RGB-only Active 3D Scene Graph Generation for Indoor Mobile Robots

Giorgia Modi[12*], Davide Buoso[2], Giuseppe Averta[2], Daniele De Martini[1]
[1]Mobile Robotics Group (MRG), University of Oxford, UK
[2]Visual and Multimodal Applied Learning Lab (VANDAL), Politecnico di Torino, Italy

***Abstract*— Current approaches to 3D scene graph generation rely on dedicated depth sensors, such as LiDAR or RGB-D cameras, for metric 3D reconstruction. This limits deployment to specialized robotic platforms and excludes settings where only RGB cameras are available, such as fixed external infrastructure. Existing pipelines also typically operate on passively collected observation trajectories, rather than selecting viewpoints based on the partially built scene representation, and therefore fail to effectively exploit the semantic and spatial information encoded within the graph during exploration. This paper presents a fully visual framework for the active, incremental construction of 3D scene graphs from RGB input only, addressing both limitations. The proposed approach unifies perception and planning around a shared structured representation that captures object semantics, 3D geometry, relational context, and information from multiple viewpoints. Because the framework is hardware-agnostic and relies only on RGB observations, it can incorporate inputs from both onboard robot cameras and fixed external cameras within the same representation. Experiments on the Replica dataset show that the RGB-only pipeline achieves F1-score parity with baselines using ground-truth depth. Active exploration experiments on ReplicaCAD further show that semantic-driven viewpoint selection detects more than twice as many objects as a geometric frontier-based baseline under the same exploration budget. Finally, the external-camera setting demonstrates that complementary RGB views can effectively bootstrap the scene graph and improve contextual understanding at no additional exploration cost.**

I. Introduction

Modern robotic systems operating in unstructured human environments—such as homes, hospitals, and offices—depend on rich contextual scene understanding to perceive, decide, and act autonomously [1]. Traditional perception pipelines either build low-level dense metric maps that capture free and occupied space but lack semantic context, or produce object-centric outputs such as bounding boxes and semantic segmentations, which isolate entities effectively yet fail to encode relational structure and distinguish semantically different but visually similar spatial configurations [2]. Reliable robot autonomy therefore requires structured scene representations that jointly capture which entities are present, where they are located in three-dimensional space, and how they relate to one another.

To address these limitations, 3D scene graphs have emerged as a foundational architecture for environment perception and contextual reasoning [3], [4]. A 3D scene graph models the environment as a graph whose nodes represent object instances with semantic labels and metric locations, while edges encode pairwise spatial and functional relationships. This compact, queryable representation bridges low-level visual perception and high-level decision-making, supporting language grounding and task planning [5], navigation [6], and manipulation in partially observed scenes [7].

Despite their theoretical utility, the deployment of 3D scene graphs is constrained by significant limitations. State-of-the-art frameworks typically require RGB-D cameras or LiDAR to acquire the dense depth required for metric 3D reconstruction [8], [9], [10]. This restricts deployment to specialized platforms, excluding scenarios where only commodity RGB cameras are available, such as surveillance infrastructure, low-cost robots, or handheld devices. Furthermore, existing 3D scene graph pipelines are predominantly passive, relying on pre-recorded observation trajectories rather than actively selecting informative viewpoints for efficient 3D scene graph construction, thus failing to fully exploit the rich information encoded within the graph.

To address these limitations, we present a fully visual perception–action framework for active 3D scene graph construction from RGB images only, combining MapAnything-based feed-forward 3D reconstruction [11], ConceptGraphs open-vocabulary semantic mapping [8], and semantic exploration [12] within a single pipeline. Our contributions are:

1) an RGB-only pipeline for 3D scene graph generation;
2) the integration of this representation into an active semantic exploration loop for next-best-view selection;
3) an evaluation of fixed external RGB cameras as complementary observations for scene graph bootstrapping and improvement.

II. Related Work

3D Scene Graphs (3DSGs). Scene graphs were introduced as structured representations coupling object semantics with relational context, addressing ambiguities in purely object-centric systems [2]. Extending this abstraction to 3D grounds entities and relationships in metric space, enabling navigation, manipulation, and language-guided planning [13]. Representative systems such as Hydra [9] and SceneGraphFusion [10] build hierarchical representations from SLAM outputs, relying on posed RGB-D data, but operate on passive trajectories and use closed-vocabulary semantics. Open-vocabulary methods, notably ConceptGraphs [8], overcome vocabulary limits using foundation models for

*Correspondent author: giorgiamodi@gmail.com

This work was supported by the EPSRC Impact Acceleration Account (IAA), and the UKRI InnovateUK project 'A general-purpose configuration management system for highly configurable systems validated in automated logistics'.

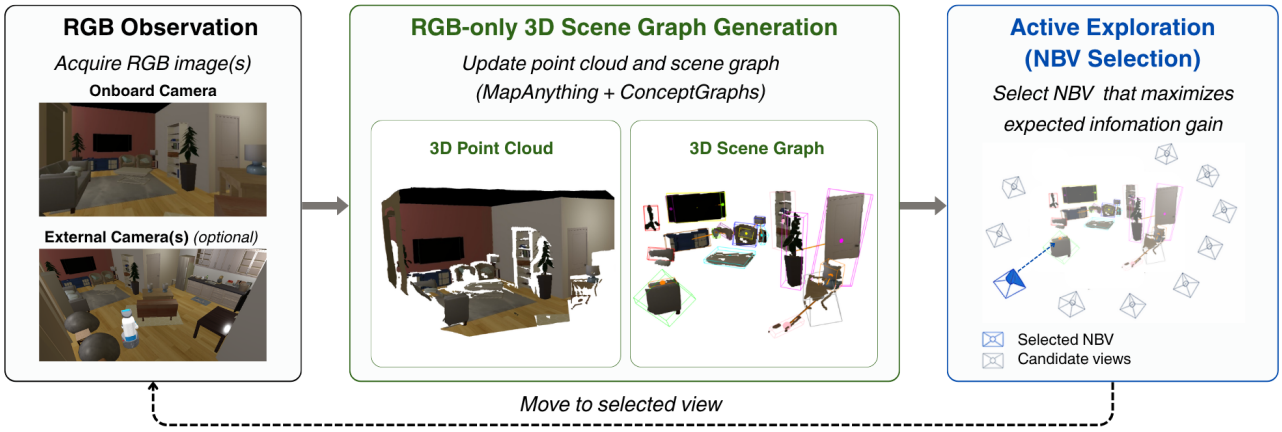

Fig. 1: RGB-only active 3D scene graph generation loop. RGB images from the onboard camera (and, optionally, from fixed external cameras) feed the RGB-only pipeline, incrementally updating the point cloud and scene graph. The active module uses the current scene graph to select the NBV. The robot moves to the selected view and the loop repeats.

flexible language-grounded representations, yet still assume RGB-D inputs. Active scene graph construction remains underexplored: most approaches operate on fixed trajectories without influencing viewpoint selection [13]. Exceptions include the embodied framework of Li et al. [14] and Active Semantic Perception (ASP) [12], which uses Large Language Model (LLM)-sampled scene graph completions to guide exploration, but still relies on depth observations and known camera poses.

Active Perception. Active perception—controlling sensing to maximize information gain—has long been studied via Next-Best-View (NBV) planning [15]. Geometric frontier-based methods [16] and density-based approaches such as Surface Edge Explorer (SEE) [17] efficiently expand coverage but ignore semantics. Recent work incorporates semantic reasoning: Ding et al. [18] use semantic scene completion to guide exploration toward uncertain regions, while Chen et al. [19] combine semantic and geometric uncertainty via 3D Gaussian Splatting. ASP [12] further enables LLM-based reasoning over scene graphs, turning them into active exploration tools. However, prior work on ASP has not considered settings with external camera inputs, nor RGB-only perception.

Multi-View Perception via External Cameras While robotic perception is typically egocentric, fixed external cameras provide complementary global context. Prior work shows their value for visual servoing [20] and VLM-guided navigation [21], but their integration into 3DSGs remains unexplored.

III. Methodology

A. System Overview

The system operates as an incremental perception–action loop (Fig. 1). At each step t, the robot captures an RGB image I_t from its current viewpoint $v_t \in \mathcal{V}$, where $\mathcal{V}$ is the set of navigable viewpoints in the environment. The RGB-only 3DSG generation pipeline processes I_t to update the reconstructed point cloud $\mathcal{P}_t$ and the scene graph $\mathcal{G}_t = (\mathcal{O}_t, \mathcal{E}_t)$, where $\mathcal{O}_t$ is the set of detected object nodes and $\mathcal{E}_t$ the set of spatial relationship edges between them. The active exploration module then selects the Next-Best-View (NBV) v_{t+1} that maximizes expected information gain. The robot navigates to v_{t+1} and the loop repeats. Fixed external cameras inject additional RGB frames into the pipeline, contributing complementary viewpoints without robot motion.

B. RGB-only 3DSG generation

Depths and Poses Estimation. The pipeline leverages MapAnything [11] to infer scene geometry from a set of RGB images. Given N inputs, the model predicts in a single forward pass a *factored* representation of per-view pixel ray directions R_i, up-to-scale depth maps $\tilde{D}_i$, poses $\tilde{P}_i$ expressed in the reference frame of the first image, and a global metric scale factor $m \in \mathbb{R}$ shared across views. These factored outputs are mathematically composed to back-project 2D pixels into a metrically scaled 3D point cloud.

Since MapAnything predicts all camera poses in the reference frame of the first input image, recovering absolute poses only requires fixing this reference: knowing the pose of the first input image (e.g., an external camera or the robot's initial viewpoint) is sufficient to express all poses in a consistent global frame.

Open-Vocabulary Scene Graph Construction. RGB images, together with estimated depths and poses, are passed to ConceptGraphs [8]: SAM [22] extracts class-agnostic instance masks, CLIP [23] embeds each region into a semantic descriptor, the masks are projected into 3D using the MapAnything-derived depth and poses, and multi-view association incrementally merges detections into 3-D object nodes $o_j \in \mathcal{O}_t$. While the original ConceptGraphs pipeline queries proprietary LLMs to infer relationships between objects [8], here spatial edges are instead derived through a deterministic, geometry-only procedure over oriented 3-D bounding boxes of the object nodes: for each ordered object pair, a fixed-priority set of predicates is evaluated - *on top of / supported by* (vertical contact with footprint overlap), *under*

/ over (vertical adjacency), *inside* (volumetric containment), and *next to* (horizontal proximity at similar height) - and at most one edge per ordered pair is assigned, yielding a reproducible scene graph. All geometric relation thresholds (e.g., maximum vertical gap, minimum footprint overlap, and horizontal proximity) are fixed across experiments and tuned empirically once per dataset.

The use of foundation models for perception, such as SAM and CLIP, is crucial in this context, as it removes the need for predefined categories, enabling recognition of unseen objects and better generalization to new environments.

C. Active Semantic Perception

To evaluate exploration efficacy, the framework incorporates two complementary strategies. As a geometric baseline, the Surface Edge Explorer (SEE) algorithm [17] uses measurement density to identify frontier points on partially observed surfaces and proposes viewpoints that maximize their visibility. In contrast, the Active Semantic Perception (ASP) framework [12] performs semantic-driven reasoning using the 3DSG produced by the RGB-only pipeline. At each step, an LLM samples plausible completions of the unobserved scene from the current graph. For each candidate viewpoint x (among the navigable poses available in Habitat), ASP computes the expected information gain as the mutual information between the predicted observation Y_{k+1} and the current graph G_k:

$$I(Y_{k+1}; G_k \mid x) = H(Y_{k+1} \mid x) - H(Y_{k+1} \mid x, G_k) \quad (1)$$

where $H(\cdot)$ denotes Shannon entropy. By selecting the viewpoint that maximizes information gain, ASP prioritizes locations likely to resolve semantic ambiguities (e.g., behind a door leading to a kitchen). In this setting, the 3DSG acts not only as the target representation, but also as the cognitive structure guiding exploration. This dual role highlights the value of scene graphs, which compactly encode entities, attributes, and spatial relations while remaining expressive enough to support higher-level scene understanding and reasoning.

D. Multi-View External Camera Integration

Because the perception framework is strictly visual, observations from uncalibrated external RGB cameras can be processed by MapAnything identically to onboard egocentric images. This allows bird's-eye views to be incorporated into the same 3DSG pipeline to provide a broad initial estimate of the scene, update it during exploration, and improve scene context for downstream planning.

IV. Experiments and Results

We evaluate the method in three settings: static pipeline validation, dynamic active exploration, and multi-view external camera integration. Across all settings, node quality is assessed automatically using joint semantic and geometric matching: predicted and ground-truth labels are embedded with a SentenceTransformer [24], and matches are accepted only when both semantic similarity and 3D localization constraints satisfy fixed thresholds. Quantitative evaluation is therefore restricted to nodes, since datasets used in this work do not provide ground-truth relational annotations, while edges are generated by a deterministic geometry-based module that was extensively tested during development.

A. Static RGB-Only 3DSG Pipeline Validation

Baseline validation was conducted on the Replica dataset [25] to quantify accuracy lost in node prediction when substituting ground-truth depths and poses with MapAnything inferences. Table I reports the results.

TABLE I: Quantitative evaluation of scene graph node accuracy across seven Replica scenes (`room0-2`, `office0-3`). CG is the original ConceptGraphs, CG-RGB is the proposed RGB-only variant, with predicted depths and poses. Values are reported as mean $\pm$ standard deviation across scenes.1

Experiment	Precision ↑	Recall ↑	F1-score ↑
CG	$\mathbf{0.686 \pm 0.05}$	0.401 ± 0.10	0.499 ± 0.08
CG-RGB (ours)	0.615 ± 0.07	$\mathbf{0.436 \pm 0.12}$	$\mathbf{0.500 \pm 0.08}$

The results demonstrate remarkable parity, with comparable F1 scores. MapAnything's noisier depth predictions introduce minor geometric inconsistencies, slightly lowering precision but preventing ConceptGraphs from overly aggressive merging of closely situated objects, thereby improving recall. Overall, these findings indicate that replacing ground-truth depth and pose with MapAnything estimates preserves the quality of scene graph nodes.

B. Active Exploration Evaluation

The RGB-only active exploration pipeline reported in Fig. 1 was evaluated within the Habitat simulator [26] using the six `apartment` scenes from the ReplicaCAD dataset [27], where local navigation to selected viewpoints is handled via Habitat's built-in navigation utilities. `gemini-2.5-pro` [28] was used as LLM for ASP. Experiments were run from 10 distinct starting positions over 30 exploration steps. Since performance trends were consistent across the six apartments, the results reported in Fig. 2 focus on `apartment 3` as a representative scene.

The semantic-driven ASP framework consistently outperformed the geometric SEE baseline. Operating exclusively on onboard cameras, ASP identified approximately 110 object nodes by step 30, approaching the ground-truth total (124). SEE stalled at roughly 45 nodes. In addition, ASP achieved double the recall (0.54 vs. 0.22). By prioritizing semantic ambiguity, ASP matches the quality of hundreds of passive trajectory frames in thirty intelligent movements.

C. Multi-View External Camera Integration Study

The efficiency of multi-view perception was tested by integrating fixed external infrastructure as bootstrapping inputs. In the active exploration trials (Fig. 2), initializing the scene graph with just *one* overhead external camera instantly jumped SEE's starting node count from 16 to 37 (+125%) and ASP's from 23 to 36 (+57%). Initial recall improved

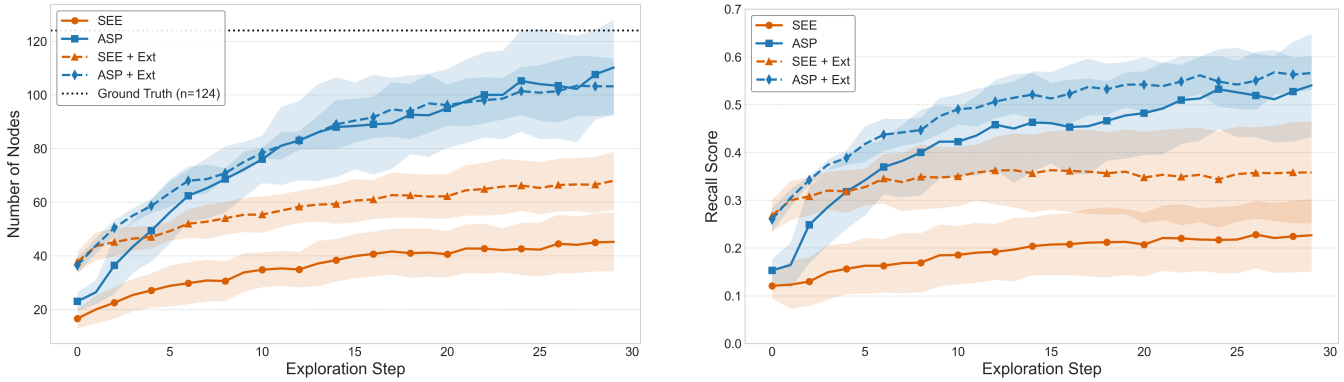

Fig. 2: Evolution of number of predicted nodes (left) and recall (right) for the geometry-based (SEE) and semantic-based (ASP) exploration algorithms, with and without an initial external camera view (+Ext), over exploration steps. Shaded areas denote standard deviation over different initial robot poses.

from 0.12 to 0.27 for SEE (+130%) and from 0.15 to 0.27 for ASP (+80%), and this was accompanied by a consistent improvement in final recall for both methods, particularly pronounced for SEE.

To evaluate standalone infrastructure capabilities, 270 experiments were conducted across 90 ReplicaCAD scenes utilizing only external cameras (from one to three) without a mobile robot. The scenes were split into complex `apartments` (dense clutter, ∼123 objects, used also in the previous experiment) and simpler `furnished rooms` (∼25 larger objects). Table II reports the results.

TABLE II: Quantitative evaluation of scene graph node accuracy using between one and three fixed external cameras (#Cam.) on ReplicaCAD `apartments` (Apt.) and `furnished rooms` (Fur.) scenes. Values are reported as mean ± standard deviation across scenes.

Scene	#Cam.	Precision ↑	Recall ↑	F1-score ↑
Apt.	1	**0.770 ± 0.054**	0.198 ± 0.030	0.315 ± 0.041
	2	0.741 ± 0.036	0.232 ± 0.030	0.353 ± 0.036
	3	0.698 ± 0.031	**0.301 ± 0.029**	**0.421 ± 0.030**
Fur.	1	**0.566 ± 0.087**	0.398 ± 0.071	0.465 ± 0.073
	2	0.540 ± 0.077	0.487 ± 0.071	0.510 ± 0.067
	3	0.486 ± 0.060	**0.555 ± 0.065**	**0.516 ± 0.055**

While 3 static viewpoints cannot match the recall of a 30-step active robotic exploration, they can successfully establish the environment's dominant structure. In complex apartments, three cameras achieved an F1-score of 0.421, while in furnished rooms, the F1-score reached 0.516, successfully recovering over 55% of the ground truth objects instantly. Precision declines slightly as cameras are added due to overlapping multi-view duplicate fragments, but the net F1-score confirms that fixed external RGB cameras can provide useful complementary observations without requiring additional onboard sensing hardware.

V. Discussion

The results suggest that active robot perception benefits more from structured semantic representations than from geometry alone. While dense geometric maps are effective for navigation, they do not explicitly encode object identity or spatial relations. By combining geometry, semantics, and structure, 3D scene graphs provide a richer and more informative perceptual model. Their structured nature also makes them naturally queryable by language models, which here act as spatial reasoning engines for inference in partially observed environments. Instead of relying only on geometric visibility, the system can use the evolving scene graph to reason about unobserved regions, infer plausible scene completions, and prioritize semantically informative viewpoints, consistent with the improved recall achieved by ASP.

Removing the dependence on depth sensors also offers practical advantages, since an RGB-only pipeline enables hardware-agnostic, multi-view perception and supports heterogeneous visual inputs. In this setting, fixed external cameras can help bootstrap and update the scene graph, reducing exploration effort while improving scene understanding. Beyond increasing scene graph completeness, these viewpoints can also support safer motion planning under semantic uncertainty by revealing obstacles, free space, and relevant structures before they become visible to the onboard camera, thereby reducing both geometric and semantic uncertainty in partially observed environments.

VI. Conclusion

This paper introduced an active perception framework for incremental 3D scene graph construction using only RGB visual inputs. By combining feed-forward reconstruction, open-vocabulary semantics, and LLM-driven exploration, the system removes depth-sensor dependency while improving exploration efficiency. Experiments demonstrate that semantic-driven perception significantly outperforms geometric baselines, while the RGB-only design enables effective integration of external cameras, opening broader possibilities for flexible, infrastructure-assisted perception in real-world robotic deployments. These results highlight the potential of semantics-first representations as a foundation for reliable and scalable robot autonomy.

REFERENCES

[1] J. Ni, Y. Chen, G. Tang, J. Shi, W. Cao, and P. Shi, "Deep learning-based scene understanding for autonomous robots: A survey," *Intelligence & Robotics*, vol. 3, no. 3, pp. 374–401, 2023.

[2] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5410–5419.

[3] H. Li, G. Zhu, L. Zhang, Y. Jiang, Y. Dang, H. Hou, P. Shen, X. Zhao, S. A. A. Shah, and M. Bennamoun, "Scene graph generation: A comprehensive survey," *Neurocomputing*, vol. 566, p. 127052, 2024.

[4] J. Bae, D. Shin, K. Ko, J. Lee, and U.-H. Kim, "A survey on 3d scene graphs: Definition, generation and application," in *International Conference on Robot Intelligence Technology and Applications*. Springer, 2022, pp. 136–147.

[5] 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," *arXiv preprint arXiv:2307.06135*, 2023.

[6] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, "Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[7] H. Jiang, B. Huang, R. Wu, Z. Li, S. Garg, H. Nayyeri, S. Wang, and Y. Li, "Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation," *arXiv preprint arXiv:2402.15487*, 2024.

[8] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa *et al.*, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5021–5028.

[9] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3d scene graph construction and optimization," *arXiv preprint arXiv:2201.13360*, 2022.

[10] S.-C. Wu, J. Wald, K. Tateno, N. Navab, and F. Tombari, "Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7515–7525.

[11] N. Keetha, N. Müller, J. Schönberger, L. Porzi, Y. Zhang, T. Fischer, A. Knapitsch, D. Zauss, E. Weber, N. Antunes *et al.*, "Mapanything: Universal feed-forward metric 3d reconstruction," *arXiv preprint arXiv:2509.13414*, 2025.

[12] H. Tang and P. Chaudhari, "Active semantic perception," *arXiv preprint arXiv:2510.05430*, 2025.

[13] I. Catalano, C. C. Zumaya, J. A. Placed, J. Civera, W. M. Bessa, and J. Peña-Queralta, "3d scene graphs in robotics: A unified representation bridging geometry, semantics, and action," *Authorea Preprints*, 2025.

[14] X. Li, D. Guo, H. Liu, and F. Sun, "Embodied semantic scene graph generation," in *Conference on robot learning*. PMLR, 2022, pp. 1585–1594.

[15] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," *Autonomous Robots*, vol. 42, no. 2, pp. 177–196, 2018.

[16] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*. IEEE, 1997, pp. 146–151.

[17] R. Border and J. D. Gammell, "The surface edge explorer (see): A measurement-direct approach to next best view planning," *The International Journal of Robotics Research*, vol. 43, no. 10, pp. 1506–1532, 2024.

[18] H. Ding, X. Liang, Y. Fang, Y. Wu, J. Shi, J. Huo, W. Li, J. Wu, Y.-K. Lai, and Y. Gao, "Sea: Semantic map prediction for active exploration of uncertain areas," *arXiv preprint arXiv:2510.19766*, 2025.

[19] L. Chen, H. Zhan, H. Yin, Y. Xu, and P. Mordohai, "Understanding while exploring: Semantics-driven active mapping," *arXiv preprint arXiv:2506.00225*, 2025.

[20] L. Robinson, M. Gadd, P. Newman, and D. D. Martini, "Robot-relay: Building-wide, calibration-less visual servoing with learned sensor handover networks," *Autonomous Robots*, vol. 50, no. 1, p. 3, 2026.

[21] D. Buoso, L. Robinson, G. Averta, P. Torr, T. Franzmeyer, and D. De Martini, "Select2plan: Training-free icl-based planning through vqa and memory retrieval," *IEEE Robotics and Automation Letters*, 2025.

[22] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[24] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[25] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[26] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9339–9347.

[27] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. S. Chaplot, O. Maksymets *et al.*, "Habitat 2.0: Training home assistants to rearrange their habitat," *Advances in neural information processing systems*, vol. 34, pp. 251–266, 2021.

[28] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.