methodology, and reported results.

ETHICS STATEMENT

This work studies efficient evidence acquisition for video understanding using existing research datasets. Video agents may produce incorrect interpretations, inherit biases from their training data and teacher models, or overlook relevant visual evidence. Their outputs therefore require appropriate human oversight in consequential applications. Any use or release of video data and derived training trajectories should respect applicable dataset licenses, privacy requirements, and usage restrictions.

REPRODUCIBILITY STATEMENT

The paper describes LARVA's architecture, training objectives, teacher-trajectory generation, and evaluation protocols. To support reproducibility, we will provide implementation details, model configurations, prompts, hyperparameters, and evaluation scripts, including the protocols for trajectory similarity, answerer-swap evaluation, and inference-time measurement. We plan to release code and permitted training artifacts, subject to the licenses and usage restrictions of the underlying datasets and models.

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, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Kerui Chen, Jinglu Wang, Jianrong Zhang, Ming Li, Yan Lu, and Hehe Fan. Scaling video understanding via compact latent multi-agent collaboration, 2026. URL https://arxiv.org/abs/2605.00444.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling RL to long videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning?, 2025. URL https://arxiv.org/abs/2505.21374.

DeepSeek-AI, :, Anyi Xu, B. Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, C. C. Yu, Chao Jin, Chaofan Lin, Chen Dong, Chenbing Wang, Chenfan Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyuan Zhang, Chenhao Xu, Chenqi Zhao, Chenze Shao, Chuhao Wang, Chuqi Zhang, Damai Dai, Dejian Yang, Deli Chen, Di Huang, Di Wu, Donghao Li, Erhang Li, Eric Fu, F. Zhou, Fangwei Zhou, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, Guangbo Hao, Guanglin Li, Guanting Chen, Guoai Cao, Guofan Fan, Guolai Meng, Guowei Li, Haichuan Zhang, Haiyang Ma, Haiyang Shen, Han Li, Han Yu, Han Zhang, Hangyuan Deng, Hanwei Xu, Hanxiang Xu, Hanxun Zhong, Hao Guo, Hao Jiang, Hao Li, Hao Qin, Haodong Wen, Haofen Liang, Haofeng Huang, Haohua Liu, Haoling Zhang, Haoming Luo, Haoran Yang, Haotian Xu, Haotian Yuan, Haoting Huang, Haowen Luo, Haoyang Cai, Haoyu Chen, Haozhe Ji, Hengran Zhang, Hengrui Wang, Hengxu Wu, Honghui Ding, Hongxuan Tang, Huadong Wang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J. Yang, J. H. Jin, J. H. Zhang, J. X. Zou, Jia Yu, Jiahui Zhou, Jiajun Chen, Jialiang Huang, Jialin Zhao, Jiamin Tang, Jian Zhou, Jianan Tong, Jianwen Li, Jiaqi Zhu, Jiarui Wang, Jiasheng Ye, Jiashi Li, Jiaxin Xu, Jiaying Ding, Jibai Lu, Jiewen Hu, Jin Yan, Jincheng Zhai, Jingchang Chen, Jingcheng Hu, Jingli Zhou, Jingsheng Xu, Jingting Xiang, Jingyan Yun, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jinpeng Wang, Jinyi Chen, Jinyi Hu, Jiping Yu, Jueliang Guo, Junbo Pei, Junbo Sun, Junguang Jiang, Junjie Qiu, Junkang Zhou, Junqi Liu, Junren Li, Junxian Li, Junxiao Song, Junyi Guo, Kai Dong, Kaifeng Chen, Kaige Gao, Kang Guan, Kangdong Yuan, Ke Hong, Ke Xu, Kefan Zhao, Kexin Ji, Kexin Zhang, Kexing Zhou, Kuai Yu, Lan Zhang, Lean Wang, Lecong Zhang, Lei Wang, Letian Gao, Liang Zhao, Liansheng Xu, Lihua Guo, Lingxiao Luo, Lingyue Fu, Litao Deng, Litong Wang, Liyue Zhang, Longhao Chen, Lu Chen, Luotian Huang, Luyao Ma, Luyao Wang, M. S. Di, Max Mei, Menghao Ye, Miao Cui, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingjing Zhang, Mingqi Wei, Mingshu Chen, Mingxing Liu, Mingxu Zhou, Mingyu Xu, Mingyu Yang, Mingze Wang, Muyang Chen, Ni Shentu, Ning Wang, Niufang Ning, Panpan Huang, Peixin Cong, Peiyi Wang, Peiyuan Xin, Pengfei Ren, Pengfei Yan, Pengle Zhang, Qi Kang, Qi Tang, Qiancheng Wang, Qiang Li, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Qizhou Guo, Rongxian Xu, Rui Ding, Rui Hu, Rui Tian, Rui Yu, Ruidong Zhu, Ruifan Xu, Ruihan Yang, Ruihang Xia, Ruijie Lu, Ruilin Geng, Ruipeng Hong, Ruiqi Ge, Ruisong Zhang, Ruize Sun, Ruizhe Pan, Runji Wang, Runqian Chen, Runxin Xu, Ruohong Tian, Ruomeng Shen, Ruoyu Zhang, Ryan X., S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoheng Nie, Shaoyuan Chen, Shengding Hu, Shengkai Lin, Shengwen Ran, Shengyu Liu, Shengyuan Jia, Shi Bai, Shi Feng, Shicheng Xu, Shichun Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shiyuan Feng, Shufan Gong, Shuhan Lin, Shuiping Yu, Shunfeng Zhou, Shuo Yang, Shuomeng Wang, Shuting Guo, Shuting Pan, Shuying Yu, Sinuo Cao, Siyi Lin, Sizhe Chen, Songyang Chen, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian Pei, Tian Ye, Tianle Lin, Tianran Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tongrui Xiong, Wangding Zeng, Wei Liu, Wei Zhang, Weibin Xu, Weihao Zeng, Weilin Zhao, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjing Yao, Wenjun Gao, Wenkai Shao, Wenkai Yang, Wenli Zhang, Wenlu Wang, Wenlve Huang, Wenqian Yan, Wentao Zhang, Xi Gao, Xiang He, Xiang Li, Xiangli Li, Xiangwen Wang, Xiangying Zhang, Xiankui Wei, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojian Qu, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaoyao Zou, Xiaoyuan Li, Xicheng Guo, Xieting Chu, Xin Cheng, Xin Liu, Xin Xie, Xinbo Xu, Xingchao Liu, Xingchen Liu, Xingkai Yu, Xingyou Li, Xintong Yao, Xinyang Chen, Xinyong Jiang, Xinyu Yang, Xinyu Yang, Xu Chen, Xuanyu Wang, Xubei Zhong, Xuecheng Su, Xuejie Liu, Xuheng Lin, Xujie Fan, Xuncheng Zhao, Xuwei Fu, Y. C. Yan, Y. H. Jiang, Y. T. Wu, Y. W. M., Y. Z. Wang, Yafei Gao, Yang Yang, Yang Zhang, Yanru Ma, Yanwen Huang, Yao Li, Yao Li, Yao Meng, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yaoyang Ye, Yehang Yin, Yexinrui Wu, Yi Qian, Yi Tao, Yi Yu, Yichao Zhang, Yichen Jiang, Yicheng Wang, Yifan Ding, Yifan Shi, Yifeng Peng, Yifeng Zhai, Yijia Wu, Yiliang Xiong, Yilun Wang, Ying He, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yiping Wang, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyao Yang, Yiyuan Liu, Yizai Cai, Yizhen Wei, Yizhi Wang, Yonglun Yang, Yongqi Zhuo, Yongqiang Guo, Yongtong Wu, Yu Wu,

Yu Zhang, Yuan Bian, Yuan Cheng, Yuan Ou, Yuan Sun, Yuanfan Xu, Yuanhang Sun, Yuanhao Li, Yuchen Liu, Yuchen Yao, Yudong Han, Yuduan Wang, Yuhan Wu, Yuhao Meng, Yuheng Zou, YuKun Li, Yunchuan Wang, Yunfan Xiao, Yunfan Xiong, Yupeng Chen, Yuqian Cao, Yuqian Wang, Yuqing Chen, Yushun Zhang, Yutong Lin, Yuwei Xiao, Yuxian Gu, Yuxiang Chen, Yuxiang Huang, Yuxiang Luo, Yuxiang You, Yuxin Chen, Yuxin Xiang, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhe Guo, Yuzhen Huang, Yuzhuo Bai, Z. Y. Z., Zanlin Ni, Zehao Wang, Zehua Zhao, Zehui Ren, Zejun Zhao, Zhangli Sha, Zhanying Wang, Zhaochen Zhang, Zhaoshuai Du, Zhe Fu, Zhean Xu, Zhenda Xie, Zheng Liu, Zhengyan Zhang, Zhenhua Dong, Zhewen Hao, Zhibang Wang, Zhibin Gou, Zhicheng Ma, Zhihao Li, Zhihong Shao, Zhihuan Huang, Zhijie Li, Zhirui Lu, Zhixian Huang, Zhixuan Chen, Zhixuan Chen, Zhixuan Pan, Zhiyu Wu, Zhizhou Ren, Zhu He, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zili Zhang, Zilin Li, Zilong Hou, Zilong Lyu, Ziqiao Wang, Ziwei Xie, Ziya Zhang, Ziyi Gao, Zizheng Pan, Zonglin Li, Zongqing Yao, Zui Chen, Zuofan Wu, Chenchen Ling, Chengyu Hou, Chong Chen, D. Li, Di Qi, Dongjie Ji, Fang Wei, Fanyi Xia, Fei Xie, Feiyi Tan, Hailong Guo, Haiyan Zhai, Hui Zhou, Huihui Tan, Huijie Li, Jia Luo, Jia Song, Jialu Cai, Jian Liang, Jiangting Zhou, Jiaqi Gao, Jiayi Shao, Jie Chen, Jieyu Yang, Jin Chen, Jingde Zhang, Jingzi Zhou, Jinqian Wang, Jinyang Liu, JinZhao Sun, Junhua Ling, Junmin Zheng, Kaicheng Yang, Ke Xu, Le Su, Leyi Xia, Liangfeng Ding, Lin Zhuo, Linwang Ma, Linyan Zhu, Liyu Cai, Luqi Yao, M. K. Zhang, Meng Li, Miao Lin, Miaojun Wang, Min Zhang, Mingming Li, Mingming Wang, Mingze Yin, Minmin Han, Nan Cao, Ning Wang, Ningxin Ma, Panpan Wang, Peihan Lin, Peng Sun, Peng Zhang, Qian Ying, Qiang Xiang, Qiao Wang, Qingmiao Mao, Qiwei Jiang, Rongli Jin, Ruyi Chen, Sha Tao, Shangmian Sun, Shaoqing Wu, Shichao Zou, Si Lei, Tianyang Zhang, Tianyu Sun, Tingting Yin, W. L. Xiao, Wei An, Wei Li, Wei Wang, Weiwei Lin, Wenqing Hou, X. Lin, Xiangfei Meng, Xianzhu Huang, Xiao Peng, Xiaoqian Li, Xiaoting Zhang, Xiaowen Sun, Xiaoxiang Wang, Xiaoyu Ye, Xinrou Zhang, Xinyu Zhang, Xue Cao, Xueyin Chen, Yanan Zhou, Yanhong Xu, Yao Xia, Yao Xu, Yi Shao, Yihong Zhang, Yiling Ma, Ying Tang, Yining Lou, Yiru Chen, Yishi Piao, Yixuan Chen, Yong Xiong, Yuchen Xuan, Yuehan Yang, Yuer Xu, Yukun Zha, Yunxian Ma, Yuping Lin, Yuting Yan, Yutong Xie, Yuwen Sheng, Yuxuan Zhu, Zekai Zhang, Zhe Ju, Zhenzhen Lin, Zheren Gao, Zheyang Sun, Zhigang Yan, Zhongyu Wu, Zi Wang, Zihua Qu, Ziling Yan, and Ziyi Wan. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026. URL https://arxiv.org/abs/2609.19969.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24108–24118. Computer Vision Foundation / IEEE, 2025.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Dongwon Jung, Peng Shi, Yi Zhang, Junshan Zhang, and Muhao Chen. Adaptive latent agentic reasoning, 2026. URL https://arxiv.org/abs/2606.02871.

Keliang Li, Yansong Li, Hongze Shen, Mengdi Liu, Hong Chang, and Shiguang Shan. Lenswalk: Agentic video understanding by planning how you see in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19518–19528, June 2026.

Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Xiaodong Yu, Hao Chen, Jiebo Luo, Zicheng Liu, and Emad Barsoum. Unleashing hour-scale video training for long video-language understanding. 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/19741c617c87a25627682e5714af1501-Abstract-Conference.html.

Jingyang Lin, Jialian Wu, Jiang Liu, Ximeng Sun, Ze Wang, Xiaodong Yu, Jiebo Luo, Zicheng Liu, and Emad Barsoum. Videoseek: Long-horizon video agent with tool-guided seeking. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5465–5475, June 2026.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. LongVideoAgent: Multi-agent reasoning with long videos. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40404–40416, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1876. URL https://aclanthology.org/2026.acl-long.1876/.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. VideoMind: A chain-of-LoRA agent for temporal-grounded video reasoning. In *International Conference on Learning Representations (ICLR)*, 2026b.

Qwen Team. Qwen3.8-27B. https://huggingface.co/Qwen/Qwen3.8-27B, 2026.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing chain-of-thought into continuous space via self-distillation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.36. URL https://aclanthology.org/2025.emnlp-main.36/.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL https://arxiv.org/abs/2407.15754.

Linquan Wu, Tianxiang Jiang, Yifei Dong, Haoyu Yang, Fengji Zhang, Shichaang Meng, Ai Xuan, Linqi Song, and Jacky Keung. Lavit: Aligning latent visual thoughts for multi-modal reasoning, 2026. URL https://arxiv.org/abs/2601.10129.

Haotian Xia, Zilin Xiao, Junbo Zou, Vicente Ordonez, and Hanjie Chen. Perception before reasoning: Dynamic latent reasoning for video understanding and question answering, 2026. URL https://arxiv.org/abs/2608.04124.

Chenghao Yang, Yuning Zhang, Zhoufutu Wen, Tao Gong, Jiaheng Liu, Qi Chu, and Nenghai Yu. When agents look the same: Quantifying distillation-induced similarity in tool-use behaviors, 2026. URL https://arxiv.org/abs/2604.21255.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24042–24051, June 2026.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. Videoexplorer: Think with videos for agentic long-video understanding, 2025. URL https://arxiv.org/abs/2506.10821.

Yaolun Zhang, Ruohui Wang, Jiahao Wang, Yepeng Tang, Xuanyu Zheng, Haonan Duan, Hao Lu, Hanming Deng, and Lewei Lu. EVA: Efficient reinforcement learning for end-to-end video agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12289–12299, June 2026.

Yijia Zheng and Marcel Worring. Latentrag: Latent reasoning and retrieval for efficient agentic rag, 2026. URL https://arxiv.org/abs/2605.06285.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

A Appendix

This appendix supplements Sections 3 and 4 with implementation and training details, observation tools, action-decoding and alignment formulations, the complete human-evaluation protocol, and additional analyses of latent-token capacity and teacher-trajectory similarity.

A.1 Observation Tools

We adopt the **Overview**, **Skim**, and **Focus** interface of VideoSeek (Lin et al., 2026), and use the same tools for our controlled textual and latent students. Overview uniformly samples 64 frames across the full video. Skim samples 16 frames from a reasoner-specified interval and asks the observer to search for query-relevant cues. Focus also samples 16 frames from a specified interval, with a query targeting a concrete sub-question. These operations support global exploration and progressively targeted inspection while keeping acquisition actions externally executable.

A.2 Projection and Explicit Action Decoding

We use θ and ϕ to denote the role-specific trainable parameters, including the final projector following the backbone's last-layer output, which we implement as one layer-norm followed by a linear layer.

The prefix $z_t^{<i}$ contains the preceding latent states in the current block. After completing the block, the model switches to textual decoding to predict the next tool call or answer. For an acquisition action,

$$p_\theta(A_t \mid Q, \mathcal{H}_{t-1}, Z_t) = \prod_j p_\theta(a_{t,j} \mid Q, \mathcal{H}_{t-1}, Z_t, a_{t,<j}), \tag{8}$$

where $a_{t,j}$ is the j-th token in the serialized tool call.

A.3 Trajectory Replay and Teacher Forcing

During latent training, we retain the teacher's acquisition actions but replace textual intermediates with generated latent blocks. At step t, the reasoner produces Z_t from the current latent context and predicts the teacher action $\widetilde{A}_t$. We then append the teacher-forced action, execute its frame-sampling operation, and compute O_t from the selected frames. The action and latent observation become available to the next decision. The teacher thus fixes which evidence is acquired during training, while the students learn how to represent and use that evidence. At inference, acquisition actions are generated by the reasoner itself.

Here Z_{T+1} is the reasoning block preceding the terminal answer, and the answer sequence includes the model's terminal response formatting. Each target token is scored using only its preceding target prefix. Teacher forcing therefore supplies the history for subsequent prediction positions without exposing the token being predicted.

A.4 Alignment Contexts and Token Distributions

Reconstructing an observation does not directly supervise how it combines with prior reasoning to support the next decision. We therefore compare the next-CoT predictions induced by textual and latent contexts under the same frozen reference $p_{\bar{\psi}}$. After acquisition t, the textual context C_t^{text} contains teacher reasoning and observations. The corresponding context C_t^{lat} replaces these fields with generated reasoning and observation latents. The question, teacher actions, and shared contextual fields are held fixed.

For every step with an available next reasoning trace, both branches receive the same teacher-forced prefix $\widetilde{R}_{t+1,<j}$. Their distributions over the text vocabulary are

$$q_{t,j}(v) = p_{\bar{\psi}}(v \mid C_t^{\text{text}}, \widetilde{R}_{t+1,<j}),$$
$$p_{t,j}(v) = p_{\bar{\psi}}(v \mid C_t^{\text{lat}}, \widetilde{R}_{t+1,<j}). \tag{9}$$

The distributions in Equation 9 are compared at corresponding token positions by the KL objective in Equation 6. The reference parameters remain frozen, while the latent-context branch retains derivatives to the states generated by the reasoner and observer.

A.5 Training and Evaluation Setup

Training data and implementation. We generate 20,000 interaction trajectories from Video-Marathon (Lin et al., 2025), the VideoHolmes training set (Cheng et al., 2025), and LongVideo-Reason (Chen et al., 2025). DeepSeek-V4.1-Flash (DeepSeek-AI et al., 2026) serves as the teacher reasoner, producing reasoning traces, tool calls, and final answers; Qwen3.8-27B (Qwen Team, 2026) serves as the teacher observer, returning textual observations of the requested frames. Teacher trajectories with an incorrect final answer or at least one invalid tool call are filtered out. These trajectories supervise explicit decisions and provide textual targets for learning the latent interface.

We instantiate LARVA with Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-4B (Bai et al., 2025a). The reasoner and observer share a backbone with separate LoRA adapters ($r = 16$, $\alpha = 32$), initialized from the distilled CoT student and textual observer, respectively. We train for one epoch with a learning rate of 10^{-4} and set $\lambda_{\text{act}} = \lambda_{\text{obs}} = \lambda_{\text{align}} = 1$. The main configuration uses eight latent tokens per reasoning and observation block and permits at most eight acquisition rounds. Tool calls and final answers are decoded as text. The acquisition tools are Overview (64 frames) and Skim/Focus (16 frames each), as defined in Section 3.1. Since we train LARVA from already trained CoT and captioner models, the CoT policy, Direct policy, and captioner in Section A are trained for two epochs instead of one to ensure fair comparison.

A.6 Human-Evaluation Protocol

We conduct a paired comparison between full LARVA and the latent-reasoning/textual-observation ablation, both using Qwen2.5-VL-7B, on 100 held-out questions. Each model receives the same 64-frame Overview and makes one additional Skim/Focus call returning 16 frames. This fixes the initial visual coverage and the additional-frame budget, focusing the comparison on the usefulness of the next acquisition.

Two independent judges inspect the question, a reference evidence checklist, the shared Overview, and anonymized acquisition sets presented in randomized order. Model identities, reasoning traces, and answers are hidden, and the checklist is not provided to either model. Judges label whether the actual sampled frames reveal a new answer-relevant clue absent from the Overview as Yes, No, or Uncertain, with a brief justification. Repeated clues do not count as new evidence. Disagreements are adjudicated.

Evidence-gain rate is the number of adjudicated Yes judgments divided by all 100 questions; Uncertain cases remain in the denominator and are reported separately in Table 5. This evaluation measures the first acquisition after a shared Overview, rather than evidence coverage across an entire trajectory.

A.7 Latent-Token Capacity

Figure 3 varies the number of observation or reasoning latents from one to eight. Accuracy generally improves with capacity on all three benchmarks, with smaller changes between six and eight tokens. The observation sweep is particularly sensitive to very short sequences, consistent with an overly narrow interface restricting the visual information available for downstream decisions. Eight tokens provide a strong accuracy setting within the evaluated range. The figure measures accuracy, so it does not by itself establish an optimal latency–accuracy trade-off.

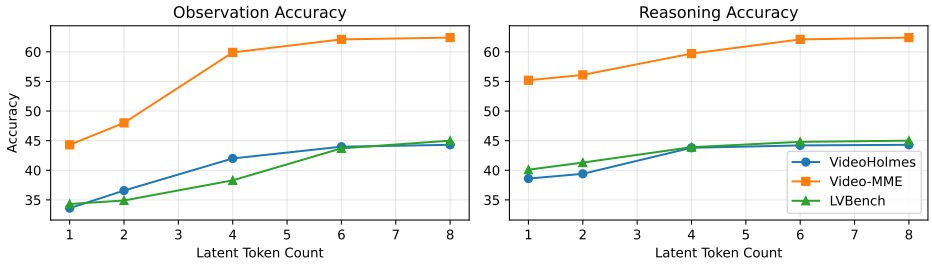


Figure 3: **Left:** Accuracy vs. latent observation token count. **Right:** Accuracy vs. latent reasoning token count. The performance increases with both hyperparameters and plateaus at around 8 tokens.

Table 6: Similarity to teacher trajectories. Answer accuracy (%) and agreement between student and teacher acquisition trajectories on Qwen2.5-VL-7B. For each benchmark, the best result is in **bold**.

Benchmark	Method	Accuracy	Similarity to teacher trajectory		
			Tool Sequence	Temporal IoU	Optional Tools
VideoHolmes	Direct	34.6	54.6	36.2	26.1
	CoT	39.2	57.3	38.2	29.5
	LARVA	**44.3**	**61.9**	**39.4**	**30.7**
Video-MME	Direct	53.4	60.6	30.4	29.0
	CoT	59.9	63.6	30.5	35.9
	LARVA	**62.4**	**64.4**	**32.8**	**40.9**
LVBench	Direct	37.6	61.5	30.1	33.3
	CoT	41.3	63.6	30.6	34.2
	LARVA	**45.0**	**64.9**	**32.6**	**46.1**

A.8 Teacher-Trajectory Similarity

Table 6 compares student trajectories with the teacher using three metrics. **Tool Sequence** is normalized Levenshtein similarity, $1 - d_{\text{Lev}}(s,t)/\max(|s|,|t|)$, where s and t are student and teacher tool-name sequences. **Temporal IoU** is $\mu(A_s \cap A_t)/\mu(A_s \cup A_t)$, where A_s and A_t are the unions of their retrieved temporal intervals and μ denotes duration. Following Yang et al. (2026), **Optional Tool Agreement** measures agreement on using or omitting tool types that appear in some but not all successful reference trajectories for the same question. LARVA outperforms both textual students on all three metrics across all benchmarks. In benchmark order, gains over CoT are 4.6/0.8/1.3 points for Tool Sequence, 1.2/2.3/2.0 for Temporal IoU, and 1.2/5.0/11.9 for Optional Tool Agreement. These results support better transfer of teacher acquisition behavior; the answerer-swap evaluation separately assesses the utility of the acquired evidence.