Event-Grounding Graph: Unified Spatio-Temporal Scene Graph from Robotic Observations

Phuoc Nguyen, Francesco Verdoja, Ville Kyrki

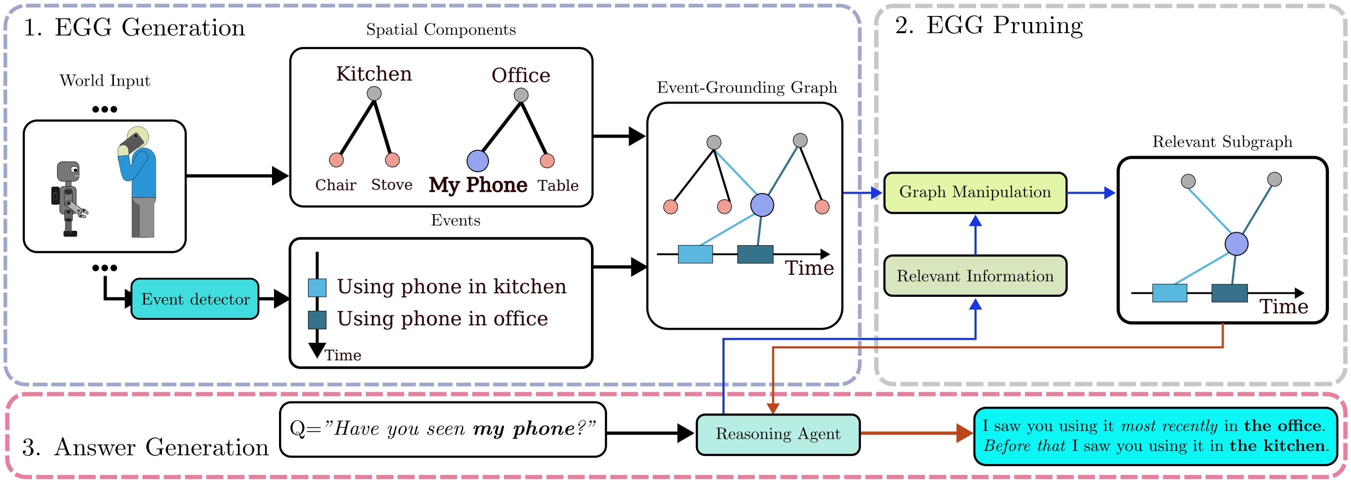

Fig. 1: We propose Event-grounding graph (EGG), a scene representation framework that grounds observed events with involved spatial elements (*e.g.*, objects) to allow robots to respond to complex queries.

***Abstract*— A fundamental aspect for building intelligent autonomous robots that can assist humans in their daily lives is the construction of rich environmental representations. While advances in semantic scene representations have enriched robotic scene understanding, current approaches lack a connection between spatial features and dynamic events; *e.g.*, connecting *the blue mug* to the event *washing a mug*. In this work, we introduce the event-grounding graph (EGG), a framework grounding event interactions to spatial features of a scene. This representation allows robots to perceive, reason, and respond to complex spatio-temporal queries. Experiments using real robotic data demonstrate EGG's capability to retrieve relevant information and respond accurately to human inquiries concerning the environment and events within. Furthermore, the EGG framework's source code and evaluation dataset are released as open-source at: `https://github.com/aalto-intelligent-robotics/EGG`.**

I. Introduction

Memory representations that encapsulate both the geometric and semantic aspects of the environment are crucial for robots to perceive and understand their surroundings while performing physical tasks to assist us in our daily lives. For instance, to answer a question *"Where is my phone?"*, a robot needs to complete three steps: identify objects that could be classified as "phones," determine which one is specifically "my phone," and then locate it. Humans typically recall an object's position based on previous experiences, such as where they *last* used it or where they *commonly* use it.

There has been a surge in research aimed at embedding semantic features into spatial scene elements. Most notable are the development of 3D scene graphs (3DSGs) [1]–[6]. By representing spatial concepts such as objects, freespace, and rooms as nodes and their inter-relationships as edges, 3DSGs boast rich semantic features while remaining compact in memory footprint. While this body of work incorporates semantic information into geometric features of the representation, *e.g.*, grounding the concept of "my phone" into a node, they do not capture dynamic events, *i.e.*, interactions that result in changes to the scene, frequently occurring within those environments.

Another line of research focuses on recording and verbalizing robot experiences and observations by leveraging the capabilities of visual-language models (VLMs) [7]–[9]. This line of research enhances robots' understanding of how the scene can be interacted with and enables them to respond to free-form spatio-temporal questions from humans. However, these works often describe the environment using free-form natural language captions, which often leads to a lack or misrepresentation of the spatial features of the environment, *e.g.*, multiple events of "using a phone" can be captured as text captions, but they do not provide information about which "phone" the person is using.

Currently, there is a gap in the representation of the connection between the scene's spatial feature changes and the interactions that cause them. Connecting the semantic attributes of these interactions (*i.e.*, their symbolic knowl-

This work was supported by the Research Council of Finland (decision 354909). The authors acknowledge the use of the MIDAS infrastructure of Aalto School of Electrical Engineering. P. Nguyen, F. Verdoja, and V. Kyrki are with School of Electrical Engineering, Aalto University, Espoo, Finland. `{firstname.lastname}@aalto.fi`

edge) with the scene's spatial features is an extension of the grounding problem [4], [10]. Grounding dynamic events through spatial features allows robots to not only perceive the environment, but also to recall and reason from historical interactions, *e.g.*, offering insights about unusual occurrences, assessing the states of routine tasks, or tracking personal items based on usage history.

In this work, we propose a novel framework, event-grounding graph (EGG), to aggregate, organize, and describe robotic observations over time, creating a unified representation that includes both spatial features and dynamic events. The rich information from the framework efficiently enables the extraction of relevant information subsets based on specific queries or tasks, ensuring insightful and contextually appropriate responses from robots, as illustrated in Fig. 1. We then demonstrate the capability of our representation through a series of experiments using real robotic data. The core contributions of this work are:

- EGG, a framework that unifies spatial features from the environment with dynamic events, allowing robots to perceive, reason, and explain what happens in a scene.
- A method that exploits the spatio-temporal queryability of EGG to retrieve a task-related compressed subgraph.
- Experimental results from real robotic data showing that EGG allows robots to retrieve relevant information and respond to human queries regarding the scene.
- An open-source implementation of EGG framework as well as the evaluation dataset, promoting further advancements in spatio-temporal scene understanding.

II. Related works

Enriching geometric robotic maps with semantic information is a vital step towards achieving deeper scene understanding. In addition to methods that integrate rich semantic data into conventional 2D and 3D maps [11]–[15], 3DSGs [1], [16] have emerged as an alternative representation and gathered considerable research interests. 3DSGs represents spatial concepts, such as rooms and objects, as nodes. As outlined in [4], 3DSGs are metric-semantic; they aim to ground the semantic information of a scene into the map's geometry, specifically through nodes and edges. Thus, they provide robots with semantic insights about the scene, enabling them to perform advanced tasks autonomously while maintaining a low memory footprint [2], [3]. Notably, Hughes *et al.* introduces Hydra [3], [4], a real-time 3DSG construction framework that solidifies the practical application of this representation on real robots. Concept-Graphs [2] and HOV-SG [5] enrich the semantic information available for object nodes in a 3DSG by incorporating open-vocabulary features into the frameworks. By leveraging the rich semantic information available in scene graphs, multiple works have utilized reasoning agents such as large language models (LLMs) to perform complex tasks such as pick-and-drop [10], embodied question-answering [17], interactive object search [6], or rearrangements [18]. However, current 3DSGs are limited to grounding semantic information into static spatial concepts and can only handle tasks that do not require awareness of external interactions within the environment. Current spatio-temporal 3DSG frameworks, such as [19]–[21], only focus on updating the geometric features of the environment, and not on the semantic context that causes scene changes.

Another significant body of research focuses on developing memory representations that incorporate interactions with the environment. Bärmann *et al.* [22] propose a framework that systematically records a robot's interaction history within a hierarchical tree-like structure. This representation includes raw robotic sensory inputs at its base layer and abstracted descriptions of the robot's experiences at higher layers. Closer to our work, ReMEmbR [7] constructs a spatio-temporal representation that aggregates video captions generated by a VLM, along with timestamps and the robot's positional data, into a vector database. Another relevant work is Embodied-RAG [9], which develops a spatio-temporal representation in the form of a semantic forest. This representation is constructed by embedding video captions captured at various points in time into the nodes of a topological graph, with a LLM summarizer generating higher-level nodes from these embeddings. While these approaches acknowledge the importance of interactions with the environment, they exhibit limitations concerning the grounding of recorded events in the spatial features of the environment. This leads to issues such as multi-view inconsistency and the unnecessary accumulation of redundant data, exemplified by the repetitive description of the same scene upon revisit. Our work addresses this gap.

III. Problem Formulation

Formally, a scene can be described by the set of N tracked spatial elements S, *e.g.*, objects, rooms, or free space. Each spatial element $S^i \in S$ with $i \in [1, N]$ is characterized by a set of attributes. The choice of attributes is task dependent, but commonly includes at least position in a known world frame and a semantic class or name. We classify attributes into two types: time-invariant, *e.g.*, semantic class, material, and color, and time-variant, *e.g.*, position and state of operation. Thus, $S^i = (\mathcal{I}^i, \mathcal{V}^i)$, with $\mathcal{I}^i$ and $\mathcal{V}^i$ being the set of time-invariant and time-variant attributes respectively. We denote with $\mathcal{V}_T^i = \{\mathcal{V}_j^i \mid t_j \in T\}$ the history of time-variant attributes of spatial element S^i in the time interval T, *i.e.*, the list of all recorded values of the attributes at any instant within the interval.

We define an event that occurs in the time interval T as a sequence of interactions executed by either the observer or an external agent within the scene, each resulting in an attribute change of at least one tracked spatial element at a particular instant $t \in T$. We denote an event by $E_T = (O_T, D_T, s_T)$. $O_T \subseteq S$ is the set of spatial elements that experience attribute changes during the event. Attribute changes, $\mathcal{V}_T = \{\mathcal{V}_T^i \mid S^i \in O_T\}$, may include changes to objects' positions, *e.g.*, someone moving the chairs, the appearance or disappearance of objects, *e.g.*, someone taking a mug outside of the robot's view, or modifications in their operational states due to interactions, *e.g.*, someone

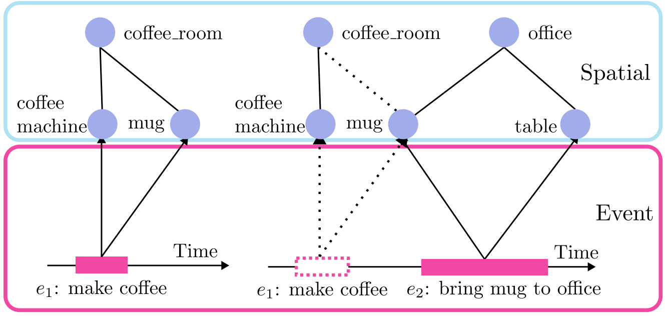

Fig. 2: Spatial and event components connections. Spatial and event edges are not permanent and are characterized by time intervals.

turning the lights on or off. D_T is the set of element-wise descriptions of each change, providing insight into the role that each spatial element plays in the event. Formally, for every attribute change $\mathcal{V}_T^i \in \mathcal{V}_T$, a corresponding description $d_T^i \in D_T$ exists that verbally describes the change. s_T is the event summary, which provides an overall description of the event. Collectively, an event E_T not only outlines the scene change, but also describes how each spatial element is affected during the interval T.

Our first objective is to create a representation $\mathcal{G}$ that aggregates, describes, and grounds observed events E with the spatial elements S of the scene. Secondly, by harnessing the information from $\mathcal{G}$, we extract task-relevant details to respond to free-form natural language queries Q. Particularly, our goal is to predict an answer A given a query Q and a representation $\mathcal{G}$.

IV. Methodology

Fig. 1 illustrates the pipeline of our framework. Initially, from observed events, we generate an EGG $\mathcal{G}$. Then, using information relevant to a query Q, we extract a subgraph $\mathcal{G}_Q$. Finally, we send $\mathcal{G}_Q$ to a reasoning agent to generate the answer A.

A. EGG generation

EGG consists of two interrelated sets of components: spatial components and event components, as illustrated in Fig. 2. Formally, EGG is the graph $\mathcal{G} = (\mathcal{N}^S, \mathcal{N}^E, \mathcal{E}^S, \mathcal{E}^E)$, with $\mathcal{N}^S$ and $\mathcal{N}^E$ being the spatial and event nodes, and $\mathcal{E}^S$ and $\mathcal{E}^E$ being the spatial and event edges.

The spatial components $\mathcal{G}^S = (\mathcal{N}^S, \mathcal{E}^S)$ follow a hierarchical 3DSG structure analogous to that of Hydra [3], [4]. Each spatial node $n_i^S \in \mathcal{N}^S$ corresponds to the attribute set of a spatial element, so $n_i^S = S^i$. In this work, the constructed 3DSG consists of two layers: rooms and objects. However, the formulation from Sec. III does not prevent the inclusion of other spatial concepts, such as buildings and free space [3], [4] or asset objects [21]. Unlike static scene graph frameworks such as [2]–[4], objects in EGG also include time variant attributes, which track changes in objects resulting from event interactions. Each spatial edge $e_{il}^S = (T_{il}, n_i^S, n_l^S)$ connecting two nodes, n_i^S and n_l^S, represents spatial relationships such as inclusion, *e.g.*, the mug is in the kitchen. Since nodes are not permanent in EGG, edges are additionally characterized by the time interval T_{il} during which the spatial relationship exists.

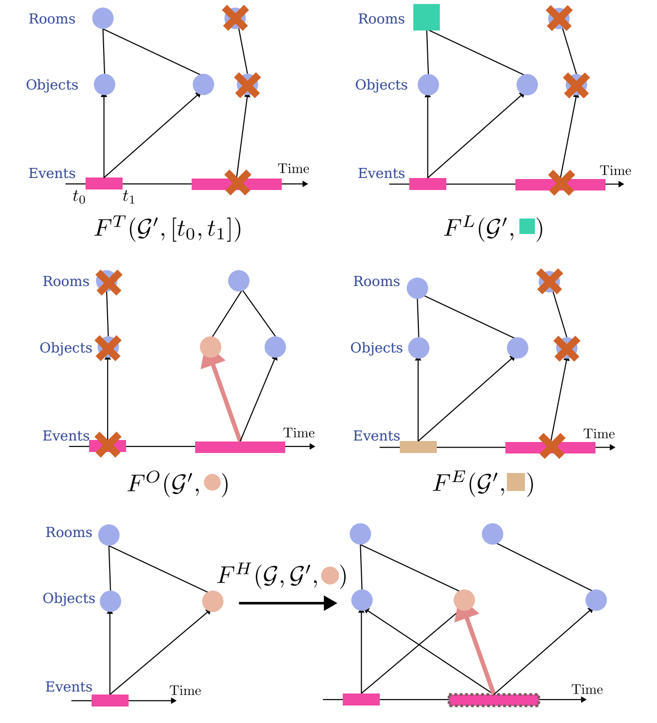

Fig. 3: Task-based graph manipulation

The event components $\mathcal{G}^E = (\mathcal{N}^E, \mathcal{E}^E)$ represent the observed events and their relationships with the interacted objects. Fig. 2 illustrates how events are grounded with multiple objects. To each event E_T corresponds an event node $n_T^E = (p_T, s_T)$ defined spatially by the observation points of the robot p_T, *i.e.*, the sensor's pose, and temporally by the respective time interval T. The event overall description is provided by the summary s_T. Each event edge $e_{Ti}^E = (T, n_T^E, n_i^S, d_T^i)$ connects the event E_T to the object $S^i \in O_T$. Similar to the spatial edges, the event edges are also timestamped and exist only in the interval T when the event occurs. The event description $d_T^i \in D_T$ describes the role of the object S^i in the event.

B. EGG pruning

After receiving a query Q, the full graph $\mathcal{G}$ is pruned into a subgraph $\mathcal{G}_Q$ which contains information relevant to that query. The pruning aims to both decrease the computational effort needed to generate the answer and to improve the quality of the answer by removing potentially distracting irrelevant input from the answer generation process.

To this end, we assume the information relevant to a query, denoted as I_Q, can be split into four types: time interval T_Q, locations L_Q, objects O_Q, and events E_Q. We will describe how to retrieve I_Q from Q in Sec. V-B. Given the set of relevant information $I_Q = (T_Q, L_Q, O_Q, E_Q)$ and an EGG $\mathcal{G}$, we retrieve the query-relevant subgraph $\mathcal{G}_Q \subseteq \mathcal{G}$ by using a series of graph manipulation functions, illustrated in Fig. 3 and detailed below. Each function takes as input a subgraph $\mathcal{G}' = \{\mathcal{N}^{S'}, \mathcal{E}^{S'}, \mathcal{N}^{E'}, \mathcal{E}^{E'}\} \subseteq \mathcal{G}$ and a subset of I_Q to return another subgraph $\mathcal{G}^* = \{\mathcal{N}^{S*}, \mathcal{E}^{S*}, \mathcal{N}^{E*}, \mathcal{E}^{E*}\} \subseteq \mathcal{G}$. This is achieved first by selecting the list of relevant nodes $\{\mathcal{N}^{S*}, \mathcal{N}^{E*}\}$ according to a set of rules, and then by adding all edges connected to the selected nodes to $\mathcal{G}^*$, *i.e.*, $\mathcal{E}^{S*} = \{e^S_{ij} \mid n^S_i, n^S_j \in \mathcal{N}^{S*}\}$ and $\mathcal{E}^{E*} = \{e^E_{Ti}, \mid n^S_i \in \mathcal{N}^{S*}, n^E_T \in \mathcal{N}^{E*}\}$. Specifically, the graph manipulation functions are:

- **Time pruning** $F^T(\mathcal{G}', T_Q)$: From a time interval T_Q, we can select a subgraph $\mathcal{G}^*$ consisting of the set of events that transpire during it. The resulting subgraph consists of event nodes $\mathcal{N}^{E*} = \{n^E_T \in \mathcal{N}^{E'} | \; T \subseteq T_Q\}$. The preserved spatial components are those with at least one connection to the events in the pertinent time window, that is, $\mathcal{N}^{S*} = \{n^S_i \in \mathcal{N}^{S'} \mid (\exists e^E_{Ti} \in \mathcal{E}^{E'}) \; T \subseteq T_Q\}$.
- **Location pruning** $F^L(\mathcal{G}', L_Q)$: From a list of locations L_Q, we select the subgraph $\mathcal{G}^*$ that consists of events and objects from those locations. Specifically for this work, the locations list $L_Q \in \mathcal{N}^{S'}$ is essentially a list of spatial nodes that exists on the room layer. Thus $\mathcal{N}^{S*} = L_Q \cup \{n^S_i \in \mathcal{N}^{S'} \mid (\exists n^S_l \in L_Q) \; e^S_{li} \in \mathcal{E}^{S'}\}$, and $\mathcal{N}^{E*} = \{n^E_T \in \mathcal{N}^{E'} \mid (\exists n^S_l \in L_Q) \; e^E_{Tl} \in \mathcal{E}^{E'}\}$.
- **Object pruning** $F^O(\mathcal{G}', O_Q)$: From a list of query-relevant object nodes $O_Q \subseteq \mathcal{N}^{S'}$, we select the subgraph $\mathcal{G}^*$ that contains relevant objects and the events that involve them. Specifically, we prune the graph such that the event nodes are connected to the relevant objects, $\mathcal{N}^{E*} = \{n^E_T \in \mathcal{N}^{E'} \mid (\exists n^S_i \in O_Q) \; e^E_{Ti} \in \mathcal{E}^{E'}\}$. Then, the preserved spatial nodes are the ones connected to at least one of those events, *i.e.*, $\mathcal{N}^{S*} = \{n^S_i \in \mathcal{N}^{S'} \mid (\exists n^E_T \in \mathcal{N}^{E*}) \; e^E_{Ti} \in \mathcal{E}^{E'}\}$.
- **Event pruning** $F^E(\mathcal{G}', E_Q)$: From a list of query-relevant event nodes $E_Q \subseteq \mathcal{N}^{E'}$, we select the subgraph $\mathcal{G}^*$ that contains relevant events and the objects that they involve. The event nodes in this case are just $\mathcal{N}^{E*} = E_Q$, and the spatial nodes are those that are connected to at least one of these events, *i.e.*, $\mathcal{N}^{S*} = \{n^S_i \in \mathcal{N}^{S'} \mid (\exists n^E_T \in E_Q) \; e^E_{Ti} \in \mathcal{E}^{E'}\}$.
- **History expansion** $F^H(\mathcal{G}, \mathcal{G}', O_Q)$: From a list of objects $O_Q \subseteq \mathcal{N}^{S'}$, we expand the graph to include the history of events that involve them. The event nodes that involve O_Q are $\mathcal{N}^{E*} = \{n^E_T \in \mathcal{N}^E \mid (\exists n^S_i \in O_Q) \; e^E_{Ti} \in \mathcal{E}^E\}$. The spatial nodes include those that are given by O_Q, and related to the expanded event history, *i.e.*, $\mathcal{N}^{S*} = O_Q \cup \{n^S_i \in \mathcal{N}^S \mid (\exists n^E_T \in \mathcal{N}^{E*}) \; e^E_{Ti} \in \mathcal{E}^E\}$.

Thus, we retrieve the subgraph:

$$\mathcal{G}_Q = \mathcal{C}(F^T, F^L, F^O, F^E, F^H)(\mathcal{G}, I_Q) \tag{1}$$

where $\mathcal{C}(\cdot)$ is any combination of the graph manipulation functions. We will present our implemented combination in Sec. V-B.

V. Implementation

In addition to the formulation presented in Sec. IV, our implementation incorporates the following assumptions:

- All objects recorded within the events are considered unique entities.
- A localization method is available to determine the robot's pose during the mapping sessions [23].
- A method exists to identify moments when human activities occur within the scene and to select the objects relevant to these activities. In the experiments, we manually identify these moments and objects for each robot mapping session.
- A method for re-identifying objects post-mapping session is established [19]. In the experiments, we use ground truth information to re-identify the objects.

A. Building EGG from robotic observations

We construct EGG from a stream of multimodal robotic data, including RGB-D images accompanied by frame-wise timestamps and the robot's pose. We derive object data from an RGB-D video stream to generate spatial nodes. In our implementation, the time-variant attribute set $\mathcal{V}^i$ for each object node includes the varying positions of the objects and the state of operation. The position of each object is retrieved by projecting its segmented point cloud from the depth image. While we do not explicitly capture the state of operation, they can be inferred from the instance-wise descriptions D_T of the event edges, which we will discuss in more detail later in this section. For simplicity, we record only the initial and final positions of each object for each recorded event. To enrich the object nodes and differentiate between instances of the same semantic class, we gather views of each object from multiple perspectives and utilize a VLM (GPT4o [24]) to generate concise captions that describe their appearance. Thus, the time-invariant attribute set $\mathcal{I}^i$ of each object node consists of its name, semantic class, and caption. The room nodes only have time-invariant attributes, which are their manually labeled names and their positions.

To form the event components $\{\mathcal{N}^E, \mathcal{E}^E\}$ we retrieve event captions using the video captioning model VideoRefer [25]. In order to retrieve the summary caption s_T, we prompt the model to describe the actions of the person in the video, alongside a list of manually selected involved objects O_T. Furthermore, to generate element-wise descriptions D_T, we ask the model to express the relationship between the person featured in the video and the corresponding objects. The position p_T of the event node is the mean of all the robot's camera positions during the observation time interval T. To evaluate the quality of automated video captioning, for the experiments presented in Sec. VI, we also manually generate all corresponding ground-truth event captions.

B. Task-based graph pruning

The entire EGG can be text-serialized into a JSON data format, which can subsequently be parsed directly by a pretrained LLM [24]. In this implementation, we deploy a multi-stage sampling strategy to approximate I_Q, along with the graph manipulation functions in Sec. IV-B to estimate the task-relevant subgraph $\mathcal{G}_Q$. In the first stage, based on the query, we prompt the LLM agent to select the relevant time interval T_Q and rooms L_Q to look for the information. We then retrieve the subgraph:

$$\mathcal{G}_Q^{(1)} = F^L(F^T(\mathcal{G}, T_Q), L_Q) \tag{2}$$

Next, we provide the LLM agent with a list of object names and captions and event summary descriptions from $\mathcal{G}_Q^{(1)}$. The agent then, based on the semantic context of the query, is tasked to extract the objects O_Q and events E_Q related to the query. If there are both objects and events related to the query, we merge them to identify the objects that are most relevant to the query and expand their history. More precisely, if $O_Q \neq \emptyset \wedge E_Q \neq \emptyset$, then $O_Q^* = \{n_i^S \in O_Q \mid (\exists n_T^E \in E_Q)\ e_{Ti}^E \in \mathcal{E}^{E(1)}\}$ and $E_Q^* = \{n_T^E \in E_Q \mid (\exists n_i^S \in O_Q)\ e_{Ti}^E \in \mathcal{E}^{E(1)}\}$, with $\mathcal{E}^{E(1)}$ being the set of event edges of $\mathcal{G}_Q^{(1)}$. *E.g.*, if $Q =$*"Where is the mug that I was drinking coffee with in the coffee room yesterday?"*, $O_Q = \{$all mugs found in the coffee room yesterday$\}$, and $E_Q = \{$all events involving a person drinking coffee$\}$. After merging, O_Q^* will contain only the mugs used for drinking coffee in the coffee room yesterday. In the case that $O_Q = \emptyset$ or $E_Q = \emptyset$, O_Q and E_Q are not merged. Thus, we retrieve the subgraph:

$$\mathcal{G}_Q^{(2)} = \begin{cases} F^E(F^O(\mathcal{G}_Q^{(1)}, O_Q^*), E_Q^*), & \text{if } O_Q \neq \emptyset \wedge E_Q \neq \emptyset \\ F^O(\mathcal{G}_Q^{(1)}, O_Q), & \text{if } O_Q \neq \emptyset \wedge E_Q = \emptyset \\ F^E(\mathcal{G}_Q^{(1)}, E_Q), & \text{if } O_Q = \emptyset \wedge E_Q \neq \emptyset \\ \emptyset, & \text{if } O_Q = \emptyset \wedge E_Q = \emptyset \end{cases} \tag{3}$$

Finally, if $O_Q \neq \emptyset$, we retrieve the task-relevant subgraph $\mathcal{G}_Q$ by expanding the event history of relevant objects and pruning the graph to the relevant time interval T_Q. Intuitively, we expand the history of the relevant objects to retrieve events involving them at all locations in EGG, since objects could move around and events that happen before the key events in E_Q could have a role in determining the answer for Q. We then prune the graph to the relevant time interval to reduce the graph size and remove the events that are too far in the past or future.

$$\mathcal{G}_Q = \begin{cases} F^T(F^H(\mathcal{G}, \mathcal{G}_Q^{(2)}, O_Q^*), T_Q), & \text{if } O_Q \neq \emptyset \wedge E_Q \neq \emptyset \\ F^T(F^H(\mathcal{G}, \mathcal{G}_Q^{(2)}, O_Q), T_Q), & \text{if } O_Q \neq \emptyset \wedge E_Q = \emptyset \\ \mathcal{G}_Q^{(2)}, & \text{if } O_Q = \emptyset \end{cases} \tag{4}$$

Finally, we text-serialize $\mathcal{G}_Q$ into JSON format, which is then integrated into the prompt for the LLM agent to respond to the user's query[1].

VI. Evaluation

A. Experimental setup

To evaluate EGG, we set up and conduct a series of experiments using the representation to perform an information retrieval task, where an LLM agent uses the information from the scene to answer queries in natural language. With these experiments, we aim to address the following research questions:

- What is the impact of the spatial and event components of EGG on the performance of information retrieval tasks, both individually and in combination?
- How does pruning EGG influence information retrieval accuracy and token usage?
- In what ways does the accuracy of generated event captioning affect scene representation, and what is its impact on downstream information retrieval tasks?

Currently, to the best of our knowledge, there is no suitable dataset to evaluate EGG's capability. We acknowledge the existence of prior datasets that document human activities in natural settings [26], [27]. However, these datasets are primarily designed for video captioning research and do not include crucial robotic data, such as depth images and poses, which are necessary for building and evaluating EGG. Another notable example is the Ego4D dataset, which contains over 3000 hours of daily-life activity videos captured from an egocentric perspective, along with object detection labels and 3D poses for some samples [28]. Yet, the videos in this dataset are not recorded within a consistent environment, nor do they provide instances where the same object is reused, rendering it inadequate for benchmarking EGG's capability to maintain a history of each object.

To address this gap, we evaluate our framework using real-world data collected from a Hello Robot Stretch 2 mobile manipulator platform with an Astra 2 RGB-D camera mounted on top. Data collection was conducted in two locations within our department building: a coffee room and an office. Our dataset includes 35 videos, captured by the robot, of an individual performing various daily activities and interacting with 21 distinct objects in a coffee room and an office. On average, an event consists of 3 to 4 objects being interacted with. Additionally, we provide frame-wise timestamps, object instance segmentation, event-relevant object identification, and local positions of the robot.

The robot task involves verbalizing its observations of a scene through different forms of scene representation. We manually label each video with ground truth summary captions and object role captions to assess the capability of VLMs in comparison to human judgment. For each event, we also manually provide segmentation masks for each object at their first and last appearances, enabling projection to capture their instance positions and incorporate them into the position history of their respective spatial nodes (refer to Sec. V-A for implementation details of the spatial graph). The dataset comprises 80 questions with ground truth answers across four modalities, *for instance*[2]:

- *text* (24 queries): Returns an answer in natural language. *E.g.*, *Which room was the phone last seen used in?*
- *binary* (28): Returns True or False. *E.g.*, *Was the red ceramic mug ever used to make coffee?*
- *node* (18): Returns a list of spatial nodes. *E.g.*, *The person left something on the office table at some point between 15:30:00 and 16:16:00. What was it?*

[1] All prompts and specific LLM used can be found in the paper codebase.

[2] The full dataset, including the events and question-answer pairs, will be made publicly available after acceptance to preserve anonymity.

TABLE I: Evaluating the capability of EGG in verbalizing robot observations against baseline methods. The results are averaged over 5 trials. VideoRefer [25] is used for automatic captioning of events for EGG and ReMEmbR.

Method	$\mathbf{S^{text}} \uparrow$	$\mathbf{F1^{binary}} \uparrow$	$\mathbf{F1^{node}} \uparrow$	$\mathbf{A^{time}} \uparrow$
EGG	**0.70**	**0.78**	**0.76**	**0.78**
*Hydra**[3], [4]	0.52	0.00	0.68	0.10
ReMEmbR†[7]	0.54	0.49	✖	0.38

* Implemented by removing the event nodes $\mathcal{N}^E$ and event edges $\mathcal{E}^E$ from EGG

† Since ReMEmbR does not have access to the scene graph, we did not evaluate it on the queries with *node* modality.

- *time* (10): Returns a point in time. *E.g., What is the earliest time the person was seen working?*

To assess the accuracy of the generated responses for the queries with *text* modality we employ LLM semantic scores, where we employ a LLM to judge each generated answer with a score from 0 to 1. This evaluation method was chosen over manual evaluation to mitigate potential human biases in judgment and leverage the established capability of LLMs to provide evaluation scores that consider context accuracy and nuance [29]. For the *time* modality, we consider an answer to be correct if it lies within 2 minutes of the ground truth answer. Additionally, we provide quantifiable F1 scores for *node* ($F1^{node}$) and *binary* ($F1^{binary}$) modalities.

In our experiments, GPT4o [24] is utilized to handle object captioning, graph pruning, and reasoning, and VideoRefer [25] is utilized for automatic video captioning.

B. Experimental results

We evaluate our method against two baselines, a purely spatial memory representation, Hydra [3], [4], and a spatio-temporal memory building method, ReMEmbR [7]. Our experimental results, presented in Tab. I, clearly demonstrate that our method outperforms both baselines across all modalities. From these findings, it can be inferred that the unified spatio-temporal features of EGG preserve the advantages of both purely spatial and temporal representations, enabling the LLM agent to reason more effectively, especially on instance-based queries.

Qualitatively, Hydra [3], [4], a purely spatial scene representation, struggles to address event-related queries. By observing the reasoning output, the LLM often rely solely on the positions of objects to infer their relationships. However, this approach is ineffective for queries that require detailed information about human interactions, which are not typically captured in a 3DSG. For instance, the query *"Describe the appearance of the most frequently used mug for making coffee at the coffee machine"* requires usage frequency of different mugs to correctly answer.

On the other hand, ReMEmbR [7] performs better on descriptive queries with its access to event data. Yet, it lacks the grounding of events with spatial features, so the LLM can only rely on video captions. Notably, it fails to address queries such as *"How many different mugs were seen in the events?"*, which could be easily answered with spatial representations. Furthermore, it struggles with queries requiring the usage history of specific objects, such as *"Describe the appearance of the most frequently used mug for making coffee at the coffee machine"*.

Overall, it is evident that each component of EGG—the spatial, event components—plays a crucial role in enriching the information provided to the LLM agent.

C. Ablation study

The results from Tab. II illustrate several ablation studies on different versions of EGG. In addition to the metrics mentioned in Sec. VI-A, we provide an additional accuracy score across all modalities (A^{all}), a tokens count and graph compression percentage, which is the ratio of the number of characters of the JSON-serialized subgraph $\mathcal{G}_Q$ with the JSON-serialized full graph $\mathcal{G}$.

1) The role of event edges: To assess the importance of event edges, we perform pruning on the original EGG and omit them when JSON-serializing the graph (*w.o edges* in Tab. II). This version of EGG performs close to the standard version, but misses some needed details for understanding specific interactions. In most cases, the pruning process essentially removes unrelated objects of the same class from the resulting graph. *E.g.*, for the query $Q_1 =$*"Which mug was most often used for coffee-making?"*, only the mugs involved in the coffee-making events are present in the subgraph $\mathcal{G}_{Q_1}$. Although complete edge information is valuable and the standard version shows higher performance, removing edges after pruning can still provide reasonable performance when token limits are a consideration.

2) Graph pruning efficiency: To assess how the subgraph $\mathcal{G}_Q$ affects reasoning performance and token usage, we compare two versions of the information search strategy: one utilizing the pruning method detailed in Sec. V-B and the other using the complete graph with the LLM agent (EGG w.o pruning in Tab. II). The version employing graph pruning outperforms the alternative and achieves approximately 65% graph compression, resulting in a 36% reduction in token usage. It should be noted that the token count includes the input prompt tokens. The pruned version of EGG demonstrated significantly higher overall semantic score in the downstream information retrieval task. Disregarding the excess input tokens, we also hypothesize that this improvement may stem from the LLM's tendency to hallucinate when presented with redundant information, as discussed in studies like [30]. Moreover, LLMs have shown positional biases towards information presented at the beginning or end of the input prompt [31]. The experimental results indicate that the pruning process improves LLM performance on downstream tasks by excluding task-irrelevant entities from the representation.

3) Video captioning capability: In this study, we construct EGG using two different video captioning methods: using ground truth labeling and using a VLM, VideoRefer [25]. As anticipated, the quality of video captioning directly affects the accuracy of subsequent tasks. As detailed in Tab. I, automatic captioning reduces the overall accuracy

TABLE II: Ablation study on EGG. The results are averaged over 5 trials.

Method	$\mathbf{A^{all}} \uparrow$	$\mathbf{S^{text}} \uparrow$	$\mathbf{F1^{binary}} \uparrow$	$\mathbf{F1^{node}} \uparrow$	$\mathbf{S^{time}} \uparrow$	Tokens (1K) ↓	Compression ↑
EGG	**0.83**	**0.85**	**0.77**	**0.87**	0.80	998	64.72 %
EGG (w.o edges)	0.79	0.81	0.72	0.82	0.80	**922**	**68.02** %
EGG (w.o pruning)	0.77	0.80	0.65	0.79	**0.88**	1576	0.00 %
EGG (auto-captions)	0.76	0.70	0.78	0.76	0.78	985	66.55 %

* Ground truth captioning is used for event nodes in all versions except for *auto-captions*, for which VideoRefer [25] is used.

by 9% compared to using ground truth event labeling. These results underscore the necessity for advancements in video captioning models.

D. Failure modes analysis

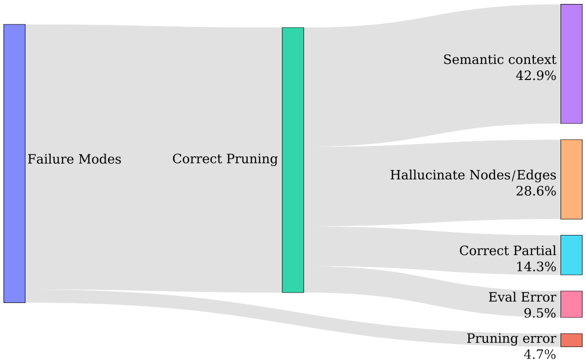

Fig. 4: Summary of failure modes of EGG with pruning and ground truth captioning.

Fig. 4 provides a summary of the potential failure modes identified during an experimental trial using ground truth event captions (shown in Tab. II). Our observations show that most errors originate from the LLM answer generation. More specifically, the primary sources of error include:

- Semantic Context: the LLM misinterprets the question or the context of an event, leading to irrelevant answers.
- Hallucinate Nodes and Edges: the LLM occasionally makes incorrect connections between nodes that do not exist in the graph or misidentifies objects and events related to the query.

In contrast, we identified only 1 instance where pruning removed a vital node related to the query, suggesting that EGG provides adequate information for reasoning tasks and effectively supports graph-based reasoning with current LLMs, and the shortcomings mostly arise from the interaction with the LLM.

To increase performance, incorporating more specialized models or integrating graph reasoning tools could help the LLM better interpret the structured information provided by EGG, thereby reducing errors and improving understanding in complex tasks.

VII. Discussions

Our experiments demonstrate that EGG's grounding of observed dynamic events with spatial features allows it to retain the capabilities of purely spatial and temporal representations while improving understanding of historical interactions within a scene. We have also shown that incorporating a graph pruning strategy to select relevant information improves the performance on the downstream information retrieval task and reduces token usage. By retaining both geometric changes and semantic context, EGG could provide valuable information for robots to execute higher level tasks, such as making more informed decisions during semantic object search [2], [6], or learning mobile manipulation skills from human demonstrations [32].

However, a notable constraint of our framework is that the subgraph $\mathcal{G}_Q$ as formulated in (1) is not guaranteed to be optimal. We observe that often the number of relevant objects typically returned can be redundant, as our heuristic approach may include irrelevant objects connected to pertinent events. This redundancy also applies to the events. Thus, this underscores the necessity for the development of more robust semantic search algorithms, especially when extending to larger, more complex environment with more events. Another constraint of EGG lies in the lack of connectivity between the recorded events. As a result, the LLM agent has to infer from scattered events and form implicit links based on the timestamps or event ordering. A hierarchical structure in the time domain, such as having summaries of multiple events [22], could further improve the performance of our framework. Moreover, the process of building EGG is currently offline, making it not yet suitable for real-time robotic applications. Currently, our method requires that any information queried be already present in the map, a limitation shared by many semantic mapping techniques. This issue is however further compounded by the addition of event memory. Efforts to overcome this limitation, such as those that focus on inferring unseen interactions and transitions [20], are crucial and hold great promise for long-term robotic operation.

VIII. Conclusion

In this work, we present EGG, a novel scene graph framework that unifies the benefits of spatial and temporal scene representations. We achieve this by grounding observed dynamic events to the scene's spatial features, resulting in a richer description of the scene that allows an off-the-shelf LLM to answer complex queries about the scene. We evaluated our method by constructing EGG from real-world robotic data, and tested its capability on a downstream information retrieval task. Our experimental results indicate that each component of EGG contributes to enriching EGG's information and allows the LLM to respond correctly to more questions. Furthermore, the experiments have shown evidence that the spatio-temporal queryability of EGG combined with a graph pruning strategy improves the result on the downstream task by removing task-irrelevant information.

Capabilities like multi-robot coordination, robot teaching by demonstration, and natural human-robot cooperation depend on the ability of robots to operate in dynamic human-inhabited environments. Real-time spatio-temporal representations supporting comprehension of physical entities, interactions between them, and other agents' actions will be instrumental in solving those tasks and deploying intelligent industrial and service robots.

IX. Acknowledgement

The authors acknowledge the use of the MIDAS infrastructure of Aalto School of Electrical Engineering. The authors also acknowledge the use of Grammarly's generative AI tool and OpenAI's GPT4o [24] for improving readability across all sections. Additionally, the authors acknowledge the role of OpenAI's GPT4o [24] in evaluating our method as detailed in Sec. VI, as well as generating documentation and improving code readability in the project's code base.

References

[1] I. Armeni *et al.*, "3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 5663–5672.

[2] Q. Gu *et al.*, "ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, May 2024, pp. 5021–5028.

[3] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A Real-time Spatial Perception System for 3D Scene Graph Construction and Optimization," in *Robotics: Science and Systems XVIII*. Robotics: Science and Systems Foundation, Jun. 2022.

[4] N. Hughes *et al.*, "Foundations of spatial perception for robotics: Hierarchical representations and real-time systems," *The International Journal of Robotics Research*, p. 02783649241229725, Feb. 2024.

[5] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, "Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation," in *Robotics: Science and Systems XX*. Robotics: Science and Systems Foundation, Jul. 2024.

[6] D. Honerkamp, M. Büchner, F. Despinoy, T. Welschehold, and A. Valada, "Language-Grounded Dynamic Scene Graphs for Interactive Object Search With Mobile Manipulation," *IEEE Robotics and Automation Letters*, vol. 9, no. 10, pp. 8298–8305, Oct. 2024.

[7] A. Anwar, J. Welsh, J. Biswas, S. Pouya, and Y. Chang, "ReMEmbR: Building and Reasoning Over Long-Horizon Spatio-Temporal Memory for Robot Navigation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, May 2025, pp. 2838–2845.

[8] L. Bärmann, F. Peller-Konrad, S. Constantin, T. Asfour, and A. Waibel, "Deep Episodic Memory for Verbalization of Robot Experience," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5808–5815, Jul. 2021.

[9] Q. Xie *et al.*, "Embodied-RAG: General Non-parametric Embodied Memory for Retrieval and Generation," Jan. 2025.

[10] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf, "SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Robot Task Planning," in *Proceedings of The 7th Conference on Robot Learning*. PMLR, Dec. 2023, pp. 23–72.

[11] A. Rosinol *et al.*, "Kimera: From SLAM to spatial perception with 3D dynamic scene graphs," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1510–1546, Dec. 2021.

[12] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual Language Maps for Robot Navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 10 608–10 615.

[13] L. Schmid *et al.*, "Panoptic Multi-TSDFs: A Flexible Representation for Online Multi-resolution Volumetric Mapping and Long-term Dynamic Scene Consistency," in *2022 International Conference on Robotics and Automation (ICRA)*. Philadelphia, PA, USA: IEEE, May 2022, pp. 8018–8024.

[14] J. Mccormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger, "Fusion++: Volumetric Object-Level SLAM," in *2018 International Conference on 3D Vision (3DV)*. Verona: IEEE, Sep. 2018, pp. 32–41.

[15] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. Davison, and S. Leutenegger, "MID-Fusion: Octree-based Object-Level Multi-Instance Dynamic SLAM," in *2019 International Conference on Robotics and Automation (ICRA)*. Montreal, QC, Canada: IEEE, May 2019, pp. 5231–5237.

[16] A. Rosinol, A. Gupta, M. Abate, J. Shi, and L. Carlone, "3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans," in *Robotics: Science and Systems XVI*. Robotics: Science and Systems Foundation, Jul. 2020.

[17] S. Saxena *et al.*, "GraphEQA: Using 3D semantic scene graphs for real-time embodied question answering," in *Proceedings of the 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Lim, S. Song, and H.-W. Park, Eds., vol. 305. PMLR, Sep. 2025, pp. 2714–2742.

[18] M. Mohammadi *et al.*, "MORE: Mobile Manipulation Rearrangement Through Grounded Language Reasoning," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2025, pp. 1805–1811.

[19] P. Nguyen, F. Verdoja, and V. Kyrki, "REACT: Real-time Efficient Attribute Clustering and Transfer for Updatable 3D Scene Graph," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2025, pp. 2209–2216.

[20] L. Schmid, M. Abate, Y. Chang, and L. Carlone, "Khronos: A Unified Approach for Spatio-Temporal Metric-Semantic SLAM in Dynamic Environments," in *Robotics: Science and Systems XX*, vol. 20, Jul. 2024.

[21] L. Ge, X. Zhu, Z. Yang, and X. Li, "DynamicGSG: Dynamic 3D Gaussian Scene Graphs for Environment Adaptation," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2025, pp. 2232–2239.

[22] L. Bärmann *et al.*, "Episodic Memory Verbalization Using Hierarchical Representations of Life-Long Robot Experience," in *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, Sep. 2025, pp. 783–790.

[23] S. Macenski and I. Jambrecic, "SLAM Toolbox: SLAM for the dynamic world," *Journal of Open Source Software*, vol. 6, no. 61, p. 2783, May 2021.

[24] OpenAI (2024), "GPT-4o System Card," Oct. 2024.

[25] Y. Yuan *et al.*, "VideoRefer Suite: Advancing Spatial-Temporal Object Understanding with Video LLM," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 18 970–18 980.

[26] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, Jun. 2015, pp. 961–970.

[27] L. Zhou, C. Xu, and J. Corso, "Towards Automatic Learning of Procedures From Web Instructional Videos," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

[28] K. Grauman *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*, 2022, pp. 18 995–19 012.

[29] H. Li *et al.*, "LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods," Dec. 2024.

[30] F. Shi *et al.*, "Large Language Models Can Be Easily Distracted by Irrelevant Context," in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Jul. 2023, pp. 31 210–31 227.

[31] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis, "Efficient Streaming Language Models with Attention Sinks," *International Conference on Representation Learning*, vol. 2024, pp. 21 875–21 895, May 2024.

[32] E. Merlo, M. Lagomarsino, E. Lamon, and A. Ajoudani, "Exploiting Information Theory for Intuitive Robot Programming of Manual Activities," *IEEE Transactions on Robotics*, vol. 41, pp. 1245–1262, 2025.