R4DSG: Relative 4D Scene Graph Memory for Object-Centric Question Answering in Long Egocentric Video

Ke Ma
College of Design and Innovation
Tongji University
Shanghai, China
make@hust.edu.cn

Yamin Mao
Samsung R&D Institute China - Beijing
Beijing, China
yamin18.mao@samsung.com

Weiming Li
Samsung R&D Institute China - Beijing
Beijing, China
weiming.li@samsung.com

Shuai Tan
Shanghai Jiao Tong University
Shanghai, China
tanshuai0219@sjtu.edu.cn

Yijie Zhong
College of Design and Innovation
Tongji University
Shanghai, China
dun.haski@gmail.com

Hao Chen
Samsung Networks
Samsung Research America
Plano, Texas, United States
hao.chen1@samsung.com

Haofen Wang
College of Design and Innovation
Tongji University
Shanghai, China
carter.whfcarter@gmail.com

Meng Wang✉
College of Design and Innovation
Tongji University
Shanghai, China
Shanghai Research Institute for Intelligent Autonomous Systems
Tongji University
Shanghai, China
mengwangtj@tongji.edu.cn

Abstract

Long-horizon egocentric video is a rich substrate for wearable AI assistants, but object-centric questions such as where an item was moved, when it last changed state, or why it was relocated remain difficult because caption- and transcript-based memories rarely preserve persistent object identity or structured spatial change. Existing long-video QA methods mainly emphasize temporal grounding and clip retrieval, while prior 3D scene-graph methods typically assume stronger geometry than free-motion wearable RGB video provides, including point clouds, RGB-D input, posed views, sparse reconstruction, or reconstructed scenes. R4DSG introduces a **relative 4D scene graph memory** for long egocentric video. Instead of storing raw graph sequences, R4DSG converts video into compact queryable memory entries indexed by time, place, persistent objects, anchor-relative change, and local interaction context. The main idea is to separate stable anchors from dynamic objects, maintain persistent object identity across frames, and represent object state through anchor-relative transitions rather than a globally aligned world model. Built on recent RGB-only advances in promptable video segmentation, temporal propagation, and relative 3D lifting, the method produces a retrieval-ready memory directly usable for long-horizon question answering. Evaluation on a 255-question object-related subset from EgoLifeQA shows, under question-only retrieval, a 6.7-point overall gain over EgoRAG-Text and a 12.5-point gain on when questions, which highlights the value of temporally organized object memory. These results position relative 4D scene graphs as a practical memory substrate for wearable assistants, AR systems, and embodied multimedia agents. GitHub Page: https://dualtransparency.github.io/R4DSG/.

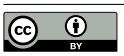
CCS Concepts

• **Computing methodologies** → **Scene understanding**; *Computer vision representations*; • **Information systems** → *Multimedia information systems*; *Question answering*.

Keywords

egocentric video, 3D scene graph, temporal memory, graph retrieval, object-state reasoning, multimodal question answering

ACM Reference Format:
Ke Ma, Yamin Mao, Weiming Li, Shuai Tan, Yijie Zhong, Hao Chen, Haofen Wang, and Meng Wang. 2026. R4DSG: Relative 4D Scene Graph Memory for Object-Centric Question Answering in Long Egocentric Video. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835995



MM '26, Rio de Janeiro, Brazil.

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835995

1 Introduction

Wearable cameras and AI glasses are moving multimedia systems from passive recognition toward persistent assistance. In that setting, a useful assistant is not limited to describing the current view. It must also answer questions grounded in past interactions with objects and spaces, such as *Where did I leave the charger?*, *When*

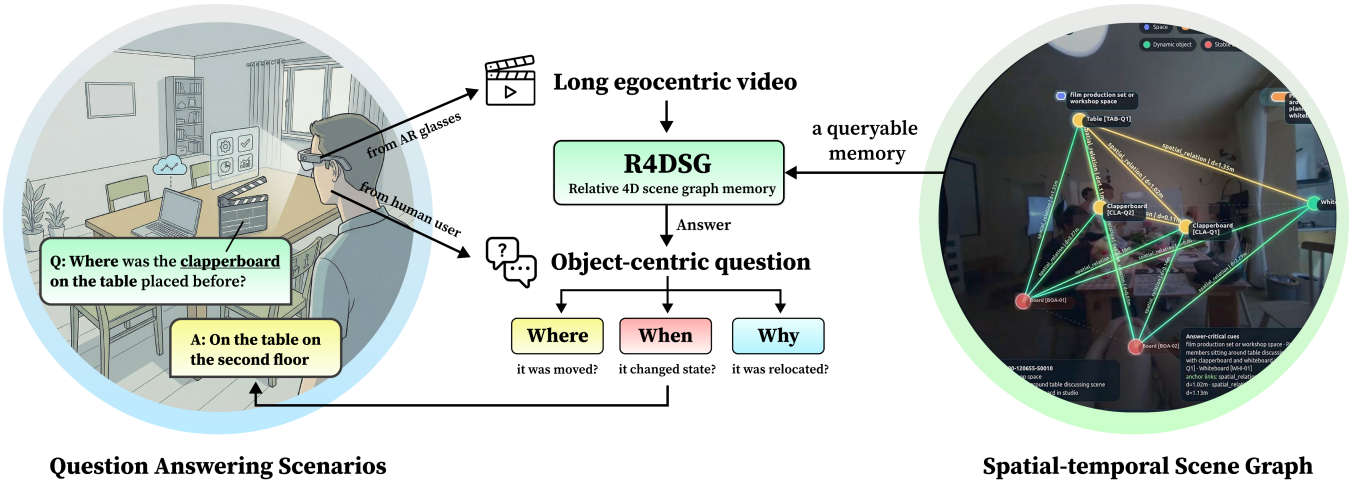


Figure 1: R4DSG maps long egocentric video to a relative scene-memory for object-centric question answering.

did I last move the mug from the desk?, or *Why did I take the bowl out of the fridge?* Recent work on egocentric devices and wearable assistants makes this trajectory increasingly explicit, spanning egocentric multimodal sensing platforms, context-aware AR assistants, proactive smart-glasses systems, and life-assistant benchmarks [1, 9, 16, 19, 20, 22, 26, 29, 33, 38]. Figure 1 shows R4DSG.

The central challenge is the need for a *queryable memory*. Object-centric questions depend on preserving object identity, the stable references surrounding the object, and the transitions that explain its current state. Flat summaries, captions, or retrieved clips are often enough for coarse event recall, but they are weak at preserving structured evidence for object-state change. A useful memory for egocentric assistance should therefore satisfy four properties. It should preserve persistent object identity across time, encode spatial relations rather than only local appearance, compress long streams into compact retrievable units, and expose those units in a form directly usable by downstream QA.

Recent progress in long egocentric video understanding sharpens this gap but does not close it. Grounded Multi-Hop VideoQA shows that many questions depend on multiple scattered evidential moments rather than one contiguous segment [3]. AMEGO argues that very long egocentric streams should be converted into an explicit reusable memory instead of being processed end to end each time [11]. EgoLife extends this agenda to a life-assistant setting and highlights memory retrieval, event recall, habit understanding, and relation reasoning as core capabilities for future egocentric assistants [38]. Yet the dominant abstractions remain caption-centric or summary-centric. They record events in language, but they do not maintain a persistent object-level account of an object's location, its relations to stable references, when those relations changed, and the local context that explains such changes.

A natural answer is to use scene graphs. At the representation level, scene graphs are appealing because they bind entities, relations, and evidence into compact structured units that can be updated, searched, and reused [6, 37, 43]. Prior work has shown that 3D scene graphs provide compact and semantically rich representations of objects and relations [35, 36, 40]. Open3DSG, VL-SAT, OpenFunGraph, and 3DGraphQA further demonstrate that open-vocabulary semantics, language guidance, and graph reasoning can make structured 3D representations more expressive downstream [15, 34, 37, 39]. These results are highly relevant, but they still assume stronger geometry than the target setting here provides, including point clouds, RGB-D input, posed views, sparse reconstruction, or reconstructed indoor scenes. In daily egocentric capture, the camera moves with the wearer, observations are partial, and a globally aligned world coordinate system is often unavailable, unreliable, or simply unnecessary [9, 10, 40].

Recent RGB-only foundation models for open-set video segmentation and tracking, temporal propagation, and monocular 3D lifting make this route technically plausible under weaker sensing conditions [2, 4, 14, 17, 27, 32]. In particular, SAM 3 strengthens promptable video segmentation with concept-aware temporal propagation, while SAM 3D, DUSt3R, and MASt3R provide RGB-only routes to relative 3D cues without requiring depth input or pre-existing point clouds. Together, these advances make it feasible to build a queryable memory that captures relative spatial change in long egocentric video. What matters instead is whether an object moved away from one stable reference and toward another, whether that transition persisted, when it occurred, and what local interaction context best explains it.

Motivated by this view, we introduce R4DSG, a **relative 4D scene graph memory**, which converts long egocentric video into queryable memory entries indexed by time, place, persistent objects, anchor-relative change, and local interaction context. The key insight in R4DSG is that long-horizon object memory can be organized around *static anchors* and *dynamic objects*. Static anchors such as tables, shelves, counters, drawers, sofas, or fridges provide stable relational references. Dynamic objects accumulate state changes by forming new spatial relations to different anchors over time. Instead of forcing all observations into one global frame, the method associates object instances across frames, retains only anchor-relative transitions that remain stable over time, and promotes those transitions into memory entries. The result is a memory that is compact enough for retrieval yet structured for object-centric reasoning.

The main contributions are as follows:

- A relative 4D scene-graph formulation for long egocentric RGB video that represents object state through persistent anchor-relative transitions rather than a globally aligned map.
- A queryable memory design that converts frame-level graph evidence into segment-level retrieval documents while preserving place, activity, object state, actor, interaction, and edge cues needed by long-horizon QA.
- A pipeline grounded in SAM3-style consistency and RGB-only 3D lifting, followed by persistent identity association, static-anchor inference, and retrieval-aware memory writing.

2 Related Work

2.1 Long egocentric video understanding, memory, and question answering

Egocentric video understanding has evolved from action recognition toward long-horizon memory and question answering. EMQA is an early formulation of episodic memory QA, where an egocentric agent answers grounded questions using an explicit scene memory [7]. EgoSchema later established very long-form multiple-choice QA as a diagnostic benchmark for temporal understanding in egocentric video [23]. Building on the Ego4D ecosystem [12], GroundVQA formalized grounded QA in long egocentric videos and coupled answer generation with temporal localization of relevant evidence [8]. Grounded Multi-Hop VideoQA further increased the reasoning burden by requiring multiple discontiguous evidence segments for a single question [3]. AMEGO proposed an active-memory representation for very long egocentric video that captures recurring locations and object interactions without repeatedly processing the full stream [11]. EgoLife broadened the task to an egocentric life assistant and introduced long-context tasks around event recall, relation reasoning, habit understanding, and memory-intensive assistance [38].

These works establish the importance of long-term memory in egocentric video, but their dominant abstractions are still clips, summaries, captions, or high-level semantic memories. They are effective for generic recall and temporal localization, yet they do not explicitly target persistent object-state memory grounded in relative spatial relations. This direction also aligns with graph-aware retrieval work, where G-Retriever, KG^{2}RAG, and GNN-RAG show that once knowledge is expressed as a graph, retrieval can operate over compact structured evidence rather than flat text chunks [13, 24, 44]. R4DSG focuses on the earlier multimedia problem: constructing such a queryable graph memory directly from long egocentric video.

2.2 3D scene graphs and open-vocabulary structured scene understanding

Scene graphs provide an explicit representation of entities and their relations, which makes them appealing for reasoning-heavy visual tasks. In 3D perception, 3DSSG introduced a benchmark and a learning framework for semantic scene graphs from indoor 3D reconstructions [31]. SceneGraphFusion then showed that 3D scene graphs can be predicted incrementally from RGB-D sequences, which enables online graph construction under partial observations [36]. Wu et al. later demonstrated that consistent 3D semantic scene graphs can also be incrementally constructed from RGB sequences by coupling sparse mapping with graph prediction [35]. VL-SAT injected visual-linguistic supervision into 3D semantic scene graph prediction from point clouds [34]. Open3DSG moved toward open-vocabulary 3D scene graphs from point clouds [15, 25]. EgoSG brought 3D scene graphs closer to the egocentric setting by learning from unposed RGB-D sequences without relying on reconstruction algorithms or camera poses [40, 41]. OpenFunGraph extended structured 3D perception to functional relationships and downstream interaction-oriented reasoning in real-world indoor scenes [21, 39]. In ACM MM, 3DGraphQA further demonstrated that explicit scene-graph reasoning can improve QA over 3D scenes by providing a more interpretable intermediate structure [37].

These methods strongly motivate graph-structured 3D reasoning, but they target settings that differ in a critical way. They typically depend on point clouds, RGB-D input, posed multi-view observations, sparse mapping, or reconstructed scenes, whereas the target input here is long free-motion monocular egocentric RGB video. Accordingly, the objective here is not globally consistent 3D scene reconstruction. The objective is a relative 3D abstraction sufficient for memory and reasoning, where pairwise distances, anchor-relative transitions, and persistent IDs matter more than metric global alignment.

2.3 Temporal scene graphs and graph-based video reasoning

A parallel line of work studies temporal or spatio-temporal graph representations for video reasoning. Cherian et al. introduced a (2.5 + 1)D spatio-temporal scene graph for video QA, combining pseudo-3D scene structure with temporal graph reasoning [5]. Urooj et al. proposed situation hyper-graphs for video QA, explicitly modeling actors, objects, and relations over time [30]. In ACM MM, HostSG constructed a holistic spatio-temporal scene graph for video semantic role labeling by merging clip-level dynamic scene graphs while preserving static and dynamic cues [43]. More recently, Action Scene Graphs introduced temporally evolving graphs tailored to long-form egocentric video [28, 42]. GraphVideoAgent showed that entity-relation graphs can also guide frame selection and reasoning in long-form video understanding [6]. At the zero-shot end, SAMJAM combined segmentation-and-tracking with VLM semantics for egocentric kitchen video scene graph generation, which demonstrates that foundation-model pipelines can produce temporally consistent graphs without task-specific training [18].

These works show that graph structure is an effective language for long-video reasoning, but they still leave open the problem addressed here. Existing temporal graph methods are often action-centric or event-centric, and even the most relevant zero-shot egocentric graph formulations focus on graph generation in narrower domains rather than on building a long-horizon memory whose atomic unit is an anchor-relative object change with explicit timestamps and retrievable contextual evidence.

3 Method

3.1 Task formulation

Given a long egocentric RGB video $X = \{I_t\}_{t=1}^{T}$, the goal is to build a queryable memory M that supports object-centric long-context QA under weak sensing conditions. The input is monocular, map-free, and only partially observed. No depth, point clouds, calibrated poses, or globally aligned world coordinates are assumed. Instead, the memory should preserve persistent object identity and object-state change over time.

The problem is written as

$$M = f(X), \qquad (\hat{y}, \hat{e}) = g(M, q), \tag{1}$$

where q is a question, $\hat{y}$ is the predicted answer, and $\hat{e}$ is the retrieved evidence. The focus is on object-centric questions whose answers depend on past state changes rather than on the current frame alone, especially `where`, `when`, and `why` questions.

Three requirements follow from this formulation. First, the same physical object should keep a stable identity across time. Second, object state should be described relative to stable references in the scene. Third, the video must be summarized into searchable segment-level units, because storing dense frame sequences would preserve redundancy rather than evidence and would make long-range retrieval expensive and brittle. Figure 2 summarizes the resulting end-to-end pipeline.

Figure 2 also clarifies that the method does not depend on a single heavy reconstruction stage. The semantic episode constructor narrows the local object inventory, the visual front-end extracts relative evidence within those windows, and the graph association stage decides persistence and anchor roles before memory writing. This staged decomposition is important for long egocentric video because errors remain local: a missed mask affects one episode, whereas failure of a global map would otherwise propagate across the whole day.

3.2 Relative 4D scene graph representation

At each sampled time step t, the method constructs a frame graph

$$G_t = (V_t, E_t), \tag{2}$$

where $V_t = \{v_t^{space}, v_t^{act}\} \cup V_t^{obj}$ contains one space node, one activity node, and a set of object nodes. Each object node stores semantic and geometric attributes,

$$a_t(v) = [c_t, m_t, b_t^{3d}, x_t^{3d}, s_t], \tag{3}$$

where c_t is the semantic class, m_t is the mask or crop reference inherited from the segmentation stage, b_t^{3d} is a coarse 3D extent, x_t^{3d} is a relative 3D location, and s_t is a textual state descriptor. Edges encode relations such as `in`, `on`, `near`, and generic `spatial_relation`, together with activity-linked interaction cues when available.

The representation becomes 4D after linking frame nodes into persistent tracks. Here, "4D" denotes a time-indexed relative 3D scene-graph memory, not a globally consistent 4D reconstruction. Let p denote a persistent object track formed by associating semantically compatible observations across neighboring graphs. For each track, the method identifies whether it behaves as a *static anchor* or a *dynamic object*. Static anchors are objects whose semantic role and motion pattern are stable, such as tables, shelves, counters, or fridges. Dynamic objects are portable or manipulated items such as bags, cups, phones, tools, or food packages.

Table 1: Schema of R4DSG and its retrieval memory.

Component	Stored fields and role
Frame node	Semantic class, mask or crop reference, relative 3D box, relative 3D location, textual state, and place association. Used to build G_t.
Frame edge	Subject, predicate, object, optional score, and relative spatial cue. Used to preserve relational structure within one frame graph.
Persistent track	Persistent ID, semantic class, temporal span, static or dynamic flag. Used to connect observations across time.
Temporal event	Persistent ID, time span, source anchor, destination anchor, local rationale. Used to write anchor-relative object change.
Retrieval-plus doc	Time span, place, activity, salient objects, object states, interaction summary, edge summary, lexical tokens, and optional explanation fields. Used for retrieval and QA.

For a dynamic track p, the anchor-relative state at time t is

$$z_t(p) = (a_t, d_t, u_t, s_t), \tag{4}$$

where a_t is the dominant nearby anchor, d_t is a coarse relative distance, u_t is a relative direction or layout cue, and s_t is the current textual state. The representation is *relative* because these quantities are defined with respect to stable anchors rather than to a global world frame.

A temporal event is written whenever a dynamic object's dominant anchor changes and the new anchor-relative state remains stable across time,

$$e_i = (p, [t_s, t_e], a^-, a^+, r_i), \tag{5}$$

where $[t_s, t_e]$ is the event span, a^- and a^+ are the source and destination anchors, and r_i is a short local rationale drawn from nearby action or interaction context. This event view explicitly records which object changed, when the change occurred, and which stable reference changed with it.

Table 1 summarizes the schema used in the representation and in the downstream memory.

Figure 3 instantiates these variables on a concrete bag-transfer case selected from the released Day1-A1-JAKE scene-graph outputs: groceries are first bagged in a supermarket, then carried into the home entryway, and finally placed near the dining table for unpacking. This case was chosen because the anchor changes are unambiguous and the number of objects remains small enough for paper visualization. The three bands in Figure 3 correspond to the same evidence at three abstraction levels: scene layout and coarse 3D extent, a persistent anchor-transition track, and the retrieval documents finally consumed by QA. The example therefore makes explicit that the memory documents are not detached summaries, but compact views written from the same anchor-relative state changes visualized in the graph.

3.3 Queryable memory on top of the 4D graph

Frame graphs and temporal events are still too local and too numerous to be queried directly over a day-scale video. We therefore

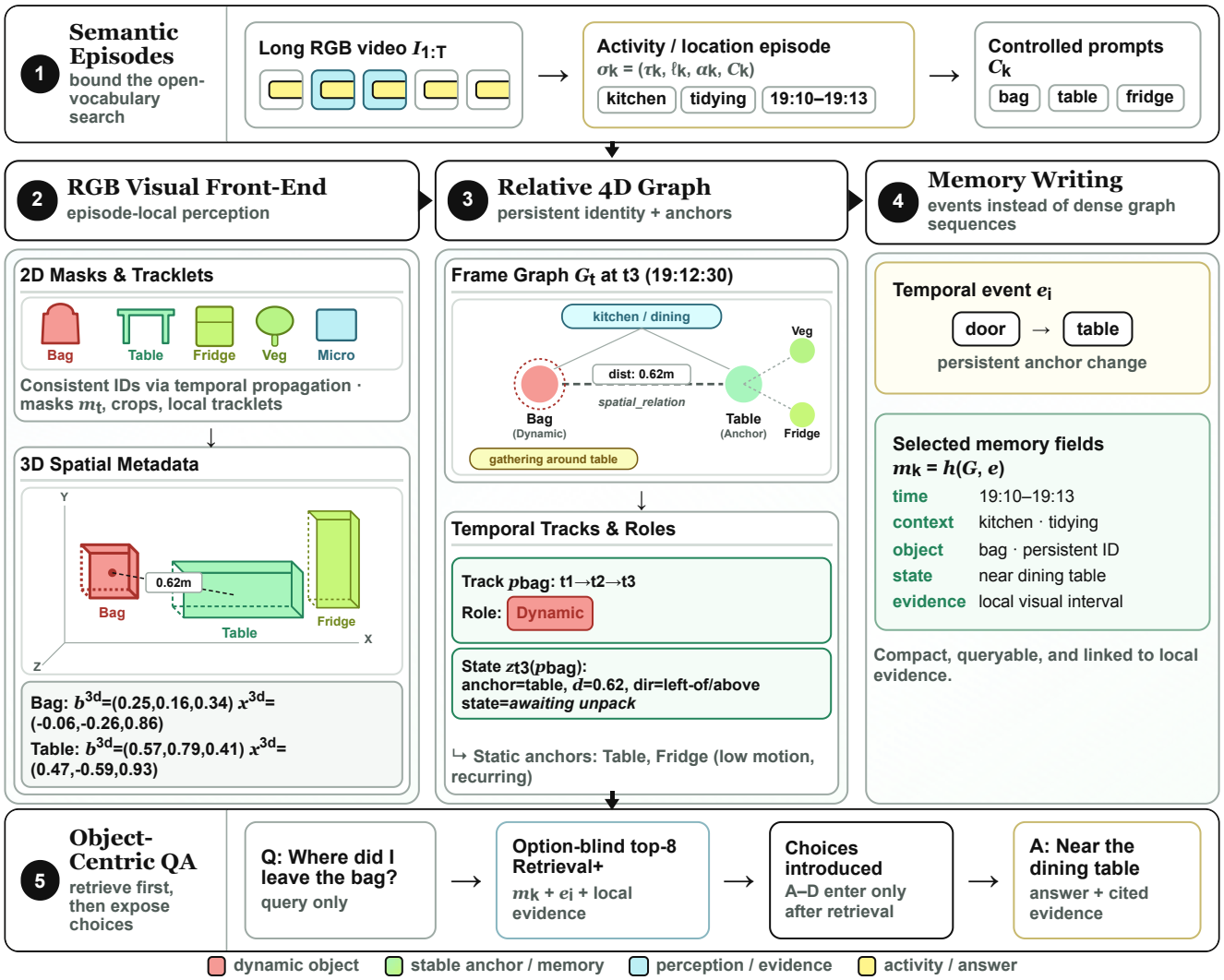


Figure 2: R4DSG pipeline. Semantic episodes guide RGB-only segmentation and lifting; persistent identities and anchor changes are then written as retrieval-ready memory for QA. The figure reads top-to-bottom, with its three core technical stages arranged left-to-right between episode parsing and downstream answer synthesis.

write them into a retrieval-ready episodic memory. For an activity-centered segment window S_k, the corresponding memory entry is represented as

$$m_k = (\tau_k, \ell_k, \alpha_k, \mathcal{O}_k, \mathcal{Z}_k, \mathcal{R}_k, \mathcal{I}_k, \mathcal{T}_k, \mathcal{Y}_k), \tag{6}$$

where τ_k is the time span, ℓ_k is the place, α_k is the activity, $\mathcal{O}_k$ is the salient object set, $\mathcal{Z}_k$ stores anchor-relative object states, $\mathcal{R}_k$ stores edge and relation summaries, $\mathcal{I}_k$ stores interaction or lexical-variant cues, $\mathcal{T}_k$ stores retrieval tokens, and $\mathcal{Y}_k$ stores optional explanation-oriented fields for `why` questions.

The document is written by the deterministic memory-writing operator h, which aggregates local frame graphs and overlapping temporal events,

$$m_k = h(\{G_t\}_{t \in S_k}, \{e_i\}_{e_i \cap S_k \neq \emptyset}). \tag{7}$$

Object nodes contribute class labels, object states, and relative 3D metadata; anchor-linked events contribute temporally compressed state transitions; and frame edges contribute the local relations that make the memory searchable by context rather than by object names alone. The memory is therefore written at the segment-window level, which keeps it compact while preserving enough relational evidence for QA.

This design directly supports downstream retrieval: `where` questions match place and anchor relations; `when` questions match time spans and anchor changes; object-state questions match salient objects, states, and edge summaries; and `why` questions can additionally use explanation fields. Because each entry can be serialized either as normalized text or as structured fields, the memory works with sparse, dense, or hybrid retrieval.

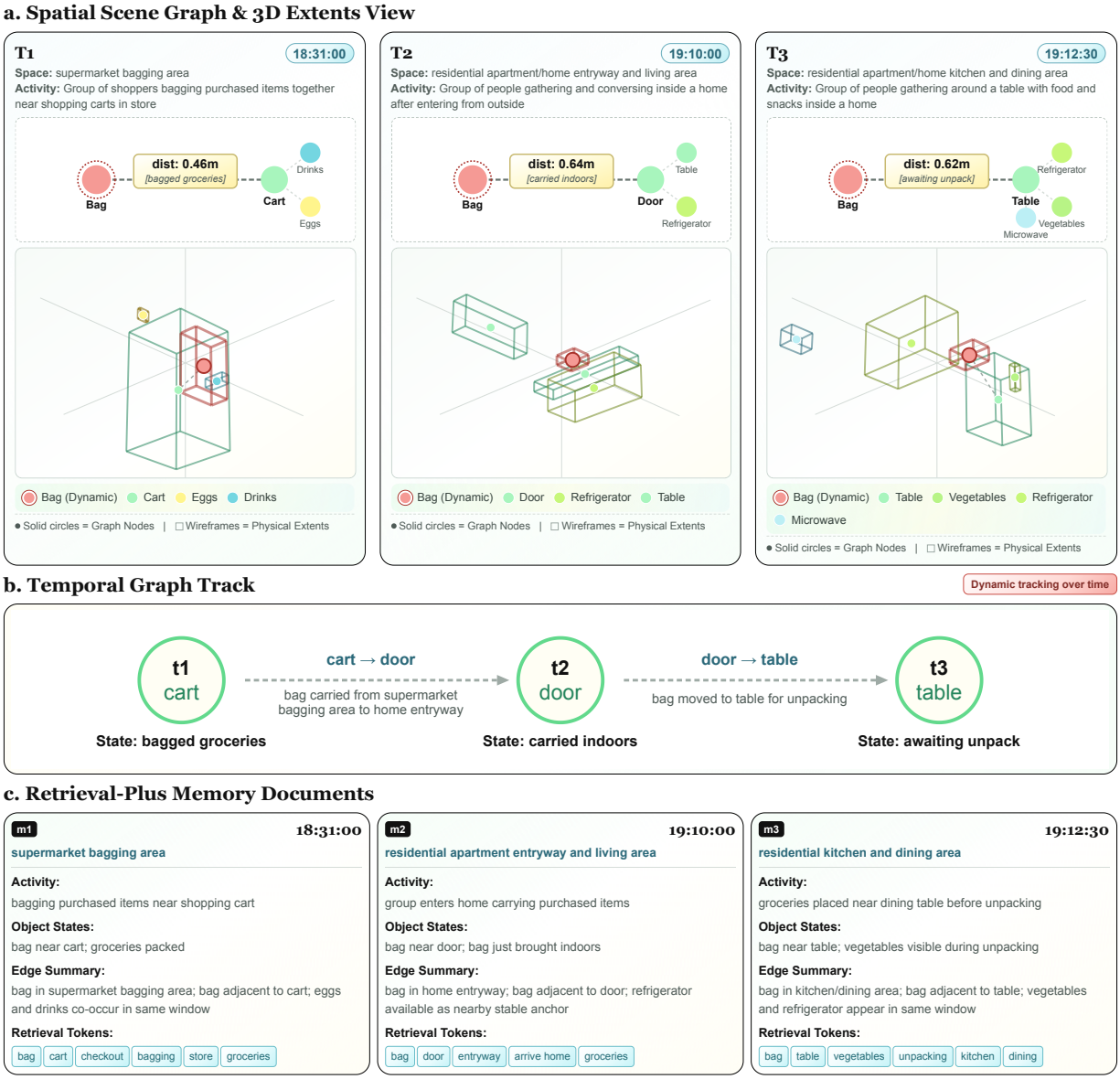


Figure 3: Bag-transfer example showing three anchor-relative states, the induced cart → door → table temporal track, and the resulting retrieval-plus memory documents. Top: scene graph and 3D extents at three timestamps; middle: the persistent track over anchors and state labels; bottom: the segment-level documents later retrieved for QA.

3.4 End-to-end pipeline

3.4.1 Semantic episode construction. A full egocentric day is too long, too redundant, and too semantically heterogeneous to be processed as one undifferentiated sequence. Before any object-centric reasoning happens, we therefore reorganize the raw video into semantic episodes. The video is segmented along two complementary axes: activity-oriented episodes and location-oriented episodes. Activity segments capture *what is being done*, whereas location segments capture *where it is taking place*. These two partitions are not redundant. Location windows stabilize anchor discovery, while activity windows provide the finer temporal units in which object states, interactions, and memory documents are written.

Each released segment record stores a timestamp span, a semantic label, and a controlled object vocabulary. We denote the resulting segment schema by

$$\sigma_k = (\tau_k, \ell_k, \alpha_k, C_k), \tag{8}$$

where τ_k is the segment span, ℓ_k is the location label, α_k is the activity label, and C_k is the controlled object set. These controlled objects become the semantic inventory for later segmentation, so the episode constructor already provides the scaffold that tells the

front-end which frames belong together and which object categories are likely to matter.

3.4.2 2D and 3D visual front-end with SAM 3-style consistency. Once semantic episodes are available, the system first seeks a stable 2D object substrate inside each episode. We use a SAM 3-style promptable video segmentation stage [2] followed by a SAM 3D-style RGB-only lifting stage [4]. The input prompts for SAM 3 are not drawn from a fixed detector vocabulary. Instead, they come from the segment semantics described above: the activity and location labels specify the local context, and the controlled object set C_k provides the target object inventory. In practice, SAM 3 is executed separately on activity-conditioned and location-conditioned windows so that the open-vocabulary search space stays bounded by episode semantics instead of ranging over the entire day.

This design is important for realistic egocentric video. Everyday scenes are open-vocabulary and highly variable, and the camera is uncalibrated and constantly moving. Segment-conditioned prompting gives SAM 3 local semantic context, while temporal mask propagation keeps the same prompted object visually consistent over neighboring frames. The output is therefore an episode-scoped bundle of prompted objects, masks, crops, and short local tracklets.

SAM 3D then lifts those observations into relative spatial metadata. The goal is not dense reconstruction, but the minimum 3D evidence needed for stable object-centric reasoning: `location_3d`, `bounding_box_3d`, and pairwise spatial relations. These metadata provide coarse object size, relative object position, and object-anchor geometry even when no global world frame is available.

3.4.3 Frame graph construction and persistent association. The 3D metadata are then compiled into frame-wise scene graphs. Each graph contains one space node from the location label, one activity node from the activity label, and a set of object nodes from the lifted object metadata. Edges connect objects to spaces and to one another through containment, support, adjacency, and generic relative spatial predicates. In the why-aware extension, activity nodes may also carry explanation-bearing attributes such as `reason`, `purpose`, or `why_summary`. This graph is the first fully symbolic interface in the pipeline: it converts visual evidence into a structure that can be aligned, compared, and accumulated over time.

Persistent identity is established in the next step. Since SAM 3 only guarantees local consistency inside an episode, the memory writer links graph nodes across neighboring frames using semantic compatibility, coarse 3D size continuity, relative 3D position continuity, and neighborhood consistency with nearby anchors. The association policy is deliberately conservative: if two observations disagree strongly, the track is split rather than force-merged; if a frame is missing an object because of blur or occlusion, the system keeps the neighboring track evidence instead of hallucinating a new state. Static anchors are inferred from objects that recur with low motion inside location-consistent windows, while dynamic objects are those whose dominant anchor or relative state changes over time. Because matching is performed in anchor-relative coordinates, the method never requires a globally aligned scene frame.

3.4.4 Memory writing. Once persistent tracks are available, the system writes two coupled memory views. The first is a temporal event stream, where a new event is emitted according to Eq. 5. The second is the retrieval-plus memory described in Sec. 3.3, written at the segment-window level. This second view aggregates the graph evidence needed for QA: place, activity, salient objects, anchor-relative object states, interaction summaries, edge summaries, lexical variants, and temporal span. When why-aware upstream annotations are available, dedicated explanation fields are written into $\mathcal{Y}_k$.

This writing step is where the relative 4D representation becomes operational for retrieval. The memory does not attempt to replay the whole video, but it also does not collapse everything into a few abstract events. Instead, it preserves exactly the intermediate granularity needed by long-horizon retrieval: enough compression to search efficiently, enough relational detail to recover the correct object-state evidence once a question arrives.

3.4.5 Question answering over retrieval-plus memory. At inference time, a question is mapped to a retrieval query over the memory entries rather than over raw frames. The main branch retrieves the top relevant retrieval-plus documents and linearizes their structured fields into evidence text. In the conservative option-blind retrieval protocol, answer options are introduced only after retrieval, in the final multiple-choice prompt to the answer model. `Why` questions can optionally route to the explanation-aware memory view when those fields are available. The experiments in Sec. 4 show that when the relevant object-state or explanation evidence has already been written into memory, retrieval-based QA becomes substantially more reliable.

Reproducibility. For each semantic episode, the pipeline (1) samples frames within the released activity/location windows, (2) constructs prompts from the episode labels and controlled object set, (3) lifts and associates observations using semantic, relative-size, relative-position, and anchor-neighborhood consistency, (4) emits persistent anchor changes using Eq. 5, and (5) writes segment documents using Eq. 7. QA retrieves the top eight documents and then applies the final multiple-choice prompt.

4 Experiments

4.1 Experimental settings

Dataset and evaluation. We evaluate on the publicly available EgoLifeQA A1_JAKE single-subject QA file, which contains 500 four-choice questions over seven recording days. Object-centric non-`who` filtering yields 255 questions, including 72 `when`, 15 `why`, and 149 with accessible target-time metadata. Accuracy is the primary metric. Our empirical claims are limited to this public single-subject split and do not imply multi-subject generalization.

Implementation and scale. All pipelines use Qwen3.5-27B as the answer model. R4DSG scene parsing from Sec. 3 is held fixed during QA, and the main branch retrieves the top $k = 8$ documents. The Day1-A1-JAKE stream used for all quantitative and qualitative experiments contains 828 clips and 06:51:50.52 of effective footage across five sessions.

Baselines. Beyond Plain RAG, EgoRAG-Text [38], and VLM-only, we evaluate adapted EMQA-style episodic [7] and AMEGO-inspired active memories [11], plus a No-Transition Object-Relation control. The adapted baselines are not official reproductions. The control uses the same cached visual evidence and local relations as

Table 2: Accuracy (%): Overall uses all 255 questions and When uses the 72-question when subset. "Question-only retrieval" uses only the question to retrieve evidence; "option-blind retrieval" withholds answer options until final answer selection; "no-why" removes optional WhyMemory fields, not why questions.

Method	Protocol	Overall	When
Plain RAG	question-only retrieval	29.4	27.8
EgoRAG-Text	question-only retrieval	32.9	30.6
VLM-only	question-only retrieval	29.8	20.8
R4DSG Retrieval+	question-only retrieval	**39.6**	**43.1**
EMQA-style episodic	option-blind retrieval	32.2	30.6
AMEGO-inspired active	option-blind retrieval	36.1	37.5
No-Transition obj.-rel.	option-blind retrieval	34.9	34.7
R4DSG Retrieval+	option-blind retrieval / no-why	**37.3**	**43.1**

Table 3: Exploratory explanation-oriented memory result (%). Overall uses all 255 questions; Why uses the 15-question why subset.

Metric	Retrieval+	WhyMemory
Overall	39.6	**40.0**
Why	33.3	**40.0**

R4DSG but removes persistent cross-segment identity and anchor-transition writing.

4.2 Experimental Results

We report overall and when accuracy under the question-only retrieval and conservative option-blind retrieval protocols, an exploratory why comparison, a memory-granularity ablation, and a compact memory-cost profile.

4.2.1 Object-centric and temporal QA. Table 2 consolidates the main results. Under question-only retrieval, R4DSG improves over EgoRAG-Text by 6.7 points overall and 12.5 points on when. Under option-blind retrieval / no-why, it remains strongest, exceeding the AMEGO-inspired baseline by 1.2 and 5.6 points, respectively. The No-Transition control is lower than R4DSG despite sharing cached visual evidence and local relations, which is consistent with a benefit from persistent identity and anchor-transition memory rather than object/relation serialization alone.

4.2.2 Explanation-sensitive evaluation with why-oriented memory. Compared with Retrieval+, WhyMemory adds four explanation fields: reason, purpose, utility, and why_summary. They are generated from local segment/activity context before QA, without ground-truth answers or answer options. Table 3 shows an increase from 33.3 to 40.0 on the 15-question why subset and a small overall lift. This result is exploratory because the subset is small; the no-why conservative result is reported separately in Table 2.

4.2.3 Memory representation evaluation across compression granularities. Event-only uses anchor-change events, Low-compression retains denser segment records, Episodic-only uses episode-level summaries, Hybrid combines event and episodic memories, and Retrieval+ uses the segment-window documents in Sec. 3.3.

Table 4: Memory-granularity ablation (%). Overall uses all 255 questions; When uses the 72-question when subset.

Metric	Event-only	Low-compression	Episodic-only	Hybrid	Retrieval+
Overall	33.7	34.5	35.7	34.1	**39.6**
When	27.8	34.7	34.7	26.4	**43.1**

Table 5: Offline memory scale profile.

Measure	Value
Clips / frame-info entries	828 / 2,476
Activity / location episodes	178 / 114
Retrieval+ documents / JSON size	134 / 0.58 MB
Average whitespace tokens per document	158.5

Table 4 shows that Retrieval+ has the highest observed overall and when accuracy. The lower observed accuracies of the alternative memory views are consistent with insufficient local context or a diluted change signal; this comparison does not isolate the causal mechanism. Retrieval+ retains the transition together with the place, activity, object, and interaction cues used for retrieval.

4.2.4 Memory scale. Table 5 summarizes the offline memory scale. The 828 clips produce 2,476 frame-info entries and 134 Retrieval+ documents in a 0.58-MB JSON memory.

5 Discussion and Limitations

Contribution and scope. R4DSG changes the memory rather than the answer model: SAM3-style consistency and RGB-only lifting supply visual evidence, while relative graphs and Retrieval+ organize it for long-range QA. The clearest gain is on when; evidence is limited to the public A1_JAKE single-subject split.

Limitations and failure modes. Errors may come from missed episodes; fragmented or duplicated tracks; cluttered anchors; retrieval misses; or insufficient causal or social context. Outputs are not yet a human-verified perception benchmark. Memory is retrospective and offline with limited cross-day persistence; online updates, stronger identity maintenance, and multi-subject validation remain future work.

6 Conclusion

R4DSG reframes long egocentric QA as memory construction over persistent objects, static anchors, and retrieval-ready documents. On object-centric EgoLifeQA, gains on when questions and the exploratory why-oriented extension suggest the value of temporal and explanatory memory. Relative scene graphs are a practical substrate for wearable assistants and embodied multimedia agents.

Acknowledgments

This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (Grant No. 2025ZD0122801) and The National Natural Science Foundation of China (Grant Nos. 62276063 and U23B2057).

References

[1] Runze Cai, Nuwan Janaka, Hyeongcheol Kim, Yang Chen, Shengdong Zhao, Yun Huang, and David Hsu. 2025. AiGet: Transforming Everyday Moments into Hidden Knowledge Discovery with AI Assistance on Smart Glasses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3706598.3713953

[2] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. 2025. SAM 3: Segment Anything with Concepts. *arXiv preprint arXiv:2511.16719* (2025). doi:10.48550/arXiv.2511.16719

[3] Qirui Chen, Shangzhe Di, and Weidi Xie. 2025. Grounded Multi-Hop VideoQA in Long-Form Egocentric Videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 2159–2167. doi:10.1609/aaai.v39i2.32214

[4] Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. 2025. SAM 3D: 3Dfy Anything in Images. *arXiv preprint arXiv:2511.16624* (2025). doi:10.48550/arXiv.2511.16624

[5] Anoop Cherian, Chiori Hori, Tim K. Marks, and Jonathan Le Roux. 2022. (2.5+1)D Spatio-Temporal Scene Graphs for Video Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 444–453.

[6] Meng Chu, Yicong Li, and Tat-Seng Chua. 2025. GraphVideoAgent: Enhancing Long-form Video Understanding with Entity Relation Graphs. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 4639–4648. doi:10.1145/3746027.3755537

[7] Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv Batra, and Devi Parikh. 2022. Episodic memory question answering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 19097–19106.

[8] Shangzhe Di and Weidi Xie. 2024. Grounded Question-Answering in Long Egocentric Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12934–12943.

[9] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, et al. 2023. Project Aria: A New Tool for Egocentric Multi-Modal AI Research. *arXiv preprint arXiv:2308.13561* (2023). doi:10.48550/arXiv.2308.13561

[10] Michael Goesele, Daniel Andersen, Yujia Chen, Simon Green, Eddy Ilg, Chao Li, Johnson Liu, Grace Kuo, Logan Wan, and Richard Newcombe. 2025. Imaging for All-Day Wearable Smart Glasses. *arXiv preprint arXiv:2504.13060* (2025). doi:10.48550/arXiv.2504.13060

[11] Gabriele Goletto, Tushar Nagarajan, Giuseppe Averta, and Dima Damen. 2024. AMEGO: Active Memory from Long EGOcentric Videos. In *Computer Vision – ECCV 2024*. doi:10.1007/978-3-031-72624-8_6

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

[13] Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. *Advances in Neural Information Processing Systems* 37 (2024).

[14] Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. 2025. CoTracker3: Simpler and Better Point Tracking by Pseudo-Labeling Real Videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 1–10.

[15] Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. 2024. Open3DSG: Open-Vocabulary 3D Scene Graphs from Point Clouds with Queryable Objects and Open-Set Relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14183–14193.

[16] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3613904.3642230

[17] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. 2024. Grounding image matching in 3d with mast3r. In *European conference on computer vision*. Springer, 71–91.

[18] Joshua Li, Fernando Jose Pena Cantu, Emily Yu, Alexander Wong, Yuchen Cui, and Yuhao Chen. 2025. SAMJAM: Zero-Shot Video Scene Graph Generation for Egocentric Kitchen Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 467–473. doi:10.1109/CVPRW67362.2025.00051

[19] Ke Ma and Jing Cao. 2019. Design Pattern as a Practical Tool for Designing Adaptive Interactions Connecting Human and Social Robots. In *International Conference on Intelligent Human Systems Integration*. Springer, 613–617. doi:10.1007/978-3-030-11051-2_93

[20] Ke Ma, Yizhou Fang, Jean-Baptiste Weibel, Shuai Tan, Xinggang Wang, Yang Xiao, Yi Fang, and Tian Xia. 2026. Phys-Liquid: A Physics-Informed Dataset for Estimating 3D Geometry and Volume of Transparent Deformable Liquids. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 7782–7790. doi:10.1609/aaai.v40i10.37721

[21] Ke Ma, Cong Fu, Jianing Wang, Yifei Wang, Wenyuan Li, Xinggang Wang, Meng Wang, and Tian Xia. 2026. WPIS: From In-the-Wild Web Images to Physics-Aware 3D Scene Graphs for Physical Reasoning. In *Proceedings of the ACM Web Conference 2026*. 1410–1421. doi:10.1145/3774904.3792591

[22] Ke Ma, Yifei Wang, Meng Wang, and Tian Xia. 2026. TransBiolab: A Real-World Multi-View Dataset of Cluttered Transparent Biomedical Objects. *arXiv preprint arXiv:2607.21071* (2026). doi:10.48550/arXiv.2607.21071

[23] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. *Advances in Neural Information Processing Systems* 36 (2023), 46212–46244.

[24] Costas Mavromatis and George Karypis. 2025. GNN-RAG: Graph Neural Retrieval for Efficient Large Language Model Reasoning on Knowledge Graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*. 16682–16699. doi:10.18653/v1/2025.findings-acl.856

[25] Zhiyu Pan, Yinpeng Chen, Jiale Zhang, Hao Lu, Zhiguo Cao, and Weicai Zhong. 2023. Find Beauty in the Rare: Contrastive Composition Feature Clustering for Nontrivial Cropping Box Regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 2011–2019. doi:10.1609/aaai.v37i2.25293

[26] Kevin Pu, Ting Zhang, Naveen Sendhilnathan, Sebastian Freitag, Raj Sodhi, and Tanya Jonker. 2025. ProMemAssist: Exploring Timely Proactive Assistance Through Working Memory Modeling in Multi-Modal Wearable Devices. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. doi:10.1145/3746059.3747770

[27] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman R"adle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Doll'ar, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714* (2024). doi:10.48550/arXiv.2408.00714

[28] Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. 2024. Action Scene Graphs for Long-Form Understanding of Egocentric Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18622–18632.

[29] Jintian Shi and Ke Ma. 2017. Digital Touchpoints in Campus Slow Traffic Service System. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 349–361.

[30] Aisha Urooj, Hilde Kuehne, Bo Wu, Kim Chheu, Walid Bousselham, Chuang Gan, Niels da Vitoria Lobo, and Mubarak Shah. 2023. Learning Situation Hyper-Graphs for Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14879–14889.

[31] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. 2020. Learning 3D Semantic Scene Graphs From 3D Indoor Reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3961–3970.

[32] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. 2024. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 20697–20709.

[33] Xiayu Wang, Ke Ma, Ruiyun Zhong, Xinggang Wang, Yi Fang, Yang Xiao, and Tian Xia. 2024. Towards Dual Transparent Liquid Level Estimation in Biomedical Lab: Dataset, Methods and Practices. In *European Conference on Computer Vision*. Springer, 198–214. doi:10.1007/978-3-031-73650-6_12

[34] Ziqin Wang, Bowen Cheng, Lichen Zhao, Dong Xu, Yang Tang, and Lu Sheng. 2023. Vl-sat: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 21560–21569.

[35] Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. 2023. Incremental 3d semantic scene graph prediction from rgb sequences. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5064–5074.

[36] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. 2021. SceneGraphFusion: Incremental 3D Scene Graph Prediction From RGB-D Sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7515–7525.

[37] Zizhao Wu, Haohan Li, Gongyi Chen, Zhou Yu, Xiaoling Gu, and Yigang Wang. 2024. 3D Question Answering with Scene Graph Reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 1370–1378. doi:10.1145/3664647.3681517

[38] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. 2025. EgoLife: Towards Egocentric Life Assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 28885–28900.

[39] Chenyangguang Zhang, Alexandros Delitzas, Fangjinhua Wang, Ruida Zhang, Xiangyang Ji, Marc Pollefeys, and Francis Engelmann. 2025. Open-Vocabulary Functional 3D Scene Graphs for Real-World Indoor Spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19401–19411.

[40] Chaoyi Zhang, Xitong Yang, Ji Hou, Kris Kitani, Weidong Cai, and Fu-Jen Chu. 2024. Egosg: Learning 3d scene graphs from egocentric rgb-d sequences. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2535–2545.

[41] Junrui Zhang, Jiaqi Li, Yachuan Huang, Yiran Wang, Jinghong Zheng, Liao Shen, and Zhiguo Cao. 2024. Towards Robust Monocular Depth Estimation in Non-Lambertian Surfaces. In *Computer Vision – ECCV 2024 Workshops*. 175–189.

[42] Qicheng Zhao, Yu Li, Qi Sun, and Zheyu Yan. 2026. ResilPhase: Plug-and-Play Phase Mapping and Noise-Resilient Macro-Trajectory Extrapolation for Diffusion Acceleration. *arXiv preprint arXiv:2606.26769* (2026). doi:10.48550/arXiv.2606.26769

[43] Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. 2023. Constructing Holistic Spatio-Temporal Scene Graph for Video Semantic Role Labeling. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5281–5291. doi:10.1145/3581783.3612096

[44] Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 8912–8924.