LOG-Nav: Efficient Layout-Aware Object-Goal Navigation with Hierarchical Planning

Jiawei Hou[1], **Yuting Xiao**[2], **Xiangyang Xue**[1,*], **Taiping Zeng**[2,*]

[1]School of Computer Science, Fudan University, Shanghai, China
[2]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China
{jwhou23, ytxiao25}@m.fudan.edu.cn, {xyxue, zengtaiping}@fudan.edu.cn

Abstract

We introduce LOG-Nav, an efficient layout-aware object-goal navigation approach designed for complex multi-room indoor environments. By planning hierarchically leveraging a global topologigal map with layout information and local imperative approach with detailed scene representation memory, LOG-Nav achieves both efficient and effective navigation. The process is managed by an LLM-powered agent, ensuring seamless effective planning and navigation, without the need for human interaction, complex rewards, or costly training. Our experimental results on the MP3D benchmark achieves 85% object navigation success rate (SR) and 79% success rate weighted by path length (SPL) (over 40% point improvement in SR and 60% improvement in SPL compared to existing methods). Furthermore, we validate the robustness of our approach through virtual agent and real-world robot deployment, showcasing its capability in practical scenarios.

Code — https://github.com/fudan-birlab/LOGNav

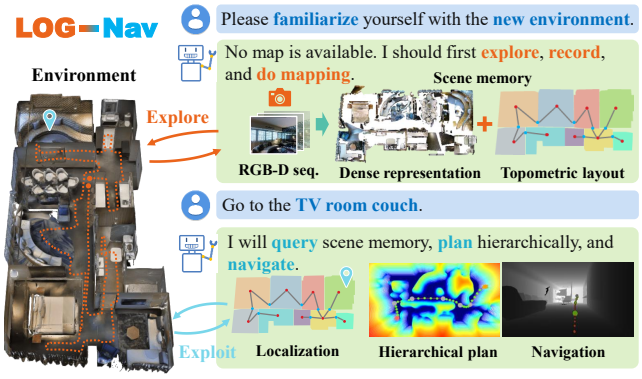

Figure 1: Our proposed LOG-Nav. An efficient object-goal navigation approach with LLM-powered agent that realizes hierarchical planning based on topology and dense scene memory. The entire process is conducted automatically without costly training or complex rewards.

Introduction

Advancements in Large Foundation Models (LFMs) and robotics research have brought household assistant robots closer to real-world deployment. A key capability for such robots is to locate a target and navigate to it based on user input (Zhu et al. 2017). Researchers have made significant progress in integrating Vision-Language Models (VLMs) to align user instructions with robot observations for prompted navigation and target localization using scene memory collected during exploration (Shah et al. 2023; Zhou et al. 2023; Dorbala et al. 2022; Chen et al. 2019; Vasudevan, Dai, and Van Gool 2021; Krantz et al. 2023). While complementary scene understanding and accurate instruction interpretation are essential for robot navigation, two mainstream approaches have emerged: process-prompted and goal-oriented navigation.

Recent advancements in robot navigation have focused on understanding and executing user instructions by leveraging Vision-Language Models (VLMs). These models enable robots to follow process-prompted instructions, allowing them to explore and navigate in unseen environments (Anderson et al. 2018b; Chen et al. 2019). However, these benchmarks are based on sparse panoramic observations, assuming known topologies, oracle navigation, and perfect localization, which do not reflect real-world deployment challenges (Krantz et al. 2020). VLN-CE has attempted to bridge this gap by removing unrealistic assumptions, translating observations into low-level controls in continuous environments (Krantz et al. 2020). The grounding of process-prompted instructions is in a step-by-step form, such as "Walk out of the bedroom. Turn right and walk down the hallway. At the end of the hallway turn left.".

While following step-by-step instructions is a necessary skill for household assistants, the goal-oriented task provides a simple optional interaction pattern for users. For instance, robots should interpret simple instructions like "Go to the couch in the living room" or a picture of the desired couch. Since ZSON (Majumdar et al. 2022) proposed the ability to locate and navigate to the object of interest in a scene based on a user inquiry, without training on this specific setting, researchers have made significant efforts to improve retrieval processes, instance recognition abilities, and scene understanding (Guan et al. 2024; Sun et al. 2024;

*Co-corresponding author

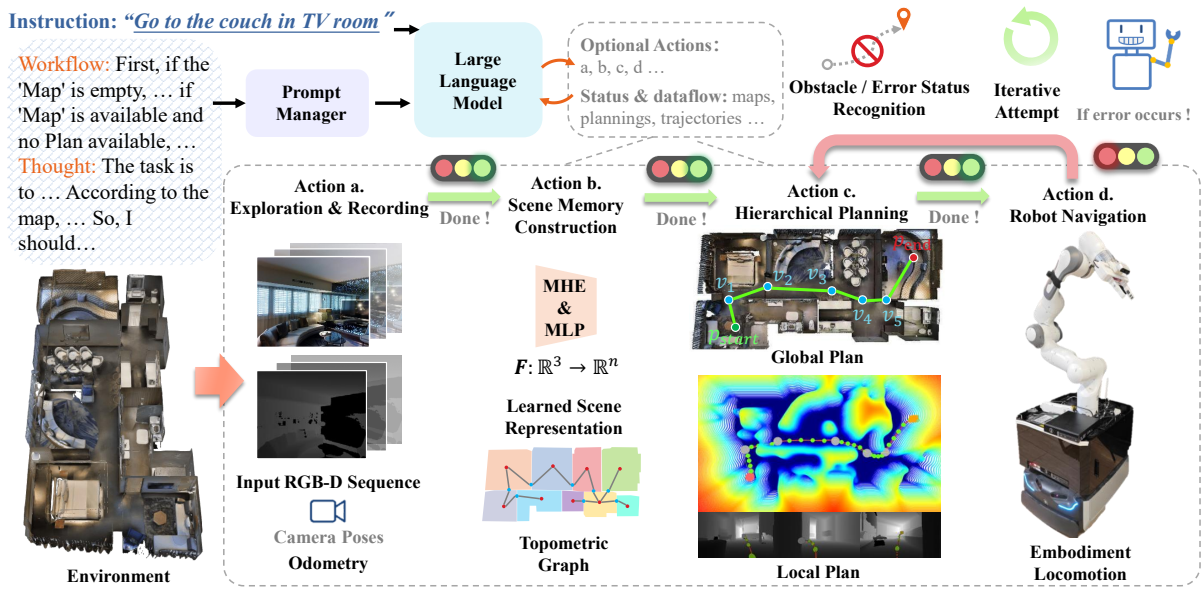

Figure 2: Overview of our proposed method. The LLM agent takes user instructions as input and manages the optional action choices according to the prompts and data flow. Optional actions include exploring and recording the scene, constructing scene memory representation, planning, and executing navigation. Obstacles, error recognition, and iterative attempts are available.

Zhou et al. 2023; Ramakrishnan et al. 2022; Yu, Kasaei, and Cao 2023). Despite these advancements, navigation in open scenes remains a challenging problem due to practical constraints. Specifically, the following challenges persist: 1) While targets may not be directly observed, planning globally efficient paths based on scene memory without unnecessary detours is important. 2) Adapting to local scene changes that may occur compared to the initial scene memory. 3) Realizing effective planning without human aid, complex rewards, or costly training.

Recent works in robot planning have demonstrated diverse approaches. Works like L3MVN (Yu, Kasaei, and Cao 2023), ESC (Zhou et al. 2023), and VLFM (Yokoyama et al. 2024) employ LFMs to respectively map the scene with language-guided frontier exploration, probabilistic commonsense constraints, and visual-language similarity scoring, bridging semantic understanding with geometric planning. Taking global layouts into consideration, HOVSG (Werby et al. 2024) constructs open-vocabulary 3D scene topological graphs through feature clustering to enable navigation. However, the static topological way-points extracted from occupancy cannot ensure efficient path planning or manage unexpected local scene changes. To solve this dual-challenge, we propose leveraging a hierarchical framework uniquely integrates a lightweight global topological map for coarse route planning with a local imperative approach that dynamically adjusts trajectories.

In this work, we introduce LOG-Nav, an efficient layout-aware object-goal navigation framework designed for complex multi-room indoor environments. LOG-Nav leverages both topological layout relationships and detailed scene representations to achieve efficient navigation in a hierarchical framework. By employing the dual-level information, the planning process first queries the open-world semantics based on user instructions to generate a global navigable topometric path at the layout level. The global waypoints are then sequentially projected into the robot's ego-centric observation space, enabling the robot to plan a dense local waypoint set that accounts for scenario changes. An LLM-powered agent is employed that operates without human interaction, complex reward design, or costly training on specific settings. The method requires only RGB-D observations as input and outputs navigation results as 3D waypoint sets, which are compatible with general platforms. The primary functions are shown in Fig.1. Evaluated on the widely-used MP3D dataset (Ramakrishnan et al. 2021), LOG-Nav demonstrates superior navigation success rates (SR) and efficiency, as measured by success weighted by path length (SPL), compared to exisiting methods. Additionally, we validate our framework through both virtual agent and real-world robots, demonstrating its applicability, efficiency and robustness against unexpected scene changes.

Our contributions are summarized as follows:

- We present LOG-Nav, an efficient layout-aware object-goal navigation approach that leverages hierarchical planning, managed by an LLM-powered agent without further rewards or training.
- We develop a novel hierarchical strategy that integrates global topological layout-aware planning with a local dynamic-aware imperative approach, ensuring efficient and effective navigation both globally and locally in complex indoor environments.
- Complex experiments are conducted on the MP3D dataset demonstrating a 85% SR and 79% SPL in object navigation. Deployments on both virtual agent and real robots are provided to verify the applicability.

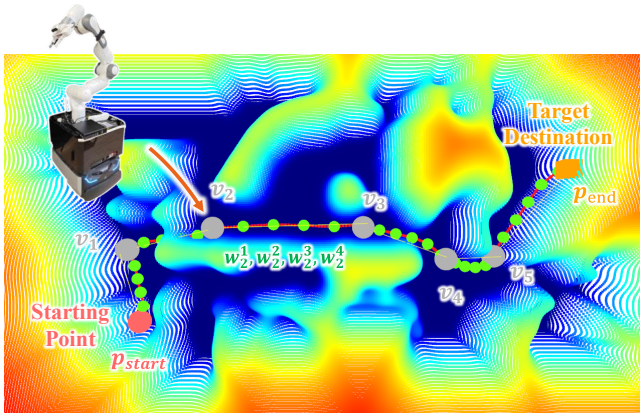

Figure 3: Hierarchical planning example. Global planning is conducted on the topological graph and generates $V = \{v_1, v_2, \ldots, v_n\}$. Local planning, realized in IL approach, generates $W_{i,i+1} = \{w_i^1, w_i^2, \ldots, w_i^m\}$.

Related Works

Process-prompted Navigation

Vision-Language navigation requires agents to follow free-form linguistic instructions to navigate in unseen environments (Anderson et al. 2018b), necessitating an understanding of complex observations and the ability to interpret instructions for effective navigation. Pioneering works were conducted under the simplified assumption that environments could be abstracted as discrete connectivity graphs (Anderson et al. 2018b; Krantz et al. 2020). To bridge the gap for more realistic modeling, researchers have dedicated efforts to address the limitations imposed by task settings (Krantz et al. 2020; Zhang et al. 2024; Krantz et al. 2021; Xu et al. 2023; Anderson et al. 2021). However, the assumption of closed-set scene priors limits the agent to pre-mapped environments. Additionally, VLN evaluation metrics often emphasize strict path fidelity over goal discovery efficiency. On the contrary, we follow a goal-oriented paradigm and propose open-set scene-specific representations built during exploration and object navigation, executed in a human-interaction-free manner with LLM agent.

Goal-oriented Navigation

Goal-oriented navigation involves interpreting multi-modal target descriptions, localizing, and navigating to the specified targets. Notable works (Yang et al. 2018; Du, Yu, and Zheng 2020; Campari et al. 2020; Liang, Chen, and Song 2021; Chaplot et al. 2020) have demonstrated impressive capabilities using large-scale reinforcement learning. However, these methods typically require substantial training data and computational resources and may suffer when generalizing across diverse environments and embodiments (Ramakrishnan et al. 2022). ZSON (Majumdar et al. 2022) trains navigation agents in target environments by mapping object goals to image-goal embeddings, emphasizing the ability to locate and navigate to objects based on inquiries without specific training in that setting. Further advancements have been made (Guan et al. 2024; Sun et al. 2024) to enhance image retrieval and semantic understanding, achieving higher success rates. Additionally, L3MVN (Yu, Kasaei, and Cao 2023) has improved exploration efficiency by constructing environment maps using frontier-based methods and leveraging LLM inference. ESC (Zhou et al. 2023) introduced soft commonsense constraints through probabilistic logic, enabling training-free zero-shot navigation by embedding object-room relationships, while VLFM (Yokoyama et al. 2024) eliminated text dependency by directly grounding visual features with language models through RGB-based similarity scoring, significantly accelerating semantic processing. Nonetheless, the absence of explicit global layout-level guidance limits their ability to ensure efficient path planning. Works like HOVSG (Werby et al. 2024) construct open-vocabulary 3D scene topological graphs through feature clustering to enable navigation. However, the static topological way-points extracted from occupancy cannot ensure efficient path planning or manage unexpected local scene changes.

Overview

We propose a hierarchical robot navigation framework for operation in unexplored indoor environments, guided by natural language instructions T, images I, or 3D locations P. The framework overview is shown in Fig.2. The planning process is conducted on a dual-level map with global topology and local dense representation built with collected RGB-D sequences i, d, where i and d denote RGB and depth observations, respectively. The scene representation consists of (1) an implicit neural function $F : \mathbb{R}^3 \rightarrow \mathbb{R}^n$ maps 3D positions to vision-language embeddings; and (2) a topometric map $G = (V, E)$, where V comprises region vertices $\mathbf{v}_r$ and entrance vertices connecting regions $\mathbf{v}_e$ and edges E connecting these vertices.

During planning, inputs T/I are encoded into embeddings $\mathbf{E}$ using a pre-trained vision-language model. These embeddings are then queried in F to identify the target position p_{end} with the highest matching similarity. A global navigable path is computed between the robot's current pose p_{start} and p_{end} using the topometric map G, generating a sparse sequence of waypoint vertices $V = \{v_1, v_2, \ldots, v_n\}$. These waypoints are progressively refined during navigation in an imperative approach, with $W_{i,i+1} = \{w_i^1, w_i^2, \ldots, w_i^m\}$ denoting the densified local waypoints between v_i and v_{i+1}. Navigation proceeds incrementally: as the robot approaches v_{i+1}, the next segment $W_{i+1,i+2}$ is planned. The framework incorporates an LLM agent designed to integrate essential data streams for task execution.

Method

Hierarchical Scene Representation

Recent works such as HOV-SG (Werby et al. 2024), CLIO (Maggio et al. 2024), and Topo-Field (Hou et al. 2024) have demonstrated progress in representing scenes as topo-maps while retaining detailed content. Although topometric maps offer computational efficiency for downstream planning and navigation tasks, high-fidelity scene

content from neural implicit representations enables semantic cue queries (Hou et al. 2024). Our approach represents 3D scenes combining a high-level topometric map and a detailed neural implicit representation.

Given posed RGB-D sequences $\{i, d\}$, we train a neural implicit function F (Hou et al. 2024) to match 3D points p derived from depth and camera parameters with the encoded vision-language embedding $\mathcal{E}_p$ (Radford et al. 2021; Reimers and Gurevych 2019). A topometric map $G = (V, E)$ is then extracted with vertices V including region vertices $\mathbf{v}_r$ and entrance vertices $\mathbf{v}_e$ and edges E connecting them. The subsequent scene representation is denoted as:

$$F : \mathbb{R}^3 \rightarrow \mathbb{R}^n, \quad (1)$$
$$G = (V, E).$$

Hierarchical Planning

Our planning design strategy encompasses two complementary objectives: 1) The planned path must be globally efficient, avoiding unnecessary deviations and minimizing twists and turns. In indoor scenes, given that the target destination may not be directly observable, robots should avoid entering irrelevant rooms or revisiting explored areas. 2) The robot must possess the capability to detect and adapt to minor changes in the environment, such as the appearance of unexpected obstacles in spaces previously identified as free.

Global Topology Plan To ensure globally efficient path planning for inter-room navigation, our approach utilizes a topological graph to generate a sparse waypoint set that the robot must traverse to reach the destination, typically located at room entrances or connecting doors. The process begins by unifying multi-modal instructions (images, text, or 3D positions) into a destination 3D position. The text/image instruction T/I is first encoded using CLIP (Radford et al. 2021) and Sentence-BERT (Reimers and Gurevych 2019) to generate the target embedding $\{C, S\}_{T/I}$. A point set P^* is then sampled uniformly from the environment, and $F(P^*) = \{(f_v, f_s)\}_{P^*}$, representing the vision and semantic embeddings of the corresponding 3D positions, is calculated. The cosine similarity between $\{C, S\}_{T/I}$ and $F(P^*)$ determines the position $p_{\text{end}} \in P^*$ with the highest similarity to the instruction query (Hou et al. 2024). The topological graph G is subsequently queried using the robot's current position p_{start} and p_{end} to identify the shortest path connecting them. This results in a sequence of vertices $V = \{v_1, v_2, \ldots, v_n\}$, where each v_i represents a structural element, such as doors or corridors, that the robot must navigate through to reach the destination.

Local Ego-centric Plan In the initial global planning phase, a sparse set of waypoints denoted as $V = \{v_1, v_2, \ldots, v_n\}$, is established. These waypoints guide the robot's inter-room navigation, ensuring efficient travel between distinct regions. To achieve intra-room movement and local navigation precision, dense waypoints, $W_{i,i+1} = \{w_i^1, w_i^2, \ldots, w_i^m\}$, are computed between consecutive global waypoints v_i and v_{i+1}. We realize local planning in an Imperative Learning (IL) approach with the target way-point v_{i+1} in the global coordinates and current depth

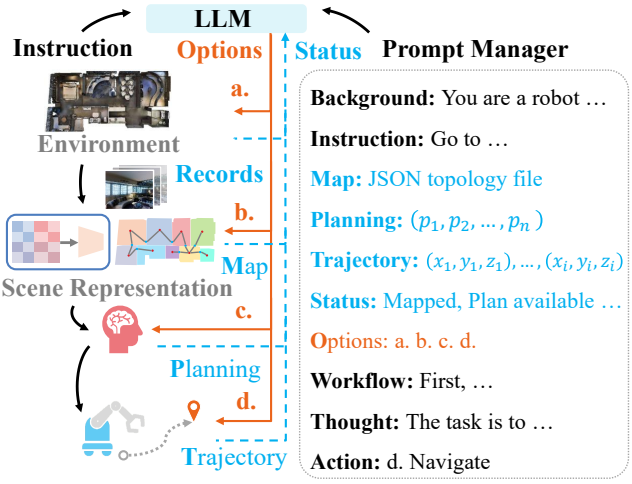

Figure 4: Data flow example of LLM Agent. Instructed by the user input, the agent conducts prompts, applies optional actions, and gets feedback from interaction.

observation d' with pose p'. First, v_{i+1} is projected to the depth frame with the pose p'. The depth input, encoded by a convolutional network (LeCun, Bengio et al. 1995) into a depth embedding $\mathcal{O}_{\lceil'}$, together with the way-point feature $\mathcal{P}'$ of p' is input into a planning network (Yang 2023) that generates a set of local key points $W_{i,i+1}$ from

$$\tau = \mathcal{F}(\mathcal{O}_{d'}, \mathcal{P}'), \quad (2)$$

where $\mathcal{F}$ is the planner and $W_{i,i+1}$ is the discrete point set sampled from the planned trajectory τ. As the robot approaches each subsequent global waypoint v_{i+1}, the next segment of dense waypoints $W_{i+1,i+2}$ is computed and executed incrementally.

The hierarchical pathway from the robot's starting position p_{start} to the destination p_{end} is structured as

$$p_{\text{start}} \rightarrow p_{\text{end}} : V = \{v_1, v_2, \ldots, v_n\}, \quad (3)$$
$$v_i \rightarrow v_{i+1} : W_{i,i+1} = \{w_i^1, w_i^2, \ldots, w_i^m\}.$$

This two-tiered approach ensures efficient navigation at both the global and local levels, allowing the robot to adapt dynamically to changes in the environment while maintaining a clear path. An illustration is shown in Fig.3.

LLM-powered Navigation Agent

Intuitively, a household robot equipped with intelligent algorithms and LLMs-powered knowledge should automatically conduct the necessary process to get into work in a new environment. To address this, an advanced LLM-powered agent is employed to manage the autonomous operation of a household robot in unexplored environments which oversees both exploration and exploitation phases. The agent's design is guided by two overarching principles: 1) It should integrate user task descriptions, representations of the scene, and the robot's current state to make informed decisions; 2) It employs a comprehensive set of strategies, including mapping, path planning, and locomotion to effectively accomplish tasks while maintaining the ability to self-update. At

Methods	Object Nav.		Instance Nav. (text-goal)		Instance Nav. (image-goal)	
	SR	SPL	SR	SPL	SR	SPL
CoW (2022) (Gadre et al. 2022)	0.061	0.039	0.018	0.011	-	-
ZSON (2022) (Majumdar et al. 2022)	0.255	0.126	0.106	0.049	0.146	0.073
L3MVN (2023) (Yu, Kasaei, and Cao 2023)	0.352	0.165	-	-	-	-
ESC (2023) (Zhou et al. 2023)	0.392	0.223	0.065	0.037	-	-
VLFM (2024) (Yokoyama et al. 2024)	0.364	0.175	-	-	-	-
PixNav (2024) (Cai et al. 2024)	0.379	0.205	-	-	-	-
HOV-SG (2024) (Werby et al. 2024)	0.404	0.236	-	-	-	-
PSL (2024) (Sun et al. 2024)	0.424	0.192	0.165	0.075	0.230	0.114
LOG-Nav(Ours)	**0.856**	**0.797**	**0.786**	**0.701**	**0.673**	**0.582**

Table 1: Quantitative comparison of object navigation and instance navigation tasks on MP3D dataset. For object navigation, text instruction is used as input. For instance navigation, text-prompted and image-prompted instructions are separately validated. The SR and SPL are employed as metrics.

Targets	Obstacles	Local-Plan Retries	Global-Plan Retries	Path Length (m)	Time Cost (s)	**SR**
chair (text)	-	0/10	0/10	16	91	10/10
chair (image)	-	2/10	0/10	16	93	7/10
sink (text)	1	2/10	0/10	20	94	9/10
sink (image)	1	3/10	1/10	20	97	6/10
sofa (text)	2	3/10	2/10	15	82	6/10
sofa (image)	3	6/10	3/10	15	80	3/10

Table 2: Quantitative object navigation evaluation in real-world scene. Each navigation task is evaluated 10 times from a similar starting position to the same target. *Obstacles* means the number of obstacles we set after robot exploration where robots have to bypass to approach targets. *Local/Global-Plan Retries* means the number of replanning called by the agent during the 10 times of navigation. *Path Length* is the average path length robot travels in 10 times navigation. *Time cost* is the average cost from starting points to destinations, where failure cases are not counted. *SR* is the navigation success rate.

each decision-making step, the agent processes a sequence of contextual tokens, denoted as:

$$\{\mathbf{B}, \mathbf{I}, \mathbf{M}, \mathbf{P}, \mathbf{T}, \mathbf{S}, \mathbf{O}, \mathbf{A}\} \tag{4}$$

where $\mathbf{B}$ represents the task background information, structured as "You are an embodied robot ...". $\mathbf{I}$ denotes the user's instruction, which can be text, an image, or a 3D position. $\mathbf{M}$ is the hierarchical map of the environment, comprising a learned implicit function F and a topological graph G in JSON format. $\mathbf{P}$ includes the robot's historical planning data. $\mathbf{T}$ refers to the robot travel trajectory. $\mathbf{S}$ indicates the robot's current status, such as its location and whether the planned path remains viable. $\mathbf{O}$ represents any optional skills the robot may use. $\mathbf{A}$ is the action decision.

The primary functions of the agent are as follows:

- Mapping. The robot captures RGB-D images to construct a comprehensive environmental representation using neural implicit functions and topometric maps.
- Planning. Based on the scene representation (F, G) and user instructions, the agent conducts planning described in Section to fulfill the task.
- Navigation. The robot follows the planned waypoints to execute the designated actions and recognize obstacles or errors to conduct iterative attempts.

The LLM directs the operational workflow by selecting the most suitable actions, primarily informed by above mentioned information. Upon entering a new environment, the robot initiates a mapping process, exploring the space frontier-based to gather RGB-D observations and build its scene representation. Once the scene is understood, the agent proceeds task planning. Continuous monitoring trajectory and status ensures that navigation proceeds as intended. If errors arise, such as repeated visits, excessive time expenditure without reaching the next waypoint, or exceeding replanning attempts, the LLM either replans or issues an error report. Fig.4 shows entire process and primary data flow.

Experimental Results

Simulation Experiments

Dataset. MP3D (Ramakrishnan et al. 2021) is a high-res photorealistic 3D reconstruction dataset of real-world scenes. We validate our approach on over 20 multi-room indoor scenarios. The dataset is formatted in a Habitat manner and defines six object categories: chairs, couches, potted plants, beds, toilets, and TVs. Notably, our approach considers layout-level information, enabling refined instance navigation experiments by differentiating objects in separate rooms, such as a chair in the living room versus one in the dining room, as distinct instances.

Setups. We employ the Habitat platform as our 3D indoor simulator. Observations resolution from the RGB-D camera is set to 480×640, as used by (Yu, Kasaei, and

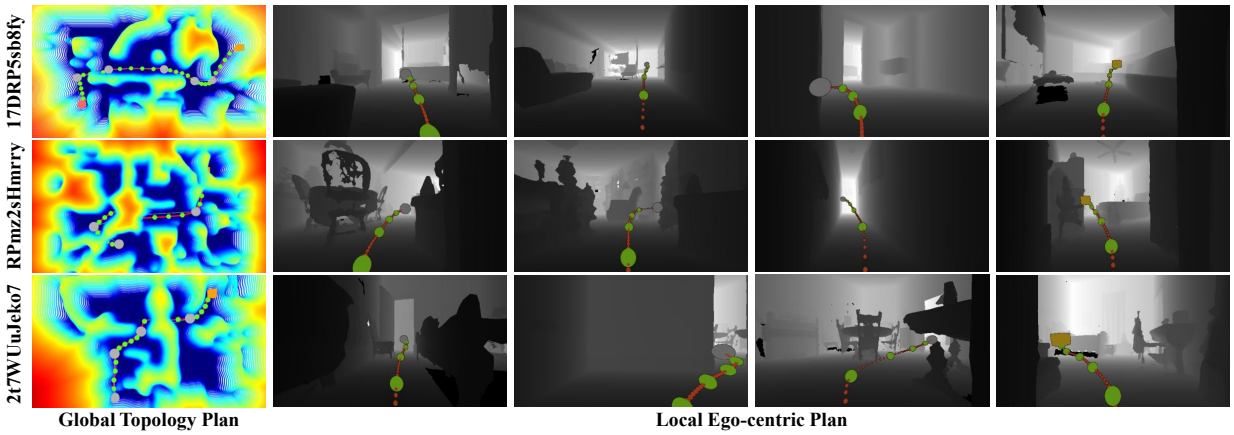

Figure 5: Navigation results on MP3D dataset. The left column shows global planning results, where gray points are entrance vertices in the topological map, red points show the starting positions and yellow points show the destinations. The right columns show local planning results on the ego-centric views.

Methods	With Map	1-time Run	10 Runs	20 Runs	30 Runs	60 Runs
ZSON(Majumdar et al. 2022)	X	0.102	0.116	0.125	0.119	0.121
PSL(Sun et al. 2024)	X	0.224	0.215	0.197	0.202	0.199
L3MVN(Yu, Kasaei, and Cao 2023)	√	0.186	0.217	0.252	0.281	0.358
ESC(Zhou et al. 2023)	√	0.249	0.286	0.337	0.351	0.389
LOG-Nav	√	0.083	0.417	0.533	0.592	0.656

Table 3: SPL comparison with ZSON approaches across different numbers of runs.

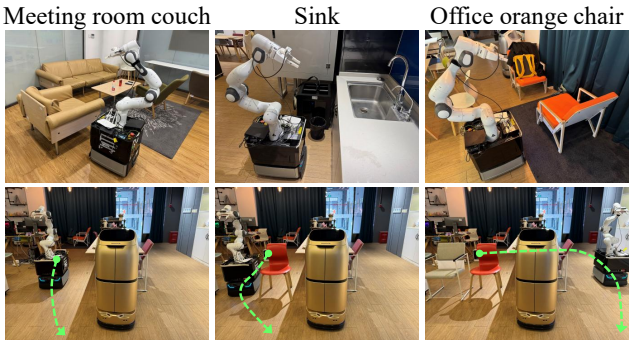

Figure 6: Real-world experiment results with mobile robots. The above row shows the navigation success examples. The bottom row shows the unexpected obstacle avoidance examples compared to the built scene memory.

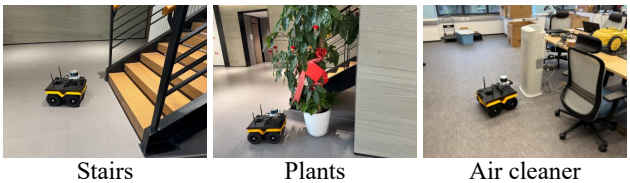

Figure 7: Real-world experiment results with Clearpath Jackal. We conducted navigation deployments on different embodied platforms to show its applicability and robustness.

Cao 2023). Pose data is provided by an odometry sensor. The automatic exploration process is implemented using a simple policy approach, where the robot drives continuously along the left frontier while keeping the camera oriented to the right. Frames frequency is set to $15Hz$, following (Yang 2023). After traversing, the camera is reset to face forward. Our implementation is based on publicly available code examples from the project repository (Puig et al. 2023).

Metrics. We evaluate our approach using Success Rate (SR) and SR weighted by inverse Path Length (SPL), following established metrics in target navigation research (Anderson et al. 2018a). Additionally, we provide instance-level navigation experiments, leveraging our approach's ability to recognize objects in different rooms as separate instances, capitalizing on the considered layout information.

Comparison Results and Discussions. As shown in Tab.1 and Fig.5, our Log-Nav outperforms existing methods in both SR and SPL. The compared methods include L3MVN (Yu, Kasaei, and Cao 2023), ESC (Zhou et al. 2023), and VLFM (Yokoyama et al. 2024), which employ LFM-powered mapping of separately frontier, probabilistic commonsense constraints, and visual-language similarity scoring. HOV-SG (Werby et al. 2024) employs global topological map, however, the static topological way-points extracted from occupancy cannot ensure efficient path planning or manage unexpected local scene changes which

cannot ensure efficiency and robustness. This underscores the advantages of our hierarchical planning strategy, ensuring both effectiveness and efficiency by incorporating refined local waypoints in a cognitive-like global layout context (Zeng, Si, and Feng 2022). Moreover, when comparing performances across object and instance navigation tasks, our method excels by setting the final target destination to the appropriate region based on related topo-vertices, overcoming limitations of previous methods that ignore layout.

Further, we compare with zero-shot-object-navigation (ZSON) methods repeatedly within the same workspace to complete object navigation tasks, shown in Tab. 3. All methods shared identical object goal sequences, and we compared the average SPL across different numbers of runs. Our 'explore-then-plan' approach constructs a complete map only once before performing planning cycles, whereas other 'mapping-during-navigation' ZSON methods accumulate mapping efforts over repeated executions. Results demonstrate that our mapping cost becomes amortized across planning iterations, particularly beneficial in scenarios with frequent deployments and extensive environments. By leveraging hierarchical environmental priors, our method outperforms incremental mapping approaches.

Real-world Deployment

Our real-world robot platforms include two types: one comprises a SLAMTEC mobile base, a Franka Panda arm, and a RealSense D435 camera mounted on the end. The camera is calibrated to the arm's base coordinates using the easy-hand-eye package. RGB-D images are recorded at $15Hz$, matching the resolution of 480×640 as in (Yu, Kasaei, and Cao 2023; Yang 2023). The other one is a Clearpath robot with the same camera setting. The environment consists of a large-scale, multi-room indoor scene covering approximately $225m^2$. It includes a small kitchen tabletop, office area, meeting room, and a hall. As demonstrated in Fig.6, the robot is tasked with navigating to different object instances through long-term paths.

To validate the robustness, after the robot explores and constructs the scene representation, we introduce obstacles and make minor adjustments to the scenario. This setup allows us to assess the navigation success rate under dynamic conditions. As shown in Fig.6, the robot successfully bypasses the obstacles or plans new paths. For further evaluation shown in Tab.2, we evaluate the path length, time cost to the destination, success rate, and number of global/local re-planning with each experiment setup 10 times. Additional results and task videos are shown in the website.

Ablations

Global Planning. The global planning process incorporates topological information, from the local IL planning. To assess the impact of global planning, we disable it and directly project the long-term final destination's 3D position into the current observation frame, relying solely on the local IL planning. As illustrated in Fig.8, the SR significantly decreases as the distance to targets increases, indicating local IL planning alone is insufficient for long-term planning.

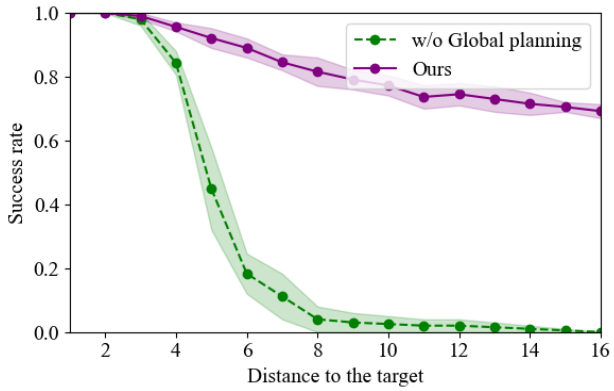

Figure 8: Ablations on Global Planning strategy. Results show how SPL decreases with distance at different settings. "w/o" means "without".

Obstacles	w/o Local Planning		LOG-Nav (Ours)	
	Cost (s)	SR	Cost (s)	SR
0	102	10/10	96	10/10
1	111	7/10	103	10/10
2	122	3/10	112	8/10
3	134	1/10	116	6/10

Table 4: Ablations on Local Planning strategy. We evaluate the navigation time cost (only counting the successful navigation) and SR under different numbers of obstacles set in the way after exploration. Each navigation task is evaluated 10 times with similar settings except for the obstacles.

Local Planning. This ablation is conducted in real-world setting. Without local planning, we directly call the point navigation API of the mobile robot base to reach the globally planned way-point. The API allows robots to navigate to the target or attempt to circumvent obstacles as they arise. The time taken by the robot to reach the destination is recorded and analyzed. As shown in Tab.4, the SR and efficiency drop rapidly when obstacle numbers grow without local planning.

Limitations

This paper introduces LOG-Nav, an efficient layout-aware object-goal navigation approach for multi-room indoor environments, hierarchically planning on both global and local levels. Our approach makes progress in addressing globally and locally efficient long-term path planning in complex indoor scenes and ease of deployment without human aid, complex rewards, or costly training. Despite these advancements, our approach currently has two limitations: 1) While local mapping is implemented through imperative learning, it does not fully leverage the potential of the local scene memory for refined obstacle avoidance and improved efficiency. 2) The detected scene changes are not dynamically integrated into the scene representation and topological map. Addressing these limitations by implementing real-time scene updates remains a priority for future work.

References

Anderson, P.; Chang, A.; Chaplot, D. S.; Dosovitskiy, A.; Gupta, S.; Koltun, V.; Kosecka, J.; Malik, J.; Mottaghi, R.; Savva, M.; et al. 2018a. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*.

Anderson, P.; Shrivastava, A.; Truong, J.; Majumdar, A.; Parikh, D.; Batra, D.; and Lee, S. 2021. Sim-to-real transfer for vision-and-language navigation. In *Conference on Robot Learning*, 671–681. PMLR.

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and Van Den Hengel, A. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3674–3683.

Cai, W.; Huang, S.; Cheng, G.; Long, Y.; Gao, P.; Sun, C.; and Dong, H. 2024. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 5228–5234. IEEE.

Campari, T.; Eccher, P.; Serafini, L.; and Ballan, L. 2020. Exploiting scene-specific features for object goal navigation. In *European Conference on Computer Vision*, 406–421. Springer.

Chaplot, D. S.; Gandhi, D. P.; Gupta, A.; and Salakhutdinov, R. R. 2020. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33: 4247–4258.

Chen, H.; Suhr, A.; Misra, D.; Snavely, N.; and Artzi, Y. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12538–12547.

Dorbala, V. S.; Sigurdsson, G.; Piramuthu, R.; Thomason, J.; and Sukhatme, G. S. 2022. Clip-nav: Using clip for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2211.16649*.

Du, H.; Yu, X.; and Zheng, L. 2020. Learning object relation graph and tentative policy for visual navigation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, 19–34. Springer.

Gadre, S. Y.; Wortsman, M.; Ilharco, G.; Schmidt, L.; and Song, S. 2022. Clip on wheels: Zero-shot object navigation as object localization and exploration. *arXiv preprint arXiv:2203.10421*, 3(4): 7.

Guan, T.; Yang, Y.; Cheng, H.; Lin, M.; Kim, R.; Madhivanan, R.; Sen, A.; and Manocha, D. 2024. LOC-ZSON: Language-driven Object-Centric Zero-Shot Object Retrieval and Navigation. *arXiv preprint arXiv:2405.05363*.

Hou, J.; Guan, W.; Liang, L.; Feng, J.; Xue, X.; and Zeng, T. 2024. Topo-Field: Topometric mapping with Brain-inspired Hierarchical Layout-Object-Position Fields. arXiv:2406.05985.

Krantz, J.; Banerjee, S.; Zhu, W.; Corso, J.; Anderson, P.; Lee, S.; and Thomason, J. 2023. Iterative vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14921–14930.

Krantz, J.; Gokaslan, A.; Batra, D.; Lee, S.; and Maksymets, O. 2021. Waypoint models for instruction-guided navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15162–15171.

Krantz, J.; Wijmans, E.; Majumdar, A.; Batra, D.; and Lee, S. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 104–120. Springer.

LeCun, Y.; Bengio, Y.; et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10): 1995.

Liang, Y.; Chen, B.; and Song, S. 2021. Sscnav: Confidence-aware semantic scene completion for visual semantic navigation. In *2021 IEEE international conference on robotics and automation (ICRA)*, 13194–13200. IEEE.

Maggio, D.; Chang, Y.; Hughes, N.; Trang, M.; Griffith, D.; Dougherty, C.; Cristofalo, E.; Schmid, L.; and Carlone, L. 2024. Clio: Real-time Task-Driven Open-Set 3D Scene Graphs. *IEEE Robotics and Automation Letters*, 9(10): 8921–8928.

Majumdar, A.; Aggarwal, G.; Devnani, B.; Hoffman, J.; and Batra, D. 2022. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *Advances in Neural Information Processing Systems*, 35: 32340–32352.

Puig, X.; Undersander, E.; Szot, A.; Cote, M. D.; Partsey, R.; Yang, J.; Desai, R.; Clegg, A. W.; Hlavac, M.; Min, T.; Gervet, T.; Vondrus, V.; Berges, V.-P.; Turner, J.; Maksymets, O.; Kira, Z.; Kalakrishnan, M.; Malik, J.; Chaplot, D. S.; Jain, U.; Batra, D.; Rai, A.; and Mottaghi, R. 2023. Habitat 3.0: A Co-Habitat for Humans, Avatars and Robots.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramakrishnan, S. K.; Chaplot, D. S.; Al-Halah, Z.; Malik, J.; and Grauman, K. 2022. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18890–18900.

Ramakrishnan, S. K.; Gokaslan, A.; Wijmans, E.; Maksymets, O.; Clegg, A.; Turner, J.; Undersander, E.; Galuba, W.; Westbury, A.; Chang, A. X.; et al. 2021. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint*

Conference on Natural Language Processing (EMNLP-IJCNLP), 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Shah, D.; Osiński, B.; Levine, S.; et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, 492–504. PMLR.

Sun, X.; Liu, L.; Zhi, H.; Qiu, R.; and Liang, J. 2024. Prioritized semantic learning for zero-shot instance navigation. In *European Conference on Computer Vision*, 161–178. Springer.

Vasudevan, A. B.; Dai, D.; and Van Gool, L. 2021. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision*, 129: 246–266.

Werby, A.; Huang, C.; Büchner, M.; Valada, A.; and Burgard, W. 2024. Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation. *Robotics: Science and Systems*.

Xu, C.; Nguyen, H. T.; Amato, C.; and Wong, L. L. 2023. Vision and Language Navigation in the Real World via Online Visual Language Mapping. *arXiv preprint arXiv:2310.10822*.

Yang, F. 2023. iPlanner: Imperative Path Planning. In *Robotics: Science and Systems (RSS 2023)*.

Yang, W.; Wang, X.; Farhadi, A.; Gupta, A.; and Mottaghi, R. 2018. Visual semantic navigation using scene priors. *arXiv preprint arXiv:1810.06543*.

Yokoyama, N.; Ha, S.; Batra, D.; Wang, J.; and Bucher, B. 2024. VLFM: Vision-Language Frontier Maps for Zero-Shot Semantic Navigation. In *International Conference on Robotics and Automation (ICRA)*.

Yu, B.; Kasaei, H.; and Cao, M. 2023. L3mvn: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3554–3560. IEEE.

Zeng, T.; Si, B.; and Feng, J. 2022. A theory of geometry representations for spatial navigation. *Progress in Neurobiology*, 211: 102228.

Zhang, J.; Wang, K.; Xu, R.; Zhou, G.; Hong, Y.; Fang, X.; Wu, Q.; Zhang, Z.; and Wang, H. 2024. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*.

Zhou, K.; Zheng, K.; Pryor, C.; Shen, Y.; Jin, H.; Getoor, L.; and Wang, X. E. 2023. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In *International Conference on Machine Learning*, 42829–42842. PMLR.

Zhu, Y.; Mottaghi, R.; Kolve, E.; Lim, J. J.; Gupta, A.; Fei-Fei, L.; and Farhadi, A. 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, 3357–3364. IEEE.