WorldMM: Dynamic Multimodal Memory Agent for Long Video Reasoning

Supplementary Material

In this supplementary material, we provide additional details on the dataset (Sec. A), additional implementation details (Sec. B) and descriptions on experiments (Sec. C). We also present detailed and additional experimental results (Secs. D and E), qualitative analyses (Sec. F), and a discussion of limitations and broader impacts (Sec. G).

A. Additional Details on Dataset

In this section, we provide additional details for each dataset used in our experiments. Tab. 5 summarizes the datasets, including the number of queries, domain categories, and the average video duration.

Table 5. Summary of benchmark datasets used in experiments.

Dataset	# Queries	Domain	Avg. Video Length
EgoLifeQA [41]	500	Egocentric	44.3h
Ego-R1 Bench [30]	300	Egocentric	44.3h
HippoVlog [19]	1,000	Vlog	0.45h
LVBench [32]	1,534	General	1.14h
Video-MME (L) [7]	900	General	0.69h

A.1. EgoLifeQA

EgoLifeQA [41] is a set of questions designed to test the capability of models to understand and remember everyday life from week-long video recordings. It includes questions that require recalling past events, tracking object locations, and reasoning over long-term activities. In our experiments, we use questions from the perspective of a single participant (A1: JAKE), along with his corresponding video stream, which spans 44.3 hours. The benchmark is organized into five distinct categories as follows.

EntityLog (Ent.) Questions that require recalling information about objects, such as their locations, states, or interactions. (Example: "Who used the screwdriver first?")

EventRecall (EvR.) Questions that ask about specific past events, including what happened, when it occurred, and relevant context. (Example: "Shure mentioned Tiramisu, when was the last time we discussed making Tiramisu?")

HabitInsight (Hab.) Questions aimed at identifying a person's recurring behaviors or long-term activity patterns. (Example: "What food does Alice love to eat?")

RelationMap (Rel.) Questions involving understanding social relationships and interactions between people. (Example: "Who usually sings when Shure plays the guitar?")

TaskMaster (Task) Questions focused on ongoing or pending tasks that require reasoning about what actions still need to be completed. (Example: "What are we planning to do in the afternoon?")

A.2. Ego-R1 Bench

Ego-R1 Bench [30] is designed as a complementary evaluation to EgoLifeQA, but with a distinct focus on model reasoning. While both benchmarks focus on the same week-long egocentric video, Ego-R1 Bench targets multi-step, tool-augmented reasoning over ultra-long video. We reorganize query types of Ego-R1 Bench to the category adopted by EgoLifeQA, as shown in Tab. 6.

Table 6. Classification of queries under the EgoLifeQA category.

Category	Ego-R1 Category
EntityLog	EntityLog, FoodLog, HealthLog, TechLog
EventRecall	EventRecall, Event Recollection, Event Memory
HabitInsight	HabitInsight, Behavior Habit(s)
RelationMap	RelationMap, Interpersonal Relationships
TaskMaster	TaskMaster, Future Plan(s)

A.3. HippoVlog

HippoVlog [19] contains 25 daily vlog videos with 1,000 multiple-choice questions for continuous audiovisual event understanding. The benchmark evaluates a model's ability to handle modality-specific information, with **Auditory (Aud.)** questions requiring reasoning over the audio stream (or transcript) and **Visual (Vis.)** questions focusing on the visual content. **Auditory+Visual (A+V)** queries test the model's ability to integrate information across both modalities, while **Summarization (Summ.)** questions assess higher-level reasoning over long temporal spans, requiring synthesis of events and semantic understanding from the continuous video.

A.4. LVBench

LVBench [32] consists of 103 long videos, typically longer than an hour, with 1,549 multiple-choice questions for extreme long video understanding. The videos cover a general and diverse set of domains. Questions include both visual perception for recognizing entities or events in short segments and summarization for higher-level reasoning across

extended sequences, evaluating models' ability to integrate information over both local and long-horizon contexts. In our experiments, we categorize questions into three groups based on their segment length, defined as the duration of video required to answer the question: **Short** (<30s), **Medium (Med.)** (30s∼5min), and **Long** (>5min). We excluded 15 questions without segment tags, leaving 1,534 questions in total for evaluation.

A.5. Video-MME

Video-MME [7] is a comprehensive video understanding benchmark with 2,700 questions and varying video durations. In this experiment, we use only the long subset (>30min), containing 900 questions, to assess the model's capability on long video reasoning. We adopt the categories provided by the benchmark, with acronyms as follows: Action Reasoning (ARES), Action Recognition (AREC), Attribute Perception (ATTR), Counting Problem (CNT), Information Synopsis (ISYN), OCR Problems (OCR), Object Reasoning (ORES), Object Recognition (OREC), Spatial Perception (SPER), Spatial Reasoning (SRES), Temporal Perception (TPER), and Temporal Reasoning (TRES).

B. Additional Implementation Details

We provide additional details on the baseline setup (Sec. B.1), the configuration of our proposed WorldMM (Sec. B.2), and the prompts used (Sec. B.3).

B.1. Baseline Setup

Base Models & Long Video LLMs For all base models and long video LLMs, the video input is uniformly sampled at 0.5 fps and capped at 768 frames due to context length limitations, as described in Sec. 1. In this setting, the models operate solely on visual frames without access to video captions or speech transcripts.

RAG-based Video LLMs For text-based RAG video models, we construct a knowledge base from video captions. Specifically, each video is segmented into 30 second chunks, and set of captions from these segments serve as retrieval pool. LightRAG [10] performs dual-level retrieval, selecting either fine-grained (low-level) or abstracted (high-level) information from the knowledge graph generated from set of captions depending on the query. HippoRAG [11], in contrast, retrieves raw captions ranked by their PPR scores, treating each caption as a separate document. For Video-RAG [21] model, retrieval is performed directly on the raw video using tools such as optical character recognition (OCR) and automatic speech recognition (ASR) to extract textual signals. Unless otherwise stated, we follow the retrieval specifications described in each model's corresponding paper or implementation.

Memory-based Video LLMs Memory-based video LLMs construct explicit memories from the video stream. For EgoRAG [41] and Ego-R1 [30], which build hierarchical textual memories, we use the same temporal granularity applied when constructing WorldMM's memory. For models that perform iterative reasoning, including Ego-R1 [30] and M3-Agent [20], we evaluate the checkpoints released by authors and set the maximum number of reasoning iterations to 5 to ensure consistent evaluation across all systems. All other implementation details follow the official specifications provided by the respective authors.

B.2. WorldMM

To construct multi-scale episodic memory, video captioning is performed at each temporal unit by passing sampled video frames along with transcripts generated using Distil-Whisper large-v3.5 [8]. Moreover, we tailor the temporal resolutions to each dataset's duration. For EgoLifeQA and Ego-R1 Bench, which contain week-long videos, we use four broad timescales: 30 seconds, 3 minutes, 10 minutes, and 1 hour. For HippoVlog, LVBench, and Video-MME, which contain shorter recordings averaging about an hour, we adopt shorter timescales of 10 seconds, 30 seconds, 3 minutes, and 10 minutes to better match their temporal structure. For semantic memory, triplets with a similarity score above 0.6 are consolidated using an LLM, and the top 10 triplets are retrieved at query time. The retrieval agent is limited to a maximum of five iterations, consistent with the baseline evaluation setting.

B.3. Prompts

To construct and retrieve memory, and to generate the final response of WorldMM, we employ carefully optimized prompts for use with an LLM. In particular, we use prompts for video captioning (Fig. 9), episodic triple extraction (Figs. 10 and 11) and multi-scale memory construction (Fig. 12), adapted from Yang et al. [41]. Furthermore, we utilize prompts for multi-scale memory retrieval (Fig. 13), semantic triple extraction (Fig. 14), semantic consolidation (Fig. 15), iterative reasoning by the retrieval agent (Fig. 16), and final response generation (Fig. 17).

C. Additional Description on Experiments

In this section, we provide additional description of the settings used in our ablation experiments.

C.1. Dynamic Temporal Scope Retrieval (Sec. 4.4)

To evaluate performance on dynamic temporal reasoning with WorldMM, we employ several approaches, including temporal grounding model, embedding-based retrieval models, hierarchical retrieval models, and keyframe selection method. For each method, we measure tIoU using ei-

Table 7. Category-wise performance breakdown of WorldMM and baselines on EgoLifeQA, Ego-R1 Bench, HippoVlog, and LVBench.

Model	EgoLifeQA						Ego-R1 Bench						HippoVlog					LVBench			
	Ent.	EvR.	Hab.	Rel.	Task	Avg.	Ent.	EvR.	Hab.	Rel.	Task	Avg.	Aud.	Vis.	A+V	Summ.	Avg.	Short	Med.	Long	Avg.
Base Models																					
Qwen3-VL-8B [1]	35.2	30.2	39.3	46.4	46.0	38.6	31.8	41.5	38.5	42.1	44.7	35.7	73.6	74.0	69.2	80.8	74.4	48.8	44.4	53.4	48.3
Gemini 2.5 Pro [4]	43.2	40.5	41.0	55.2	52.4	46.4	43.9	56.1	53.9	47.4	47.4	46.7	69.2	75.2	63.6	80.0	72.0	57.1	52.2	65.2	57.0
GPT-5 [23]	47.2	42.1	47.5	53.6	55.6	48.6	41.8	58.5	53.9	52.6	50.0	46.3	73.6	75.6	69.2	**84.4**	75.7	**59.1**	59.1	69.1	60.4
Long Video LLMs																					
VideoChat-Flash [18]	28.8	32.5	37.7	37.6	38.1	34.2	43.4	43.9	38.5	31.6	44.7	42.7	60.8	59.2	56.4	55.6	58.0	34.9	23.1	44.6	33.2
Time-R1 [35]	39.2	50.8	65.6	48.8	47.6	48.8	49.2	48.8	46.2	42.1	44.7	48.0	58.2	58.2	49.4	52.4	54.6	32.1	23.6	40.2	31.1
Video-RTS [36]	40.8	48.4	62.3	48.8	47.6	48.2	47.6	46.3	53.9	52.6	47.4	48.0	58.8	62.0	56.8	58.4	59.0	43.4	25.7	49.5	39.8
RAG-based Video LLMs																					
LightRAG [10]	40.8	48.4	67.2	50.4	44.4	48.8	54.0	61.0	46.2	42.1	42.1	52.3	51.6	46.0	44.8	47.2	47.4	30.2	28.6	34.3	30.4
HippoRAG [11]	48.8	60.3	70.5	60.8	66.7	59.6	54.5	65.9	69.2	52.6	50.0	56.0	72.4	53.2	54.0	73.2	63.2	54.9	47.5	62.3	54.0
Video-RAG [21]	49.6	56.3	67.2	55.2	54.0	55.4	48.7	58.5	53.9	47.4	44.7	49.7	63.2	64.8	63.6	68.8	65.1	32.9	30.2	39.7	33.1
Memory-based Video LLMs																					
EgoRAG [41]	40.0	56.3	62.3	54.4	52.4	52.0	46.6	56.1	46.2	47.4	55.3	49.0	64.8	53.2	47.6	64.4	57.5	32.4	32.0	31.9	32.2
Ego-R1 [30]	51.2	53.2	63.9	50.4	50.8	53.0	50.8	63.4	38.5	36.8	57.9	52.0	57.2	58.8	52.0	67.2	58.8	32.5	36.5	37.3	34.1
HippoMM [19]	45.6	53.2	70.5	55.2	58.7	54.6	51.9	56.1	46.2	52.6	57.9	53.0	68.8	77.6	59.2	82.0	71.9	40.7	33.3	35.8	38.2
M3-Agent [20]	44.4	54.8	62.3	56.8	54.0	53.5	52.4	58.5	38.5	42.1	52.6	52.0	68.4	72.4	50.8	70.4	65.5	53.0	40.7	48.5	49.3
WorldMM (Ours)																					
WorldMM-8B	49.6	56.4	63.9	58.4	58.7	56.4	48.2	63.4	53.9	52.6	57.9	52.0	69.6	73.6	65.2	70.4	69.7	55.0	54.1	59.8	55.4
WorldMM-GPT	**62.4**	**64.3**	**75.4**	**62.4**	**71.4**	**65.6**	**64.6**	**70.7**	**76.9**	**57.9**	**63.2**	**65.3**	**75.6**	**81.6**	**73.2**	82.8	**78.3**	58.3	**65.4**	**72.1**	**61.9**

ther the returned timestamps or the timestamps of the selected content. For the temporal grounding model, we use Time-R1 [35], with a slightly modified prompt that enables it to return both the evidence timestamps and the corresponding grounded responses. We sample videos at 0.5 fps and provide up to 768 frames. For embedding-based and hierarchical retrieval models, we follow the configurations described in Sec. B.1. Additionally, we include Qwen3 Emb., which applies the Qwen3-Embedding-4B [46] text encoder for caption retrieval, and InternVideo2, which encodes each segment using InternVideo2 [34] as an video encoder with uniform 16 frame averaging to enable segment-level retrieval. Both methods retrieve 30 second segments based on similarity search. For key frame selection, we apply AKS [29], which selects keyframes from the 0.5 fps sampled sequence. For tIoU evaluation, we interpret frames as representing their corresponding 30 second segments.

C.2. Efficacy of Memory Modules (Sec. 4.7)

To assess the contribution of each component within WorldMM's multimodal memory system, we evaluate several ablated variants in Sec. 4.7. In this section, we detail each variant of WorldMM created by selectively disabling a specific component. For episodic memory variants, we first construct a **fixed timescale** variant by replacing hierarchical episodic memory with a single fixed timescale memory. Specifically, we use the episodic memory of the finest granularity timescale. We also experiment an **embedding retrieval** variant in which the model's graph-based episodic retrieval is replaced with an embedding-based similarity search using Qwen-Embedding-4B. To examine the effect of semantic consolidation, we use a **w/o consolidation** version that bypasses the consolidation procedure to update the memory and instead store the raw extracted triplets without any update to existing memory. Finally, for visual memory, we ablate components of dual-retrieval mechanism by evaluating systems that rely exclusively on either **feature retrieval** through natural-language keyword search or **timestamp retrieval** based purely on temporal indices.

D. Detailed Experimental Results

In this section, we provide extended results and analyses of the experimental results.

Main results Tabs. 7 and 8 present the category-wise performance breakdown of WorldMM and baseline methods. Beyond overall benchmark averages, WorldMM consistently outperforms existing approaches across most categories. Notably, the gains are particularly pronounced in categories that rely on visual information. For instance, in the EntityRecall category of EgoLifeQA, where visual cues can help answering, WorldMM exceeds the previous best method, Ego-R1, by a substantial 11.2%. Similarly, on HippoVlog, our model achieves a 4% improvement in the Aud. and A+V categories, both of which require visual reasoning. These margins are greater than those observed in categories that do not explicitly depend on visual content, highlighting the strong advantage of our multimodal multi-memory architecture.

Table 8. Category-wise performance breakdown of WorldMM and baselines on Video-MME (L).

Model	ARES	AREC	ATTR	CNT	ISYN	OCR	ORES	OREC	SPER	SRES	TPER	TRES	Avg.
Base Models													
Qwen3-VL-8B [1]	62.2	54.0	51.9	43.8	68.1	42.9	62.9	57.4	33.3	45.5	33.3	67.0	61.0
Gemini 2.5 Pro [4]	56.9	47.6	66.7	41.7	71.8	**57.1**	53.3	40.7	0.0	**72.7**	**66.7**	48.4	55.7
GPT-5 [23]	71.1	69.8	**70.4**	47.9	**88.3**	**57.1**	**75.8**	74.1	33.3	**72.7**	50.0	75.8	74.3
Long Video LLMs													
VideoChat-Flash [18]	35.0	42.9	37.0	31.3	34.4	42.9	60.0	46.3	33.3	54.5	33.3	46.2	44.1
Time-R1 [35]	20.6	28.6	25.9	35.4	31.9	35.7	53.3	48.2	33.3	36.4	50.0	44.0	37.6
Video-RTS [36]	43.3	52.4	40.7	39.6	33.7	42.9	60.8	53.7	33.3	45.5	50.0	49.5	47.9
RAG-based Video LLMs													
LightRAG [10]	41.7	30.2	40.7	35.4	54.0	50.0	46.7	61.1	33.3	45.5	50.0	52.8	46.6
HippoRAG [11]	45.6	47.6	40.7	37.5	52.2	42.9	52.9	64.8	**66.7**	54.5	50.0	70.3	52.1
Video-RAG [21]	51.7	47.6	37.0	39.6	49.7	**57.1**	62.1	68.5	**66.7**	45.5	50.0	68.1	55.4
Memory-based Video LLMs													
EgoRAG [41]	31.1	55.6	33.3	22.9	41.1	28.6	44.6	48.2	33.3	54.5	**66.7**	48.4	41.1
Ego-R1 [30]	37.2	52.4	40.7	35.4	38.0	35.7	42.1	51.9	**66.7**	63.6	50.0	52.8	42.7
HippoMM [19]	41.1	42.9	55.6	35.4	38.7	35.7	37.9	53.7	33.3	54.5	50.0	47.3	41.6
M3-Agent [20]	52.2	57.1	59.3	45.8	51.5	42.9	54.6	64.8	33.3	45.5	50.0	71.4	55.3
WorldMM (Ours)													
WorldMM-8B	65.0	66.7	59.3	41.7	72.4	42.9	67.5	72.2	33.3	54.5	**66.7**	69.2	66.0
WorldMM-GPT	**81.1**	**73.0**	**70.4**	**54.2**	85.3	42.9	75.0	**77.8**	33.3	**72.7**	**66.7**	**79.1**	**76.6**

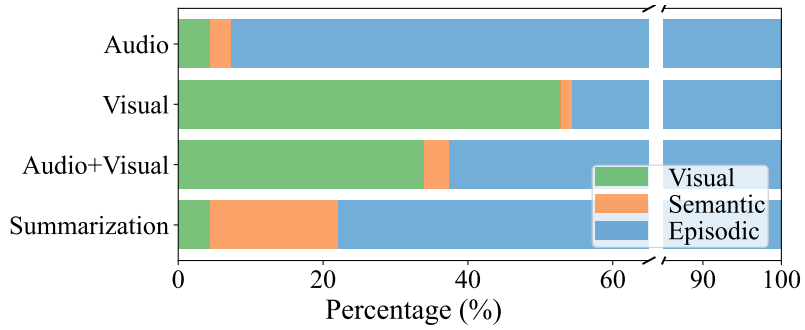

Figure 8. Memory type utilization of WorldMM on four distinctive categories in HippoVlog.

Efficacy of multimodal memory Fig. 8 shows memory type utilization of our model on HippoVlog benchmark, where categories are grouped by their modality requirements. The Audio category requires reasoning over spoken content and therefore is expected to depend primarily on textual memory derived from caption transcripts, while the Visual category focuses on visual understanding and correspondingly is designed to rely more on visual memory. Our results clearly support these expectations, showing that the Audio category predominantly activates textual memory while the Visual category relies heavily on visual memory, indicating that each category effectively leverages the required memory. Moreover, the Summarization category, which requires long-term reasoning, utilizes semantic memory more than any other category, demonstrating the complementary roles and effectiveness of each memory module in handling different reasoning demands. Together with this distribution of memory usage and the demonstrated performance gains in Tab. 2, these underscore the effectiveness of our multimodal multi-memory framework.

Dynamic temporal scope retrieval Tabs. 9 and 10 detail the per-category tIoU and accuracy results for WorldMM and baseline methods. While WorldMM significantly outperforms existing baselines on average, the results on LVBench particularly highlight the effectiveness of our dynamic episodic memory. In LVBench's Long category, where answering requires reasoning over more than five minutes of video, WorldMM outperforms the baselines by a notably larger margin than in categories that require shorter timescale, underscoring its ability to flexibly retrieve and integrate information over diverse temporal spans.

Table 9. Category-wise average tIoU (%) breakdown of WorldMM and dynamic temporal scope retrieval baselines.

Model	EgoLifeQA						Ego-R1 Bench						LVBench			
	Ent.	EvR.	Hab.	Rel.	Task	Avg.	Ent.	EvR.	Hab.	Rel.	Task	Avg.	Short	Med.	Long	Avg.
Time-R1 [35]	0.34	0.72	1.07	0.52	0.41	0.58	0.27	0.84	0.71	1.15	1.58	0.59	3.10	2.60	1.00	2.70
Qwen3 Emb. [46]	2.87	4.31	5.58	2.98	8.91	4.35	2.68	2.74	3.85	2.74	3.70	2.87	4.48	6.20	1.75	4.54
HippoRAG [11]	3.02	4.19	4.99	2.12	8.36	4.00	3.32	2.85	3.28	2.23	4.07	3.28	4.23	5.76	1.88	4.30
InternVideo2 [34]	2.09	4.42	6.04	2.00	3.88	3.36	2.71	2.55	3.09	1.85	2.32	2.60	3.66	4.71	0.87	3.55
EgoRAG [41]	3.20	3.38	4.62	3.10	4.82	3.60	2.40	3.07	4.08	2.19	3.78	2.73	4.10	3.38	0.91	3.50
Ego-R1 [30]	3.31	3.52	5.03	2.87	5.18	3.70	2.57	2.83	4.13	2.83	4.12	2.89	4.08	3.72	1.14	3.60
AKS [29]	2.42	2.77	3.08	2.93	2.67	2.75	2.03	2.48	2.99	2.58	3.04	2.30	3.81	4.11	1.10	3.52
WorldMM (Ours)	**9.79**	**10.43**	**11.85**	**7.73**	**12.97**	**10.09**	**8.91**	**9.85**	**8.86**	**9.63**	**9.58**	**9.17**	**7.53**	**14.41**	**10.02**	**9.57**

Table 10. Category-wise performance breakdown of WorldMM and dynamic temporal scope retrieval baselines.

Model	EgoLifeQA						Ego-R1 Bench						LVBench			
	Ent.	EvR.	Hab.	Rel.	Task	Avg.	Ent.	EvR.	Hab.	Rel.	Task	Avg.	Short	Med.	Long	Avg.
Time-R1 [35]	39.2	50.8	65.6	48.8	47.6	48.8	49.2	48.8	46.2	42.1	44.7	48.0	32.1	23.6	40.2	31.1
Qwen3 Emb. [46]	44.0	59.5	70.5	58.4	68.3	57.8	51.9	65.9	61.5	57.9	47.4	54.0	52.9	49.1	62.3	53.2
HippoRAG [11]	48.8	60.3	70.5	60.8	66.7	59.6	54.5	65.9	69.2	52.6	50.0	56.0	54.9	47.5	62.3	54.0
InternVideo2 [34]	40.8	54.0	60.7	51.2	52.4	50.6	50.3	56.1	46.2	47.4	52.6	51.0	47.4	37.3	53.4	45.7
EgoRAG [41]	40.0	56.3	62.3	54.4	52.4	52.0	46.6	56.1	46.2	47.4	55.3	49.0	32.4	32.0	31.9	32.2
Ego-R1 [30]	51.2	53.2	63.9	50.4	50.8	53.0	50.8	63.4	38.5	36.8	57.9	52.0	32.5	36.5	37.3	34.1
AKS [29]	41.6	51.6	63.9	51.2	52.4	50.6	51.3	63.4	46.2	36.8	50.0	51.7	43.3	33.9	39.2	40.4
WorldMM (Ours)	**62.4**	**64.3**	**75.4**	**62.4**	**71.4**	**65.6**	**64.6**	**70.7**	**76.9**	**57.9**	**63.2**	**65.3**	**58.3**	**65.4**	**72.1**	**61.9**

E. Additional Experimental Results

We additionally present experimental results supporting the design of WorldMM, including ablation studies on backbone configurations (Sec. E.1) and the impact of temporal scales (Sec. E.2).

E.1. Generalization to Different Backbones

To evaluate the flexibility and robustness of WorldMM across different backbone models, we conduct experiments with a diverse set of configurations. Specifically, in addition to the setup based on the GPT-5 model series and VLM2Vec-V2, we further incorporate Gemini 3 Flash [9] and Qwen3-VL-Embedding-2B [17]. As shown in Tab. 11, WorldMM demonstrates strong robustness to backbone selection, with the Gemini-based variant even outperforming others on EgoLifeQA. These results highlight that WorldMM generalizes well across different backbone architectures and can be seamlessly integrated with a wide range of state-of-the-art models without requiring architecture-specific modifications.

Table 11. Performance of WorldMM with various backbones.

Model	EgoLifeQA	LVBench	VideoMME (L)
WorldMM-Gemini + Qwen3-VL-Emb	67.4	61.5	74.9
WorldMM-Gemini + VLM2Vec-V2	**68.2**	61.7	75.8
WorldMM-GPT + Qwen3-VL-Emb	66.0	61.4	75.8
WorldMM-GPT + VLM2Vec-V2	65.6	**61.9**	**76.6**

E.2. Impact of Temporal Scales

While multiscale episodic memory improves overall performance, we verify that these gains result from the multiscale architecture rather than specific temporal constraints. The temporal scales used in our experiments are chosen based on empirical statistics of real-world event durations. To assess the sensitivity of WorldMM to these specific values, we introduce perturbations to the temporal scales and report the performance on EgoLifeQA in Tab. 12. The results demonstrate that WorldMM maintains consistent performance across these variations, indicating that the improvements stem from the multiscale memory design itself, rather than a reliance on precisely calibrated temporal windows.

Table 12. Performance with different episodic timescales.

Temporal Scale	Acc
20s/2m/5m/50m	65.2
30s/3m/10m/1h	65.6
1m/5m/15m/1.5h	64.8

F. Qualitative Results

In this section, we qualitatively analyze WorldMM's memory construction (Sec. F.1) and its multi-turn reasoning and refinement capabilities (Sec. F.2).

F.1. Memory Construction

Tab. 13 presents an example of episodic triplet extraction. Given a caption generated from sampled frames of a segment along with its corresponding transcript, an LLM is prompted (using the prompt in Fig. 11) to extract episodic triplets. Semantic triplets are extracted using a different prompt (Fig. 14), designed to focus on long-term dependencies and capture more abstract relationships across the segments, as shown in Tab. 14. To better capture persistent knowledge across segments, we introduce semantic consolidation, which incrementally updates the semantic graph by integrating new triplets and resolving conflicts. Using embedding-based matching and an LLM, duplicated or conflicting triplets are removed, and new or revised ones are added, generating an evolving semantic memory, as shown in Tab. 15. For instance, the new triplet "[I, uses WeChat for, money transfers]" is merged with the existing triplet to consolidate redundant information, and conflicting triplets, such as "[Lucia, dislikes, overly sweet food]" versus "[Lucia, likes, sweet desserts]", are removed to ensure consistency in the semantic memory.

F.2. Multi-turn Refinement

WorldMM demonstrates the effectiveness of multi-turn reasoning by progressively refining its retrieval strategy to answer questions, as shown in Tab. 16. In this example, the first round retrieves episodic memory using a narrow keyword focused on the "discussion" of the air conditioning, but it provides insufficient detail about the activity. In the second round, the model expands to a more general keyword, "air conditioning", which enables retrieval of every scene where the air conditioning is involved to obtain sufficient textual evidence. Moreover, in the third round, since the textual evidence fails to capture specific visual details of the scene, WorldMM refines its strategy to retrieve video frames corresponding to the relevant timestamp. Through this stepwise process, WorldMM effectively refines its search strategy with different keyword strategies and memory types to respond to the question.

G. Limitation and Broader Impact

While WorldMM serves as an effective multimodal memory agent for long video reasoning, it still requires careful preprocessing, including video captioning, triplet extraction, and semantic consolidation. Yet, this limitation is not unique to our approach but a broader constraint shared by existing memory-based video LLMs. For example, M3-

Agent [20] incurs even heavier preprocessing due to its reliance on entity recognition, and many other approaches operate with offline preprocessing. In contrast, WorldMM is designed for online operation. Memories are updated at fixed intervals (e.g., every 10 seconds), and the required preprocessing for each segment can be performed within these windows. Moreover, new information can be seamlessly integrated into the knowledge graph, and our consolidation mechanism efficiently refines the knowledge base without requiring the reconstruction of memory from scratch.

With strong long-term reasoning capabilities and support for real-time updates, WorldMM serves as a practical solution for streaming scenarios such as egocentric assistants and embodied agents. This foundation enables richer and more persistent assistance for everyday tasks and accessibility. However, the continuous accumulation of structured knowledge over periods of time raises serious privacy and security concerns. Real-world deployments must therefore enforce safeguard policies, including strict access controls, secure data handling, and privacy protections.

Table 13. Example of episodic triplet extraction.

Caption	I stand and walk to the other side of the dining table. Katrina asks, "Is this for tomorrow's game?" "Yes—let's think about what to do tomorrow," I say. I raise my right hand as Katrina walks toward me. Lucia asks, "Using ancient poems? Or what else?" Katrina says, "I'm not good with ancient poems." Tasha asks, "Then what else to use?" Katrina says, "I'll be out in the first round. My room is already cleaned up." "Okay," I say. I turn toward the stairs, put down my phone, look back at the living room door, and walk into the second-floor living room. Lucia adds, "For example, not coming out." Katrina says, "Let me check that place we're going to." Tasha asks, "I just want to ask which fields it has expanded into." Lucia says, "Okay."
Extracted Triplets	[I, stand at, dining table] [I, walk to, other side of the dining table] [Katrina, asks about, tomorrow] [I, confirm, tomorrow] [I, raise, right hand] [Katrina, walks toward, I] [Lucia, asks about, using ancient poems] [Katrina, says, not good with ancient poems] [Tasha, asks, what else to use] [Katrina, says, I will be out in the first round] [Katrina, has, room already cleaned up] [I, turn toward, stairs] [I, put down, phone] [I, look back at, living room door] [I, walk into, second-floor living room] [Lucia, adds, not coming out as an example] [Katrina, says, let me check that place we're going to] [Lucia, says, Okay]

Table 14. Example of semantic triplet extraction.

Caption	I got up, moved my phone, and checked it before turning it off. Alice expressed her feelings towards me, and I responded by checking my phone's chat interface. Alice then questioned her appearance, and I turned off the phone, looking around at the snacks and utensils on the table. I stood up, grabbed a pack of snacks, and proceeded to my room to enjoy them. Alice asked about something being fancy, and I fetched my glasses, placing them on the table. ... I managed my phone, swiping through pages, and interacted with others as I went about my tasks. I observed Alice and Tasha, discussing what to feed a cat, and continued interacting with my phone. As the environment darkened, I engaged with the surroundings, noting the layout and structures. Finally, I moved towards a house with blue-green walls, managing my power bank and surveying the area.
Extracted Triplets	[I, assigns tasks to, Katrina] [I, handles reimbursements for, Alice] [I, uses WeChat for, money transfers] [I, often eats, snacks] [I, wears, glasses] [Lucia, dislikes, overly sweet food] [Alice, expresses romantic feelings toward, I] [Katrina, helps with, expense tracking] [I, requires PDFs for, reimbursement] [Tasha, participates in, house demolition tasks] [Lucia, participates in, house demolition tasks]

Table 15. Example of semantic consolidation.

Original Triplets	[I, uses WeChat to send money]	
	[I, wears, glasses]	
	[I, often eats, fruits]	
	[Lucia, likes, sweet desserts]	
	[Tasha, participates in, household projects]	
New Triplets	[I, assigns tasks to, Katrina]	
	[I, handles reimbursements for, Alice]	
	[I, uses WeChat for, money transfers]	
	[I, often eats, snacks]	
	[I, wears, glasses]	
	[Lucia, dislikes, overly sweet food]	*% conflicts with existing "likes sweet desserts"*
	[Alice, expresses romantic feelings toward, I]	
	[Katrina, helps with, expense tracking]	
	[I, requires PDFs for, reimbursement]	
	[Tasha, participates in, house demolition tasks]	
	[Lucia, participates in, house demolition tasks]	
Consolidated Triplets	[I, assigns tasks to, Katrina]	
	[I, handles reimbursements for, Alice]	
	[I, uses, WeChat to send money]	*% merged with "uses WeChat for money transfers"*
	[I, often eats, fruits and snacks]	*% merged and generalized "fruits" and "snacks"*
	[I, wears, glasses]	*% merged with duplicate in new triples*
	[Alice, expresses romantic feelings toward, I]	
	[Katrina, helps with, expense tracking]	
	[I, requires PDFs for, reimbursement]	
	[Tasha, participates in, house demolition tasks]	*% refined general triple "household projects"*
	[Lucia, participates in, house demolition tasks]	

Table 16. Example of the multi-turn refinement of WorldMM.

Question	What were we doing last time we discussed the air conditioning temperature? (A) Eating hot pot (B) Shopping (C) Playing piano (D) Picking up a package
Answer	A
Round 1	Decision: Search // Memory: Episodic Search Query: discussing the air conditioning temperature *Retrieved:* *[DAY2 13:36:00-13:39:00]* *I joined my colleagues at the dining table and engaged in a discussion about the food while preparing to eat. ... During the meal, we discussed the air conditioner situation in the rooms. Tasha and Katrina talked about where the remotes were and who had access, with Shure mentioning that his room's AC had been on continuously due to a borrowed remote from me. I listened to their conversation, ...* *[DAY2 17:48:00-17:51:00]* *I coordinated with Shure and Lucia to find a suitable spot for my power bank, initially placing it on a flower pot as Shure suggested, but realizing it was unstable. ... We discussed the room environment, including turning off the air conditioner that was blowing cold air. Shure clarified that the remote we had was for the TV, not the air conditioner, prompting me to inquire about the correct remote. ...*
Round 2	Decision: Search // Memory: Episodic Search Query: air conditioning *Retrieved:* *[DAY2 17:48:30-17:49:00]* *"Flip it and put it underneath," I say. ... "Wow, this, this..." Lucia murmurs. "How do you turn off this air conditioner? It's blowing so cold up here." I adjust the power bank for a better position. Shure says, "Just tell it to turn off, it'll turn off." "Really? It's voice-controlled?" Lucia asks. I secure the power bank and turn to Shure and Lucia. Shure adds, "No, this is the TV remote." "Where is it?" I ask, and I chat with them.* *[DAY2 18:34:01-18:34:29]* *I watch Shure set the air conditioner to 26 degrees. "Save the Earth," I say. "Everyone's responsibility," I laugh. I kneel, then sit on the sofa. Shure asks, "Seriously, you think it's good to record this?" I look at my friends chatting and watch the movie. Shure says, "We're still using the AC." Shure says, "He really can't hear a thing." I watch Shure take a few steps. I laugh. Shure says, "Hahaha, it's like we're watching a silent movie." I smile and keep watching the movie. "Whoa, this is classic," I say. I stand up. "Come on, let's have a drink," I say.* *[DAY2 18:33:00-18:33:30]* *I walk in front of the fill light, pass behind Choiszt, and move toward the curtain. ... "The popularity is always high," I say. Shure: "Ha." Choiszt: "Turn it on." "Turn on the air conditioner?" I ask. Lucia: "I'll just, I suggest I drink water directly from a bowl." Lucia: "So later I'll eat with this bowl."*
Round 3	Decision: Search // Memory: Visual Search Query: DAY2 18:34:01-18:34:29 *Retrieved:*  *(Author Note: They are gathered around a table cooking and sharing a hot pot meal while watching a projection on the wall.)*
Round 4	Decision: Answer
Response	A

You are an expert video captioner.

You will receive a short video segment represented by ordered frames and optional transcript lines with timestamps. Write a caption describing both the visual content and the audible content of the segment.

Guidelines:
- Describe visible actions, people, objects, and environment.
- Include relevant speech, sounds, or audio events.
- Keep the caption factual and neutral.
- Do not mention frames, timestamps, or that the input came from frames.
- Avoid speculation about emotions or intentions unless clearly visible or stated in speech.

Output only the final caption text.

Figure 9. Prompt for video caption generation.

Your task is to extract named entities from the given paragraph. Respond with a JSON list of entities.

Example:

Radio City is India's first private FM radio station and was started on 3 July 2001. It plays Hindi, English and regional songs. Radio City recently forayed into New Media in May 2008 with the launch of a music portal - PlanetRadiocity.com that offers music related news, videos, songs, and other music-related features.

{ "named_entities":
["Radio City", "India", "3 July 2001", "Hindi", "English", "May 2008", "PlanetRadiocity.com"]
}

Figure 10. Prompt for named entity recognition (NER). Recognized named entities are used to extract episodic triplets as shown in Fig. 11.

Your task is to construct an RDF (Resource Description Framework) graph from the given passages and named entity lists. Respond with a JSON list of triples, with each triple representing a relationship in the RDF graph.

Pay attention to the following requirements:
- Each triple should contain at least one, but preferably two, of the named entities in the list for each passage.
- When resolving pronouns, if the pronoun refers to the first-person (e.g., I, me, my), keep it as "I" instead of replacing with terms like "speaker" or "narrator". For other pronouns, clearly resolve them to their specific names to maintain clarity.

Convert the paragraph into a JSON dict, it has a named entity list and a triple list.

Example:

Radio City is India's first private FM radio station and was started on 3 July 2001. It plays Hindi, English and regional songs. Radio City recently forayed into New Media in May 2008 with the launch of a music portal - PlanetRadiocity.com that offers music related news, videos, songs, and other music-related features.

{ "named_entities":
    ["Radio City", "India", "3 July 2001", "Hindi", "English", "May 2008", "PlanetRadiocity.com"]
}

{ "triples": [
    ["Radio City", "located in", "India"],
    ["Radio City", "is", "private FM radio station"],
    ["Radio City", "started on", "3 July 2001"],
    ["Radio City", "plays songs in", "Hindi"],
    ["Radio City", "plays songs in", "English"],
    ["Radio City", "forayed into", "New Media"],
    ["Radio City", "launched", "PlanetRadiocity.com"],
    ["PlanetRadiocity.com", "launched in", "May 2008"],
    ["PlanetRadiocity.com", "is", "music portal"],
    ["PlanetRadiocity.com", "offers", "news"],
    ["PlanetRadiocity.com", "offers", "videos"],
    ["PlanetRadiocity.com", "offers", "songs"]
]}

Figure 11. Prompt for episodic triplet extraction.

As an Event Summary Documentation Specialist, your role is to systematically structure and summarize event information, ensuring that all key actions of major characters are captured while maintaining clear event logic and completeness. Your focus is on concise and factual summarization rather than detailed transcription.

Specific Requirements

1. Structure the Events Clearly
- Merge related events: Consolidate similar content into major events and arrange them in chronological order to ensure a smooth logical flow.
- Logical segmentation: Events can be grouped based on location, task, or theme. Each event should have a clear starting point, progression, and key turning points without any jumps or fragmentation in the information.

2. Retain Key Information
- All subjects' decisions and actions must be fully presented, including all critical first-person activities. Transitions between different parts, such as moving between floors or starting/ending a task, should be seamless.
- Any discussions, decisions, and task execution involving the primary character and other key individuals that impact the main storyline must be reflected. This includes recording, planning, and confirming matters, but in a concise manner.
- The purpose and method of key actions must be recorded, such as "ordering takeout using a phone" or "documenting a plan on a whiteboard."

3. Concise Expression, Remove Redundancies
- Keep the facts clear, avoiding descriptions of atmosphere, emotions, or abstract content.
- Remove trivial conversations and extract only the core topics and conclusions of discussions. If a discussion is lengthy, summarize it into task arrangements, decision points, and specific execution details.

4. Strictly Adhere to Facts, No Assumptions
- Do not make assumptions or add interpretations—strictly organize content based on available information, ensuring accuracy. Every summarized point must have a basis in the original information, with no unnecessary additions.
- Maintain the correct chronological order of events. The sequence of developments must strictly follow their actual occurrence without any inconsistencies.

Output Format
Each paragraph should represent one major event, structured in a summary-detail-summary format. Strictly output below {word_limit} words in total. Do not report the word count in the output.

Figure 12. Prompt for episodic memory construction to generate coarser-level caption.

You are an expert assistant that helps filter and select relevant video captions based on a given query. Your task is to analyze the retrieved video captions and determine which ones are most relevant to answer the question.

Given the following question and retrieved video captions, select and rank the most relevant captions that should be used to answer the question.

Instructions:
1. Consider the nature of the question when selecting captions:
   - e.g., for queries about specific events, focus on finer granularities; for habitual, relationship, or general queries, consider coarser granularities.
   - Note that coarser granularity captions may provide broader context, but finer granularity captions often contain more specific details.
2. Each caption shows its time range (start_time to end_time)
3. Analyze each caption for relevance to the question
4. Select captions that directly help answer the question
5. Return the IDs in ranked order (most relevant first)
6. Only include captions that are truly relevant

Return ONLY a JSON array of caption IDs in order of relevance (most relevant first), without additional justification.

Figure 13. Prompt for episodic memory retrieval to select from multiple timescales.

You are tasked with extracting semantic knowledge from episodic triples. Your goal is to infer generalizable information that extends beyond the specific episode. Focus on capturing valid semantic triples that can guide reasoning about behavior, relationships, or preferences.

What to Extract
1. Relationships: social bonds or roles between entities that persist over time
   (e.g., “Alice is a friend with Bob”, “Jason is a teacher of Alice”).
2. Attributes & Preferences: tendencies, likes/dislikes, personality-like traits, or behavioral habits
   (e.g., “Alice prefers not having dessert”, “Bob enjoys music”).
3. Habits & Capabilities: actions or patterns that suggest what an entity often does, can do, or tends to do
   (e.g., “Alice often helps friends”, “Jason can give advice”).
4. Conceptual Knowledge: directly useful facts that support reasoning, but avoid overly broad taxonomic statements
   (e.g., “Alice’s office is near Cafe X”, “Bob’s gym is closed on Sundays”).

What to Avoid
- One-off events or transient states (e.g., “ate pizza yesterday”, “was late once”) unless explicitly declared as a preference/role
- Broad taxonomy or trivia unrelated to behavior (e.g., “a laptop is electronics”, “Paris is in France”)
- Speculative or mind-reading inferences without textual support (e.g., motives, beliefs not evidenced)

Important Notes
- Prefer to base semantic triples on multiple supporting episodes.
- BUT if a single episode clearly reflects a role, preference, habit, or capability, it is valid to include it.
- Each semantic triple MUST have at least one supporting episodic triple.
- Reduce duplication. If multiple episodic triples support the same or very similar semantic knowledge, merge them into one semantic triple rather than repeating.
- The ‘episodic_evidence[i]’ list must always point to the indices that support ‘semantic_triples[i]’.
- Aim for broad coverage: extract as many valid semantic triples as reasonably supported by the input.

Output Format
- Return ONLY a JSON object with the following two keys:
  - ‘semantic_triples’ (List[List[str]]): Each item is a triple [subject, predicate, object].
  - ‘episodic_evidence’ (List[List[int]]) : Each item is a list of 0-based indices pointing to the input episodic triples that support the corresponding semantic triple at the same position.
- The two lists MUST have the same length and aligned order.
- If no semantic knowledge is inferable, return: {“semantic_triples”: [], “episodic_evidence”: []}

Example:

Episodic triples:
0. [“Alice”, “talks to”, “Bob”],
1. [“Alice”, “laughs with”, “Bob”],
2. [“Alice”, “doesn’t eat cake”, “at restaurant”],
3. [“Alice”, “shares personal stories with”, “Bob”],
4. [“Alice”, “brings coffee to”, “Bob”],
5. [“Jason”, “talks to”, “Alice”],
6. [“Alice”, “declines dessert”, “at friend’s house”]

Output:
```
{
    “semantic_triples”: [
        [“Alice”, “is a friend with”, “Bob”],
        [“Alice”, “prefers”, “not having dessert”]
    ],
    “episodic_evidence”: [
        [0, 1, 3],
        [2, 6]
    ]
}
```

Figure 14. Prompt for semantic triplet extraction.

You are tasked with consolidating semantic knowledge by processing a new semantic triple against relevant existing knowledge from previous timestamps.

Your job is to make two decisions:
1. Which existing triples to remove/pop — those that should be merged with the new triple or conflict with it
2. How to update the new triple — to capture merged information or resolve conflicts

Consolidation Rules
1. Merge Similar Information: If existing triples express very similar information to the new triple, remove them and update the new triple to contain the most complete/accurate form.
2. Resolve Conflicts: If the new triple conflicts with existing ones, decide which is more accurate/recent and remove outdated ones.
3. Update with Context: Use information from existing triples to make the new triple more specific or more accurate.
4. Preserve Unique Information: Only remove existing triples when they are redundant or conflicting.

Output Format
Return ONLY a JSON object with the following two keys:
- 'updated_triple' (List[str]): The new triple, possibly updated [subject, predicate, object].
- 'triples_to_remove' (List[int]): Indices of existing triples to remove (empty list if none).

Example:

New triple: ["Alice", "enjoys", "coffee"]

Existing triples:
0. ["Alice", "likes", "beverages"]
1. ["Alice", "favors", "to have coffee after dinner"]
2. ["Alice", "prefers", "hot drinks"]
3. ["Alice", "likes to drink", "coffee"]

Output:
{
    "updated_triple": ["Alice", "likes", "coffee"],
    "triples_to_remove": [1, 3]
}

Figure 15. Prompt for semantic memory consolidation.

You are a reasoning agent for a video memory retrieval system. Your job is to decide whether to stop and answer, or to search memory for more evidence. When searching, you must select exactly one memory type and form a query.

Decision Modes
1. search: Retrieve memory to begin, continue, or extend progress toward the answer
   - Choose one memory type and form a keyword(phrase)-style search query.
2. answer: Stop searching because the accumulated results are sufficient.
   - No memory type selection is needed.

Memory Types
1. Episodic: Specific events/actions. Stores memories of past events and actions. Query by EVENT/ACTION.
2. Semantic: Entities/relationships. Stores factual knowledge about entities and their relationships, roles, and habits. Query by ENTITY/CONCEPT.
3. Visual: Scene/setting snapshots. Stores visual snapshots of scenes and settings. Query by SCENE/SETTING or TIMESTAMP RANGE.
   - For timestamp range queries, return in the format: DAY X HH:MM:SS - DAY Y HH:MM:SS.
Context Inputs
- Current Query
- Round History: Log of past retrieval rounds. Each round is written in this format:

    ### Round N
    Decision: <search|answer>
    Memory: <episodic|semantic|visual>
    Search Query: <query text>
    Retrieved: <retrieved items>

Strict Output Rules
- If decision = "search": Must include "selected_memory" with exactly one memory type and one query.
- If decision = "answer": Do NOT include "selected_memory".
- Always output in valid JSON only, no extra commentary.

Output Format
```
{
    "decision": "search" | "answer",
    "selected_memory": {
        "memory_type": "episodic" | "semantic" | "visual",
        "search_query": <str>
    } # Omit if decision = "answer"
}
```

(Few-shot examples given)

Figure 16. Prompt for retrieval agent to decide retrieval strategy.

You are an AI assistant that answers questions about video using retrieved memory context. Your task is to answer multiple choice questions based on this accumulated context. Always choose the most relevant answer from the given choices based on the evidence provided.

Guidelines
- Analyze all provided context carefully.
- Choose the answer that best matches the evidence.
- If evidence is unclear, make the most reasonable inference.

Output Format
Provide your answer as a single letter (A, B, C, or D) based on the evidence.

Figure 17. Prompt for response agent to generate response based on retrieved results.