# LIEREx: Language-Image Embeddings for Robotic Exploration

Felix Igelbrink[1*†], Lennart Niecksch[2,1†], Marian Renz[2,1], Martin Günther[1], Martin Atzmueller[2,1]

[1]German Research Center for Artificial Intelligence (DFKI), Research Department Cooperative and Autonomous Systems (CAS), Osnabrück, Germany.
[2]Osnabrück University, Semantic Information Systems Group, Osnabrück, Germany.

*Corresponding author(s). E-mail(s): felix.igelbrink@dfki.de;
Contributing authors: lennart.niecksch@dfki.de; marian.renz@dfki.de;
martin.guenther@dfki.de; martin.atzmueller@uos.de;
†These authors contributed equally to this work.

**Abstract**

Semantic maps allow a robot to reason about its surroundings to fulfill tasks such as navigating known environments, finding specific objects, and exploring unmapped areas. Traditional mapping approaches provide accurate geometric representations but are often constrained by pre-designed symbolic vocabularies. The reliance on fixed object classes makes it impractical to handle out-of-distribution knowledge not defined at design time. Recent advances in Vision-Language Foundation Models, such as CLIP, enable open-set mapping, where objects are encoded as high-dimensional embeddings rather than fixed labels. In LIEREx, we integrate these VLFMs with established 3D Semantic Scene Graphs to enable target-directed exploration by an autonomous agent in partially unknown environments.

**Keywords:** 3D Semantic Scene Graphs, Vision-Language Models, Open-Set Semantic Mapping, Active Perception, Robotic Exploration

## 1 Introduction

Autonomous mobile robotic agents operating within partially or entirely unknown environments require a high degree of scene understanding to ensure effective and safe operation. This fundamental capability relies primarily on the construction of a semantic map from sensor data, combining information about the geometry of the environment with details about the objects contained therein, such as their classes and properties [1]. Traditional semantic mapping systems predominantly rely upon rigid, closed symbolic vocabularies, typically defined by a fixed set of object classes. This limitation restricts the system's flexibility and scalability, making it impractical to handle generic knowledge or concepts not explicitly predefined at design time.

A significant methodological advancement has emerged through the development of Vision-Language Foundation Models (VLFMs), enabling the effective combination of open vocabularies with multimodal visual data. The Contrastive Language-Image Pretraining (CLIP) model [2] is a prominent example of such VLFMs. These models are trained on vast datasets of image-text pairs, allowing them to learn rich multimodal features and map visual concepts and natural language into a joint feature space. Consequently, this capability enables open-set semantic mapping, where objects are represented not by fixed labels, but by high-dimensional feature embeddings.

The integration of these VLFMs into semantic mapping architectures introduces sophisticated querying capabilities that significantly surpass the limitations of conventional, vocabulary-constrained methods. This fosters novel applications, particularly in target-directed exploration and persistent surveillance within dynamically changing or partially explored environments. Utilizing open-set semantic queries, robotic agents gain the capacity to interpret and search for arbitrary objects and abstract concepts using natural language.

Within the LIEREx[1] project (Language-Image Embeddings for Robotic Exploration), we are investigating the integration of these VLFM advancements with hybrid machine learning methods to advance semantic mapping and autonomous exploration in mobile robotics. LIEREx leverages the advantages of open-set map representations alongside the existing spatial reasoning capabilities of popular 3D Semantic Scene Graph (3DSSG)-based semantics. Specifically, the project focuses on dynamically generating exploration strategies and deriving optimal observation poses using a neural network-based estimation system. This approach facilitates the efficient verification of search and exploration results, allowing the system to move beyond relying solely on the language query itself.

This endeavor is closely coupled with the ExPrIS project (Knowledge-Level Expectations as Priors for Object Interpretation from Sensor Data), sharing infrastructure and the critical adoption of the 3DSSG as the foundational environment representation.

The remainder of this report is organized as follows: Section 2 presents the open-set mapping approach as the foundation of the architecture, followed by the exploration planning system in Section 3. Section 4 describes the robotic demonstrator. Finally, we conclude with a discussion of key technical insights in Section 5.

## 2 Open Set Semantic Mapping

Both LIEREx and ExPrIS contribute towards a common, unified semantic framework. While ExPrIS focuses on deriving a structured representation for integration with existing structured knowledge, LIEREx extends and generalizes this goal by incorporating modern VLMs. This allows the system to exploit common sense knowledge already embedded within language concepts in a semantic map. Specifically, LIEREx aims to enhance object retrieval capabilities by not only providing the location or geometry corresponding to a query but also directly generating suitable observation poses where the queried object is likely to be encountered. These poses can then be utilized by a robotic agent to search for the requested object or location. The basic pipeline of LIEREx is shown in Fig. 1. Both projects share the 3DSSG as their foundational representation and are designed to be interoperable, combining symbolic knowledge (ExPrIS) with multimodal open-set perception (LIEREx).

The structure of our 3DSSG representation is inspired by other popular approaches in the robotics field [3, 4]. It is implemented as a heterogeneous graph organized as a dynamic hierarchy of multiple layers. These layers represent different levels of semantic concepts, ranging from low-level concepts (e. g., individual objects) to higher-level concepts (e. g., rooms). A core contribution of LIEREx is the extension of this 3DSSG structure to incorporate vision-language features inferred from a VLFM.

These models, most notably the pioneering CLIP model [2], learn rich multimodal feature spaces combining natural language with visual concepts. The initial CLIP model already showed remarkable performance in zero-shot image classification tasks, even matching supervised models trained on benchmark datasets [5, 6]. This, along with the model's comparative simplicity and straightforward integration into downstream tasks, has led to the growing popularity of VLFMs for many computer vision tasks. Recently, *Large Language Models* (LLMs) have also emerged as a promising complementary tool in these applications [7]. Empowered by the integration of VLFMs, the new family of *Large Multimodal Models* (LMMs) also allows visual input from images or videos to be processed directly [8].

[1] https://www.dfki.de/en/web/research/projects-and-publications/project/lierex

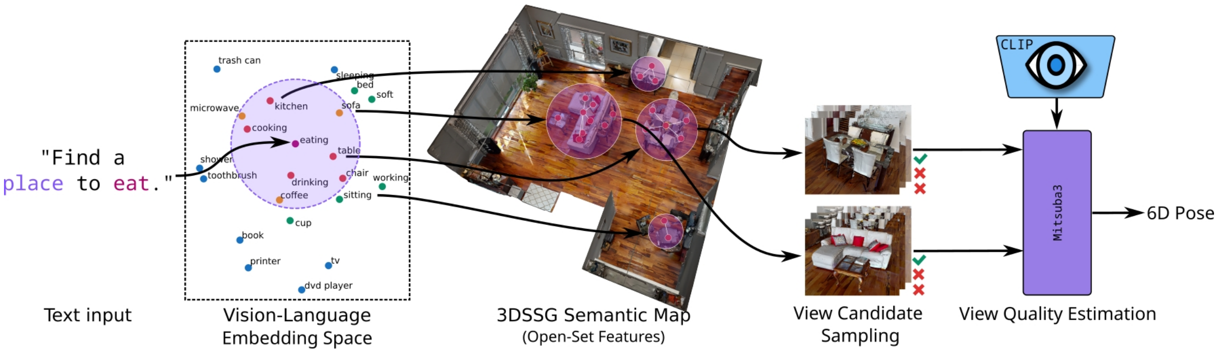

**Fig. 1**: Overview of the LIEREx pipeline. A textual query retrieves the best matching instances from the 3DSSG, taking into account both the scene context provided by the graph and proximity in the Vision-Language Model (VLM) embedding space. Candidate view poses are then sampled from the 3DSSG and scored by the estimation module based on their semantic similarity to the query.

However, CLIP and similar models, e. g., SigLIP [9], BLIP [10], and BLIPv2 [11] only produce feature vectors for the entire image. This prevents their direct applicability in semantic and instance segmentation tasks required for semantic mapping, as these require fine-grained features to localize detected objects within an image. Pixel- and region-based approaches [12–15] address this shortcoming by implementing 2D open-vocabulary semantic segmentation for each pixel or larger regions. Using baseline VLFMs and recent instance-agnostic segmentation models [16] as foundations, these approaches learn to generate individual feature vectors for pixels or regions, albeit at the cost of significantly higher run times compared to image-based VLFMs and traditional models [17].

Unlike traditional segmentation models, these methods do not generate a segmentation mask associated with a specific label unless requested at run-time. Instead, they produce high-dimensional feature maps that encode semantic information directly into the visual representation. This enables a flexible mapping process where visual features are preserved in a latent space and can be retrieved flexibly.

In LIEREx, we extend our hybrid 3DSSG structure with additional CLIP feature vectors to allow for open-set semantic queries and language-driven reasoning atop the spatial reasoning capabilities provided by the graph. Rather than relying on distilled models that project CLIP features onto dense pixels or 3D points, we infer feature vectors using a two-step approach aligned with other popular methods [18–21]. First, potential object candidates are segmented from the incoming RGB-D frames using a class-agnostic segmentation model [16, 22, 23]. Individual CLIP feature vectors are then inferred using the obtained masks and systematically integrated into the corresponding 3DSSG nodes, preferring views where the object of interest is clearly visible [20]. The final 3DSSG can be queried using arbitrary text by comparing the query feature vector with the nodes using cosine similarity and returning all matching objects (see Figure 2).

Additionally, we integrate the spatial structure of the environment with these open-set semantic capabilities. Unlike existing methods that often treat scene graphs as flat collections of objects or a strict hierarchy, our approach aims at inferring the inherent hierarchy of indoor environments automatically. By aligning the multi-scale feature maps from the vision-language model with the 3DSSG structure [18, 21], we ensure that semantic embeddings are not only grounded in visual appearance but also in their spatial context and low-level objects can be clustered dynamically into higher-level objects. This hierarchical architecture transcends static categorization, facilitating complex reasoning for abstract queries that typically exceed the capacity of standard VLFMs.

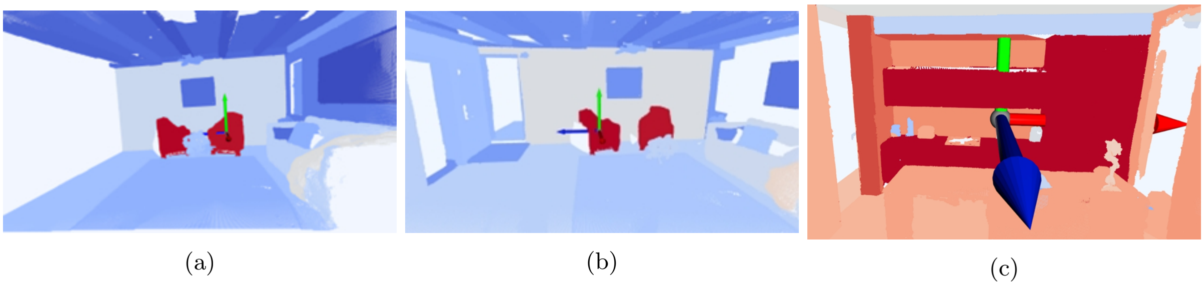

**Fig. 2**: Example queries in the VL map. (a) and (b) show the top 2 results for the text query `chair`, including cropped regions of the surroundings. (c) depicts a query for the higher-order concept `kitchenette`, which comprises multiple object instances.

# 3 Exploration Planning System

In the domain of open-vocabulary exploration, recent approaches predominantly focus on augmenting classical frontier-based strategies with semantic guidance derived from VLFMs or VLMs. These methods typically maintain the fundamental logic of exploring boundaries between known and unknown space but weight these frontiers based on semantic relevance.

In CoW [24], the authors enhance standard frontier selection by integrating CLIP-based relevancy maps to prioritize regions of interest. Similarly, GOAT [25] employs an image-based object instance memory coupled with a global 2D semantic map to navigate towards target object locations. To further guide this process, approaches such as Ren et al. [26] project current RGB frames into 3D voxel maps, utilizing VLMs to score potential exploration directions based on the identified free space.

A significant limitation common to the majority of these works is their reliance on 2D map representations or 2D projections for planning. A notable exception is the work by Laina et al. [19], which operates directly in 3D space by leveraging CLIP-based cosine similarity to modulate sampling near frontiers. However, to estimate the information gain of candidate poses, this method relies on ray-casting over TSDF submaps—a process that is computationally expensive and difficult to scale on resource-constrained platforms.

In LIEREx, we address these limitations by moving away from purely geometric evaluations of candidate views. Instead of performing costly online ray-casting for every potential observation pose, we propose a *learned view quality estimation* system. By training a model to predict the quality of a view relative to a semantic query—based on partially mapped objects and their context—we can efficiently predict quality scores for candidate poses directly from the 3DSSG structure. This allows the agent to verify search results and explore likely object locations with significantly reduced computational overhead. We integrate this targeted view estimation with a semantically guided frontier-based exploration strategy: the system prioritizes learned view proposals for resolving known or hypothesized semantic targets, while reverting to weighted frontier exploration to uncover the unknown parts of the environment.

## 3.1 View Quality Estimation

To enable object-goal navigation, observation poses and trajectories have to be provided in the map's coordinate system. Traditionally, this has been done using heuristic sampling (e. g., FLAP for CAOS [27]) or handcrafted cost functions (e. g., OK-Robot [28]).

The high computational costs for ray-casting and rendering of all poses (compare, e. g., Linok et al. [21]) often makes this infeasible on resource-constrained systems. A data-driven approach can be much more efficient, but requires appropriate training data.

To the best of our knowledge, there is no public dataset available for the task of view pose estimation in cluttered scenes. One insight is that a ground truth *best* observation pose is often very hard to define.

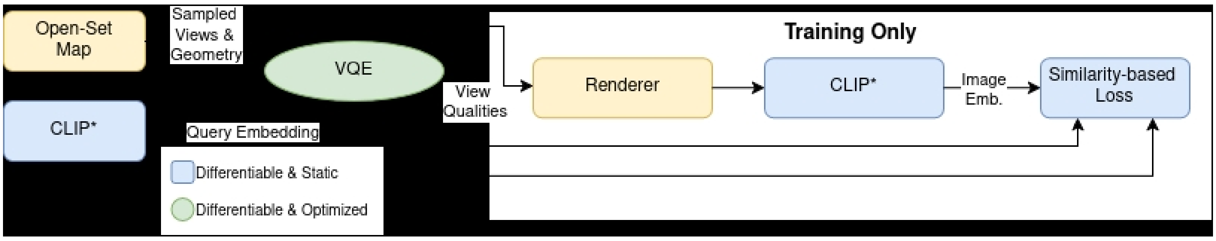

**Fig. 3**: The self-supervised training pipeline for the View Quality Estimation (VQE) module. The model learns to predict quality scores by comparing CLIP embeddings of rendered views against the query embedding.

In LIEREx, we propose a semi-supervised approach based on the evaluation of view quality instead. Inspired by previous works in grasping [29, 30], we propose an approach that estimates the view quality of uniformly sampled poses around potential targets rather than directly regressing poses. Instead of focusing on the visibility of single instances, poses are rendered during training and the inferred CLIP embedding of the rendering is compared to the embedding of the query (see Fig. 3). The core idea is that the model should learn to prefer observation poses that clearly distinguish the queried object or concept in the CLIP feature space.

### 3.2 Data Generation

To facilitate the necessary large-scale training and evaluation of the view quality estimation models, we employ the Habitat simulator [31]. In combination with the Matterport3D [32] and Matterport-Habitat (HM3D) [33] datasets this provides a large and diverse set of realistic indoor environments created from high-resolution 3D photogrammetry.

Unlike traditional robotic simulation environments such as Gazebo, Habitat is optimized for efficient testing of perception and planning algorithms without the overhead of simulating full robot dynamics. This setup enables rapid iteration and large-scale data generation for the self-supervised training of our view pose models. We utilize the HM3D dataset and generate semantic maps based on ground truth trajectories and semantic annotations using our vision-language mapping pipeline.

### 3.3 Training

During training, query-map pairs are randomly sampled from the set of maps and an expected vocabulary, and the top-k regions are extracted based on cosine similarity of the CLIP embeddings (see Fig. 2). The views at the evaluated poses are then rendered using the ground truth meshes from the dataset. CLIP features of the renderings are computed using a frozen version of the same CLIP model used in the VL mapping pipeline. Finally, a variant of a cosine loss between the renderings and the query is computed and back-propagated to the model. Fig. 3 provides a high-level overview of the proposed data flow.

## 4 Indoor Demonstrator

Since our focus is on the development of spatio-semantic environment representations rather than a full SLAM system, we integrated the MICP-L approach [34] on the TIAGo[2] platform. This enables us to utilize high-resolution pre-recorded maps (e. g., from terrestrial LiDARs or CAD models) to obtain ground-truth-like pose estimations without requiring an external tracking system or running full SLAM. This allows for the controlled transfer of our approach from simulated into real-world environments and enables the evaluation of the system with respect to noisy sensor and perception data. The robot is equipped with an Ouster OS0 LiDAR for localization and a Femto Bolt ToF RGB-D camera. Extrinsic calibration was performed by aligning the camera's point cloud to the 3D LiDAR reference frame via ICP registration.

[2] https://pal-robotics.com/es/robot/tiago/

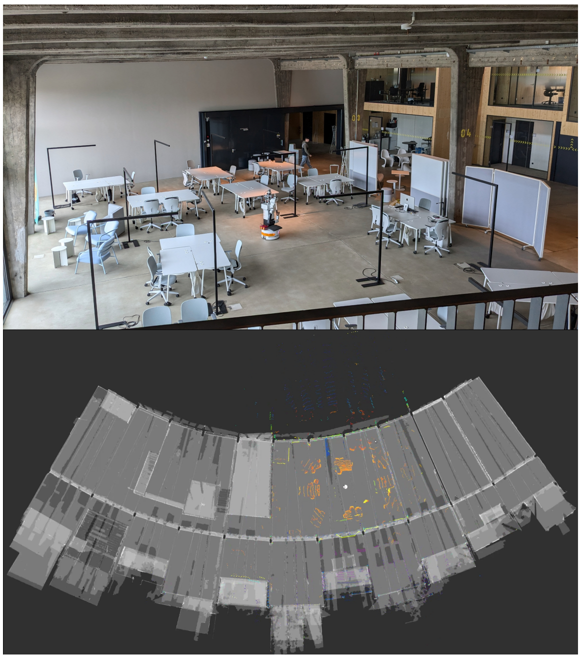

**Fig. 4**: Localization of the TIAGo robot in a large-scale environment. The bottom image shows the pre-recorded polygonal map of the building (excluding furniture) overlaid with LiDAR points that were not associated with the map geometry during registration.

Additionally, associating measurements with the known map geometry allows for the effective filtering of common panoptic classes (e. g., walls, ceilings) that do not necessarily have to be tracked as separate instances in the semantic map. This prunes unnecessary instances from the map and reduces uncertainty, thereby increasing overall efficiency (see Fig. 4).

The robotic platform will be used to evaluate the full mapping and exploration pipeline, including 3DSSG construction, view quality estimation, semantic querying, and spatial reasoning in future work.

## 5 Discussion & Outlook

The development of LIEREx yielded several critical insights regarding the integration of VLFMs into 3D exploration.

We observed that generating *good* observation poses is rarely a function of geometry alone. Simple geometric heuristics (e. g., viewing angle or distance to centroid) fail in cluttered scenes where occlusion and semantic orientation (e. g., the front of an object) are crucial. Consequently, 2D map representations are insufficient; dense 3D information is essential to evaluate visibility and semantic relevance effectively.

Additionally, our initial experiments revealed that directly regressing optimal view poses, e. g., via differentiable rendering, is impractical. The non-convex nature of the optimization landscape frequently leads to convergence in local minima where views are geometrically valid but semantically meaningless. This necessitated a pivot to a *View Quality Estimation* approach, where the system regresses quality scores for sampled candidates rather than optimizing pose parameters directly.

Finally, we note a significant tension between instance-based and unified map representations. While our 3DSSG facilitates complex symbolic reasoning and open-set queries, maintaining geometric consistency for individual instances under sensor noise is significantly harder than in unified voxel-based representations (e. g., TSDFs). While unified maps offer superior efficiency for ray-casting and visibility checks, they often lack the flexibility required for the high-level semantic manipulation afforded by the graph structure.

In future work, we aim to address these trade-offs by integrating the learned view quality estimation into a hybrid, semantically guided frontier-based exploration planner. This complete pipeline will be evaluated on the TIAGo demonstrator to validate the efficacy of open-set 3DSSG reasoning in real-world scenarios.

**Acknowledgements.** This work has been partially supported by the LIEREx project through a grant from the German Federal Ministry of Research, Technology and Space (BMFTR) with Grant Number 16IW24004. The DFKI Niedersachsen (DFKI NI) is supported by the Ministry of Science and Culture of Lower Saxony and the Volkswagen Stiftung.

## Declarations

The authors have no competing interests to declare that are relevant to the content of this article.

# References

[1] Nüchter A, Hertzberg J. Towards Semantic Maps for Mobile Robots. Robotics and Autonomous Systems. 2008 Nov;56(11):915–926. https://doi.org/10.1016/j.robot.2008.08.001.

[2] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models from Natural Language Supervision. In: International Conference on Machine Learning (ICML). PMLR; 2021. p. 8748–8763.

[3] Rosinol A, Gupta A, Abate M, Shi J, Carlone L. 3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans. Robotics: Science and Systems (RSS). 2020;.

[4] Hughes N, Chang Y, Carlone L. Hydra: A Real-time Spatial Perception System for 3D Scene Graph Construction and Optimization. In: Robotics: Science and Systems (RSS); 2022. p. Article 50.

[5] Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In: European Conference on Computer Vision (ECCV). Springer; 2024. p. 38–55.

[6] Ren T, Jiang Q, Liu S, Zeng Z, Liu W, Gao H, et al. Grounding DINO 1.5: Advance the "Edge" of Open-Set Object Detection. arXiv preprint arXiv:240510300. 2024;https://doi.org/10.48550/ARXIV.2405.10300. 2405.10300.

[7] Zeng F, Gan W, Wang Y, Liu N, Yu PS. Large Language Models for Robotics: A Survey. arXiv preprint arXiv:231107226. 2023;https://doi.org/10.48550/ARXIV.2311.07226. 2311.07226.

[8] Huang D, Yan C, Li Q, Peng X. From Large Language Models to Large Multimodal Models: A Literature Review. Applied Sciences. 2024;14(12):5068.

[9] Zhai X, Mustafa B, Kolesnikov A, Beyer L. Sigmoid Loss for Language Image Pre-Training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023. p. 11975–11986.

[10] Li J, Li D, Xiong C, Hoi S. BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation. In: International Conference on Machine Learning (ICML). PMLR; 2022. p. 12888–12900.

[11] Li J, Li D, Savarese S, Hoi S. BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In: International Conference on Machine Learning (ICML). PMLR; 2023. p. 19730–19742.

[12] Zhou C, Loy CC, Dai B. Extract Free Dense Labels from CLIP. In: European Conference on Computer Vision (ECCV). Springer; 2022. p. 696–712.

[13] Ghiasi G, Gu X, Cui Y, Lin TY. Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. In: European Conference on Computer Vision (ECCV). Springer; 2022. p. 540–557.

[14] Ding Z, Wang J, Tu Z. Open-Vocabulary Universal Image Segmentation with MaskCLIP. In: International Conference on Machine Learning (ICML). vol. 202. PMLR; 2023. p. 8090–8102.

[15] Lüddecke T, Ecker AS. Image Segmentation Using Text and Image Prompts. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2022. p. 7076–7086.

[16] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023. p. 4015–4026.

[17] Yamazaki K, Hanyu T, Vo K, Pham T, Tran M, Doretto G, et al. Open-Fusion: Real-Time Open-Vocabulary 3D Mapping and Queryable Scene Representation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2024. p. 9411–9417.

[18] Maggio D, Chang Y, Hughes N, Trang M, Griffith D, Dougherty C, et al. Clio: Real-Time Task-Driven Open-Set 3D Scene Graphs. IEEE Robotics Autom Lett. 2024;9(10):8921–8928. https://doi.org/10.1109/LRA.2024.3451395.

[19] Laina SB, Boche S, Papatheodorou S, Schaefer S, Jung J, Leutenegger S. FindAnything: Open-Vocabulary and Object-Centric Mapping for Robot Exploration in Any Environment. arXiv preprint arXiv:250408603. 2025;.

[20] Kassab C, Mattamala M, Morin S, Büchner M, Valada A, Paull L, et al. The Bare Necessities: Designing Simple, Effective Open-Vocabulary Scene Graphs. arXiv preprint arXiv:241201539. 2024;.

[21] Linok S, Zemskova T, Ladanova S, Titkov R, Yudin D, Monastyrny M, et al. Beyond Bare Queries: Open-Vocabulary Object Grounding with 3D Scene Graph. In: IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2025. p. 13582–13589.

[22] Zhao X, Ding W, An Y, Du Y, Yu T, Li M, et al. Fast Segment Anything. arXiv preprint arXiv:230612156. 2023;.

[23] Zhang C, Han D, Qiao Y, Kim JU, Bae SH, Lee S, et al. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. arXiv preprint arXiv:230614289. 2023;.

[24] Gadre SY, Wortsman M, Ilharco G, Schmidt L, Song S. Cows on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. p. 23171–23181.

[25] Chang M, Gervet T, Khanna M, Yenamandra S, Shah D, Min SY, et al. GOAT: GO to Any Thing. In: Proceedings of Robotics: Science and Systems (RSS). Delft, Netherlands; 2024. p. Article 73.

[26] Ren AZ, Clark J, Dixit A, Itkina M, Majumdar A, Sadigh D. Explore until Confident: Efficient Exploration for Embodied Question Answering. In: Proceedings of Robotics: Science and Systems (RSS). Delft, Netherlands; 2024. p. Article 89.

[27] Gedicke T, Günther M, Hertzberg J. FLAP for CAOS: Forward-Looking Active Perception for Clutter-Aware Object Search. IFAC-PapersOnLine. 2016;49(15):114–119.

[28] Liu P, Orru Y, Vakil J, Paxton C, Shafiullah NMM, Pinto L. Demonstrating OK-Robot: What Really Matters in Integrating Open-Knowledge Models for Robotics. In: Proceedings of Robotics: Science and Systems (RSS). Delft, Netherlands; 2024. p. Article 91.

[29] Wang C, Fang HS, Gou M, Fang H, Gao J, Lu C. Graspness Discovery in Clutters for Fast and Accurate Grasp Detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021. p. 15964–15973.

[30] Fang HS, Wang C, Fang H, Gou M, Liu J, Yan H, et al. AnyGrasp: Robust and Efficient Grasp Perception in Spatial and Temporal Domains. IEEE Transactions on Robotics. 2023;39(5):3929–3945.

[31] Savva M, Kadian A, Maksymets O, Zhao Y, Wijmans E, Jain B, et al. Habitat: A Platform for Embodied AI Research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2019. p. 9339–9347.

[32] Chang A, Dai A, Funkhouser T, Halber M, Niessner M, Savva M, et al. Matterport3D: Learning from RGB-D Data in Indoor Environments. arXiv preprint arXiv:170906158. 2017;.

[33] Ramakrishnan SK, Gokaslan A, Wijmans E, Maksymets O, Clegg A, Turner J, et al. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-Scale 3D Environments for Embodied AI. arXiv preprint arXiv:210908238. 2021;.

[34] Mock A, Wiemann T, Pütz S, Hertzberg J. MICP-L: Mesh-Based ICP for Robot Localization Using Hardware-Accelerated Ray Casting. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE; 2024. p. 10664–10671.