OMNIMEMORY: TOWARDS UNIFIED OMNIMODAL MEMORY FOR LONG-HORIZON INTERACTIVE ASSISTANTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon interactive assistants must repeatedly reason over audiovisual experiences that accumulate beyond finite context windows. Existing memory methods often prioritize visual observations and speech transcripts, underrepresenting richer auditory cues. Fragmented context, persistent interpretation errors, and perceptual information loss further limit memory reliability. We introduce OmniMemory, a training-free framework that builds evolving audiovisual memory with a unified omnimodal model. Progressive construction propagates scene and entity context across clips to build Episodic Memory, explicitly preserving visual events, speaker associations, utterance-aligned paralinguistic attributes, environmental sounds, and music. Retrospective auditing uses later evidence to revise earlier identity assignments and update affected records, while consolidation distills reusable knowledge into Core Memory. Query-aware routing reuses textual memories and selectively revisits source audiovisual clips in Native Memory when additional perceptual evidence is needed. Evaluations on M3-Bench-Robot, M3-Bench-Web, and VideoOdyssey-AV demonstrate the highest overall accuracy among evaluated systems, with substantial gains over memory baselines on Sound and Music questions. Ablations demonstrate the benefits of auditory fields, retrospective auditing, and complementary memory sources, while efficiency analysis shows a favorable trade-off among accuracy, construction overhead, and query costs. Anonymous code is available at: `https://anonymous.4open.science/r/OmniMemory-94E2`

1 INTRODUCTION

An interactive omnimodal agent does not experience the world one query at a time. It continuously observes people, conversations, events, sounds, and their interactions, while future questions may refer to any of these observations long after they occurred. Advanced omnimodal models (Team, 2026; Cui et al., 2026) provide a foundation for audiovisual assistants in embodied and egocentric settings (Fan et al., 2025; Yang et al., 2026). This setting fundamentally differs from conventional video understanding, where the relevant evidence is typically contained within a bounded input. As interactions accumulate over potentially unbounded horizons, an agent must transform a continuous stream of audiovisual experience into memory that remains useful for future reasoning despite finite context windows and limited computation (Yang et al., 2025; Long et al., 2026). The central challenge is therefore not simply how to retrieve past information, but *how to construct, evolve, and access memory from continuously arriving omnimodal experience*.

This challenge is particularly important for audiovisual interaction because the information required by future queries is not known when an observation is first encountered. Visual observations may identify people and events, while speech conveys explicit content; meanwhile, speaker identity, vocal tone, laughter, background sounds, and music can provide information that is equally important for interpreting what happened. For example, a later query may ask who made an utterance, whether a statement was delivered sarcastically, or what instrument was being played in the background (Figure 1). A memory restricted to visual descriptions and speech transcripts can preserve semantic content while losing the perceptual evidence needed to answer such questions. Effective omnimodal

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.

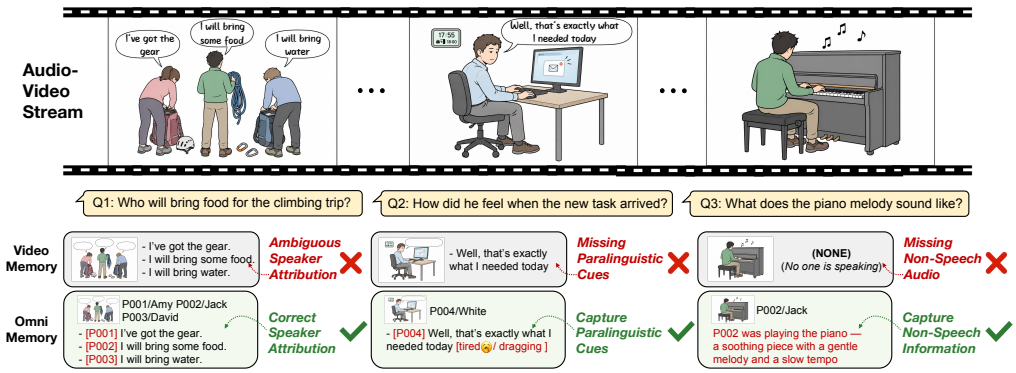


Figure 1: **OmniMemory preserves auditory information beyond transcripts.** The examples contrast a memory representation limited to visual observations and ASR transcripts with OmniMemory, which explicitly retains speaker associations (left), paralinguistic attributes paired with utterances (middle), and non-speech acoustic information (right) for subsequent recall.

memory therefore needs to preserve not only *what was said or seen*, but also *who said it, how it was said, what was happening around it, and the underlying audiovisual evidence*.

Existing approaches to long-horizon video understanding primarily address memory through visual caches, feature banks, or external retrieval stores (Chen et al., 2026b; Luo et al., 2025; Zhang et al., 2024b; Fan et al., 2024). Recent memory-based systems such as WorldMM and HippoMM (Yeo et al., 2026; Lin et al., 2026) organize visual observations and speech transcripts into retrievable memories. M3-Agent (Long et al., 2026) still relies on separate ASR and speaker identification, exposing memory construction to upstream errors. More broadly, existing memory systems remain vulnerable to fragmented cross-clip context, persistent interpretation errors, and perceptual information loss during textualization, highlighting three properties of continuous audiovisual experience. First, *context is accumulated over time*: an entity or event may span multiple observations and cannot always be fully understood from an individual clip. Second, *memory is revisable*: later observations may provide evidence that clarifies an earlier transcription, speaker assignment, or event interpretation. Third, *memory is inherently lossy*: textualized memories may omit perceptual details that are irrelevant to one query but essential to another. These properties suggest that effective omnimodal memory should not be a static archive, but a continuously evolving representation that can be refined and, when necessary, grounded back to its original audiovisual evidence.

We introduce **OmniMemory**, a training-free framework for building such evolving memory from continuous audiovisual interactions. We treat memory as a *progressive, revisable, and evidence-grounded representation* rather than a collection of independent retrieval records. OmniMemory maintains three complementary forms of memory. *Episodic Memory* preserves detailed audiovisual experiences as the primary evidence layer, explicitly recording visual events, entity references, utterance-aligned paralinguistic attributes, environmental sounds, and music. *Core Memory* consolidates stable information across episodes so that frequently reused knowledge can be accessed without repeatedly reconstructing it from individual observations. *Native Memory* retains the original audiovisual clips, providing perceptual evidence when textual memories are insufficient.

Building such memory requires more than storing these representations. We therefore develop three mechanisms corresponding to the evolving nature of audiovisual experience. First, *Progressive Memory Construction* propagates contextual state and entity information across incoming observations, allowing fragmented evidence to be accumulated into coherent memories as interactions unfold. Second, *Retrospective Consolidation & Auditing* allows later evidence to revise and validate earlier records, addressing ambiguities that cannot be resolved at the moment an observation is first processed. This makes memory *revisable rather than append-only*. Third, *Query-Aware Memory Routing* dynamically determines the appropriate level of evidence for each query: consolidated knowledge is used when sufficient, detailed episodic observations are retrieved when necessary, and the original audiovisual content is replayed only when the available textual memories cannot support reliable reasoning. In this way, OmniMemory separates the persistence of experience from the cost of repeatedly perceiving it, while retaining access to the original evidence when needed.

We evaluate OmniMemory on M3-Bench-Robot, M3-Bench-Web, and VideoOdyssey-AV. We compare against agentic, retrieval-augmented, and memory-based approaches, with full-video base models serving as references on the finite benchmark inputs. OmniMemory achieves the highest overall accuracy among the evaluated methods across all three benchmarks, with particularly pronounced improvements on Sound and Music questions. These results demonstrate that explicitly preserving auditory information beyond transcripts is important for long-horizon omnimodal understanding. Further ablations demonstrate the performance benefits of progressively constructed audiovisual memories, retrospective auditing, and query-aware access to complementary memory sources. We analyze the trade-off between memory construction overhead and online audiovisual replay.

Our main contributions are summarized as follows:

- We formulate long-horizon omnimodal memory as a *progressive, revisable, and evidence-grounded* construction problem, and introduce **OmniMemory**, a training-free framework that transforms continuous audiovisual experience into evolving, reusable memories.
- We develop a unified memory pipeline that progressively accumulates contextual evidence, retrospectively audits and consolidates earlier memories, and adaptively routes queries from consolidated knowledge to detailed episodes and native audiovisual evidence.
- We preserve auditory evidence beyond transcripts, including speaker associations, utterance-aligned paralinguistic attributes, environmental sounds and music, and show gains on three audiovisual benchmarks, especially in non-speech auditory understanding.

2 Related Work

Long-Horizon Multimodal Understanding. Long-horizon video understanding requires preserving task-relevant evidence under limited context and computation. Compact representations and streaming memory expand temporal coverage by reducing the information processed or retained at each step (Song et al., 2024; Shen et al., 2024; Li et al., 2026; Zhang et al., 2024b). Retrieval and active perception complement these strategies by making detailed evidence selectively accessible according to the current query (Luo et al., 2025; Jeong et al., 2025; Xue et al., 2025; Xing et al., 2026). Recent studies extend this problem to day-, week-, and month-scale recordings (Tian et al., 2025; Chen et al., 2026a). For potentially unbounded streams, however, bounded representations can lose details needed by future queries, while query-driven retrieval alone does not determine how an ever-growing history is retained and organized. This motivates persistent external memory that incrementally transforms observations into records reusable across contexts and queries.

Memory-Augmented Multimodal Agents. Agent memory systems organize accumulated information for reuse across interactions (Zhong et al., 2024; Chhikara et al., 2025; Xu et al., 2025b). Multimodal systems extend this principle by combining semantic abstractions with entity-linked or modality-specific evidence (Long et al., 2026; Yeo et al., 2026; Lin et al., 2026). The usefulness of such memory depends on both the audiovisual information it retains and the connections it establishes across observations. Speech content alone cannot capture paralinguistic cues or non-speech sounds that may be relevant to subsequent questions, making the representation of auditory experience an important part of memory construction. Meanwhile, persistent entity representations and contextual memory allow observations from different clips to contribute to coherent identities and narratives (Long et al., 2026; Fan et al., 2025; Zhang et al., 2024a; He et al., 2024). OmniMemory brings these considerations together through entity-linked episodic records that explicitly describe speech content, paralinguistic cues, and non-speech sounds, while retaining source audiovisual clips for direct inspection when textual memories provide insufficient evidence.

Native Omnimodal Models. Native omnimodal models such as Qwen3.5-Omni (Team, 2026) and MiniCPM-o 4.5 (Cui et al., 2026) jointly process text, vision, and audio, providing a foundation for audiovisual assistants. Recent benchmarks further evaluate joint audiovisual understanding across diverse tasks and temporal scales (Li et al., 2025; Geng et al., 2025; He et al., 2026). For continuously operating assistants, these perceptual capabilities need to be complemented by mechanisms that retain and update relevant experiences beyond the active context. OmniMemory builds on native omnimodal perception with a persistent memory layer that preserves auditory cues, connects observations across clips, and supports revision of stored records.

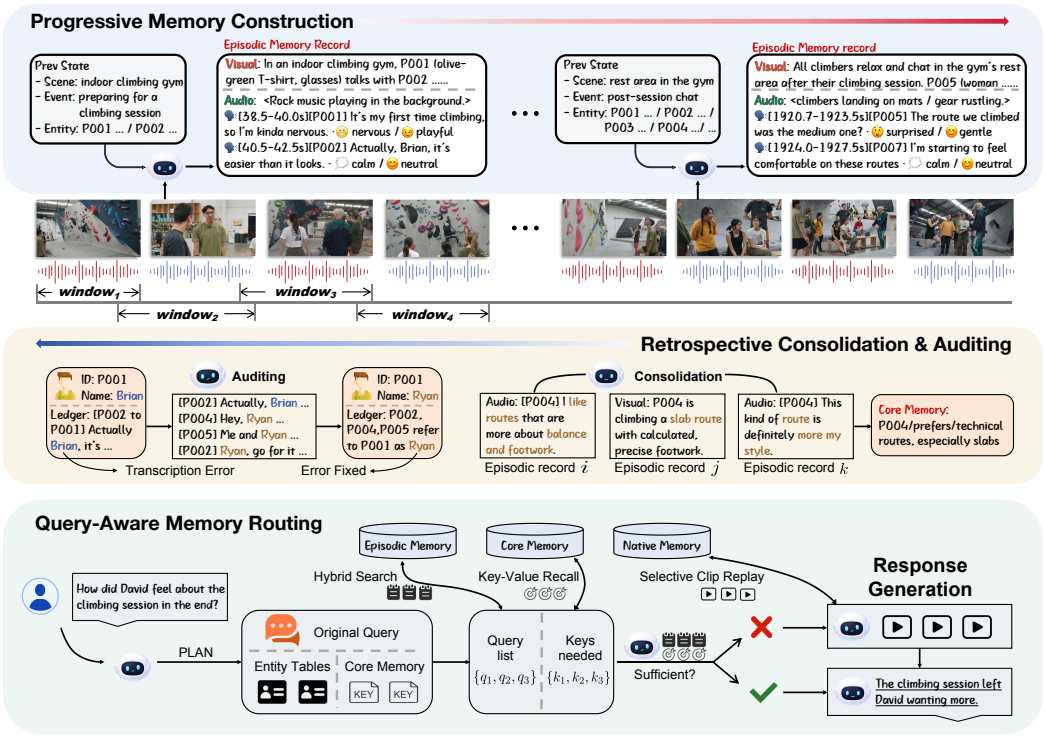


Figure 2: **Overview of OmniMemory. Progressive Memory Construction** builds Episodic Memory through joint audiovisual extraction from overlapping clips with propagated context. **Retrospective Consolidation & Auditing** uses cross-clip evidence to refine records and distill stable knowledge into Core Memory. **Query-Aware Memory Routing** retrieves relevant episodic and core memories and replays source clips from Native Memory when stored evidence is insufficient.

3 OMNIMEMORY

As illustrated in Figure 2, we maintain three complementary memory banks: Episodic Memory $\mathcal{M}^{\text{episodic}}$ stores contextual audiovisual observations as textual records, explicitly preserving auditory information beyond speech transcripts; Core Memory $\mathcal{M}^{\text{core}}$ stores stable knowledge distilled from accumulated experiences; and Native Memory $\mathcal{M}^{\text{native}}$ retains source audiovisual clips. As new clips arrive, *Progressive Memory Construction* builds episodic records while propagating contextual state and entity information across clips. Periodic *Retrospective Consolidation & Auditing* revises existing records using cross-clip evidence and consolidates stable knowledge into Core Memory. When a query arrives, *Query-Aware Memory Routing* retrieves relevant episodic observations and core knowledge, invoking replay from Native Memory when the textual evidence is judged insufficient. This separates incremental memory maintenance from query-driven access, allowing accumulated experience to be reused across subsequent interactions. We denote the omnimodal model by Φ and use subscripts to distinguish the task-specific operations described below.

3.1 PROGRESSIVE MEMORY CONSTRUCTION

We incrementally build Episodic Memory from an audiovisual stream V. Using a sliding window, we obtain a temporally ordered sequence of clips $(C_i)_{i=1}^{N}$, where N denotes the number of clips processed so far and grows as new observations arrive. Adjacent clips overlap by a duration δ. This overlap provides shared audiovisual evidence around clip boundaries, mitigating the effects of truncated utterances and sound events while supporting continuity across clips.

To interpret each clip in the context of preceding observations, we maintain a summary state $\mathcal{S}_{i-1}$ and a global entity roster $\mathcal{G}_{i-1}$. The summary state captures the current scene and ongoing events, while the roster maintains persistent entity identifiers and associated information for recognizing entities across clips. Conditioned on this context, the omnimodal model jointly interprets the visual

and auditory content of C_i, producing an episodic record and updated state:

$$(m_i, \mathcal{S}_i, \mathcal{G}_i) = \Phi_{\text{build}}(C_i \mid \mathcal{S}_{i-1}, \mathcal{G}_{i-1}), \tag{1}$$

where m_i is the clip's textual record, and $\mathcal{S}_i$ and $\mathcal{G}_i$ propagate context and entity information.

Each episodic record organizes three types of observations. Visual descriptions capture events and entity actions. Speech records associate each utterance with its temporal span, speaker identity, transcript, and corresponding paralinguistic attributes, such as emotion and tone. Non-speech descriptions explicitly capture environmental sounds and music. Entity references in visual descriptions and speech records are linked to the global roster, connecting what is seen and heard to consistent identities. Transcription, speaker attribution, and the extraction of paralinguistic and acoustic information are performed within the joint audiovisual construction operation.

We accumulate these records as Episodic Memory $\mathcal{M}^{\text{episodic}} = \{m_1, m_2, \ldots, m_N\}$ and retain the source clips as Native Memory $\mathcal{M}^{\text{native}} = \{C_1, C_2, \ldots, C_N\}$. Each episodic record links to its source clip, enabling later retrieval of the original audiovisual evidence for replay. The evolving summary state and entity roster provide context for interpreting subsequent clips.

3.2 Retrospective Consolidation & Auditing

OmniMemory separates persistent entity tracking from revisable identity assignments: person IDs anchor observations across clips, while their associated identities can be corrected as evidence accumulates. Every R clips, we audit history $\mathcal{H}_k = \{m_i\}_{i=1}^{kR}$ and roster $\mathcal{G}_{kR}$ using a bounded dialogue sample and active entity set (Appendix C). The audit re-infers identity assignments through cross-clip self-introductions and address–response cues, distinguishing speakers from addressees to resolve ambiguities in isolated name mentions.

Evidence and entity links preserved in episodic records and the roster allow later information to be traced back to earlier observations. When new evidence supports a revised interpretation, auditing overrides earlier assignments and propagates the corrections to the roster and affected records:

$$(\widetilde{\mathcal{H}}_k, \widetilde{\mathcal{G}}_{kR}) = \Phi_{\text{audit}}(\mathcal{H}_k, \mathcal{G}_{kR}), \tag{2}$$

where tildes denote revised versions. Errors introduced much earlier in the stream remain correctable beyond the latest batch. This complements progressive construction: past context guides the interpretation of new observations, while new evidence revises the interpretation of past observations.

We then consolidate the revised recent batch $\widetilde{\mathcal{B}}_k = \{\widetilde{m}_{(k-1)R+1}, \ldots, \widetilde{m}_{kR}\}$ into reusable knowledge about entity roles, preferences, and relationships:

$$\Delta\mathcal{M}_k^{\text{core}} = \Phi_{\text{consolidate}}(\widetilde{\mathcal{B}}_k, \widetilde{\mathcal{G}}_{kR}). \tag{3}$$

Consolidation combines related observations and merges redundant information into Core Memory as key–value entries, such as `<David/role>-<Engineer>`, incorporating the revised entity associations into the resulting facts. Episodic records remain available for contextual retrieval.

3.3 Query-Aware Memory Routing

Textual memories enable efficient reuse of past observations and consolidated knowledge, but may omit perceptual details needed by a query. Native audiovisual replay provides access to such details at a higher token cost. We therefore first retrieve relevant textual evidence and invoke replay only when the model judges the available information insufficient.

Given a query q, we plan memory access from bounded working sets of entities $\mathcal{G}$ and Core Memory keys $\mathcal{M}_{\text{keys}}^{\text{core}}$, maintained by activity- and recency-weighted eviction:

$$(\mathcal{Q}_q, \mathcal{G}_q, \mathcal{K}_q) = \Phi_{\text{plan}}(q, \mathcal{G}, \mathcal{M}_{\text{keys}}^{\text{core}}), \tag{4}$$

where $\mathcal{Q}_q = \{q_1, \ldots, q_n\}$ contains reformulated or decomposed queries for episodic retrieval, $\mathcal{G}_q$ contains relevant entity information, and $\mathcal{K}_q$ specifies the core keys to access.

For Episodic Memory, we perform dense embedding retrieval and BM25 retrieval (Robertson & Zaragoza, 2009) for each subquery q_j, combining their rankings through reciprocal rank fusion

(RRF) (Cormack et al., 2009). Dense retrieval captures semantic similarity beyond exact wording, while BM25 adds lexical evidence. We merge and deduplicate the results across subqueries, retaining at most Λ episodic records to form $\mathcal{R}_q^{\text{episodic}}$. For Core Memory, we directly retrieve the entries associated with the selected keys $\mathcal{K}_q$ to form $\mathcal{R}_q^{\text{core}}$. We retain episodic retrieval rankings and source clip links for replay.

Conditioned on retrieved memories and entity information, the model answers directly or requests replay, optionally providing a retrieval hint q_{hint} specifying audiovisual content to inspect:

$$\Phi_{\text{answer}}(q, \mathcal{G}_q, \mathcal{R}_q^{\text{episodic}}, \mathcal{R}_q^{\text{core}}) = \begin{cases} a, & \text{if judged sufficient,} \\ (\text{REPLAY}, q_{\text{hint}}), & \text{otherwise,} \end{cases} \tag{5}$$

where a denotes the generated answer and q_{hint} may be empty without additional retrieval guidance.

Upon a replay request, we use q_{hint}, when available, to retrieve candidate clips relevant to the requested audiovisual evidence. We supplement these candidates with source clips selected according to the episodic retrieval ranking. We merge and deduplicate the candidates, selecting up to K distinct source clips from Native Memory to form $\mathcal{C}_q \subseteq \mathcal{M}^{\text{native}}$. The omnimodal model then inspects the selected audiovisual content and generates the answer:

$$a = \Phi_{\text{replay}}(q, \mathcal{C}_q, \mathcal{G}_q, \mathcal{R}_q^{\text{episodic}}, \mathcal{R}_q^{\text{core}}) \tag{6}$$

4 Experiments

4.1 Datasets and Evaluation

We evaluate OmniMemory on M3-Bench-Robot, M3-Bench-Web, and VideoOdyssey-AV. M3-Bench (Long et al., 2026) assesses memory-based reasoning using robot-perspective interaction recordings and diverse web videos, closely matching our target assistant scenarios. VideoOdyssey-AV (He et al., 2026) provides a stress test through ultra-long videos and explicitly assesses non-speech audio through its Sound and Music categories. We follow the official LLM-as-a-judge protocol for M3-Bench (Long et al., 2026) and report exact-match accuracy for VideoOdyssey-AV's multiple-choice questions. Further dataset and evaluation details are provided in Appendix B.

4.2 Baselines

We compare OmniMemory with three groups of baselines. (1) *Base models* directly answer questions from visual or audiovisual inputs; we evaluate Qwen2.5-VL-7B$_{+\text{ASR}}$ (Bai et al., 2025), Qwen2.5-Omni-7B (Xu et al., 2025a), Qwen3.5-Omni-Plus (Team, 2026), and Gemini-3.1-Pro (Google DeepMind, 2026). (2) *Trained agentic and memory methods* incorporate trained components for long-video processing, memory, or agentic reasoning; we include VideoChat-Flash$_{+\text{ASR}}$ (Li et al., 2026), Flash-VStream$_{+\text{ASR}}$ (Zhang et al., 2024b), M3-Agent (Long et al., 2026), and OmniAgent (Xing et al., 2026). (3) *Training-free RAG and memory methods* augment pretrained models through retrieval and memory mechanisms; we consider Video-RAG (Luo et al., 2025), VideoARM (Yin et al., 2026), HippoMM (Lin et al., 2026), and WorldMM (Yeo et al., 2026). The subscript $_{+\text{ASR}}$ denotes variants supplemented with speech transcripts. We include full-video inference with base models as a reference, using the entire video at query time. Our primary comparisons are against agentic, retrieval-augmented, and memory-based methods, reflecting our focus on organizing and reusing an expanding audiovisual history. More details are provided in Appendix C.

4.3 Implementation Details

To reflect realistic omnimodal interaction settings, we use Qwen3.5-Omni-Plus (Team, 2026) as the backbone of OmniMemory and `text-embedding-v4` (Zhang et al., 2025) for dense retrieval. For memory construction, we use 30-second clips with an overlap of $\delta = 5$ seconds and trigger retrospective processing every $R = 10$ clips. We set the episodic retrieval budget to $\Lambda = 20$ records and the number of replayed clips to $K = 3$. For fair comparison, we also adopt Qwen3.5-Omni-Plus as the backbone for the main pipelines of all training-free RAG and memory baselines, while retaining their specialized components, such as OCR models. All methods requiring ASR use transcripts generated by Qwen3.5-Omni-Plus. We rerun all baselines on the three benchmarks under these settings. Further implementation details are provided in Appendix C.

Table 1: **Main results on M3-Bench-Robot and M3-Bench-Web.** ME, MH, CM, PU, and GK denote multi-evidence reasoning, multi-hop reasoning, cross-modal reasoning, person understanding, and general knowledge extraction, respectively. The subscript $_{+\mathrm{ASR}}$ indicates that speech transcripts generated by Qwen3.5-Omni-Plus are provided as additional input.

Method		M3-Bench-Robot						M3-Bench-Web					
		ME	MH	CM	PU	GK	**All**	ME	MH	CM	PU	GK	**All**
Base Models (Full-Video References)													
Qwen2.5-VL-7B$_{+\mathrm{ASR}}$	7B	44.7	43.5	40.3	59.5	30.6	44.7	52.1	43.7	44.6	63.3	57.1	54.5
Qwen2.5-Omni-7B	7B	35.4	29.4	32.4	48.7	21.4	33.8	37.5	22.5	33.8	51.9	30.0	37.2
Qwen3.5-Omni-Plus	–	51.5	41.2	52.3	66.4	33.6	50.9	67.5	67.5	67.1	76.4	65.5	69.2
Gemini-3.1-Pro	–	55.8	50.6	52.7	68.4	36.1	54.1	69.0	66.4	69.5	75.5	71.4	71.3
Trained Agentic & Memory Methods													
VideoChat-Flash$_{+\mathrm{ASR}}$	7B	47.6	45.9	45.4	63.5	28.1	45.8	51.9	44.6	49.6	63.4	55.4	54.6
Flash-VStream$_{+\mathrm{ASR}}$	7B	42.5	40.0	41.0	54.4	26.9	41.0	45.2	34.9	39.1	54.7	47.9	46.7
M3-Agent	7B/32B	31.7	28.2	30.9	43.1	22.3	31.8	42.5	28.0	44.8	59.0	54.7	48.7
OmniAgent	7B	36.8	28.2	35.1	48.4	19.6	34.2	45.0	38.0	46.3	57.5	47.3	48.4
Training-Free RAG & Memory Methods													
Video-RAG	–	54.6	61.2	53.8	71.5	34.3	54.2	63.7	62.7	60.2	75.2	64.7	66.2
VideoARM	–	26.5	25.9	26.9	35.6	17.7	26.6	48.5	32.7	49.6	59.0	45.6	48.0
HippoMM	–	42.3	48.2	40.3	60.6	24.2	41.5	51.4	40.8	51.6	66.0	52.2	54.0
WorldMM	–	45.4	48.2	43.9	61.5	31.5	45.8	54.1	45.7	51.8	63.2	52.0	54.5
OmniMemory	**–**	57.0	54.1	55.0	63.9	49.2	**56.7**	65.3	66.9	68.8	77.5	75.4	**71.5**

Table 2: **Main results on VideoOdyssey-AV.** We report performance across four high-level capability dimensions (Perception, Cognition, Summarization, and Temporal), three audio categories (Speech, Sound, and Music), and the overall score. The subscript $_{+\mathrm{ASR}}$ indicates that speech transcripts generated by Qwen3.5-Omni-Plus are provided as additional input.

Method		Capability				Audio Type			All
		Perception	Cognition	Summ.	Temporal	Speech	Sound	Music	
Base Models (Full-Video References)									
Qwen2.5-VL-7B$_{+\mathrm{ASR}}$	7B	34.2	34.8	44.0	31.1	38.8	27.9	24.8	34.3
Qwen2.5-Omni-7B	7B	30.1	30.0	44.0	41.0	32.7	32.2	29.5	31.6
Qwen3.5-Omni-Plus	–	56.1	49.8	72.0	39.3	58.8	46.5	47.7	53.6
Gemini-3.1-Pro	–	56.4	56.3	72.0	34.4	60.8	49.6	45.6	55.5
Trained Agentic & Memory Methods									
VideoChat-Flash$_{+\mathrm{ASR}}$	7B	42.4	40.8	50.0	39.3	45.6	36.4	39.6	42.2
Flash-VStream$_{+\mathrm{ASR}}$	7B	35.2	33.8	56.0	36.1	39.6	31.4	28.9	35.7
M3-Agent	7B/32B	34.9	31.5	34.0	41.0	36.5	31.8	26.8	34.3
OmniAgent	7B	35.4	32.3	54.0	29.5	36.9	31.0	38.9	34.7
Training-Free RAG & Memory Methods									
Video-RAG	–	54.0	51.0	72.0	42.6	59.0	43.8	44.3	52.4
VideoARM	–	38.3	36.3	58.0	52.5	41.5	37.6	38.3	38.9
HippoMM	–	40.5	37.8	42.0	41.0	42.9	35.7	32.9	39.3
WorldMM	–	46.7	45.5	52.0	54.1	51.5	41.9	37.6	46.7
OmniMemory	–	60.9	54.0	76.0	60.7	64.9	50.4	51.0	**58.9**

4.4 Main Results

Tables 1 and 2 show that OmniMemory achieves the highest overall accuracy among the evaluated agentic, retrieval-augmented, and memory-based methods, with base models included as full-video inference references. OmniMemory reaches 56.7% on M3-Bench-Robot, 71.5% on M3-Bench-Web, and 58.9% on VideoOdyssey-AV, outperforming Video-RAG, the strongest baseline in this group, by 2.4, 5.3, and 6.4 percentage points, respectively. The comparisons with training-free baselines use Qwen3.5-Omni-Plus as the backbone of their main pipelines, keeping backbone choice fixed when evaluating the different frameworks.

The improvements extend to non-speech auditory understanding. On VideoOdyssey-AV, OmniMemory achieves 50.4% on Sound and 51.0% on Music, exceeding the strongest baseline in each category by 6.6 and 6.7 percentage points, respectively. These gains show that the overall advantage

Table 3: **OmniMemory ablations** (ACC, %): memory components (left); input modalities, memory mechanisms, and retrieval strategies (right). P&A: paralinguistic and acoustic fields.

Memory Component			M3-Robot	M3-Web	VO-AV
Episodic	Core	Native			
✓			45.2	65.2	55.6
	✓		16.6	28.0	37.6
		✓	42.4	54.8	49.9
✓	✓		45.5	66.0	57.0
✓		✓	55.2	70.3	56.9
	✓	✓	44.6	57.5	48.9
✓	✓	✓	**56.7**	**71.5**	**58.9**

Variants	M3-Robot	M3-Web	VideoOdyssey-AV			
			Speech	Sound	Music	All
Input Modalities						
Visual only	36.7	50.3	46.3	40.7	40.3	43.1
Audio only	50.4	64.9	58.7	47.7	46.3	53.0
Memory Mechanism						
w/o P&A fields	54.2	68.9	63.7	45.3	43.0	55.2
w/o audit	52.8	70.8	63.9	48.8	47.7	57.3
Retrieval Strategies						
BM25 search	53.8	70.5	63.0	45.3	**55.0**	57.3
embedding search	54.6	**71.6**	60.4	50.0	46.3	55.5
OmniMemory	**56.7**	71.5	**64.9**	**50.4**	51.0	**58.9**

extends to questions involving acoustic information beyond spoken content, an important consideration for memory built from audiovisual experience. Temporal reasoning and general knowledge extraction also benefit. OmniMemory achieves 60.7% on the Temporal dimension of VideoOdyssey-AV, surpassing the strongest baseline on this dimension by 6.6 percentage points. On general knowledge extraction, it reaches 49.2% on M3-Bench-Robot and 75.4% on M3-Bench-Web, improving over the strongest baseline on each benchmark by 14.9 and 10.7 percentage points, respectively.

These gains are consistent with OmniMemory's combination of detailed audiovisual records, cross-clip entity context, and reusable semantic knowledge. By retrieving relevant memories and selectively revisiting source clips, OmniMemory makes both accumulated context and fine-grained perceptual evidence available for reasoning, supporting questions that require auditory detail, temporal integration, or knowledge aggregated across observations.

4.5 Ablation Studies

We examine memory components, input modalities, memory mechanisms, and retrieval strategies in Table 3. **(1) Memory components.** Removing any component from the full system reduces overall accuracy across all three benchmarks, with the largest drops caused by removing Episodic Memory. Compared with Episodic Memory alone, adding Core or Native Memory further improves performance, supporting the value of consolidated knowledge and direct access to source audiovisual evidence. Combining all three achieves the best results. **(2) Input modalities.** We restrict the available input modalities during memory construction and native replay. Full audiovisual input consistently outperforms both visual-only and audio-only variants. These results support jointly preserving visual and auditory evidence in reusable memory for long-horizon interaction, where future questions may depend on different aspects of earlier experiences. **(3) Memory mechanisms.** Compared with the full system, removing paralinguistic and acoustic (P&A) fields reduces Sound and Music accuracy on VideoOdyssey-AV by 5.0 and 8.0 percentage points, respectively, highlighting the importance of preserving auditory information beyond speech transcripts. Disabling auditing reduces overall accuracy on M3-Bench-Robot by 3.9 points. This benchmark includes many questions about people and their interactions, for which consistent entity information and speaker attribution across clips are important. The result supports retrospective auditing for maintaining reliable memories about people over time. **(4) Retrieval strategies.** We compare hybrid retrieval for Episodic Memory with BM25-only and embedding-only search. Compared with either strategy alone, hybrid retrieval provides more consistently strong overall performance across benchmarks, suggesting better generalization through the combination of lexical and semantic matching. **Overall**, these ablations support OmniMemory's integrated design for preserving, refining, and retrieving audiovisual evidence for long-horizon reasoning.



Figure 3: **Efficiency comparison of OmniMemory and baselines on M3-Bench-Robot. (a) Accuracy versus query latency**, measured from query submission to final response completion.[1] **(b) Accuracy versus query token consumption.** In (a,b), the red star marks OmniMemory, and shaded regions indicate lower accuracy and higher latency or token consumption. **(c) Memory construction overhead for memory-based methods**, showing construction time (top) and token consumption (bottom), with input and output tokens distinguished. K denotes thousands of tokens.

4.6 Efficiency Analysis

Figure 3 evaluates OmniMemory's accuracy–efficiency trade-off on M3-Bench-Robot, considering both online query costs and memory construction overhead. Query latency and token consumption are averaged per question, while construction time and token consumption are averaged per video. **(1) Query efficiency.** Figure 3(a,b) shows that OmniMemory achieves the highest accuracy among the compared agentic, RAG, and memory methods while maintaining low query latency and moderate token consumption. Some baselines offer faster responses or consume fewer query tokens, but achieve lower accuracy. OmniMemory controls query costs by answering from retrieved textual memories when sufficient and selectively invoking more token-intensive audiovisual replay when additional perceptual evidence is needed. This routing strategy supports efficient reuse of stored knowledge while retaining access to fine-grained audiovisual details. **(2) Memory construction.** Figure 3(c) shows that OmniMemory builds memory faster than WorldMM and M3-Agent. Although construction consumes more tokens than in WorldMM and HippoMM, OmniMemory achieves substantially higher accuracy with far fewer query tokens and markedly lower query latency than both methods. These results reflect a trade-off between greater token expenditure during memory construction and more accurate, efficient subsequent queries. For potentially unbounded interaction histories and repeated queries, the constructed memories can be reused across interactions, allowing construction costs to be amortized through repeated access.

5 Conclusion

We presented OmniMemory, a training-free memory framework designed for long-horizon, repeated interactions. Through a unified omnimodal model, we jointly capture visual observations, speech with associated paralinguistic cues, environmental sounds, and music. We organize these experiences into complementary Episodic, Core, and Native Memory, coordinated through progressive construction, retrospective consolidation and auditing, and query-aware routing. This design maintains contextual records across clips, consolidates reusable knowledge, refines inconsistent observations, and selectively revisits source audiovisual evidence when textual memories are insufficient. Evaluations on M3-Bench-Robot, M3-Bench-Web, and VideoOdyssey-AV demonstrate the highest overall accuracy among the evaluated methods. Ablations support the complementary roles of the three memory components and the value of auditory information beyond transcripts, while efficiency analysis shows a favorable trade-off among accuracy, query costs, and upfront construction overhead. These findings highlight the value of combining reusable audiovisual memories with selective access to source evidence for repeated reasoning over an expanding interaction history.

[1] Query latency excludes ASR time for VideoChat-Flash (+ASR) and Flash-VStream (+ASR).

AI USE STATEMENT

The authors designed the algorithmic framework of OmniMemory and implemented the method. Generative AI tools were subsequently used to assist in reviewing the existing framework and implementation, identifying and correcting errors, and making refinements. The authors evaluated the AI-assisted changes, decided which changes to incorporate, and reviewed and tested the resulting implementation.

The manuscript was drafted by the authors. Generative AI tools were used to refine the draft's wording, organization, and technical presentation, and to improve grammar, clarity, and readability. The authors reviewed and revised the AI-assisted edits to ensure consistency with their intended meaning, the implemented method, and the reported experimental results.

Generative AI tools were not used to formulate the research premises or hypotheses, which were established by the authors. The authors made all final scientific decisions regarding the method, experiments, and conclusions. We take full responsibility for the final content of this work, including its methods, analyses, results, claims, and AI-assisted materials.

ETHICS STATEMENT

All experiments in this work are conducted on publicly available benchmark datasets: M3-Bench-Robot, M3-Bench-Web, and VideoOdyssey-AV. We use their existing audiovisual recordings and annotations to evaluate memory construction and question answering. We acknowledge the original dataset creators and refer readers to the respective dataset publications for details on data collection and annotation.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide an anonymous code repository linked at the end of the abstract. Section 3 describes the complete algorithmic pipeline, including progressive memory construction, retrospective consolidation and auditing, and query-aware memory routing. Section 4.3 specifies the implementation details and parameter settings used in our experiments.

REFERENCES

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

Guo Chen, Lidong Lu, Yicheng Liu, Liangrui Dong, Lidong Zou, Jixin Lv, Zhenquan Li, Xinyi Mao, Baoqi Pei, Shihao Wang, et al. Towards multimodal lifelong understanding: A dataset and agentic baseline. *arXiv preprint arXiv:2603.05484*, 2026a.

Tao Chen, Kun Zhang, Qiong Wu, Xiao Chen, Chao Chang, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Scaling the long video understanding of multimodal large language models via visual memory mechanism. *arXiv preprint arXiv:2603.29252*, 2026b.

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

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 758–759, 2009.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

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

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

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18959–18969. IEEE, 2025.

Google DeepMind. Gemini 3.1 Pro Model Card, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Haichen He, Jiayi Zhou, Sifeng Shang, Yihan Hu, Yuanhan Zhang, and Kaiyang Zhou. Videoodyssey: A benchmark for ultra-long-context and omni-modal video understanding. *arXiv preprint arXiv:2605.22907*, 2026.

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

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. Videorag: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21278–21298, 2025.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, volume 2026, pp. 109089–109117, 2026.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Yidan Wen, Yanghai Wang, Shihao Li, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. Omnibench: Towards the future of universal omni-language models, 2025. URL https://arxiv.org/abs/2409.15272.

Yueqian Lin, Jingyang Zhang, Qinsi Wang, Hancheng Ye, Yuzhe Fu, Yudong Liu, Hai Helen Li, and Yiran Chen. Hippomm: Hippocampal-inspired multimodal memory for long audiovisual event understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5968–5977, 2026.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In *International Conference on Learning Representations*, volume 2026, pp. 146197–146246, 2026.

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, 2025. URL https://arxiv.org/abs/2411.13093.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009.

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

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

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

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

Zhenghao Xing, Ruiyang Xu, Yuxuan Wang, Jinzheng He, Ziyang Ma, Qize Yang, Yunfei Chu, Jin Xu, Junyang Lin, Chi-Wing Fu, et al. Native active perception as reasoning for omni-modal understanding. *arXiv preprint arXiv:2606.19341*, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents, 2025b. URL https://arxiv.org/abs/2502.12110.

Zhucun Xue, Jiangning Zhang, Xurong Xie, Yuxuan Cai, Yong Liu, Xiangtai Li, and Dacheng Tao. Adavideorag: Omni-contextual adaptive retrieval-augmented efficient long video understanding. *arXiv preprint arXiv:2506.13589*, 2025.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28885–28900. IEEE, 2025.

Sicheng Yang, Yukai Huang, Weitong Cai, Shitong Sun, Fengyi Fang, You He, Yiqiao Xie, Jiankang Deng, Hang Zhang, Jifei Song, et al. Egocentric co-pilot: Web-native smart-glasses agents for assistive egocentric ai. In *Proceedings of the ACM Web Conference 2026*, pp. 8862–8873, 2026.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026.

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

Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Mm-narrator: Narrating long-form videos with multimodal in-context learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13647–13657. IEEE, 2024a.

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

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

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

A Limitations and Future Work

OmniMemory is designed for unbounded audiovisual streams, but maintaining consistent person identities over an expanding history remains challenging, particularly without a dedicated face recognition module. Our global roster and cross-clip auditing largely mitigate identity drift: the

roster maintains persistent identity references across clips, while auditing uses accumulated evidence to correct inconsistent attributions and revise earlier assignments. These mechanisms provide substantial robustness to local identification errors, although ambiguous observations can still lead to identity confusion over prolonged interactions.

Directly incorporating face recognition would not fully resolve this limitation. Back-facing views, partial profiles, and occlusions can make facial evidence unavailable or unreliable, and treating uncertain matches as definitive may introduce incorrect identity merges or splits that propagate through memory. Future work will develop more reliable person alignment that selectively incorporates facial cues alongside appearance, voice, and temporal context. This mechanism will explicitly account for uncertainty and work with the existing roster and auditing processes to defer or revise ambiguous identity assignments.

B DATASETS

B.1 M3-BENCH-ROBOT

Dataset construction and statistics. M3-Bench-Robot (Long et al., 2026) evaluates memory-based reasoning in everyday service-robot scenarios. It contains 100 videos and 1,276 manually annotated, open-ended question–answer pairs, with an average video duration of 2,039.9 seconds (approximately 34.0 minutes). The recordings cover seven indoor scenarios: living rooms, kitchens, bedrooms, studies, offices, meeting rooms, and gyms. These scenarios involve interactions through which an agent can observe people's preferences and relationships, track objects and events, and acquire knowledge about its environment.

The videos were produced from manually designed scripts containing sequences of events and questions linked to earlier reference events. A human actor playing the robot wore a head-mounted camera to capture the simulated robot's first-person audiovisual observations. The collection involved 67 actors across 51 locations. Audio was recorded by the head-mounted device, with supplementary voice recordings collected using individual lapel microphones.

After filming, annotators checked the scripted questions against the recorded content, retained or revised applicable questions, discarded unsuitable ones, and added further questions.

Temporal evaluation protocol. For every M3-Bench-Robot question, we truncate the video at its annotated timestamp and evaluate OmniMemory using the memory state available at that point. Memory construction, auditing, consolidation, and native replay use only preceding observations. All baselines follow the same cutoff for preprocessing, memory construction, and inference. Accordingly, "full-video" base-model inputs on this dataset are sampled only from the permitted prefix. All reported results, including ablations, follow this protocol to prevent future-information leakage.

Question types. M3-Bench-Robot and M3-Bench-Web share five capability categories defined by the original benchmark (Long et al., 2026). A question may receive multiple labels, reflecting the combination of capabilities needed to answer it.

- **Multi-evidence Reasoning (ME).** These questions require collecting and aggregating multiple pieces of information distributed across a video. The emphasis is on combining observations to support a conclusion, such as recalling attributes of several objects or events and comparing them.
- **Multi-hop Reasoning (MH).** These questions require step-by-step reasoning across different video segments. For example, answering a question about what happened after a particular visit requires first identifying that visit and then following subsequent events. This category assesses the ability to connect observations through a sequence of reasoning steps.
- **Cross-modal Reasoning (CM).** These questions require integrating evidence from different modalities, particularly visual content and spoken language. For example, determining which folder should contain confidential documents requires associating a spoken instruction with the color of the folder shown on screen.

- **Person Understanding (PU).** These questions assess reasoning about people's identities, emotions, personalities, abilities, and relationships. Answers may require integrating behavior and interactions across multiple scenes, such as inferring a person's cooking ability from their observed performance.
- **General Knowledge Extraction (GK).** These questions assess whether an agent can derive general knowledge from specific experiences shown in the video. For example, observations of groceries being placed on different refrigerator shelves can support an inference about storage rules. The target knowledge must be grounded in the observed events.

B.2 M3-BENCH-WEB

Dataset construction and statistics. M3-Bench-Web (Long et al., 2026) extends the benchmark to a broader range of audiovisual content collected from YouTube. It contains 920 videos and 3,214 manually annotated, open-ended question–answer pairs. The average video duration is 1,630.7 seconds (approximately 27.2 minutes). The collection spans 46 video types, covering documentaries, exploration, interviews, reviews, street interactions, tutorials, variety shows, and vlogs.

Video collection follows a question-driven process: annotators select videos whose narratives and entity relationships support questions about memory and reasoning. A reference category list guides collection to promote diversity and avoid overrepresenting content that is easy to annotate. The annotator who selects a video also creates its associated question–answer pairs. Annotations undergo staged quality control, including trial tasks, sampled review by the dataset authors, feedback, and revision.

Question types and annotation requirements. M3-Bench-Web uses the same five capability labels: Multi-evidence Reasoning, Multi-hop Reasoning, Cross-modal Reasoning, Person Understanding, and General Knowledge Extraction. These capabilities are evaluated across the broader web-video scenarios. Examples in the original paper include comparing the starting prices of several collectibles, identifying the next shop visited after a specified stop, and inferring a person's cooking ability from multiple scenes. These examples illustrate evidence aggregation, sequential reasoning, and person understanding, respectively.

All questions are posed at the end of the video. The annotation guidelines require specific questions with objective, unambiguous answers supported by video evidence. They emphasize character attributes and video-grounded commonsense reasoning, while discouraging questions answerable solely from external knowledge, direct matching of spoken lines, or an isolated brief moment. The guidelines require the relevant context to span more than 10 seconds and include questions that depend on visual information. As in M3-Bench-Robot, individual questions may carry multiple capability labels.

B.3 VIDEOODYSSEY-AV

Dataset construction and statistics. VideoOdyssey-AV is the audiovisual subset of VideoOdyssey (He et al., 2026), designed to evaluate long-context understanding that requires integrating visual and auditory information. It contains 100 synchronized audiovisual videos collected from YouTube, spanning 11 domains and 54 subcategories. The collection covers diverse content, including documentaries, movies, television programs, cartoons, sports, games, tutorials, daily-life recordings, and egocentric and surveillance footage. Videos range from 60 to 255 minutes in duration, with an average of 109 minutes. The collection criteria require a minimum resolution of 720p, dynamic visual scenes, and rich audio content.

The final subset contains 1,062 human-annotated multiple-choice questions, each with one correct answer and three plausible distractors. Annotators could repeatedly inspect videos and navigate their timelines to construct questions emphasizing long-context dependencies and joint audiovisual evidence. Quality control combined automated filtering with expert review. Specifically, Gemini-2.5-Pro and Qwen3-Omni were used to identify questions answerable from video alone or audio alone, and human experts subsequently checked the remaining questions for modality dependence, clarity, and answer validity. This process removed 79 questions from the initial 1,141 candidates.

Temporal scope and audio coverage. Each question is annotated with a *continuous certificate length* (CCL), defined as the duration of video that a human must continuously watch to establish its answer. This measures the temporal extent of the required evidence rather than the duration of the complete source video. The average CCL for VideoOdyssey-AV is 12.8 minutes, with questions covering evidence durations from seconds to more than an hour. The benchmark also characterizes questions along three audio categories: *Speech*, *Sound*, and *Music*, covering spoken information, non-speech acoustic events, and musical content, respectively. These categories provide an additional view of auditory coverage alongside the question types described below.

Question types. VideoOdyssey-AV covers 18 question types organized under four capability dimensions: *Perception*, *Cognition*, *Summarization*, and *Temporal Grounding*. The following descriptions follow the original task definitions (He et al., 2026). All types, including Captioning and Summarization, are evaluated through multiple-choice questions.

Perception comprises nine question types that assess the recognition and retrieval of observable audiovisual information:

- **Counting (Count).** Counts occurrences of specified entities or activities using audiovisual cues. Questions may require tracking repeated participation by a particular person, combining identity recognition with the accumulation of occurrences over time.
- **Object Recognition (ObRec).** Identifies or classifies entities, including people, from audiovisual context. For example, answering may require identifying the student who approaches a teacher by connecting the visible interaction with spoken information.
- **Action Recognition (AcRec).** Recognizes actions performed by people or other entities. Questions require resolving the relevant actor and occasion before identifying the action, such as what a student does upon first leaving their seat.
- **Visual Attribute Recognition (VAR).** Identifies visual properties of people, objects, or scenes, with auditory information helping specify the target or context. For example, a question may ask about the clothing color of a person referenced during a particular conversation.
- **Optical Character Recognition (OCR).** Recognizes text appearing in the video and connects it to the relevant audiovisual context. This can involve identifying a displayed place name associated with a location discussed in speech.
- **Acoustic Event Recognition (AER).** Identifies and classifies sound events in their visual context. Questions may ask which animal sound occurs, or does not occur, during a sequence delimited by visible events.
- **Acoustic Attribute Recognition (AAR).** Recognizes acoustic properties of sounds or music, such as tempo and loudness. This includes comparing background music across visually distinct scenes.
- **Speech Fact Retrieval (SFR).** Retrieves facts explicitly stated in speech, while using visual context to identify the relevant speaker, event, or utterance. The central requirement is accurate recovery of spoken information and its association with the correct context.
- **Captioning (Cap).** Selects a detailed description of observable audiovisual content, including actions, speech, environmental sounds, and scene dynamics. Correct answers must preserve the relationships among these observations within the relevant interval.

Cognition comprises seven question types that require inference or relational reasoning over audiovisual observations:

- **Causal Reasoning (CaRea).** Infers the causes or consequences of an event from audiovisual evidence. For example, answering may require determining which sound triggers a character's visible reaction.
- **Emotional Reasoning (EmRea).** Infers emotional states, their causes or evolution, or the atmosphere of a scene. Relevant evidence can include musical cues, such as whether background music conveys tension or relaxation during a conversation.

- **Intentional Reasoning (InRea).** Infers the purposes or motivations behind a person's actions. Questions require interpreting behavior together with dialogue and surrounding events to determine what the person intends to achieve.
- **Object Reasoning (ObRea).** Infers which entity satisfies a condition, or reasons about an entity's function, attributes, or relationships. For example, identifying the most active student requires integrating comments and observed behavior across a classroom sequence.
- **Speech Content Reasoning (SCR).** Infers information that is implicit rather than explicitly stated in speech. Questions require interpreting dialogue in its visual and situational context to recover implications or explain interactions.
- **Spatial Reasoning (SpRea).** Reasons about the spatial locations and relative positions of entities using audiovisual context. Answering requires identifying the relevant moment and participants before resolving their spatial relationship.
- **Temporal Ordering (Order).** Arranges multiple events or entity appearances in chronological order. This tests temporal tracking and comparison across observations, such as ordering participants by their last appearance.

Summarization contains **Summarization (Sum)**, which assesses high-level understanding of audiovisual content. Questions ask for the central topic or meaning of a relevant interaction or sequence, requiring abstraction over individual observations. For example, a question may ask for the topic of a visually identified participant's spoken question.

Temporal Grounding contains **Temporal Grounding (TeGro)**, which assesses the localization of specified audiovisual events in time. Questions require selecting the appropriate timestamps or intervals, such as the first and last appearances of a particular person, thereby testing the association between event identity and temporal evidence.

C Baseline and Implementation Details

Base models. We evaluate four base models: Qwen2.5-VL-7B, Qwen2.5-Omni-7B, Qwen3.5-Omni-Plus, and Gemini-3.1-Pro (Bai et al., 2025; Xu et al., 2025a; Team, 2026; Google DeepMind, 2026). All follow a full-video inference setting, where each question is answered directly from inputs sampled across the source video at query time. Repeatedly processing these inputs incurs substantial token costs and does not scale to unbounded video streams; we include these models as references for performance on the finite benchmark videos. Under the 32K context budget used for the 7B models, we uniformly sample 64 frames for Qwen2.5-VL-7B and supplement them with ASR transcripts generated by Qwen3.5-Omni-Plus. For Qwen2.5-Omni-7B, we uniformly select 30 audiovisual windows of 10 seconds each, retaining their audio and sampling 64 frames across the resulting 300-second input. VideoOdyssey-AV requires additional sampling because its videos exceed the one-hour audiovisual input limit of Qwen3.5-Omni-Plus. We therefore uniformly select 360 windows of 10 seconds each, yielding one hour of audiovisual content, and sample 768 frames from these windows. We use the same sampling configuration for Gemini-3.1-Pro.

VideoChat-Flash. We use VideoChat-Flash as a trained long-video understanding baseline (Li et al., 2026). Its Hierarchical video token Compression (HiCo) combines clip-level compression, which aggregates spatiotemporal information and merges similar visual tokens, with video-level compression, which progressively drops visual tokens within the language model. Together with multi-stage short-to-long training, this design reduces the computational cost of processing long video sequences while retaining relevant visual information. In our experiments, we evaluate the 7B model and supplement its video inputs with speech transcripts generated by Qwen3.5-Omni-Plus. These transcripts provide spoken content as textual context for question answering, yielding the VideoChat-Flash$_{+\mathrm{ASR}}$ baseline.

Flash-VStream. We use Flash-VStream as a trained baseline for streaming video understanding (Zhang et al., 2024b). Its Spatial-Temporal-Abstract-Retrieved (STAR) memory combines recent spatial details, clustered temporal context, semantic summaries, and representative frame features to maintain compact visual context. The method separates continuous frame processing from

question answering: a frame handler updates the memory as new frames arrive, while an asynchronous question handler accesses the stored features for response generation. In our experiments, we evaluate the 7B model and provide speech transcripts generated by Qwen3.5-Omni-Plus as additional textual context, yielding the Flash-VStream$_{+\mathrm{ASR}}$ baseline.

M3-Agent. We use M3-Agent as a trained multimodal agent with long-term memory (Long et al., 2026). Its memorization module constructs episodic and semantic memories in an entity-centric multimodal graph, linking observations and knowledge to persistent face and voice identities. Native audiovisual understanding is complemented by separate ASR, speaker diarization, and speaker verification components. A control policy trained through reinforcement learning performs multi-turn reasoning and iterative memory retrieval before generating an answer. In our experiments, we evaluate its 7B memorization model and 32B control model, based on Qwen2.5-Omni and Qwen3, respectively, and use Qwen3.5-Omni-Plus to generate the ASR transcripts.

OmniAgent. We use OmniAgent as a trained baseline for active audiovisual understanding (Xing et al., 2026). It follows an Observation–Thought–Action cycle, selectively requesting frames, audio segments, or audiovisual clips according to the question and accumulated evidence. Observations are distilled into textual memory retained across reasoning turns, while processed media are removed from the active context. The agent undergoes supervised fine-tuning followed by reinforcement learning to coordinate perception, reasoning, and action selection. In our experiments, we evaluate its 7B model based on Qwen2.5-Omni, which directly interprets the requested audiovisual evidence and generates answers through iterative interactions.

Video-RAG. We use Video-RAG as a training-free baseline for retrieval-augmented long-video comprehension (Luo et al., 2025). The method decomposes each question into retrieval requests and gathers relevant auxiliary text from ASR, OCR, and object detection through a single-turn retrieval pipeline. The selected evidence is organized temporally and supplied alongside sampled video frames and the original question for answer generation. In our implementation, we use Qwen3.5-Omni-Plus for query analysis, answer generation, and ASR transcription, while retaining specialized visual components for OCR and object detection. The retrieved auxiliary text complements the visual input within this unified response model.

VideoARM. We use VideoARM as a training-free baseline for agentic long-video understanding (Yin et al., 2026). Its controller follows an observe–think–act–memorize loop, invoking tools for temporal localization, scene inspection, clip analysis, and speech transcription to explore videos from coarse context to fine-grained details. The framework maintains hierarchical memory comprising Sensory Memory for visual observations, Result Memory for tool outputs, and Working Memory for reasoning traces and tool-use intentions. These memories guide subsequent evidence collection and answer generation. In our implementation, we use Qwen3.5-Omni-Plus for the controller, visual analysis modules, and ASR transcription within the VideoARM framework.

HippoMM. We use HippoMM as a training-free baseline for memory-based video understanding (Lin et al., 2026). It segments videos into episodes and consolidates them into semantic summaries associated with multimodal features and temporal information. Its hierarchical retrieval first searches these summaries; when the evidence is insufficient, detailed recall identifies relevant seed segments through feature or semantic search and expands temporal windows to gather additional context. Speech content is incorporated through ASR transcripts. In our implementation, we use Qwen3.5-Omni-Plus for visual description, memory summarization, ASR transcription, and answer generation, while retaining HippoMM's segmentation and hierarchical retrieval mechanisms.

WorldMM. We use WorldMM as a training-free baseline for long-video reasoning with multimodal memory (Yeo et al., 2026). It maintains episodic memory indexing events at multiple temporal scales, semantic memory accumulating high-level knowledge, and visual memory linking segment features to source frames for detailed inspection. An adaptive retrieval agent iteratively selects memory sources and temporal granularities according to the question and accumulated evidence. In our implementation, we use Qwen3.5-Omni-Plus for memory construction, retrieval planning, and answer generation. Speech information is incorporated through ASR transcripts generated by the same model.

Table 4: **Performance Comparison of VL Methods With and Without ASR on M3-Bench-Robot and M3-Bench-Web.** The subscript $_{+\mathrm{ASR}}$ indicates that speech transcripts generated by Qwen3.5-Omni-Plus are provided as additional input.

Method		M3-Bench-Robot						M3-Bench-Web					
		ME	MH	CM	PU	GK	**All**	ME	MH	CM	PU	GK	**All**
VL Methods w/o ASR													
Qwen2.5-VL-7B	7B	27.7	28.2	25.8	35.0	21.4	26.9	34.7	18.3	31.4	43.9	24.0	31.2
VideoChat-Flash	7B	32.8	34.1	30.7	44.0	18.3	31.1	35.4	22.3	34.8	48.7	23.7	33.7
Flash-VStream	7B	27.6	25.9	23.3	35.2	15.9	25.5	29.0	12.8	26.1	40.4	18.3	26.8
VL methods w/ ASR													
Qwen2.5-VL-7B$_{+\mathrm{ASR}}$	7B	44.7	43.5	40.3	59.5	30.6	44.7	52.1	43.7	44.6	63.3	57.1	54.5
VideoChat-Flash$_{+\mathrm{ASR}}$	7B	47.6	45.9	45.4	63.5	28.1	45.8	51.9	44.6	49.6	63.4	55.4	54.6
Flash-VStream$_{+\mathrm{ASR}}$	7B	42.5	40.0	41.0	54.4	26.9	41.0	45.2	34.9	39.1	54.7	47.9	46.7

Table 5: **Performance Comparison of VL Methods With and Without ASR on VideoOdyssey-AV.** The subscript $_{+\mathrm{ASR}}$ indicates that speech transcripts generated by Qwen3.5-Omni-Plus are provided as additional input.

Method		Capability				Audio Type			All
		Perception	Cognition	Summ.	Temporal	Speech	Sound	Music	
VL Methods w/o ASR									
Qwen2.5-VL-7B	7B	27.3	30.3	32.0	39.3	29.6	26.7	28.9	28.7
VideoChat-Flash	7B	30.2	29.8	36.0	39.3	30.6	31.4	31.5	30.7
Flash-VStream	7B	25.1	25.3	32.0	36.1	28.0	22.1	20.1	25.8
VL Methods w/ ASR									
Qwen2.5-VL-7B$_{+\mathrm{ASR}}$	7B	34.2	34.8	44.0	31.1	38.8	27.9	24.8	34.3
VideoChat-Flash$_{+\mathrm{ASR}}$	7B	42.4	40.8	50.0	39.3	45.6	36.4	39.6	42.2
Flash-VStream$_{+\mathrm{ASR}}$	7B	35.2	33.8	56.0	36.1	39.6	31.4	28.9	35.7

OmniMemory working-set budgets. Each audit uses a uniform sample of accumulated dialogue, capped at approximately 500 lines, together with at most 40 active entities. A sliding-window mode can instead restrict the dialogue pool to recent clips. The query planner receives at most 80 Core Memory keys and 40 entities. Entity and key working sets are maintained using a weighted combination of activity and recency, with lower-ranked entries evicted from the active context as new entries arrive. These limits apply to the working context rather than the stored episodic history. Identity assignments already propagated into stored records persist after entities leave the active roster. Audit and planning use bounded contexts as external memory grows.

D Additional Experimental Results and Analysis

D.1 Effect of ASR Transcripts on Vision–Language Methods

We examine the contribution of speech information by comparing Qwen2.5-VL-7B, VideoChat-Flash, and Flash-VStream with and without full-video ASR transcripts generated by Qwen3.5-Omni-Plus. Tables 4 and 5 report the results. Without ASR, all three methods, including both the VL base model and the trained video models, achieve lower overall accuracy than Qwen2.5-Omni-7B on all three benchmarks, as reported in Tables 1 and 2.

Adding ASR improves overall accuracy in all nine method–benchmark comparisons, with gains ranging from 5.6 to 23.3 percentage points. These improvements highlight the value of spoken information alongside visual observations, consistent with the audiovisual focus of these benchmarks. The ASR-augmented variants also surpass Qwen2.5-Omni-7B in overall accuracy. This comparison involves different temporal coverage, however: the VL methods receive transcripts spanning the entire video, whereas Qwen2.5-Omni-7B processes sampled audiovisual windows.

The benefits are not uniform across question categories. On VideoOdyssey-AV, Speech exhibits the largest improvement among the three audio categories for every method, while Qwen2.5-VL-7B shows decreases on Music and Temporal despite its overall gain. Thus, transcript augmentation pro-

Table 6: **Question-type results on VideoOdyssey-AV.** Scores are reported as accuracy (%).

Method	Size	Perception									Cognition									Overall
		Count	ObRec	AcRec	VAR	AER	AAR	OCR	SFR	Cap	CaRea	EmRea	InRea	ObRea	SCR	SpRea	Order	Sum	TeGro	
Base Models (Full-Video References)																				
Qwen2.5-VL-7B$_{+\mathrm{ASR}}$	7B	23.5	34.9	23.3	34.5	25.6	26.2	31.3	46.8	48.9	43.1	35.5	28.8	32.1	40.3	25.0	38.0	44.0	31.1	34.3
Qwen2.5-Omni-7B	7B	19.8	36.5	30.0	39.7	27.9	32.3	33.3	26.6	38.3	32.8	35.5	30.8	32.1	22.6	20.0	38.0	44.0	41.0	31.6
Qwen3.5-Omni-Plus	–	37.0	49.2	48.3	62.1	46.5	50.8	47.9	69.9	83.0	60.3	61.3	57.7	42.9	62.9	30.0	30.0	72.0	39.3	53.6
Gemini-3.1-Pro	–	27.2	54.0	50.0	62.1	45.3	47.7	60.4	72.8	78.7	69.0	62.9	80.8	42.9	66.1	36.7	34.0	72.0	34.4	55.5
Trained Agentic & Memory Methods																				
VideoChat-Flash$_{+\mathrm{ASR}}$	7B	37.0	41.3	35.0	39.7	26.7	36.9	37.5	55.5	59.6	44.8	46.8	46.2	33.9	56.5	26.7	28.0	50.0	39.3	42.2
Flash-VStream$_{+\mathrm{ASR}}$	7B	28.4	27.0	21.7	39.7	23.3	27.7	33.3	52.6	40.4	43.1	33.9	25.0	28.6	45.2	25.0	34.0	56.0	36.1	35.7
M3-Agent	7B/32B	27.2	33.3	23.3	37.9	31.4	35.4	27.1	43.9	42.6	34.5	24.2	36.5	28.6	37.1	33.3	26.0	34.0	41.0	34.3
OmniAgent	7B	24.7	33.3	30.0	34.5	25.6	38.5	37.5	42.2	51.1	36.2	40.3	38.5	26.8	33.9	31.7	16.0	54.0	29.5	34.7
Training-Free RAG & Memory Methods																				
Video-RAG	–	37.0	54.0	43.3	48.3	30.2	49.2	52.1	74.6	80.9	62.1	54.8	65.4	39.3	74.2	31.7	26.0	72.0	42.6	52.4
VideoARM	–	22.2	34.9	48.3	20.7	32.6	30.8	37.5	45.1	76.6	39.7	35.5	36.5	42.9	40.3	25.0	34.0	58.0	52.5	38.9
HippoMM	–	25.9	46.0	35.0	44.8	23.3	30.8	35.4	49.1	78.7	44.8	40.3	38.5	39.3	41.9	30.0	28.0	42.0	41.0	39.3
WorldMM	–	24.7	47.6	53.3	34.5	36.0	35.4	58.3	57.8	72.3	56.9	51.6	53.8	41.1	58.1	28.3	26.0	52.0	54.1	46.7
OmniMemory	–	34.6	61.9	55.0	65.5	46.5	44.6	56.3	82.1	83.0	69.0	58.1	76.9	42.9	72.6	28.3	28.0	76.0	60.7	**58.9**

vides useful evidence but does not consistently improve every aspect of audiovisual understanding. Moreover, all ASR-augmented VL methods remain below OmniMemory in overall accuracy across the three benchmarks, showing that access to full-video speech transcripts alone does not close the performance gap.

D.2 Accuracy by Question Type on VideoOdyssey-AV

Table 6 reports detailed results across the 18 question types of VideoOdyssey-AV. Among the evaluated agentic, RAG, and memory methods, OmniMemory achieves the best or joint-best accuracy in 12 question types. The largest gains over the strongest competing baseline in each question type occur in Visual Attribute Recognition, Intentional Reasoning, and Acoustic Event Recognition, with improvements of 17.2, 11.5, and 10.5 percentage points, respectively. These advantages span both perception and cognition. On Speech Fact Retrieval and Temporal Grounding, OmniMemory also exceeds the strongest competing baselines by 7.5 and 6.6 percentage points, respectively, demonstrating strong performance in recovering spoken facts and locating relevant events. Leading results on Captioning and Summarization further indicate that these advantages extend to both detailed descriptions and higher-level understanding.

The gains are not uniform across question types. Within the same comparison group, OmniMemory remains below the strongest baselines on Acoustic Attribute Recognition, Spatial Reasoning, and Temporal Ordering. Its strong temporal grounding performance coexists with weaker results on ordering multiple events, highlighting distinct challenges in event localization and sequence comparison. These remaining gaps identify fine-grained acoustic understanding and the representation of spatial and sequential relationships as areas for further improvement.

D.3 Query Routing, Accuracy, and Online Costs

Table 7 examines the behavior and online costs of query-aware memory routing. Across all three benchmarks, most queries are answered using retrieved textual memories. Episodic and Core Memory thus provide the primary evidence for answering, while source clips are selectively revisited when the model judges the retrieved descriptions insufficient.

Queries following the replay route incur higher average token consumption and latency across all three benchmarks. On M3-Bench-Robot, text answering consumes 8.0K tokens per query, compared with 32.4K for the replay route, which also has approximately twice the latency. Combining both routes yields an average of 19.4K tokens and 10.7 seconds per query. These measurements highlight the practical role of textual memory: a substantial portion of the query workload follows the less expensive text route, while additional audiovisual evidence remains accessible through replay. The resulting query costs depend on both the cost of each route and how frequently it is selected.

Table 7: **Query routing statistics and online costs on three benchmarks.** Metrics for each route are computed over queries following that route, while overall averages are computed over all queries. Costs cover the complete online query pipeline, including planning and replay when invoked, but exclude memory construction. K denotes thousands of tokens.

Benchmark	Replay rate (%)	Text answering			Native replay			Overall average	
		ACC (%)	Tokens (K)	Latency (s)	ACC (%)	Tokens (K)	Latency (s)	Tokens (K)	Latency (s)
M3-Bench-Robot	46.6	68.3	8.0	7.1	43.5	32.4	14.7	19.4	10.7
M3-Bench-Web	30.6	80.9	11.0	6.2	50.3	38.5	13.5	19.4	8.5
VideoOdyssey-AV	26.2	65.6	15.2	21.0	39.9	40.1	29.5	21.7	23.2

The replay route also exhibits lower accuracy than text answering. This observation should be interpreted in light of the routing policy: replay is requested for queries that the model judges difficult to answer from the available textual evidence. These queries may require perceptual details that are missing or insufficiently represented in the retrieved memories, making them more challenging for the current memory pipeline. The two routes therefore operate on different query subsets, and their accuracy difference does not directly measure the effect of replay on answer quality.

E Memory Visualization

Figure 4 presents a memory example from `meeting_room_01` in M3-Bench-Robot. The episodic panel shows a 30-second window spanning 200–230 s. It records visible actions, non-speech sounds, and utterances linked to speaker identities and paralinguistic attributes. For example, Annie offers to order food, while Matthew mentions a piece of bread. Their utterances are associated with P003 and P001, respectively, alongside emotion and tone labels. Footsteps and chair movement provide additional acoustic context for the interaction.

The core panel presents six selected entries from the 52 active entries accumulated over the full video. These entries organize information about roles, preferences, relationships, habits. They may draw on observations outside the displayed window. Together, the panels illustrate how detailed episodic records preserve the context of individual interactions, while core entries provide a compact key–value representation of information accumulated across the video.

F Prompts

Prompt for Episodic Memory Extraction

```
You are a rigorous, objective **video + audio analyst** operating inside a **sliding-
    window memory pipeline** with an **accumulative scene-environment memory** and **
    canonical entity (incl. name) grounding**. You are given one ~30-second video clip (
    frames **and** an audio track) that is part of a longer video processed as
    overlapping windows. In a **single pass** produce BOTH parts below, output as **one**
     JSON object with exactly two top-level keys `visual` and `audio`:

- **(A) VISUAL**  detached third-person record of people/actions/objects/space, plus
    sliding-window continuity / de-duplication / entity-alignment / accumulative-
    environment / **name-grounding** signals.
- **(B) AUDIO**  verbatim transcription **with timestamps**, then paralinguistics and
    environmental sounds.

Division of labour: spoken *words* belong ONLY to audio (verbatim); visual records *
    visible communicative behavior* but never the words.

================================================================
============== SLIDING-WINDOW INPUT (read this first) ===========
================================================================
This clip is one window of a longer video; windows overlap. Time layout in **absolute
    original-video seconds**:
- **window_range** (full clip span): {{WINDOW_RANGE}}
- **context_range** (leading overlap already covered by PREVIOUS clip  continuity/
    alignment ONLY): {{CONTEXT_RANGE}}
- **target_range** (the genuinely NEW portion  PRIMARY source of new memory): {{
    TARGET_RANGE}}
```

(a) Episodic Memory win_008 · 200–230 s

Entities in view. **P001** / Matthew: navy "wassup" T-shirt, glasses; **P002** / Lucy: dark T-shirt, hair tied back; **P003** / Annie: light-yellow T-shirt, entering.

Visual description. A new person (P003 / Annie) enters the meeting room from the right carrying a white object, walks to the head of the table, and stands there. P001 / Matthew, seated at the left, turns to face her and gestures while speaking. P002 / Lucy, seated at the far end, looks up from her laptop.

Actions. Annie enters → walks to the head of the table; Matthew turns and gestures; Lucy looks up from her laptop.

Non-speech acoustics. Footsteps; chair movement. **Acoustic scene:** quiet meeting room.

Speaker-attributed dialogue and paralinguistics

#	Speaker	Utterance	Emotion / tone
1	**P003** / Annie	"Hi everyone."	happy / gentle
2	**P001** / Matthew	"Hi Annie."	neutral / calm
3	**P003** / Annie	"You are quite early."	neutral / calm
4	**P001** / Matthew	"Yeah."	neutral / calm
5	**P003** / Annie	"Do you have some breakfast?"	neutral / gentle
6	**P003** / Annie	"If you don't have, I can order some food."	neutral / gentle
7	**P001** / Matthew	"Oh, that's great."	happy / calm
8	**P001** / Matthew	"I just have a piece of bread."	neutral / calm

(b) Core Memory 6 of 52 active entries · full video

Key	Value	Type	Conf.
`Annie/role`	team lead / manager	identity	high
`Lucy/role`	new team member	identity	high
`Matthew/prefers/food`	Chinese food (Beijing roast duck, Sichuan)	preference	high
`Matthew/relates_to/Annie`	colleagues, same project team	relation	high
`Matthew/habit/photography`	takes photos outside on weekends	habit	high
`Mary/task_assignment/booking`	book the Italian restaurant	task	high

Entries are selected across the full video and may refer to other windows, people, or events.

Figure 4: **Memory visualization on M3-Bench-Robot** (`meeting_room_01`). **(a)** One 30-second episodic record combines visual observations, actions, non-speech acoustics, and speaker-attributed utterances with emotion and tone. Utterances are shown in their recorded order. **(b)** Six selected core entries illustrate identity, preference, relation, habit and task information accumulated across the video.

```
Clip internal time 0 = start of window_range. **Convert every timestamp to absolute
    original-video seconds by adding the window-start offset.**
Rules:  context_range = continuity only, do NOT re-commit.  target_range = where new
    memory comes from.  if an utterance/action straddles the boundary, record it in full.
      if context_range is `none`, this is the FIRST clip  whole window is new.

================================================================
========== CONTINUITY PRIOR  previous_state (optional) =========
================================================================
Carried over from previous clip(s). MAY contain: `scene_id`/`scene_summary`, `
    scene_env_known` (env elements already stored), `known_entities` (each `person_id`,
    **`name`**, `appearance`, `last_location`, `last_action`), `active_event_id`/`
    active_event_summary`, `last_processed_until`, and `global_context` (a rolling high-
    level summary of the whole video so far  use it ONLY as background to resolve
    references/continuity; do NOT re-commit it as new memory).

previous_state
{{PREVIOUS_STATE}}

Use it ONLY as a prior to align scene/entity/name/event IDs and to know what is already
    stored. **Current frames+audio are the only source of facts**; do NOT copy prior
```

```
     facts into your caption/transcript. If a person already has a `name` in `
     known_entities`, **reuse that name** (don't relabel) unless strong conflicting
     evidence. If `none`, cold start.

============================================================
============== PART A  VISUAL (top-priority rules) ==============
============================================================
**A1. Only what is truly visible; never hallucinate.** No carry-over of unseen objects/
     people. POV described objectively ("visible on screen ..."), not first person.
**A2. Hedge/omit easily-confused specifics** (colors, counts, identity, text, clock)
     unless sure.
**A3. Person-reference:** stable visual descriptors; no bare pronouns. (Canonical IDs/
     names in A7.)
**A4. No dialogue words here; DO record visible communicative behavior** (who speaks,
     faces whom, expression, gestures).
**A5. Density first; no filler; no hollow summaries.**

**A6. Sliding-window delta, de-dup & ACCUMULATIVE environment.**
- already-stored & unchanged  suppress (do not re-commit).
- newly-revealed DURABLE environment not in `scene_env_known`  `visual.scene_env_update` (
     even when `same_scene`; e.g. "whiteboard on left wall", "water dispenser in corner").
      `scene_env_update`  transient action.
- transient new happenings in target_range  `visual.target_range_delta`.
Also set `scene_continuity` (same_scene|new_scene|scene_transition|uncertain); on
     same_scene reuse prior `scene_id` and keep `visual_caption` focused on what's new.
     Set `memory_commit_recommendation` (create_event|extend_event|update_entity_state|
     no_op|uncertain).

**A7. Canonical entity alignment + NAME grounding.** For each person in `key_entities`:
- `person_id`: matched known id (P001); or `null` + `match_status=new_entity` if clearly
     new; or `null` + `match_status=uncertain_candidate` if unsure.
- **NAME grounding**  bind a real name to a person ONLY with explicit evidence; output `
     name` (or null), `name_confidence` (high/medium/low), `name_evidence`:
  - `self_introduction`  someone says "I'm X / my name is X"  that name belongs to the
     SPEAKER (cross-check who is speaking). HIGH.
  - `addressed_and_confirmed`  "are you X?" answered "yes", or repeatedly addressed as X
     and responds  that person. HIGH/MEDIUM.
  - `on_screen`  name tag / slide text. HIGH.
  - `mentioned_uncertain`  a name merely mentioned about someone possibly **absent** (e.g.
      "Mary will help")  do NOT bind to a visible person; set `name=null`.
  - If `known_entities` already gives this person a name, reuse it (`name_evidence="prior
     "`).
  - Never bind a name just because it's the only name heard; require the link to be
     evidenced.
- `attribute_change` (compatible|conflict|not_observed|new_detail)  flag only; don't
     overwrite stored attrs.

VISUAL cover when present appearance; actions in temporal order; person-object/person-
     person interaction; key objects w/ state+location; space/setting.

============================================================
============== PART B  AUDIO (top-priority rules) ==============
============================================================
**B1. Verbatim transcription overrides all**  fillers/hesitations kept; no summarize/
     paraphrase/translate; original language; `[inaudible]` when unsure (never guess); no
     speech  `utterances`=[].
**B2. Paralinguistics from closed sets** (below); `unknown` if undeterminable; judge from
     sound, not words.
**B3. Acoustics open-set**  `acoustic.events` short lowercase phrases ([] if none); `
     acoustic.scene` short phrase (`unknown` if undeterminable). Speech is not an acoustic
      event.
**B4. Transcribe first, label second.**
**B5. Timestamps required**  `start_sec`/`end_sec` absolute video seconds. One `utterances
     ` item per utterance, temporal order.
**B6. Audiovisual speaker alignment**  `speaker_id`=canonical person_id if identifiable
     else unknown/offscreen; keep `speaker_hint` (descriptor/name). Don't force a binding.

Paralinguistic closed sets emotion: neutral|happy|sad|angry|anxious|excited|tired|fearful|
     surprised|frustrated|affectionate|other|unknown  tone: calm|urgent|gentle|harsh|
     playful|serious|impatient|commanding|pleading|comforting|other|unknown

============================================================
==================== OUTPUT (strict, single JSON) ================
============================================================
Output ONLY the JSON below in one ```json fence. Exactly two top-level keys `visual`,`
     audio`. Cold-start defaults: `scene_continuity=new_scene`, `match_status=new_entity`,
      `person_id=null`, `name=null`, full initial environment listed in `scene_env_update
     `.
```

```
```json
{
  "visual": {
    "visual_caption": "<third-person; on same_scene focus on what's NEW in target_range;
     no dialogue words; flag uncertainties inline>",
    "scene_id": "<reused prior scene_id if same_scene; else null>",
    "scene_continuity": "same_scene | new_scene | scene_transition | uncertain",
    "scene_env_update": ["<durable env newly revealed, not in scene_env_known; on first
     clip list full initial env; [] if none>"],
    "key_entities": [
      {
        "ref": "<stable visual descriptor>",
        "person_id": "<P001 if matched; else null>",
        "match_status": "matched | new_entity | uncertain_candidate",
        "name": "<grounded real name, or null>",
        "name_confidence": "high | medium | low",
        "name_evidence": "self_introduction | addressed_and_confirmed | on_screen | prior
     | mentioned_uncertain | none",
        "type": "person | object | animal | other",
        "attributes": "<salient attrs, only if sure>",
        "attribute_change": "compatible | conflict | not_observed | new_detail",
        "state": "<current state/doing; '' if none>",
        "location": "<position; '' if undeterminable>"
      }
    ],
    "actions": ["<key actions temporal order; no dialogue words>"],
    "target_range_delta": ["<NEW transient happenings in target_range>"],
    "memory_commit_recommendation": "create_event | extend_event | update_entity_state |
     no_op | uncertain"
  },
  "audio": {
    "utterances": [
      {
        "speaker_id": "<person_id if identifiable; else unknown | offscreen>",
        "speaker_hint": "<descriptor / name / unknown speaker>",
        "start_sec": "<abs seconds, number>",
        "end_sec": "<abs seconds, number>",
        "text": "<verbatim, original language, [inaudible] where unsure>",
        "paralinguistic": {"emotion": "<...>", "tone": "<...>"}
      }
    ],
    "acoustic": {"events": ["<non-speech phrases; [] if none>"], "scene": "<env phrase;
     unknown if undeterminable>"}
  }
}
```

Edge cases First clip: new_scene, full env in scene_env_update, persons new_entity/
     person_id null/name null unless self-intro present. No people: no person item. No
     speech: utterances []. Blurred/black: say so in visual_caption.

Example (illustrative; non-first clip; window=[40,70], context=[40,45], target=[45,70];
     prior has S001, scene_env_known=["circuit-pattern wall","conference table"], P001(man
     ,navy t-shirt,name=null), P002(woman,white blouse,name=null). In target a self-
     introduction is heard.
```json
{
  "visual": {
    "visual_caption": "In target_range, P001 turns toward P002 and speaks while gesturing;
      P002 nods. No change to room layout.",
    "scene_id": "S001",
    "scene_continuity": "same_scene",
    "scene_env_update": [],
    "key_entities": [
      {"ref": "the man in a navy t-shirt", "person_id": "P001", "match_status": "matched",
      "name": "Matthew", "name_confidence": "high", "name_evidence": "self_introduction",
     "type": "person", "attributes": "navy t-shirt,short dark hair", "attribute_change": "
     compatible", "state": "speaking toward P002", "location": "head of table"},
      {"ref": "the woman in a white blouse", "person_id": "P002", "match_status": "matched
     ", "name": "Lucy", "name_confidence": "medium", "name_evidence": "
     addressed_and_confirmed", "type": "person", "attributes": "white blouse", "
     attribute_change": "compatible", "state": "nodding", "location": "near side of table
     "}
    ],
    "actions": ["P001 turns to P002 and speaks", "P002 nods"],
    "target_range_delta": ["P001 introduces himself to P002"],
    "memory_commit_recommendation": "extend_event"
  },
  "audio": {
```
```

```
    "utterances": [
      {"speaker_id": "P002", "speaker_hint": "Lucy", "start_sec": 45.2, "end_sec": 47.0, "
     text": "are you the new member, Matthew?", "paralinguistic": {"emotion": "neutral", "
     tone": "calm"}},
      {"speaker_id": "P001", "speaker_hint": "Matthew", "start_sec": 47.5, "end_sec":
     49.6, "text": "hi, I'm Matthew.", "paralinguistic": {"emotion": "neutral", "tone": "
     gentle"}}
    ],
    "acoustic": {"events": [], "scene": "quiet meeting room"}
  }
}
```

Build environment accumulatively, suppress already-stored, surface newly-revealed env, and **ground names only with explicit evidence**. Commit new memory only from target_range. Output only the single JSON object above.

Prompt for Core Memory Consolidation

You are a rigorous **multimodal knowledge inducer** for an entity-centric memory system. You are given (a) an ENTITY TABLE (canonical people/objects with ids + grounded names), (b) the EXISTING semantic memory accumulated so far (may be empty), and (c) a NEW BATCH of first-stage episodic memories (visual descriptions + verbatim audio with speaker & paralinguistic tags). Your task: **distill stable, reusable knowledge / preferences / relationships, and task-level aggregated facts**, as triples with a compact retrieval KEY, reconciling against existing memory. Output strict JSON.

================ TOP-PRIORITY RULES ================

1. Induce stable knowledge & aggregates, do NOT retell one-off events.
- Target: information that **still holds across time** intrinsic attributes, role/identity, relationships, habits, **preferences/likes/dislikes**, and **task-level aggregates/decisions** (budgets, who-does-what, plans, scheduled things, required resources).
- A single transient action is NOT a triple ("opened a folder at 12:01" = episodic). A repeated/declared tendency IS ("prefers Chinese food", said he loves photography).

2. Stable-from-one vs needs-repetition (avoid over-generalization).
- Stable from a single clear statement (role, a stated preference, a self-introduced name, a decided budget number) may output with `confidence=medium` (or `high` if explicit & unambiguous, e.g. "I'm Matthew", "prize: 3000 dollars").
- A *habit* claim ("usually/always") needs 2 evidences; with one instance, downgrade to a plain attribute/preference or omit.

3. Evidence required; no hallucination; no mind-reading.
- Every triple MUST be supported by the batch (or by updating an existing one); list supporting episodic ids (clip indices / timestamps) in `evidence`.
- No guessing motives. When uncertain, omit or mark `confidence=low`.

4. Entity-centric subjects (use the ENTITY TABLE).
- For people, set `subject` to the grounded **name** if known (else the descriptor) and `subject_id` to the canonical **person_id** (P00x). Keep consistent with the entity table.
- For task-level facts, use a stable event subject `subject="event:<short_name>"` (e.g. `event:promotion_campaign`), `subject_id` = same string.

5. predicate ONLY from the ontology; type ONLY from the enum.

6. KEY convention `subject/short-category[/disambiguator]`, e.g. `Matthew/prefers/food`, `Annie/role`, `event:promotion_campaign/budget/prize`. Short, unique, readable; merge same subject+category under one key.

7. Reconcile with EXISTING semantic memory (CRUD signals).
- If this batch SUPPORTS an existing triple (same key/meaning) `op="update"`, reuse its `key`, add new evidence (the driver will merge & may raise confidence).
- If NEW `op="create"`.
- If this batch CONTRADICTS an existing triple (same subject+predicate, different object) still output your best current triple, set `op="update"`, and set `conflicts_with` = the existing key, with a one-line `conflict_note`. (The driver resolves by confidence > evidence-count > recency; loser is soft-superseded.)
- Do not duplicate an unchanged existing triple that this batch says nothing new about omit it.

==
predicate ontology (English key = meaning)
- attributes: `has_appearance` | `has_attribute` | `has_color`
- location: `usually_located_at` | `belongs_to_place` | `layout_is`

```
- possession: `owns` | `belongs_to`
- habit: `habitually_does` | `usually_at_time`
- preference: `prefers` | `dislikes` | `allergic_to`
- identity/ability: `role_is` | `can_do`
- relation: `relates_to` | `family_of` | `friend_of` | `colleague_of` | `manages` | `
    reports_to`
- task/event: `requires` | `budget_item` | `decided` | `plans_to` | `scheduled_at`
- fallback: `other` (clarify in statement)

type enum `attribute` | `possession` | `location` | `habit` | `preference` | `identity` |
    `relation` | `task` | `decision` | `budget` | `other`

ENTITY TABLE (canonical people/objects: id, name, appearance)
{{ENTITY_TABLE}}

EXISTING semantic memory (current triples; may be empty)
{{EXISTING_SEMANTIC}}

NEW BATCH of first-stage episodic memories (id = clip idx; audio lines are '[t][person_id/
    name][emotion/tone if salient] text')
{{MEMORY_BATCH}}

Output format: strict JSON in one ```json fence, only this object
```json
{
  "triples": [
    {
      "key": "<compact stable retrieval key>",
      "subject": "<grounded name or stable descriptor; or event:...>",
      "subject_id": "<P00x / event:... ; null if unknown>",
      "predicate": "<an English key from the ontology>",
      "object": "<object/value>",
      "statement": "<one natural-language sentence for retrieval & reading>",
      "type": "<one of the type enum>",
      "confidence": "low | medium | high",
      "evidence": ["<supporting clip idx / timestamps from the batch (and existing
    evidence if updating)>"],
      "op": "create | update",
      "conflicts_with": "<existing key it contradicts, or null>",
      "conflict_note": "<one line if conflicts_with set, else empty>"
    }
  ]
}
```

Edge cases If the batch yields no distillable stable knowledge: return `{"triples": []}`.
    Merge multiple batch evidences for the same knowledge into ONE triple. Prefer the
    grounded name (e.g. "Matthew") over a descriptor whenever the entity table provides
    it.

Mini example (illustrative) Batch shows P001/Matthew saying he loves photography and (
    earlier) wanting Chinese food; Annie (P003) handed out materials & set the prize at
    3000 dollars.
```json
{
  "triples": [
    {"key": "Matthew/prefers/hobby", "subject": "Matthew", "subject_id": "P001", "
    predicate": "prefers", "object": "photography", "statement": "Matthew likes
    photography (taking photos of people, scenery, animals).", "type": "preference", "
    confidence": "high", "evidence": ["54","55"], "op": "create", "conflicts_with": null,
     "conflict_note": ""},
    {"key": "Matthew/prefers/food", "subject": "Matthew", "subject_id": "P001", "predicate
    ": "prefers", "object": "Chinese food", "statement": "Matthew prefers Chinese food.",
     "type": "preference", "confidence": "medium", "evidence": ["48"], "op": "create", "
    conflicts_with": null, "conflict_note": ""},
    {"key": "Annie/role", "subject": "Annie", "subject_id": "P003", "predicate": "role_is
    ", "object": "team leader/manager", "statement": "Annie is the leader: hands out
    meeting materials and sets the agenda.", "type": "identity", "confidence": "high", "
    evidence": ["8","12"], "op": "create", "conflicts_with": null, "conflict_note": ""},
    {"key": "event:promotion_campaign/budget/prize", "subject": "event:promotion_campaign
    ", "subject_id": "event:promotion_campaign", "predicate": "budget_item", "object": "
    prize money: 3000 USD", "statement": "The campaign's prize money budget is 3000 USD
    .", "type": "budget", "confidence": "high", "evidence": ["37"], "op": "create", "
    conflicts_with": null, "conflict_note": ""}
  ]
}
```
```

```
Induce only stable knowledge & task aggregates, ground subjects to entity ids/names,
     support every triple with evidence, reconcile with existing memory. Output only the
     JSON object above.
```

Before instantiating the audit prompt below, we apply the dialogue sampling and entity limits in Appendix C. The template's "full cumulative conversation" therefore refers to the supplied sample, rather than every stored utterance.

Prompt for Attribution Auditing

```
You are resolving the character ROSTER of a long video's memory. A prior pass processed
     the video clip-by-clip and detected people as ANONYMOUS entities (P001, P002, )
     without binding names. Your job: determine each entity's real name from the
     cumulative dialogue transcript.

You are given:
1. PERSONS  canonical people, each an id (P001, P002, ) + visual appearance description.
     NO names are shown  you derive them purely from dialogue evidence.
2. DIALOGUE  the full cumulative conversation in time order. Each line is labeled ONLY by
     the speaker's person id (P001, P002, ), NEVER by name. "offscreen" and "unknown" are
     non-person speakers (usually the robot assistant).

Decide each person's name using ONLY in-dialogue evidence, in this PRIORITY order:
(a) SELF-INTRODUCTION (strongest): "I'm X" / "I am X" / "my name is X"  that speaker is X.
(b) VOCATIVE + RESPONSE: someone says "X, <request>" and the next responder is X. NOTE: a
     bare GREETING ("Hi X") is the WEAKEST vocative  trust only when nothing stronger
     contradicts.
(c) CHARACTERIZATION / ROLE DESCRIPTION: "X leads the meeting"  map to the person whose
     role matches.

CRITICAL RULES:
- **A name said BY a person to ADDRESS someone else is the LISTENER's name, never the
     speaker's.** Example: P001 says "Hello Emma"  "Emma" is NOT P001's name; it likely
     refers to whoever P001 is talking to (check who responds next).
- **"Robot" when addressing the assistant is NOT a person name.** The robot/assistant
     often responds as "offscreen". If someone says "Hello Robot" or "Thank you Robet" and
      "offscreen" answers, this is the robot  do NOT assign "Robot" to any human person.
- **COUNT THE PEOPLE.** Never assign the same name to two different person_ids.
- If TWO different names point to the SAME person, one is a mis-transcription. Keep the
     one with STRONGER evidence (more occurrences, clearer context).
- Do NOT invent names. If no clear evidence exists  name = null.
- This is an INCREMENTAL process  you may be called multiple times as the video progresses
     .
Base your decision on ALL available evidence. If you previously assigned a name but new
     evidence contradicts it, correct it.

PERSONS:
{{ENTITY_TABLE}}

DIALOGUE (labeled by id only):
{{DIALOGUE}}

Output STRICT JSON only, in a ```json fence:
```json
{
  "roster": [
    {
      "person_id": "P001",
      "name": "Matthew",
      "confidence": "high",
      "evidence": "self-intro: 'Hi, I am Matthew' at 43s",
      "candidates": [
        {"name": "Matthew", "score": 0.9, "reason": "self-introduction"},
        {"name": "Matt", "score": 0.1, "reason": "possible nickname variant"}
      ]
    },
    {
      "person_id": "P002",
      "name": null,
      "confidence": "low",
      "evidence": "no reliable naming evidence yet",
      "candidates": []
    }
  ]
}
```
```

```
```
Rules for candidates:
- scores should sum to ~1.0 for each person (approximate is fine)
- include ALL plausible name candidates, even if score < 0.2
- "name" field = the top candidate (highest score), or null if no candidate scores >= 0.3
- confidence: "high" if top candidate score >= 0.7, "medium" if >= 0.4, "low" otherwise

Return a roster entry for EVERY person id listed in PERSONS. Output EXACTLY ONE json
    object.
Keep each "evidence" to ONE short sentence. One name maps to at most one person id.
```

**Prompt for Retrieval Plan Generation**

```
You are the retrieval PLANNER for a video-memory QA system. Given the QUESTION
and the RESIDENT memory (known entity keys + available semantic KEYS + a SCENE key),
    decide what to
retrieve BEFORE answering. Do NOT answer the question. Output STRICT JSON in a ```json
    fence:
```json
{
  "reasoning": "<one line: what info the question needs>",
  "need_entities": ["<entity id like P001 that is relevant>"],
  "need_keys": ["<pick relevant items from the listed SEMANTIC KEYS; ALSO include \"
     SCENE_ENV\" if the question is about the environment / layout / where objects or
     people are located>"],
  "episode_queries": ["<free-text queries to semantically search video moments for details
     >"],
  "time_ranges": [[start_sec, end_sec]],
  "sufficient": true
}
```
Rules: need_keys MUST be chosen from the listed keys only (SEMANTIC KEYS, plus "SCENE_ENV"
     when the
question concerns environment/layout/locations). Use episode_queries for details not
    covered by keys. Set time_ranges only if the question is about a time span. Set
    sufficient=false if the text memory clearly cannot answer and the raw clip must be re
    -watched; put WHAT to look for (the object/detail) into episode_queries the system
    uses them to pick the most relevant clip. Output only the JSON object.

RESIDENT:
{{RESIDENT}}

QUESTION: {{QUERY}}
```

**Prompt for Answer Generation**

```
You are answering the user's question about a video, using the memory below (RESIDENT is
    always true; RETRIEVED is what search pulled for THIS question). The memory includes
    per-utterance speech / emotion / tone and third-person behavior captions.

Decide honestly which of the two to do:
- If the RETRIEVED evidence clearly CONTAINS what the question needs, answer directly from
     it. Answer in the SAME language as the question; concise, direct, natural; prefer
    grounded names; evidence lines carry absolute timestamps like [12.0-14.0s] use them
    for time questions.
- If you are NOT confident the evidence actually supports the answer e.g. it does not
    state the specific fact, or the question needs something text captions cannot
    reliably carry: exact COUNTS of actions/objects, precise SPATIAL locations (where
    something/someone is), WHICH specific objects were involved, fine on-screen text /
    color / appearance, or aggregating across moments you were not given then DO NOT
    GUESS. Output EXACTLY one line and nothing else: NEED_REPLAY | <the specific thing to
     look for> Rule of thumb: prefer NEED_REPLAY over a low-confidence guess. But if the
    evidence genuinely answers the question, just answer it do NOT over-trigger replay (
    the memory exists to be used).

{{EVIDENCE}}

QUESTION: {{QUERY}}
```
```

Prompt for Answer via Replay

```
You are answering the user's question about a video. The stored text
memory was insufficient, so you are now RE-WATCHING the original video clip(s) provided
     above. Use what you SEE and HEAR in the clip(s) (plus the text memory below as
     context) to answer. If multiple clips are given, they are in chronological order
     combine evidence across them (e.g. for counting or "which objects"). Answer in the
     SAME language as the question; be concise and direct.

{{EVIDENCE}}

QUESTION: {{QUERY}}
```