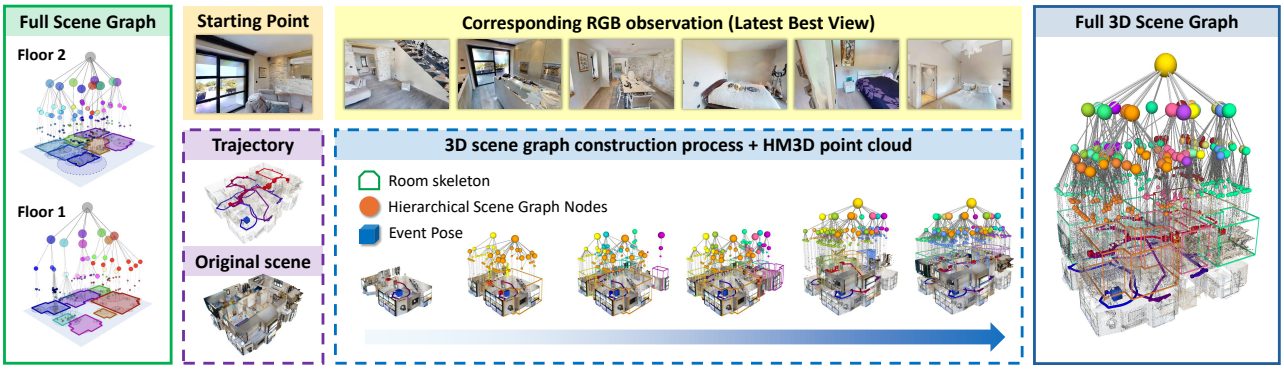
INHerit-SG: Incremental Hierarchical Semantic Scene Graphs with RAG-Style Retrieval

YukTungSamuel Fang[1], Zhikang Shi[1], Jiabin Qiu[1], Zixuan Chen[1], Jieqi Shi[1*], Hao Xu[2], Jing Huo[1*], Yang Gao[1]
[1]Nanjing University, Nanjing, China [2]Beihang University, Beijing, China
[*]Corresponding authors: isjieqi@nju.edu.cn, huojing@nju.edu.cn



***Abstract*—Driven by recent advancements in foundation models, semantic scene graphs have become a promising paradigm for high-level 3D environmental abstraction in robot mapping. However, existing frameworks struggle to handle complex embodied queries while maintaining continuous semantic graph construction. We present INHerit-SG, an asynchronous dual-stream framework that structures 3D environments into a RAG-ready knowledge base. It integrates hierarchical node representations, event-triggered asynchronous updates, and structured retrieval. Geometric segmentation is decoupled from semantic reasoning for mapping efficiency, while semantic nodes store natural language summaries for text-based retrieval. For complex queries, multi-role LLMs reason over the scene graph, followed by visual verification to mitigate false positives. Experiments on a newly constructed benchmark, HM3DSem-SQR, and real-world environments show state-of-the-art performance, especially on queries involving negations and chained spatial constraints. Project Page: https://fangyuktung.github.io/INHeritSG.github.io/**

I. Introduction

Robotic mapping has evolved from high-precision metric reconstruction toward semantic understanding and embodied AI-driven hierarchical and topological representations. Traditional systems mainly focus on geometric reconstruction for safe navigation [1], [2], [3]. However, agents operating in human environments must understand complex instructions rather than only coordinate goals. This requires an interactive and interpretable map that can resolve ambiguity, negation, and chained spatial constraints in human commands.

To serve embodied intelligence tasks, especially complex natural language queries, a robotic mapping system should be structured, semantically rich, on-the-fly, and interpretable. Existing methods still struggle to satisfy these requirements. Flat open-vocabulary maps [4], [5] lack explicit multi-level structure for complex reasoning. Structured 3D scene graphs [6], [7] often rely on heavy offline processing and are difficult to maintain on-the-fly. Online 3D scene graph systems [8], [9] usually lack language-aligned semantic richness. Recent embodied RAG (Retrieval-Augmented Generation) methods [10], [11], [12] retrieve task-relevant scene elements from memory banks, but mainly depend on dense embedding similarity, which is fragile under negation, compositional constraints, and complex logical structures. Graph prompting and QA methods such as SG-Nav [13] and GraphEQA [14] also lack an interpretable, structured retrieval mechanism for explicitly locating intermediate candidates under full instruction constraints.

To address these limitations, we propose INHerit-SG, an incremental hierarchical semantic scene graph framework designed as a systematic RAG-ready architecture. INHerit-SG organizes the environment into a four-level hierarchy of floor, room, area, and object nodes, encoding both interlayer hierarchical relations and intralayer spatial topology. Each node stores visual features, 3D information, and natural language descriptions to support text-based retrieval. For map maintenance, an asynchronous dual-stream process decouples geometric segmentation from semantic reasoning, while an event-triggered mechanism updates high-level nodes only upon significant topological changes. For retrieval, multi-role LLMs parse complex instructions into structured logical constraints and Graph-RAG [15] resolves spatial dependencies. Then, a scoring mechanism is used to filter nodes, and finally a visual verification module reduces false positives.

In summary, we make the following contributions:

1) We propose INHerit-SG, a systematic RAG-oriented architecture that bridges 3D environments and foundation models through hierarchical node representation, event-triggered updates, and structured retrieval.

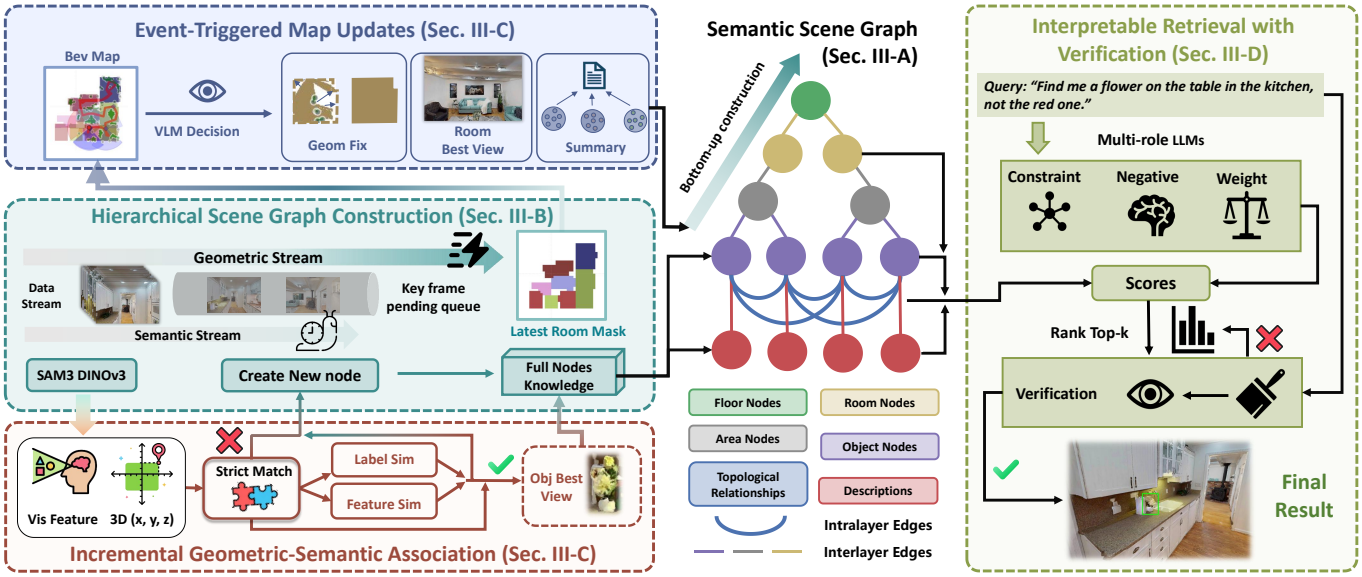


Fig. 1. The INHerit-SG Framework.

2) We present a score-based, interpretable retrieval pipeline that couples LLM-based constraint parsing, GraphRAG-driven spatial dependency retrieval, and VLM-based visual verification.
3) We construct HM3DSem-SQR, a benchmark for fine-grained retrieval, featuring logical negations, multiple spatial relationships, and complex attribute constraints.

II. Related Work

A. From Traditional Mapping to Semantic Scene Graphs

Traditional mapping and localization systems, such as KinectFusion [2], provide robust foundations for real-time spatial tracking. However, their reliance on low-level geometric primitives limits the semantic precision and object-level understanding required for complex human-robot interactions. Open-vocabulary semantic mapping addresses this limitation by projecting high-dimensional visual-language embeddings into 3D space [4], [7], [16], [17], with later works integrating instance segmentation, 3D Gaussian Splatting, spatio-temporal consistency, and metric-semantic formulations [5], [18], [19], [20], [21]. While these dense representations excel at zero-shot recognition, their lack of topological structure limits complex spatial reasoning. Structured 3D scene graphs therefore organize environments into hierarchical and actionable nodes for open-world navigation and multi-modal reasoning [8], [9], [13], [22], [23], [24].

B. RAG-Guided Retrieval and Instruction Reasoning

The utility of a semantic map depends on the agent's ability to retrieve targets and perform reasoning. Inspired by RAG in NLP [25], recent works adapt this paradigm to embodied AI by retrieving task-relevant scene elements from environmental memory [10], [26]. Other frameworks further enrich long-horizon multimodal memory, affordance-aware functional understanding, active Embodied Question Answering, and hierarchical task analysis [27], [28], [29], [30], [31]. Closely related methods apply reasoning over graph structures, using semantic graphs to ground LLMs for question answering, task planning, and logical traversal [12], [14], [15], [32], [33]. However, most existing methods fail to tightly couple perceptual mapping with retrieval, limiting their performance on complex semantic queries.

III. Technical Approach

A. Framework Overview

To support complex reasoning, we define the environment as a RAG-ready hierarchical scene graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. As shown in Fig. 2, $\mathcal{V}$ is organized into four semantic levels: floor (L_0), room (L_1), area (L_2), and object (L_3). The edge set $\mathcal{E}$ contains *interlayer edges* for hierarchical dependencies and *intralayer edges* for spatial topologies among same-level entities.

As shown in Fig. 1, INHerit-SG incrementally builds and maintains $\mathcal{G}$ from visual and depth observations. The geometric stream constructs floor and room structures on-the-fly, while the semantic stream asynchronously instantiates object nodes and updates higher-level semantic nodes. The maintained graph is then used for efficient indexing, multi-role instruction parsing, and VLM-verified 3D target localization.

B. Hierarchical Scene Graph Construction

Embodied robotic systems need to balance real-time performance and high-level accuracy, while LLM-based semantic understanding often struggles to guarantee efficiency. Therefore, inspired by the fast-slow system strategy [34], we design a dual-stream construction process (Fig. 3). The geometric stream builds the structural layers of the environment, while the semantic stream identifies object-level nodes from visual observations and geometric context.

Geometric Stream: Dense Topology and Keyframe Gating (L_1, L_0). The geometric stream continuously integrates RGB-D observations into a 2D Bird's-Eye View occupancy grid via depth raycasting. Masks of room nodes (L_1) are segmented from accumulated free space using Euclidean Distance

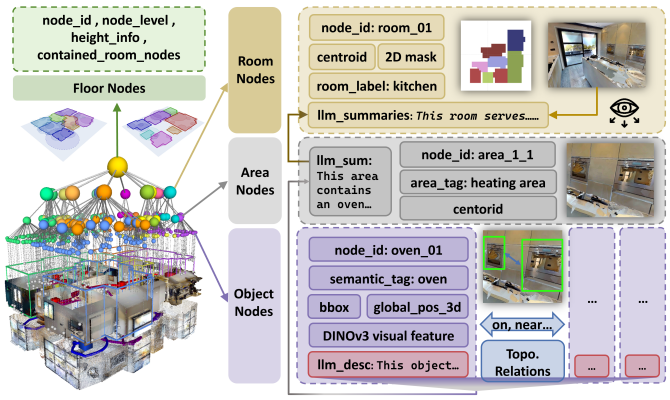


Fig. 2. INHerit-SG semantic scene graph.

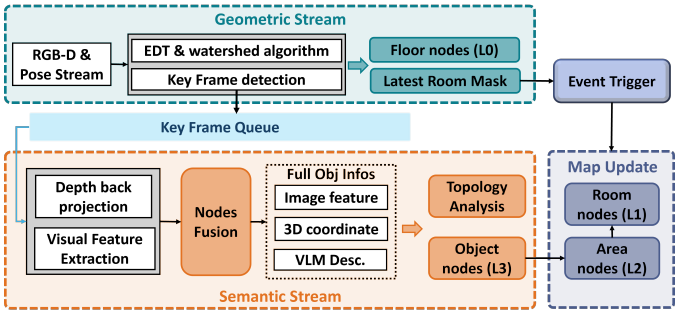


Fig. 3. Dual-Stream Construction Pipeline.

Transform (EDT) and watershed, while detected doors and smoothed distance fields reduce room fragmentation. Vertical motion is monitored to instantiate floor nodes (L_0).

To regulate the computational load of the semantic stream, we introduce visual gating. DINOv3 [35] features are extracted from incoming frames and compared with the last keyframe. Once the similarity drops below a predefined threshold, the frame is pushed into a pending queue. This allows the semantic stream to analyze only informative keyframes, while keeping geometric tracking continuous at about 2 Hz.

Semantic Stream: Object, Area, and Room Node Instantiation (L_3, L_2, L_1). The semantic stream operates asynchronously on the pending queue to instantiate object nodes (L_3). For each keyframe, SAM3 [36] generates instance masks, whose centroids are back-projected into 3D space. DINOv3 features are extracted from the masks for object matching across frames. After node fusion (Sec. III-C), each object node stores its ID, label, visual features, 3D coordinates, and description.

Higher-level area and room nodes are constructed bottom-up only upon global update events. During an update (Sec. III-C), object coordinates are aligned with the latest room masks to determine room affiliations. Within each room, objects are clustered into area nodes (L_2), and an LLM derives their functional labels from textual semantics. For room-level descriptions, the system selects a geometrically optimal viewpoint and combines it with area summaries for VLM-based summarization.

The final hierarchy is serialized into directed graphs and structured tables, providing a lightweight representation compatible with language-driven RAG retrieval.

C. Node Association and Map Update

After graph initialization, INHerit-SG maintains the scene graph on-the-fly through incremental node association and event-triggered structural updates.

Incremental Geometric-Semantic Association. Merging new observations into stable graph nodes is critical to prevent semantic drift and redundancy. Inspired by human cognition, INHerit-SG associates objects using spatial distance, visual appearance, and semantic similarity, measured by 3D Euclidean distance, DINOv3 feature similarity, and SentenceTransformer embedding similarity, respectively.

We adopt a two-stage fusion strategy. Confident observations with high spatial and visual consistency are directly matched to existing nodes, while ambiguous cases are resolved according to object type, semantic consistency, and spatial distance. Matched nodes update their visual features and 3D coordinates using a weighted moving average, while unmatched observations instantiate new nodes with VLM-generated descriptions.

After object association, nearby objects are clustered into area nodes to support relational queries. Pairwise spatial relations are inferred using either a geometric mode based on 3D bounding box offsets and vertical proximity, or a VLM mode based on annotated RGB images. The validated relations are inserted into the global spatial graph.

Event-Triggered Map Updates. To avoid redundant computation, INHerit-SG updates high-level summaries only when meaningful scene changes are detected. A *hard trigger* is activated by discrete state changes such as floor transitions, while a *soft trigger* is invoked by a VLM-based supervisor.

The soft trigger operates on a dynamically generated BEV map containing room masks, the current trajectory, and historical update points. The VLM uses this map to determine whether the previous summary is obsolete or whether a newly explored area requires summarization. Once triggered, the system optimizes global room masks by removing noisy regions, regularizing room boundaries, and resolving overlapping boundaries via EDT, and summarizes the room details based on the optimal room perspective. Objects are then reassigned, and higher-level area and room nodes are reconstructed in a bottom-up manner.

D. RAG-Style Retrieval with Verification

Robust retrieval is essential for compositional queries, especially with negation and relational constraints. Inspired by neuro-symbolic reasoning and GraphRAG [8], [12], INHerit-SG integrates instruction parsing, intent-weighted scoring, and graph topology traversal to resolve complex queries.

Instruction Parsing. We parse each user query into structured constraints using a multi-role LLM architecture. The parser decomposes the instruction into target objects, reference landmarks, and spatial requirements, while extracting negative constraints and assigning dynamic importance scores.

Candidate Retrieval. The parsed constraints are used to rank candidate nodes. Floor ID is treated as a binary hard

TABLE I
QUANTITATIVE COMPARISON ON HM3DSEM-SQR AND REAL-WORLD DATA. - DENOTES UNSUPPORTED TESTS.

Method	HM3DSem-SQR Accuracy (%) ↑										Semantic Acc (%) ↑			Real-World Exp Acc (%) ↑		
	Within 1m					Within 0.5m					Human Evaluation			Human Evaluation		
	ABC	D	E	F	Avg	ABC	D	E	F	Avg	QLCR	AHAR	ALMA	Simple	Complex	Avg
ConceptGraphs	22.84	14.79	21.54	20.22	19.95	21.62	14.30	21.38	18.99	19.03	66.1	54.2	59.8	27.3	44.4	35.0
Embodied-RAG	24.80	19.33	25.33	21.29	22.58	18.28	15.09	18.92	15.9	16.95	62.2	50.4	49.9	18.2	44.4	20.0
HOV-SG	27.0	31.6	34.7	28.5	29.40	20.32	23.07	25.33	22.01	21.94	–	–	–	–	–	–
DualMap	36.52	25.89	36.02	33.88	33.02	**30.78**	22.21	**31.58**	28.34	28.01	–	–	–	–	–	–
INHerit-SG (Ours)	**37.7**	**32.3**	**41.1**	**36.6**	**36.3**	30.1	**25.6**	30.9	**29.6**	**28.9**	**79.7**	**72.9**	**70.6**	**54.5**	**66.7**	**60.0**

filter ($H_{floor} \in \{0, 1\}$), while other constraints are combined as soft filters:

$$S(n) = H_{floor} \cdot \sum_{i=1}^{K} p_i \cdot w_i \cdot \text{Sim}(n, c_i), \tag{1}$$

where n is the candidate node, K is the number of constraints, w_i denotes the importance weight of constraint c_i, and p_i controls its polarity for penalizing negated constraints. $Sim(n, c_i)$ denotes the semantic similarity score between the constraints of candidate node n and constraint c_i. Specifically, for spatial constraints, INHerit-SG integrates GraphRAG techniques to traverse the spatial topology graph, thereby generating scores based on the relative spatial relationships between objects.

Visual Verification. To reduce false positives, top-ranked candidates undergo visual verification. The target object label is used as a text prompt for SAM3, and the best-view image of the object with the SAM3 bounding box is then passed to a verification VLM. It validates whether the candidate satisfies the query constraints. Once verified, the system outputs the 3D centroid coordinates for downstream application.

IV. EXPERIMENTAL EVALUATION

We evaluate INHerit-SG on HM3DSem-SQR and real-world trajectories, focusing on retrieval accuracy under complex semantic constraints. The evaluation compares INHerit-SG with recent semantic mapping and retrieval baselines using both geometric and human-evaluated semantic metrics.

A. Dataset and Baselines

Simulation Dataset. To test whether semantic maps can support complex logical queries, we construct **HM3DSem-SQR** from HM3D-Sem [37]. The dataset focuses on compositional reasoning rather than simple object recall. We use human expert operations to generate realistic exploration trajectories with synchronized sensor streams, and manually construct 36 trajectories and 6084 indexed instructions. These instructions cover basic spatial relations (A–C, ≈40.4%), negation queries (D, ≈26.8%), descriptive queries (E, ≈10.0%), and ambiguous instructions (F, ≈22.8%). They evaluate whether a map is structured, semantically rich, and interpretable.

Real-world Dataset. We also collect data from three real-world environments and design 80 queries. Success rates are evaluated through manual assessment. Camera trajectories are obtained from the front-end SLAM system, while depth is computed from a Livox LiDAR to provide RGB-D streams and poses.

Baselines. We compare INHerit-SG with four state-of-the-art methods: **ConceptGraphs** [4], **Embodied-RAG** [11], **HOV-SG** [6], and **DualMap** [5]. These baselines cover flat point-cloud maps, retrieval-based memory for RAG, offline hierarchical graphs, and SLAM-centric semantic mapping. All map constructions are performed on a single RTX 4090 GPU with cloud-based GPT-4o inference.

B. Retrieval Accuracy

This experiment evaluates whether our representation and retrieval design improve reliability under complex semantic constraints. We use two metrics: **Geometric Accuracy**, which checks whether the retrieved object lies within a distance threshold of the ground truth, and **Semantic Accuracy**, which checks whether the object satisfies the instruction, and is evaluated through a human study with 120 participants.

As shown in Table I, INHerit-SG outperforms all baselines under the 1.0m geometric threshold, following the embodied task metric in [38]. It also maintains clear advantages at 0.5m on challenging queries such as negation and ambiguous semantics, while remaining competitive on easier cases. These results show that decomposing structured constraints improves retrieval reliability even without storing dense point clouds.

For human evaluation, we report three metrics: **QLCR**, which measures majority-vote correctness per query; **AHAR**, which measures the global human acceptance rate; and **ALMA**, which averages annotator-level pass rates. INHerit-SG achieves stronger human-evaluated semantic accuracy, showing that its retrieved targets better match user intent. On real-world data, it also demonstrates clear advantages, indicating its adaptability to noisy real environments.

V. CONCLUSION

We presented **INHerit-SG**, an incremental hierarchical semantic scene graph with RAG-style retrieval. By formulating the hierarchical scene graph as RAG-ready memory, INHerit-SG bridges geometric mapping with language-driven reasoning. It maintains the graph through event-triggered updates and retrieves targets through logical parsing, RAG-style scoring, and visual verification. Experiments on HM3DSem-SQR and real-world sequences show that INHerit-SG handles complex instructions more reliably than existing baselines. Future work will extend the framework to highly dynamic layouts, frequent object rearrangements, life-long scenarios, and complex mobile manipulation tasks.

References

[1] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, et al., "Elasticfusion: Dense slam without a pose graph," 2015.

[2] R. A. Newcombe, S. Izadi, O. Hilliges, et al., "Kinectfusion: Real-time dense surface mapping and tracking," 2011, pp. 127–136.

[3] C. Campos, R. Elvira, J. J. G. Rodriguez, et al., "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[4] Q. Gu, A. Kuwajerwala, S. Morin, et al., *Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning*, 2023.

[5] J. Jiang, Y. Zhu, Z. Wu, et al., "Dualmap: Online open-vocabulary semantic mapping for natural language navigation in dynamic changing scenes," *IEEE Robotics and Automation Letters*, vol. 10, no. 12, pp. 12 612–12 619, Dec. 2025.

[6] A. Werby, C. Huang, M. Büchner, et al., "Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation," ser. RSS2024, Robotics: Science and Systems Foundation, Jul. 2024.

[7] C. Huang, O. Mees, A. Zeng, et al., "Visual language maps for robot navigation," London, UK, 2023.

[8] N. Hughes, Y. Chang, and L. Carlone, *Hydra: A real-time spatial perception system for 3d scene graph construction and optimization*, 2022.

[9] D. Maggio, Y. Chang, N. Hughes, et al., *Clio: Real-time task-driven open-set 3d scene graphs*, 2024.

[10] M. Booker, G. Byrd, B. Kemp, et al., *Embodiedrag: Dynamic 3d scene graph retrieval for efficient and scalable robot task planning*, 2024.

[11] Q. Xie, S. Y. Min, P. Ji, et al., *Embodied-rag: General non-parametric embodied memory for retrieval and generation*, 2025.

[12] K. Rana, J. Haviland, S. Garg, et al., *Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning*, 2023.

[13] H. Yin, X. Xu, Z. Wu, et al., *Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation*, 2024.

[14] S. Saxena, B. Buchanan, C. Paxton, et al., *Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering*, 2025.

[15] D. Edge, H. Trinh, N. Cheng, et al., *From local to global: A graph rag approach to query-focused summarization*, 2025.

[16] S. Peng, K. Genova, C. M. Jiang, et al., "Openscene: 3d scene understanding with open vocabularies," 2023.

[17] J. Kerr, C. M. Kim, K. Goldberg, et al., "Lerf: Language embedded radiance fields," 2023.

[18] A. Takmaz, E. Fedele, R. W. Sumner, et al., "OpenMask3D: Open-Vocabulary 3D Instance Segmentation," 2023.

[19] P. D. A. Nguyen, T. D. Ngo, E. Kalogerakis, et al., *Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance*, 2024.

[20] Y. Deng, Y. Yue, J. Dou, et al., *Omnimap: A general mapping framework integrating optics, geometry, and semantics*, 2025.

[21] L. Schmid, M. Abate, Y. Chang, et al., *Khronos: A unified approach for spatio-temporal metric-semantic slam in dynamic environments*, 2024.

[22] C. Kassab, M. Mattamala, S. Morin, et al., *The bare necessities: Designing simple, effective open-vocabulary scene graphs*, 2024.

[23] J. Loo, Z. Wu, and D. Hsu, "Open scene graphs for open-world object-goal navigation," *The International Journal of Robotics Research*, Oct. 2025.

[24] X. Zhou, T. Xiao, L. Liu, et al., *Fsr-vln: Fast and slow reasoning for vision-language navigation with hierarchical multi-modal scene graph*, 2025.

[25] P. Lewis, E. Perez, A. Piktus, et al., *Retrieval-augmented generation for knowledge-intensive nlp tasks*, 2021.

[26] Y. Chang, R. Chen, Z. Zhang, et al., *Rag-3dsg: Enhancing 3d scene graphs with re-shot guided retrieval-augmented generation*, 2026.

[27] M. Yuan, H. Zhang, M. Mohammadi, et al., *Star: Scalable task-conditioned retrieval for long-horizon multi-modal robot memory*, 2026.

[28] A. Anwar, J. Welsh, J. Biswas, et al., *Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation*, 2024.

[29] R. Korekata, Q. Xie, Y. Bisk, et al., *Affordance rag: Hierarchical multimodal retrieval with affordance-aware embodied memory for mobile manipulation*, 2025.

[30] M. F. Ginting, D.-K. Kim, X. Meng, et al., *Enter the mind palace: Reasoning and planning for long-term active embodied question answering*, 2025.

[31] Y. Chang, L. Fermoselle, D. Ta, et al., *Ashita: Automatic scene-grounded hierarchical task analysis*, 2025.

[32] Z. Guo, L. Xia, Y. Yu, et al., "Lightrag: Simple and fast retrieval-augmented generation," 2024.

[33] A. Ray, J. Arkin, H. Biggie, et al., *Structured interfaces for automated reasoning with 3d scene graphs*, 2025.

[34] T. Anthony, Z. Tian, and D. Barber, "Thinking fast and slow with deep learning and tree search," vol. 30, 2017.

[35] O. Siméoni, H. V. Vo, M. Seitzer, et al., *DINOv3*, 2025.

[36] N. Carion, L. Gustafson, Y.-T. Hu, et al., *Sam 3: Segment anything with concepts*, 2025.

[37] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, et al., "Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI," 2021.

[38] X. Puig, E. Undersander, A. Szot, et al., *Habitat 3.0: A co-habitat for humans, avatars and robots*, 2023.