# ViSAGE: Constructing Self-Correcting Memories for Long-Form Video Understanding

Anonymous ACL submission

## Abstract

Multimodal agents operating in long-horizon environments must build and continually update memory to support entity-consistent, temporally grounded reasoning. However, existing agentic memory approaches often discard fine-grained identity cues under aggressive compression and segment-wise processing. They also over-trust vector-similarity retrieval, surfacing semantically related yet identity-mismatched evidence and thereby causing identity errors and hallucinations. We propose ViSAGE, a multimodal agentic memory framework that constructs self-correcting, entity-centric memories. Specifically, ViSAGE anchors entity identity via cross-modal binding over long temporal ranges. It then applies bidirectional memory refinement to propagate delayed identity evidence, retroactively unifying historical records and stabilizing future reasoning. We also introduce multi-agent cross-verification to adjudicate retrieved evidence under an identity–evidence alignment constraint, enabling verified refusals instead of hallucinations when evidence is missing. Extensive results demonstrate that ViSAGE consistently outperforms the strongest baseline, achieving 5.9% higher accuracy.

## 1 Introduction

With the rapid progress of multimodal large language models (MLLMs) (Google DeepMind, 2025; Lin et al., 2024b; Li et al., 2023), agents can now perceive and communicate through vision, audio, and language (Fan et al., 2024; Fung et al., 2025). However, deploying such agents in long-horizon settings requires memory that persists beyond a single context window. Therefore, prior work typically uses retrieval-augmented memory to extend MLLMs beyond finite context windows (Park et al., 2023; Zhong et al., 2024; Chhikara et al., 2025; Liu et al., 2024; Hu et al., 2025). In this paradigm, the agent writes observations to external memory and retrieves relevant entries to guide generation and decisions. Many systems use vector-store retrieval (Zhong et al., 2024; Chhikara et al., 2025; Fan et al., 2024, 2025), while hierarchical designs organize memories across multiple time scales for long-horizon planning (Liu et al., 2024; Hu et al., 2025). Optimus-1 (Li et al., 2024) further explores hybrid multimodal memory, combining a hierarchical directed knowledge graph with an abstracted experience pool to encode world knowledge and past multimodal experience. More recently, M3-Agent structures memory as an entity-centric graph to better support long-horizon reasoning over entities (Long et al., 2025).

Although these efforts have shown promising performance in long-horizon multimodal agentic memory tasks, we identify several limitations in current paradigms:

- **Spatio-Temporal Detail Loss.** Existing systems face a trade-off between span and detail. To ingest extensive multimodal streams, they often rely on aggressive down-sampling or compression, which erases critical micro-dynamics and thus removes fine-grained identity cues needed for reliable entity disambiguation.

- **Segmentation Dilemma.** Many long-horizon pipelines process inputs in isolated chunks. Identity evidence often arrives late and must be linked back to earlier events. Chunk-wise processing blocks cross-segment revision, so late cues cannot fix earlier memories.

- **Similarity–Veracity Gap.** Retrieval-based memory conflates semantic relevance with factual correctness. Vector-similarity retrieval can surface contextually related yet identity-mismatched or incorrect evidence; without verification, agents are prone to answering based on noisy retrieval.

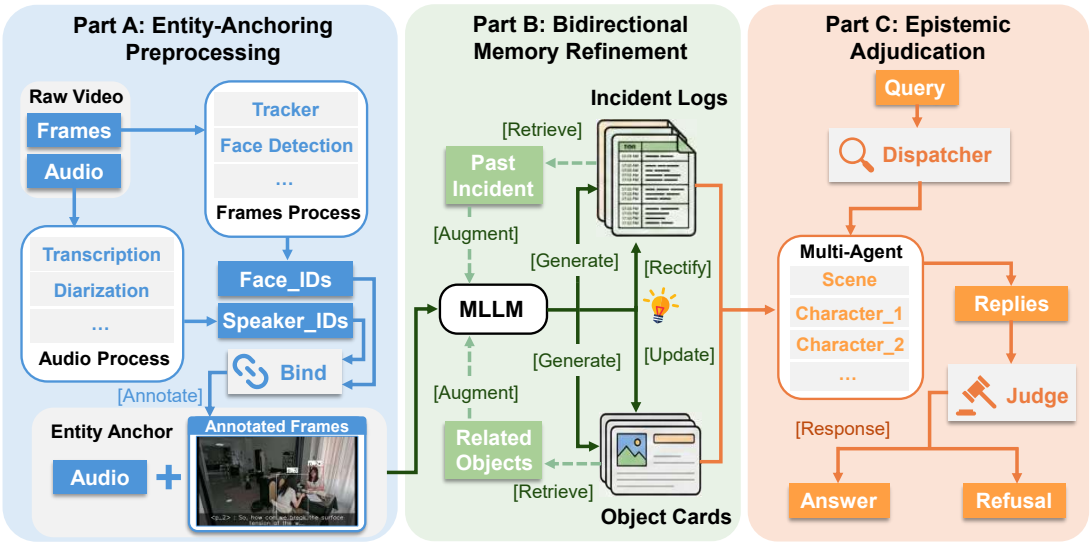


Figure 1: Architecture of ViSAGE. The framework proceeds in three stages: (A) Entity-Anchoring Preprocessing binds visual and audio signals to annotated frames; (B) Bidirectional Context Refinement maintains self-correcting memory (Incident Logs and Object Cards) via retrieval and rectification; and (C) Epistemic Adjudication employs multi-agent cross-verification to ensure reliable responses.

To address these challenges, we introduce ViSAGE, a **Vi**sual **S**elf-correcting **AGE**ntic memory framework that constructs *entity-centric, self-correcting* memories beyond linear, chunk-wise processing. ViSAGE maintains a dual-structured memory: *Incident Logs* record sequential narrative events, while *Object Cards* maintain evolving entity profiles that accumulate and revise identity attributes. To align identity cues across spatial degradation and temporal sparsity, we perform **Cross-Modal Entity Binding** by binding long-range visual tracklets with audio signals. To overcome isolated segmentation, ViSAGE further introduces **Bidirectional Memory Refinement** to propagate delayed identity evidence, retroactively unifying historical records and stabilizing future reasoning. Finally, we propose **Multi-Agent Cross-Verification** to adjudicate retrieved evidence under an identity–evidence alignment constraint, enabling verified refusals instead of hallucinations when evidence is missing.

We evaluate ViSAGE through extensive experiments on long-horizon multimodal benchmarks, including M3-Bench-robot, M3-Bench-web (Long et al., 2025), and Video-MME-long (Fu et al., 2025). Experimental results show that ViSAGE consistently outperforms strong baselines across diverse reasoning categories by 5.9%, demonstrating robust gains on identity-critical queries. Moreover, our results verify that Multi-Agent Cross-Verification improves reliability by enforcing identity-evidence alignment, enabling verified refusals instead of hallucinated answers when supporting evidence is missing or noisy. Our contributions are summarized as follows:

- **Framework Design:** We introduce *ViSAGE*, a visual self-correcting agentic memory framework that enables *retroactive, entity-centric* updates beyond chunk-wise processing. ViSAGE separates memory into sequential *Incident Logs* and structured *Object Cards*, allowing delayed identity evidence to be aligned with historical events for long-horizon entity consistency.
- **Algorithmic Novelty:** We propose **Bidirectional Memory Refinement** to resolve the Segmentation Dilemma. This mechanism transcends sequential processing via a feedback loop that retroactively unifies fragmented narratives using delayed identity evidence, ensuring long-term entity consistency.
- **Epistemic Safety:** To bridge the Similarity-Veracity Gap, we introduce **Multi-Agent Cross-Verification**. This framework enforces identity-constrained reasoning to mitigate hallucinations, prioritizing Verified Epistemic Refusals over plausible fabrications in high-stakes embodied scenarios.
- **Empirical Impact:** ViSAGE sets new state-of-the-art on authoritative benchmarks. Our framework consistently outperforms baselines,

achieving accuracy of 45.5% on M3-Bench-robot, 58.4% on M3-Bench-web, and 79.1% on Video-MME-long, validating its robustness in long-form video understanding.

## 2 Related Work

### 2.1 Long Video Understanding

Recent advancements in Multimodal Large Language Models (MLLMs), such as GPT-4o (Hurst et al., 2024), Gemini-3-Pro (Google DeepMind, 2025), and the open-source Qwen3-VL (Bai et al., 2025), have revolutionized visual perception. While methods like Video-LLaVA (Lin et al., 2024a) and VILA (Lin et al., 2024b) excel in short clips, extending them to hour-long videos remains a challenge due to context limits. Existing solutions typically follow two paradigms:

Early works, inspired by Socratic Models (Zeng et al., 2022), decompose video reasoning into text-based tasks. Methods like Video Recap (Islam et al., 2024) and AutoAD (Han et al., 2023) generate hierarchical captions or summaries to fit textual RAG systems (Lewis et al., 2020). Recent benchmarks like EgoSchema (Mangalam et al., 2023) and LongVideoBench (Wu et al., 2024) highlight that while efficient, these methods suffer from severe information loss, discarding fine-grained visual cues essential for reasoning.

Recently, another line of work has focused on expanding the model to retain visual features within limited context windows. Specifically, token compression methods such as LongVILA (Chen et al., 2024b) and LongVU (Shen et al., 2024), utilize spatiotemporal pooling to reduce redundancy, while VidCompress (Lan et al., 2024) and Video-XL (Liu et al., 2025) employ adaptive selection to fit hour-scale videos. In contrast, memory-centric models like MovieChat (Song et al., 2024), MA-LMM (He et al., 2024), LifelongMemory (Wang et al., 2023) and MeMViT (Wu et al., 2022) introduce explicit external banks or buffers, with Flash-VStream (Zhang et al., 2024) specifically optimizing for real-time updates.

### 2.2 Memory Systems in AI Agents

Recently, AI Agents have evolved from simple instruction followers to autonomous entities capable of long-horizon planning, heavily relying on robust memory architectures.

Foundational works like Generative Agents (Park et al., 2023) established the importance of retrieving past experiences. Recent frameworks have formalized this: Mem0 (Chhikara et al., 2025) and MemoryBank (Zhong et al., 2024) offer production-ready vector memory layers. HiAgent (Hu et al., 2025) and AgentLite (Liu et al., 2024) extend this with hierarchical memory for complex planning, while AIOS (Mei et al., 2024) explores operating-system-level memory management for resource optimization.

Extending agents to the visual domain, VideoAgent (Fan et al., 2024) and its embodied variant Embodied VideoAgent (Fan et al., 2025) utilize tools for iterative video retrieval. Jarvis-1 (Wang et al., 2024) integrates memory with multimodal planning. A-Mem (Xu et al., 2025c) explores agentic memory for evolving environments. Most notably, M3-Agent (Long et al., 2025) structures memory into an entity-centric graph.

## 3 Method

As shown in Fig. 1, to solve the challenges of identity fragmentation and hallucination, ViSAGE introduces an integrated framework composed of three distinct specialized phases:

### 3.1 Memory Bank

Inspired by Tulving's multiple memory systems theory (Tulving, 2002), which distinguishes time- and context-bound experience from decontextualized knowledge about entities and emphasizes a mechanistic link "from mind to brain", we design ViSAGE's memory bank as an incident–entity pair (Table 1). Specifically, *Incident Logs* store an append-only record of *what happened*, including timestamps, actions, and dialogue. This event content is immutable, but the *who* field remains refinable. In parallel, *Object Cards* maintain continuously updated entity knowledge. This separation enables retroactive identity unification without corrupting the historical event record.

### 3.2 Entity-Anchored Preprocessing

**Countering Spatial Degradation via Sequence-Level Entity Anchoring.** The representative agent-based approach (Long et al., 2025) adopts a partition-and-process paradigm that slices long videos into independent segments. This confines identity recognition to each local window, making it brittle under occlusion, pose changes, or compression artifacts. As a result, long videos often attenuate fine-grained spatio-temporal cues, leading

| Memory Track | Schema Definition | Instance Example (Snapshot) |
|---|---|---|
| **Incident Logs** | *A chronological sequence of narrative events:* $\mathcal{L} = \{(t, \mathcal{P}_{\text{ids}}, \text{Act}, \text{Dial})\}$ | `[00:07]` <p_1=Adam=Adan> crouches down to examine the details of the bright yellow wing structure.<br>`[00:28]` <p_1=Adam=Adan> gestures towards the mechanism and asks: "What does this platform represent?" |
| **Object Cards** | *Entity-centric profile and state containers:* $\mathcal{O} = \{(\text{ID}, \mathcal{N}, \mathcal{V}, \mathcal{P}, \mathcal{S})\}$ | **ID:** 1 **Names:** [Adam, Adan]<br>**Vectors:** $\{\mathbf{v}_{\text{face}}, \mathbf{v}_{\text{voice}}, \mathbf{v}_{\text{name}}\}$<br>**Profile:** {Role: Presenter, Gender: Male}<br>**State:** {Loc: Workshop, Action: Gesturing} |

Table 1: Data Structure of the Dual-Track Memory System. Examples are simplified for visualization.

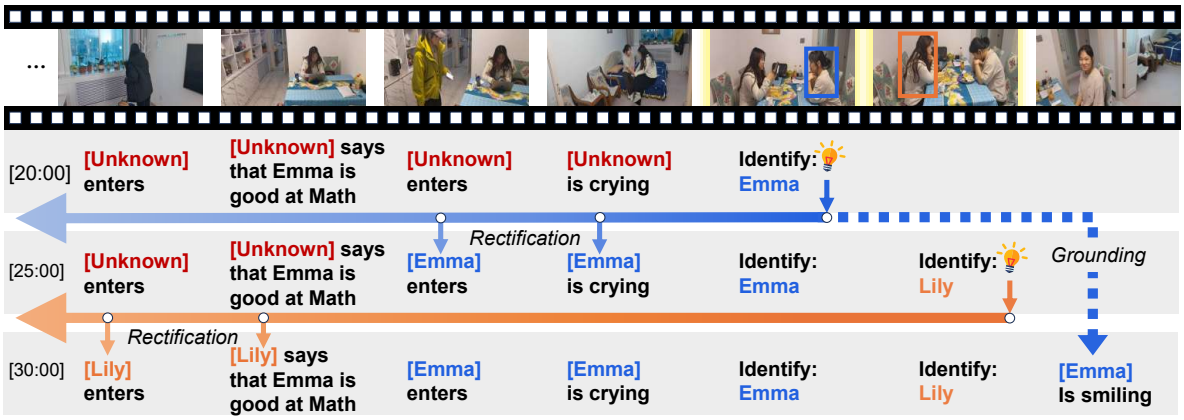


Figure 2: Visualizing Bidirectional Memory Refinement. Late-appearing Identity Triggers activate a dual loop: Backward Rectification retroactively fills historical uncertainties, while Forward Grounding propagates these resolved identities to future timestamps. This ensures actions are anchored to the correct character profile.

to identity switches and noisy event–entity associations during memory construction. Such errors propagate to both *Incident Logs* and *Object Cards*, degrading retrieval and consolidation. We therefore perform entity-anchored preprocessing to stabilize identity grounding at the sequence level.

To this end, we introduce *Sequence-Level Entity Anchoring*. Rather than relying on unreliable per-frame detections, we adopt a prototype-based strategy over long-range tracks. Specifically, we run continuous multi-object tracking over the entire video to obtain long-range trajectories, and collect all face detections within each trajectory. We then select the highest-quality face instance as a *cluster prototype* and treat it as the canonical reference. All remaining potentially low-quality face observations are matched to this prototype, clustering transient frames around a reliable anchor. This design yields a one-to-one mapping from each identity cluster to an entity in the semantic layer: the prototype initializes the corresponding *Object Card*, while the cluster assignment provides a consistent entity ID that is propagated across frames. Importantly, partial cues observed in fleeting or low-quality frames can be attributed to the correct entity by referencing the prototype, enabling robust event–entity association in *Incident Logs* and consistent attribute consolidation into *Object Cards*. Robust identity is thus maintained throughout the duration of each track, ensuring the subject identity remains invariant despite transient visual noise.

**Bridging Temporal Sparsity via Speaker Binding.** Parallel to visual anchoring, we leverage audio cues to improve *who-did-what* attribution, where knowing *who is speaking* is decisive. A practical challenge is that standard MLLM pipelines operate on temporally sparsified video inputs, which discard the rapid lip dynamics needed for reliable speaker grounding. Meanwhile, relying solely on speaker embeddings (Long et al., 2025) can be brittle in crowded scenes due to acoustic similarity and intra-speaker variability. Motivated by the McGurk effect (McGurk and MacDonald, 1976), which illustrates how vision can disambiguate am-

biguous audio, we introduce *Dual-Criteria Validation* for voice–face binding. Concretely, we run Active Speaker Detection (ASD) on the high-FPS stream to obtain per-face articulation/synchrony evidence, and combine it with acoustic matching from voice embeddings. We bind a `Voice_ID` to a `Face_ID` whenever the audio is *confidently identifiable* with high voice-similarity or the video is *physically corroborative* with ASD-confirmed audio–visual synchrony within the same time window; otherwise, the speaker remains unassigned.

### 3.3 Bidirectional Memory Refinement

To resolve the Segmentation Dilemma, where strict sequential processing fails to link asynchronous identity evidence with prior appearances, ViSAGE implements Bidirectional Memory Refinement. Unlike unidirectional approaches that freeze past event–entity bindings once written to memory, this mechanism establishes a symmetric feedback loop: it combines Forward Grounding to stabilize current perception with Retroactive Rectification, which propagates late-arriving identity clues back into the Incident Logs to unify disjointed narrative segments, as illustrated in Fig. 2.

**Forward Grounding via Identity-Keyed Retrieval.** To counteract local contextual drift in the current timeframe, we implement an anchor-based retrieval strategy. Prior to memory generation, the system utilizes the stable `Face_ID`, derived from the tracklet anchoring phase, to retrieve memory from two parallel knowledge streams: the immediately preceding events from Incident Logs (for narrative continuity) and the character's accumulated profile from Object Cards (for persona consistency). By priming the MLLM with this dual-source memory, we transform the interpretation of the current clip from an isolated fragment into a continuous narrative extension, ensuring the "Now" is grounded in the "Past".

**Backward Refinement via Logical Merging.** Despite the robust physical continuity established in Section 3.2, where the system correctly binds the audio information to the entity, the agent remains blind to the character's nominal identity. The mapping between a face and a name typically relies on complex dialogue comprehension, specifically determining the addressee in a conversation. To handle this, we treat every newly appearing name as a provisional semantic identity. The LLM performs memory-aware inference: if it deduces that a specific visual subject is the target of a naming utterance, such as "Hey, Mario!", it generates a Refinement Signal ($\text{ID}_{\text{sem}} \equiv \text{ID}_{\text{vis}}$). Crucially, this signal triggers a Global Backward Update: the system traverses the historical Incident Logs $\mathcal{L}$, retroactively replacing all prior instances of the anonymous $\text{ID}_{\text{vis}}$ with the resolved identity. This mechanism ensures that if an anonymous subject recorded on Day 1 is identified as "Mario" on Day 2, their entire history is unified, repairing the structural fragmentation.

### 3.4 Epistemic Adjudication

To bridge the Similarity–Veracity Gap, ViSAGE introduces Object Cards as a high-confidence knowledge anchor, supplementing the traditional retrieval from Incident Logs.

**Noise Filtering via Identity Resolution Dispatcher.** To identify the entity most relevant to the query, the Dispatcher functions as a sanitization layer against ASR-induced noise. Automatic transcription frequently introduces phonetic inconsistencies, such as erroneously transcribing the target "Mario" as "Malloy", which causes naive vector retrieval to fail or return irrelevant distractors. Capitalizing on the fact that all valid character names are *indexed* in the database, the Dispatcher employs LLM-based logical inference to perform Identity Resolution. It explicitly maps the noisy query entity to its correct counterpart within this registry, effectively correcting errors like "Malloy" to "Mario". It then activates only the relevant Character Agent, ensuring that retrieval is strictly targeted and robust to phonetic mismatches.

**Evidence Extraction via Persona-Driven Agents.** Crucially, ViSAGE employs a complementary retrieval strategy specialized by query type. The Scene Agent interrogates Incident Logs to answer factual, event-driven questions. It utilizes the objective chronological record to resolve dynamic narrative details, such as "who stood up first", which rely on precise temporal causality. Complementing this, the Character Agent leverages Object Cards to resolve static attribute inquiries. Without this solidified profile, answering identity-related questions like "What is Lily's profession?" would necessitate on-the-fly inference based on disjointed event fragments in the Incident Logs. In contrast, ViSAGE continuously accumulates and validates identity traits throughout the video processing pipeline.

| Method | M3-Bench-robot | | | | | | M3-Bench-web | | | | | | Video-MME-long |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ME | MH | CM | PU | GK | ALL | ME | MH | CM | PU | GK | ALL | |
| *Video-Native Socratic method* | | | | | | | | | | | | | |
| Qwen2.5-Omni-7b | 2.1 | 1.4 | 1.5 | 1.5 | 2.1 | 2.0 | 8.9 | 8.8 | 13.7 | 10.8 | 14.1 | 11.3 | 42.2 |
| Gemini-1.5-Pro | 6.5 | 7.5 | 8.0 | 9.7 | 7.6 | 8.0 | 18.0 | 17.9 | 23.8 | 23.1 | 28.7 | 23.2 | 38.1 |
| *Online Video Understanding Methods* | | | | | | | | | | | | | |
| MovieChat | 13.3 | 9.8 | 12.2 | 15.7 | 7.0 | 11.2 | 12.2 | 6.6 | 12.5 | 17.4 | 11.1 | 12.6 | 19.4 |
| MA-LMM | 25.6 | 23.4 | 22.7 | 39.1 | 14.4 | 24.4 | 26.8 | 10.5 | 22.4 | 39.3 | 15.8 | 24.3 | 17.3 |
| Flash-Vstream | 21.6 | 19.6 | 19.3 | 24.3 | 14.1 | 19.4 | 24.5 | 10.3 | 24.6 | 32.5 | 20.2 | 23.6 | 25.0 |
| *Discrete Socratic method* | | | | | | | | | | | | | |
| Qwen3-VL-8b | 39.0 | 29.7 | 39.3 | 52.9 | 18.8 | 36.1 | 38.8 | 22.0 | 40.0 | 40.5 | 45.1 | 36.9 | 60.1 |
| Qwen3-VL-30b | 36.8 | 36.1 | 37.1 | 53.3 | 22.2 | 36.4 | 43.0 | 26.0 | 40.0 | 50.4 | 45.1 | 40.9 | 67.0 |
| GPT-4o | 34.0 | 32.8 | 32.9 | 42.2 | 16.7 | 29.9 | - | - | - | - | - | - | 65.3 |
| Gemini-3-Pro-preview | 42.7 | 38.3 | 38.0 | 55.1 | 23.3 | 39.6 | 51.9 | 40.3 | 62.9 | 61.0 | 59.0 | 53.8 | 74.2 |
| *Agent-based Method* | | | | | | | | | | | | | |
| M3-Agent | 32.8 | 29.4 | 31.2 | 43.3 | 19.1 | 30.7 | 45.9 | 28.4 | 44.3 | 59.3 | 53.9 | 48.9 | 61.8 |
| **ViSAGE(Ours)** | **45.2** | **47.2** | **47.1** | **56.9** | **30.9** | **45.5** | **62.3** | **47.8** | **71.4** | **63.3** | **62.2** | **58.4** | **79.1** |

Table 2: Performance Comparison on M3-Bench and Video-MME-long. We report results across different reasoning types in M3-Bench: Multi-Evidence (ME), Multi-Hop (MH), Cross-Modal (CM), Person Understanding (PU), and General Knowledge (GK).

**Epistemic Adjudication via Judge Agent.** In the final phase, the Judge Agent functions as the ultimate arbiter of veracity. Instead of simply aggregating outputs, it rigorously evaluates the sufficiency and consistency of the evidence retrieved from both Incident Logs and Object Cards. Crucially, when valid evidence is absent in both sources, or when irreconcilable conflicts arise, the Judge triggers a Verified Epistemic Refusal. This strict rejection mechanism ensures safety by preventing the system from hallucinating answers to unanswerable queries.

# 4 Experiments

## 4.1 Experimental Setup

| Dataset Split | robot | web | V-MME-long |
|---|---|---|---|
| **# Videos** | 100 | 920 | 300 |
| **Avg. Length** | 34m | 27m | 40m |

Table 3: Statistics of evaluation benchmarks. We utilize the robot and web splits from M3-Bench and the long-video subset of Video-MME.

**Evaluation Datasets.** To comprehensively analyze ViSAGE, we utilize three authoritative benchmarks (summarized in Table 3). First, the robot split of M3-Bench (Long et al., 2025) focuses on general-purpose robots, featuring egocentric videos that test memory-guided reasoning such as inferring human personalities, interpersonal relationships, and object affordances. Complementing this, the web split of M3-Bench (Long et al., 2025) and Video-MME (Fu et al., 2025) provide content with high information density from online platforms. These videos cover diverse topics like documentaries and movies, challenging the agent to process complex narratives and open-world knowledge relevant to practical multimodal applications.

**Baselines.** We benchmark ViSAGE against a suite of methods across three paradigms: (1) **Socratic Models**, encompassing Video-Native systems (Gemini-1.5-Pro (Team et al., 2024), Qwen2.5-Omni (Xu et al., 2025b)) that process holistic streams, and Discrete variants (GPT-4o (Hurst et al., 2024), Qwen3-VL (Bai et al., 2025), Gemini-3-Pro-preview (Google DeepMind, 2025)) utilizing frames sampled at 0.5 fps with ASR; (2) **Online Video Understanding Methods** representing diverse memory compression strategies, including MovieChat (Song et al., 2024), MA-LMM (He et al., 2024), and Flash-VStream (Zhang et al., 2024); and (3) **Agent-**

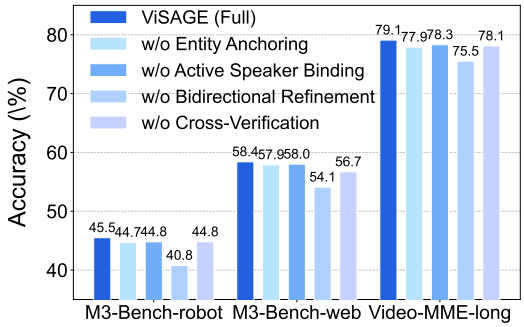


Figure 3: Ablation study on key components. The chart illustrates the impact of removing each module across three benchmarks (M3-Bench-robot, M3-Bench-web, and Video-MME-long).

**based Frameworks**, specifically the previous SOTA M3-Agent (Long et al., 2025), which employs an entity-centric memory graph and serves as the primary structural benchmark.

**Implementation Details.** We utilize Qwen3-Omni-Flash (Xu et al., 2025a) for speech transcription, followed by Gemini-3-Pro-preview as the backbone MLLM for reasoning and memory synthesis. See Appendix A for more details.

## 4.2 Performance Analysis

Table 2 presents the quantitative evaluation. ViSAGE consistently demonstrates superior efficacy across all benchmarks.

Crucially, regarding the Cross-Modal (CM) metric, ViSAGE achieves a decisive breakthrough. This metric directly validates the resolution of the Spatio-Temporal Detail Loss, measuring the system's ability to bind fragmented visual and auditory cues into coherent entities. On this front, ViSAGE significantly outperforms the agentic baseline M3-Agent (improving +15.9% on robot and +27.1% on web), proving that our Entity Anchoring mechanism successfully counters signal degradation where standard agents fail.

Furthermore, ViSAGE effectively grounds the raw capabilities of the Gemini-3-Pro-preview backbone. While the backbone exhibits strong parametric knowledge, it remains vulnerable to the sampling-induced information loss, which scores only 38.0% on robot CM. By shifting from passive perception to entity-anchored reasoning, ViSAGE bridges this gap, boosting the score to 47.1%. This robust grounding capability extends to open-world long-horizon scenarios, where ViSAGE achieves 79.1% on Video-MME-long, surpassing both the ungrounded backbone (74.2%) and the fragmentation-prone M3-Agent (61.8%).

| Backbone Model | Accuracy (%) |
|---|---|
| ViSAGE w/ Qwen3-VL-30b | 42.1 |
| ViSAGE w/ GPT-5 | 44.5 |
| ViSAGE w/ Gemini-3-Pro-preview | 45.5 |

Table 4: Impact of backbone scaling. We report the average accuracy on M3-Bench-robot (Long et al., 2025).

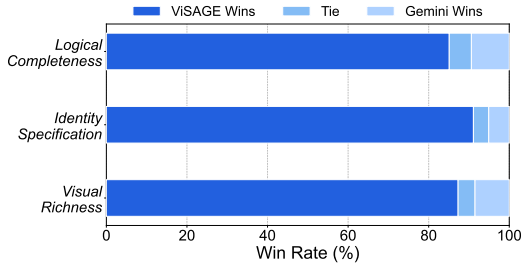


Figure 4: Quality Assessment on Correct Samples. Even when the final answer is correct, ViSAGE generates responses with significantly higher visual richness and logical completeness compared to baselines.

## 4.3 Ablation Study

Fig. 3 and Table 4 summarize our component analysis. Structurally, Bidirectional Memory Refinement proves critical (Fig. 3); its removal triggers the sharpest drop ($-4.7\%$ on robot), confirming that Backward Rectification is essential for unifying the fragmented narratives caused by the Segmentation Dilemma. Other modules also contribute distinctively to the system's robustness: Entity Anchoring (Part A) provides consistent gains by securing identity stability against the spatio-temporal detail loss. Multi-Agent Cross-Verification (Part C) further boosts precision by filtering semantic noise and enforcing epistemic safety. Crucially, Table 4 highlights the universality of our framework. ViSAGE maintains robust performance across distinct architectures ranging from Qwen3-VL-30b to GPT-5 (OpenAI, 2025). This validates that our Dual-Track design is a model-agnostic framework: it successfully decouples memory structure from specific reasoning backends, enhancing any foundation model with long-horizon reasoning capabilities.

## 4.4 Qualitative and Safety Analysis

**Case Study: Resolving Pronominal Ambiguity.** As detailed in Table 5, the baseline fails on pronominal resolution, succumbing to misleading hearsay. In contrast, ViSAGE successfully disentangles this ambiguity. By retroactively grounding

**User Query:** "What is Emma's major?"
**Ground Truth:** Programming.

**Gemini Response (Baseline):**
"Emma's major is **Linear Algebra** ✗."
***Retrieved Memory:*** A young woman says: "*I know Emma is good at Linear Algebra.*"
***Error Analysis:*** The baseline relies on surface-level name matching, prioritizing the explicit mention of "Emma" by a third party while ignoring the pronoun "I" spoken by the target character (who was initially anonymous).

**ViSAGE (Ours) Reasoning Process:**
**1. Bidirectional Refinement (Resolving "I"):**
• *Initial State:* At [00:19], <p_7> says "**I** study Programming." (Speaker unknown).
• *Trigger & Update:* At [05:30], identity revealed (<p_7> → Emma). System **retroactively propagates** this identity, modifying the historical log at [00:19] to: "**p_7=Emma** says 'I study Programming.'"
**2. Cross-Verification:**
• **Character Agent (Emma):** "My major is **Programming**. I recall stating this at the start of the conversation."
• **Scene Agent (Objective):** Retrieves two relevant logs:
1. [00:19] **Emma** says: "I study Programming."
2. [02:40] **Lily** says: "...I know Emma is good at Linear Algebra."
• **Judge Verdict:** "Evidence 1 is a **direct first-person statement**, whereas Evidence 2 is **hearsay**. Direct claim takes precedence. **Final Answer: Programming** ✓"

Table 5: Qualitative Case Study. ViSAGE successfully resolves the ambiguous pronoun "I" via retroactive refinement, whereas the baseline fails due to surface-level matching.

the first-person pronoun ("I") to the target entity, our system enables the Judge Agent to prioritize direct testimony over third-person claims, effectively preventing hallucination.

**Beyond Accuracy: Response Granularity.** Fig. 4 reveals that correct labels do not imply full comprehension. Employing GPT-5 as a judge, we evaluated three dimensions: (1) **Visual Richness** (density of scene details), where ViSAGE achieves a win rate of 87.3%; (2) **Identity Specification** (pinpointing who performs the action), achieving our highest margin of 91.1%; (3) **Logical Completeness** (providing transparent reasoning steps), with an 85.1% lead. These gaps confirm that while baselines rely on shallow pattern matching (correct label, sparse evidence), ViSAGE constructs grounded narratives with precise entity binding.

**Failure Mode Analysis: The Safety of "Knowing What You Don't Know."** Finally, we analyze failure boundaries by categorizing errors into Hazardous Hallucinations and Safe Refusals.

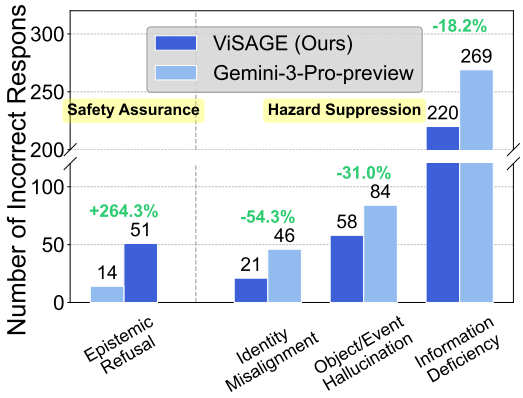


Figure 5: Safety Analysis on M3-Bench-robot. ViSAGE significantly reduces hazardous hallucinations and increases safe epistemic refusals.

Fig. 5 reveals a critical shift in ViSAGE, prioritizing safety in high-stakes scenarios:
(1) **Mitigating Hazardous Failures.** The most dangerous errors in embodied agents are Identity Misalignment (wrongly attributing actions to specific individuals) and Object/Event Hallucination (fabricating non-existent entities). ViSAGE significantly suppresses these risks, reducing Identity Misalignment by 54.3% and Hallucinations by 31.0%, thereby preventing **execution** based on false premises.
(2) **Embracing Safe Refusals.** Instead of hallucinating, ViSAGE exhibits a dramatic 264.3% increase in Epistemic Refusal: the explicit acknowledgment of information deficit. In robotics, this represents a *graceful degradation*: preferring an honest "I don't know" over a confident error is vital for trustworthiness and safety.

## 5 Conclusion

In this paper, we argue that long-form video understanding requires a paradigm shift from strictly sequential processing to a self-correcting memory architecture. By enabling agents to retroactively align historical memory with evolving identity evidence, ViSAGE demonstrates that post-hoc rectification is a structural necessity for maintaining narrative coherence. Beyond accuracy, our framework establishes a new standard for epistemic safety in embodied AI: rather than prioritizing plausible fabrication, it enforces rigorous evidence verification. This transition from hazardous hallucinations to verified refusals is critical for deploying trustworthy agents in real-world environments, ensuring that future systems "know what they don't know."

## Limitations

We acknowledge two primary limitations in the current system. First, our Entity-Anchored Preprocessing is explicitly tailored for human-centric interactions, utilizing `Face_ID` and `Speaker_ID` bindings to resolve complex social dynamics. Consequently, the system presently treats non-human elements, such as plot-critical objects or animals, as background context rather than actively tracked profiles. To address this, our future work aims to generalize the scope of entity registration by incorporating open-vocabulary object tracking and re-identification. This will evolve ViSAGE from a character-centric assistant into a broadly entity-aware agent. Second, relying on commercial APIs implies that strictly reproducing exact numerical outputs may be constrained by the model's closed-source nature. However, to ensure methodological reproducibility, we release all prompt templates and pipeline code. Furthermore, strictly relying on APIs limits local deployment; thus, we plan to address this by fine-tuning open-source models in future work to achieve a fully reproducible and self-contained system.

**AI Assistance Statement.** We acknowledge the use of Large Language Models (LLMs) to assist with text refinement and LaTeX formatting. The authors have verified all content for accuracy and maintain full responsibility for the scientific validity of this work.

## Ethical considerations

This work utilizes public video datasets (M3-Bench, Video-MME) strictly for academic research to enhance the consistency of multimodal agents. Although the proposed framework incorporates face detection and speaker identification modules, these components are employed *solely* to ground entities within the narrative context of the video. We do not collect any new personal data. Furthermore, we explicitly disclaim any intent for this technology to be used for real-world surveillance or biometric profiling without explicit consent and privacy safeguards.

## References


Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. 2022. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*.

Shuai Bai and 1 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631. ArXiv preprint arXiv:2511.21631.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, Shiliang Zhang, and Junjie Li. 2024a. Eres2netv2: Boosting short-duration speaker verification performance with computational efficiency. *arXiv preprint arXiv:2406.02167*.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, and 1 others. 2024b. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer.

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

Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, and 1 others. 2025. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*.

Google DeepMind. 2025. Gemini 3: A new era of intelligence. `https://deepmind.google/models/gemini/`. Accessed: 2025-12-13.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023. Autoad: Movie description in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18930–18940.

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

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. 2025. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32779–32798.

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

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18198–18208.

Glenn Jocher and Jing Qiu. 2024. Ultralytics yolo11. https://github.com/ultralytics/ultralytics. Version 11.0.0.

Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. 2024. Vidcompress: Memory-enhanced temporal compression for video understanding in large language models. *arXiv preprint arXiv:2410.11417*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. 2024. Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. *Advances in neural information processing systems*, 37:49881–49913.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024a. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024b. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699.

Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. 2025. Video-xl-pro: Reconstructive token compression for extremely long video understanding. *arXiv preprint arXiv:2503.18478*.

Zhiwei Liu, Weiran Yao, Jianguo Zhang, Liangwei Yang, Zuxin Liu, Juntao Tan, Prafulla K Choubey, Tian Lan, Jason Wu, Huan Wang, and 1 others. 2024. Agentlite: A lightweight library for building and advancing task-oriented llm agent system. *arXiv preprint arXiv:2402.15538*.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. 2025. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *arXiv preprint arXiv:2508.09736*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244.

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Kai Mei, Xi Zhu, Wujiang Xu, Wenyue Hua, Mingyu Jin, Zelong Li, Shuyuan Xu, Ruosong Ye, Yingqiang Ge, and Yongfeng Zhang. 2024. Aios: Llm agent operating system. *arXiv preprint arXiv:2403.16971*.

OpenAI. 2025. Gpt-5 technical report. *arXiv preprint arXiv:25xx.xxxxx*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, and 1 others. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*.

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

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. 2021. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3927–3935.

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

Endel Tulving. 2002. Episodic memory: From mind to brain. *Annual review of psychology*, 53(1):1–25.

Ying Wang, Yanlai Yang, and Mengye Ren. 2023. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. *arXiv preprint arXiv:2312.05269*.

Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, and 1 others. 2024. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 2022. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13587–13597.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.

Jin Xu, Z Guo, H Hu, Y Chu, X Wang, J He, Y Wang, X Shi, T He, and X Zhu. 2025a. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025b. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025c. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, and 1 others. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. 2024. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*.

Shiliang Zhang, Cong Liu, Hui Jiang, Si Wei, Lirong Dai, Yu Hu, and Lihe Niu. 2018. Deep feedforward sequential memory networks for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6358–6362. IEEE.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

## A Implementation Details

Here, we provide the implementation details of the tools for representation extraction introduced in Section 4.2. Unless otherwise stated, hyperparameters were determined empirically based on preliminary experiments to balance performance and efficiency.

**Entity-anchored Preprocessing.** We process the raw video at its original frame rate. For person detection, we utilize the YOLO11l model (Jocher and Qiu, 2024) with a confidence threshold of 0.2 and an Intersection over Union (IoU) threshold of 0.2 for Non-Maximum Suppression. Subsequently, we employ BoT-SORT (Aharon et al., 2022) for multi-object tracking. The tracker is configured with high and low detection thresholds of 0.5 and 0.1, respectively, and a new track initiation threshold of 0.6. We set the track buffer to 60 frames and the match threshold to 0.8. The system enables score fusion and utilizes sparse optical flow (`sparseOptFlow`) for Global Motion Compensation (GMC), with proximity and appearance thresholds set to 0.5 and 0.8, respectively.

**Face Processing and Identity Embedding.** For each generated tracking sequence (tracklet), we employ the `buffalo_l` predefined model suite from the InsightFace library (Deng et al., 2019)[1] to extract facial attributes, including bounding box coordinates, identity embeddings, and detection/quality scores. To optimize computational efficiency while ensuring data quality, we implement an early-stopping mechanism: processing for a tracklet halts once three frontal faces with quality scores exceeding 0.8 are detected. The identity embedding corresponding to the face with the highest quality score among these candidates is then selected as the representative embedding for the entire tracklet. For the extracted visual identity embedding, we perform a nearest-neighbor search in the Object Database using a cosine similarity threshold of 0.6. If the highest similarity score exceeds this threshold, the embedding is linked to the existing entity; otherwise, a new entity is initialized.

**Audio Pipeline.** Our audio pipeline begins with Voice Activity Detection (VAD) using the `FSMN-VAD` model (Zhang et al., 2018)[2] to isolate valid speech segments. For each segment, we employ `Qwen-3-Omni-Flash` (Xu et al., 2025a) for Automatic Speech Recognition (ASR) to refine the timestamp boundaries of the spoken utterances (the prompt is shown below). Subsequently, we extract speaker embeddings (voiceprints) using the `ERes2NetV2` model (Chen et al., 2024a)[3].

To assign these audio embeddings to the correct entity, we adopt a **hierarchical matching strategy**:

1. We first attempt to match the embedding against the Object Database with a cosine similarity threshold of 0.6.

2. If no match is found, we utilize TalkNet (Tao et al., 2021) to perform Active Speaker Detection (ASD). We evaluate whether any visual entity appearing within the audio timeframe is speaking (confidence $> 0.5$). If confirmed, the audio is attributed to that visual entity.

3. If neither method yields a match, a new entity is created.

**Search.** For textual memory retrieval, we perform Maximum Inner Product Search (MIPS) to match the input query against all stored text nodes. We utilize Qwen's `text-embedding-v4` (Yang et al., 2025) as the embedding backbone and retrieve the top-5 candidates to capture comprehensive context.

**The Prompt for Voice Processing**

```
Please perform Speaker Diarization and
Automatic Speech Recognition (ASR) on
this audio.
Requirements:
1. Identify different speakers in the
audio
2. Mark the speaking time period for
each speaker (start_time and end_time)
as numeric seconds with one-decimal
precision (e.g., 83.2 for 1 minute 23.2
seconds). Provide seconds only (no
hour/minute string fields).
3. Transcribe the speech content of each
speaker (transcription)
4. Assign a unique ID to each speaker
(speaker_id, starting from 0)
5. Extract all personal names mentioned
anywhere in the transcriptions (across
all segments) and list them under
"mentioned_names" (deduplicated).
IMPORTANT: If no personal names are
mentioned in any transcription,
"mentioned_names" must be an empty array
```

*(Continued...)*

[1] https://github.com/deepinsight/insightface

[2] https://www.modelscope.cn/models/iic/speech_fsmn_vad_zh-cn-16k-common-pytorch/summary

[3] https://www.modelscope.cn/models/iic/speech_eres2netv2_sv_zh-cn_16k-common/summary

```
[].
6. IMPORTANT: output English only
7. If there is no information, output
empty arrays []
8. When a single speaker is talking,
prefer breaking into shorter sentences;
split long utterances into multiple
sentences when reasonably possible.
Avoid unnecessarily long run-on
sentences. However, ensure each segment
duration is at least 3.0 seconds; if a
split would produce a segment shorter
than 3.0 seconds, merge it with adjacent
content to satisfy the minimum length.
Please return the results strictly in
the following JSON format (do not add
any extra text):
{
    "segments": [
        {
            "start_time": 0.0,
            "end_time": 5.0,
            "speaker_id": 0,
            "transcription": "John's
favorite fruit is mango"
        },
        {
            "start_time": 6.0,
            "end_time": 10.0,
            "speaker_id": 1,
            "transcription": "Hi Mary,
how are you?"
        }
    ],
    "mentioned_names": [
        "John",
        "Mary"
    ]
}
```

# B Prompt Templates

## B.1 Memory Construction Prompt

This prompt guides the MLLM to convert annotated frames into structured logs.

**System Prompt: Memory Generation**

```
You are the Central World Model for an
advanced video understanding system. You
are processing a specific segment (clip)
of a longer video to build a structured,
evolving memory database.

INPUT DATA:
1. Visual Frames: Chronological images
with person bounding boxes. Tags like
<p_1*> indicate p_1 is currently
speaking.
2. Subtitles: Real-time speech
transcription.
3. Diarization Timeline:
{available_person_ids_str}
{diarization_timeline}
```

*(Continued...)*

```
4. Timing Context: This specific video
clip starts at timestamp
[{current_clip_start_time}] (HH:MM:SS)
relative to the beginning of the full
video.
   - The diarization timeline shows
absolute times (global time from the
beginning of the video) in "HH:MM:SS"
format.
   - You should use these absolute times
directly in your incident_memory
timestamps.
5. Previous Knowledge:
{person_profile_and_state}
   - Recent Context: {recent_incidents}

YOUR TASK:
Analyze the provided images and text to
generate a JSON output containing THREE
distinct memory blocks.

### 1. incident_memory (Chronological
Event Log)
Generate a strictly chronological list
of events for the "Scene Agent".
- Format: An array of strings. EVERY
line must start with a standardized
timestamp "[HH:MM:SS]".
- Timing Logic: CRITICAL: The
diarization timeline already shows
absolute times. Use these times directly.
  - Example: If timeline shows
[00:01:08], use [00:01:08].
- Content: - Merge Action & Dialogue.
  - Visual Grounding: EXPLICITLY
describe visual details.
  - Objective Tone: Describe WHAT
happened, WHO did it, and WHERE.

### 2. object_memory (Structured
Incremental Updates)
Generate structured updates for each
person.
- p_id: The ID (e.g., "<p_2>").
- state_snapshot (The "Now"):
  - Describe Visual Appearance,
Location, and Action.
  - Incremental Logic: If identical to
previous state, output "".
- profile_delta (The "New Knowledge"):
  - Extract ONLY NEW long-term traits.
  - Incremental Logic: If trait is
already known, output "".

### 3. merge_memory (Person Identity
Merging)
Identify which person IDs refer to the
same real-world person.
- Strict Format: "Merge: <p_x>, <p_y>"
(comma separated).

STRICT CONSTRAINTS:
1. JSON Only: Output valid JSON. No
markdown formatting.
2. Entity Consistency: Always use <p_X>
format.
3. No Hallucination: If you don't see an
object, don't invent it.
```

*(Continued...)*

```
EXPECTED JSON OUTPUT FORMAT:
{
  "incident_memory": [
    "[{current_clip_start_time}] <p_2>
enters the room.",
    "[HH:MM:SS] <p_2> says 'The meeting
starts now.'"
  ],
  "object_memory": [
    {
      "p_id": "<p_2>",
      "state_snapshot": "Wearing a black
suit.",
      "profile_delta": "Revealed he is
the team leader."
    }
  ],
  "merge_memory": [
    "Merge: <p_2>, <p_5>"
  ]
}
```

### B.2 Scene Agent Prompt

This prompt guides the Scene Agent to generate an objective answer based strictly on the retrieved visual logs. It serves as the "Antithesis" (Visual Fact) in the dialectic verification process.

**System Prompt: Objective Scene Testimony**

```
Based on the following video
description, answer the question.
Please provide both your reasoning and
the final answer.

Question: {question}
Relevant Incident log:
{retrieved_incidents}

Please provide your answer in the
following format:

Reasoning: [Your step-by-step reasoning
process based on the incident log]
Answer: [Your direct answer to the
question]

Reasoning:
```

### B.3 Character Agent Prompt

This prompt instantiates a specific Character Agent (serving as the "Thesis") during the dialectic verification stage.

**System Prompt: Character Agent Simulation**

```
### ROLEPLAY INSTRUCTION
You are NOT an AI assistant. You are
{character_name} (ID: {primary_id}).
You are currently in a "Simulated
Witness Interview".

### YOUR PROFILE (Who you are)
{merged_profile_text}

### YOUR MEMORY STREAM (What you
experienced)
Read these logs. These are your DIRECT
experiences. You cannot 'see' what
happened when you were not there.
{filtered_incident_logs}

### VISUAL EVIDENCE (Your Eyes)
At the time of the event, you were:
{current_visual_state}

### THE QUESTION
The Judge asks: "{user_question}"

### YOUR ANSWER
Answer in the first person ("I").
- Stick strictly to your memory logs.
- If the question uses a wrong name for
you (e.g., "Malloy" instead of "Mario"),
assume they mean you, but politely
correct them if needed.
- If you don't know, say "I didn't see
that."
- BE CONCISE. Just state the facts.
```

### B.4 Judge Agent Prompt

This prompt guides the Judge Agent to synthesize the conflicting perspectives (Thesis vs. Antithesis) into a final verified verdict.

**System Prompt: Judicial Adjudication**

```
User Question: "{user_question}"

{witness_testimonies}

### TASK
Synthesize a final answer for the user
by conducting a Dialectic Verification.

1. Resolve Conflicts (Cross-Examination):
   - Compare the testimonies.
   - IF there is a conflict regarding
physical actions, appearance, or
location, PRIORITIZE the Scene Agent, as
it represents objective visual evidence.
   - IF the question is about internal
intent or specific dialogue nuances,
consider the Character Agent's
perspective but verify it against visual
context.
```

*(Continued...)*

```
2. Answer the Question: Directly address
the user's prompt.

3. Show Your Work: Briefly mention
"According to..." or "Visual evidence
confirms..." to make the answer
convincing.

Final Answer:
```

### B.5 Prompt for Qualitative Side-by-Side Comparison

In addition to error taxonomy, we evaluated the relative quality of correct responses. To do this, we employed a pairwise comparison prompt where an expert evaluator assesses two correct answers based on Comprehensiveness, Diversity, and Empowerment. This ensures that even when baselines do not hallucinate, we can quantify the superior richness and helpfulness of ViSAGE's responses. Crucially, to ensure fairness, we address positional bias (where LLMs may favor the first-appearing text): we systematically swap the order of the answers within the prompt to prevent the evaluator from showing preference based on presentation sequence.

**Evaluation Prompt: Qualitative Side-by-Side Comparison**

```
You are an expert evaluator comparing
two answers to the same question. Both
answers are correct, but you need to
evaluate which one is better in each of
the following 3 dimensions:

**Question:**
{question}

**Answer 1 :**
{answer1}

**Answer 2 :**
{answer2}

Please evaluate and compare the two
answers in the following 3 dimensions:

1. **Visual Richness**:
   - **Definition**: This measures the
**density of scene details**.
   - **Criteria**: Which answer provides
a more vivid, detailed, and descriptive
account of the visual elements (e.g.,
objects, environment, textures, colors)?
Look for the answer that goes beyond
sparse evidence to capture the richness
of the scene.
```

*(Continued...)*

```
2. **Identity Specification**:
   - **Definition**: This measures the
precision in **pinpointing who performs
the action**.
   - **Criteria**: Which answer better
identifies specific entities or agents?
Look for precise **entity binding**
(e.g., distinguishing between "a man"
and "the man in the red shirt" or
correctly attributing actions to
specific characters) rather than using
vague or ambiguous references.

3. **Logical Completeness**:
   - **Definition**: This measures the
quality of **providing transparent
reasoning steps**.
   - **Criteria**: Which answer
constructs a **grounded narrative** with
a clear chain of thought? Look for the
answer that explains the "why" and "how"
behind the conclusion, rather than just
performing shallow pattern matching or
giving a label without justification.

For each dimension, choose which answer
is better: "Answer 1" or "Answer 2", or
"Tie" if they are equally good.

Please respond in the following JSON
format:
{{
    "Visual Richness": "Answer 1" or
"Answer 2" or "Tie",
    "Identity Specification": "Answer 1"
or "Answer 2" or "Tie",
    "Logical Completeness": "Answer 1"
or "Answer 2" or "Tie",
    "reasoning": {{
        "Visual Richness": "Brief
explanation (1-2 sentences)",
        "Identity Specification": "Brief
explanation (1-2 sentences)",
        "Logical Completeness": "Brief
explanation (1-2 sentences)"
    }}
}}

Respond ONLY with valid JSON, no
additional text.
```

### B.6 Prompt for Automated Error Diagnosis

To ensure a systematic and reproducible evaluation of failure modes, we designed a rigorous prompt for automated error diagnosis. This prompt instructs an evaluator LLM to audit the discrepancy between the Model Prediction and the Ground Truth given the Video Context. It enforces a clear and mutually exclusive taxonomy of errors, categorizing failures into distinct classes such as Identity Misalignment, Object Hallucination, or Epistemic Refusal. This structured classification minimizes ambiguity between error types. We pro-

vide the full text of this diagnostic prompt below to facilitate future benchmarking of agent trustworthiness.

**Evaluation Prompt: Automated Error Taxonomy**

```
You are an expert in analyzing video
understanding errors. Given the Ground
Truth and a Model's Prediction, classify
the error into ONE of the following 4
categories:

1. Epistemic Refusal: The model admits
uncertainty or inability to answer
(e.g., "I don't know", "impossible to
determine", "no information available",
"cannot be determined from the video").

2. Identity Misalignment: The answer
contains correct actions but attributes
them to the WRONG person/character. This
includes confusing entities, mixing up
names, or misunderstanding "who did
what."

3. Object/Event Hallucination: The
answer describes objects or events that
are physically non-existent or
completely factually incorrect (e.g.,
describing a car crash when none
happened).

4. Information Deficiency: The answer is
partially correct but misses key details
or specific keywords required by the
Ground Truth (e.g., correct action but
missing the object acted upon).

**Ground Truth:**
{ground_truth}

**Model Prediction:**
{model_prediction}

Please analyze the error and respond in
the following JSON format:
{{
    "category": "1" or "2" or "3" or "4",
    "category_name": "Epistemic Refusal"
or "Identity Misalignment" or
"Object/Event Hallucination" or
"Information Deficiency",
    "reasoning": "Brief explanation of
why this category was chosen (1-2
sentences)"
}}

Respond ONLY with valid JSON, no
additional text.
```

### B.7 Prompt for Answer Correctness Evaluation

To assess the veracity of the generated responses, we move beyond rigid string matching, which often penalizes valid but differently phrased answers. Instead, following the established evaluation protocol of M3-Agent (Long et al., 2025), we employ a Logic-Based Entailment strategy. Our prompt is aligned with the rigorous standards of M3-Agent, instructing the evaluator to determine whether the ground truth can be reasonably inferred from the agent's response. This validation mechanism focuses on semantic equivalence, ensuring that the agent is credited for correct reasoning even if the surface realization differs from the reference.

**Evaluation Prompt: Semantic Correctness**

```
You are provided with a question, a
ground truth answer, and an answer from
an agent model. Your task is to
determine whether the ground truth
answer can be logically inferred from
the agent's answer, in the context of
the question.

Do not directly compare the surface
forms of the agent answer and the ground
truth answer. Instead, assess whether
the meaning expressed by the agent
answer supports or implies the ground
truth answer. If the ground truth can be
reasonably derived from the agent
answer, return "Yes". If it cannot,
return "No".

Important notes:
- Do not require exact wording or
matching structure.
- Semantic inference is sufficient, as
long as the agent answer entails or
implies the meaning of the ground truth
answer, given the question.
- Only return "Yes" or "No", with no
additional explanation or formatting.

Input fields:
- question: the question asked
- ground_truth_answer: the correct answer
- agent_answer: the model's answer to be
evaluated

Now evaluate the following input:

Input:
- question: {question}
- ground_truth_answer: {expected_answer}
- agent_answer: {student_answer}

Output ('Yes' or 'No'):
```

### B.8 Reliability and Robustness Analysis

We address concerns regarding the reliance on LLM-based evaluators and the statistical robust-

**Algorithm 1** ViSAGE: Bidirectional Memory Refinement Pipeline

**Require:** Video stream $\mathcal{V}$, Pre-trained MLLM
**Ensure:** Incident Logs $\mathcal{L}$, Object Cards $\mathcal{O}$

Initialize $\mathcal{L} \leftarrow \emptyset$, $\mathcal{O} \leftarrow \emptyset$
**for** each time step $t$ in $\mathcal{V}$ **do**
  **// Step 1: Forward Grounding**
  Extract visual cues $v_t$ and audio cues $a_t$
  Retrieve relevant history: $ctx \leftarrow \text{Retrieve}(\mathcal{L}, \mathcal{O}, v_t)$
  Generate current memory: $\mathcal{L}_t, \mathcal{O}_t, \mathcal{M}_{merge} \leftarrow \text{MLLM}(v_t, a_t, ctx)$
  Append new records: $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{L}_t$
  Update profiles: $\mathcal{O} \leftarrow \text{Update}(\mathcal{O}, \mathcal{O}_t)$
  **// Step 2: Retroactive Rectification**
  **if** $\mathcal{M}_{merge}$ is not empty **then**
    RETROACTIVEUNIFICATION($\mathcal{L}, \mathcal{O}, \mathcal{M}_{merge}$)
  **end if**
**end for**
**return** $\mathcal{L}, \mathcal{O}$

**Algorithm 2** Retroactive Identity Unification

**Require:** Incident Logs $\mathcal{L}$, Object Cards $\mathcal{O}$, Merge Signals $\mathcal{M}$

**Input Parsing:** Extract equivalence groups $\mathcal{G}$ from $\mathcal{M}$ (transitive closure of $ID_i \equiv ID_j$)
**for** each group $G \in \mathcal{G}$ **do**
  **// Selection Strategy**
  Sort IDs in $G$ to find canonical anchor $id^*$ (e.g., earliest appearance)
  Let $id_{sources} \leftarrow G \setminus \{id^*\}$
  **// 1. Consolidate Object Cards (Data Merging)**
  **for** each $id_{src} \in id_{sources}$ **do**
    Retrieve source card $O_{src} \leftarrow \mathcal{O}[id_{src}]$
    Merge attributes: $\mathcal{O}[id^*].\text{Visual} \leftarrow \mathcal{O}[id^*].\text{Visual} \cup O_{src}.\text{Visual}$
    Merge aliases: $\mathcal{O}[id^*].\text{Names} \leftarrow \text{Unique}(\mathcal{O}[id^*].\text{Names} \cup O_{src}.\text{Names})$
    Remove source: Delete $\mathcal{O}[id_{src}]$
  **end for**
  **// 2. Global Backward Update (Incident Logs)**
  **for** each event $e \in \mathcal{L}$ from $t = 0$ to $t_{now}$ **do**
    **if** $e.\text{participants} \cap id_{sources} \neq \emptyset$ **then**
      Replace occurrences of $id_{src}$ with canonical $id^*$ (and update alias string)
      $e.\text{content} \leftarrow \text{Rewrite}(e.\text{content}, id^*)$
    **end if**
  **end for**
**end for**
**return** Updated $\mathcal{L}, \mathcal{O}$

ness of our comparisons. While we acknowledge the challenges of evaluating open-ended generation, our experimental design explicitly mitigates these risks through the following three methodological standards:

**1. Alignment with Established Protocols.** To ensure the validity of the LLM-as-a-Judge paradigm, we strictly adhered to the Logic-Based Entailment protocol established by the previous state-of-the-art, M3-Agent (Long et al., 2025). Unlike subjective scoring (e.g., rating 1-10), our evaluator (Prompt B.7) performs a binary semantic verification: it checks if the ground truth is *logically inferable* from the prediction. This reduces the complexity of the judgment task to objective fact-checking, minimizing the hallucination rate of the

| Phase | System Process & Evidence |
|---|---|
| **User Query** | *"On which coat rack should Emma's coat be placed, the taller one or the shorter one?"* |
| **Dispatcher Logic** | The system correctly identified the intent and recruited relevant agents based on the query entities:<br>• **Intent:** Determine the correct coat rack (taller or shorter) for Emma's coat.<br>• **Target Agents:** <p_7> (Emma), Scene_Agent.<br>• **Reasoning:** The question concerns the placement of Emma's (<p_7>) coat. Emma is the primary agent regarding her own belongings and actions. The Scene_Agent is included to verify environmental details related to the coat racks. |
| **Agent Execution** | Agents queried their memory banks, revealing the lack of fine-grained visual details:<br>• **Agent Emma (<p_7>):** "Based on my memory, I walked from the dining room to the bedroom around 00:05:30. I removed my yellow hooded jacket and then lay down on the bed. My logs do not mention me placing the coat on a rack, nor do I recall seeing a taller or shorter coat rack. I simply took the jacket off before lying down."<br>• **Scene Agent:** "I saw the Robot hang the yellow jacket on a white coat rack near the window [00:07:13]. However, my memory logs do not specify if it was a taller or shorter rack, or if there were multiple racks to choose from." |
| **System Output** | **Epistemic Refusal:** "Based on the available testimonies, it is impossible to determine..." (The Judge correctly identified that neither agent provided the comparative attribute 'taller/shorter'). |
| **Failure Analysis** | **Modality Information Loss.** The text captioning process successfully encoded the object class ("coat rack") but discarded the fine-grained visual attribute (relative height) required to distinguish between the two racks. |

Table 6: Failure Case Study: Granularity Loss in Textual Memory.

judge itself. Since our method outperforms M3-Agent by a significant margin (+5.9% on average and up to +27.1% on Cross-Modal tasks) using the exact same evaluation protocol, the relative ranking remains robust regardless of potential evaluator noise.

**2. Standardization and Fairness.** To ensure fair comparison, we addressed the "different pipeline" concern by standardizing the input constraints. For all Discrete Socratic baselines (e.g., Qwen3-VL, GPT-4o), we utilized the same sampling rate (0.5 fps) and the same ASR-generated transcripts as the base input. Our performance gains primarily stem from ViSAGE's architectural contribution-specifically, the *Entity-Anchoring Preprocessing* and *Bidirectional Refinement*, rather than unfair advantages in raw data quality. The comparison is thus "System vs. System", verifying that our structured memory processing is superior to standard context-window approaches under identical video inputs.

**3. Deterministic Reproducibility.** Regarding statistical variance, we eliminated generation stochasticity by enforcing a deterministic decoding strategy (Temperature $= 0$) for both the Agent's reasoning and the Judge's evaluation across all runs. Given the high cost of long-video benchmarks, this deterministic setting is the standard practice in recent literature (Fu et al., 2025; Long et al., 2025) to ensure reproducibility. Furthermore, the magnitude of our improvement (e.g., +15.9% on robot-CM) far exceeds the typical variance range of LLMs (usually $\pm 1 - 2\%$), statistically confirming that the gains are non-trivial and not artifacts of random sampling.

## C Algorithmic Details

In this section, we provide the formal algorithms for the Bidirectional Memory Refinement process described in Section 3.3. Algorithm 1 outlines the overall framework, while Algorithm 2 details the retroactive identity unification mechanism.

## D Failure Case Analyses

**Analysis: The Limitation of Human-Centric Anchoring.** While there is an inherent bottleneck in compressing visual details into text (i.e., the caption omitted the relative height "taller" or "shorter"), the root cause of this failure lies in the scope of our Entity-Anchored Preprocessing. Currently, our system is explicitly tailored for human-centric interactions, strictly utilizing `Face_ID` and `Speaker_ID` bindings to resolve complex social dynamics between characters. Consequently, the system treats non-human elements, such as the coat rack in this case, as ephemeral background context rather than actively tracked profiles. Because the coat rack was not registered as a distinct entity with its own lifecycle, its static attributes (height comparisons) were not preserved in the memory logs, leaving the agents unable to resolve the specific object reference required by the user's query (as illustrated in Table 6).

**Future Direction: Towards Generalized Entity Awareness.** To bridge this gap, future research must evolve ViSAGE from a character-centric assistant into a broadly Entity-Aware Agent.

- **Generalizing Registration:** We plan to expand the scope of entity registration beyond human actors to incorporate open-vocabulary object tracking and re-identification.
- **Object Lifecycle Management:** By maintaining distinct memory profiles for any significant narrative element (e.g., plot-critical objects or animals), the system will be able to attribute and retrieve fine-grained properties for inanimate objects just as effectively as it currently does for human behaviors.