2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

DIPP: A Diffusion-based Potential Planner for Synergistic Navigation and Mapping

Yiqing Zhang[1], Tao Wang[1], Miaoxin Pan[1], Yi Yang[1,*], Mengyin Fu[1,2]

***Abstract*— Object-Goal Navigation (ObjectNav) requires an embodied agent to search for and reach a target object category in previously unseen environments using only onboard egocentric observations, which is a fundamental capability for long-horizon autonomous robots. Current Object-Goal Navigation methods typically discard environmental knowledge after each episode, limiting their ability to operate autonomously over long horizons. To overcome this limitation, we introduce DIPP, a diffusion-based potential planner that unifies navigation and mapping. DIPP generates two complementary potential fields: a navigation potential that directs the agent toward the target and a topological potential that captures the environment's structural skeleton. The topological potential serves a dual purpose: it acts as an implicit structural prior for waypoint selection when fused directly with the navigation potential and, more importantly, enables the incremental construction of a persistent, explicit topological graph. This graph enables a hierarchical policy to select strategic, long-horizon waypoints, elevating planning from a tactical search to a strategic decision. We evaluate DIPP in the Habitat simulator on the Gibson dataset. Results show that DIPP achieves strong performance on standard ObjectNav metrics (SR, SPL) while constructing structurally accurate maps, evidenced by a high Node Recall score. Furthermore, leveraging the explicit persistent graph for hierarchical planning significantly boosts navigation performance. These findings demonstrate the effectiveness of DIPP in enabling embodied agents to build and exploit persistent spatial knowledge for long-term operation in unseen environments.**

I. INTRODUCTION

Embodied agents for Object-Goal Navigation (ObjectNav) have achieved remarkable success in navigating complex, unseen indoor environments to find specific objects [1], [2]. However, the vast majority of current systems are designed for single-episode execution, operating without persistent memory. They learn to solve a given navigation problem but discard the valuable environmental understanding acquired upon completion. This is fundamentally inefficient and limits their potential for long-term autonomy, as they must re-learn an environment's layout from scratch for every new task. This core limitation motivates a departure from purely reactive navigation policies towards agents that can build and leverage a persistent, abstract understanding of their surroundings. We argue that for true long-term autonomy, an agent must not only find a path to a goal but also concurrently construct a reusable map of the environment's underlying structure, a task that current ObjectNav formulations do not explicitly require.

To address this gap, we propose a paradigm shift towards simultaneous navigation and topological mapping. Our solution **DIPP**, a **DI**ffusion-based **P**otential **P**lanner, actualizes this by learning not only *where* to go, but also *how* an environment is structured, thereby building a valuable spatial prior during its task. At the core of DIPP is a dual-channel conditional diffusion model, conditioned on an aggregated semantic map, that jointly generates two synergistic potential fields: a navigation potential (U_{nav}) for goal-directed movement and a topological potential (U_{topo}) identifying the environment's structural skeleton. This topological potential is pivotal: while it can serve as an *implicit* exploration prior, its primary function is to enable the incremental construction of a persistent, *explicit* topological graph. This graph elevates planning from a tactical, pixel-level search to a strategic, node-level decision process, creating a closed loop that intelligently balances goal-seeking with strategic map-building.

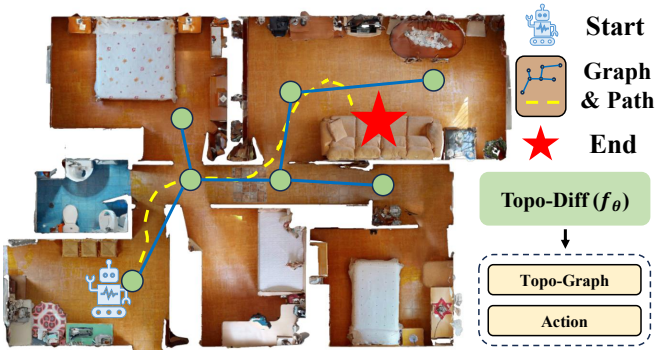

Fig. 1: An example of DIPP, our framework for ObjectNav framed as a dual objective. Our core diffusion model, Topo-Diff, synergizes navigation and mapping by simultaneously generating a structural topological graph (blue) and a goal-oriented action plan (realized as the yellow path). This synergy enables the agent to build an abstract map while efficiently navigating from the start to the goal (red star).

Our main contributions are as follows: (1) We propose DIPP, a novel framework that reframes ObjectNav as a dual objective of simultaneous navigation and topological mapping, enabling an agent to build a persistent and abstract map of its environment, as illustrated in Fig. 1; (2) We introduce a dual-channel conditional diffusion model as the core of our high-level policy. This is the first work to synergize the generation of a goal-directed navigation

This work was supported by National Natural Science Foundation of China (Grant No. NSFC 62233002, 92370203) and National Key R&D Program of China (2022YFC2603600).

[1]The authors are with the School of Automation, Beijing Institute of Technology, Beijing 100081, P.R.China.
[2]The author is also with Nanjing University of Science and Technology, Nanjing 210014, P.R.China.
*Corresponding author: Yi Yang (email: yang_yi@bit.edu.cn).

potential and a structural topological potential within a single generative model, allowing the two tasks to mutually enhance one another; (3) We conduct extensive experiments in the Habitat simulator on the Gibson dataset, demonstrating that DIPP not only achieves top-tier performance on the standard ObjectNav task but also constructs high-quality, structurally-accurate topological maps, validated by both standard and newly proposed metrics.

II. RELATED WORKS

Our research intersects object-goal navigation, topological representations, and generative diffusion models. We review the most relevant literature in these domains.

A. Object-Goal Navigation

Object-Goal Navigation (ObjectNav) is a standard embodied navigation task requiring an agent to locate an object in an unseen environment [3], [4]. Early solutions combined classical components like SLAM-based mapping [5] with exploration strategies [6]. Modern learning-based methods are typically divided into end-to-end and modular systems.

Seminal end-to-end RL methods like DD-PPO [7] established strong baselines, later improved by techniques such as sophisticated memory architectures [8] and auxiliary objectives [9] to enhance long-horizon reasoning. However, these methods can suffer from sample inefficiency and produce policies that are difficult to interpret.

Modular approaches, in contrast, decouple perception, mapping, and planning [10]. A prominent direction uses an explicit top-down map for planning, from frontier-based planners on semantic maps [6] to learning potential fields that guide exploration [1]. Our work follows this modular, potential-field-based paradigm but makes a crucial departure: instead of treating each navigation episode as a self-contained task, we introduce a concurrent objective of building a persistent, abstract map of the environment, a task more aligned with long-term autonomy.

B. Topological Representations for Robot Navigation

While precise, metric maps can be computationally expensive. Topological maps offer a more compact, abstract representation of an environment's structure as a graph of nodes (key places) and edges (connectivity) [11], supporting efficient long-horizon planning.

More relevant to our work are methods that construct the topological graph online. For example, OVG-Nav [12] constructs a graph of candidate locations and predicts a goal-conditioned value for each node to guide high-level planning. Other approaches learn to build and navigate topological abstractions from visual inputs [13], [14]. However, these methods often treat map generation and navigation as distinct modules. A key distinction of our work is the tight integration of topological node identification and goal-directed navigation within a single, unified generative model. Our DIPP model synergistically co-generates both potentials from a shared latent space, allowing the prediction of topological structure to directly inform the navigation policy in real-time.

C. Generative Diffusion Models in Robotics

Denoising Diffusion Probabilistic Models (DDPMs) [15] are a powerful class of generative models that have achieved state-of-the-art results in domains like image synthesis [16]. In robotics, diffusion models have been applied to imitation learning for manipulation [17] and policy learning, where they generate action sequences or future waypoints, as seen in Diffusion Policy [18] and T-Diff [2].

While these works generate low-level control sequences, our approach leverages diffusion for a more abstract task: generating high-level potential fields. By training a model to produce these 2D utility maps, we decouple high-level spatial reasoning from low-level control, allowing our model to focus on the semantic and structural aspects of the problem. To our knowledge, this is the first work to propose a dual-channel diffusion model for jointly generating synergistic topological and navigation potentials for embodied AI tasks.

III. PROBLEM FORMULATION

We address the task of Object-Goal Navigation (ObjectNav) [4], a widely used embodied AI benchmark where an agent is initialized at a random starting location within a previously unseen environment and must navigate to an instance of a specified object category (e.g., 'chair'). At each time-step t, the agent receives an egocentric RGB-D observation o_t and its 3-DoF pose $p_t = (x, y, \theta)$ relative to its starting position. The discrete action space A consists of *MOVE_FORWARD*, *TURN_LEFT* $(30°)$, *TURN_RIGHT* $(30°)$, and *STOP*. An episode is considered successful if the agent invokes the stop action within a threshold distance $(1.0m)$ of any instance of the target object before an episode step limit of $T = 500$ is exceeded.

In standard ObjectNav, the sole objective is efficient goal-reaching. Distinct from this formulation, we augment the task with a concurrent objective: the incremental construction of a persistent topological graph, $G_t = (V_t, E_t)$, where V_t is the set of topological nodes and E_t is the set of edges representing connectivity between nodes. To enable this dual objective, the agent operates on an aggregated top-down semantic map m_t constructed from the observation sequence, from which both goal-directed navigation and graph construction must be simultaneously addressed.

IV. METHODOLOGY

We propose DIPP, a modular framework designed for Simultaneous Navigation and Topological Mapping (Fig. 2). DIPP is composed of three principal components: a Semantic Mapper, a Dual-Channel Diffusion Model (the core of the potential planner), and a Motion Planner, which are framed with a blue dotted line in Figure. At each high-level decision step, the Semantic Mapper fuses incoming sensory data and odometry into a unified, top-down semantic map m_t. Conditioned on the target object category g, this map is processed by our diffusion model to generate two complementary potential fields: a navigation potential U_{nav} and a topological potential U_{topo}. The topological potential is incrementally refined into a persistent graph G_t that

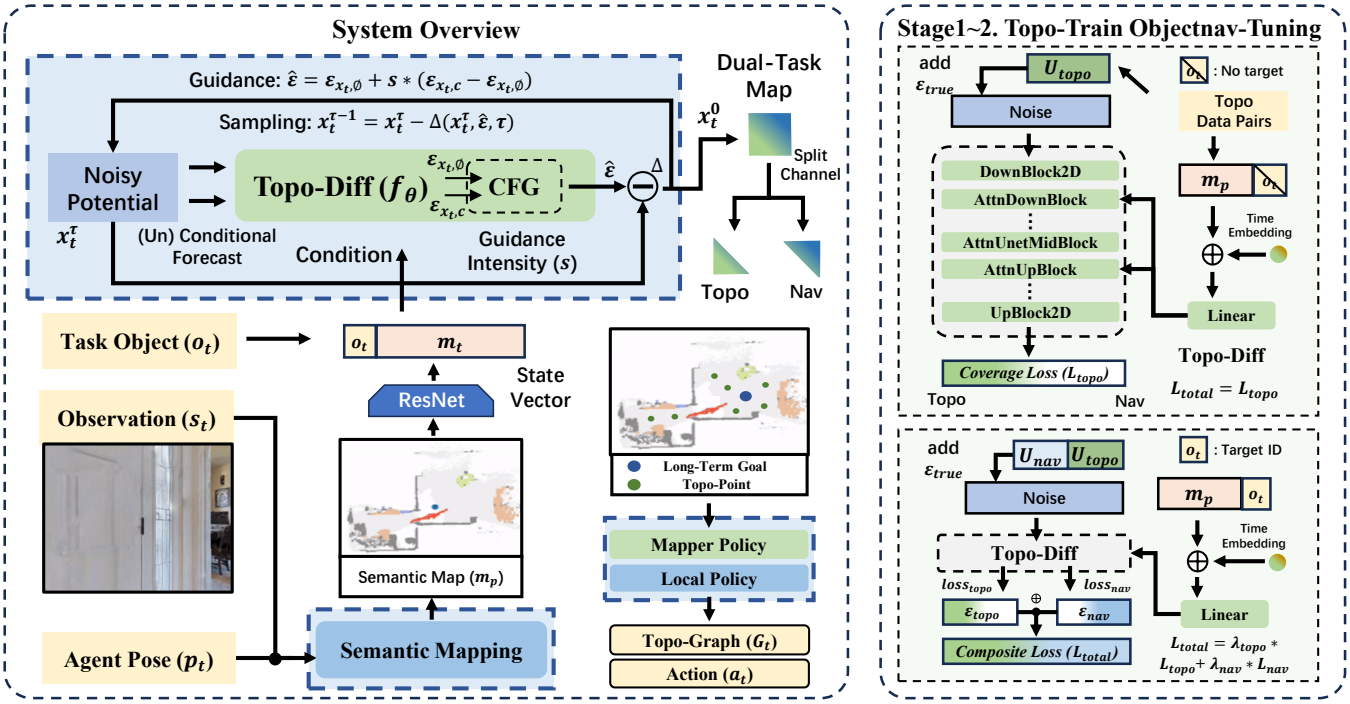

Fig. 2: Overview of DIPP. **Left (online inference; repeated each replanning step):** mapping updates an allocentric semantic map from observation and pose, then Topo-Diff generates a *navigation potential* and a *topological potential* conditioned on the map and target category (with classifier-free guidance using a null-goal branch). The topological potential updates the persistent graph, and the planner selects a long-horizon waypoint that the local motion planner executes into action. **Right (offline supervised training):** a two-stage curriculum pre-trains on topology-only targets to learn a structural prior, then fine-tunes jointly on navigation and topology targets.

captures the abstract structure of the environment. Both fields are then combined to select a long-term waypoint, balancing goal-directed navigation with map construction. Finally, the Motion Planner executes low-level actions to reach the selected waypoint. The key novelty of DIPP lies in its unified generative model, which jointly reasons about navigation and environmental structure, enabling these dual tasks to reinforce each other in a cohesive manner.

A. Preliminaries: Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) [15] are a class of generative models that learn to approximate a data distribution $q(x)$ through a gradual denoising process. The forward process progressively corrupts a clean sample $x_0 \sim q(x)$ by adding Gaussian noise across T steps, yielding noisy samples $x_1, \ldots, x_T$. This Markov chain can be expressed as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}), \tag{1}$$

where $\beta_{t_{t=1}}^{T}$ is a variance schedule controlling the noise magnitude.

The reverse process aims to invert this corruption by learning a parameterized denoising distribution $p_\theta(x_{t-1} \mid x_t)$. Rather than modeling it directly, DDPMs train a neural network ϵ_θ to predict the injected noise, using the objective:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]. \tag{2}$$

At inference, a sample is generated by iteratively denoising from Gaussian noise $x_T \sim \mathcal{N}(0, I)$ back to x_0.

To enable task-specific control, conditional diffusion models incorporate side information c (e.g., semantic maps or goals) into $\epsilon_\theta(x_t, t, c)$. Classifier-Free Guidance (CFG) [19] further enhances controllability by interpolating between conditional and unconditional predictions, yielding more goal-aligned generations.

B. Ground Truth Potential Field Generation

Our framework, DIPP, leverages a conditional diffusion model to generate dual potential fields that provide high-level guidance for simultaneous navigation and mapping. The methodology encompasses three key stages: generation of ground truth potential fields for training, the architecture and training of our potential diffusion model, and the inference procedure for online planning. This section details the first stage.

To construct a supervised dataset for training our model, we generate ground truth potential fields from complete, ground-truth environment maps. For each environment, we define two distinct fields, the topological potential U_{topo} and the navigation potential U_{nav}. These fields jointly constitute the target data $x_0 = [U_{\text{topo}}, U_{\text{nav}}]$ for the diffusion process.

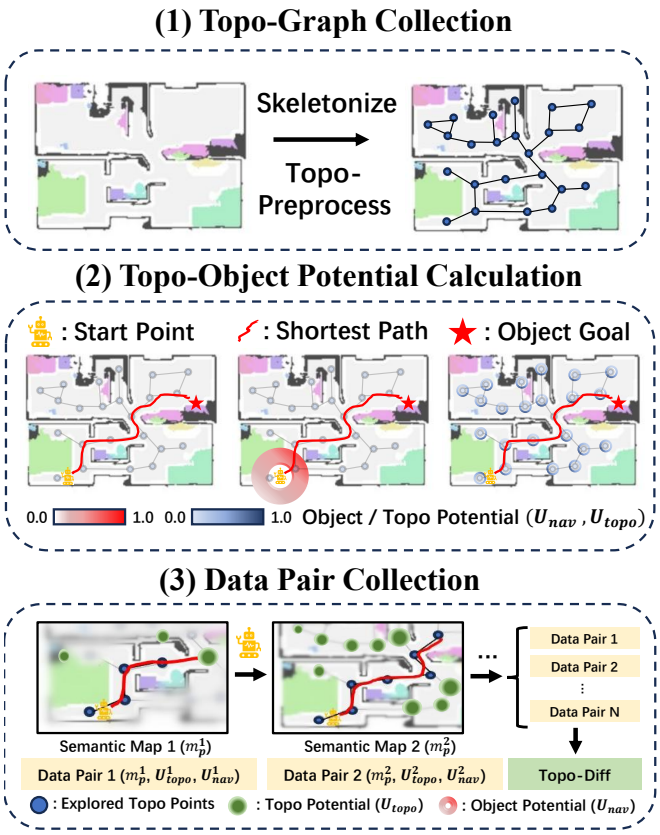

Fig. 3: Overview of the ground truth data generation pipeline. (1) A topological graph is extracted from the complete, ground-truth semantic map. (2) Object and topological potential fields are calculated based on geodesic distances to goals and topological nodes, respectively. (3) These components are used to formulate training data pairs for the diffusion model.

Our formulation is inspired by heuristic functions from classical planning and recent learning-based methods, but is specifically adapted for our dual objective of simultaneous navigation and mapping.

*1) **Topological Potential**(U_{topo}):* The primary purpose of this field is to train the model to identify locations of high structural importance within an environment. To generate this ground truth, we extract and smooth the traversable space from a complete semantic map [6]. We then generate a refined topological graph by skeletonizing this area and subsequently pruning and merging nodes via clustering—a process inspired by Voronoi-based methods [20]. This yields the final set of ground-truth topological nodes, V_{topo}, representing structurally significant locations such as room centroids and hallway intersections. The potential at any navigable location x is then defined as:

$$U_{\text{topo}}(x) = \max\left(1 - \frac{d(x, V_{\text{topo}})}{d_{\text{cutoff}}^{\text{topo}}}, 0\right) \tag{3}$$

where $d_{\text{cutoff}}^{\text{topo}}$ is a predefined hyperparameter. This formulation encourages the model to learn the underlying structural features that define topological nodes, creating a potential field with high values centered on these nodes and thereby guiding the model to perceive the environment's abstract structure.

*2) **Navigation Potential**(U_{nav}):* The navigation potential field provides a direct, goal-oriented signal for the navigation task. Inspired by the effective formulation in PONI [1], we define this potential based on the geodesic distance to the target's success zone rather than its exact location.

Let S_g denote the union of all 1.0m radius success zones surrounding the instances of a target object category g. The navigation potential at any location x is computed based on its geodesic distance, $d_g(x, S_g)$, to the closest point within this combined success region. The potential is formally defined as:

$$U_{\text{nav}}(x, g) = \max\left(1 - \frac{d_g(x, S_g)}{d_{\text{cutoff}}^{\text{nav}}}, 0\right) \tag{4}$$

where $d_{\text{cutoff}}^{\text{nav}}$ is a hyperparameter determined through validation experiments. This approach creates a smooth gradient towards the target's vicinity, rather than to a single point, providing a more robust signal for navigation.

Finally, these two fields are stacked along the channel dimension to form the 2-channel ground truth target, $x_0 = [U_{\text{topo}}, U_{\text{nav}}]$, for our potential diffusion model.

C. Potential Diffusion Model

We model the joint distribution of the potential fields, $p(U_{\text{topo}}, U_{\text{nav}} | m_t, g)$, with a conditional Denoising Diffusion Probabilistic Model (DDPM). Our model, DIPP, is trained to reverse the diffusion process, as detailed in the Preliminaries section. An overview of the architecture and training procedure is provided in Figure 2.

1) Architecture and Conditioning: The core of DIPP is a U-Net-based architecture [21], which accepts the noisy potential fields as input and is trained to predict the added noise. A key aspect of our design is the explicit separation of the data undergoing diffusion from the conditioning information.

The input to the U-Net is the 2-channel noisy potential field, $x_t = [U_{\text{topo},t}, U_{\text{nav},t}]$. The conditional inputs—comprising the partial semantic map m_t, the target object category g, and the diffusion timestep t—are processed to form a shared embedding. Specifically, the map m_t is first encoded into a feature vector using a ResNet18 encoder. This vector is then concatenated with a learned embedding of the object category g (or a special null token $\emptyset$ for unconditional generation). This combined vector, along with the sinusoidal embedding of the timestep t, is projected by an MLP and subsequently integrated into the intermediate blocks of the U-Net via cross-attention mechanisms. This allows the model to modulate its noise prediction based on the agent's comprehensive state.

2) Two-Stage Training Curriculum and Objective: To effectively learn the dual objectives, we introduce a two-stage training curriculum. The model is optimized using a composite objective function, which is a weighted sum of the losses from each potential field channel:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{topo}}\mathcal{L}_{\text{topo}} + \lambda_{\text{nav}}\mathcal{L}_{\text{nav}}, \quad (5)$$

where $\mathcal{L}_{\text{topo}}$ and $\mathcal{L}_{\text{nav}}$ represent the mean-squared error losses between the true and predicted noise on the topological and navigation channels, respectively (as in Eq. 2), and λ_{topo} and λ_{nav} are balancing hyperparameters.

a) Stage 1: Topological Prior Pre-training.: The initial stage focuses exclusively on learning the environment's structure. The model is trained to generate only the topological potential field. In this stage, the navigation channel of the ground truth x_0 is masked (i.e., set to zero), the navigation loss weight is set to zero ($\lambda_{\text{nav}} = 0$), and the object condition g is consistently set to the null token $\emptyset$. The objective simplifies to minimizing $\mathcal{L}_{\text{topo}}$. This initial stage provides the model with a robust, task-agnostic foundation for spatial reasoning.

b) Stage 2: Joint Navigation and Mapping Fine-tuning.: Subsequently, the model is initialized with the pre-trained weights and fine-tuned on the complete dual-objective task. It is now provided with the complete 2-channel ground-truth fields $[U_{\text{topo}}, U_{\text{nav}}]$ and valid object category conditions g. The model is optimized using the full composite loss from Eq. 5 with $\lambda_{\text{nav}} > 0$. This stage adapts the learned structural prior to the specific demands of goal-directed navigation, enabling the model to synergistically reason about both exploration (via the topological map) and goal-directed navigation.

D. Navigation and Mapping with Generated Potentials

Our navigation framework operates in a closed-loop manner, where the agent continuously updates its map of the world and replans its path toward the goal. The process at each global replanning step consists of three main stages: semantic map construction, potential field generation via our DIPP model, and online topological mapping.

a) Semantic Map Construction: The agent first updates its persistent, top-down allocentric map, m_t. Following the established procedure of Chaplot et al. [6], this is achieved by projecting the current egocentric semantic segmentation and depth observation into a 3D point cloud, transforming it into the global coordinate frame using the agent's current pose p_t, and aggregating it with the map from the previous step, m_{t-1}. The resulting map m_t contains channels representing obstacles, explored areas, and semantic object categories, serving as the conditional input for our generative model.

b) Potential Field Generation and Topological Mapping: Conditioned on the updated map m_t and the target object category g, the trained DIPP model generates two distinct potential fields: a navigation field $\hat{U}_{\text{nav}}$ indicating promising directions toward the target, and a topological field $\hat{U}_{\text{topo}}$ highlighting structurally significant locations. This generation is an iterative denoising process, starting from pure Gaussian noise and enhanced by Classifier-Free Guidance (CFG) for stronger goal-conditioning.

The core of our method is the incremental construction of a robust topological graph, G_t, using these generated fields. This process leverages a hybrid strategy to ensure both semantic relevance and structural robustness. Candidate nodes are derived from two complementary sources:

(1) Learned Candidates from $\hat{U}_{\text{topo}}$: The primary source is the topological potential field, $\hat{U}_{\text{topo}}$, which is refined by a *geometric centrality prior*. Specifically, we compute a distance transform of the traversable area and use it to create a weighting map that penalizes locations near walls or obstacles. This encourages the selection of more representative nodes in the center of rooms and corridors.

(2) Geometric Candidates from Skeletonization: During initial exploration (when map coverage is below a threshold τ_{exp}), we supplement the learned candidates with nodes derived from *skeletonizing* the traversable map to provide a robust initial graph structure.

All candidate nodes (new and historical) are then globally refined. First, we apply *DBSCAN clustering* to merge redundant, proximate nodes into single representative centroids while preserving isolated ones as outliers. Second, we reconstruct the graph's connectivity using a *Relative Neighborhood Graph (RNG)* approach, where each potential edge is validated by an A* path search to ensure traversability. This two-step refinement ensures the final graph G_t is both sparse and structurally coherent.

c) Waypoint Selection: The resulting persistent graph, G_t, enables strategic, long-term planning. To validate its contribution, we compare two planners. Our main *Hierarchical Planner* identifies the semantic goal location in $\hat{U}_{\text{nav}}$, finds the closest nodes in G_t to the agent and the goal, plans a path on the graph using A*, and selects the next node on this path as a subgoal. For our ablation studies, a simpler *Reactive Planner* bypasses the graph, relying on a direct pixel-wise fusion $U_{\text{final}} = (1-\alpha)\hat{U}_{\text{nav}} + \alpha\hat{U}_{\text{topo}}$ to select a waypoint. In both cases, the final waypoint is passed to a low-level planner for collision-free navigation, which uses the Fast Marching Method (FMM) [22].

V. EXPERIMENTS

A. Experimental Setup

We conduct experiments in the Habitat simulator [23] on the Gibson dataset [24], following the standard ObjectNav setup from the CVPR 2021 Challenge. For the Gibson experiments, we utilize the ObjectNav dataset, following the settings of [6], which includes six object classes: "chair", "couch", "potted plant", "bed", "toilet", and "tv".

Evaluation Metrics: We assess ObjectNav performance using three standard metrics: Success Rate (**SR**), the proportion of successful trials; Success weighted by Path Length (**SPL**) , which evaluates path efficiency relative to an agent; and Distance to Target (**DTS**), the distance (in meters) of the agent from the success threshold of the goal object at the end of the episode [4]. To evaluate the structural accuracy of the generated map, we introduce Node Recall (NR). A ground-truth topological node is considered "recalled" if there is at least one generated node within a predefined distance threshold (we use 1.5m in our experiments). NR is then calculated as the fraction of ground-truth nodes that are successfully recalled.

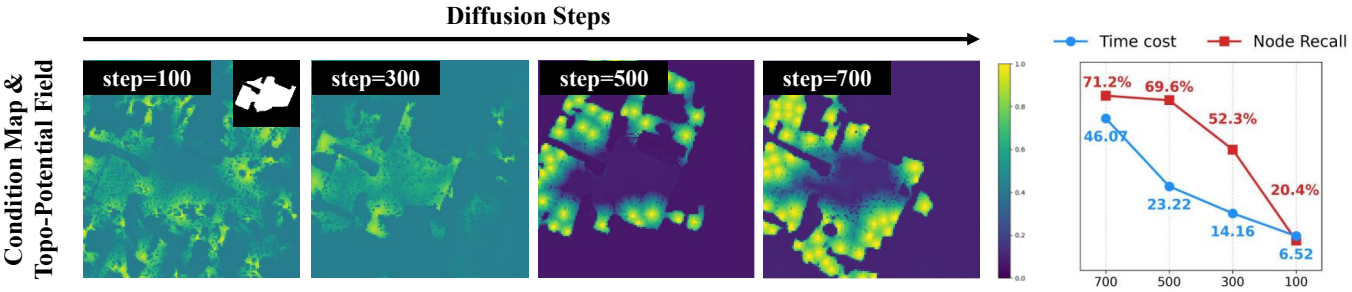

Fig. 4: **Analysis of Denoising Steps at Inference.** The left panel visualizes the generated topological potential field at different denoising steps, conditioned on the partial map (top-left inset). The potential field becomes progressively more defined as the step count increases. The right panel quantitatively shows the trade-off between computational cost (Time cost, in s) and mapping quality (Node Recall). A significant reduction in computation time can be achieved with only a minor drop in performance by reducing steps from 700 to 500.

B. Baselines

We compare DIPP against a comprehensive suite of modular and end-to-end baselines, spanning classical techniques and state-of-the-art methods.

a) Modular Methods: This category includes classical, learning-based, and generative approaches. We start with the foundational **Frontier-Based Exploration (FBE)** [25], a seminal robotics pipeline that navigates to the nearest frontier on an occupancy map. In the realm of reinforcement learning, we include **SemExp** [6], which learns a goal-sampling policy. More recent methods have shifted towards supervised, interaction-free paradigms. **PONI** [1], for instance, reframes target search as a perception task trained with supervised learning. The latest approaches leverage generative models and Large Language Models (LLMs). **L3MVN** [26] utilizes the common-sense knowledge of LLMs to infer semantically relevant frontiers; **T-Diff** [2] employs a diffusion model to generate future trajectory sequences; and **Imagine-Before-Go** [27] hallucinates unobserved map regions to infer the target's likely location.

b) End-to-End Methods: For end-to-end comparisons, we include **DD-PPO** [7], a widely-used distributed reinforcement learning baseline. Additionally, we include **Self-Supervised In-Situ Finetuning (SSIF)** [28], an approach that uses location consistency as a self-supervision signal for fine-tuning in new environments without 3D ground truth.

C. Implementation Details

We implement DIPP using a 'UNet2DConditionModel' [29] with attention mechanisms, which takes the noisy potential fields as input. The model is conditioned on the partial semantic map and a learned object embedding, which are processed into a context vector and injected into the U-Net's cross-attention layers. We train for 1000 diffusion steps with a linear beta schedule ($\beta_0 = 10^{-4}, \beta_T = 0.02$) using the AdamW optimizer [30] at a learning rate of 1.0×10^{-4}. Our two-stage curriculum consists of 200 epochs of topological pre-training to establish a structural prior, followed by 150 epochs of joint fine-tuning with balanced loss weights ($\lambda_{\text{topo}} = \lambda_{\text{nav}} = 0.5$). To facilitate Classifier-Free Guidance (CFG), the object condition is replaced by a null token with a probability of 0.1 during fine-tuning. We use an Exponential Moving Average (EMA) of the model weights for all evaluations.

At inference, we employ CFG with a guidance scale of 7.5 to enhance goal-directed generation. While the model is trained for 1000 steps, we use an accelerated 500-step sampling strategy for all experiments. As justified by our analysis in Fig. 4, this strikes an optimal balance, retaining over 97% of the peak Node Recall (a 1.6% drop from 71.2%) while reducing computation time by approximately 50%. The agent operates within the Habitat simulator with a 30° turning angle, a 0.25 m forward step distance, and a maximum episode length of 500 steps. All models are trained on NVIDIA RTX 3090 GPUs.

D. Evaluation Results

We present the empirical evaluation of DIPP, first comparing its performance against state-of-the-art baselines. Subsequently, we conduct in-depth ablation studies to validate our key design choices.

TABLE I: Main results on the Gibson ObjectNav validation set. The table compares our direct **potential fusion** approach against the **full hierarchical system**. Best results are in **bold**.

Method	SR ↑	SPL ↑	DTS ↓
Modular Methods			
Frontier Exploration[25]	0.643	0.283	1.78
SemExp [6]	0.717	0.396	1.39
PONI [1]	0.736	0.410	1.25
L3MVN [26]	0.769	0.388	1.01
T-Diff [2]	0.796	0.449	**1.00**
Imagine-Before-Go [27]	0.780	0.440	1.11
End-to-End Method			
DD-PPO [7]	0.150	0.107	3.24
SSIF [28]	0.600	0.312	1.89
Our Method			
DIPP (Potential Fusion)	0.748	0.454	1.18
DIPP (Hierarchical System)	**0.807**	**0.476**	1.08

1) Main Results: Navigation and Mapping Performance: The primary navigation results are summarized in Table I. Our 'DIPP (Potential Fusion)' variant, using only direct

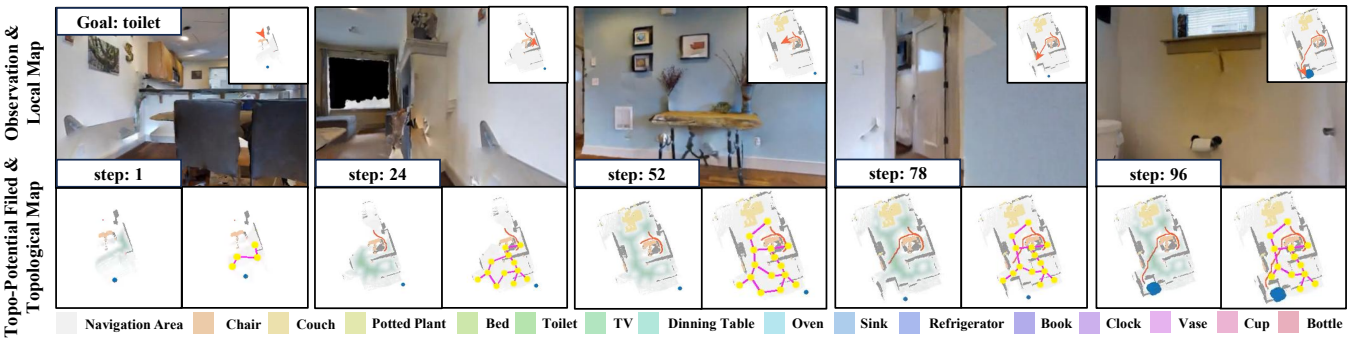

Fig. 5: A qualitative visualization of navigation with DIPP, tasked with finding a 'toilet'. The main images show the agent's first-person view, with top-right insets displaying the local map and agent trajectory. The bottom row visualizes our model's core outputs: the generated Topological Potential Field ($\hat{U}_{\text{topo}}$) and the resulting topological graph. **(Steps 1-52)** During initial exploration, our model generates a dynamic Topological Potential Field ($\hat{U}_{\text{topo}}$, green heatmaps), whose peaks directly guide the placement and adjustment of new topological nodes (yellow dots). **(Steps 52-96)** The value of this synergy becomes evident at the hallway junction (step 52), where the agent leverages the accumulated topological graph—itself a product of the potential field—to make an informed decision. This demonstrates our framework's core feedback loop, where intelligent map construction provides the global awareness for efficient navigation to the goal.

pixel-wise fusion, already outperforms strong baselines like PONI. This validates that the generated topological potential ($\hat{U}_{\text{topo}}$) serves as an effective implicit structural prior for navigation.

Second, our full hierarchical system achieves state-of-the-art performance, surpassing all baselines in both Success Rate (SR) and Success weighted by Path Length (SPL) with an SR of 0.807 and an SPL of 0.476. The significant performance gain over the potential fusion baseline demonstrates the value of our hierarchical approach. By structuring the raw topological potential into an explicit, persistent graph, the agent's decision-making is elevated from a tactical, pixel-level choice to a strategic, node-level selection. This abstraction provides enhanced robustness and long-term memory, enabling the agent to navigate complex layouts more coherently and achieve a higher success rate. Qualitative examples of this intelligent, graph-guided exploration are shown in Fig. 5.

A core tenet of our work is that the agent should not only navigate successfully but also build a meaningful representation of the environment. We evaluate this capability via Node Recall (NR), where our DIPP model achieves a score of **0.61**. This drastically outperforms key non-learning baselines, such as a **Random Baseline** (NR: 0.11), which places nodes randomly in explored space, and a **Frontier Baseline** (NR: 0.22), which places nodes at exploration frontiers. This result validates that our model, trained with the objective defined in Eq. 3, has successfully learned to identify locations of high structural importance rather than simply memorizing exploration patterns. The high NR score demonstrates that the generated potential field $\hat{U}_{\text{topo}}$ accurately reflects the underlying geometric skeleton of the environment.

2) Ablation Studies: To validate the specific design choices within DIPP, we conduct two critical ablation studies.

a) The Role of Topological Potential in Navigation.: First, we investigate whether the topological potential $\hat{U}_{\text{topo}}$ merely serves the mapping objective or if it actively contributes to the navigation task itself. To this end, we perform an ablation on the fusion weight α in the reactive planner, sweeping $\alpha \in 0.0, 0.3, 0.5, 0.7$, where $\alpha = 0.0$ fully disables topological guidance during inference and larger α places more emphasis on $\hat{U}_{\text{topo}}$. As shown in Table II, removing $\hat{U}_{\text{topo}}$ ($\alpha = 0.0$) leads to a substantial degradation across all metrics: SR drops from 0.748 to 0.641 and SPL decreases from 0.454 to 0.347, while DTS increases from 1.18 to 1.98. Introducing a moderate amount of topological guidance already yields clear gains (e.g., $\alpha = 0.3$ achieves SR/SPL of 0.701/0.412 and reduces DTS to 1.49). Performance peaks at the balanced setting $\alpha = 0.5$, whereas further increasing α to 0.7 slightly hurts SR/SPL (0.724/0.435) and increases DTS (1.24), suggesting that overweighting the structural prior can dilute the goal-directed signal from $\hat{U}_{\text{nav}}$. Overall, these results empirically confirm that $\hat{U}_{\text{topo}}$ is not redundant: it functions as an essential exploration prior that guides the agent to traverse the environment in a structurally coherent manner, which is crucial for efficiently locating objects in complex, unseen scenes.

TABLE II: Ablation on the topological potential fusion weight α. Removing the topological guidance ($\alpha = 0.0$) significantly degrades navigation performance, confirming its utility.

α	SR ↑	SPL ↑	DTS ↓
0.0	0.641	0.347	1.98
0.3	0.701	0.412	1.49
0.5	**0.748**	**0.454**	**1.18**
0.7	0.724	0.435	1.24

Finally, we validate our two-stage training curriculum by comparing our full model with a variant trained from a randomly initialized U-Net directly on the joint objective (Stage 2 only). The results in Table III show that removing

topological pre-training leads to a clear drop in performance. We hypothesize that jointly learning a structural prior and a goal-directed navigation policy from scratch is a difficult optimization problem. Topological pre-training provides a foundation for spatial reasoning, allowing the subsequent fine-tuning stage to adapt this knowledge for object navigation and yielding a more effective and efficient policy.

TABLE III: Ablation on the two-stage training curriculum. Pre-training on the topological prior (Stage 1) is crucial for achieving optimal performance.

Training Strategy	SR ↑	SPL ↑	DTS ↓
Ours (Stage 1 + Stage 2)	**0.748**	**0.454**	**1.18**
Ours (Only Stage 2)	0.542	0.291	2.53

b) The Necessity of Pre-training: Our "Only Stage 2" variant (SR: 0.542, SPL: 0.291) is close to PONI using only its object potential function (SR: 0.588, SPL: 0.349) [1], which is unsurprising because both are effectively trained with a navigation-only objective without a structural prior. This suggests that the improvement of our full model mainly comes from the structural knowledge learned in Stage 1.

VI. CONCLUSIONS

We presented DIPP, a diffusion-based potential planner that jointly generates navigation and topological potentials and incrementally builds a persistent graph for hierarchical ObjectNav. Experiments in Habitat on Gibson show improved SR/SPL and accurate topological maps. Future work will focus on faster diffusion sampling and scalable lifelong map management.

REFERENCES

[1] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "Poni: Potential functions for objectgoal navigation with interaction-free learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 890–18 900.

[2] X. Yu, S. Zhang, X. Song, X. Qin, and S. Jiang, "Trajectory diffusion for objectgoal navigation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 110 388–110 411, 2024.

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

[4] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "Objectnav revisited: On evaluation of embodied agents navigating to objects," *arXiv preprint arXiv:2006.13171*, 2020.

[5] S. Thrun, "Probabilistic robotics," *Communications of the ACM*, vol. 45, no. 3, pp. 52–57, 2002.

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

[7] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," *arXiv preprint arXiv:1911.00357*, 2019.

[8] K. Fang, A. Toshev, L. Fei-Fei, and S. Savarese, "Scene memory transformer for embodied agents in long-horizon tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 538–547.

[9] J. Ye, D. Batra, A. Das, and E. Wijmans, "Auxiliary tasks and exploration enable objectnav," *arXiv preprint arXiv:2104.04112*, 2021.

[10] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural slam," *arXiv preprint arXiv:2004.05155*, 2020.

[11] N. Correll, B. Hayes, C. Heckman, and A. Roncone, *Introduction to autonomous robots: mechanisms, sensors, actuators, and algorithms*. Mit Press, 2022.

[12] H. Yoo, Y. Choi, J. Park, and S. Oh, "Commonsense-aware object value graph for object goal navigation," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4423–4430, 2024.

[13] E. Beeching, J. Dibangoye, O. Simonin, and C. Wolf, "Learning to plan with uncertain topological maps," in *European Conference on Computer Vision*. Springer, 2020, pp. 473–490.

[14] D. S. Chaplot, R. Salakhutdinov, A. Gupta, and S. Gupta, "Neural topological slam for visual navigation," in *CVPR*, 2020.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[17] T. Pearce, T. Rashid, A. Kanervisto, D. Bignell, M. Sun, R. Georgescu, S. V. Macua, S. Z. Tan, I. Momennejad, K. Hofmann *et al.*, "Imitating human behaviour with diffusion models," *arXiv preprint arXiv:2301.10677*, 2023.

[18] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[19] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[20] P. Wu, Y. Mu, B. Wu, Y. Hou, J. Ma, S. Zhang, and C. Liu, "Voronav: Voronoi-based zero-shot object navigation with large language model," *arXiv preprint arXiv:2401.02695*, 2024.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[22] J. A. Sethian, "Fast marching methods," *SIAM review*, vol. 41, no. 2, pp. 199–235, 1999.

[23] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra, "Habitat 2.0: Training home assistants to rearrange their habitat," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[24] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson env: Real-world perception for embodied agents," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9068–9079.

[25] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*. IEEE, 1997, pp. 146–151.

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

[27] S. Zhang, X. Yu, X. Song, X. Wang, and S. Jiang, "Imagine before go: Self-supervised generative map for object goal navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 414–16 425.

[28] S. Y. Min, Y.-H. H. Tsai, W. Ding, A. Farhadi, R. Salakhutdinov, Y. Bisk, and J. Zhang, "Self-supervised object goal navigation with in-situ finetuning," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7119–7126.

[29] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, D. Nair, S. Paul, W. Berman, Y. Xu, S. Liu, and T. Wolf, "Diffusers: State-of-the-art diffusion models," https://github.com/huggingface/diffusers, 2022.

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.