[1]Human Centered Artificial Intelligence Master [2] Participating Universities [3]Budapest Univeristy of Technology and Economics [4]HUN-REN Institute for Computer Science and Control [5]SatiNav Kft.

SATINav: Semantic Navigation with Visual Robot Foundation Models Agnostic to Map-Robot Variations [4]

Domokos Kiss, PhD
Supervisor, BME, Senior Lecturer

Márk Czimber
Author, BSc Student, HUN-REN Sztaki

Kornél Czimber, PhD
Supervisor, SOE, Associate Professor, Vice Dean

Abstract

This thesis is the continuation of a prior thesis conducted by László Freund and myself [4]. Building autonomous systems that scale in the real world, adapt to human interaction, and evolve with technological progress remains a significant challenge even today. This encouraged us to explore the possibilities. We took on robotic navigation as a multidisciplinary challenge that requires technological solutions from various domains, including spatial computing, control theory, and semantic understanding. We developed a semantic navigation framework for robots, independent from spatial representation and robot configuration. Once a map is constructed, semantic features provided by a foundation model are added to the detected objects to handle complex natural language queries, for example: "Identify the object where cold beverages are kept". We created a sparse, skeleton graph of the map capturing the underlying structure of the scene. We introduced three novel contributions: (a) a post-processing algorithm that enhances sparse skeleton graphs to produce cleaner, more structured representations, (b) a raycasting-based method that enables safe path generation specifically for ground robots, and (c) an algorithm facilitating efficient exploration of unknown environments, while incrementally building the topological graph. We partition the graph into semantically meaningful regions (e.g., rooms) and assign contextual labels using a Large Language Model. The graph's nodes serve as strategic waypoints, providing navigational guidance for the agent through the environment. I extend this with visual foundation models (general-purpose goal-conditioned visual navigation policies) trained and fine-tuned on diverse, cross-embodiment training data to facilitate zero-shot control invariant to different robot configurations. To reduce the iteration cycle during development, we implement and showcase our system in Nvidia Isaac ROS, a photorealistic simulation environment with an emphasis on seamless transfer to real-world deployment. We demonstrated that our skeleton graph construction approach represents a viable solution for both comprehensive topological representation and efficient ground robot path planning, while achieving competitive performance for aerial vehicle navigation. To comprehensively evaluate our pipeline, we conducted two distinct case studies that assess both its simulated and real-world performance. Our novelty lies in the unique integration and improved collaboration of distinct, state-of-the-art research achievements. Our future plan is to advance the field of robotics by creating flexible, modular, and future-extensible solutions that enable robots to safely and efficiently support humans in diverse environments.

1. Introduction

How can we design an autonomous robot navigation framework, that efficiently solves complex tasks, explores unknown scenes, or collaborates with humans in shared confined spaces, while remaining agnostic to system-level components such as environment representation or robot type?

In a rapidly evolving world, autonomous robots play a crucial role in aiding humans across various industries, from offices and hospitals to production lines. Many of these are mobile robots operating in dynamic, crowded environments, necessitating an understanding of natural language to interpret the world as humans do. Our prior research aimed to develop an end-to-end framework for semantic robot navigation enabling seamless interoperability between humans and robots by structuring semantic information hierarchically, facilitating effective collaboration without any loss of contextual understanding, as illustrated in fig. 1. Building on our prior work, I aim to enhance this versatile framework to enable collision free navigation in unseen environments by utilizing visual foundation models [3, 1, 2].

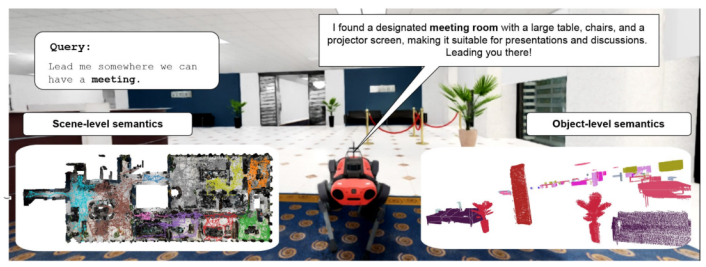

Figure 1: A comprehensive semantic understanding of indoor scenes necessitates both scene-level and object-level semantic information (fig. 2). Semantic navigation allows robots to be guided by complex, indirect natural language commands. The image illustrates how this system could be effectively deployed in a human-centered office setting.

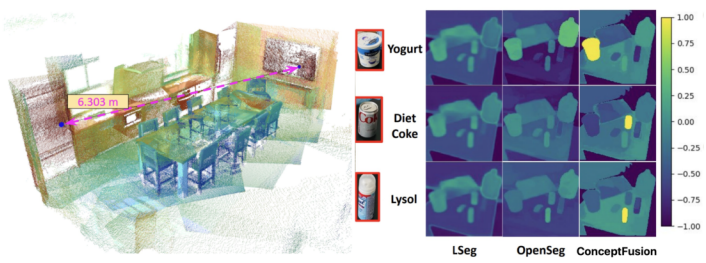

Figure 2: A key difficulty lies in the hard-to-obtain nature of high-quality semantic data, which often requires extensive manual labeling or sophisticated sensor systems. Open vocabulary systems have gained importance due to their ability to recognize and interpret a wide range of concepts, extending beyond a fixed set of predefined labels without the need for retraining on every new task or object encountered. State-of-the-art foundation models enable semantic spatial queries, such as locating specific objects or measuring distances between items with semantic awareness of the surroundings.

2. Background

How to enable robots to perform tasks based on high-level instructions to achieve navigation in complex environments considering object significance, spatial relationships, and contextual cues?

Research on semantic robot navigation has produced diverse approaches. The solution typically involves constructing a map representation enriched with semantic features, identifying navigation waypoints, associating these waypoints with relevant semantic entities, conducting waypoint searches, and guiding the robot along the generated path. We depend on these foundational components, as they are indispensable due to the inherent complexity of the task, as demonstrated in fig. 3.

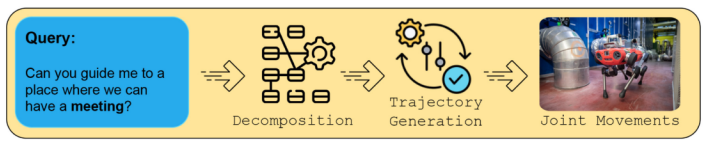

Figure 3: The image explains how difficult is to obtain actuation control from a complex natural language command. It also highlithgs there are several decompositon steps are involved to understand the surroundings and later on plan a trajectoy to the goal.

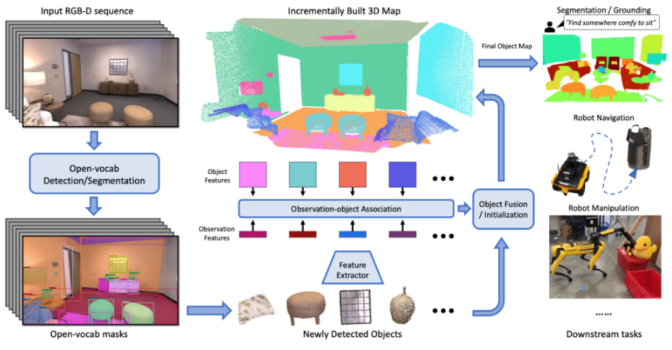

Figure 4: **ConceptFusion** integrates foundation models, combined with SLAM, to generate semantic understanding of environments through open-set, multimodal 3D mapping. First, RGB and depth images are processed, then object masks and feature embeddings aligned at the pixel level are mapped to 3D points to create a dense 3D map. Second, images, text, and audio are encoded into query vectors for language queries without requiring additional training or fine-tuning. With dense 3D mapping and multimodal querying capabilities, it supports zero-shot reasoning across a wide range of concepts.

Acknowledgments: The HCAIM (the Human-Centred AI Master's Programme) Project is co-financed by the Connecting Europe Facility of the European Union under Grant No. CEF-TC-2020-1 Digital Skills 2020-EU-IA-0068. This poster was created as part of the Blended Intensive Programme organised under the Erasmus+ Programme of the European Union.

3. Methodology

In our workflow (fig. 5), a map representation is constructed using ground truth poses from a simulation engine, which is replaced by Simultaneous Localization and Mapping in real environments. Each component in the pipeline is invariant to changing any other. For the semantic navigation problem we propose a Semantics-Augmented Topologic Inference framework for Navigation, **SatiNav** fig. 5.

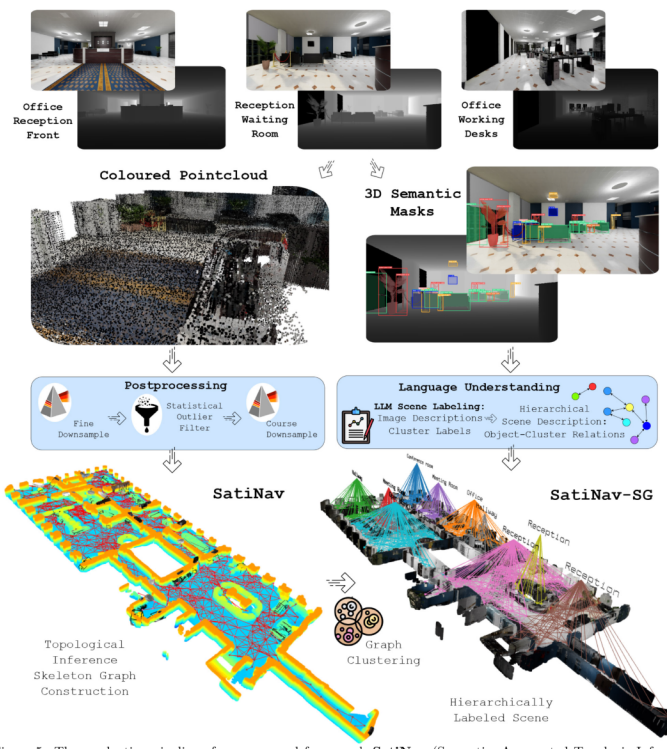

Figure 5: The production pipeline of our proposed framework **SatiNav** (Semantics-Augmented Topologic Inference for Navigation). First, RGB-D frames, along with semantic segmentation masks are projected into 3D space. Next, a topology graph is constructed. The nodes of the graph are organized into clusters, which are then augmented to construct **SatiNav-Scene Graph** for hierarchical description of the scene.

The topological map of free space points is constructed using an enhancement of SkeletonFinder; a novel algorithm **SkeletonProcessor** and **SkeletonExplorer** to generate efficient and coherent graphs for navigation or online exploration. We employ **Llama 3 8b** as the core LLM, configured with a system prompt defining its task (fig. 6). With these graphs Visual foundation models offer a strong alternative for local navigation (fig. 7) and collision avoidance. While reinforcement learning can struggle with high-dimensional latent spaces, they can achieve robust generalization upon sufficient trajectory data.

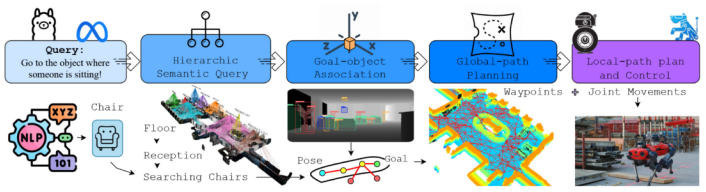

Figure 6: SatiNav-LI (Language Instructions): The figure illustrates the query pipeline, demonstrating a hierarchical search for chairs to identify objects and locate the nearest required object.

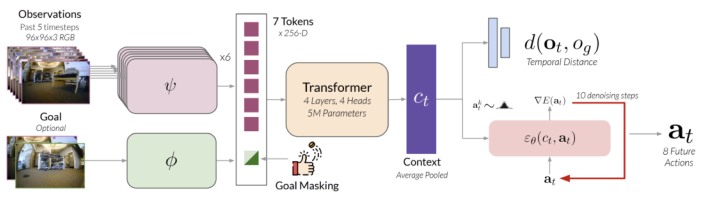

Figure 7: NoMaD [1] is the first flexibly conditioned diffusion model that can perform both goal-conditioned navigation and undirected exploration in previously unseen environments. It uses goal masking to condition on an optional goal image, and a diffusion policy to model complex, multimodal action distributions in challenging real-world environments.

4. Results

We show that **SkeletonProcessor** and **SkeletonExplorer** provide efficient, comprehensive graphs for indoor navigation in both 2D (ground) and 3D (aerial) settings, outperforming three baselines: **grid-based**, **probabilistic roadmap**, and **SkeletonFinder**. Evaluations used pathfinding success rates, normalized path length, and computational cost, along with key metrics like reachable object pairs and path-search times. Both achieved higher success rates and better connectivity, particularly for ground navigation, while remaining competitive for aerial navigation.

	Nodes	Neighbors	Success %	Find path Rel.Length	Find path Time ($\frac{ms}{m}$)	Shorten path Rel.Length	Shorten path Time ($\frac{ms}{m}$)
PRM 3D (2k points)	1437	11	14.75%	1.62	3,	1.52	1.09
PRM 3D (4k points)	3990	23	90.25%	**1.64**	2.43	**1.59**	1.57
SkeletonFinder3D	1203	**2**	**91.90%**	1.89	**0.29**	1.74	1.60
SkeletonProcessor3D	**830**	12	90.87%	**1.63**	0.65	**1.58**	**1.18**
Grid 2D (0.5m)	4293	6	85.46%	**1.66**	7.94	**1.59**	4.57
Grid 2D (0.75 m)	613	6	6.01%	1.25	0.23	1.17	1.95
PRM 2D (1k points)	584	10	61.4%	2.19	0.34	2.13	1.89
PRM 2D (2k points)	1496	20	85.51%	1.89	0.31	1.75	**1.51**
SkeletonProcessor2D	**820**	8	90.72%	**1.72**	**0.14**	1.69	2.15
SkeletonExplorer2D	968	**2**	**92.17%**	2.13	**0.18**	1.99	3.16

Table 1: **Comparison of graph construction algorithms in 3D above and 2D below.**

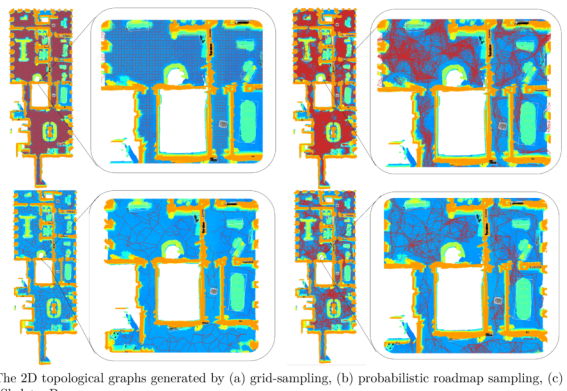

Figure 8: The 2D topological graphs generated by (a) grid-sampling, (b) probabilistic roadmap sampling, (c) SkeletonExplorer, (d) SkeletonProcessor.

There is a demand for navigation graphs that are able to robustly capture complex concepts like rooms, buildings, and hallways. Comprehensive topologic graphs allow us to extract such features using graph theory or data mining. A straightforward benchmark for assessing structural robustness is clustering, where each waypoint in the navigation graph is grouped based on a measure of similarity (fig. 9).

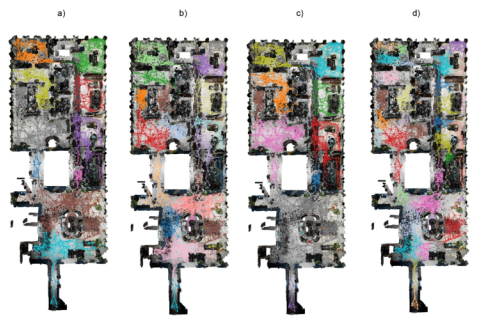

Figure 9: Topologic graphs produced by SkeletonProcessor clustered with different methods: (a) edge betweenness partitioning, (b) Louvain community clustering, (c) spectral clustering, (d) infomap clustering.

4.1 Case study: Semantic Navigation in Simulated Office Scene

The environment was well-aligned and sparsely populated, making object detection straightforward. Large language model (LLM) queries sometimes yielded spurious answers, but generally succeeded. Pathfinding produced suboptimal or collision-prone routes, underscoring the need for an improved local planner. Despite these issues, sequential queries (fig. 10) ran reliably, confirming the system's potential for future refinements while noting it is not yet fully ready for real-world deployment.

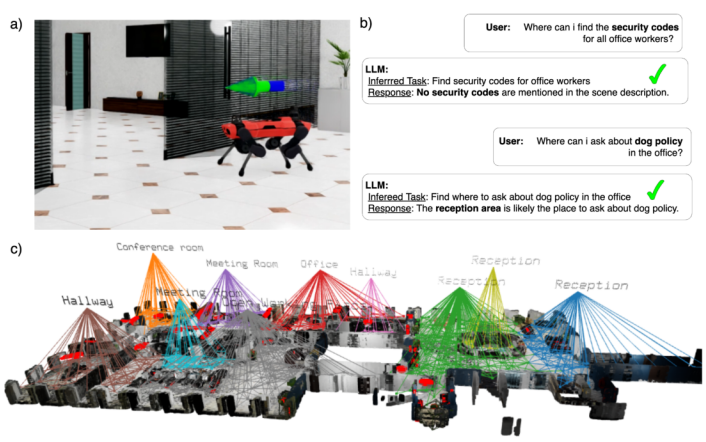

Figure 10: (a) ANYmal-C navigating in the Office environment, with a green arrow indicating its current heading direction (velocity command). (b) Example queries to the LLM: irrelevant queries are understood and denied, while achievable ones are accepted and executed. (c) SatiNav Scene Graph representation of the simulated Office environment.

4.2 Case study: Scene Graph Construction of Real World Scene

Data (fig. 11) from a real building level exhibited noise and misalignment, though ray-walking managed these drifts. Inconsistent door states added complexity, and redundant objects (e.g., many identical chairs) inflated memory. Misdetections by vision-language models and LLMs highlighted reliability gaps, underscoring the need for improved preprocessing and error handling.

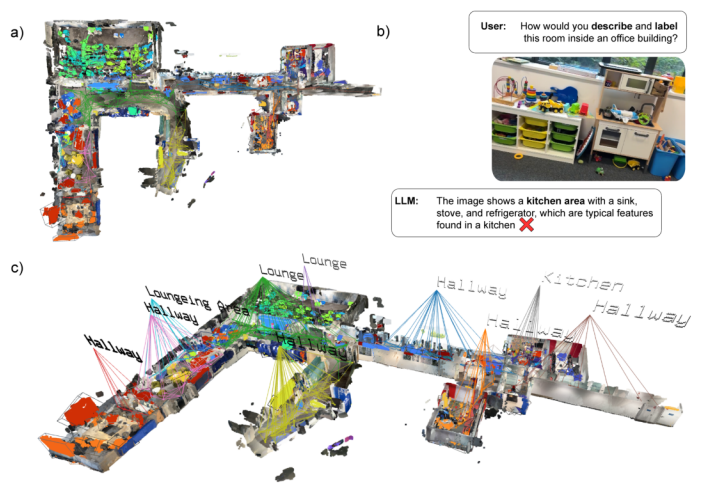

Figure 11: (a) Top-view of the scene graph generated by SatiNav, showing high object redundancy in the top left due to multiple identical green chairs. The topological graph successfully achieves full room coverage, even in real-world scenarios. (b) Image description LLM struggling to recognize a complex scene involving a toy kitchen set. (c) SatiNav Scene Graph representation of the Real World Office environment, visualized.

4.3 Ethics and Outlook

A human-centered approach ensures that robots augment rather than replace human capabilities, promoting inclusive solutions that respect individual autonomy and dignity. In practice, this involves careful data governance, meaningful stakeholder engagement, and robust fail-safes to prevent harmful outcomes. Beyond safeguarding personal data through strict encryption and compliance with relevant regulations, developers can implement "people filters" that exclude human figures from pointcloud data, preserving privacy. Thorough prototype testing in both static and human-in-the-loop dynamic environments is critical, complemented by transparent communication of system capabilities and continuous user feedback to ensure ongoing ethical compliance and adaptability.

This research focuses on collision avoidance using visual foundation models, while ongoing work addresses topology-graph optimization, object-approach point determination, and robot configuration from scratch. Large real-world maps, effective domain randomization and assembling a complete robot system opens up opportunities for sim-to-real transfer and real-world testing of the proposed framework in concrete scenarios. We at **SatiNav** plan to launch a startup around this general navigation framework (fig. 12), which adapts to a wide range of robot configurations to provide safe, efficient everyday assistance for humans. In pursuit of affordability and universal access, we aim to develop cost-effective, potentially 3D-printed robots for housecleaning, kitchen work, waitering, gardening, and food delivery. Ultimately, our goal is a *future-extensible*, *modular*, and *scalable* navigation system that can integrate seamlessly into the Industrial Metaverse.

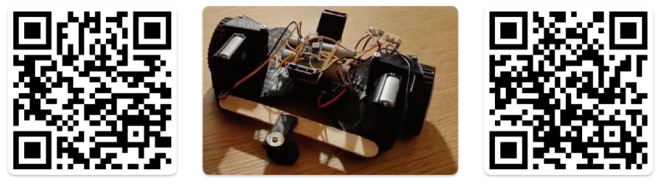

Figure 12: The first QR code points to our SatiNav "first look" demonstration, the middle image shows our initial prototype SatiBot-v1, which is designed as a waiter robot base. The last QR code links to our previous research [4], where additional references and supporting materials are available.

References

[1] Ajay Sridhar et al. Nomad: Goal masked diffusion policies for navigation and exploration, 2023.
[2] Dhruv Shah et al. GNM: A General Navigation Model to Drive Any Robot. In *International Conference on Robotics and Automation (ICRA)*, 2023.
[3] Dhruv Shah et al. ViNT: A foundation model for visual navigation. In *7th Annual Conference on Robot Learning*, 2023.
[4] László Freund and Márk Czimber. Semantic navigation agnostic to map-robot variations, 2024.