EgoGraph: Temporal Knowledge Graph for Egocentric Video Understanding

Shitong Sun
Queen Mary University of London

Ke Han
University of Trento

Yukai Huang

Weitong Cai
Queen Mary University of London

Jifei Song
University of Surrey

Zhensong Zhang

Abstract

Ultra-long egocentric videos spanning multiple days present significant challenges for video understanding. Existing approaches still rely on fragmented local processing and limited temporal modeling, restricting their ability to reason over such extended sequences. To address these limitations, we introduce EgoGraph, a training-free and dynamic knowledge-graph construction framework that explicitly encodes long-term, cross-entity dependencies in egocentric video streams. EgoGraph employs a novel egocentric schema that unifies the extraction and abstraction of core entities, such as people, objects, locations, and events, and structurally reasons about their attributes and interactions, yielding a significantly richer and more coherent semantic representation than traditional clip-based video models. Crucially, we develop a temporal relational modeling strategy that captures temporal dependencies across entities and accumulates stable long-term memory over multiple days, enabling complex temporal reasoning. Extensive experiments on the EgoLifeQA and EgoR1-bench benchmarks demonstrate that EgoGraph achieves state-of-the-art performance on long-term video question answering, validating its effectiveness as a new paradigm for ultra-long egocentric video understanding.

1. Introduction

Daily activity logging is becoming increasingly essential for both humans and embodied agents, driven by the widespread adoption of wearable cameras in augmented reality devices and robotic platforms. These egocentric videos provide continuous first-person visual streams that naturally capture daily experiences, playing a crucial role in downstream tasks such as episodic memory retrieval and question answering [4, 5]. Despite significant progress in video understanding, existing methods primarily focus on short clips of less than one hour [8, 11], and consequently struggle to generalize to real-world long-term scenarios where relevant information is embedded within day-long contexts [22]. This challenge of ultra-long video understanding is particularly crucial for egocentric videos, where the core difficulty lies not merely in recognizing what happens, but in capturing when events occur and how they temporally relate across extended time spans.

However, research on this topic remains very limited, with [22] representing the pioneering attempt to tackle ultra-long egocentric video understanding. This work segments long videos into short clips and converts them into textual captions. The captions are then organized hierarchically through multi-level summarization (hourly, daily), leveraging LLMs' strong textual reasoning capabilities for retrieval. Despite its effectiveness, this paradigm exhibits two fundamental limitations. 1) It processes short video clips separately, overlooking inter-clip dependencies and long-range temporal dynamics. Consequently, semantically related events that occur across distant time spans are fragmented into disconnected textual pieces, making it difficult to reason about temporal relationships among people, actions, and events. 2) Enormous yet fragmented captions are continuously produced, resulting in a large-scale but unstructured information space that potentially limits the model's scalability and retrieval efficiency. These limitations motivate an alternative representation: rather than hierarchically aggregating information, a structured knowledge graph can explicitly preserve entity relationships and temporal dependencies. As illustrated in Figure 1, graph-based representation maintains fine-grained connections that are lost in hierarchical summarization.

Building on this insight, we introduce EgoGraph, a structured temporal knowledge graph that dynamically stores and evolves over time to represent condensed egocentric information for long-term video understanding. Our key insight is to frame a human-like structured memory based

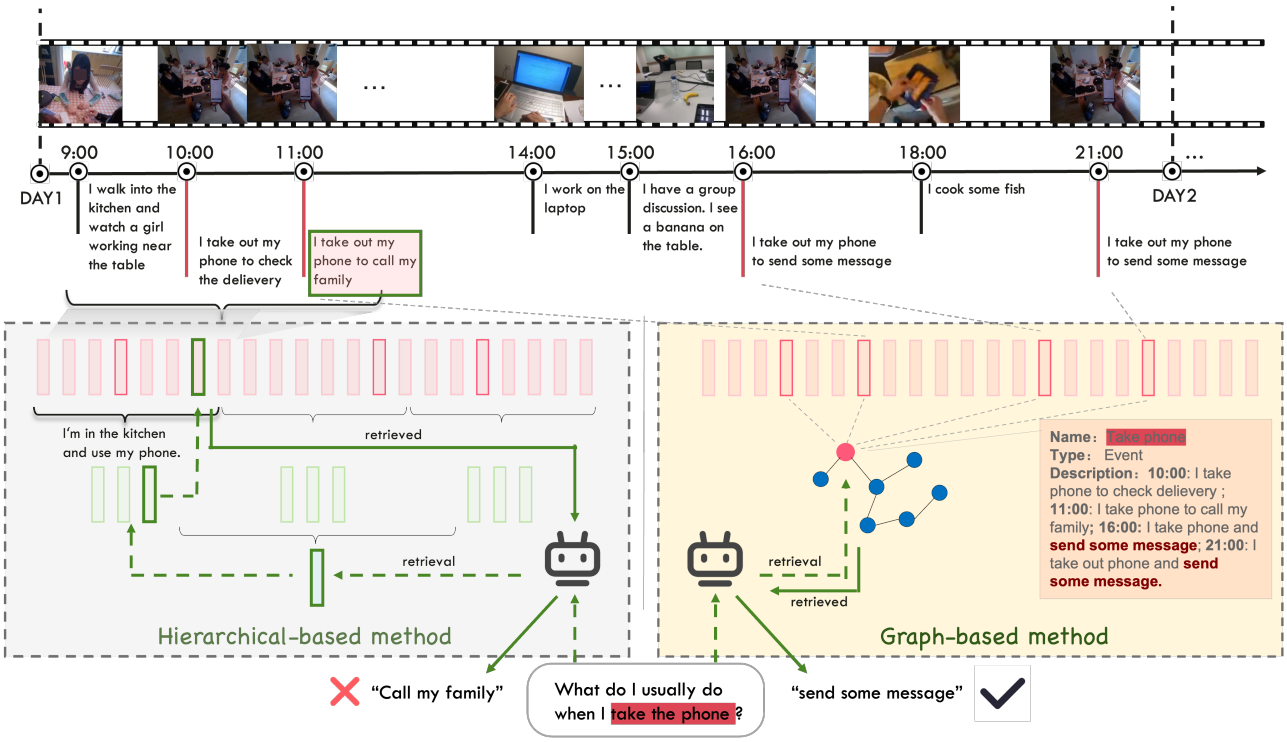

Figure 1. Comparison between hierarchical and graph-based methods for ultra-long egocentric video understanding. The hierarchical method summarizes each clip-level segment independently, whereas our graph-based approach constructs an entity-centric memory that models long-term dependencies across temporally separated events.

on events, which analyzes inter-entity relationships at specific moments and explicitly connects temporal relationships of entities across time through consistent inference over video streams. Unlike existing static graphs in other domains (e.g., knowledge inference, course understanding [2, 15, 16]), which struggle to capture the dynamic, person- and event-driven nature of daily life, EgoGraph is specifically designed for egocentric videos that involve long-term, multi-person interactions from a first-person viewpoint.

To construct structured graphs, we design an egocentric schema that defines node types for key semantic elements in egocentric videos, such as persons, events, objects, and their associated attributes, analogous to maintaining condensed personal profiles and event records. This schema enables a comprehensive yet informative representation of egocentric video content. More importantly, EgoGraph consistently models temporal connections across multiple levels, including nodes, edges, and graph chunks, and performs reasoning over the graph, such as inferring a person's habits from multiple observations, tracking changes in an object's location. It further forms relational hubs that connect participating entities with their temporal context while maintaining the graph compact through node incorporation. This design also allows temporally constrained queries to retrieve relevant subgraphs directly, rather than scanning the entire video history, thereby enhancing retrieval efficiency. Consequently, video contexts are incrementally stored and updated within a structured and compact graph, substantially improving both the efficiency and accuracy of video understanding.

Our main contributions are as follows: (1) We introduce EgoGraph, a training-free temporal knowledge graph framework for ultra-long egocentric video understanding, effectively overcoming the fragmented processing and limited temporal modeling of existing approaches. (2) We propose an egocentric schema that constructs structured entities and cross-entity relationships, together with a temporal relational modeling strategy that captures long-range dependencies across days, enabling efficient and coherent long-term reasoning. (3) We conduct extensive experiments on ultra-long egocentric benchmarks EgoLife and EgoR1-bench, demonstrating that EgoGraph achieves state-of-the-art performance on video question answering, significantly outperforming existing models.

2. Related Work

Egocentric Video Understanding. Egocentric videos captured from wearable cameras record the first-person experiences of daily activities, presenting unique challenges for video understanding due to their long-form nature, frequent viewpoint changes, and dense object interactions [5]. The

Ego4D benchmark [5] established large-scale tasks for egocentric video analysis, with episodic memory retrieval as a core challenge—requiring systems to answer queries such as "Where did I last see X?" by localizing relevant temporal windows in past video streams [4]. Recent advances leverage multi-modal large language models (MLLMs) for egocentric question answering [21, 23], demonstrating the potential of vision-language models for this domain. However, these approaches face fundamental limitations when processing long videos: EgoTempo [14] reveals that current models often rely on static frames or commonsense reasoning rather than temporal dynamics, while QaEgo4D [1] highlights the difficulty of maintaining constant-sized memory representations for extended video sequences. Action Scene Graphs [16] introduce graph-based representations to capture evolving object relationships, yet remain limited to static structural modeling without explicit temporal attributes. These challenges motivate structured representations that can simultaneously capture relational knowledge and temporal dynamics inherent to egocentric video understanding.

Graph-based Visual Understanding. Recent methods introduce structured knowledge graphs to capture relational information in videos. Scene graphs [9] represent objects as nodes and their spatial or semantic relationships as edges, enabling explicit reasoning over visual content. Early work applied scene graphs to video captioning [13] and visual question answering [18], demonstrating improved performance through structured relational modeling. Dynamic scene graph generation methods [3, 12] extend this paradigm to videos by constructing frame-level graphs and modeling temporal evolution of relationships across adjacent frames. Recent advances integrate knowledge graphs with video retrieval: VideoRAG [15] constructs static knowledge graphs for lecture videos, while GraphVideoAgent [2] tracks object appearance status to build temporal graphs. However, these approaches face critical limitations for egocentric video understanding. VideoRAG's static graphs cannot capture the dynamic, event-driven nature of first-person experiences, while GraphVideoAgent's object-centric tracking lacks global context to inter-connect entities through shared events. More fundamentally, existing methods treat temporal information either implicitly through sequential construction or as localized object trajectories, rather than embedding time as first-class graph attributes with filtering capabilities. This gap becomes critical for episodic memory queries that demand knowing not just what happened, but precisely when events occurred and how they temporally relate.

3. Methodology

Current methods for egocentric video understanding suffer from fragmented information processing and insufficient temporal modeling in ultra-long videos. To address these limitations, we propose EgoGraph, a structured temporal knowledge graph that explicitly encodes long-term dependencies among entities, to improve memory efficiency and video understanding accuracy. We first formalize EgoGraph as a temporal knowledge graph and present its construction pipeline from egocentric videos (Section 3.1). Then, we introduce our question answering framework that leverages EgoGraph for temporal reasoning and answer generation (Section 3.2).

3.1. Temporal-aware EgoGraph

To enable temporal reasoning over long-form egocentric videos, we propose EgoGraph, a temporal knowledge graph that captures entities, relations, and their temporal dynamics from first-person observations. This section presents the formal definition and construction pipeline of EgoGraph. We define the temporal-aware EgoGraph as:

$$\mathcal{G} = (V, E), \tag{1}$$

where each entity $v \in V$ is represented as:

$$v = (\sigma, n, a, \mathcal{T}_v, \mathcal{D}_v), \tag{2}$$

where σ donates entity type; n is entity name; a represent the predefined attributes for the entity type σ; $\mathcal{T}_v, \mathcal{D}_v$ are a set of timestamps and descriptions, respectively. Each edge $e \in E$, represented as

$$e = (v_s, v_t, \mathcal{T}_e, \mathcal{D}_e), \tag{3}$$

connects entities v_s and v_t through relation description $d \in \mathcal{D}$ at timestamps $t \in \mathcal{T}$.

Given an egocentric video $\mathcal{V}$ spanning multiple days, our goal is to construct a temporal knowledge graph $\mathcal{G}$. For a query q at timestamp t_q, we retrieve answers from the temporally-filtered graph:

$$\mathcal{G}^{\leq t_q} = \{(v, e) \in \mathcal{G} : \forall t \in \mathcal{T}_v \cup \mathcal{T}_e, t \leq t_q\}. \tag{4}$$

3.1.1. EgoGraph Construction

For temporal modeling in graph construction, we begin by establishing temporal grounding for video content descriptions $\mathcal{S}$, which include dense captions and video transcripts, following prior works [22]. To process the potentially unbounded length of descriptions, we partition descriptions into chunks $\{c_1, c_2, \ldots, c_M\}$ subject to a maximum token constraint L_{max}. Critically, each chunk is labeled with a timestamp t_i, which indicates its temporal anchor from the earliest timestamp within its span. This anchoring strategy ensures that extracted entities are attributed to their first observed occurrence, maintaining temporal causality in downstream graph construction.

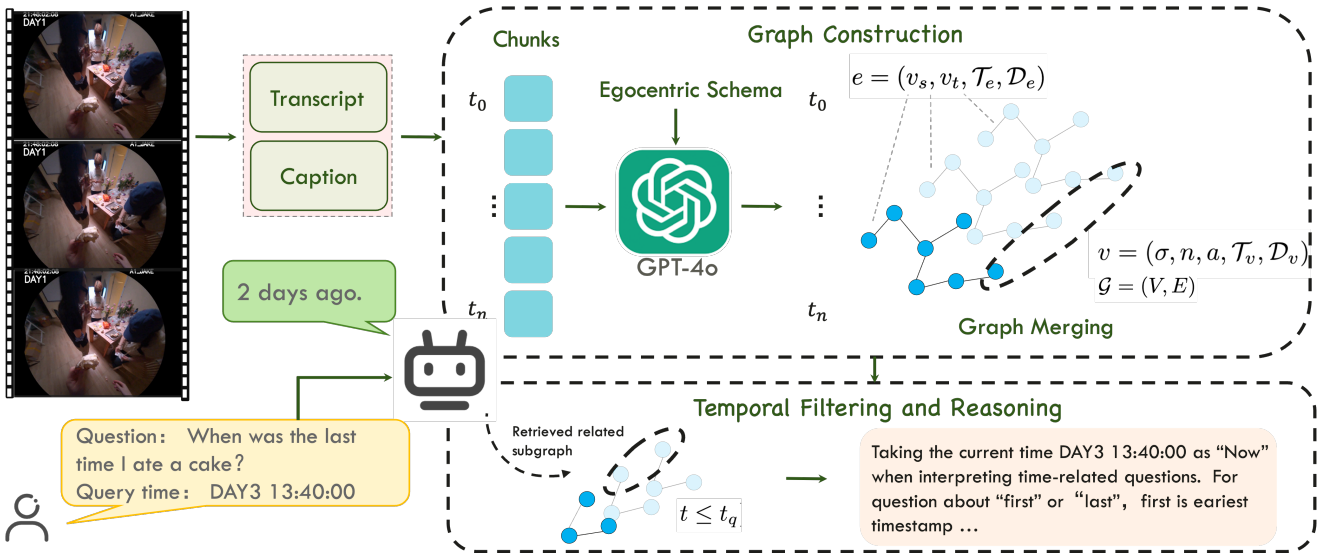

Figure 2. The pipeline of our proposed EgoGraph. EgoGraph encode ultra-long video into temporal-aware event knowledge graph, which imitate human brain memory processing.

For each temporally anchored chunk (c_i, t_i), we leverage large language models (LLMs) to extract entities and relations for graph construction. Prior graph-based methods often impose few constraints on entity types, which can result in trivial or redundant semantics in egocentric scenarios, thereby reducing graph effectiveness. To address this, we propose an egocentric schema that guides the extraction of nontrivial semantics and enforces semantic consistency across the entire graph.

Egocentric Schema. Personal episodic memory typically involves the who, where, and what of past experiences, e.g., the people encountered, locations visited, and contexts of events. Inspired by how the human brain organizes memory, our proposed egocentric schema captures the essential entity types and their attributes in egocentric scenarios. We define four core entity types, i.e., `Person`, `Location`, `Object`, and `Event`, each associated with specific attributes, as summarized in Table 1. This attribute-enriched representation ensures that the resulting graph is not merely a collection of mentions but a semantically rich and structurally organized knowledge abstraction.

Moreover, this schema ensures semantic consistency through well-defined entity types that enable type-aware reasoning, captures the subject–object–location–action interaction patterns characteristic of egocentric videos, and maintains scalability by preventing unbounded growth of entity types in long-term video streams.

Temporal Modeling and Reasoning. To enable temporal modeling over ultra-long videos, we associate each entity and relation with a list of timestamps indicating when they were observed. Each timestamp records both the day and the time of day, allowing the system to track entity evolution and retrieve historical context at any point along the video timeline. As formulated in Eq. (2) and (3), $\mathcal{T}_v$ or $\mathcal{T}_e = t_1, t_2, \ldots, t_n$ denotes an ordered list of timestamps, each following the format `[DAY`d `HH:MM:SS]`, where d represents the recording day and `HH:MM:SS` specifies the time within that day. Whenever a new timestamp t_i is added to the graph, new temporal connections can be established, triggering temporal reasoning in which the LLM infers higher-level conclusions such as personal habits, interpersonal relationships, and person–object interactions. For instance, if the nodes "Jack" and "playing the piano" are repeatedly connected across multiple days, the model can infer that Jack enjoys playing the piano.

For timestamped queries, our method can use the query timestamp as a temporal reference point, which not only localizes the relevant temporal context but also filters out

Table 1. Entity Type Definitions with Attributes

Type	Attributes	Example
PERSON	name, gender, appearance, preferences, dislikes, habits, hometown	`Person("John", gender="male", hometown="London")`
LOCATION	name, description	`Location("kitchen", desc="office kitchen area")`
OBJECT	name, type, color, size, condition, owner, purchase_information	`Object("coffee_mug", color="yellow", owner="John")`
EVENT	name, description, start_time, subject, object, location	`Event("meeting", loc="kitchen", subject="John")`

future time steps, thereby improving reasoning efficiency.

EgoGraph Merging and Update. A fundamental challenge in graph construction lies in the rapid accumulation of redundant nodes over time, which expands the graph scale while diminishing its efficiency. Unlike existing methods that either discard temporal information through aggregation or create duplicate nodes for each observation [citation], our temporal graph: 1) continuously merges redundant nodes based on their textual embedding similarity, and 2) appends new temporal information ($\mathcal{T}$ and $\mathcal{D}$) to preserve the complete evolutionary trajectory, where

$$\mathcal{T} = \{t_1, t_2, \ldots, t_n\}, \quad \mathcal{D} = \{d_1, d_2, \ldots, d_n\}, \tag{5}$$

and each timestamp t_i is paired with a corresponding description d_i that records new information about node states or edge connections. For example, attributes gathered from the same entity across different time points are consolidated into a single unified node, where the most recent non-empty attribute values are updated and existing values are retained. This design enables the graph to dynamically capture long-term dependencies among entities and relations in a scalable and semantically consistent way, rather than expanding the graph scale through redundant node accumulation.

3.2. Question Answering with EgoGraph

Question answering represents an important application of video understanding. Given a question q with an associated timestamp t_q about the egocentric video, and following the retrieval-augmented generation process [6], we leverage the constructed graph to retrieve relevant information for answering the question. Specifically, we extract semantic keywords from the query q and perform vector similarity search over the entity embeddings to retrieve the top-k most relevant entities. In parallel, we retrieve the top-k related edges and combine their contextual information to form a comprehensive reasoning context.

Temporal Filtering. Since questions can be raised at any point in the video stream, we further define a temporal filtering operation to improve answering efficiency by extracting a subgraph that includes only temporally aligned information:

$$\mathcal{G}(t \leq t_q) = (V^{\leq t_q}, E^{\leq t_q}), \tag{6}$$

where $V^{\leq t_q}$ and $E^{\leq t_q}$ denote the nodes and edges that satisfy the temporal constraint $t \leq t_q$. This operation filters entities and their associated information based on the query timestamp, ensuring temporally coherent reasoning.

LLM-based Temporal Reasoning. Given retrieved entities and relations with their timestamps, we leverage the temporal reasoning capabilities of LLMs through structured prompting. The LLM performs temporal reasoning by comparing timestamps against the query reference point and applying the provided temporal rules. It receives the retrieved context along with explicit temporal reasoning instructions that define how to interpret time-related expressions relative to query timestamp t_q. The prompt specifies that all timestamps follow format `[DAYd HH:MM:SS]` where d indicates the date, with total ordering by day then time-of-day, and establishes t_q as the reference point ("NOW") for all temporal interpretation. Relative time expressions are resolved through rules provided in the prompt: "yesterday" maps to day$(t_q) - 1$, "last time" selects $\max\{t \in \tau \mid t < t_q\}$, "first time" selects $\min\{t \in \tau\}$, and expressions like "2 hours ago" compute $t_q - 2$h. The prompt explicitly requires timestamp citations for all referenced events, enabling answer verification and temporal grounding. This approach handles diverse temporal expressions in natural language without exhaustive rule enumeration, leveraging the LLM's language understanding while maintaining temporal accuracy through explicit instruction and structured output requirements.

4. Experiments

4.1. Datasets and Evaluation Metrics

We adopt zero-shot evaluation to test on the following two datasets: **EgoLifeQA** [22] constructed a super-long egocentric video dataset in which six individuals wore camera-equipped glasses to capture their daily experiences over seven days in a shared house, recording their daily activities such as cooking, shopping, and socializing. It publishes 500 public question-answering pairs for evaluating ultra-long video question answering. **EgoR1-Bench** [19] is a reasoning based benchmark for ultra-long egocentric video understanding, which comprises 300 QAs evenly distributed across six first-person perspectives.

4.2. Implementation Details

We leverage `InternVL-3.5-8B-Instruct` [20] as the visual captioner to extract visual information from videos. The long videos are segmented into two-minute clips, with each clip processed alongside its corresponding transcript as input to the captioner. For graph construction, we employ `gpt-4o` to perform entity and relation extraction within our designed constraints, and use the same model for question answering. To retrieve relevant nodes and edges, we implement `text-embedding-3-small` from OpenAI as the text encoder and calculate cosine similarity to identify the top-k nodes and edges, where $k = 40$. Each entity is stored with its description at each timestamp, which are automatically summarized when the number of timestamps exceeds 100 to maintain conciseness.

4.3. Comparison with State-of-the-Arts

Baselines. We compare EgoGraph with state-of-the-art methods for ultra-long video understanding:

Table 2. Performance comparison of EgoGraph with state-of-the-art models on EgoLifeQA and EgoR1-bench benchmarks.

Model	EgoLifeQA						EgoR1-Bench
	EntityLog	EventRecall	HabitInsight	RelationMap	TaskMaster	Average	
MLLMs							
Gemini-1.5-Pro [17]	36.0	37.3	**45.9**	30.4	34.9	36.9	38.3
GPT-4o [7]	34.4	42.1	29.5	30.4	44.4	36.2	-
LLaVA-OV[10]	36.8	34.9	31.1	22.4	28.6	30.8	31.6
EgoGPT [22] (EgoIT-99K+D1)	39.2	36.5	31.1	33.6	39.7	36.0	-
EgoGPT [22] (InternVL-3.5-8B)	28.0	31.7	31.1	34.4	38.1	32.2	31.3
Graph-based methods							
LightRAG [6]	40.8	40.4	36.0	32.0	50.7	39.2	31.1
VideoRAG [15]	19.2	15.9	19.7	28.0	14.3	20.0	19.8
EgoGraph	**46.4**	**46.8**	**45.9**	**35.2**	**60.3**	**45.8**	**41.3**

(1) **EgoGPT** [22] constructs hierarchical memory banks over long-term egocentric videos and leverages retrieval-augmented generation for question answering. It generates summaries based on video captions at both hour-level and day-level granularities, and retrieves temporal information through a top-down process from day to specific hours. (2) **VideoRAG** [15] integrates static knowledge graphs with videos. It is designed for lecture videos, where visual variation is significantly less pronounced compared to egocentric scenarios, in which head movements lead to substantial viewpoint changes. (3) **LightRAG** [6] is a text-based method from NLP that first segments long contexts into chunks and then extracts nodes and edges from each chunk. This graph-based retrieval-augmented generation approach is particularly suitable for query-focused summarization [24]. Our scenario shares a similar philosophy, as we track the same entities appearing across temporally separated segments of ultra-long video contexts. Note that all existing methods lack explicit temporal modeling and are not suitable for ultra-long egocentric video scenarios, which EgoGraph addresses through temporal knowledge graph sequences with explicit temporal reasoning. For fair comparison, we employ InternVL-3.5-8B as the captioner for all methods unless otherwise specified. We also implement the same question answering model, `gpt-4o`, across all compared methods.

Results. We provide a comprehensive evaluation of EgoGraph against state-of-the-art (SOTA) methods on the EgoLifeQA and EgoR1-Bench benchmarks, with detailed results presented in Table 2. Our competitors include leading MLLMs (e.g., Gemini-1.5-Pro, GPT-4o) and prior graph-based approaches (e.g., LightRAG, VideoRAG). On the primary EgoLifeQA benchmark, EgoGraph achieves a new SOTA average accuracy of 45.8%, significantly surpassing all other methods. This represents a substantial lead of +6.6 points over the best-performing graph-based competitor, LightRAG (39.2%), and +8.9 points over the strongest MLLM, Gemini-1.5-Pro (36.9%). This superiority is consistent across all five sub-categories: EgoGraph achieves the highest or tied-for-highest score in every single task. Notably, it establishes a commanding lead in complex reasoning tasks such as TaskMaster (60.3%) and EventRecall (46.8%), while also matching the powerful Gemini-1.5-Pro in HabitInsight (45.9%).

Furthermore, we validate our model's robustness on the EgoR1-Bench. EgoGraph again achieves the top score of 41.3%, outperforming the formidable Gemini-1.5-Pro (38.3%) by a clear margin of +3.0 points. While powerful MLLMs like Gemini-1.5-Pro are strong generalists, our results demonstrate their limitations in specialized, long-term egocentric reasoning compared to a dedicated, structured approach. The poor performance of VideoRAG (20.0%) also suggests that naive graph construction is insufficient for this challenge. In contrast, EgoGraph's consistent top-tier performance across both benchmarks unequivocally establishes its superior capability to structure, retrieve, and reason over complex, long-duration egocentric video.

4.4. Ablation Study

4.4.1. Hierarchical vs Graph

We compare hierarchical-based methods with our graph-based approach on a series of subsets from the EgoLife dataset. We extract three subsets from the EgoLifeQA dataset based on temporal characteristics: (1) temporal aggregation, which includes questions containing "usually"; (2) temporal dependency, which includes questions containing "after"; and (3) entity tracking, which includes questions containing "where". We focus on these question types because they involve entities that are temporally separated across long contexts. As shown in Figure 3, EgoGraph substantially outperforms EgoGPT across all three categories, achieving an average improvement of 29.3%. This strongly supports our claim that graph-based methods with instance-centric representations, compared to hierarchical methods with continuous temporal information, can better capture long-term dependencies.

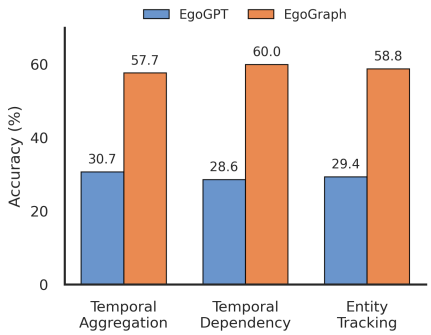

Figure 3. Quantitative comparison on temporal reasoning tasks. EgoGraph outperforms EgoGPT across temporal aggregation, temporal dependency, and entity tracking, achieving an average improvement of 29.3%.

Table 3. Evaluating component effectiveness on EgoLifeQA.

Model	Overall (%)
Baseline	39.2
– w/ Egocentric Schema	41.4
– w/ Time Filter	43.0
– w/ Temporal Reasoning **EgoGraph (Full)**	45.8

4.4.2. Vanilla Graph vs. EgoGraph

Recognizing the advantages of graph-based methods over hierarchical methods, we further explore the essential constraints needed to adapt graphs to egocentric scenario. As shown in Table 3, we conduct incremental ablations on techniques in EgoGraph. Starting from LightRAG [6], a vanilla graph-based method that treats egocentric video captions and transcripts as static knowledge graphs, we progressively introduce our temporal modeling components. We adopt LightRAG [6], a vanilla graph-based method that directly applies static knowledge graph construction to egocentric video captions and transcripts, as our baseline. We progressively add our proposed components to this baseline. The baseline achieves only 39.2% accuracy, revealing a fundamental limitation: static graph representations cannot directly transfer to temporal egocentric video understanding. This gap stems from the inherent mismatch between static knowledge graphs and the temporal nature of first-person experiences—without temporal modeling, the system cannot distinguish between events that occurred at different times or enforce causal consistency in reasoning. Adding the Egocentric Schema (+w/ Egocentric Schema) brings a modest improvement to 41.4% by introducing structured entity types (Person, Object, Location, Event) and an accumulated attribute profile for each entity. To further modeling temporal causality, we incorporate timestamp information in graph constuction stage and temporal filtering and reasoning in question answering stage. Temporal filtering eables the system to construct causally-consistent subgraphs by retrieving only information from before the query timestamp. This mechanism is essential for avoiding temporal leakage and mirrors how human episodic memory naturally constrains retrieval to past experiences. The full EgoGraph model, which integrates temporal reasoning capabilities on top of the structured schema, achieves 45.8% accuracy. The cumulative improvement of 6.6% over the baseline validates our statement: temporal awareness is not an optional enhancement but a fundamental requirement for graph-based egocentric video understanding.

4.4.3. Analysis on Temporal Robustness

To evaluate the robustness of EgoGraph for long-term video understanding, we conduct two experiments analyzing its robustness to temporal dynamics. We compare EgoGraph against the baseline model (EgoGPT) and Plain-text, where GPT-4o retrieves information directly from captions, with results presented in Figure 5. To explore model scalability as total video context grows, we design experiments measuring cumulative query time versus accuracy. As shown in Figure 5 on the top, longer query times correspond to longer contexts, with values representing average accuracy up to the positioned query time. As cumulative duration increases from 1 to 7 days, Plain-text performance drops dramatically from 43.1% to 8.8%, as the long context exceeds the maximum token limit and fails to respond after the first day. EgoGPT accuracy remains stagnant at approximately 30-31%. In contrast, EgoGraph achieves 51% accuracy on first-day data alone. When handling the full 7-day context (500 questions), its performance demonstrates remarkable stability, decreasing only slightly to 45.80%. This shows that EgoGraph effectively indexes and retrieves from growing contexts, consistently maintaining a significant 15% performance margin over EgoGPT and 29% over Plain-text, demonstrating graceful scaling without being overwhelmed by increasing information volume. Second, we analyze a more challenging scenario: the model's robustness to the temporal gap between the query time (t_q) and the target event time (t_{target}). Longer temporal gaps impose higher retrieval demands. A robust model should not degrade significantly when queries refer to distant past events. Figure 5 on the bottom demonstrates our superiority over both comparison methods.

4.4.4. Analysis on Graph Retrieval Components

To identify the key components in our knowledge graph, we conduct a detailed ablation study on the retrieval strategy. After constructing the knowledge graph, we freeze its structure and systematically evaluate different retrieval components. For the complete EgoGraph, we retrieve information through all components including nodes, edges, and plain text chunks. As shown in Table 4, when retrieving with a single component, node-based retrieval achieves the best result of 40.8%, outperforming plain text chunks (39.6%) and edges (35.6%). This is expected as nodes store

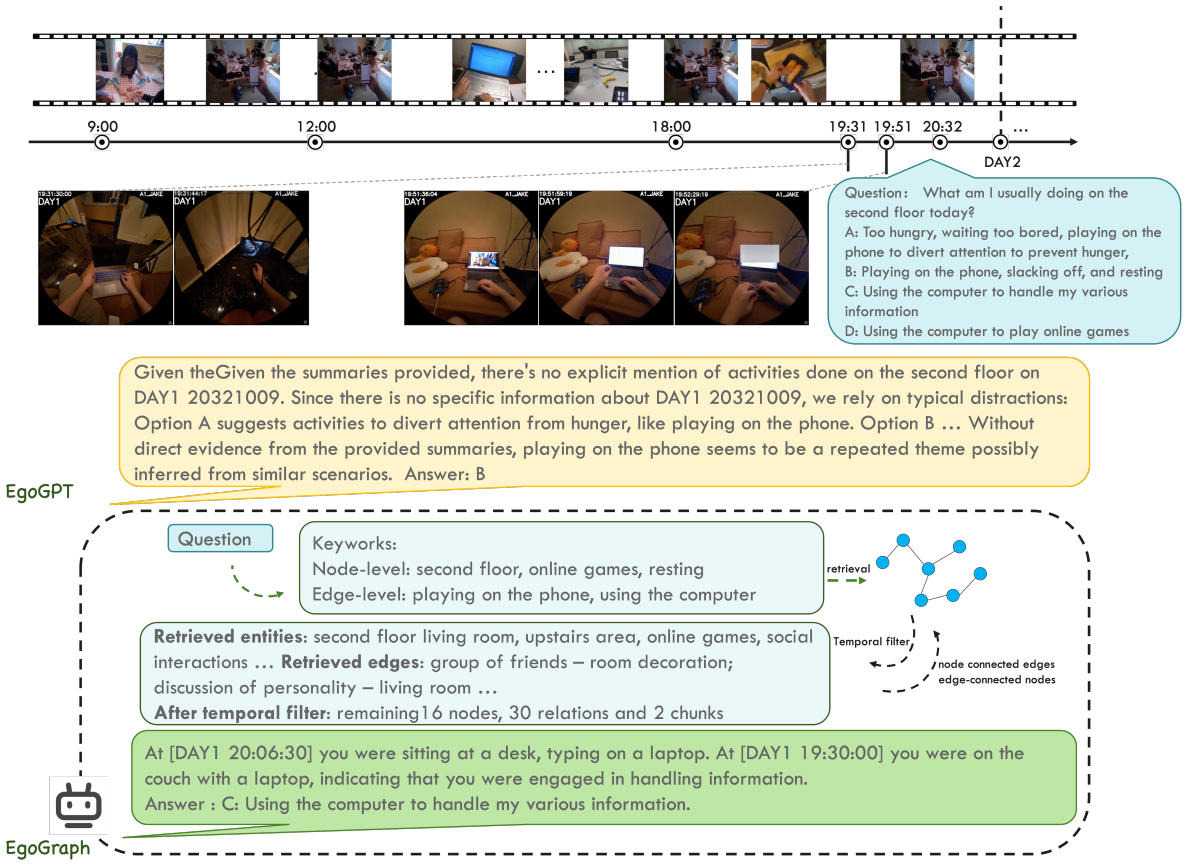

Figure 4. Qualitative comparison of question answering on multi-day egocentric videos. EgoGPT lacks explicit temporal grounding and incorrectly infers activities. EgoGraph retrieves temporally-filtered knowledge graph relations with specific timestamps, enabling accurate answers about daily routines.

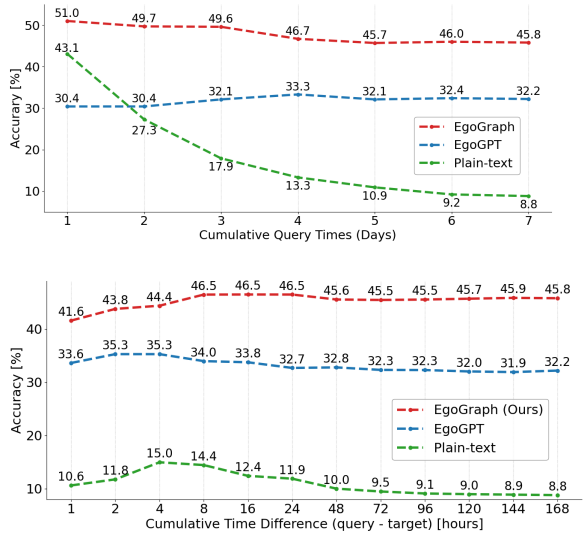

Figure 5. Robustness analysis of EgoGraph on long-term video understanding. (Top) Scalability with growing context. and (Bottom) Temporal gap sensitivity.

the primary contextual information, including accumulated profiles for each event, object, person, and location. Combining both nodes and edges yields a notable improvement of 2.8% over chunk-based retrieval, as they provide complementary information to each other, achieving 42.4% accuracy. Finally, the full implementation of EgoGraph leverages all three components, with their complementary information leading to the best performance of 45.8%.

Table 4. Evaluation on retrieval components. We evaluate different combinations of nodes, edges, and text chunks for knowledge graph retrieval on a frozen graph structure.

Node	Edge	Chunk	Accuracy
		✓	39.6
✓			40.8
	✓		35.6
✓	✓		42.4
✓	✓	✓	45.8

5. Conclusion

In this work, we propose EgoGraph, a novel temporal knowledge graph framework for ultra-long egocentric video understanding. The key insight is to establish long-term dependencies across temporally-separated video segments through structured entities and relationships. First, we introduce an egocentric schema that defines entity types and

their corresponding attributes, enabling the cumulative construction of a structurally organized and semantically consistent long-term memory. Second, we propose a temporal relational modeling strategy that tracks entity evolution and retrieves historical context at any point along the video timeline. Third, we demonstrate the effectiveness of our framework on video question answering through temporal filtering and reasoning processes, enabling accurate and efficient temporal retrieval. Overall, extensive experiments on the EgoLifeQA and EgoR1-bench dataset demonstrate that EgoGraph significantly outperforms existing methods in capturing long-term dependencies for ultra-long egocentric video understanding.

References

[1] Leonard Bärmann and Alex Waibel. Where did i leave my keys? episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 3

[2] Meng Chu, Yicong Li, and Tat-Seng Chua. Understanding long videos via llm-powered entity relation graphs. *arXiv preprint arXiv:2501.15953*, 2025. 2, 3

[3] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16372–16382, 2021. 3

[4] Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv Batra, and Devi Parikh. Episodic memory question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19119–19128, 2022. 1, 3

[5] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 1, 2, 3

[6] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. 2024. 5, 6, 7

[7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6

[8] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 1

[9] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678, 2015. 3

[10] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6

[11] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 1

[12] Sayak Nag, Xinhang Zhu, Yi-Zhe Song, and Tao Xiang. Unbiased scene graph generation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22803–22813, 2023. 3

[13] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10870–10879, 2020. 3

[14] Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24129–24138, 2025. 3

[15] Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*, 2025. 2, 3, 6

[16] Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. Action scene graphs for long-form understanding of egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18622–18632, 2024. 2, 3

[17] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6

[18] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2017. 3

[19] 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. 5

[20] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 5

[21] Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. *arXiv preprint arXiv:2312.05269*, 2023. 3

[22] 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 *CVPR*, pages 28885–28900, 2025. 1, 3, 5, 6

[23] Haoyu Zhang, Meng Liu, Zixin Liu, Xuemeng Song, Yaowei Wang, and Liqiang Nie. Multi-factor adaptive vision selection for egocentric video question answering. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 59310–59328. PMLR, 2024. 3

[24] Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*, 2024. 6