Long-Term Planning Around Humans in Domestic Environments with 3D Scene Graphs

1st Ermanno Bartoli
KTH Royal Institute of Technology
bartoli@kth.se

2nd Dennis Rotondi
University of Stuttgart
dennis.rotondi@ki.uni-stuttgart.de

3rd Kai O. Arras
University of Stuttgart
kai.arras@ki.uni-stuttgart.de

4th Iolanda Leite
KTH Royal Institute of Technology
iolanda@kth.se

***Abstract*—Long-term planning for robots operating in domestic environments poses unique challenges due to the interactions between humans, objects, and spaces. Recent advancements in trajectory planning have leveraged vision-language models (VLMs) to extract contextual information for robots operating in real-world environments. While these methods achieve satisfying performance, they do not explicitly model human activities. Such activities influence surrounding objects and reshape spatial constraints. This paper presents a novel approach to trajectory planning that integrates human preferences, activities, and spatial context through an enriched 3D scene graph (3DSG) representation. By incorporating activity-based relationships, our method captures the spatial impact of human actions, leading to more context-sensitive trajectory adaptation. Preliminary results demonstrate that our approach effectively assigns costs to spaces influenced by human activities, ensuring that the robot's trajectory remains contextually appropriate and sensitive to the ongoing environment. This balance between task efficiency and social appropriateness enhances context-aware human-robot interactions in domestic settings. Future work includes implementing a full planning pipeline and conducting user studies to evaluate trajectory acceptability.**

***Index Terms*—long term planning, 3d semantic scene graphs, aware motion planning**

I. Introduction

Long-term Human-Robot Interaction (HRI) aims to create robots that continuously adapt their behavior by learning from ongoing interactions with humans [1]. This capability is essential for assistive robots operating in domestic environments, where they must not only execute tasks but do so in a way that is sensitive to human preferences. Effective robot behavior in these settings requires understanding not just spatial constraints but also the activities humans engage in and how these activities influence the environment.

A fundamental challenge in such environments is motion planning, which extends beyond simple obstacle avoidance to ensuring socially appropriate navigation. Robots must account for human activities and preferences, making decisions that respect explicit instructions, such as "watch out for the glass table; it could break," and implicit contextual cues, such as "don't spill the glass of wine," suggesting caution in the navigation. For example, a robot performing a cleaning task should avoid obstructing the human's line of sight while watching television or interfering with their ongoing activities. Achieving this requires a fine-grained understanding of human behavior and its spatial implications.

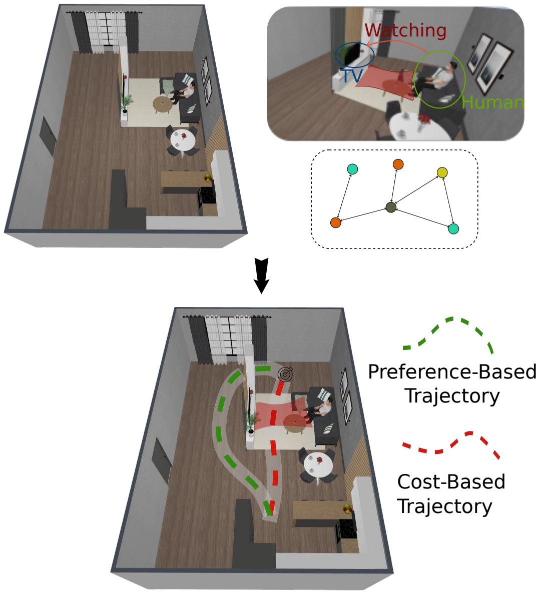

Fig. 1. Overview of the method. Starting from a 3D map and its 3D scene graph representation, our approach computes a preferred based trajectory which is socially aware of the human presence in the scene.

This complexity is further amplified by static objects in the environment, whose significance changes dynamically based on human engagement. For example, if a person watches TV, the space between them and the screen becomes socially significant, requiring the robot to avoid passing through it unless necessary. Similarly, an armchair that is currently unoccupied differs in relevance from one actively used by a human. These cases suggest the need for an approach that evaluates an object's relevance based on the human's current activities and involvement with it.

In the context of social navigation, this presents several challenges: balancing task efficiency with social appropriateness, adapting to dynamic human behaviors, and ensuring long-term generalization. Traditional motion planning in human-shared environments primarily focused on respecting personal space, but respecting social norms is equally important [2], [3]. Moreover, as human activities continuously reshape the environment, static objects take on varying importance [4], [5], requiring robots to interpret these dynamics accurately. Learning implicit human preferences over time adds another layer of complexity, necessitating structured representations that can store and infer relevant information [6], [7]. A key challenge remains generalization, as models trained in controlled settings often fail to transfer effectively to real-world scenarios [8], [9].

To address these issues, we propose an approach that integrates human preferences, activities, and spatial configurations into a 3D scene graph representation. Unlike purely vision-language-based models, which provide rich semantic understanding but lack structured spatial reasoning, 3D scene graphs enable both real-time interpretation and long-term planning. By capturing human activities as graph relationships, robots can dynamically adjust their motion plans based on human presence, ensuring socially aware navigation that respects both the immediate and evolving context. This structured yet adaptable representation allows for more intuitive and personalized human-robot interactions in shared environments, bridging the gap between spatial reasoning and socially intelligent behavior.

II. Related Work

A. 3D Scene Graph Resources

3D scene graphs (3DSGs) [10], [11] are designed for robotics, featuring a hierarchical structure where nodes represent scene parts like rooms and floors. At the lowest level, objects are connected by spatial (e.g., object1 *is next to* object2) and comparative relationships (e.g., object1 *is larger than* object2). Currently, only two datasets include 3D scene graphs [11], [12], focusing on standardizing object classes via WordNet [13]. However, they don't incorporate active base relationships, which are present in semantic inventories [14]–[16]. These datasets are static, lacking human presence, which limits robotic applications in dynamic, human-populated environments. To address this, 3D Dynamic Scene Graphs [17]–[21] have been introduced to account for dynamic scenes and agents like humans. However, the focus has mostly been on tasks like trajectory prediction and autonomous driving, with limited exploration of reasoning over dynamic 3D scene graphs.

B. Planning with 3D Scene Graphs

The 3D scene graph structure, often built on SLAM or processed images and point clouds, is as powerful as any pre-learned representation, making it ideal for robotics applications like localization [22], [23], navigation [24]–[26], and planning [27]–[31]. Planning typically involves using a large language model (LLM) to interpret a text-based 3D scene graph with object positions, bounding boxes, labels, and relationships. This helps robots understand goals and plan tasks, such as moving objects to reach their goal. A less explored but important area is incorporating object affordances and attributes into decision-making. For example, this could guide navigation to avoid fragile objects or stepping on valuable surfaces like carpets or clothes. Current 3DSG planners do not account for dynamic agents and their interactions with the environment, limiting their use to static scenes. This is suitable for environments with minimal human involvement but inadequate for dynamic settings like homes or offices.

C. Social Navigation

Social navigation is a broad research area that has evolved significantly over the years. Early works primarily focused on learning the social use of space through the lens of proxemics, studying how humans naturally maintain spatial boundaries and personal space [32]–[34]. A key advancement in the field has been the incorporation of richer environmental representations, which allow robots to better understand and navigate spaces in a socially acceptable manner [2], [3]. Approaches to social navigation have been developed using both supervised and unsupervised learning techniques [4]. The problem has been explored in both outdoor [9] and indoor [8] settings, with a particular focus on crowded spaces [5] and human-robot encounters [35]. However, a persistent challenge across these approaches has been generalization, ensuring that learned behaviors remain effective across diverse environments and interactions.

With the advent of LLMs and vision-language models (VLMs), significant progress has been made toward richer semantic understanding, enabling more complex robot behaviors informed by contextual cues [36]. Early works identify navigation targets [37], and recent advancements have focused on using them to guide low-level navigation behaviors, ensuring that robots adapt their movements in a socially appropriate manner based on the specific scenario [38].

While LLMs and VLMs enhance contextual understanding by capturing human presence and activities, they lack an explicit, structured representation that supports long-term planning. We foresee that this structure can be provided by extending 3DSG representations to include humans and relate them to their surroundings.

III. Proposed Methodology

A. Preliminary Considerations

We consider a scene represented by a 3DSSG as provided in [12], where nodes correspond to objects in the environment and edges define spatial relationships between them (e.g., ”on top of”, ”next to”, ”hanging on”). However, existing 3DSSG datasets lack human presence. As a preliminary step, we manually introduce one or more humans into the scene and integrate them into the scene graph based on their chosen activities. This is done by creating a new node labeled ”human” and associating it with relevant objects through both spatial and activity-based relationships. Spatial relationships (e.g., ”sitting on”, ”standing next to”) define the human's

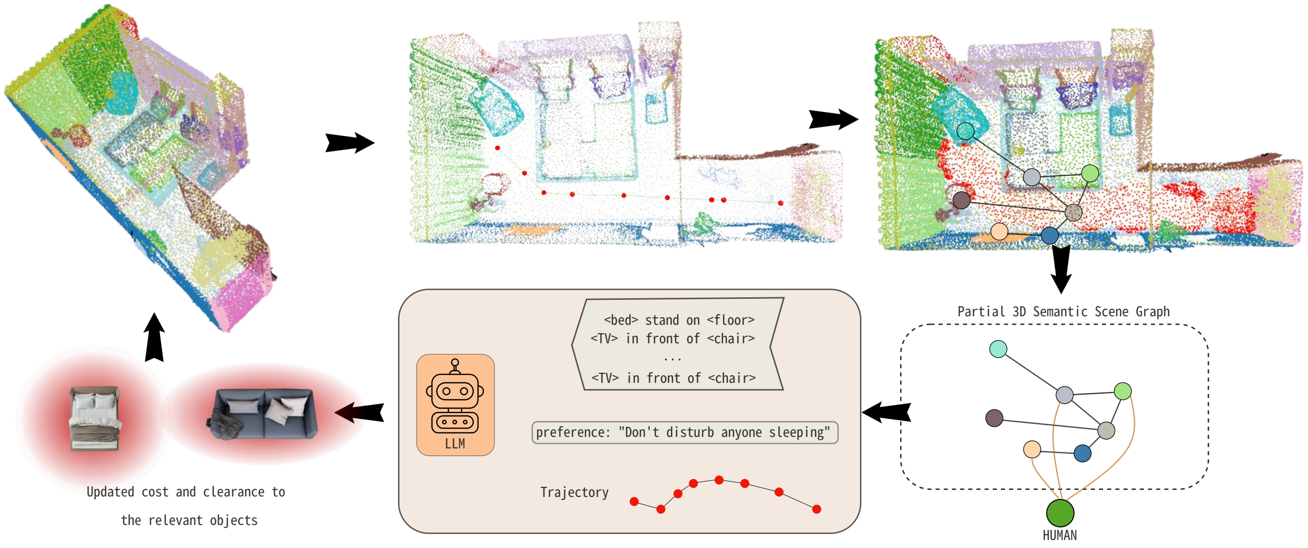

Fig. 2. Our proposed approach constructs an object-centric description of impact factors, considering both objects and spaces, through a human-centered investigation of the scene. Starting from a 3D map, a trajectory, and a set of preferences, we: (a) extract the Partial 3D Semantic scene graph with objects that could potentially impact the trajectory; (b) enrich the graph by incorporating the human and their activities, integrating with existing nodes into the 3D scene graph; and (c) feed the enriched graph representation, along with the trajectory and preferences, into a Large Language Model (LLM). The LLM then calculates for each object of interest a cost, combined with a clearance value, that describes how the cost decreases with distance.

physical interaction with the environment, while activity-based relationships (e.g., ”reading”, ”watching”, ”speaking”) capture contextual interactions with objects. The result is a 3DSSG that integrates human presence.

B. Planning Appropriate Trajectories

We define the problem of trajectory planning in a shared human-robot environment as a scenario where the robot must navigate from a starting point (A) to a goal point (B) while accounting for human activities that may impact the scene. Additionally, the robot may be provided with explicit or implicit preferences that it must incorporate during the planning process.

Given the 3DSSG, along with the robot’s task, starting and goal points, and relevant preferences, the robot’s objective is to plan a trajectory that aligns with these preferences and is appropriate for the human presence and activities within the environment.

Given a trajectory, which consists of a set of waypoints described by $T = \{p_1, p_2, p_3, \ldots, p_n\}$ for $i = 1, 2, 3, \ldots, n$ where n is the total number of waypoints along the trajectory, we investigate the potential objects that could impact the trajectory. This is done by searching within a defined radius around each waypoint along the trajectory, which can be adjusted as needed, as shown in Fig. 2. Once the relevant objects are identified, they are processed, and the following description is generated for each object. Each object is represented by:

- ``object_id``: the id of the object
- ``object_tag``: the label of the object
- ``bbox_center``: centroid of the 3D bounding box for the object
- ``bbox_extent``: extents of the 3D bounding box for the object
- ``affordances``: The set of the affordances of the object
- ``attibutes``: The set of the attributes of the object
- ``relations``: a set of tuples (name of relation, tail entity)

Consequently, the Partial 3DSSG, enriched with the human’s information, the trajectory and the preferences are inputted to an LLM which is responsible to return for each relevant object a cost and a clearance. The cost is a value equal or greater than 1 (1 if no impact at all), and reflects the impact factor of the object in the trajectory, while the clearance is a value equal or greater than 0 (0 if no impact at all) and acts as a diminishing factor for the cost, reducing the impact as the robot moves farther from the object. This ensures that the robot adjusts its trajectory based on both the proximity to objects and human preferences, maintaining safety and efficiency.

The computed costs and clearances can be integrated into a cost-based planner to generate an optimal, yet human-aware trajectory. Since these values reflect activity-based influences and spatial relationships, the resulting trajectory would inherently respect human presence and preferences. Our approach focuses on cost assignment via an LLM, while planning remains an extension.

IV. Preliminary Findings

We evaluated our approach using the scene shown in Fig. 2, where a human was manually placed in the scene, sitting on the bed and watching TV. The costs of the relevant nodes along the trajectory were extracted, focusing on the following objects: bed, human, armchair. We compared our approach—incorporating human information and both spatial and activity-based relationships—against two baseline planners: one using the 3DSSG without human information, and another using the 3DSSG with human information but excluding activity-based relationships. The preliminary results are presented in Table I.

TABLE I
TABLE TYPE STYLES

	Cost (Clearance)*		
	Bed	***Human***	***armchair***
No Human	1 (0.5)	-	2 (1.5)
Human w/out relations	2 (0.5)	10 (2)	3 (1)
Human** w/ relations	3 (1.5)	5 (2)	1 (0)

*Preference: Don't disturb anyone watching a football match

**The human is sitting on the bed, watching TV.

TABLE II
THIS TABLE SHOWS THE COSTS AND CLEARANCE OF THE NODES BED,HUMAN, AND ARMCHAIR FOR THE TRAJECTORY SHOWED IN FIG. 2, GIVEN THE HUMAN POSITION AND PREFERENCES

In this scenario, the human is sitting on the bed and watching TV. Notably, without considering activity-based relationships, the armchair is assigned an unnecessarily high cost, despite no one occupying it. This occurs because, without relational context, the system cannot infer that the human is already seated elsewhere. When relations are incorporated, the cost of the armchair is minimized, as the system recognizes that the human is sitting on the bed instead. Although the human's cost decreases from 10 to 5 when relations are considered, it remains high, with the same clearance value, meaning that the planner's behavior would be unaffected.

V. CONCLUSION

In this work, we explored how augmenting 3D scene graphs of static scenes (without dynamic agents) plays a key role in formulating plans that are socially aware of human behaviors and produce results that would otherwise be impossible without human-centered considerations. To evaluate this, we manually generated new 3D scene graph structures, placing humans in coherent relationships with objects in the scene, and observed how an LLM-based planner adapts under these conditions. In future extensions of this work, we aim to integrate the computed costs into a planner to generate trajectories and evaluate their effectiveness. Additionally, we plan to conduct a user study to assess the acceptability of these trajectories by comparing our approach with baseline methods. Participants will evaluate the generated trajectories in context, providing insights into human preferences and the perceived social appropriateness of different planning strategies. This approach lays the foundation for integrating 3D scene graph relationships into planning, enabling more context-aware and socially intelligent robot navigation.

REFERENCES

[1] Khadija Shaheen, Muhammad Abdullah Hanif, Osman Hasan, and Muhammad Shafique. Continual learning for real-world autonomous systems: Algorithms, challenges and frameworks. *Journal of Intelligent & Robotic Systems*, 105(1):9, 2022.

[2] Thibault Kruse, Amit Kumar Pandey, Rachid Alami, and Alexandra Kirsch. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, 61(12):1726–1743, 2013.

[3] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020.

[4] Matthias Luber, Luciano Spinello, Jens Silva, and Kai O Arras. Socially-aware robot navigation: A learning approach. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 902–907. IEEE, 2012.

[5] Adarsh Jagan Sathyamoorthy, Utsav Patel, Moumita Paul, Nithish K Sanjeev Kumar, Yash Savle, and Dinesh Manocha. Comet: Modeling group cohesion for socially compliant robot navigation in crowded scenes. *IEEE Robotics and Automation Letters*, 7(2):1008–1015, 2021.

[6] Hong Jun Jeon, Smitha Milli, and Anca D. Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning, 2020.

[7] Simon Holk, Daniel Marta, and Iolanda Leite. Predilect: Preferences delineated with zero-shot language-based reasoning in reinforcement learning. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 259–268. ACM, March 2024.

[8] Mark Pfeiffer, Ulrich Schwesinger, Hannes Sommer, Enric Galceran, and Roland Siegwart. Predicting actions to act predictably: Cooperative partial motion planning with maximum entropy models. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2096–2101. IEEE, 2016.

[9] Fabian Schilling, Xi Chen, John Folkesson, and Patric Jensfelt. Geometric and visual terrain classification for autonomous mobile navigation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2678–2684. IEEE, 2017.

[10] Ue-Hwan Kim, Jin-Man Park, Taek-jin Song, and Jong-Hwan Kim. 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. *IEEE Transactions on Cybernetics*, 50(12):4921–4933, December 2020.

[11] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R. Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera, 2019.

[12] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions, 2020.

[13] George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[14] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.

[15] Andrea Di Fabio, Simone Conia, and Roberto Navigli. VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China, November 2019. Association for Computational Linguistics.

[16] Roberto Navigli, Marco Lo Pinto, Pasquale Silvestri, Dennis Rotondi, Simone Ciciliano, and Alessandro Scirè. NounAtlas: Filling the gap in nominal semantic role labeling. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16245–16258, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[17] Antoni Rosinol, Andrew Violette, Marcus Abate, Nathan Hughes, Yun Chang, Jingnan Shi, Arjun Gupta, and Luca Carlone. Kimera: from slam to spatial perception with 3d dynamic scene graphs, 2021.

[18] Antoni Rosinol, Arjun Gupta, Marcus Abate, Jingnan Shi, and Luca Carlone. 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans, 2020.

[19] Nicolas Gorlo, Lukas Schmid, and Luca Carlone. Long-term human trajectory prediction using 3d dynamic scene graphs, 2024.

[20] Daniel Honerkamp, Martin Büchner, Fabien Despinoy, Tim Welschehold, and Abhinav Valada. Language-grounded dynamic scene graphs for interactive object search with mobile manipulation. *IEEE Robotics and Automation Letters*, 9(10):8298–8305, October 2024.

[21] Elias Greve, Martin Büchner, Niclas Vödisch, Wolfram Burgard, and Abhinav Valada. Collaborative dynamic 3d scene graphs for automated driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, page 11118–11124. IEEE, May 2024.

[22] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning, 2023.

[23] Yang Miao, Francis Engelmann, Olga Vysotska, Federico Tombari, Marc Pollefeys, and Dániel Béla Baráth. Scenegraphloc: Cross-modal coarse visual localization on 3d scene graphs, 2024.

[24] Zachary Ravichandran, Lisa Peng, Nathan Hughes, J. Daniel Griffith, and Luca Carlone. Hierarchical representations and explicit memory: Learning effective navigation policies on 3d scene graphs using graph neural networks, 2022.

[25] Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation, 2024.

[26] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In *Robotics: Science and Systems XX*, RSS2024. Robotics: Science and Systems Foundation, July 2024.

[27] Yuchen Liu, Luigi Palmieri, Sebastian Koch, Ilche Georgievski, and Marco Aiello. Delta: Decomposed efficient long-term robot task planning using large language models, 2024.

[28] Christopher Agia, Krishna Murthy Jatavallabhula, Mohamed Khodeir, Ondrej Miksik, Vibhav Vineet, Mustafa Mukadam, Liam Paull, and Florian Shkurti. Taskography: Evaluating robot task planning over large 3d scene graphs, 2022.

[29] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning, 2023.

[30] Zhirui Dai, Arash Asgharivaskasi, Thai Duong, Shusen Lin, Maria-Elizabeth Tzes, George Pappas, and Nikolay Atanasov. Optimal scene graph planning with large language model guidance, 2024.

[31] Aaron Ray, Christopher Bradley, Luca Carlone, and Nicholas Roy. Task and motion planning in hierarchical 3d scene graphs, 2024.

[32] Ross Mead and Maja J Matarić. Autonomous human–robot proxemics: socially aware navigation based on interaction potential. *Autonomous Robots*, 41(5):1189–1201, 2017.

[33] Jonathan Mumm and Bilge Mutlu. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 331–338, 2011.

[34] Leila Takayama and Caroline Pantofaru. Influences on proxemic behaviors in human-robot interaction. In *2009 IEEE/RSJ international conference on intelligent robots and systems*, pages 5495–5502. IEEE, 2009.

[35] Alexis Linard, Ilaria Torre, Ermanno Bartoli, Alex Sleat, Iolanda Leite, and Jana Tumova. Real-time rrt* with signal temporal logic preferences. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8621–8627, 2023.

[36] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications. *Challenges, and the Future*, 2023.

[37] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pages 492–504. PMLR, 2023.

[38] Adarsh Jagan Sathyamoorthy, Kasun Weerakoon, Mohamed Elnoor, Anuj Zore, Brian Ichter, Fei Xia, Jie Tan, Wenhao Yu, and Dinesh Manocha. Convoi: Context-aware navigation using vision language models in outdoor and indoor environments. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13837–13844. IEEE, 2024.