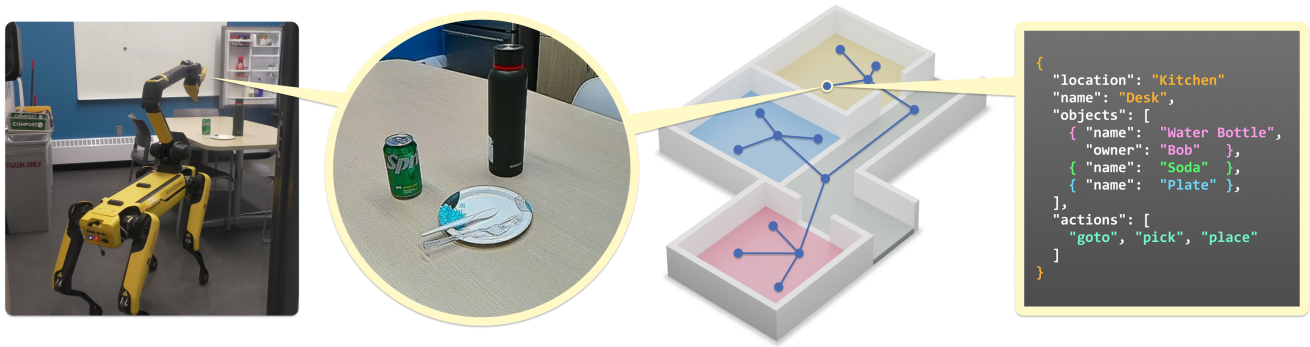
SENT-Map: Semantically Enhanced Topological Maps with Foundation Models

Raj Surya Rajendran Kathirvel[1], Zach Chavis[2], Stephen J. Guy[2], and Karthik Desingh[1,2]



(a) Operator's perspective **(b)** Robot's perspective **(c)** Topological map **(d)** Semantic node JSON

Fig. 1: **SENT-Map overview.** (**a**) An operator guides a robot throughout an environment, (**b**) stopping at task-relevant locations to take a snapshot, (**c**) which is tagged with a location and appended to a graph, and (**d**) fed to a foundation model alongside human annotation to generate a JSON description of the semantic context.

I. Introduction

We introduce SENT-Map, a semantically enhanced topological map for representing indoor environments, designed to support autonomous robot navigation and manipulation by leveraging advancements in foundational models (FMs). The general semantic information and planning capabilities found in modern FMs provide exciting potential for robots operating in complex human environments to have powerful capabilities in unconstrained, open-world environments. However, these FMs also come with notable risks related to hallucination, false confidence, and other assurance-related issues. We propose to address these issues by grounding plans from FMs in the real-world locations as represented through a topological map [1]. Additionally, we rely on both vision-language models (VLMs) and human operators to further enhance these topological maps with semantic information, allowing a robot to build robust plans to account for its navigation capabilities, manipulation affordances, and other semantic information about objects, rooms, and people in the robot's environment. By incorporating these resulting SENT-Maps with a Large-Language Model (LLM) for planning, the resulting system is able to execute a variety of navigation-manipulation tasks in complex environments specified through natural language commands.

At a high level, the proposed framework operates in two stages: a mapping stage and a planning/execution stage. The mapping stage begins with a navigational map build through an environmental walkthrough followed by the use of an FM to construct the SENT-Map. The SENT-Map construction process is human-guided in two ways: first, a human operator is responsible for walking the robot through the environment, and second, the operator highlights various semantically interesting locations and objects. The result of the mapping stage is a structured SENT-Map in JSON format, representing the environment as a topological graph with nodes containing robot affordances, navigability links, and any additional environmental information useful for mobile manipulation tasks such as people's locations or objects' ownership. Critically, these node-based maps can be easily visualized, checked, and edited by a human for not only correctness, but also supplementary semantic information not privy to FMs. The planning and execution stage is executed by a Planning FM, in this case an LLM, which takes the SENT-Map in JSON format, a description of the robot's known skills, and a natural language command as input to generate and execute task plans. These actions generated by the FM can be limited at plan creation time to those allowed by the SENT-Map, eliminating the possibility of hallucinations at plannign time related to affordances or capabilities.

In summary, our contributions are as follows:

- SENT-Map, a Semantically-Enhanced Topological Map in human-interpretable JSON for autonomous robot navigation and manipulation
- A framework for constructing and planning over SENT-Maps using existing foundation models.
- Experimental results showing SENT-Maps improve FM planning success even on locally-deployable FMs.

[1]Minnesota Robotics Institute (MnRI), and [2]Department of Computer Science and Engineering (CS&E), University of Minnesota, Minneapolis, MN 55414 US.
(rajen064|chavi014|sjguy|kdesingh)@umn.edu

II. Related Works

Liu et al. proposed FM-fusion [2], an instance-aware semantic mapping framework combining vision-language models with SLAM for camera pose estimation. While it enables open-set labeling and dense segmentation, it suffers from high computational demands, lacks embodiment-specific affordance reasoning, and provides no assurances for task execution. For building SENT-Maps, we utilize FMs to generate a map in JSON, which is human-editable to further enhance or correct the map to allow for assured task execution.

Object-centric mapping approaches [3], [4] integrate rich semantic information about objects into maps, making them highly effective for manipulation tasks. While these methods excel in supporting intricate manipulation scenarios, current foundational models are not yet sufficiently advanced for reliable open-set 3D reconstruction or pose estimation, though promising progress has been made [5]. Our proposed framework seeks to address this limitation by leveraging FM models for 2D visual semantic understanding combined with natural language utterances to construct sparse representations, as seen in SENT-Maps.

Several works leverage FMs to semantically represent environments for visual navigation. CLIP-Fields [6] maps 3D spatial locations to high-dimensional feature vectors embedding CLIP-based language and visual features, while NLMap-SayCan [7] uses a 2D grid-based map with discrete object representations derived from a region proposal network and VLM features. VLMaps [8] represents fixed object sets in a 2D grid as top-down projections, enabling spatial goal navigation using language commands. Techniques like 3D-LLMs [9] and the real-time OpenFusion [10] offer open-vocabulary 3D mapping and queryable scene representations using RGB-D data. While these methods are effective, they represent semantic information in feature space, making their maps non-verifiable by humans and limiting their ability to incorporate affordances. In contrast, our proposed framework uses FM during mapping phase to achieve open-set semantic enhancement, enabling the creation of human-verifiable editable maps.

Graph-based methods represent scenes as 3D graphs, embedding geometric and semantic information into nodes for high-level task planning. Concept Graphs [11] constructs 3D graphs by leveraging 2D foundation models and fusing their outputs into 3D through multi-view association, enabling open-vocabulary representations without extensive 3D datasets. CLIO [12] builds task-driven 3D object-centric maps, clustering object primitives into semantic regions based on task specifications. While these methods create object-centric topological maps, they require task specifications during map creation, and are not designed for editability, limiting their flexibility and usability. by enabling open-set semantic enhancement during the mapping phase and task specification through natural language interaction in the execution phase, thereby improving adaptability and task assurance.

III. Semantically Enhanced Topological Map

A. Problem Definition

Consider a topological map $\mathcal{M}$ represented as a graph $\mathcal{M} = G(V, E)$, where V denotes the nodes the robot can navigate to and E denotes the navigation actions it can perform to move between the vertices. The SENT-Map embeds additional information into a subset of the vertices $V_{SE} \subseteq V$, as illustrated in Fig. 1. Our framework has two phases and hence two problems: one, constructing the map $\mathcal{M}$, where a human walks the robot in the indoor environment, facilitating the capture of RGB observations I_v of the semantic node locations v, which are then passed through an FM (Scene Representation FM) $\mathcal{M} \leftarrow \{S_{FM}(I_v) \mid v \in V_{SE}\}$ to output the JSON-structured SENT-Map $\mathcal{M}$; and two, planning with the constructed $\mathcal{M}$, where given the SENT-Map and a natural language query, an FM (Planning FM) outputs the navigation and manipulation sequence to complete the task. Given a target node v_{target}, a current node $v_{current}$, and the map $\mathcal{M}$, we assume that there exists a path planner that gives the shortest path the robot can successfully execute to reach v_{target}.

B. Scene Representation (SENT-Map)

To enable reasoning by language models over physical environments, the SENT-Map represents the environment as a JSON-structured topological graph. It encodes spatial and semantic information in a hierarchical format, grounded in the robot's navigation and interaction capabilities. At the top level, the SENT-Map consists of navigation nodes, which represent traversable waypoints in free space. Each navigation node specifies its connectivity to other nodes via directed edges, forming the graph structure $\mathcal{M} = G(V, E)$ as shown in Fig 2 (c). Each navigational node has the potential to be a semantic node, representing nearby stationary entities such as fridges, drawers, cabinets, tables, or desks that serve as potential locations for robot interaction. As entities represented by semantic nodes may require additional context (e.g., a fridge must be opened to retrieve its contents), the semantic node can indicate how the entity's state can be changed by the robot through manipulation. These semantic nodes may also contain objects, which are movable, graspable entities such as mugs, tissue boxes, bottles, or cans; these objects are targets for high-level tasks such as "get," "move," or "clean." Each semantic node may further contain additional metadata relevant for downstream tasks, such as tagging individual objects with ownership, e.g. "Bob's mug." This structured and interpretable representation enables foundational models to reason over physical spaces using natural language, while supporting efficient navigation and task planning.

C. Foundation Models Maps with Human Grounding

To construct a SENT-Map, an operator guides a robot around an environment as shown in Fig 1 (a). As the robot moves, navigational nodes are created, memorizing the free-space of the environment. During mapping, the operator stops at desired points of interest, which become semantic nodes. To construct the semantic node, the robot first takes an

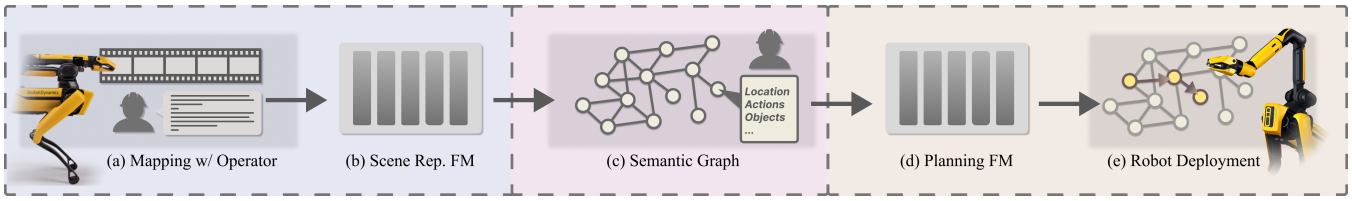


Fig. 2: **SENT-Map Framework.** (**a**) An operator defines a map alongside a robot. (**b**) Images and operator prompt are given to a Scene Representation FM, which outputs a node in JSON. (**c**) A collection of nodes defines our semantic graph. Due to the interpretability of JSON, the operator is free to make additions or corrections within the JSON. (**d**) The full JSON graph is fed to a planning FM alongside a query, and a skill sequence is output. (**e**) The robot then executes the skill sequence within the environment.

RGB snapshot using onboard cameras as shown in Fig 1 (b). The image is then passed to a Vision-FM along with a prompt defining the JSON template, i.e. the structure of a semantic node, as shown in Fig 1 (d). After mapping, the collection of semantic nodes is represented as a Scene JSON, a compressed-textual representation of the map.

As the JSON format is human-interpretable, during this mapping phase the user can also modify/rectify the JSON or add other additional information such as ownership tags ("this is person A's mug"), associated with the semantic node or any of the objects it contains to enrich the SENT-Map. Any hallucination or incorrect inference made by the mapping FM can be corrected by the operator before the planning stage. Hence, the JSON used during the planning stage will be representative of the ground truth. An example segment from a SENT-Map JSON file is shown in Fig 2 (c).

D. Planning with Foundation Models

For planning, we employ a text-only foundation model as a Planning FM, which is tasked with determining the series of steps that satisfy a task given the Scene JSON and user-defined query as shown in Fig 2 (d). Following [13], we ground the output actions to a skill API, which defines the actions the robot can take, and a description of the robot's physical constraints (i.e. single arm). The final prompt for the FM contains the scene, the skill API, the physical robot constraints, and the user query. Once the FM returns a plan, the robot executes the series of commands using the navigation nodes for global motion planning, and off-the-shelf methods for local collision avoidance and object manipulation.

IV. Experimental Results

To evaluate the impact of semantic enhancement for planning, we compare the performance on the SENT-Map environment represented in Fig 3, consists of nine different semantic nodes defined over three major zones (office, lounge and kitchen). Each room has various items for a total of 23 objects as inferred by the Vision-FM (Llama 3.2 90B Vision Instruct [14]) and verified by the operator. To evaluate the impact of semantic enhancement for planning, we compare the performance of 5 large language models: Llama 3.1 8B and 405B Instruct [14], GPT 4o mini and o3 mini [15], and Gemma 3 27B and Gemini Flash 2.0 [16], across three object retrieval scenarios: *Get-Sponge*, *Get-Coffee*, and *Get-Tissue*. The *Get-Sponge* task is an unambiguous reasoning instance where the target object (a sponge) is placed in its logically and semantically appropriate location—i.e., in the kitchen by the sink. This makes it relatively straightforward for LLMs to infer that the kitchen sink is the obvious place to search for the sponge. In the *Get-Coffee* task, we introduce a misleading association: the coffee powder is placed on a tray in the office rather than in the lounge or kitchen. This reflects a common real-world scenario where an object is not in a semantically expected location, requiring the agent to rely on a prior mapping phase to identify its placement. Finally, the *Get-Tissue* task presents a many-to-one mapping scenario, where multiple tables (at least one per location) could plausibly contain a tissue box, but only one actually does. This setting reflects a frequent occurrence in household environments, where several semantically valid locations might exist for a given object. In such cases, an accurate map with the object's current location is essential for efficient task completion. Queries are given in natural language, and the model is asked to reply with the skill sequence most likely to solve the task.

A. Baseline Performance

We define our baseline map as a Scene JSON with no semantic enhancement, meaning only top-level location information is provided (e.g., office desk, kitchen fridge), inspired by techniques such as [13]. This consists of our nine semantic nodes over three zones, but without object context. As the baseline method lacks contextual information about the objects present, the LLM is forced to infer where an object may be located based on semantic cues commonly associated with each location. Usually in large indoor environments, multiple similar locations exist, introducing semantic ambiguity. For example, when asked to retrieve a tissue, the language model understands that tables are a common location for tissues to be located, but must guess between the various tables within the environment as pictured in Fig 3, leading to inconsistent task performance. This trend is seen across all LLMs tested as shown in Table I.

In case of indirect queries – where the target object is not explicitly stated (e.g., *"I have a cold and I'm feeling a bit sniffly."*, tissue implied) – the LLM was unable to answer correctly using the baseline map as shown in Table II.

Model	Baseline				Semantic Enhancement			
	Sponge	Coffee	Tissue	*Average*	Sponge	Coffee	Tissue	*Average*
Gemma 3 27B	✓	×	✓	66.7%	✓	✓	✓	100%
Gemini 2.0 Flash	∅	∅	∅	0.0%	✓	✓	✓	100%
Llama 3.1 8B	×	×	✓	33.3%	✓	✓	✓	100%
Llama 3.1 405B	×	×	✓	33.3%	✓	✓	✓	100%
GPT 4o mini	×	×	✓	33.3%	✓	✓	✓	100%
GPT o3 mini	✓	✓	×	66.7%	✓	✓	✓	100%
Average	38.9%				100%			

TABLE I: Task success across several LLMs. A "✓" denotes task success, an "×" denotes task failure, and a "∅" denotes the model's refusal to output a solution due to requiring additional context.

Task	Baseline		Semantic Enhancement	
	Direct	Indirect	Direct	Indirect
Watch TV	×	×	✓	✓
Runny Nose	✓	×	✓	✓
Private listening	×	×	✓	✓
Sanitization	×	×	✓	✓
Call a friend	×	×	✓	✓
Flavor Coffee	✓	×	✓	✓
	SE		SE + Ownership	
Store Bob's leftovers	✓	✓	✓	✓
Get Bob his drink	✓	×	✓	✓
Bob's things to Alice	✓	×	✓	✓

TABLE II: Direct-query and indirect-query task success for small foundation model. Gemma 3 27B was prompted with two types of queries, one directly asking for the objects, and one indirectly suggesting the object without naming it. Results indicate that semantic enhancement enables even a small FM to reason about complex tasks.

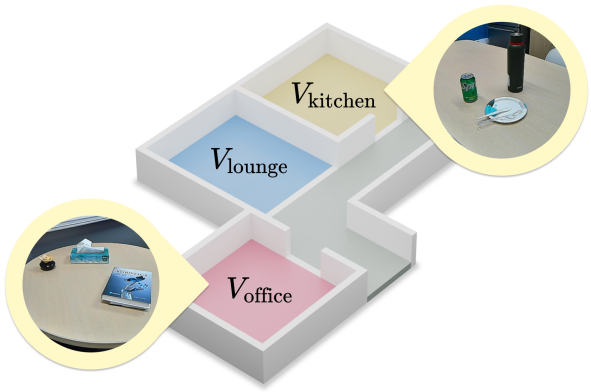


Fig. 3: **Semantic Ambiguity.** The topological map of our indoor environment contains several instances of drinks and tables, two nodes of which are pictured here. When given a task asking for a "tissue", each FM knows that a desk or table is a likely location for a tissue box, but is forced to make a guess without additional semantic context. Similarly, the FM must guess between drinks when queried for a beverage with no context of who a drink belongs to.

B. Semantic Enhancement

When provided with semantic context about the objects present at each node, all LLMs are able to correctly infer the target skill sequence even in adversarial cases like the *Get-Coffee* or *Get-Tissue* scenarios. The semantic context also allows even a relatively small LLM (like Gemma 3 27 B) to accurately plan the correct skill sequence even in cases of tricky indirect queries as shown in Table II. We also explore queries that require knowledge of object ownership. By including two people, Bob and Alice, into the Scene JSON and tagging two of Bob's items, we can plan for tasks involving someone's items or current locations. While the FM can plan for direct object queries on the SENT-Map, lack of ownership and location tagging causes hallucinations in ambiguous scenarios, such as assuming anything near Bob's is owned by him. Results are shown in Table II.

V. Conclusion and Future Work

We introduced SENT-Maps, Semantically-Enhanced Topological Maps for autonomous robot navigation and manipulation. SENT-Maps represent the environment in JSON format, enabling humans to edit and foundation models to parse the environment for downstream planning. We demonstrate a framework for constructing a SENT-Map using an operator guided mapping phase, and a planning phase, both assisted by foundation models. Through semantic enhancement, we showed foundation models were able to plan more effectively for locally-deployable (27B param) FMs and from indirect queries.

An important limitation of this work is the effort required from an operator to attain sufficient semantic enhancement. This may be especially limiting in large or complex environments where the required operator effort may not scale well with the environment. Additionally, the SENT-Maps resulting from complex environments may result in long or complex JSON files that smaller LLMs could have trouble parsing correctly or that may confuse the human operator. To address these limitations, we hope to explore a tradeoff between operator-level semantic enhancement and scalable mapping methods for robots such as Hydra [5], as well as the semantic complexity tradeoff. We also hope to further investigate the ability for users to interpret and augment the SENT-Map, allowing them to alter the map through a user interface during and after mapping, as well as while planning.

References

[1] B. Kuipers and Y.-T. Byun, "A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations," *Robotics and Autonomous Systems*, vol. 8, no. 1, pp. 47–63, 1991, special Issue Toward Learning Robots.

[2] C. Liu, K. Wang, J. Shi, Z. Qiao, and S. Shen, "Fm-fusion: Instance-aware semantic mapping boosted by vision-language foundation models," *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2232–2239, 2024.

[3] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1352–1359.

[4] Z. Zeng, Y. Zhou, O. C. Jenkins, and K. Desingh, "Semantic mapping with simultaneous object detection and localization," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE Press, 2018, p. 911–918.

[5] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3D scene graph construction and optimization," 2022.

[6] N. Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam, "Clip-fields: Weakly supervised semantic fields for robotic memory," 10 2022.

[7] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. S. Ryoo, A. Stone, and D. Kappler, "Open-vocabulary queryable scene representations for real world planning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 11 509–11 522.

[8] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," 2023.

[9] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3d-llm: Injecting the 3d world into large language models," in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 20 482–20 494.

[10] K. Yamazaki, T. Hanyu, K. Vo, T. Pham, M. Tran, G. Doretto, A. Nguyen, and N. Le, "Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation," in *ICRA*, 2024, pp. 9411–9417.

[11] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, C. Gan, C. M. de Melo, J. B. Tenenbaum, A. Torralba, F. Shkurti, and L. Paull, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," 2023.

[12] D. Maggio, Y. Chang, N. Hughes, M. Trang, D. Griffith, C. Dougherty, E. Cristofalo, L. Schmid, and L. Carlone, "Clio: Real-time task-driven open-set 3d scene graphs," 2024.

[13] R. Shah, A. Yu, Y. Zhu, Y. Zhu, and R. Martín-Martín, "Bumble: Unifying reasoning and acting with vision-language models for building-wide mobile manipulation," *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[14] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[15] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[16] S. Pichai, D. Hassabis, and K. Kavukcuoglu, "Introducing gemini 2.0: our new ai model for the agentic era," 2024.