WorldMM: Dynamic Multimodal Memory Agent for Long Video Reasoning

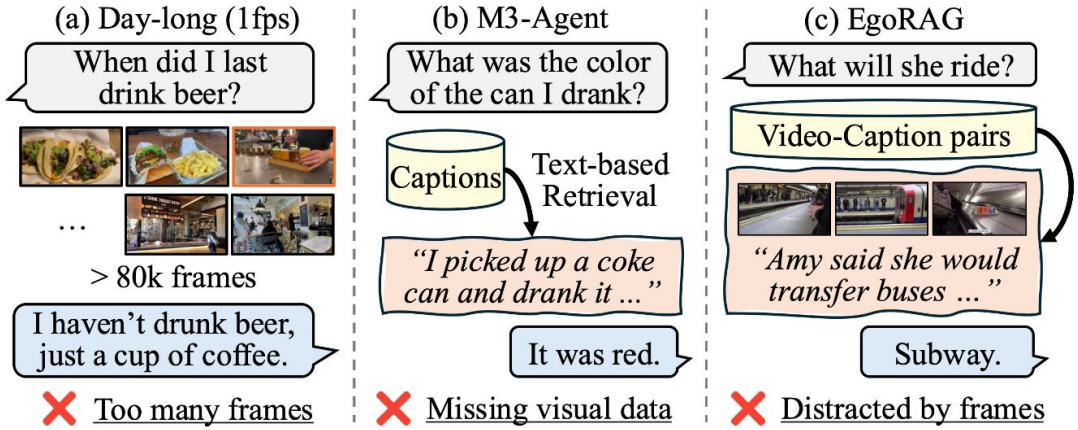
Woongyeong Yeo[1]*, Kangsan Kim[1]*, Jaehong Yoon[2†], Sung Ju Hwang[1,3†]

KAIST[1], NTU Singapore[2], DeepAuto.ai[3]

*: Equal contribution; †: Equal advising

Understanding Long Video

Processing every frame is impractical and often ineffective due to excessive context, and recent approaches focus on converting **videos into textual** representations.



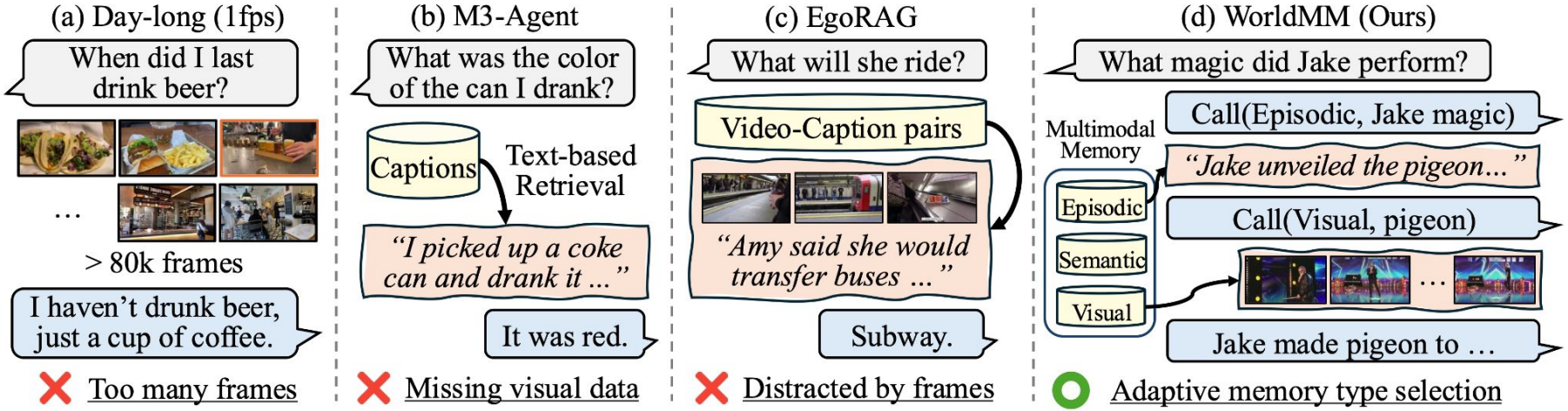
Yet key questions remain:

- Can visual details be fully represented via text?
- Does retrieving both text and visuals always help?

[1] EgoLife: Towards Egocentric Life Assistant, CVPR 2025.
[2] Seeing, Listening, Remembering, and Reasoning: A Multimodal Agent with Long-Term Memory, ICLR 2026.

Multimodal Memory

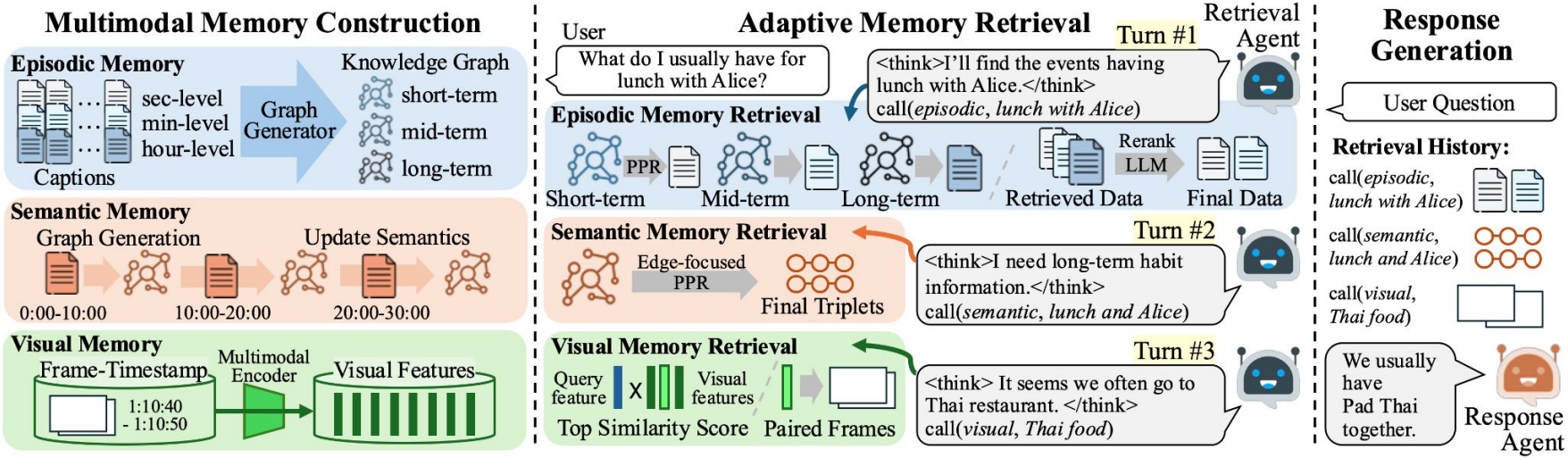
To fully exploit multimodal information, rather than relying on text or exhaustive input context, we propose to leverage **multiple types of memories**, consisting of both textual and visual representations, with **adaptive memory retrieval**.



WorldMM

WorldMM is a dynamic multimodal memory agent that

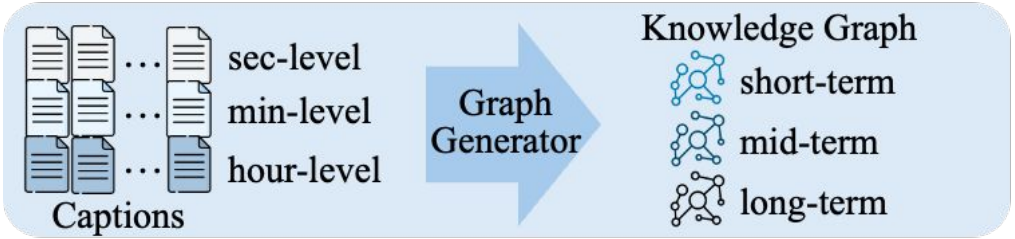
- builds separate multimodal memory at **multiple modalities and granularities**;
- and **iteratively selects** the most relevant memory.



Episodic Memory

Episodic memory **encodes events into knowledge graphs** generated from video captions.

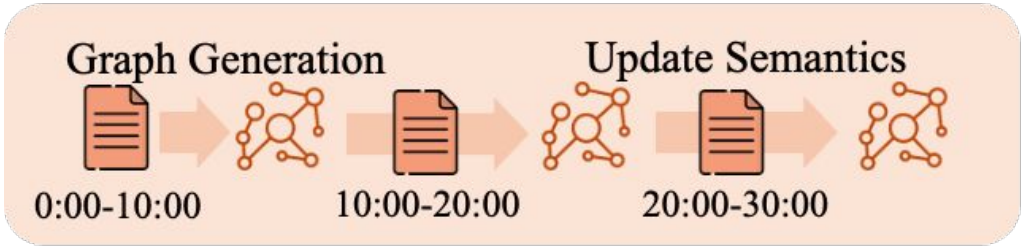
As relying on a single fixed temporal scale is impractical, we formulate a **multi-scale memory** with multiple temporal resolutions to encode events at varying scales.



e.g., *“Where did I leave my glasses?”* vs. *“What happened in the second half of the soccer match?”*

Semantic Memory

Semantic memory captures **long-term, evolving** knowledge about relationships and habits.



We introduce a **consolidation** process to continuously update and compress the knowledge.

Existing Semantic Memory
[I, wears, glasses]
[I, often eats, fruits]
[Lucia, likes, sweet desserts]
[Tasha, participates in, household projects]

I stood up, grabbed a pack of snacks, and proceed to my room to enjoy them. Alice asked about something being fancy, and I fetched my glasses, …

→

[I, often eats, snacks]
[I, wears, glasses]
[Lucia, dislikes, overly sweet food]
[Alice, expresses romantic feelings toward, I]
[Tasha, participates in, house demolition tasks]

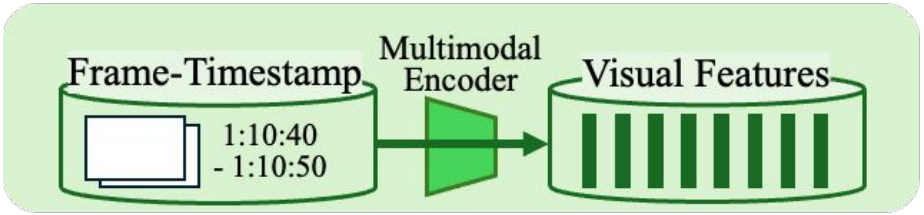
→

Updated Semantic Memory
[I, wears, glasses]
[I, often eats, fruits and snacks]
[Alice, expresses romantic feeling toward, I]
[Tasha, participates in, household projects]

Visual Memory

Visual memory captures **rich visual details** that cannot be fully conveyed through text by storing selected frames. We support two scenarios:

1. Keyword-based Search : Encode segments into visual features with multimodal encoder



2. Timestamp-based Search : Retrieves a clip for the specified time range when required
   e.g., *"What ingredients did Alice use to make the cookies?"*

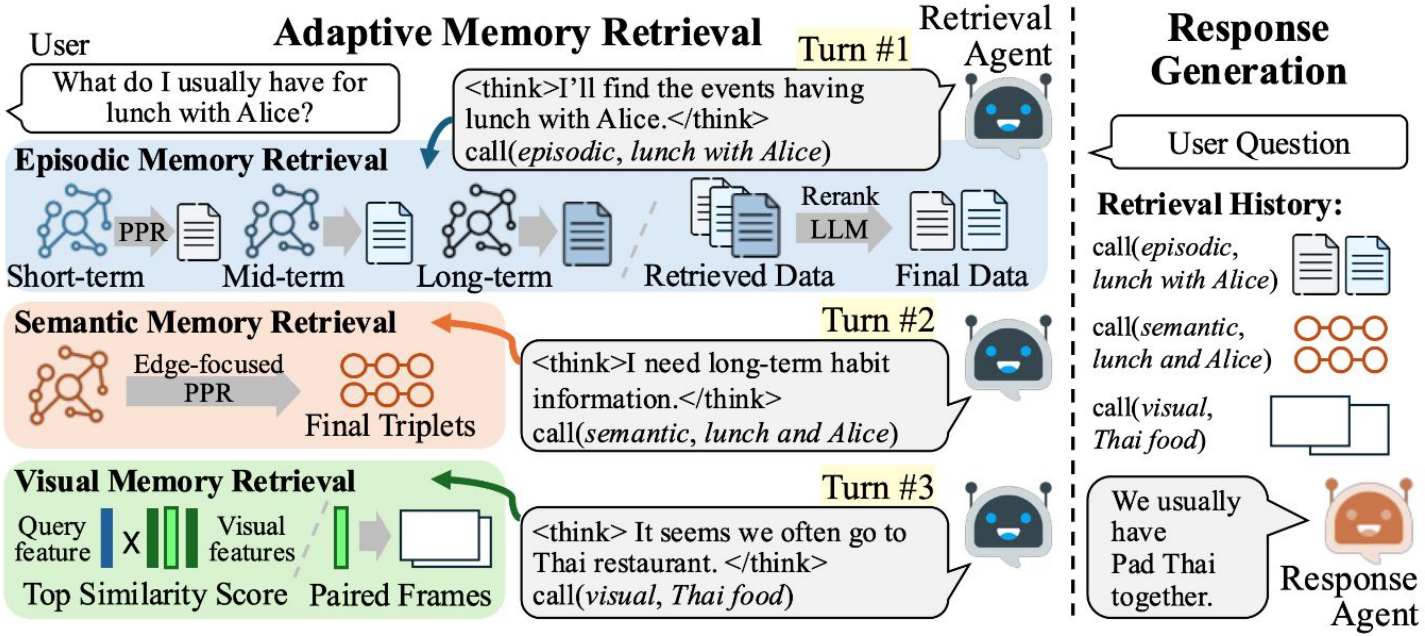
[DAY 5 17:00:00 - 17:30:00]

Alice baked cookies. She combined several ingredients and mixed them together, then shaped the dough. After that, she preheated the oven and placed the cookies …

<think> We need a visual of Alice combining ingredients. </think>
~~call(visual, Alice combining ingredients)~~
call(visual, DAY5 17:00:00-17:10:00)

Adaptive Memory Retrieval

The **retrieval agent** iteratively decides **which memory to access and what query to issue**, conditioned on the user question and retrieval history.



Main Results

Table 1. Performance of WorldMM with various baselines across long video QA benchmarks. "–" denotes a proprietary backbone.

Model		EgoLife QA	Ego-R1 Bench	Hippo Vlog	LV Bench	Video-MME (L)	Avg.
Base Models							
Qwen3-VL-8B [1]	8B	38.6	35.7	74.4	48.3	61.0	51.6
Gemini 2.5 Pro [3]	–	46.4	46.7	72.0	57.0	55.7	55.6
GPT-5 [16]	–	48.6	46.3	75.7	60.4	74.3	61.1
Long Video LLMs							
VideoChat-Flash [11]	7B	34.2	42.7	58.0	33.2	44.1	42.4
Time-R1 [28]	3B	48.8	48.0	54.6	31.1	37.6	44.0
Video-RTS [29]	7B	48.2	47.4	59.0	39.8	47.9	48.6
RAG-based Video LLMs							
LightRAG [6]	–	48.8	52.3	47.4	30.4	46.6	45.1
HippoRAG [7]	–	59.6	56.0	63.2	54.0	52.1	57.0
Video-RAG [14]	–	55.4	49.7	65.1	33.1	55.4	51.7
Memory-based Video LLMs							
EgoRAG [33]	–	52.0	49.0	57.5	32.2	41.1	46.4
Ego-R1 [23]	3B	53.0	52.0	58.8	34.1	42.7	48.1
HippoMM [12]	–	54.6	53.0	71.9	38.2	41.6	51.8
M3-Agent [13]	7B	53.5	52.0	65.5	49.3	55.3	55.1
WorldMM (Ours)							
WorldMM-8B	8B	56.4	52.0	69.7	55.4	66.0	59.9
WorldMM-GPT	–	**65.6**	**65.3**	**78.3**	**61.9**	**76.6**	**69.5**

- WorldMM-GPT achieves an average score of 69.5%, exceeding the strongest baseline by **8.4%.**
- Retrieval- and memory-based methods score higher, showing **selective retrieval works better**.
- Outstanding performance of WorldMM highlights **integrating textual and visual memory** and **adaptively selecting temporal scopes** are crucial.

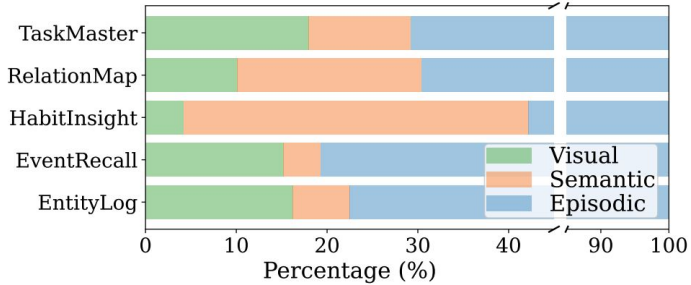
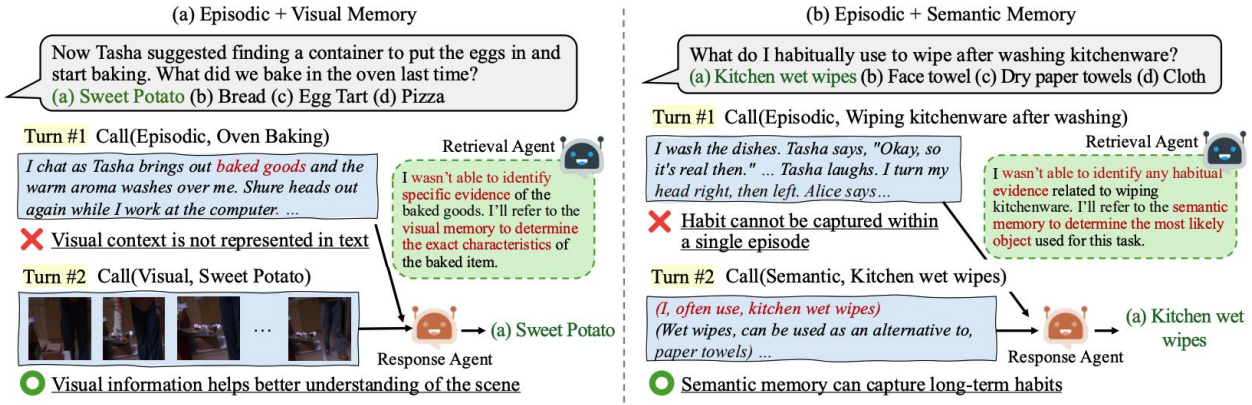


Figure 3. Memory type utilization of WorldMM on five distinctive categories in EgoLifeQA.

Effect of Memories

Visual memory improves **perceptual understanding** by preserving visual details, while semantic memory enables **long-term and relational reasoning** across events.

Model	EgoLifeQA						Ego-R1 Bench						HippoVlog					LVBench	Video-MME (L)	Avg.
	Ent.	EvR.	Hab.	Rel.	Task	Avg.	Ent.	EvR.	Hab.	Rel.	Task	Avg.	Aud.	Vis.	A+V	Summ.	Avg.			
E	57.6	61.1	70.5	61.6	69.8	62.6	54.5	70.7	53.9	52.6	57.9	57.0	72.4	73.2	68.4	80.4	73.6	60.6	72.7	64.9
V	40.8	35.7	36.1	34.4	39.7	37.2	36.5	34.1	23.1	31.6	28.2	34.2	35.2	66.4	54.8	48.8	51.3	47.4	64.2	44.9
E+S	56.8	61.9	73.8	62.4	71.4	63.4	59.3	68.3	69.2	57.9	60.5	61.0	70.8	75.2	68.8	80.4	73.8	58.8	74.1	66.8
E+V	59.2	63.5	70.5	60.8	68.8	63.3	65.1	68.3	53.9	47.4	57.9	63.0	73.2	77.2	70.8	79.6	75.2	59.8	76.0	66.9
E+S+V	**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**	**61.9**	**76.6**	**69.5**



Efficiency Analysis

WorldMM achieves the **best accuracy–latency trade-off** among all methods.
By adaptively retrieving **only relevant video segments**, it delivers much higher accuracy with lower inference time than long-video LLMs and RAG-based approaches.

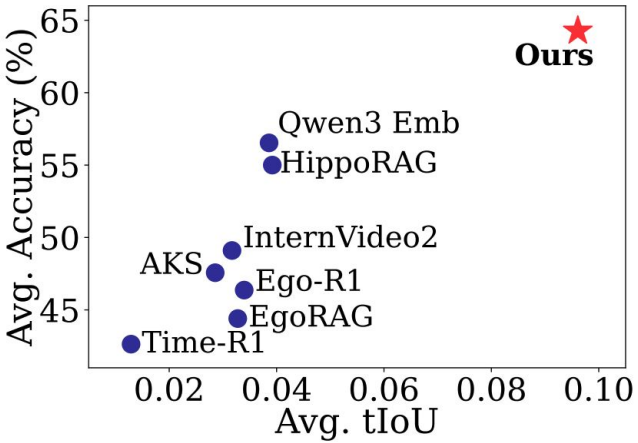


Figure 5. Average tIoU and performance of WorldMM and baselines.

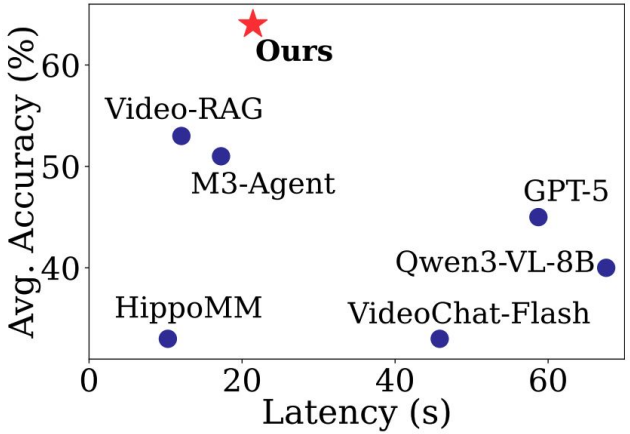


Figure 6. Average latency and performance of WorldMM and baselines.

Thank you!

We'd be happy to discuss our research further.

Woongyeong Yeo (wgcyeo@kaist.ac.kr)

Kangsan Kim (kangsan.kim@kaist.ac.kr)

Project Page

Paper