Interactive Open-Set Semantic Mapping with a 3D Scene Graph Backend

Felix Igelbrink[1] , **Lennart Niecksch**[2,1] , **Martin Günther**[1] , **Marian Renz**[2,1] , **Oscar Lima**[1] and **Martin Atzmueller**[2,1]

[1]German Research Center for Artificial Intelligence (DFKI), Osnabrück, Germany
[2]Osnabrück University, Semantic Information Systems Group, Osnabrück, Germany
{felix.igelbrink, lennart.niecksch, martin.guenther, marian.renz, oscar.lima}@dfki.de,
martin.atzmueller@uos.de

Abstract

While Open-Set Semantic Mapping and 3D Semantic Scene Graphs (3DSSGs) have become established paradigms in robotic perception in recent years, most existing works are limited to small environments or sacrifice geometric detail and instance granularity for scalability. Deploying these systems at scale for large multi-room environments remains a major challenge due to the computational overhead of high-dimensional feature integration and the maintenance of the 3DSSG structure. In this paper, we demonstrate a modular mapping architecture that establishes 3D Semantic Scene Graphs (3DSSGs) as its foundational backend. Unlike approaches that generate scene graphs as a post-processing step, our system maintains the graph as the primary, incrementally updated knowledge representation. Our architecture is optimized for GPU-accelerated operations, enabling the dense representation of extensive environments containing thousands of unique object instances, supporting open-vocabulary queries via CLIP features without requiring any additional post-processing steps. In this live demonstration, we showcase our pipeline processing large-scale data from the Habitat Matterport 3D (HM3D) dataset as well as live data collected from a handheld device. Attendees will interact with the generated maps by performing real-time, open-set queries (e. g., "find the vintage wooden chair") across complex, multi-story environments, highlighting the system's capability to represent dynamic, human-aligned environmental understanding suitable for downstream robotic tasks.

Demo video: https://youtu.be/ZpUPamQJ18c

1 Introduction

Autonomous mobile robots operating in complex, large-scale environments require a semantic understanding that extends beyond pure geometric occupancy to perform high-level tasks [Nüchter and Hertzberg, 2008; Igelbrink *et al.*, 2024]. While volumetric representations effectively model surface geometry, they often lack the structural information needed for actual scene understanding. 3D Semantic Scene Graphs (3DSSGs) have emerged as a powerful solution to bridge this gap between sub-symbolic sensor data and symbolic reasoning by representing environments as hierarchical graphs of entities and relationships [Armeni *et al.*, 2019; Rosinol *et al.*, 2020]. Recent works combine the structure of 3DSSGs with the flexibility of open-set embeddings like Contrastive Language-Image Pretraining (CLIP) [Radford *et al.*, 2021], representing instances as feature embeddings rather than a closed semantic vocabulary. However, the field of open-set mapping currently faces a significant scalability gap. Most existing works [Gu *et al.*, 2024; Linok *et al.*, 2025; Peng and Genova, 2023; Jatavallabhula *et al.*, 2023] target small-scale, single-room environments, often requiring significant offline processing. Conversely, frameworks that scale to larger environments typically sacrifice geometric detail or the granularity of individual object instances [Hughes *et al.*, 2022; Maggio *et al.*, 2024; Laina *et al.*, 2025]. Furthermore, many systems treat the scene graph as a mere post-hoc derivative layer, preventing real-time interaction with the map in e. g., exploration or continuous mapping scenarios.

In this system demonstration, we present a mapping architecture with a 3DSSG as the foundational backend for the entire mapping process – based on our recent work [Günther *et al.*, 2026; Igelbrink *et al.*, 2026], which details the theoretical underpinnings. Our system is specifically designed to address the challenges of large-scale open-set mapping via the following key features: (1) An architecture optimized for GPU-accelerated operations, capable of managing thousands of unique, spatially grounded object instances across large, multi-story environments. (2) The system maintains the 3DSSG as its primary, incrementally updated knowledge representation – valid at any time step without requiring post-processing. (3) We automatically integrate well localized CLIP features directly into the graph nodes during the mapping process, enabling immediate natural language queries.

Our live demonstration focuses on our system's interactive capabilities by processing extensive data from the Matterport-Habitat (HM3D) dataset [Ramakrishnan *et al.*, 2021] alongside real-time streams from a handheld device, showcasing how it fuses high-bandwidth sensor data into a consistent semantic knowledge base that attendees can query on the fly.

2 System Architecture

The core of our demonstration is a modular, GPU-accelerated pipeline designed to build and maintain the underlying 3DSSG as the foundational representation for the mapping life-cycle and all integrated data. The system takes as input a stream of RGB-D frames and their poses. The system is engineered for high throughput on the GPU, enabling the processing of large-scale environments like HM3D in real-time.

2.1 Scalable Scene Graph Backend

To manage thousands of unique instances across large-scale environments, the backend is organized into three hierarchical abstraction layers:

1. **Frames Layer**: Tracks 6D poses and raw 2D segmentation masks for all processed frames.
2. **Segments Layer**: An intermediate GPU-resident layer tracking active 3D object fragments and their visual descriptors used for associating 2D detections to 3D segments.
3. **Objects Layer**: The persistent global representation where fragments are consolidated into distinct object instances with accumulated geometric and CLIP features.

This structure maintains additional efficient spatial indices ensuring it remains computational efficient and queryable even as the number of nodes grows very large.

2.2 Incremental Open-Set Perception

Our pipeline avoids expensive post-processing by inferring and integrating open-vocabulary features directly during the mapping process instead of inferring them post-hoc from all processed frames.

Segmentation and Feature Extraction. For each frame, we use *FastSAM* [Zhao *et al.*, 2023] for segmentation and extract a dense feature maps using the *DINOv2* [Oquab *et al.*, 2024] and *openCLIP* [Ilharco *et al.*, 2021] models. Normally, both models generate only a single feature vector per image, which requires a costly cropping strategy to obtain per-mask features [Kassab *et al.*, 2024]. To avoid this, we modify both models based on the MaskCLIP [Dong *et al.*, 2023] approach: we compute local per-patch features from earlier transformer layers rather than the final output, allowing the models to run just once per frame.

In order to ensure high-quality semantic grounding, we modulate the local CLIP features with their global feature embeddings. The resulting features are then weighted using a *view quality score* based on geometric properties, ensuring that only good observations of the instance contribute to the persistent node feature.

Consistent Integration and Refinement. The transition from raw sensor data to a topologically valid scene graph is performed via a two-stage matching process optimized for the GPU:

1. **Greedy Association**: Local segments are matched against the global graph using 3D IoU and feature similarity, allowing for early integration of clearly visible segments and creation of new instances.
2. **Active Refinement**: As a greedy merging approach alone often leads to over-segmentation and the accumulation of spurious artifacts due to temporal inconsistencies in the SAM predictions, we re-evaluate modified nodes within their spatial neighborhood in an additional refinement stage with respect to their more accurate voxel overlaps. This process effectively resolves conflicts in the instances due to over- or under-segmentation and fuses fragmented segments into coherent object instances, maintaining topological consistency during exploration of the scene.

The incremental design of our mapping pipeline ensures that the 3DSSG is valid at any discrete time step, providing a stable foundation for interactive open-set queries without requiring further refinement steps.

3 Demonstration Scenario

The primary goal of this demonstration is to showcase the real-time capabilities and scalability of our presented 3DSSG backend in large, multi-story environments. The demonstration setup consists of a high-performance mobile workstation (NVIDIA RTX 4090 GPU) and a handheld RGB-D device (e. g., Orbbec Femto Bolt, Azure Kinect or iPad). Our demonstration is divided into three parts in order to highlight and showcase its main features, including large-scale processing, semantic interactivity, and live flexibility.

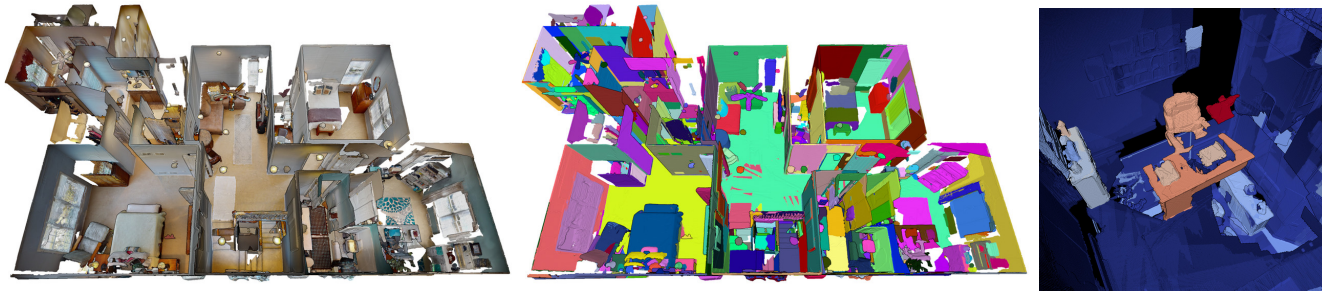
3.1 Large-Scale Incremental Mapping

In the first part, we demonstrate the system's ability to handle the high density of instances found in the HM3D dataset. Our pipeline processes multi-room sequences from HM3D at 3–5 Hz, incrementally building a scene graph that remains consistent and fast even as thousands of objects are instantiated and tracked, see Figures 1-2 for illustrating examples. Attendees will observe:

- **2D Segmentation and Feature Extraction:** The raw RGB-D feed from the dataset is segmented in real time using the FastSAM Model and sharp local CLIP and DINO features are extracted.
- **Real-Time Instance Refinement:** A split-screen visualization shows how over-segmented fragments are fused into coherent semantic objects through our GPU-accelerated active refinement stage.
- **Scalable Visualization:** The 3D view displays the global 3DSSG, where color-coded instances represent the system's current "source of truth" regarding the environment's structure.

3.2 Interactive Querying

Once the environment is mapped, we invite attendees to interact directly with the generated representation, enabled by interactive and customized querying. Attendees can type natural language queries, such as *"vintage wooden chair"* or *"fire extinguisher"*, see Figure 1c for an example. Attendees will observe:



(a) RGB Reconstruction (b) Object instances (c) Query: “A workplace”

Figure 1: Demonstration of the mapping results on the HM3D dataset. (a) Fully reconstructed geometric 3D map. (b) Individually tracked instances in the map. (c) Visualization of the system’s response to an example natural language query “A workplace”. Brighter colors indicate higher cosine similarity between the query and the graph nodes, illustrating the system’s ability to ground semantic concepts in real-time.

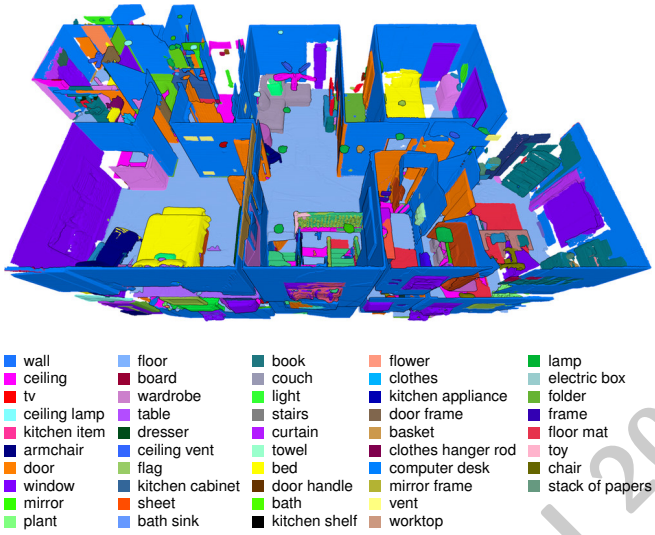


Figure 2: Open-Set semantic segmentation results based on the tracked CLIP features.

- **Instant Semantic Grounding:** The system computes the cosine similarity between the query embedding and the CLIP features stored in the graph nodes, instantly highlighting matching entities in the 3D map.
- **Graph Inspection:** Users can click on individual 3D objects to inspect the underlying scene graph structure, including connectivity and visibility of the object during the mapping process.

3.3 Handheld Live Mapping

To showcase the robustness of the pipeline beyond static, synthetic datasets, we will showcase our mapping pipeline on real-world data, which we capture on site. This part demonstrates the system’s readiness for real-world robotic deployment, where perception and open-set reasoning must operate in tandem. Attendees will observe:

- **Live Data Fusion:** Using a handheld RGB-D device, we capture a live dataset of the demonstration area, which is integrated into a scene graph.
- **Dynamic Updates:** Attendees can see how new observations update the tracked segments and objects in the 3DSSG in real-time during the mapping.

4 Conclusion

In this demonstration, we presented a mapping architecture that bridges the gap between raw sensor data and semantic reasoning by establishing the 3D Semantic Scene Graph (3DSSG) as its foundational backend. By focusing on incremental integration instead of post-processing workflows and utilizing a GPU-accelerated pipeline, we show that open-set semantic mapping can scale to large, multi-story environments containing thousands of unique object instances without sacrificing real-time performance. The three-part demonstration highlights the practical utility of our approach:

1. Efficiently processing large-scale datasets,
2. enabling immediate natural language interaction through integrated CLIP features, and
3. maintaining robustness on real-world data.

Overall, our system ensures that the 3DSSG remains the incrementally updated “source of truth”, providing a consistent and human-aligned knowledge base at every time step. We believe that this high-performance, explicit structure is a crucial enabler for next-generation hybrid intelligence systems. Here, perception and reasoning must operate in tandem to tackle unlabeled, complex real-world scenarios with enhanced interpretability and alignment to human concepts, as well as to connect to knowledge-grounded representations, e.g., [Renz *et al.*, 2025].

Acknowledgements

This work is supported by the ExPrIS, LIEREx and AMENABLE projects through grants from the German Federal Ministry of Research, Technology and Space (BMFTR) with grant numbers 16IW23001, 16IW24004 and 16IW26002. The DFKI Niedersachsen (DFKI NI) is sponsored by the Ministry of Science and Culture of Lower Saxony and the Volkswagen Foundation.

References

[Armeni *et al.*, 2019] Iro Armeni, Zhi-Yang He, Junyoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3D scene graph: A structure for unified semantics, 3D space, and camera. In *ICCV*, pages 5664–5673, 2019.

[Dong *et al.*, 2023] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. MaskCLIP: Masked self-distillation advances contrastive language-image pretraining. In *CVPR*, pages 10995–11005. IEEE, 2023.

[Gu *et al.*, 2024] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In *ICRA*, pages 5021–5028, May 2024.

[Günther *et al.*, 2026] Martin Günther, Felix Igelbrink, Oscar Lima, Lennart Niecksch, Marian Renz, and Martin Atzmueller. A scene graph backed approach to open set semantic mapping. In *AAAI Spring Symposium: Machine Learning and Knowledge Engineering for Knowledge-Grounded Semantic Agents (MAKE)*, 2026.

[Hughes *et al.*, 2022] Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception system for 3D scene graph construction and optimization. In *Robotics: Science and Systems (RSS)*, 2022.

[Igelbrink *et al.*, 2024] Felix Igelbrink, Marian Renz, Martin Günther, Piper Powell, Lennart Niecksch, Oscar Lima, Martin Atzmueller, and Joachim Hertzberg. Online knowledge integration for 3D semantic mapping: A survey. *arXiv preprint arXiv:2411.18147*, 2024.

[Igelbrink *et al.*, 2026] Felix Igelbrink, Lennart Niecksch, Marian Renz, Martin Günther, and Martin Atzmueller. LIEREx: Language-image embeddings for robotic exploration. *Künstl Intell*, 2026.

[Ilharco *et al.*, 2021] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. https://doi.org/10.5281/zenodo.5143773, 2021.

[Jatavallabhula *et al.*, 2023] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd. Omama, Ganesh Iyer, Soroush Saryazdi, Tao Chen, Alaa Maalouf, Shuang Li, Nikhil Varma Keetha, et al. ConceptFusion: Open-set multimodal 3D mapping. In *Robotics: Science and Systems (RSS)*, 2023.

[Kassab *et al.*, 2024] Christina Kassab, Matías Mattamala, Sacha Morin, Martin Büchner, Abhinav Valada, Liam Paull, and Maurice F Fallon. The bare necessities: Designing simple, effective open-vocabulary scene graphs. *arXiv preprint arXiv:2412.01539*, 2024.

[Laina *et al.*, 2025] Sebastián Barbas Laina, Simon Boche, Sotiris Papatheodorou, Simon Schaefer, Jaehyung Jung, and Stefan Leutenegger. FindAnything: Open-vocabulary and object-centric mapping for robot exploration in any environment. *arXiv preprint arXiv:2504.08603*, 2025.

[Linok *et al.*, 2025] Sergey Linok, Tatiana Zemskova, Svetlana Ladanova, Roman Titkov, Dmitry Yudin, Maxim Monastyrny, and Aleksei Valenkov. Beyond bare queries: Open-vocabulary object grounding with 3D scene graph. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 13582–13589. IEEE, 2025.

[Maggio *et al.*, 2024] Dominic Maggio, Yun Chang, Nathan Hughes, Matthew Trang, Dan Griffith, Carlyn Dougherty, Eric Cristofalo, Lukas Schmid, and Luca Carlone. Clio: Real-time task-driven open-set 3D scene graphs. *IEEE Robotics Autom. Lett.*, 9(10):8921–8928, 2024.

[Nüchter and Hertzberg, 2008] Andreas Nüchter and Joachim Hertzberg. Towards semantic maps for mobile robots. *Robot Auton Syst*, 56(11):915–926, 2008.

[Oquab *et al.*, 2024] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024.

[Peng and Genova, 2023] Songyou Peng and Kyle Genova. OpenScene: 3D Scene Understanding With Open Vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[Ramakrishnan *et al.*, 2021] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-Matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI. *arXiv preprint arXiv:2109.08238*, 2021.

[Renz *et al.*, 2025] Marian Renz, Felix Igelbrink, and Martin Atzmueller. Integrating prior observations for incremental 3D scene graph prediction. In *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2025.

[Rosinol *et al.*, 2020] Antoni Rosinol, Arjun Gupta, Marcus Abate, Jingnan Shi, and Luca Carlone. 3D dynamic scene graphs: Actionable spatial perception with places, objects, and humans. *Robotics: Science and Systems (RSS)*, 2020.

[Zhao *et al.*, 2023] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast Segment Anything. *arXiv preprint arXiv:2306.12156*, 2023.