# VILAMAS: LATENT STATE HANDOFFS FOR MULTI-AGENT LONG-VIDEO UNDERSTANDING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multi-agent systems (MAS) for long-video question answering face a key bottleneck: progressive accuracy degradation across successive agent handoffs. Textual handoffs can omit visual evidence and intermediate reasoning, which may cause additional collaboration stages to undermine earlier understanding and limit the effective use of more elaborate MAS. We introduce ViLaMAS, a training-free framework for reducing this degradation within existing MAS workflows. Agents sharing a compatible frozen backbone transfer their complete accumulated multimodal KV caches and continuation metadata, allowing subsequent agents to continue autoregressive latent reasoning. Evidence-guided revisiting augments the inherited state with newly sampled frames and selected original frames, followed by latent updates that integrate the additional evidence. We also introduce Self-Verification Inference (SVI) as a baseline for long-video question answering. With Qwen3-VL-32B, SVI exceeds the reported accuracy of GPT-4o on EgoSchema and LongVideoBench. SVI with two textual agent handoffs reduces accuracy from 66.0% to 47.5% with Qwen3-VL-8B on LongVideoBench. ViLaMAS raises accuracy to 61.8% within the same collaboration structure. Across six MAS architectures, ViLaMAS reduces this gap by more than 60% on average compared with textual handoffs. These results support preserving and extending multimodal context to retain video understanding accuracy across compatible MAS.

## 1 INTRODUCTION

Long-video question answering requires identifying brief visual details, relating events separated in time, and integrating them into an answer. Context extension, memory mechanisms, and frame selection improve the evidence available for this task (Zhang et al., 2025a; He et al., 2024; Wang et al., 2025). Multi-agent systems (MAS) organize video understanding into cooperating stages of observation, reasoning, and synthesis (Liu et al., 2026b;a). The effectiveness of these workflows depends on whether later agents can retain and use the understanding established by earlier ones. If accuracy deteriorates during collaboration, additional reasoning stages can undermine the visual understanding on which the workflow depends.

We investigate progressive accuracy degradation across agent handoffs as a barrier to applying MAS workflows to video understanding. Our controlled evaluations reveal accuracy degradation across successive textual handoffs (Figure 1). These observations motivate our central question: *how can multimodal understanding be preserved through successive agent handoffs for a multi-agent system?* We take the workflow's role assignments and collaboration structure as given and focus on reducing the performance degradation associated with its communication stages. Textual handoffs may omit visual attributes, temporal relations, or competing interpretations, limiting the context available to subsequent agents. This motivates preserving accumulated multimodal context across handoffs and extending it when further visual evidence is needed. Latent communication offers a foundation: $L^2$-VMAS separates perception and thinking memories (Yu et al., 2026), while MACF learns compact latent messages between video observers and a coordinator (Chen et al., 2026b). We study multimodal state continuation coupled with source-video revisiting within existing workflows.

We introduce **ViLaMAS**, a training-free framework for preserving and extending multimodal reasoning state across compatible agents. Each agent transfers its complete accumulated KV cache and continuation metadata to the receiving agent, which extends the inherited state through autoregres-



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.

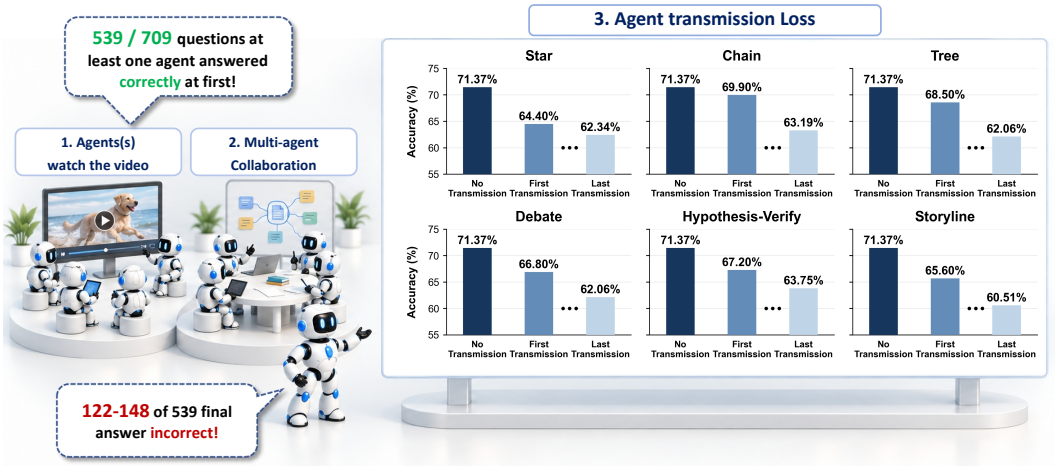


Figure 1: Accuracy across communication stages in six multi-agent architectures. Qwen3-VL-8B is evaluated on 709 Perception Test questions. No Transmission denotes the video-reading observer; First Transmission reports the first-stage mean. Last Transmission reports the Aggregator's accuracy.

sive latent updates. ViLaMAS retains the workflow's role assignments and handoff structure, with state transfer restricted to agents sharing a compatible frozen backbone. To incorporate evidence missing from the inherited context, **evidence-guided revisiting** uses the final agent's current state to formulate a visual question that distinguishes competing answers and localize a relevant video interval. Revisited observations enter the accumulated state and are integrated through further latent updates. We also introduce **Self-Verification Inference (SVI)**, combining independent candidate generation, visual verification, and answer fusion, as a system-level performance reference.

We evaluate ViLaMAS across six MAS architectures and both ViLaMAS and SVI on three benchmarks, with SVI exceeding reported GPT-4o accuracy on multiple benchmarks. ViLaMAS improves accuracy from 47.5% with textual handoffs to 61.8% under the same collaboration structure on LongVideoBench. We also report inference costs alongside accuracy and conduct ablation studies to assess the contribution of each component. Our contributions are summarized as follows:

- We characterize **accuracy degradation across agent handoffs** in different MAS architectures. Our analysis shows that successive textual handoffs can undermine video understanding established by earlier agents.
- We develop **ViLaMAS**, a training-free framework for reducing this degradation within existing MAS workflows. It combines multimodal KV state transfer, autoregressive latent updates, and evidence-guided revisiting to preserve and extend context across agents sharing a compatible frozen backbone.
- We evaluate ViLaMAS on three video benchmarks and assess its components through controlled sequential ablations, using Self-Verification Inference (SVI) as a baseline. With Qwen3-VL-8B on LongVideoBench, textual handoffs reduce accuracy from 66.0% to 47.5%, while ViLaMAS raises it to 61.8% within the same collaboration structure.

## 2 Related Work

### 2.1 Long-Video Understanding

Long-video understanding requires retaining visual context and selecting informative observations under limited input budgets. LongVA transfers extended language context to visual inputs (Zhang et al., 2025a), while MovieChat and MA-LMM retain historical information in memory (Song et al., 2024; He et al., 2024). LongVU reduces visual redundancy through text-guided spatiotemporal compression (Shen et al., 2025). Complementary approaches focus on acquiring relevant evidence. VideoTree and VideoMiner organize video content hierarchically to guide evidence selec-

tion (Wang et al., 2025; Cao et al., 2025). Flexible Frame Selection learns task-dependent sampling policies (Buch et al., 2025), while Adaptive Keyframe Sampling balances relevance and temporal coverage (Tang et al., 2025). Video-RAG supplements visual inputs with retrieved text (Luo et al., 2025), and LongVT revisits relevant clips through tool use (Yang et al., 2026b). Together, these directions motivate reasoning that retains prior context while incorporating additional observations.

### 2.2 Multi-Agent Video Understanding

Agentic video systems organize observation and reasoning through tool use and specialized roles. VideoAgent and VideoMind coordinate tools and reasoning stages (Wang et al., 2024b; Liu et al., 2026b), while MR. Video aggregates clip analyses through MapReduce (Pang & Wang, 2025). VideoChat-M1 learns collaborative policies through multi-agent reinforcement learning (Chen et al., 2026a). LongVideoAgent and Symphony coordinate specialized agents for visual grounding and reasoning (Liu et al., 2026a; Yan et al., 2026). Alongside workflow design, communication mechanisms govern the context transferred between agents. $L^2$-VMAS separates perception and thinking memories (Yu et al., 2026) and LatentMAS supports language-agent collaboration (Zou et al., 2026), while MACF learns compact, query-conditioned messages between video observers and a coordinator (Chen et al., 2026b). ViLaMAS focuses on reducing accuracy degradation across agent handoffs within existing MAS workflows. Its central idea is to preserve multimodal reasoning context while acquiring additional visual evidence, helping agents retain and refine understanding established.

## 3 Method

We introduce **ViLaMAS**, a **training-free framework for latent collaboration in video multi-agent systems**. It combines KV cache transfer with continuous latent updates to let downstream agents continue reasoning from the accumulated multimodal computational context. Evidence-guided revisiting appends visual observations to the accumulated state, allowing reasoning to continue across agents and source-video revisits.

We first define the shared video input and the SVI inference procedure (Section 3.1). We then specify the state exchanged between compatible agents and how receiving agents continue inference from it (Section 3.2). Finally, we describe how this continuation process incorporates observations obtained through video revisiting (Section 3.3).

### 3.1 Self-Verification Inference

Self-Verification inference implements a test-time scaling strategy using a Qwen3-VL backbone $f_\theta$ for independent candidate generation, visual verification, and answer fusion (Figure 2).

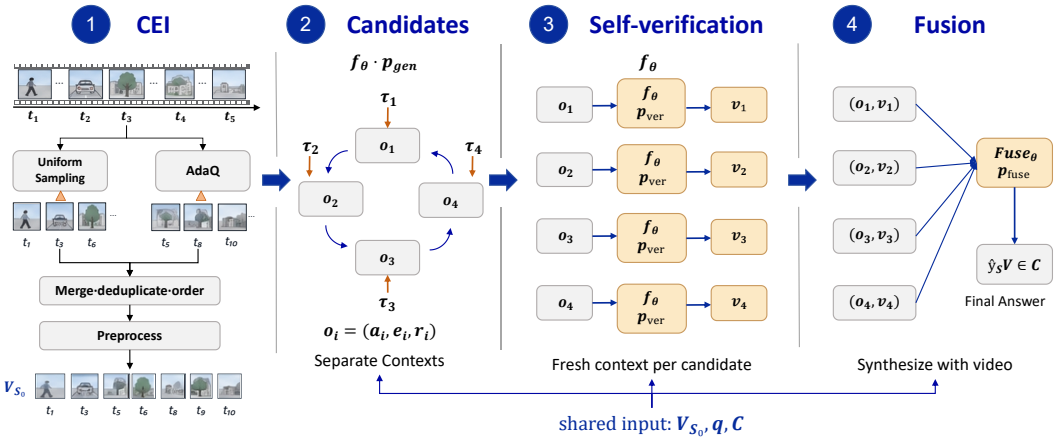


Figure 2: Overview of SVI. All inference stages use the same Qwen3-VL backbone without parameter updates. We set $N = 4$ to run four candidate-generating agents in parallel.

**Complementary Evidence Integration.** Complementary Evidence Integration (CEI) constructs the initial video input by combining uniformly sampled frames for global temporal coverage with question-adaptive frames selected using AdaQ (Zhang et al., 2026). The two frame sets are merged, deduplicated, and ordered chronologically while preserving their source timestamps. The merged frame sequence is then processed using the same model-native video preprocessing pipeline across all configurations and serves as the shared visual input for subsequent inference. We denote this sequence by $V_{\mathcal{S}_0}$, the question by $q$, and its answer options by $\mathcal{C}$.

**Multi-instance inference.** To obtain diverse candidate predictions for subsequent verification, we run $N$ instances in separate contexts using a shared prompt $p_{\text{gen}}$ and independent stochastic decoding seeds $\tau_i$:

$$o_i = f_\theta(V_{\mathcal{S}_0}, q, \mathcal{C}; p_{\text{gen}}, \tau_i), \qquad i = 1, \ldots, N. \tag{1}$$

Each output $o_i$ is represented as $o_i = (a_i, e_i, r_i)$, where $a_i \in \mathcal{C}$ is the predicted answer, $e_i$ denotes the supporting frame or timestamp references, and $r_i$ is a brief explanation.

**Self-Verification and fusion.** Inspired by criterion-decomposed verification (Kwok et al., 2026), we independently verify each candidate in a fresh context using the verification prompt $p_{\text{ver}}$. For candidate $o_i$, the verifier evaluates its visual grounding, answer support, and temporal consistency:

$$\begin{aligned} v_i &= f_\theta(V_{\mathcal{S}_0}, q, \mathcal{C}, o_i; p_{\text{ver}}), \\ \widehat{y}_{\text{SV}} &= \text{Fuse}_\theta\big(V_{\mathcal{S}_0}, q, \mathcal{C}, \{(o_i, v_i)\}_{i=1}^{N}; p_{\text{fuse}}\big)\,. \end{aligned} \tag{2}$$

For fusion, $p_{\text{fuse}}$ instructs the same backbone to synthesize supported observations and check disagreements against the video, without treating repeated claims as independent evidence. The output belongs to $\mathcal{C}$ but need not match an individual proposal.

For controlled sequential evaluation, we append a Reasoner and an Aggregator to SVI fusion, treating the fused answer as an intermediate proposal. In ViLaMAS-Text, the Reasoner assesses event relations and competing interpretations using the candidates, verification feedback, and fused proposal. The Aggregator combines this analysis with the proposal to produce the final answer. ViLaMAS follows the same role sequence while inheriting the fusion instance's complete context, including its processed initial visual context, and subsequent extensions. Candidate and verifier outputs enter fusion as text; their independent caches are not merged. Neither configuration re-prefills the initial video at handoffs. Using SVI as the strong reference, we compare accuracy retention under textual and latent communication through the same downstream roles.

### 3.2 KV CACHE TRANSFER AND AUTOREGRESSIVE LATENT REASONING

We combine KV cache transfer with autoregressive latent updates to preserve multimodal computational context across agent handoffs.

**State transfer and prefix alignment.** All agents share a frozen backbone, system prompt, tool definitions, and prompt template. Role-specific instructions are appended to the accumulated context without modifying the inherited prefix. For a handoff from agent $i$ to agent $j$, we represent the transferred state as $S_i = (\mathcal{M}_i, \pi_i)$, where $\mathcal{M}_i = \{(K_i^{(\ell)}, V_i^{(\ell)})\}_{\ell=1}^{L}$ contains the KV entries at all processed positions across the $L$ decoder layers. The metadata $\pi_i$ records the cache positions and multimodal rotary-position information required for autoregressive continuation. Cache reuse requires matching prefix inputs and compatible attention masks and positional metadata. The handoff operator $\mathcal{H}_{i\to j}$ transfers this state to agent $j$, which then processes its continuation prompt $p_j$ (Figure 3(b)):

$$(h_{j,0}, S_{j,0}) = \text{Prefill}_\theta(p_j; \mathcal{H}_{i\to j}(S_i))\,. \tag{3}$$

Here, $h_{j,0} \in \mathbb{R}^d$ is the final-layer hidden state at the last position of $p_j$, and $S_{j,0}$ contains the inherited cache extended with the KV entries for $p_j$, along with updated continuation metadata.

**Autoregressive latent continuation.** Agent $j$ performs $m_j$ latent updates using the frozen decoder $F_\theta$ (Figure 3(a)). We precompute a fixed ridge alignment $W_a$ to map hidden states into the input embedding space:

$$W_a = (W_{\text{out}}^{\mathsf{T}} W_{\text{out}} + \lambda I)^{-1} W_{\text{out}}^{\mathsf{T}} W_{\text{in}}, \tag{4}$$

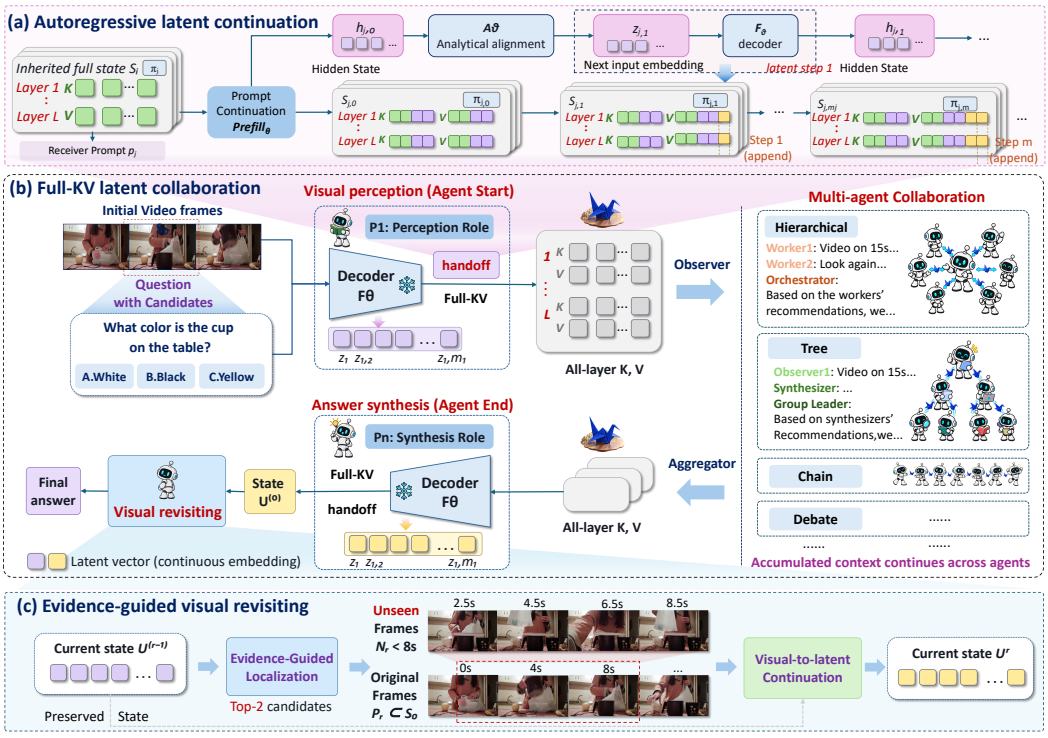


Figure 3: Overview of ViLaMAS. (a) **Autoregressive latent continuation:** aligned hidden states drive frozen-decoder updates that extend the KV cache. (b) **Full-KV latent collaboration:** complete multimodal KV states carry accumulated context across compatible agents within existing MAS workflows. (c) **Evidence-guided visual revisiting:** newly sampled and selected original frames augment the inherited state for further latent reasoning and answer revision.

where $W_{\text{in}}, W_{\text{out}} \in \mathbb{R}^{|\mathcal{V}|\times d}$ are the pretrained input embedding and output projection matrices, respectively, $\lambda$ controls ridge regularization, and $I$ is the $d$-dimensional identity matrix.

At each step, the aligned hidden state is rescaled using the mean input-embedding norm $\bar{\rho}$ and fed back into the decoder:

$$(h_{j,k}, S_{j,k}) = F_\theta\left(\bar{\rho}\,\frac{h_{j,k-1}W_a}{\max(\|h_{j,k-1}W_a\|_2, \epsilon)}; S_{j,k-1}\right), \tag{5}$$

for $k = 1, \ldots, m_j$, where $\epsilon$ ensures numerical stability. We set $\lambda = 10^{-5}$ and $\epsilon = 10^{-6}$. Each update appends layer-wise KV entries and updates the continuation metadata. Both the backbone and the alignment remain fixed, requiring no task-specific optimization. Conditions for exact computational equivalence to uninterrupted continuation are detailed in Appendix A.

**Instantiation and applicability.** Our sequential instantiation uses Observer, Reasoner, and Aggregator roles. The interface also applies to individual handoff links in other topologies with compatible backbones and state representations. Textual and KV-only controls retain the same roles and connections, using textual reports or KV transfer without latent updates, respectively.

### 3.3 Latent Visual Reasoning with Evidence-Guided Revisiting

Evidence-guided revisiting complements context preservation by acquiring additional visual evidence. The final agent integrates revisited observations into its inherited state and continues latent reasoning before revising its answer.

**Evidence-guided localization.** The final agent formulates a visual question to distinguish its two highest-scoring answers. Question generation and hierarchical interval routing use disposable state

copies, leaving the main reasoning state unchanged. Routing selects an interval $I_r$ of at most eight seconds containing at least four previously unseen frames (Figure 3(c)).

**Combined visual observation.** Each revisit combines four new frames from $I_r$ with four original frames selected from $\mathcal{S}_0$. New frames are evenly sampled from indices not previously observed along the current reasoning path. Original frames are selected by temporal proximity to $I_r$ and may lie outside it. The combined frame indices $\mathcal{R}_r$ define a chronologically ordered native video input $V_{\mathcal{R}_r}$ that preserves source timestamps. This input supplies new evidence alongside earlier observations for renewed inspection.

**Visual-to-latent continuation.** Let $U = (h, S)$ denote the reasoning state and $\mathcal{L}_\theta^m(U)$ denote $m$ updates from Equation (5). Starting from the final agent's post-collaboration state $U^{(0)}$, each successful revisit updates the state and prediction:

$$\begin{aligned} U^{(r)} &= \mathcal{L}_\theta^{m_r}\Big(\mathrm{AppendVideo}_\theta\big(U^{(r-1)}, V_{\mathcal{R}_r}\big)\Big), \\ \widehat{y}^{(r)} &= \underset{j\in\{1,\dots,K\}}{\arg\max}\ s_\theta(j \mid U^{(r)}). \end{aligned} \tag{6}$$

Here, $\mathrm{AppendVideo}_\theta$ processes the combined video input using the inherited cache, appends the resulting KV entries, and returns the updated reasoning state, including the new hidden state and continuation metadata. The subsequent $m_r$ latent updates integrate the revisited observations before answer selection using the constrained answer-letter score $s_\theta$, where $K = |\mathcal{C}|$.

# 4 Experiments

## 4.1 Experimental Setup

**Tasks and Datasets.** We evaluate ViLaMAS and Self-Verification Inference on three video understanding benchmarks: (i) Egocentric Video Understanding, including EgoSchema (Mangalam et al., 2023); (ii) Long-Video Understanding, including LongVideoBench (Wu et al., 2024); and (iii) Video Understanding across Different Durations, using Video-MME (Fu et al., 2025). We report accuracy on EgoSchema and LongVideoBench. For Video-MME, we report accuracy on short, medium, and long videos, together with overall accuracy, under both subtitle-free and subtitle-assisted settings. We additionally use Perception Test (Patraucean et al., 2023) for the cross-architecture analysis.

**Models and Baselines.** We instantiate ViLaMAS and Self-Verification Inference with Qwen3-VL-8B and Qwen3-VL-32B backbones (Bai et al., 2025) as shown in Table 1. Comparisons cover: (i) off-the-shelf video MLLMs, including GPT-4o, InternVL-2.5, and LLaVA-Video; (ii) single and multi-agent methods, including VideoAgent, SVAgent, Video-RAG, MapReduce, and MACF; and (iii) controlled variants, including ViLaMAS-Text, which communicates through textual handoffs, and ViLaMAS, which transfers accumulated KV caches with latent updates and evidence-guided revisiting. Each MAS architecture must include three functional roles: **Observer** extracts question-relevant visual evidence; **Reasoner** analyzes event relations and evaluates competing interpretations; **Aggregator** integrates intermediate assessments into the final answer. Architectures differ in their interaction structures, while role assignments and connections remain fixed within each architecture when comparing textual and latent communication.

**Implementation details.** We evaluate ViLaMAS across six multi-agent architectures, comparing it with textual handoffs under each architecture's existing role assignments and collaboration structure on Perception Test. CEI combines 64 AdaQ-selected frames with 32 uniformly sampled frames. Revisiting is limited to two rounds, with the first localization opportunity provided uniformly to revisit-enabled configurations. After the first successful revisit, the agent may stop, inspect another event, or examine the previous interval more densely. Each generated textual report is limited to 64 tokens. In ViLaMAS, each of the $H \geq 2$ receiving roles per path performs $K_s = 10$ latent updates. Latent update budgets and the six MAS architectures are detailed in Appendices A.2 and A.4, respectively. All inference experiments are conducted on 4 × NVIDIA 80 GB H100 GPUs.

Table 1: Comparison with other methods on three benchmarks. Scores are accuracy (%). Paired Video-MME scores are reported without / with subtitles; single values retain the reporting format of their source. * Qwen3-VL-8B backbone; † Qwen3-VL-32B backbone.

| Model | EgoSchema | LongVideoBench | Video-MME | | | |
|---|---|---|---|---|---|---|
| | | | Short | Medium | Long | Average |
| GPT-4o (OpenAI, 2024) | 72.2 | 66.7 | 80.0 / 82.8 | 70.3 / 76.6 | 65.3 / 72.1 | 71.9 / **77.2** |
| ShareGPT4Video-8B (Chen et al., 2024) | – | 39.7 | 48.3 / 53.6 | 36.3 / 39.3 | 35.0 / 37.9 | 39.9 / 43.6 |
| VideoChat2-7B (Li et al., 2024) | 56.7 | 39.3 | 48.3 / 52.8 | 37.0 / 39.4 | 33.2 / 39.2 | 39.5 / 43.8 |
| InternVL-2.5-8B (Chen et al., 2025) | – | 60.0 | – | – | – | 64.2 / 66.9 |
| Qwen2-VL-7B (Wang et al., 2024a) | 66.7 | 55.6 | – | – | – | 63.3 / 69.0 |
| LongVA-7B (Zhang et al., 2025a) | – | 51.3 | 61.1 / 61.6 | 50.4 / 53.6 | 46.2 / 47.6 | 52.6 / 54.3 |
| LLaVA-Video-7B (Zhang et al., 2025b) | 57.3 | 58.2 | – | – | – | 63.3 / 69.7 |
| Aria-28B (Li et al., 2025) | – | 64.2 | 76.9 / 78.3 | 67.0 / 71.7 | 58.8 / 66.3 | 67.6 / 72.1 |
| Oryx-34B (Liu et al., 2025) | – | – | 77.3 / 80.6 | 65.3 / 74.3 | 59.3 / 69.9 | 67.3 / 74.9 |
| LLaVA-Video-72B (Zhang et al., 2025b) | 65.6 | 64.9 | 81.4 / 82.8 | 68.9 / 75.6 | 61.5 / **72.5** | 70.6 / 77.0 |
| InternVL-2.5-78B (Chen et al., 2025) | – | 63.6 | **82.8** / 83.2 | 70.9 / 74.1 | 62.6 / 64.8 | 72.1 / 74.0 |
| VideoLLaMA-2-72B (Cheng et al., 2024) | 63.9 | – | 69.8 / 72.0 | 59.9 / 63.0 | 57.6 / 59.0 | 62.4 / 64.7 |
| VideoAgent (Wang et al., 2024b) | 54.1 | – | – | – | – | – |
| SVAgent* (Yang et al., 2026a) | 68.2 | 61.0 | 73.9 / 78.0 | 61.5 / 71.1 | 53.9 / 60.5 | 63.1 / 69.8 |
| Video-RAG-72B (Luo et al., 2025) | – | 65.4 | 81.1 / – | **72.9** / – | **73.1** / – | **75.7** / – |
| MapReduce* (Pang & Wang, 2025) | 40.8 | 38.2 | – | – | – | 46.7 |
| LatentMAS* (Zou et al., 2026) | – | 48.5 | – | – | – | 55.7 |
| MACF* (Chen et al., 2026b) | 53.7 | 56.8 | – | – | – | 60.4 |
| **ViLaMAS-Text*** | 65.3 | 47.5 | 62.8 / 66.2 | 49.3 / 55.4 | 45.6 / 48.9 | 52.6 / 56.9 |
| **ViLaMAS*** | 72.0 | 61.8 | 73.9 / 77.2 | 61.7 / 66.3 | 54.1 / 59.9 | 63.2 / 67.8 |
| **Self-Verification Inference*** | 72.3 | 66.0 | 78.0 / 81.9 | 67.3 / 73.3 | 58.9 / 64.9 | 68.1 / 73.4 |
| **Self-Verification Inference†** | **76.8** | **70.2** | 81.2 / **83.6** | 69.1 / **76.9** | 62.2 / 71.1 | 70.9 / **77.2** |

## 4.2 Results

We first quantify accuracy loss following textual handoffs from SVI and the recovery achieved by ViLaMAS (Section 4.2.1). We then assess accuracy across six MAS architectures and report cost (Section 4.2.2). Next, we present the ablation studies and additional experimental analyses (Section 4.2.3). Finally, we discuss the limitations of ViLaMAS (Section 4.2.4).

### 4.2.1 Q1: How do agent handoffs affect accuracy, and how much does ViLaMAS recover?

SVI provided a strong performance reference before additional handoffs (Table 1). The results also show that strong upstream predictions do not necessarily translate into strong final system performance. The 32B configuration achieved 76.8% on EgoSchema and 70.2% on LongVideoBench, exceeding reported GPT-4o accuracy by 4.6 and 3.5 percentage points, respectively. Specialized video workflows use complementary grounding, perception, and verification roles (Liu et al., 2026a; Yan et al., 2026). We use the subsequent Reasoner and Aggregator stages as a controlled setting for studying multimodal state preservation across role-specific handoffs. With the 8B backbone on LongVideoBench, two additional textual handoffs reduced accuracy from 66.0% to 47.5%, yielding ViLaMAS-Text. ViLaMAS raised accuracy to 61.8% within the same collaboration structure, narrowing the gap to SVI from 18.5 to 4.2 points. The other 8B comparisons also showed recovery. On EgoSchema, ViLaMAS improved accuracy from 65.3% to 72.0%, numerically close to SVI's 72.3%. On Video-MME, average accuracy improved over ViLaMAS-Text by 10.6 and 11.0 percentage points without and with subtitles. These results show consistent gains over textual handoffs across all benchmarks, although the extent of degradation and recovery varies by benchmark.

### 4.2.2 Q2: How effective is ViLaMAS across different MAS architectures, and what is its computational cost?

Within each architecture, textual and latent variants share the same roles and predecessor dependencies. At multi-predecessor nodes, independently computed branches are combined by retaining the shared prefix once and concatenating branch-specific KV suffixes in a fixed order with unchanged

positional indices. ViLaMAS narrowed the accuracy gap to original multi-observers in all six architectures on Perception Test (Figure 4). These improvements support applicability beyond a single sequential workflow, although the extent of recovery varied across architectures. Generation and memory measurements showed fewer generated text tokens and we also evaluated sensitivity to the textual report cap (Table 2).

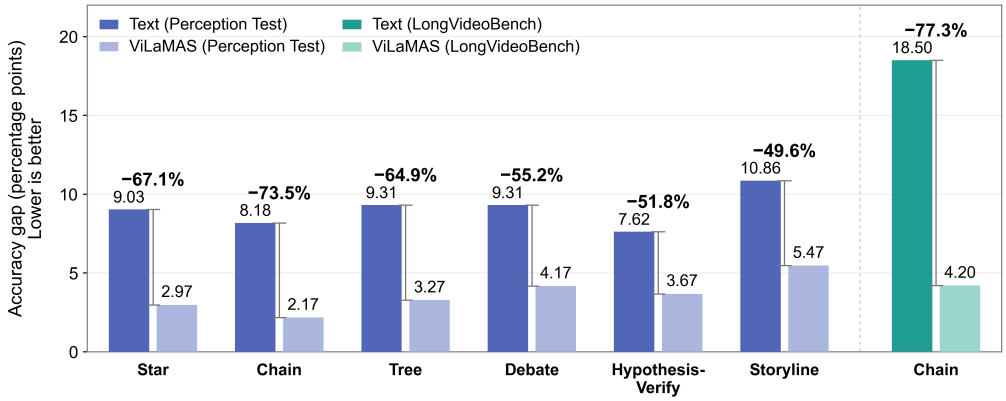


Figure 4: Accuracy gaps for ViLaMAS vs Text-handoff, using Qwen3-VL-8B. Accuracy is evaluated across six multi-agent architectures on Perception Test and Chain (SVI) on LongVideoBench.

Table 2: Generation and memory costs (left) and report-cap sensitivity on a sampled Video-MME subset using Qwen3-VL-8B (right). Mean gen. tokens reports the mean number of intermediate text tokens generated per question during the added collaboration stages and revisiting, including textual reports and planning or control outputs. Cap-hit denotes reaching the report cap without EOS.

| Method | Mean gen. tokens | Peak GPU memory (GiB) | | |
|---|---|---|---|---|
| | | Mean | P95 | Max |
| ViLaMAS-Text | 116.7 | 40.98 | 45.12 | 52.32 |
| w/ revisiting | 165.7 | 40.98 | 45.12 | 52.32 |
| ViLaMAS | 54.3 | 42.83 | 45.65 | 57.68 |
| w/o revisiting | 0.0 | 41.51 | 45.27 | 56.63 |

| Report cap (tokens) | 64 | 128 | 256 |
|---|---|---|---|
| Short (%) | 68.66 | 70.15 | 70.15 |
| Medium (%) | 59.70 | 58.21 | 58.21 |
| Long (%) | 59.09 | 59.09 | 59.09 |
| Tokens | 55.52 | 68.58 | 69.92 |
| Cap-hit (%) | 47.50 | 4.50 | 0.25 |

#### 4.2.3 Q3: How do individual components and handoff configurations affect accuracy?

Table 3 reports the sequential ablations on LongVideoBench. With the 8B backbone, adding KV transfer increased accuracy from 47.5% to 52.2%, and subsequent latent updates raised it to 58.9%. Evidence-guided revisiting further increased accuracy to 61.8% and 64.1% for the 8B and 32B backbones, respectively. These results indicate that additional reasoning and visual evidence can partially compensate for degradation, while preserving context across agent handoffs remains important for accuracy recovery. Text accuracy progressively declined from 76% at $H = 0$ to 66% at $H = 6$ (Fig. 6). Figure 5 shows that our method outperforms textual handoffs with varied report content and context available to receivers. These results support multimodal state transfer for accuracy retention, with latent updates providing a further gain. To examine answer correction and regression for EGL, we compared original-frame rereading, random eligible intervals, original-question localization, and EGL in Appendix B.4.

#### 4.2.4 Q4: What are the boundaries and limitations of ViLaMAS?

ViLaMAS reduced accuracy loss but did not fully recover SVI performance (Table 3). The remaining degradation may involve changes in accessible evidence, role-specific processing, or answer

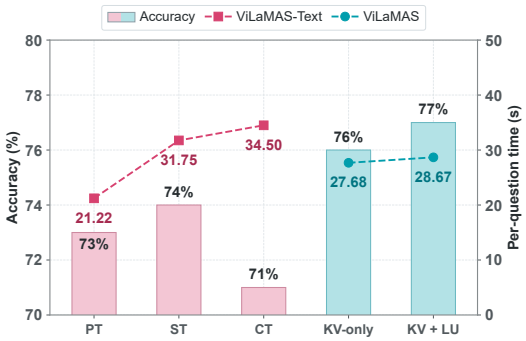


Figure 5: Accuracy (bars) and per-question time (lines) across handoff configurations on a VideoMME subset. PT: Plain text; ST: structured text; CT: Cumulative text with original video.

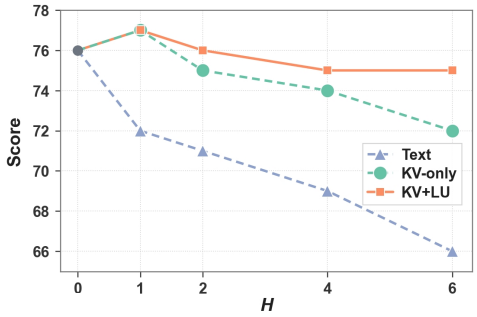


Figure 6: Effect of handoff depth on VideoMME subset. $H$ denotes the number of handoffs, with $H = 0$ representing the shared SVI starting point.

Table 3: ViLaMAS ablations and post-SVI controls. $A_1$ and $A_2$ denote the post-SVI Reasoner and Aggregator roles. Full-KV denotes complete accumulated state retention through local cache reuse or explicit state transfer. LU: latent updates; EGL: latent visual reasoning with evidence-guided revisiting. Single-instance controls use matched role prompts or one generic role, with matched LU budgets and revisiting opportunities. For cumulative ablations from SVI through full ViLaMAS, $\Delta$ is the change from the preceding row in percentage points. The two ViLaMAS-Text controls use ViLaMAS-Text as their reference; the two single-instance controls use full ViLaMAS.

| **Configuration** | **Post-SVI roles** | | **Components** | | | **LongVideoBench** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $A_1$ | $A_2$ | Full-KV | LU | EGL | 8B | $\Delta$ | 32B | $\Delta$ |
| Self-Verification Inference | | | | | | 66.0 | – | 70.2 | – |
| + Text-only agent 1 | ✓ | | | | | 55.7 | −10.3 | 60.4 | −9.8 |
| **+ Text-only agent 2 (ViLaMAS-Text)** | ✓ | ✓ | | | | 47.5 | −8.2 | 51.4 | −9.0 |
| + KV transfer | ✓ | ✓ | ✓ | | | 52.2 | +4.7 | 55.9 | +4.5 |
| + Latent updates | ✓ | ✓ | ✓ | ✓ | | 58.9 | +6.7 | 61.0 | +5.1 |
| **+ EGL (ViLaMAS)** | ✓ | ✓ | ✓ | ✓ | ✓ | 61.8 | +2.9 | 64.1 | +3.1 |
| ViLaMAS-Text + LU only | ✓ | ✓ | | ✓ | | 47.7 | +0.2 | 52.1 | +0.7 |
| ViLaMAS-Text + LU + EGL | ✓ | ✓ | | ✓ | ✓ | 51.8 | +4.3 | 56.3 | +4.9 |
| Single instance (matched roles) | ✓ | ✓ | ✓ | ✓ | ✓ | 61.4 | −0.4 | 63.6 | −0.5 |
| Single instance (generic role) | | | ✓ | ✓ | ✓ | 60.4 | −1.4 | 62.8 | −1.3 |

readout. And our claim is limited to better preserving video question-answering accuracy relative to textual handoffs within the evaluated MAS workflows. The current transfer interface's applicability to heterogeneous agents has not been established. Complete cache inheritance also retains the accumulated context as additional observations and latent reasoning steps are appended, increasing memory requirements over longer trajectories. Effective cache management for more extended collaboration therefore remains an important direction for further investigation.

## 5 Conclusion

We present ViLaMAS, a training-free framework for reducing accuracy degradation across agent handoffs in video MAS. It combines KV cache transfer, autoregressive latent updates, and evidence-guided revisiting to preserve and extend multimodal context within existing workflows. Experiments across six multi-agent architectures show substantially reduced accuracy loss versus textual handoffs. We also introduce Self-Verification Inference, which achieves strong performance across video benchmarks. Future work could explore state transfer between heterogeneous agents and cache management over longer trajectories for multi-agent systems.

## AI USE STATEMENT

We used generative AI tools for drafting and editing limited portions of this manuscript. We did not use generative AI tools for generating synthetic datasets, processing datasets, or interpreting results. All experimental results and manuscript content were independently verified by at least two authors and checked for potential plagiarism. All AI-assisted outputs were reviewed and verified by the authors. The authors take full responsibility for the final manuscript.

## 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. *arXiv preprint arXiv:2511.21631*, 2025.

Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29071–29082, 2025.

Xinye Cao, Hongcan Guo, Jiawen Qian, Guoshun Nan, Chao Wang, Yuqi Pan, Tianhao Hou, Xiaojuan Wang, and Yutong Gao. VideoMiner: Iteratively grounding key frames of hour-long videos via tree-based group relative policy optimization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23773–23783, 2025.

Boyu Chen, Zikang Wang, Zhengrong Yue, Kainan Yan, Chenyun Yu, Yi Huang, Zijun Liu, Yafei Wen, Xiaoxin Chen, Yang Liu, Peng Li, and Yali Wang. VideoChat-M1: Collaborative policy planning for video understanding via multi-agent reinforcement learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33772–33783, 2026a.

Kerui Chen, Jinglu Wang, Jianrong Zhang, Ming Li, Yan Lu, and Hehe Fan. Scaling video understanding via compact latent multi-agent collaboration. *arXiv preprint arXiv:2605.00444*, 2026b.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. ShareGPT4Video: Improving video understanding and generation with better captions. In *Advances in Neural Information Processing Systems*, pp. 19472–19495, 2024.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2025.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. 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)*, pp. 24108–24118, 2025.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13504–13514, 2024.

Jacky Kwok, Shulu Li, Pranav Atreya, Yuejiang Liu, Yixing Jiang, Chelsea Finn, Marco Pavone, Ion Stoica, and Azalia Mirhoseini. LLM-as-a-Verifier: A general-purpose verification framework. *arXiv preprint arXiv:2607.05391*, 2026.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, Chongyan Zhu, Xiaoyi Ren, Chao Li, Yifan Ye, Peng Liu, Lihuan Zhang, Hanshu Yan, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2025.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. LongVideoAgent: Multi-agent reasoning with long videos. In *Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 40404–40416, 2026a.

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.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx MLLM: On-demand spatial-temporal understanding at arbitrary resolution. In *International Conference on Learning Representations (ICLR)*, 2025.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-RAG: Visually-aligned retrieval-augmented long video comprehension. In *Advances in Neural Information Processing Systems*, 2025.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, pp. 46212–46244, 2023.

OpenAI. GPT-4o system card. Technical report, 2024.

Ziqi Pang and Yu-Xiong Wang. MR. Video: MapReduce as an effective principle for long video understanding. In *Advances in Neural Information Processing Systems*, pp. 6640–6672, 2025.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception Test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, pp. 54582–54599, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From dense token to sparse memory for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232, 2024.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision (ECCV)*, 2024b.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3283, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, pp. 28828–28857, 2024.

Haiyang Yan, Hongyun Zhou, Peng Xu, Xiaoxue Feng, and Mengyi Liu. Symphony: A cognitively-inspired multi-agent system for long-video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24031–24041, 2026.

Zhongyu Yang, Zuhao Yang, Shuo Zhan, Tan Yue, Wei Pang, and Yingfang Yuan. SVAgent: Storyline-guided long video understanding via cross-modal multi-agent collaboration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24062–24072, 2026a.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. LongVT: Incentivizing "thinking with long videos" via native tool calling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33816–33826, 2026b.

Xinlei Yu, Chengming Xu, Zhangquan Chen, Bo Yin, Cheng Yang, Yongbo He, Yihao Hu, Jiangning Zhang, Cheng Tan, Xiaobin Hu, and Shuicheng Yan. Dual latent memory for visual multi-agent system. In *International Conference on Machine Learning (ICML)*, 2026.

Kun Zhang, Chenxin Fang, Tao Chen, Baiyang Song, Yunhang Shen, Yiyi Zhou, and Rongrong Ji. Towards fast and effective long video understanding of multimodal large language models via adaptive quasi-gaussian sampling. *arXiv preprint arXiv:2606.24187*, 2026.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Transactions on Machine Learning Research*, 2025a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025b.

Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent collaboration in multi-agent systems. In *International Conference on Machine Learning (ICML)*, 2026.

# APPENDIX

This appendix provides additional method details, experimental settings, and reporting conventions for ViLaMAS. The code is available in the supplementary materials. The contents below link to the corresponding sections.

**Contents** **Page**


**A.** Model 13
- A.1 Inference Configurations 13
- A.2 State Transfer and Latent Continuation 14
- A.3 Evidence-guided Revisiting 15
- A.4 Multi-agent Architectures Workflows 16

**B.** Experiments 17
- B.1 Benchmarks and Settings 17
- B.2 Observation and Computational Budgets 18
- B.3 Gap Reduction 19
- B.4 Additional Analysis of Evidence-guided Revisiting 19

**C.** Miscellaneous 21
- C.1 Theoretical Justification for Equations 21
- C.2 Prompt Template 23


# A MODEL

## A.1 INFERENCE CONFIGURATIONS

**Backbones and initial observations.** The main experiments use frozen Qwen3-VL-8B and Qwen3-VL-32B backbones on four NVIDIA H100 GPUs with 80 GB memory each. Both backbones use $N = 4$ independent candidate-generation instances, with GPU allocation determined separately for each backbone. Complementary Evidence Integration (CEI) combines 64 question-adaptive frames selected with AdaQ and 32 uniformly sampled frames. Their union is deduplicated and ordered by source timestamp. Let $\mathcal{S}_0$ denote the resulting source indices and $n_0 = |\mathcal{S}_0| \leq 96$ their number. All matched configurations use the same initial frame sequence and native video preprocessing. Video-MME includes evaluations with and without official subtitles; the subtitle condition is specified for each reported comparison.

**Self-Verification Inference.** SVI runs $N$ candidate-generation instances with the same input and generation instruction, using independent sampling seeds. Each candidate contains an answer, supporting frame or timestamp references, and an explanation. The same frozen backbone then verifies each candidate in a fresh context. One fusion instance receives the initial video, question, answer options, candidates, and verification feedback, and produces the SVI prediction. These stages comprise four candidate-generation instances, four verification calls, and one fusion call. Candidate generation does not exchange hidden states between instances. Independent sampling does not guarantee independent errors.

**Controlled sequential extension.** The Reasoner and Aggregator follow the completed SVI fusion stage. The SVI answer becomes an intermediate proposal, which either added role may retain. In ViLaMAS-Text, the Reasoner receives the candidates, verification feedback, and fused answer. The Aggregator receives the Reasoner's textual analysis and the fused answer. Neither added role rereads the initial video during the handoffs. Revisiting, when enabled, is a separate source-video inspection stage. The KV-based conditions retain these roles and their order. The Reasoner inherits the fusion instance's accumulated multimodal state; the Aggregator then inherits the Reasoner's updated state. Thus, *complete accumulated state* refers to the state of this inherited trajectory. It does not denote a merge of the four candidate instances' internal caches or the verification instances' caches. Their generated outputs enter the fusion context as text.

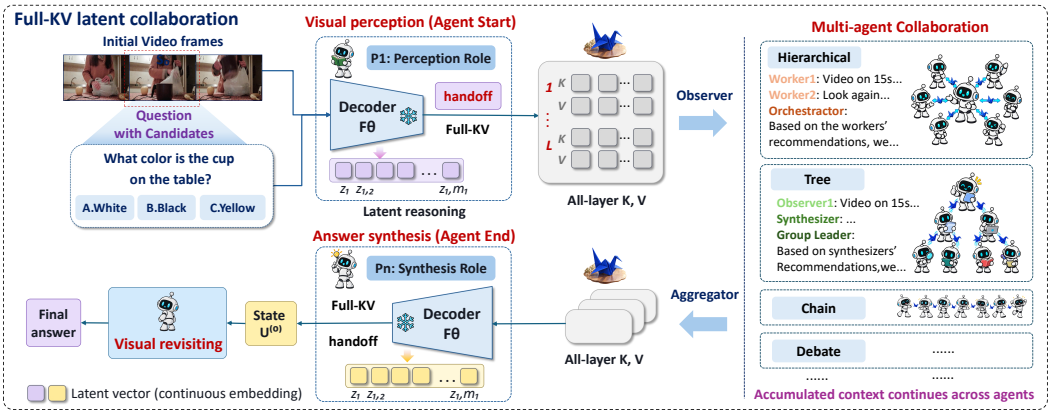


Figure 7: Full-KV latent collaboration.

## A.2 State Transfer and Latent Continuation

### A.2.1 Transferred State and Receiver Initialization

As shown in Figure 7, for a scheduled handoff from agent $i$ to agent $j$, the transferred state is

$$S_i = (\mathcal{M}_i, \pi_i), \qquad \mathcal{M}_i = \{(K_i^{(\ell)}, V_i^{(\ell)})\}_{\ell=1}^{L}. \tag{A.1}$$

The cache includes all processed positions in the sender's current context at every decoder layer. The continuation metadata $\pi_i$ includes the next cache position and multimodal rotary-position information. Any additional state affecting the decoder transition must be restored or reconstructed consistently. Compatible agents share frozen weights, cache layouts, attention conventions, system instructions, tool definitions, and the prompt template. Role-specific instructions are appended without modifying the inherited prefix. Shared-prefix reuse requires matching token and visual inputs, attention masks, and positional metadata. Matching prefix length alone is insufficient.

The receiver obtains its initial hidden state by processing its own continuation prompt $p_j$:

$$(h_{j,0}, S_{j,0}) = \mathrm{Prefill}_\theta\big(p_j; \mathcal{H}_{i\to j}(S_i)\big). \tag{A.2}$$

The sender's final hidden state is not directly substituted for $h_{j,0}$. The receiver continues from the restored cache and its newly processed prompt.

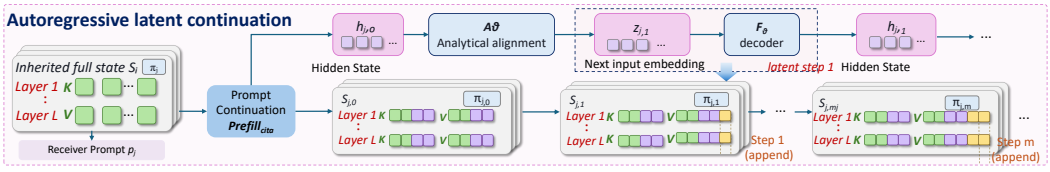


Figure 8: Autoregressive latent continuation.

### A.2.2 Alignment and the Continuation Property

As shown in Figure 8, for pretrained input and output embedding matrices $W_{\text{in}}, W_{\text{out}} \in \mathbb{R}^{|\mathcal{V}|\times d}$, define

$$\begin{aligned} W_a &= (W_{\text{out}}^{\mathsf{T}} W_{\text{out}} + \lambda I)^{-1} W_{\text{out}}^{\mathsf{T}} W_{\text{in}}, \\ \mathcal{A}_\theta(h) &= \bar{\rho}\,\frac{hW_a}{\max(\|hW_a\|_2, \epsilon)}, \\ (h_{j,k}, S_{j,k}) &= F_\theta\big(\mathcal{A}_\theta(h_{j,k-1}); S_{j,k-1}\big). \end{aligned} \tag{A.3}$$

Here, $\bar{\rho}$ is the mean pretrained input-embedding norm, $\lambda = 10^{-5}$, and $\epsilon = 10^{-6}$. The alignment is computed from frozen pretrained weights without fitting a task-specific communication adapter. Each update appends layer-wise KV entries.

**Continuation property.** Consider uninterrupted execution and execution restored at one handoff. Assume identical weights, cached tensors, continuation metadata, attention rules, receiver prompt, and subsequent inputs. For a deterministic decoder in exact arithmetic,

$$(h_{j,k}^{\text{restored}}, S_{j,k}^{\text{restored}}) = (h_{j,k}^{\text{uninterrupted}}, S_{j,k}^{\text{uninterrupted}}) \qquad \text{for all } k \geq 0. \tag{A.4}$$

*Proof.* Identical restored tensors, positional metadata, masks, and prompt embeddings produce identical receiver-prefill outputs, establishing the base case $k = 0$. Suppose the hidden state and continuation state agree after step $k$. The fixed mapping $\mathcal{A}_\theta$ produces identical next embeddings. The deterministic decoder therefore returns identical hidden states and appended cache entries at step $k + 1$. Induction establishes the result. The same argument applies to visual appends when their processed embeddings, auxiliary visual features, and positional inputs are identical. This property concerns restoration of an individual trajectory. It does not establish equivalence for merging independently computed branch states, equivalence to SVI, or an accuracy benefit from additional reasoning. Finite-precision implementations may also differ because of numerical rounding or kernel choices.

### A.2.3 Latent Update Budgets and Numerical Settings

Each of the $H \geq 2$ receiving roles on an evaluated path performs $K_s = 10$ latent updates. Each successful revisit adds $m_r = 4$ further updates. The path-level latent budget is

$$B_{\text{latent}} = HK_s + \sum_{r=1}^{R} m_r = 10H + 4R, \qquad R \leq 2. \tag{A.5}$$

For the two-handoff configuration, $H = 2$, giving 20 updates before revisiting and at most 28 updates after two successful revisits. Latent updates are decoder steps and are not counted as generated text tokens.

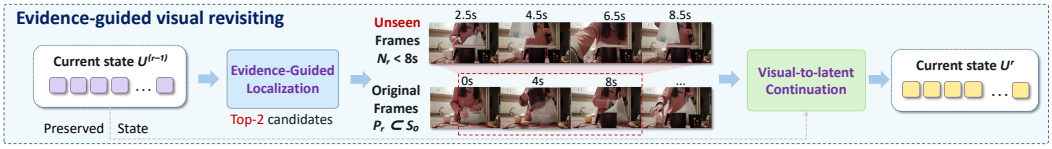


Figure 9: Evidence-guided visual revisiting.

## A.3 Evidence-guided Revisiting

### A.3.1 Planning and Hierarchical Localization

The final agent formulates a visual question to distinguish its two highest-scoring answer candidates. Planning questions are limited to 96 generated tokens. Question generation and interval routing operate on disposable copies of the current state. Generated planning text therefore does not enter the main trajectory as an observed fact. Hierarchical routing considers eight candidate intervals at each level, using at most four levels. Each candidate supplies eight uniformly spaced source indices and is eligible only when at least four are unseen along the current reasoning path. The selected interval $I_r$ spans at most eight seconds. The first localization attempt is mandatory, and failure to select an eligible interval ends revisiting with the latest prediction.

### A.3.2 Combining New and Original Frames

Let $u_0 < \cdots < u_{m-1}$ be the eligible unseen source indices in the selected interval, where $m \geq 4$. Four new frames are selected at zero-based positions

$$j_i = \left\lfloor \frac{(m-1)i}{3} \right\rfloor, \qquad i \in \{0, 1, 2, 3\}. \tag{A.6}$$

All indices already observed on the current path are excluded from this pool. The selected frames are combined with four frames from the initial CEI input.

For interval $I_r = [a_r, b_r]$, define its center $c_r = (a_r + b_r)/2$ and the distance of source timestamp $t$ from the interval as

$$d(t, I_r) = \max\{a_r - t,\, 0,\, t - b_r\}. \tag{A.7}$$

Original frames are ranked lexicographically by $d(t, I_r)$, $|t - c_r|$, and source index. The first four are reused with their original RGB values. All eight frames are ordered by source timestamp and supplied as one native video input. The duration constraint applies only to the new-frame interval; selected original frames may lie outside it.

### A.3.3 State Update and Action Rules

The native visual input is processed using the inherited cache. The resulting KV entries augment the accumulated state, after which $m_r$ latent updates integrate the additional evidence. The updated prediction is read out using constrained answer-letter scores. This operation preserves the main trajectory while extending its available observations. After the first successful revisit and latent update, the final agent chooses one of the actions in Table 4. Each additional interval must satisfy the unseen-frame eligibility rule. At most two revisits are completed; any localization failure retains the most recent prediction.

Table 4: Actions after the first successful revisit. The preceding interval is the one selected in the previous completed round.

| Action | Effect or interval constraint |
|---|---|
| `STOP` | Return the latest prediction without another revisit. |
| `OTHER_EVENT` | Select an interval whose overlap with the preceding interval is at most half the new interval's width. |
| `DENSIFY` | Select an interval contained in the preceding interval and having half its width. |

## A.4 Multi-agent Architectures Workflows

The cross-architecture evaluation uses Star, Chain, Tree, Debate, Hypothesis–Verify, and Storyline workflows on 709 Perception Test questions with Qwen3-VL-8B. Each workflow includes Observer, Reasoner, and Aggregator functions.

All text-report baselines start from four fixed observer reports $P = [p_0, p_1, p_2, p_3]$. Each intermediate call $\mathsf{G}_B(\phi, \cdot; Q)$ receives the question and specified predecessor reports, and outputs a report capped at $B$ tokens. $\mathsf{A}$ denotes the final Aggregator.

**Algorithm 1** Star.

$s \leftarrow \mathsf{G}_B(\phi_{\text{star}}, P; Q)$
$\hat{y} \leftarrow \mathsf{A}([s]; Q)$
**return** $\hat{y}$

**Algorithm 2** Chain.

$s_0 \leftarrow p_0$
**for** $i = 1, \ldots, 3$ **do**
    $s_i \leftarrow \mathsf{G}_B(\phi_{\text{merge}}, [s_{i-1}, p_i]; Q)$
**end for**
$\hat{y} \leftarrow \mathsf{A}([s_3]; Q)$
**return** $\hat{y}$

**Algorithm 3** Tree.

$u_{01} \leftarrow \mathsf{G}_B(\phi_{\text{pair}}, [p_0, p_1]; Q)$
$u_{23} \leftarrow \mathsf{G}_B(\phi_{\text{pair}}, [p_2, p_3]; Q)$
$\hat{y} \leftarrow \mathsf{A}([u_{01}, u_{23}]; Q)$
**return** $\hat{y}$

**Algorithm 4** Debate.

**for** $i = 0, \ldots, 3$ **do**
    $d_i \leftarrow \mathsf{G}_B(\phi_{\text{debate},i}, [p_i] \oplus [p_j : j \neq i]; Q)$
**end for**
$\hat{y} \leftarrow \mathsf{A}([d_0, d_1, d_2, d_3]; Q)$
**return** $\hat{y}$

**Algorithm 5** Hypothesis–Verify.

$h \leftarrow \mathsf{G}_B(\phi_{\text{propose}}, P; Q)$
$v \leftarrow \mathsf{G}_B(\phi_{\text{verify}}, P \oplus [h]; Q)$
$\hat{y} \leftarrow \mathsf{A}(P \oplus [h, v]; Q)$
**return** $\hat{y}$

**Algorithm 6** Storyline.

$g \leftarrow \mathsf{G}_B(\phi_{\text{timeline}}, [p_0]; Q)$
**for** $i = 1, \ldots, 3$ **do**
    $c_i \leftarrow \mathsf{G}_B(\phi_{\text{local}}, [p_i, g]; Q)$
**end for**
$\hat{y} \leftarrow \mathsf{A}([g, c_1, c_2, c_3]; Q)$
**return** $\hat{y}$

# B Experiments

## B.1 Benchmarks and Settings

We evaluate video question answering on EgoSchema, LongVideoBench, and Video-MME, and use Perception Test for the cross-architecture analysis. The main benchmark comparisons use frozen Qwen3-VL-8B and Qwen3-VL-32B backbones, with the backbone specified for each configuration.

**EgoSchema.** EgoSchema (Mangalam et al., 2023) evaluates temporal understanding of egocentric videos. It contains approximately 5,000 human-curated multiple-choice questions based on three-minute clips from Ego4D, covering diverse everyday activities. Each question has five answer options and is designed to require information distributed across the video. The benchmark therefore emphasizes sustained understanding of activities and their temporal context. We report question-answering accuracy to assess performance on egocentric video understanding before and after additional agent handoffs.

**LongVideoBench.** LongVideoBench (Wu et al., 2024) evaluates long-context video-language understanding using videos with accompanying subtitles and durations of up to one hour. Its multiple-choice questions span 17 categories and require identifying referenced video contexts and reasoning about the associated details. We evaluate on a 1,337-question set and report question-answering accuracy. We use this benchmark for the controlled sequential comparison and component ablations, examining how state transfer, latent updates, and revisiting affect performance after agent handoffs.

Table 5: Evaluation splits and execution settings. $N$ counts unique questions per condition. All backbones are frozen Qwen3-VL models.

| Dataset | Split | $N$ | Subtitles | Backbone |
|---|---|---|---|---|
| EgoSchema | Official full | 5,031 | Off | 8B/32B |
| LongVideoBench | Public validation | 1,337 | on | 8B/32B |
| Video-MME | Public test | 2,700 | Off/on | 8B/32B |

*Notes.* $H$ counts additional handoffs after SVI fusion; $H = 0$ denotes the SVI reference. On denotes available official subtitles; missing inputs remain empty.

**Video-MME.** Video-MME (Fu et al., 2025) evaluates video understanding across diverse content and temporal scales. It comprises 900 videos and 2,700 multiple-choice questions, covering six broad domains and 30 subcategories. Video durations range from 11 seconds to one hour, with short, medium, and long categories supporting analysis across different temporal scales. The benchmark also provides subtitles and audio as additional modalities. Our experiments report accuracy for each duration category and overall, under subtitle-free and subtitle-assisted conditions. Official subtitles are used when available in the subtitle-assisted setting.

**Perception Test.** Perception Test (Patraucean et al., 2023) is a diagnostic benchmark for multimodal perception and reasoning. It contains approximately 11,600 real-world videos with an average duration of 23 seconds. Its annotations support six tasks, including multiple-choice and grounded video question answering, object and point tracking, and temporal localization of actions and sounds. The benchmark probes memory, abstraction, physical understanding, and semantic interpretation. We use its multiple-choice question-answering task for the cross-architecture analysis, evaluating Qwen3-VL-8B on a 709-question subset. We report accuracy across six MAS architectures and their communication stages.

### B.2 Observation and Computational Budgets

#### B.2.1 Budgets Along the Inherited Trajectory

Observation counts distinguish repeated frame presentations from unique source frames. For the combined revisiting policy, let $R \leq 2$ denote the number of completed revisits. With two-frame temporal patching, the counts along one trajectory containing the initial video input are

$$\begin{aligned} B_{\text{obs}}^{\text{path}} &= 2\lceil n_0/2 \rceil + 8R \leq 112, \\ B_{\text{unique}}^{\text{path}} &= n_0 + 4R \leq 104. \end{aligned} \tag{B.1}$$

The first term in $B_{\text{obs}}^{\text{path}}$ includes duplication of the final initial frame when $n_0$ is odd. Each successful revisit adds eight presentations, comprising four previously unseen and four original frames. The latter increase the presentation count without increasing the unique-frame count.

In the controlled sequential extension, the inherited trajectory starts from the fusion context. Its cached initial visual context is reused by the Reasoner and Aggregator. Equation (B.1) does not include the separate candidate-generation and verification contexts preceding fusion. The latent budget is given independently in Equation (A.5).

#### B.2.2 System Computation and Communication Payload

SVI contains $2N + 1 = 9$ video-conditioned inference stages before either added serial role. These logical stages must be distinguished from the number of physical visual-prefill executions, which depends on valid prefix reuse. If all nine stages independently process the initial video, their initial frame-presentation count is

$$B_{\text{obs}}^{\text{SVI,unshared}} = (2N + 1)\, 2\lceil n_0/2 \rceil. \tag{B.2}$$

This is an accounting expression for unshared execution, not a measured call count. Sharing decoded RGB frames or processor tensors alone does not remove independently executed model computa-

tion. The four candidate instances may run concurrently; concurrency changes elapsed time without eliminating their computation.

For a handoff with $T_i$ cached positions, let $d_{\mathrm{KV}}^{(\ell)}$ be the number of key or value scalars per position at layer $\ell$. With $b$ bytes per scalar, the raw tensor payload is

$$C_i^{\mathrm{KV}} = 2bT_i \sum_{\ell=1}^{L} d_{\mathrm{KV}}^{(\ell)}. \tag{B.3}$$

This expression excludes metadata, serialization overhead, and additional implementation-specific copies. The payload includes accumulated visual, textual, and latent positions. Few newly generated text tokens therefore do not imply a small transferred state.

### B.3 Gap Reduction

For a fixed benchmark, evaluation split, backbone, and subtitle condition, let $A_{\mathrm{reference}}$, $A_{\mathrm{Text}}$, and $A_{\mathrm{ViLaMAS}}$ denote the accuracies of the designated reference, ViLaMAS-Text, and ViLaMAS, respectively. All three accuracies are evaluated on the same questions; the textual and latent variants use the same roles and collaboration structure. Their accuracy gaps to the reference are

$$\begin{aligned} G_{\mathrm{Text}} &= A_{\mathrm{reference}} - A_{\mathrm{Text}}, \\ G_{\mathrm{ViLaMAS}} &= A_{\mathrm{reference}} - A_{\mathrm{ViLaMAS}}. \end{aligned} \tag{B.4}$$

When accuracies are expressed in percent, these gaps are measured in percentage points. We define *gap reduction* as the fraction of the textual baseline's gap closed by ViLaMAS:

$$\begin{aligned} \mathrm{GapReduction} &= \frac{G_{\mathrm{Text}} - G_{\mathrm{ViLaMAS}}}{G_{\mathrm{Text}}} \\ &= \frac{A_{\mathrm{ViLaMAS}} - A_{\mathrm{Text}}}{A_{\mathrm{reference}} - A_{\mathrm{Text}}}. \end{aligned} \tag{B.5}$$

The metric is defined for $A_{\mathrm{reference}} > A_{\mathrm{Text}}$. It is dimensionless and is reported as a percentage by multiplying by $100$. A value of $0$ indicates no improvement over ViLaMAS-Text, and $1$ indicates complete closure of the reference gap. Negative values indicate a larger gap, while values above $1$ indicate accuracy exceeding the reference.

**LongVideoBench example.** In the Qwen3-VL-8B sequential comparison, the reference is SVI before the two additional post-fusion handoffs. Using the accuracies reported in Table 1, $A_{\mathrm{reference}} = 66.0$, $A_{\mathrm{Text}} = 47.5$, and $A_{\mathrm{ViLaMAS}} = 61.8$ give

$$\mathrm{GapReduction} = \frac{61.8 - 47.5}{66.0 - 47.5} = \frac{14.3}{18.5} \approx 0.773. \tag{B.6}$$

Thus, the reported gap reduction is $77.3\%$: ViLaMAS recovers $14.3$ percentage points of the original $18.5$-point gap, leaving a $4.2$-point gap to SVI.

### B.4 Additional Analysis of Evidence-guided Revisiting

**Experimental setup.** We compared four revisiting strategies on two 100-question and 300-question Video-MME subsets using Qwen3-VL-8B. The subset comprised 34/100 short-, 33/100 medium-, and 33/100 long-video questions, respectively. Official subtitles were provided when available. Revisit-enabled conditions used no more than two rounds, with four latent updates (LU) after each successful revisit.

**Revisiting strategies.** The four strategies differed in their localization query or visual input:

1. **Original-frame rereading.** This condition reused the localization trajectory of the evidence-guided condition and reread eight frames from the initial observation sequence at each revisit, without adding new source frames.
2. **Random eligible interval.** This condition randomly selected an interval satisfying the revisiting eligibility constraints.

3. **Original-question localization.** This condition used the original video question as the localization query.
4. **Evidence-guided revisiting (EGL).** This condition used a state-conditioned visual question to distinguish the current competing answers, following Appendix A.3.

For the random-interval, original-question, and evidence-guided conditions, each successful revisit combined four new frames with four frames from the initial observation sequence. Thus, all four strategies presented eight frames per successful revisit, while the original-frame control acquired no new visual observations. Its reuse of the EGL localization trajectory held the selected intervals fixed for the comparison with EGL.

Table 6: Revisiting results on 100 sampled Video-MME questions with Qwen3-VL-8B and available official subtitles. Transition entries give counts and conditional percentages. Accuracy and ΔAcc. are derived from these counts.

| Method | Wrong→correct ↑ | Correct→wrong ↓ | Acc. (%) ↑ | ΔAcc. (pp) ↑ |
|---|---|---|---|---|
| Initial diagnostic state | 0/24 (0.00%) | 0/76 (0.00%) | 76.0 | 0.0 |
| Original-frame rereading | 2/24 (8.33%) | 4/76 (5.26%) | 74.0 | −2.0 |
| Random eligible interval | **3/24 (12.50%)** | 4/76 (5.26%) | 75.0 | −1.0 |
| Original-question localization | 2/24 (8.33%) | **2/76 (2.63%)** | 76.0 | 0.0 |
| Evidence-guided (EGL) | **3/24 (12.50%)** | **2/76 (2.63%)** | **77.0** | **+1.0** |

Table 7: Revisiting results on 300 sampled Video-MME questions with Qwen3-VL-8B and available official subtitles. Transition entries give counts and conditional percentages. Accuracy and ΔAcc. are derived from these counts.

| Method | Wrong→correct ↑ | Correct→wrong ↓ | Acc. (%) ↑ | ΔAcc. (pp) ↑ |
|---|---|---|---|---|
| Initial diagnostic state | 0/93 (0.00%) | 0/207 (0.00%) | 69.0 | 0.0 |
| Original-frame rereading | 6/93 (6.45%) | 18/207 (8.70%) | 65.0 | −4.0 |
| Random eligible interval | 3/93 (3.23%) | 14/207 (6.76%) | 65.3 | −3.7 |
| Original-question localization | 3/93 (3.23%) | 14/207 (6.76%) | 65.3 | −3.7 |
| Evidence-guided (EGL) | **10/93 (10.75%)** | **10/207 (4.83%)** | **69.0** | **0.0** |

**Results and interpretation.** EGL achieved the highest post-revisit accuracy in both evaluations, offering the best observed balance between correcting errors and preserving correct answers. On the 100-question subset (Table 6), EGL corrected three initially incorrect answers while introducing only two new errors, increasing accuracy from 76.0% to 77.0%. It matched the highest correction rate and the lowest rate of harmful revisions among the revisiting strategies, outperforming the alternatives by 1.0–3.0 percentage points.

On the 300-question subset (Table 7), EGL corrected ten errors, compared with three to six for the alternatives, while introducing ten new errors, compared with fourteen to eighteen. This balance preserved the pre-revisit accuracy of 69.0%, whereas the alternatives reduced accuracy to 65.0–65.3%. EGL therefore achieved a 3.7–4.0 percentage-point advantage over the other revisiting strategies. These results support evidence-guided localization as a more effective strategy for answer correction and accuracy retention, with EGL most effectively limiting post-handoff accuracy degradation among the evaluated revisiting strategies.

# C MISCELLANEOUS

## C.1 THEORETICAL JUSTIFICATION FOR EQUATIONS

We derive the ridge alignment in Equation 4 and justify the norm calibration used in the autoregressive update of Equation 5. We distinguish the optimality of this alignment objective from the properties of the subsequent decoder update.

**Notation and assumptions.** Let $W_{\text{in}}, W_{\text{out}} \in \mathbb{R}^{|\mathcal{V}|\times d}$ denote the frozen input embedding and output projection matrices. Their rows $w_{\text{in},v}, w_{\text{out},v} \in \mathbb{R}^{1\times d}$ correspond to the same vocabulary item $v \in \mathcal{V}$. Hidden states and latent inputs are represented as row vectors, so the alignment matrix acts on the right. We assume $\lambda > 0$ and $\epsilon > 0$; the reported configuration uses $\lambda = 10^{-5}$ and $\epsilon = 10^{-6}$.

**Derivation of Equation 4.** The paired rows provide a vocabulary-level criterion for mapping output vectors to input embeddings. For $W \in \mathbb{R}^{d\times d}$, consider the regularized least-squares objective

$$\begin{aligned}\mathcal{J}(W) &= \sum_{v\in\mathcal{V}} \|w_{\text{out},v}W - w_{\text{in},v}\|_2^2 + \lambda\|W\|_F^2 \\ &= \|W_{\text{out}}W - W_{\text{in}}\|_F^2 + \lambda\|W\|_F^2. \end{aligned} \tag{C.1}$$

This objective treats vocabulary items equally and penalizes the magnitude of the alignment matrix. Expanding the squared Frobenius norm gives

$$\begin{aligned}\mathcal{J}(W) = \operatorname{tr}\big(W^\top W_{\text{out}}^\top W_{\text{out}} W\big) - 2\operatorname{tr}\big(W^\top W_{\text{out}}^\top W_{\text{in}}\big) \\ + \operatorname{tr}\big(W_{\text{in}}^\top W_{\text{in}}\big) + \lambda \operatorname{tr}(W^\top W). \end{aligned} \tag{C.2}$$

Since $W_{\text{out}}^\top W_{\text{out}}$ is symmetric, differentiation with respect to $W$ yields

$$\nabla_W \mathcal{J}(W) = 2W_{\text{out}}^\top (W_{\text{out}}W - W_{\text{in}}) + 2\lambda W. \tag{C.3}$$

Setting this gradient to zero gives the normal equation

$$\big(W_{\text{out}}^\top W_{\text{out}} + \lambda I\big) W = W_{\text{out}}^\top W_{\text{in}}. \tag{C.4}$$

For every nonzero $x \in \mathbb{R}^d$,

$$x^\top \big(W_{\text{out}}^\top W_{\text{out}} + \lambda I\big) x = \|W_{\text{out}}x\|_2^2 + \lambda\|x\|_2^2 > 0. \tag{C.5}$$

The coefficient matrix is therefore positive definite and invertible, even when $W_{\text{out}}$ is rank deficient. Moreover, for any nonzero perturbation $\Delta \in \mathbb{R}^{d\times d}$,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{J}(W + t\Delta)\bigg|_{t=0} = 2\|W_{\text{out}}\Delta\|_F^2 + 2\lambda\|\Delta\|_F^2 > 0. \tag{C.6}$$

Thus, $\mathcal{J}$ is strictly convex, and its stationary point is the unique global minimizer:

$$W_a = \operatorname*{arg\,min}_{W\in\mathbb{R}^{d\times d}} \mathcal{J}(W) = \big(W_{\text{out}}^\top W_{\text{out}} + \lambda I\big)^{-1} W_{\text{out}}^\top W_{\text{in}}. \tag{C.7}$$

**Norm calibration in Equation 5.** For agent $j$ at latent step $k$, first form the aligned vector

$$u_{j,k} = h_{j,k-1}W_a. \tag{C.8}$$

The linear map determines an aligned direction, but its objective does not constrain the norm of every transformed hidden state. We therefore calibrate the input magnitude using the mean norm of the pretrained input embeddings,

$$\bar{\rho} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \|w_{\text{in},v}\|_2 > 0. \tag{C.9}$$

For a nonzero row vector $u$, norm calibration can be characterized by the constrained problem

$$z^\star(u) = \operatorname*{arg\,min}_{z\in\mathbb{R}^{1\times d},\ \|z\|_2=\bar{\rho}} \|z - u\|_2^2. \tag{C.10}$$

For any feasible $z$, the Cauchy–Schwarz inequality gives

$$\begin{aligned}\|z-u\|_2^2 &= \bar{\rho}^2 + \|u\|_2^2 - 2zu^\top \\ &\geq \bar{\rho}^2 + \|u\|_2^2 - 2\bar{\rho}\|u\|_2 = (\bar{\rho} - \|u\|_2)^2. \end{aligned} \tag{C.11}$$

For $u \neq 0$, equality holds precisely when $z$ has the same direction as $u$ and norm $\bar{\rho}$. Consequently, the unique solution is

$$z^\star(u) = \bar{\rho}\frac{u}{\|u\|_2}. \tag{C.12}$$

This construction preserves the aligned direction while matching the reference norm. The denominator floor $\epsilon$ regularizes the rule near the origin, giving the implemented latent input

$$z_{j,k} = \bar{\rho}\frac{u_{j,k}}{\max(\|u_{j,k}\|_2, \epsilon)} = \begin{cases} \bar{\rho}\dfrac{u_{j,k}}{\|u_{j,k}\|_2}, & \|u_{j,k}\|_2 \geq \epsilon, \\ \dfrac{\bar{\rho}}{\epsilon}u_{j,k}, & \|u_{j,k}\|_2 < \epsilon. \end{cases} \tag{C.13}$$

The two branches agree at $\|u_{j,k}\|_2 = \epsilon$, and the rule returns zero when $u_{j,k} = 0$. It is therefore continuous and avoids division by zero. Its norm satisfies

$$\|z_{j,k}\|_2 = \bar{\rho}\frac{\|u_{j,k}\|_2}{\max(\|u_{j,k}\|_2, \epsilon)} = \bar{\rho}\min\left\{1, \frac{\|u_{j,k}\|_2}{\epsilon}\right\} \leq \bar{\rho}. \tag{C.14}$$

Hence, the latent input has norm $\bar{\rho}$ whenever $\|u_{j,k}\|_2 \geq \epsilon$; below this threshold, its norm decreases linearly to zero. Exact norm matching follows from the constrained problem, whereas the denominator floor is an explicit regularization choice near zero.

**Autoregressive state update.** Let $S_{j,k-1}$ contain the accumulated KV cache and the continuation metadata before step $k$. The frozen decoder $F_\theta$ consumes the continuous input $z_{j,k}$ and returns the next hidden state and updated cache state:

$$(h_{j,k}, S_{j,k}) = F_\theta(z_{j,k}; S_{j,k-1}). \tag{C.15}$$

Substituting Equations (C.8) and (C.13) into this transition yields

$$(h_{j,k}, S_{j,k}) = F_\theta\left(\bar{\rho}\,\frac{h_{j,k-1}W_a}{\max(\|h_{j,k-1}W_a\|_2, \epsilon)}\,;\, S_{j,k-1}\right), \qquad k = 1, \ldots, m_j. \tag{C.16}$$

### C.2 Prompt Template

Select only the matching benchmark, architecture, and stage instructions below. Video and available subtitles are supplied through the native input interface.

**System prompt for all agents**

You are Qwen, created by Alibaba Cloud. You are a helpful assistant.

#### C.2.1 Prompt for the Observer Agents

Star, Chain, Tree, Debate, Hypothesis–Verify, and Storyline share independent Observers with identical prompts and initial video inputs.

You are acting as the observer now. Focus on `{benchmark_focus}`. Identify relevant visual evidence, using subtitles when provided. Return a candidate option, supplied frame or timestamp references, and a brief explanation. Preserve uncertainty and do not invent unseen events.
**Question:** `{question}`
**Options:** `{options}`

**Benchmark focus**

| | |
|---|---|
| EgoSchema | actions, object interactions, and event order |
| LongVideoBench | referenced scenes, events, and relevant details |
| Video-MME | local details or overall video content |
| Perception Test | object attributes, actions, and spatial or temporal cues |

#### C.2.2 Prompt for the Reasoner Agent

You are acting as the reasoner now. `{reasoning_task}` Focus on `{benchmark_focus}`. Compare candidate answers using the supplied context. Resolve contradictions where possible, retain supported conclusions, and preserve uncertainty. Do not give the final answer.
**Question**: `{question}`
**Options**: `{options}`

**Benchmark focus**

| | |
|---|---|
| EgoSchema | action sequences and actor-object relations |
| LongVideoBench | event identity, temporal order, and subtitle alignment when available |
| Video-MME | question-relevant details at the required temporal scope |
| Perception Test | object-state changes, event relations, and physical consistency |

**Architecture-specific `{reasoning_task}`**

| | |
|---|---|
| Main SVI | Check the fused proposal using evidence and verification feedback. |
| Star | Integrate all observers' findings. |
| Chain | Merge the accumulated assessment with the next observer's findings. |
| Tree | Reconcile the assigned observer pair's findings. |
| Debate | Reassess your assigned proposal against the other observers' evidence. |
| Hypothesis–Verify: propose | Propose a hypothesis from all observers' findings. |
| Hypothesis–Verify: verify | Verify the hypothesis against all observers' findings. |
| Storyline: timeline | Build a timeline from the first observer's findings. |
| Storyline: local | Check your assigned findings against the timeline. |

### C.2.3 PROMPT FOR THE AGGREGATOR AGENT

You are acting as the aggregator now. `{aggregation_task}` Focus on `{benchmark_focus}`. Resolve disagreements using evidence, including any revisited frames provided. Retain an earlier answer when justified. Return only the best-supported option letter.
**Question**: `{question}`
**Options**: `{options}`

**Benchmark focus**

| | |
|---|---|
| EgoSchema | activity consistency and action order |
| LongVideoBench | referenced scenes and temporal context |
| Video-MME | evidence matching the question's temporal scope |
| Perception Test | consistency of the answer with question-relevant perceptual evidence |

**Architecture-specific `{aggregation_task}`**

| | |
|---|---|
| Main SVI | Review the reasoner's assessment and fused proposal. |
| Star | Review the central reasoner's synthesis. |
| Chain | Review the final accumulated assessment. |
| Tree | Combine the two pairwise assessments. |
| Debate | Compare revised assessments by evidence. |
| Hypothesis–Verify | Combine observer findings, hypothesis, and verification feedback. |
| Storyline | Combine the timeline and three local assessments. |