Where Did I Leave My Glasses? Open-Vocabulary Semantic Exploration in Real-World Semi-Static Environments

Benjamin Bogenberger, Lukas Brunke, Siqi Zhou, Angela P. Schoellig

***Abstract*— Robots deployed in real-world environments, such as homes, must not only navigate safely but also understand their surroundings and adapt to changes in the environment. Existing research on semantic exploration largely focuses on static scenes without persistent object-level instance tracking. In this work, we propose an open-vocabulary, semantic exploration system for semi-static environments. Our system maintains a consistent map by building a probabilistic model of object instance stationarity, systematically tracking semi-static changes, and actively exploring areas that have not been visited for an extended period. In addition to active map maintenance, our approach leverages the map's semantic richness with large language model (LLM)-based reasoning for open-vocabulary object-goal navigation. Evaluated on state-of-the-art baselines using publicly available object navigation and mapping datasets, our method outperforms the baselines in both success rate in object-goal navigation tasks and in handling scene changes during mapping. A video of full (including real-world) experimental results can be found at https://tiny.cc/sem-explor-semi-static and on our website https://utiasdsl.github.io/semi-static-semantic-exploration/.**

I. Introduction

Autonomous robots executing everyday tasks, like object-goal navigation—locating objects in unknown or partially known spaces—require the same tight integration of semantic understanding, spatial reasoning, and adaptability that humans use to navigate and adapt to scene changes (e.g., moved furniture). Most existing approaches target static scenes [1]–[3], failing to address unobserved, long-term semi-static changes [4], [5]. Maintaining a consistent spatio-temporal environment representation is crucial for efficient long-term autonomy but remains underexplored in object-goal navigation.

II. Problem Formulation

Our system takes RGB-D frames $\mathbf{F}_t$ along with camera poses $\mathbf{T}_t^{\mathrm{CW}}$, which are assumed known and obtained here via a visual inertia odometry (VIO) system [6]. It incrementally builds and updates a semantic map of a semi-static environment comprising an object library $\mathcal{O}_t$, a missing object library $\mathcal{O}_{\mathrm{mis},t}$, and a background point cloud $\mathbf{P}_{\mathrm{bg},t}$, all of which can be initialized as empty sets or based on a prior map. Given a user query $\mathbf{q}$ (e.g., "Find my glasses!" or "Maintain the map!"), we generate an exploration priority map which allows the robot to either search for the requested object or actively revisit likely outdated map regions to maintain a spatial-temporal consistent 3D map.

The authors are with the Learning Systems and Robotics Lab and the Munich Institute of Robotics and Machine Intelligence, Technical University of Munich, 80333 Munich, Germany. This work has been supported by the Robotics Institute Germany, funded by BMBF grant 16ME0997K, and the European Union's Horizon Europe project under the Marie Skłodowska-Curie grant agreement No. 101155035 (SSDM). Email: firstname.lastname@tum.de

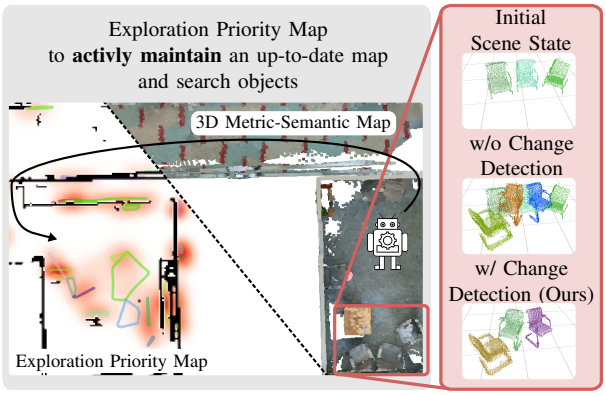


Fig. 1. Our proposed open-vocabulary semantic exploration approach for semi-static environments, where objects can be shifted, removed, or reintroduced. To account for such changes, the system explicitly maintains a stationarity score for each object instance and actively revisits regions of the map that are likely outdated. This enables the construction of an up-to-date metric-semantic map, which we use to prioritize contextually relevant areas during (unseen) object-goal navigation in semi-static scenes.

III. Methodology

At each timestep t, we process an RGB-D frame $\mathbf{F}_t$ with its camera pose $\mathbf{T}_t^{\mathrm{CW}}$ to detect visible object candidates $\mathcal{Y}_t$. These candidates are used to incrementally update the scene belief, which comprises the object library $\mathcal{O}_t$, the missing object library $\mathcal{O}_{\mathrm{mis},t}$, and the accumulated background point cloud $\mathbf{P}_{\mathrm{bg},t}$. Each object comprises a point cloud, a visual feature vector, and an open-vocabulary class label. Given a user query $\mathbf{q}$ (e.g., search or maintenance), we construct an exploration priority map $f_{\mathrm{task}}(\cdot \,|\, \mathcal{O}_t, \mathbf{q})$ to sample target waypoints for navigation.

A. Current View Object Candidates and Scene Belief Update

We extract object candidates $\mathcal{Y}_t$ from the RGB-D frame using SAM [7] for segmentation masks and CLIP [8] for visual features and open-vocabulary class labels. The background point cloud $\mathbf{P}_{\mathrm{bg},t}$ is updated with points not covered by any object mask.

1) Expected-View Association: We update the scene belief by associating current candidates $\mathcal{Y}_t$ with objects expected to be visible in the current view, $\mathcal{O}_{t,\mathrm{exp}} \subseteq \mathcal{O}_{t-1}$. An object $\mathbf{O}_i$ is expected if the visible fraction of its point cloud exceeds a threshold τ_{expected}. Matching relies on *(i)* semantic similarity (S_{sem}), which is the cosine similarity between the visual feature vectors, and *(ii)* geometric similarity (S_{geo}), which is the intersection over union of the point clouds.

The association proceeds in two steps: first, by greedily pairing objects under a stationarity assumption using S_{geo}

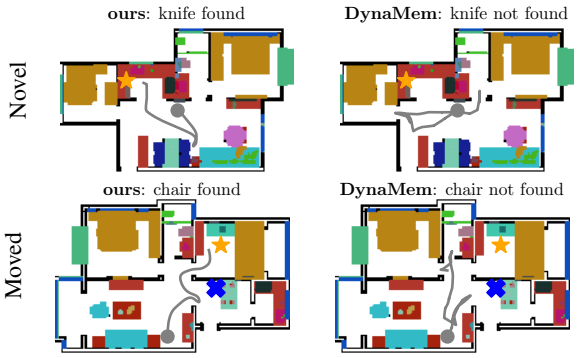


Fig. 2. Example of our method (left) and DynaMem [9] (right) searching for an unseen knife (top) and a moved chair (bottom). Our method checks the dining table, then kitchen and bedroom, while DynaMem explores randomly. Robot path and start shown in gray. Goal object marked with a yellow star; prior location with a blue cross.

and S_{sem} thresholds; second, by matching remaining objects, which may have moved, by prioritizing S_{sem} and verifying alignment via ICP. Matched candidates are merged with their corresponding objects; unmatched candidates are added as new objects.

2) Stationarity Score Update and Object Management: Each object $\mathbf{O}_i$ maintains a probabilistic belief of its stationarity score $v_i \in [0,1]$ using a model [4] that updates the score based on the measured change $\Delta_{t,i}$ and an large language model (LLM)-derived prior stationarity label $\mathbf{s}_i$ (static/dynamic), according to the Bayesian update $p(v|\Delta_{1:t}, \mathbf{s}) \propto p(\Delta_t|v, \Delta_{1:t-1})p(v|\Delta_{1:t-1}, \mathbf{s})$.

For out-of-view objects, we inject a synthetic change whose magnitude is scaled by the stationarity prior $\mathbf{s}_i$ to decay their stationary score as uncertainty increases (dynamic objects are more likely to be outdated than static ones).

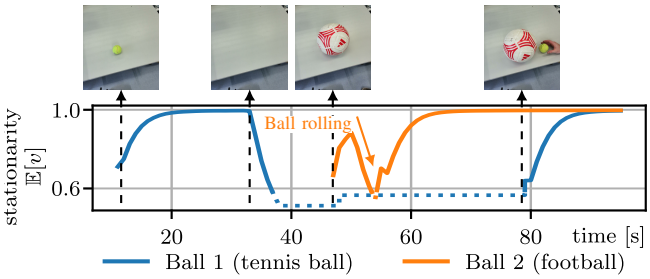
Objects $\mathbf{O}_i$ with $\mathbb{E}[v_i] \leq \theta_{\text{r}}$ are moved to the missing object library $\mathcal{O}_{\text{mis},t}$. Missing objects are reidentified and merged with newly observed objects in $\mathcal{O}_t$ within a temporal window $\pm\overline{\tau}$ of their disappearance using a similar matching process as above.

B. Exploration Priority Map

The 2D exploration priority map $f_{\text{task}}(\mathbf{x} \,|\, \mathcal{O}_t, \mathbf{q})$ (with 2D position $\mathbf{x}$) is computed as a weighted superposition of per-object floor projections $f(\mathbf{x} \,|\, \mathbf{O}_i)$, i.e., $f_{\text{task}}(\mathbf{x} \,|\, \mathcal{O}_t, \mathbf{q}) \propto \sum_{\mathbf{O} \in \mathcal{O}_t} \lambda(\mathbf{O} \,|\, \mathbf{q}) f(\mathbf{x} \,|\, \mathbf{O})$. The semantic relevancy score $\lambda(\mathbf{O}_i \,|\, \mathbf{q})$ prioritizes objects for the current task. For map maintenance, relevancy is computed with a custom designed weighting function $f(\mathbb{E}[v_i]) : [0,1] \mapsto [0,1]$ peaking for low-stationarity objects. For object search, an LLM estimates the likelihood of finding the target near each object $\mathbf{O}_i$ based on its class label.

C. Planning and Control

We select navigation goals with a sampling-based strategy from the exploration priority map f_{task}. The sampled waypoints are navigated to using a global path planner and executed via model predictive control (MPC), relying on a 2D occupancy map derived from the current scene belief.



(a) The stationarity rises and drops when objects appear and disappear.

Enabled (i.e., $\tau_\cdot > 0$) Similarity Measures \ Objects to Reidentify	Different Sem. & Geom.	Different Semantics	Different Geometries
			
$\tau_{\text{geo}} > 0, \tau_{\text{sem}} = 0$	✓	✗	✓
$\tau_{\text{geo}} = 0, \tau_{\text{sem}} > 0$	✓	✓	✗
$\tau_{\text{geo}} > \mathbf{0}, \tau_{\text{sem}} > \mathbf{0}$	✔	✔	✔

(b) The table shows which similarity measures are necessary to successfully reidentify the shown object instances. Only when both measures are used all three cases are covered.

Fig. 3. Example of removal, reintroduction, and translation. Our system can distinguish between objects of the same class and reidentify previously shown object instances.

TABLE I

RESULTS ON THE KHRONOS DATASET. KHRONOS' RESULTS FROM [11]
$^+$: upper bounded by $\text{F1} \leq \frac{1}{2}(\text{Pre} + \text{Rec})$

	Method	Dynamics Pre	Dynamics Rec	Dynamics F1	Changes Pre	Changes Rec	Changes F1
apartment	Khronos [11]	90.4	78.6	84.1	31.3	69.1	$\leq 50.2^+$
	Ours (default)	**92.1**	**86.1**	**88.9**	**94.8**	**84.8**	**89.3**
office	Khronos [11]	**96.0**	59.7	73.2	24.5	**54.2**	$\leq 39.4^+$
	Ours (default)	93.8	**71.3**	**80.2**	**66.5**	47.0	**53.0**

IV. EXPERIMENTAL RESULTS

We evaluate our approach in terms of object-goal navigation performance and change detection quality. In object-goal navigation, we design 60 closed-loop tasks on the InteriorAgent dataset [10], introducing semi-static changes by hiding and revealing objects; our method achieves an average of 25% increased success weighted by path length (SPL) compared to DynaMem [9]. For change detection quality, we compare against Khronos [11] on its own dataset, demonstrating an average of 26.35% higher change-detection F1 score, primarily because our system incorporates semantic checks in addition to geometric analysis (Tab. I). Furthermore, an ablation study on object instance tracking shows that while either semantic (S_{sem}) or geometric (S_{geo}) similarity may suffice for tracking distinct instances of the same class, the robust tracking of objects that differ in only one of these properties necessitates the integration of both similarity measures (Fig. 3). We confirm the real-world transferability of our method in three real-world environments.

V. CONCLUSION

We propose a novel open-vocabulary semantic exploration approach for robots operating in semi-static environments. Beyond traditional object-goal navigation, our approach actively targets map regions likely to be outdated. We verified its effectiveness in object-goal navigation and mapping changing scenes against state-of-the art methods.

REFERENCES

[1] H. Shah, J. Xing, N. Messikommer, B. Sun, *et al.*, "ForesightNav: Learning scene imagination for efficient exploration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2025.

[2] Y. Goel, N. Vaskevicius, L. Palmieri, N. Chebrolu, K. O. Arras, and C. Stachniss, "Semantically informed MPC for context-aware robot exploration," pp. 11218–11225, 2023.

[3] G. Georgakis, B. Bucher, K. Schmeckpeper, S. Singh, *et al.*, "Learning to map for active semantic goal navigation," *arXiv: 2106.15648*, 2022.

[4] J. Qian, V. Chatrath, J. Yang, J. Servos, A. P. Schoellig, *et al.*, "POCD: Probabilistic object-level change detection and volumetric mapping in semi-static scenes," in *Proc. Proc. Robot., Sci. Syst.*, 2022.

[5] J. Qian, V. Chatrath, J. Servos, A. Mavrinac, W. Burgard, S. L. Waslander, and A. P. Schoellig, "POV-SLAM: Probabilistic object-aware variational SLAM in semi-static environments," in *Proc. Proc. Robot., Sci. Syst.*, 2023.

[6] O. Seiskari, P. Rantalankila, J. Kannala, J. Ylilammi, *et al.*, "HybVIO: Pushing the limits of real-time visual-inertial odometry," in *Proc. IEEE/CVF Wint. Conf. Applic. of Comput. Vis.*, pp. 287–296, 2022.

[7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, *et al.*, "Segment Anything," in *Proc. IEEE/CVF Int. Conf. Com. Vis.*, pp. 4015–4026, 2023.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Intl. Conf. on ML.*, pp. 8748–8763, 2021.

[9] P. Liu, Z. Guo, M. Warke, S. Chintala, *et al.*, "Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation," in *Proc. IEEE Int. Conf. Robot. Automat.*, pp. 13346–13355, 2025.

[10] M. T. I. SpatialVerse Research Team, "Interioragent: Interactive usd interior scenes for isaac sim-based simulation." `https://huggingface.co/datasets/spatialverse/InteriorAgent`, 2025.

[11] L. Schmid, M. Abate, Y. Chang, and L. Carlone, "Khronos: A unified approach for spatio-temporal metric-semantic SLAM in dynamic environments," in *Proc. Proc. Robot., Sci. Syst.*, 2024.