Article

Not peer-reviewed version

Learning Traversable Scene Structures for Embodied Navigation with Movable Object Constraints

James Walker , Emily Carter , Thomas Bennett , Sophie Hughes *

Posted Date: 26 January 2026

doi: 10.20944/preprints202601.1852.v1

Keywords: traversable scene learning; movable objects; embodied navigation; graph neural networks; indoor environments

Article

Learning Traversable Scene Structures for Embodied Navigation with Movable Object Constraints

James Walker, Emily Carter, Thomas Bennett and Sophie Hughes *

Department of Electronic and Electrical Engineering, Imperial College London, London SW7 2AZ, United Kingdom

* Correspondence: sophie.hughes@imperial.ac.uk

Abstract

Understanding how movable objects affect navigability is critical for embodied agents operating in realistic environments. This study proposes a learning-based approach to infer traversable scene structures under object mobility constraints. A neural graph encoder is trained to predict passability relations between spatial regions conditioned on object states, using RGB-D observations and interaction feedback. The model is trained on 15,000 simulated navigation trajectories generated in rearranged indoor scenes. Quantitative evaluation shows that the learned scene structure reduces navigation failure due to blocked paths by 28.4% and improves average navigation efficiency by 16.7% compared with static scene graph representations.

Keywords: traversable scene learning; movable objects; embodied navigation; graph neural networks; indoor environments

1. Introduction

Embodied navigation has advanced substantially in recent years, driven by progress in visual perception models, large-scale simulation platforms, and improved training strategies for long-horizon decision making in indoor environments [1]. Modern navigation agents achieve notable gains in localization accuracy, exploration efficiency, and goal-directed planning compared with earlier reactive or map-free baselines, particularly when trained and evaluated on visually diverse, standardized benchmarks [2]. These improvements demonstrate the effectiveness of learning-based approaches for spatial reasoning and control under partial observability. However, despite steady progress on benchmark tasks, a persistent gap remains between typical evaluation settings and the complexity of real-world indoor environments. In practical settings such as homes, offices, hospitals, and warehouses, navigation conditions change frequently due to human activity and object movement. Doors may be opened or closed, chairs and carts may be repositioned, and temporary obstacles may appear in narrow corridors or at junctions. Under such conditions, whether a path is passable depends not only on static geometry but also on the current configuration and state of movable objects [3,4]. Recent work has begun to acknowledge this challenge by introducing hierarchical 3D scene representations that explicitly reason about traversability in the presence of movable obstacles, showing that navigation performance degrades sharply when object-induced constraints are not modeled [5]. These findings highlight that reliable indoor navigation requires reasoning beyond static layouts, toward representations that capture how object states condition feasible transitions between regions.

To address partial observability and long-horizon planning, many studies have focused on learning structured representations of indoor scenes. Hierarchical models, spatial memory systems, and map-based architectures construct intermediate representations that summarize explored regions and support planning over extended time horizons [6]. Among these, scene graphs have emerged as a common abstraction because they encode spatial layout together with semantic entities and their relations, enabling compact reasoning over rooms, objects, and connectivity [7]. Recent

 doi:10.20944/preprints202601.1852.v1

approaches use scene graphs either as explicit belief states or as auxiliary supervision, linking egocentric observations to structured nodes and edges to improve stability, generalization, and data efficiency [8,9]. Extensions to open-vocabulary and open-world navigation further enrich these representations by allowing flexible schemas that support reasoning over unseen categories and tasks [10]. Collectively, these works demonstrate that structured scene representations can reduce navigation errors caused by ambiguity, occlusion, and limited exploration [11].Despite these advances, most existing representations implicitly assume that traversability is a fixed property of space or can be approximated through repeated replanning over a static or slowly changing map. This assumption is often violated in indoor environments where movable objects temporarily block or open passages, particularly in narrow or high-traffic areas [12]. In many scene graph formulations, connectivity is defined using geometric proximity or visibility, which does not directly encode whether a transition remains feasible when an object obstructs a passage. Even when dynamic replanning is employed, failures often arise because the underlying representation does not explicitly model how object states alter passability relations between regions.Related research in embodied rearrangement and mobile manipulation further illustrates the tight coupling between navigation and object interaction [13,14]. These systems show that agents must adapt to layout changes introduced by object movement and, in some cases, actively manipulate objects to reach goals. However, most rearrangement-focused approaches emphasize task completion or object placement accuracy rather than learning a reusable representation of navigability that conditions on object states and can generalize across navigation tasks. As a result, the learned policies are often tied to specific action sequences or task distributions, rather than providing an explicit, transferable description of passable structure under dynamic conditions. Limitations in data and supervision further constrain existing navigation models. Many datasets emphasize static scenes or include only limited variability in object placement, reducing exposure to blocked-path failures during training [15]. When scene changes are introduced, they are often treated as episode-level randomness rather than as structured, labeled changes in passability relations. Consequently, agents may overfit to default layouts or adopt conservative behaviors that avoid cluttered regions, even when those regions are in fact traversable. Memory-based models can track changes over time, but they are typically designed for object localization or relational prediction rather than for estimating conditional traversability between spatial regions [16]. Evaluation scale and scenario coverage also remain limited. Although recent simulators increase visual diversity and scene count, many benchmarks still involve a narrow range of object categories or rearrangement patterns, making it difficult to assess robustness under rare but critical failures, such as a single movable object blocking a key junction [17]. Emerging embodied AI benchmarks increasingly emphasize interaction and dynamic environments, yet consistently report sharp performance degradation when object-induced constraints are not explicitly represented [18]. These observations suggest that progress in embodied navigation will depend on learning representations that treat traversability as a relational, state-dependent property rather than a static attribute of space.

In this study, we address these challenges by learning traversable scene structures under movable object constraints. We model passability as a conditional relation between spatial regions that depends on the current configuration of movable objects, rather than assuming fixed connectivity. A neural graph encoder is trained to predict region-level passability using RGB-D observations together with interaction feedback that reflects whether attempted transitions succeed or fail. This interaction-driven supervision enables the model to separate stable structural cues from state-dependent constraints introduced by object movement. As a result, the learned representation allows an agent to anticipate blocked routes, reason about alternative paths, and select actions that remain feasible at execution time. We evaluate the proposed approach using a large set of simulated trajectories in rearranged indoor environments with diverse object configurations. Compared with static scene graph representations, our method significantly reduces navigation failures caused by unexpected obstructions and improves path efficiency under dynamic conditions. By explicitly modeling conditional traversability, this work provides a reusable scene-level abstraction that

 doi:10.20944/preprints202601.1852.v1

complements policy learning, supports robust decision making under partial observability, and moves embodied navigation closer to the requirements of real-world indoor environments.

2. Materials and Methods

2.1. Samples and Study Environment

Experiments were carried out in simulated indoor environments that represent common residential and office settings. A total of 15,000 navigation trajectories were collected from 120 distinct scenes, including apartments, offices, corridors, and mixed indoor layouts. Each scene contained movable objects such as chairs, tables, carts, and doors that could change position during an episode. For each trajectory, the agent's starting pose, target location, and object configuration were randomly assigned. This sampling strategy ensured coverage of different layouts and object-induced changes in passability. RGB-D observations were recorded at every step using a fixed camera height and field of view. Spatial regions were defined by dividing the explored free space into connected cells based on geometric adjacency and line-of-sight constraints.

2.2. Experimental Design and Control Settings

The proposed method was evaluated against two baseline configurations. In the experimental setting, the model learned passability relations between spatial regions while accounting for object states. The first control setting used a static scene graph derived from the initial free-space geometry and assumed fixed connectivity throughout navigation. The second control setting applied online replanning on the same static graph without learning object-dependent relations. All methods used the same perception module, action space, and navigation policy. This design isolated the effect of scene structure modeling. Evaluation was conducted on unseen scenes with novel object rearrangements to test generalization beyond the training conditions.

2.3. Measurement Procedures and Quality Control

Passability between two regions was determined through interaction outcomes. A transition was marked as passable if the agent moved between regions without collision or deadlock. Otherwise, it was labeled as non-passable. Navigation performance was measured using success rate, failure rate caused by blocked paths, and average navigation efficiency. Efficiency was defined as the ratio between the shortest feasible path and the executed path. Each navigation task was repeated five times with different random seeds, and results were averaged to reduce variance. Simulation parameters, including sensor noise and depth resolution, were kept constant across all experiments. Training, validation, and test sets were strictly separated to avoid data leakage.

2.4. Data Processing and Model Formulation

RGB-D inputs were converted into region-level feature vectors using a shared visual encoder. These features formed a graph in which nodes represented spatial regions and edges represented possible transitions. For each edge, the model predicted a passability probability conditioned on the current object configuration. Model training minimized a binary cross-entropy loss between predicted passability and observed outcomes,

$$\mathcal{L}=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i)+(1-y_i)\log(1-p_i)\right],$$

where p_i is the predicted probability and y_i denotes the observed label. Navigation efficiency was computed as

$$E=\frac{L_{shortest}}{L_{executed}},$$

 doi:10.20944/preprints202601.1852.v1

where $L_{shortest}$ is the length of the shortest feasible path under the current object configuration, and $L_{executed}$ is the path length taken by the agent.

2.5. Training Procedure and Implementation

Training was performed using mini-batch optimization with a fixed learning rate and weight decay to control overfitting. Trajectories were shuffled at each epoch. Early stopping was applied based on validation loss. The graph encoder was trained from random initialization, while the visual encoder was shared across all experimental settings. During training, object positions were continuously varied to expose the model to both blocked and unblocked transitions. All experiments were conducted under identical hardware and software conditions to ensure consistency. Hyperparameters were selected using the validation set and kept unchanged during testing.

3. Results and Discussion

3.1. Navigation Performance Under Object-Related Layout Changes

The proposed traversable scene structure improved navigation reliability in indoor environments where object movement altered route feasibility. Compared with a static scene-graph representation, the method reduced failures caused by blocked transitions by 28.4% and increased average navigation efficiency by 16.7%. These results show that a large fraction of navigation errors originates from incorrect assumptions about passability rather than from perception noise alone. When passability is modeled as a function of object state, transitions that become blocked are less likely to be selected during planning [19]. Similar observations have been reported in object-aware navigation studies, where explicit structural reasoning improves robustness in cluttered indoor layouts (Figure 1).

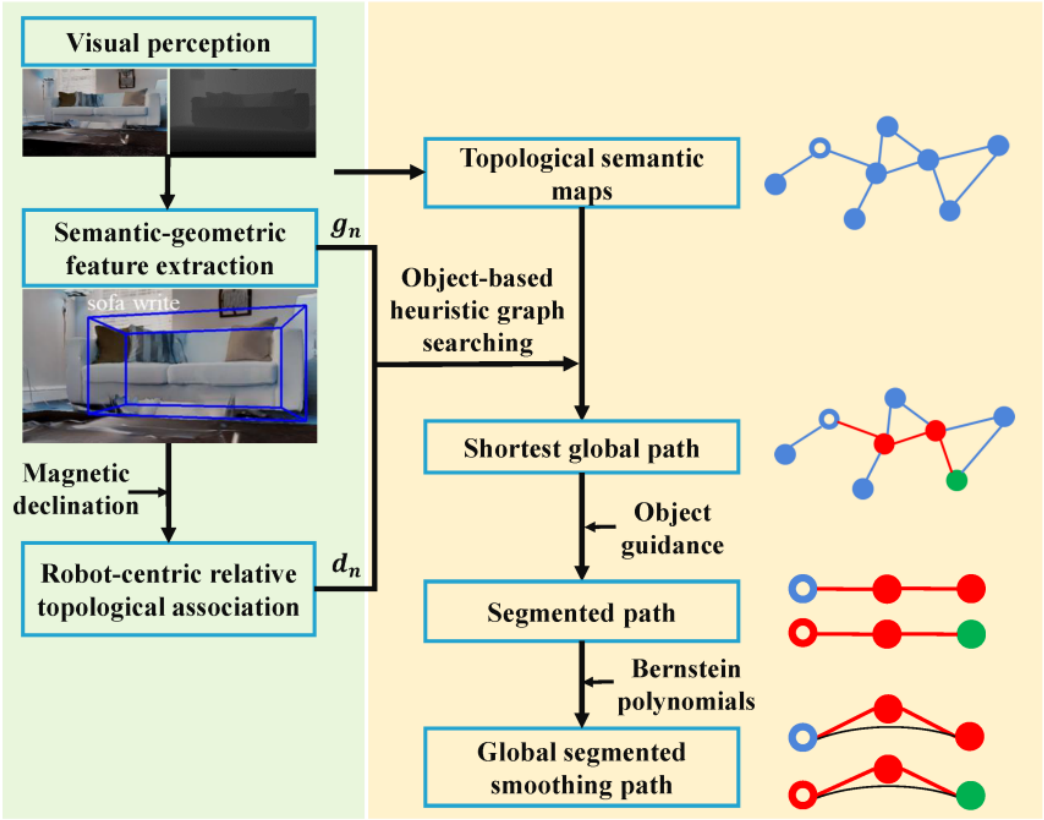

Figure 1. Learned scene structure for navigation, in which passability between regions is predicted based on the current state of movable objects.

 doi:10.20944/preprints202601.1852.v1

3.2. Comparison with Static Graphs and Replanning-Based Baselines

Replanning over a fixed connectivity graph reduced minor detours but did not prevent repeated failures near narrow passages and doorways. This limitation is inherent to static representations. If an edge is assumed to be feasible, replanning cannot correct that assumption when the underlying connectivity does not change. In contrast, the proposed method updated region-to-region feasibility using predictions conditioned on object configuration and interaction outcomes. This reduced repeated collision attempts and shortened backtracking segments [20,21]. The difference was most evident at corridor junctions and doorway regions, where early route choice has a strong effect on total path length (Figure 2).

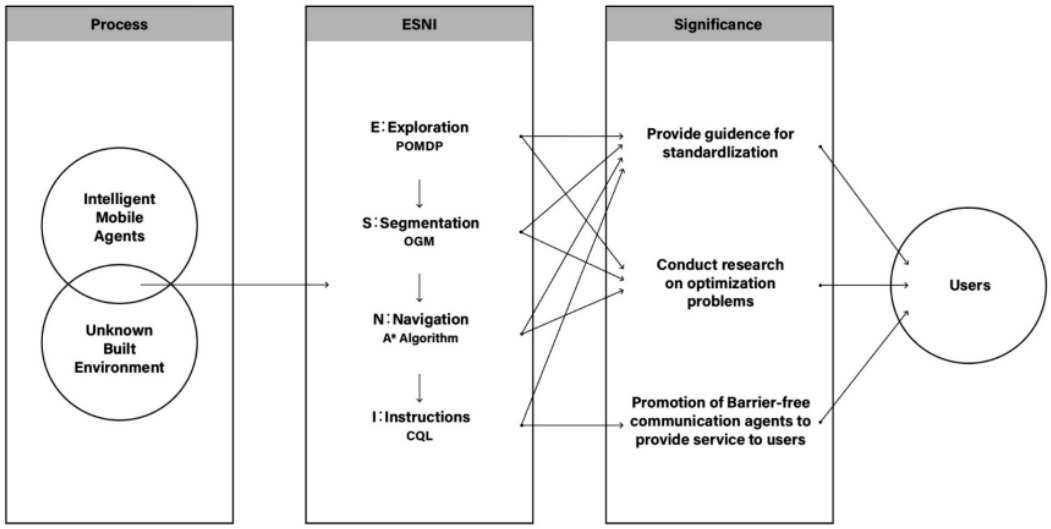

Figure 2. Comparison of navigation results showing fewer blocked-path failures and higher efficiency when using state-dependent passability instead of fixed connectivity.

3.3. Sources of Efficiency Improvement

The improvement in navigation efficiency was mainly due to fewer late-stage reversals. In dynamic scenes, static graphs often guide the agent along a geometrically short route that becomes blocked near the end of the trajectory [22]. This forces the agent to backtrack after substantial progress has already been made. By contrast, conditional passability discouraged early commitment to transitions with a high risk of obstruction. As a result, the agent followed more stable routes and showed fewer oscillations between neighboring regions. This behavior also reduced the variance of episode length across test scenes, especially in layouts with frequent object rearrangement.

3.4. Implications and Remaining Limitations

These results indicate that assuming fixed connectivity is often unsuitable for indoor navigation in environments shaped by everyday object movement. Learning a conditional traversability structure provides a reusable planning component that complements policy learning by making current feasibility explicit. However, limitations remain. Rare object configurations that are poorly represented during training can still lead to uncertain feasibility estimates. In addition, sudden object motion may temporarily invalidate predictions until new interaction feedback is obtained [23]. Future work should increase object diversity and examine calibration, assessing whether predicted passability matches observed success rates under unseen rearrangement patterns.

 doi:10.20944/preprints202601.1852.v1

4. Conclusion

This study examined how movable objects affect navigation in indoor environments and proposed a learning-based method to represent traversability as a state-dependent scene structure. The model predicts passability between spatial regions using object configuration and interaction feedback. Experiments showed that this representation reduced navigation failures caused by blocked paths and improved route efficiency compared with static connectivity. These results indicate that many navigation errors arise from incorrect assumptions about feasibility rather than from sensing limits alone. From a methodological view, the study treats traversability as a relational property that changes with object state, which links navigation and interaction at the representation level instead of within the policy. The learned structure can be reused as a planning component for embodied agents operating in cluttered indoor spaces such as homes and offices. However, the method depends on sufficient coverage of object configurations during training, and uncommon or fast-changing layouts may still reduce prediction reliability. Future work should expand object diversity, improve calibration under unseen rearrangements, and support online updates to better handle sudden environmental changes.

References

1. Raychaudhuri, S., & Chang, A. X. (2025). Semantic mapping in indoor embodied ai–a comprehensive survey and future directions. arXiv preprint arXiv:2501.05750, 3.
2. Yang, J., Chen, T., Qin, F., Lam, M. S., & Landay, J. A. (2022, April). Hybridtrak: Adding full-body tracking to vr using an off-the-shelf webcam. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-13).
3. Radhakrishnan, S., & Gueaieb, W. (2024). A state-of-the-art review on topology and differential geometry-based robotic path planning—part I: planning under static constraints. International Journal of Intelligent Robotics and Applications, 8(2), 435-454.
4. Bai, W., Wu, Q., Wu, K., & Lu, K. (2024). Exploring the Influence of Prompts in LLMs for Security-Related Tasks. In Workshop on Artificial Intelligence System with Confidential Computing (AISCC 2024)(San Diego, CA). USA. https://dx. doi. org/10.14722/aiscc.
5. Wang, Y., Feng, Y., Fang, Y., Zhang, S., Jing, T., Li, J., ... & Xu, R. (2025). HERO: Hierarchical Traversable 3D Scene Graphs for Embodied Navigation Among Movable Obstacles. arXiv preprint arXiv:2512.15047.
6. Ebrahimi Soorchaei, B., Razzaghpour, M., Valiente, R., Raftari, A., & Fallah, Y. P. (2022). High-definition map representation techniques for automated vehicles. Electronics, 11(20), 3374.
7. Mao, Y., Ma, X., & Li, J. (2025). Research on API Security Gateway and Data Access Control Model for Multi-Tenant Full-Stack Systems.
8. Saucedo, M. A., Patel, A., Saradagi, A., Kanellakis, C., & Nikolakopoulos, G. (2024, May). Belief Scene Graphs: Expanding Partial Scenes with Objects through Computation of Expectation. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 9441-9447). IEEE.
9. Mao, Y., Ma, X., & Li, J. (2025). Research on Web System Anomaly Detection and Intelligent Operations Based on Log Modeling and Self-Supervised Learning.
10. Firoozi, R., Tucker, J., Tian, S., Majumdar, A., Sun, J., Liu, W., ... & Schwager, M. (2025). Foundation models in robotics: Applications, challenges, and the future. The International Journal of Robotics Research, 44(5), 701-739.
11. Sheu, J. B., & Gao, X. Q. (2014). Alliance or no alliance—Bargaining power in competing reverse supply chains. European Journal of Operational Research, 233(2), 313-325.
12. Wen, L., Kenworthy, J., & Marinova, D. (2020). Higher density environments and the critical role of city streets as public open spaces. Sustainability, 12(21), 8896.
13. Batra, D., Chang, A. X., Chernova, S., Davison, A. J., Deng, J., Koltun, V., ... & Su, H. (2020). Rearrangement: A challenge for embodied ai. arXiv preprint arXiv:2011.01975.

 doi:10.20944/preprints202601.1852.v1

14. Hu, W. (2025, September). Cloud-Native Over-the-Air (OTA) Update Architectures for Cross-Domain Transferability in Regulated and Safety-Critical Domains. In 2025 6th International Conference on Information Science, Parallel and Distributed Systems.
15. Kessens, C. C., Kaplan, M., Rocks, T., Osteen, P. R., Rogers, J., Stump, E., ... & Srinivasa, S. S. (2022). Human-scale mobile manipulation using roman. Field Robotics, 2, 1232-1262.
16. Yang, M., Wang, Y., Shi, J., & Tong, L. (2025). Reinforcement Learning Based Multi-Stage Ad Sorting and Personalized Recommendation System Design.
17. Serban, A., Poll, E., & Visser, J. (2020). Adversarial examples on object recognition: A comprehensive survey. ACM Computing Surveys (CSUR), 53(3), 1-38.
18. Liu, S., Feng, H., & Liu, X. (2025). A Study on the Mechanism of Generative Design Tools' Impact on Visual Language Reconstruction: An Interactive Analysis of Semantic Mapping and User Cognition. Authorea Preprints.
19. Bertoli, P., Cimatti, A., Roveri, M., & Traverso, P. (2006). Strong planning under partial observability. Artificial intelligence, 170(4-5), 337-384.
20. Du, Y. (2025). Research on Deep Learning Models for Forecasting Cross-Border Trade Demand Driven by Multi-Source Time-Series Data. Journal of Science, Innovation & Social Impact, 1(2), 63-70.
21. Lagriffoul, F., Dimitrov, D., Saffiotti, A., & Karlsson, L. (2012, October). Constraint propagation on interval bounds for dealing with geometric backtracking. In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems(pp. 957-964). IEEE.
22. Ivanovic, B., & Pavone, M. (2019). The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 2375-2384).
23. Liu, M., Tang, S., Li, Y., & Rehg, J. M. (2020, August). Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In European conference on computer vision (pp. 704-721). Cham: Springer International Publishing.