# Learning to Explore Efficiently: Heterogeneous Topological Graphs and Lightweight Global Reasoning for Robotic Exploration

Zhi Li, Kairao Zheng, Yiqing Yuan, Junlong Huang, Xiaoxun Zhang, *Student Member, IEEE*, Jinze Wu, and Hui Cheng, *Member, IEEE*

***Abstract*—Autonomous exploration in large-scale, unknown environments remains a significant challenge in mobile robotics. In this letter, we propose a scalable exploration framework that integrates heterogeneous topological representations, lightweight global-local graph reasoning, and reinforcement learning. Our framework is computationally efficient, comprehensively considers global spatial context, and generalizes effectively across diverse environmental scenarios. We introduce a compact topological abstraction to encode crucial spatial and semantic information, substantially reducing map complexity. A novel hybrid inference module, combining linear global attention with local graph convolutions, effectively integrates long-range exploration with comprehensive local coverage. We further design a minimalist reward function paired with a curriculum learning to ensure stable training and enhanced generalization. A viewpoint-based action masking mechanism further refines the action space, accelerating learning convergence. Extensive simulations demonstrate our method consistently surpasses state-of-the-art baselines, achieving up to 14.5% shorter exploration time and 18.8% reduced path length while maintaining low computational overhead. Real-world experiments further validate the practical effectiveness of our approach for robotic exploration tasks.**

***Index Terms*—Autonomous systems, motion and path planning, reinforcement learning.**

## I. Introduction

AUTONOMOUS exploration in large, unstructured environments is a fundamental challenge in mobile robotics, with applications in search and rescue, industrial inspection, and environmental monitoring [1], [2]. Despite decades of progress, designing exploration systems that are both computationally efficient and behaviorally robust remains a challenging task. Classical frontier-based methods [3], [4], [5] address this challenge by decomposing the task into frontier detection, viewpoint selection, and path planning, providing an interpretable framework for exploration. However, these methods incur substantial computational overhead as the environment expands.

To improve scalability, recent research explores learning-based approaches, particularly Deep Reinforcement Learning (DRL) [6], [7], which optimizes exploration policies through interaction with the environment. DRL demonstrates strong performance in long-horizon decision making under partial observability. However, sparse and delayed rewards and the absence of structured priors hinder policy learning in large environments [8]. Furthermore, existing methods often overlook geometric information embedded in environmental topology, leading to inefficient exploration and limited generalization. Bridging the gap between the structure-aware reasoning and the flexibility of learning-based policies thus remains a pressing challenge.

In this letter, we propose a scalable graph-based exploration framework that integrates heterogeneous graph representations with global-local reasoning and RL. The main contributions are summarized as follows.

- *Heterogeneous Graph Abstraction:* We design a compact yet expressive topological representation of the environment, with diverse node types (frontiers, borders, skeletons, viewpoints) that encode distinct spatial and semantic roles. This abstraction reduces the complexity of large-scale maps while preserving critical cues such as connectivity, visibility, and coverage gain.
- *Global-Local Graph Inference Module:* We develop a hybrid graph neural network (GNN) that combines linear global attention with local graph convolution. This enables efficient $\mathcal{O}(N)$ reasoning while preserving structural bias, allowing the policy to balance long-range exploration with low computation.
- *Sparse Reward Optimization via Curriculum Learning:* We design a minimalist reward function with path cost and terminal completion terms. This avoids complex

Received 25 May 2025; accepted 7 October 2025. Date of publication 17 October 2025; date of current version 24 October 2025. This article was recommended for publication by Associate Editor P. Falco and Editor J. Kober upon evaluation of the reviewers' comments. This work was supported by the National Key R&D Program of China under Grant 2022ZD0119602. *(Zhi Li and Kairao Zheng contributed equally to this work.) (Corresponding author: Hui Cheng.)*

Zhi Li is with the School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China.

Kairao Zheng, Yiqing Yuan, Xiaoxun Zhang, and Hui Cheng are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China (e-mail: chengh9@mail.sysu.edu.cn).

Junlong Huang is with the School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510006, China.

Jinze Wu is with the Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy, and also with the School of Automation and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen 518055, China.

This article has supplementary downloadable material available at https://doi.org/10.1109/LRA.2025.3622906, provided by the authors.

Digital Object Identifier 10.1109/LRA.2025.3622906

reward shaping and hand-tuned weights, preventing reward hacking. To improve training efficiency and generalization, we adopt a curriculum learning that gradually increases decision frequency.

- *Viewpoint-Based Action Masking:* We apply classical viewpoint extraction to constrain the action space to informative viewpoints. This mitigates over-squashing effects from neighbor-based selection. It also ensures task completion and significantly accelerates learning by reducing variance and improving sample efficiency.

Extensive experiments in diverse large-scale scenarios show that our framework improves exploration performance, responsiveness, and generalization over state-of-the-art methods. Our method reduces exploration time by up to 14.5% and path length by up to 18.8% in complex environments, while maintaining high-frequency decision-making capabilities with low computational overhead, highlighting its efficiency and practical effectiveness for real-time exploration.

## II. Related Work

### A. Exploration-Oriented Map Representations

Efficient spatial representation is crucial to robotic exploration. Dense occupancy grids are costly in large-scale environments [9], leading to sparse alternatives such as Euclidean Signed Distance Fields (ESDF) [10], topological graphs [11], and skeleton-based maps [12]. Skeleton-based methods, especially those from Hamilton-Jacobi formulations, capture navigable pathways with minimal redundancy. Frontier-based exploration leverages topological abstractions to simplify viewpoint selection and planning [4], [5], [13]. However, homogeneous representations cannot distinguish between frontiers, borders, and viewpoints, limiting their expressiveness. To address this, we introduce a heterogeneous representation that explicitly encodes different node types, improving both structural clarity and semantic precision.

### B. Efficient Graph-Based Reasoning Methods

GNNs, including Graph Convolutional Networks (GCNs) [14], Graph Attention Networks (GATs) [15], and Transformer-based models [16], have shown promise in navigation and exploration. However, traditional GNNs often encounter quadratic complexity and issues such as over-smoothing and over-squashing [17]. Recent variants like Linear Transformers and Graph Linear Attention [18], [19] reduce complexity to linear order. However, these approaches may compromise local structural bias or remain inefficient in complex real-world scenarios. To address this, we propose a hybrid graph reasoning architecture that combines linear-complexity global attention with local graph convolutions, capturing long-range dependencies while preserving local structure, improving inference efficiency and scalability for real-time exploration in large-scale environments.

### C. Action Space Design for Exploration

The design of the action space strongly affects the efficiency of RL-based exploration methods. Neighbor-based action selection, which uses adjacent nodes in a topological map [7]. leads to long decision sequences and unstable policy training under sparse rewards. In dense graphs, it further causes over-squashing [17], reducing learning efficiency. In contrast, viewpoint-based action selection samples a limited set of informative viewpoints, reducing decision sequence length and ensuring exploration completion [13], [20], which alleviates sparse rewards and over-squashing. Inspired by these advantages, we introduce a viewpoint-based action mask that integrates classical viewpoint extraction into RL, constraining the action space to ensure task completeness and improve stability by addressing challenges related to sparse rewards and information compression.

### D. Sparse Reward Optimization and Curriculum Learning

Sparse rewards pose a significant challenge in RL-based exploration when feedback is rare [21]. Go-Explore [22] mitigates this by revisiting promising states, while other approaches use coverage rewards [23], curiosity [24], or hand-tuned reward functions [25], which risk reward hacking and limited generalization [26]. Curriculum learning, which incrementally increases task complexity during training, has shown effectiveness in improving RL stability and generalization [27]. To address these issues, we propose a minimalist reward function consisting only of a path cost and a terminal completion reward, avoiding intricate reward shaping and minimizing reward hacking. Additionally, we introduce a straightforward curriculum that gradually increases decision frequency to enhance training stability, policy generalization, and overall exploration performance.

## III. Preliminaries

We formulate autonomous exploration as a Partially Observable Markov Decision Process (POMDP) and describe the scene representation and reward design.

### A. POMDP Formulation

Autonomous exploration task is to find the shortest path that covers all traversable areas of an unknown environment while incrementally updating a belief map. We formulate this as a POMDP $\langle \mathcal{S}, \mathcal{A}, P, \mathcal{O}, R, \gamma \rangle$, where $\mathcal{S}$ is the full state space, $\mathcal{O}$ the partial observations from sensors (e.g., LiDAR), $\mathcal{A}$ the discrete action set of candidate waypoints, $P(s'|s,a)$ the state-transition model, $R(s,a,s')$ the reward function, and $\gamma \in [0,1)$ the discount factor.

We represent the environment as a topological graph $G = (\mathcal{V}, \mathcal{E})$ to bridge the sim-to-real gap when processing raw sensor data. The state $s$ corresponds to the global topological map that captures the complete spatial information. This graph efficiently encodes the environment where nodes $v \in \mathcal{V}$ capture key attributes (e.g., frontier values, traversability metrics) and edges $e \in \mathcal{E}$ represent feasible paths. The observation $\mathcal{O}$ at each decision step is the current topological graph built from the robot's partial knowledge. The action space $\mathcal{A}$ consists of selecting the next node to visit, simplifying planning compared to dense map representations. The reward function combines distance cost and completion bonus:

$$R(s,a,s') = -\lambda \cdot R_c(s,s') + R_f(s'), \tag{1}$$

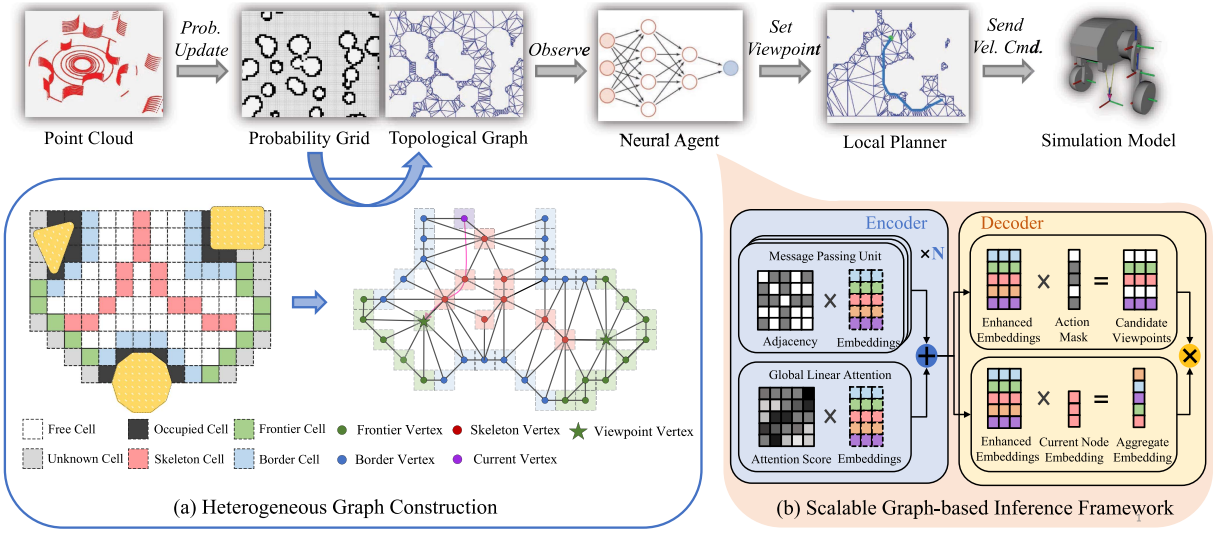



Fig. 1. Overview of the proposed exploration framework. (a) *Heterogeneous graph construction:* Incremental abstraction from occupancy grid to topological graph through boundary definition, skeleton extraction, and viewpoint insertion, producing semantically distinct node types (frontier, border, skeleton, viewpoint). (b) *Graph-based inference framework:* A lightweight encoder-decoder architecture combining linear global attention and local graph convolution efficiently generates embeddings, aggregates global context, and selects optimal viewpoints via an attention-guided action mask.

where $R_c(s, s')$ is the path cost, and $R_f(s')$ is the final completion reward. Parameter $\lambda$ ensures $\lambda \cdot R_c < \gamma^T R_f$, prioritizing exploration completion over minimizing path length, where $T$ denotes the truncation horizon.

### B. Heterogeneous Graph-Based Topological Mapping

To preserve rich environmental information while minimizing redundant nodes, we represent the environment as a heterogeneous graph $G = (\mathcal{V}, \mathcal{E}, \mathcal{T})$, where $\mathcal{V}$ denotes vertices, $\mathcal{E}$ denotes edges $(v_i, v_j)$ connecting these vertices, and $\mathcal{T}$ denotes predefined vertex types. Each vertex $v_i \in \mathcal{V}$ is associated with an attribute vector $\mathbf{u}$ tailored to its specific role. Here, we focus explicitly on distinguishing vertex types and ignore edge attributes.

*1) Vertex Extraction:* Starting from a 2D occupancy grid map classified into free, occupied, and unknown cells, we build a heterogeneous graph by extracting key vertices $\mathcal{V}$ in three steps: *defining boundaries*, *extracting skeleton*, and *inserting viewpoints*. First, we define two types of boundary vertices (see Fig. 1): **Frontier vertices** $v_f$, extracted from free cells adjacent to unknown cells, marking exploration frontiers, and **Border vertices** $v_b$, extracted from free cells adjacent to occupied cells, indicating traversability boundaries. Next, inspired by [12], we extract **Skeleton vertices** $v_k$ by deriving a Hamilton-Jacobi skeleton (see Fig. 1) from the occupancy grid's ESDF, which identifies essential pathways in free space. Finally, following the classical frontier-based exploration strategy [20], we compute **Viewpoints** $v_v$ that strategically observe unexplored regions. This structured approach ensures compactness and semantic clarity for effective exploration planning.

*2) Edge Construction:* With $\mathcal{V}$ defined, we establish the edges $\mathcal{E}$ to ensure connectivity. We generate candidate edges by performing a Delaunay triangulation over the entire node set using the CGAL library [28], providing a well-distributed set of connections. Each edge is then validated by collision checking against the occupancy grid, and those intersecting occupied or unknown cells are removed. This deterministic filtering procedure eliminates invalid or redundant connections while preserving essential pathways, resulting in a sparse yet well-connected roadmap suitable for exploration.

*3) Vertex Attributes:* Each vertex type is assigned a concise attribute vector $\mathbf{u}$ tailored for decision-making. Frontier $(v_f)$ and border $(v_b)$ store spatial coordinates $\mathbf{u} = [x, y]$, sufficient for marking transitions to unknown or obstructed areas. Skeleton vertices additionally include ESDF-based clearance information $\mathbf{u} = [x, y, d]$. Lastly, viewpoint vertices $(v_v)$ encode observation utility as $\mathbf{u} = [x, y, \epsilon]$, where $\epsilon$ quantifies potential exploration gain. This tailored attribute structure enriches the graph representation without redundancy, enhancing exploration effectiveness and planning efficiency.

## IV. Scalable Graph-Based Inference Framework

### A. Framework Overview

We propose a scalable encoder–decoder framework for robotic exploration (Fig. 1). The encoder combines a linear global attention mechanism [19] with local graph convolutions to capture both long-range and local spatial contexts in $\mathcal{O}(N)$ time, while the decoder applies attention-based action selection constrained by a viewpoint-based action mask. This design ensures linear scalability and robust decision-making in complex environments.

### B. Encoder

We first encode vertex attributes $\mathbf{u}_v$ into a latent embedding $\mathbf{h}_v^{(0)} = f_t(\mathbf{u}_v)$, where $t = \mathcal{T}(v)$, and $f_t$ is a type-specific MLP. The resulting embeddings, $\mathbf{h}^{(0)} \in \mathbb{R}^{N \times d}$, form the basis for subsequent computations.

To efficiently capture global interactions, we adopt a linear global attention mechanism. Unlike conventional softmax attention with $\mathcal{O}(N^2)$ complexity [29], our method operates in linear time, $\mathcal{O}(N)$, enabling scalable reasoning on large graphs. Specifically, given initial embeddings $\mathbf{h}^{(0)}$, we compute normalized query, key, and value matrices:

$$\mathbf{Q} = f_Q\left(\mathbf{h}^{(0)}\right), \quad \tilde{\mathbf{Q}} = \frac{\mathbf{Q}}{\|\mathbf{Q}\|_F},$$
$$\mathbf{K} = f_K\left(\mathbf{h}^{(0)}\right), \quad \tilde{\mathbf{K}} = \frac{\mathbf{K}}{\|\mathbf{K}\|_F},$$
$$\mathbf{V} = f_V\left(\mathbf{h}^{(0)}\right), \tag{2}$$

where $f_Q, f_K, f_V$ are linear transformations, and $\|\cdot\|_F$ denotes the Frobenius norm. Node embeddings are updated efficiently using a diagonal normalization matrix $\mathbf{D}$:

$$\mathbf{D} = \operatorname{diag}\left(\mathbf{1} + \frac{1}{N}\tilde{\mathbf{Q}}(\tilde{\mathbf{K}}^\top \mathbf{1})\right), \tag{3}$$

$$\mathbf{h}_g = \beta \mathbf{D}^{-1}\left[\mathbf{V} + \frac{1}{N}\tilde{\mathbf{Q}}(\tilde{\mathbf{K}}^\top \mathbf{V})\right] + (1-\beta)\mathbf{h}^{(0)}, \tag{4}$$

where $\mathbf{1} \in \mathbb{R}^N$ is an all-ones vector, and $\beta$ is a hyperparameter balancing global attention and residual connections. This design effectively captures long-range dependencies while preserving local structural integrity, making it particularly suitable for large-scale graph-based exploration.

To effectively incorporate local structural constraints from the graph adjacency matrix $\mathbf{A}$, we fuse the global embeddings $\mathbf{h}_g$ with local graph convolutional updates. Specifically, we employ a simple yet efficient approach:

$$\mathbf{h}_o = (1-\alpha)\,\mathbf{h}_g + \alpha\,\mathrm{GN}\big(\mathbf{h}^{(0)}, \mathbf{A}\big), \tag{5}$$

where $\mathrm{GN}(\cdot)$ denotes a scalable GCN [14], $\alpha$ balances global attention and local connectivity, and the resulting embeddings $\mathbf{h}_o = [\mathbf{h}_v]_{v=1}^N$ serve as inputs to the decoder. This integration ensures the model simultaneously captures global structural relationships and local graph connectivity, enhancing inference quality for exploration tasks.

### C. Decoder

The decoder aims to produce a global graph-level embedding from the node-level embeddings $\mathbf{h}_o$. To intuitively aggregate global information, we adopt a global attention mechanism [29], in which the embedding of the robot's current vertex $\mathbf{h}_c \in \mathbf{h}_o$ naturally serves as the query:

$$\mathbf{h}_{\text{agg}} = \mathrm{Attn}(\mathbf{h}_c, \mathbf{h}_o). \tag{6}$$

This approach effectively summarizes the global context by weighting node embeddings according to their relevance to the robot's current position, capturing essential structural and semantic information for subsequent decision-making. The resulting graph-level embedding $\mathbf{h}_{\text{agg}}$ is then jointly utilized by the policy head to determine the next action and by the value head to estimate expected returns, ensuring coherent exploration strategies.

**Algorithm 1:** Curriculum-Based PPO Training.

**Input:** total episodes $N$, PPO hyperparameters
**Stage allocation:** $N \times (0.625, 0.125, 0.125, 0.125)$
**Frequencies:** (1: full, 2: 0.125 Hz, 3: 0.5 Hz, 4: 2 Hz)
**for** episode $= 1$ to $N$ **do**
  determine stage $s$ by index; run episode with frequency of $s$
  collect trajectories $\mathcal{D}$, compute advantages with GAE
  update policy $\pi_\theta$ and value $V_\phi$ using PPO
**end for**
**return** trained policy $\pi_\theta$

### D. Action Selection and Value Estimation

Given the graph-level embedding $\mathbf{h}_{\text{agg}}$ from the decoder, we employ a single-head attention mechanism to determine the next action. Specifically, we compute relevance scores between $\mathbf{h}_{\text{agg}}$ and candidate vertex embeddings:

$$\mathrm{logits}(v) = \mathbf{h}_{\text{agg}}^\top \mathbf{h}_o[v], \quad v \in \mathcal{A}, \tag{7}$$

where $\mathcal{A}$ denotes the set of feasible actions. Action probabilities are then obtained via softmax normalization.

To improve learning efficiency and alleviate over-squashing in dense graphs, we introduce a viewpoint-based action mask. Unlike conventional adjacency-based masks, our method selectively retains only informative viewpoints—vertices expected to significantly reveal unexplored areas—resulting in a compact and semantically meaningful action set. This targeted pruning accelerates policy convergence and enhances exploration effectiveness.

Simultaneously, the value head estimates expected returns from the shared embedding $\mathbf{h}_{\text{agg}}$ using a lightweight MLP. This shared representation facilitates joint optimization of immediate action quality and long-term exploration performance, improving both coherence and sample efficiency.

### E. RL With Curriculum Learning Strategy

We train our exploration policy using Proximal Policy Optimization (PPO) [30], a stable and efficient reinforcement learning algorithm. To improve training stability and policy generalization, we employ a curriculum learning approach that progressively increases decision frequency during training. Initially, actions correspond to full viewpoint traversals (Stage 1), simplifying the action space and stabilizing early value estimation. Subsequently, we incrementally raise decision frequency (Stages 2–3: 0.125 Hz to 0.5 Hz), allowing the policy to smoothly adapt to finer-grained control. Finally, training transitions to the target high-frequency condition (Stage 4: 2 Hz), closely aligning with real-world robotic operations. This structured curriculum enhances convergence, robustness, and overall exploration effectiveness.

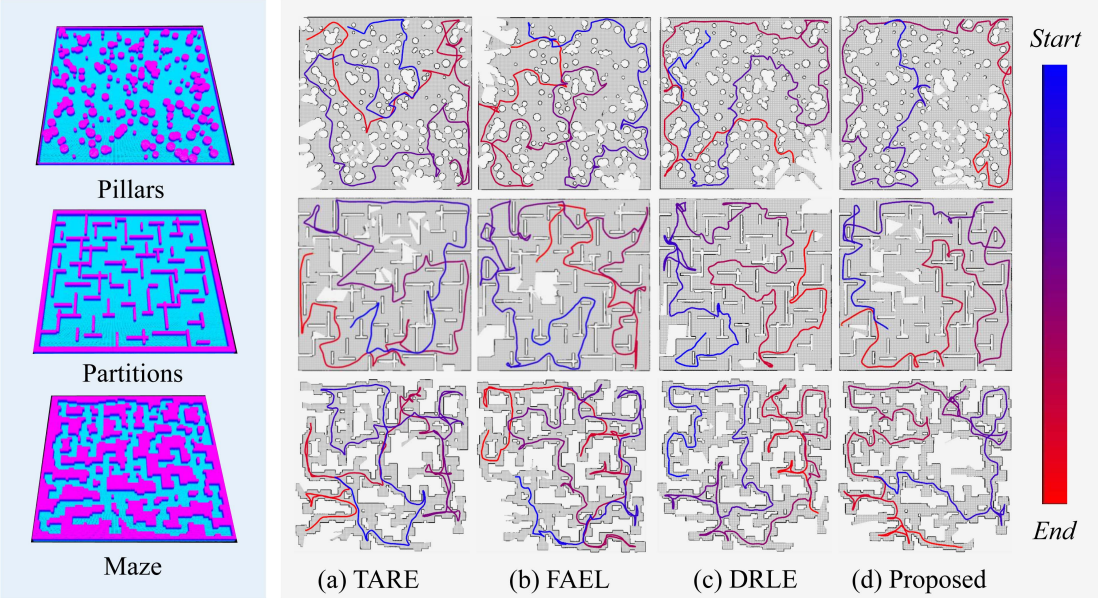



Fig. 2. Scene illustration. (*Left*) Side views of the point clouds for the three random scenes—Pillars, Partitions, and Maze; (*Right*) Topdown views of typical exploration trajectories by (a) TARE [4], (b) FAEL [5], (c) DRLE [7], and (d) our proposed method. The color gradient from blue to red indicates the progression of each trajectory.

## V. Experimental Results

### A. Implementation Details

We implement PPO using CleanRL [31] to ensure reproducibility and conduct training on a server with an AMD EPYC 7773X 64-Core CPU and two NVIDIA RTX 4090 GPUs. Simulations were performed within MARSIM [32], a lightweight LiDAR-based simulator. The policy and value networks share an encoder over heterogeneous graphs, where each node has a 6-dimensional feature embedded into a 64-dimensional latent space via a two-layer MLP. To improve generalization and stabilize training, coordinates are normalized by Min–Max scaling to [0,1] and further converted to local frames relative to the robot's position, which mitigates distribution shift and facilitates policy transfer.

TABLE I
Comparison of Different Methods on Random Scenes

| Scene & Expl. Area & Complexity | Method | Runtime (s) | Exploration Time (s) | | Movement Distance (m) | | Velocity (m/s) |
|---|---|---|---|---|---|---|---|
| | | | avg | std | avg | std | |
| Pillars $5491m^2$ 1.100 | TARE | 0.492 | 436.6 | 66.8 | 690.2 | 95.1 | 1.581 |
| | FAEL | 0.116 | 682.4 | 132.1 | 688.6 | 79.0 | 1.010 |
| | DRLE | 1.565 | 497.5 | 59.0 | 558.7 | 69.2 | 1.123 |
| | Proposed | 0.336 | **373.4** | **38.8** | **453.7** | **51.9** | 1.215 |
| Partitions $5462m^2$ 1.643 | TARE | 0.385 | 497.6 | 72.1 | 784.5 | 75.4 | 1.576 |
| | FAEL | 0.120 | 531.3 | 56.7 | 678.7 | 72.8 | 1.277 |
| | DRLE | 1.041 | 539.1 | 77.8 | 683.2 | 86.9 | 1.267 |
| | Proposed | 0.232 | **431.6** | **26.5** | **601.7** | **45.4** | 1.394 |
| Maze $3888m^2$ 1.500 | TARE | 0.330 | 563.2 | 37.9 | 868.1 | 58.9 | 1.542 |
| | FAEL | 0.061 | 607.2 | 86.7 | 773.9 | 67.1 | 1.275 |
| | DRLE | 0.484 | 544.4 | 87.8 | 721.9 | 85.2 | 1.445 |
| | Proposed | 0.217 | **540.3** | **36.8** | **669.8** | **51.9** | 1.240 |

### B. Simulation Benchmark and Analysis

We evaluate exploration efficiency in three randomized environments (*Pillars*, *Partitions*, and *Maze*), shown in Fig. 2. These scenarios differ in complexity, measured by the ratio of the shortest collision-free path to the straight-line distance.

Our method is compared with state-of-the-art LiDAR-based exploration algorithms: TARE [4], FAEL [5], and DRLE [7]. TARE employs dual-resolution mapping with dense local maps for feasible trajectories and sparse global maps for coverage guidance. FAEL uses rapid environment preprocessing and heuristic path optimization, supporting high-frequency replanning in large environments. DRLE utilizes deep reinforcement learning with attention-based policies, adaptively selecting viewpoints for reactive exploration.

We quantify the performance differences among the evaluated methods using three metrics: run time, exploration time, and movement distance. In all experiments, the robot is modeled as a differential-drive vehicle with a maximum linear velocity of $2\,\mathrm{m/s}$, linear acceleration of $2.5\,\mathrm{m/s^2}$, and a LiDAR sensor range of 40 m. In addition, the parameters for modules, including viewpoint sampling, topological graph construction, and trajectory optimization, were configured according to the officially recommended settings of the baseline methods for the respective scenarios. The exploration terminates when coverage reaches 95%. Each method is executed 10 times from identical random starting positions in each scenario.

Table I summarizes the averaged performance across 10 trials, with representative trajectories shown in Fig. 2. We further analyze these results with respect to versatility, exploration efficiency, and responsiveness.

*a) Versatility:* All methods complete the exploration tasks in each scenario. However, DRLE degrades in *Partitions* and *Pillars* due to large graph sizes, while our method maintains consistent effectiveness and generalizes better across varied, unseen environments.

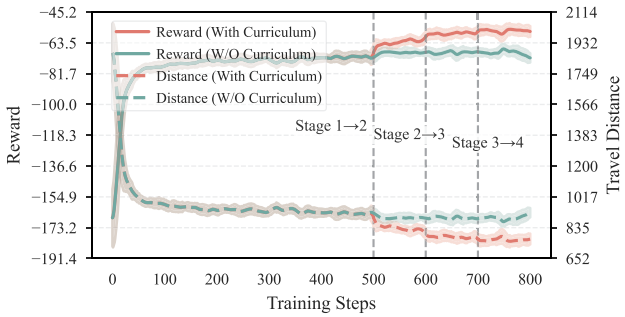



Fig. 3. Ablation study of curriculum learning. Performance comparison across four decision-frequency stages (Stage 1: full traversal; Stages 2–4: 0.125, 0.5, 2 Hz). A progressive curriculum leads to faster convergence and improved final performance.

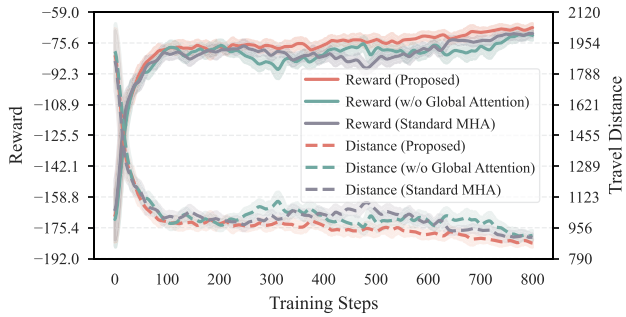



Fig. 4. Ablation study of global attention. Performance comparison of three variants: *GCN-only*, *MHA-only*, and *Proposed*. Our hybrid approach achieves the best balance of efficiency and exploration effectiveness.

*b) Exploration Efficiency:* As shown in Table I, our method achieves the best exploration time and path length across all scenarios. Compared to the second-best planner, our method reduces exploration time and path length by 14.5% and 18.8% in *Pillars*, 13.3% and 11.4% in *Partitions*, and 0.8% and 7.2% in *Maze*. Our method also yields the lowest variance across scenarios, showing stable performance. Unlike heuristic planners (TARE, FAEL), which optimize locally and produce redundant crossings, learning-based methods (ours and DRLE) achieve more uniform coverage, as illustrated in Fig. 2.

*c) Responsiveness:* Runtime comparisons in Table I highlight responsiveness differences. FAEL achieves fast updates through local planning but sacrifices path optimality. DRLE's transformer-based policy becomes less efficient as its topological graph grows. In contrast, the proposed method runs substantially faster than DRLE and remains comparable to heuristic baselines, supporting real-time deployment.

### C. Ablation Study of Curriculum

To evaluate the effectiveness of the proposed curriculum learning strategy, we conduct an ablation study across four decision-frequency stages, as shown in Fig. 3. Specifically, Stage 1 executes a full viewpoint traversal per decision (i.e., extremely low frequency). Stages 2–4 introduce intermediate and high-frequency decisions at 0.125 Hz, 0.5 Hz, and the target frequency of 2 Hz, respectively. Quantitative results indicate that progressively increasing decision frequency improves both convergence speed and final exploration performance. Training solely at the lowest frequency (Stage 1) simplifies the critic's task and enables rapid initial convergence, but lacks exposure to the fine-grained decision dynamics required at higher frequencies, limiting generalization. Conversely, training directly at the highest frequency (Stage 4) slows learning and causes instability in value estimation due to dense temporal feedback. Gradually transitioning through intermediate frequencies (Stages 2–3) allows stable adaptation of the critic, leading to improved policy generalization and performance under realistic high-frequency conditions.

### D. Ablation Study of Global Attention

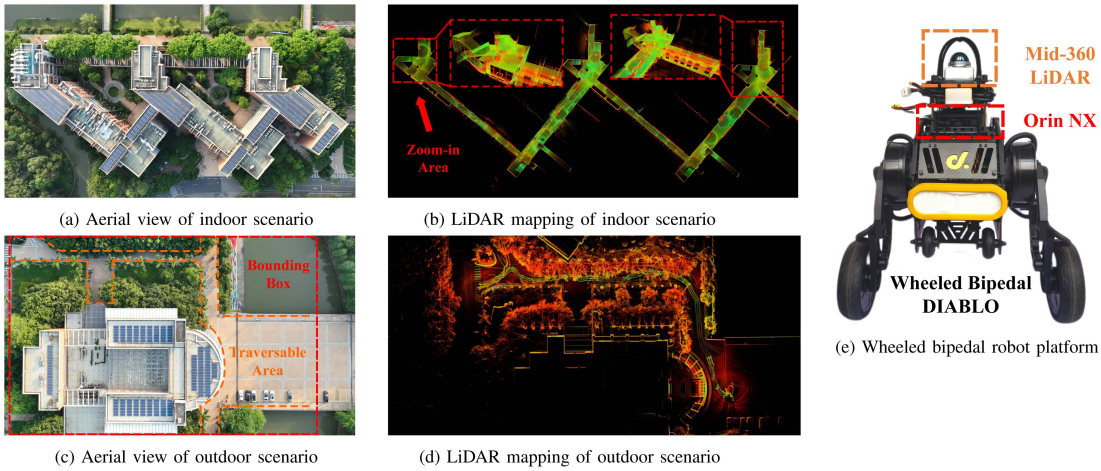
To evaluate the role of global attention, we compare three architectural variants, as shown in Fig. 4: (1) the Proposed model, which integrates linear global attention and local GCN-based message passing; (2) the GCN-only variant, which removes global attention; and (3) the Multi-Head Attention (MHA)-only variant, which replaces both modules with standard MHA constrained by graph-based masking. Experimental results indicate that the proposed hybrid architecture achieves the best balance among convergence speed, exploration efficiency, and consistency. The GCN-only variant struggles with large environments due to its limited ability to capture global dependencies. Although the MHA-only variant effectively models global context, its computational complexity significantly increases memory usage and slows training, reducing practicality.

### E. Ablation Study of Heterogeneous Graph

To evaluate the role of heterogeneous graph abstraction, we compare our full model with a homogeneous variant where all nodes share the same minimal attribute vector $[x, y, \epsilon]$, where $(x, y)$ is the spatial coordinate and $\epsilon$ quantifies potential exploration gain. In contrast, the heterogeneous graph assigns type-specific attributes to frontier, border, skeleton, and viewpoint nodes, preserving their distinct semantics. As shown in Fig. 6, the heterogeneous variant achieves consistently higher rewards and shorter travel distances after convergence, while maintaining stable training. The homogeneous variant still learns reasonable policies, but suffers from less efficient exploration due to the loss of structural priors. These results confirm that the heterogeneous abstraction is not only semantically interpretable and computationally lightweight, but also brings tangible improvements in both exploration efficiency and policy performance.

### F. Ablation Study of Viewpoint-Based Action Mask

To evaluate the effect of the viewpoint-based action mask, we compare our full model against a baseline where the mask is removed and all nodes in the heterogeneous graph are allowed as candidate actions. Experimental results are shown in Figs. 7 and 8. When the mask is removed, the agent initially achieves a higher cumulative reward and shorter travel distance since it learns to minimize movement cost. However, this behavior is




(a) Aerial view of indoor scenario (b) LiDAR mapping of indoor scenario

(c) Aerial view of outdoor scenario (d) LiDAR mapping of outdoor scenario

(e) Wheeled bipedal robot platform

Fig. 5. Overview of real-world experimental setups and exploration results. (a, c) Aerial views of the two test scenarios, with (c) showing the manually defined bounding box (red dashed outline). (b, d) Corresponding LiDAR-generated pointcloud maps; the red dashed boxes in (b) indicate zoom-in regions for clarity. (e) The wheeled bipedal robot platform equipped with Mid-360 LiDAR and Orin NX computing module.

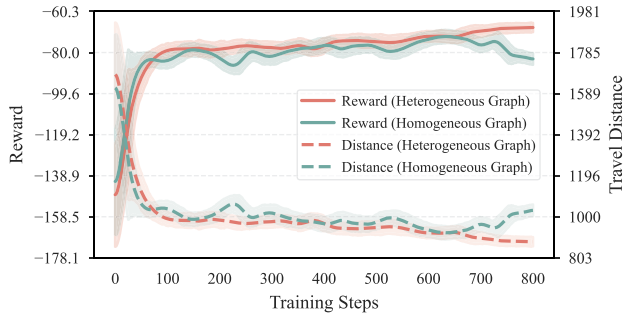



Fig. 6. Ablation study of heterogeneous graph. Comparison of heterogeneous and homogeneous graph representations in terms of reward and travel distance.

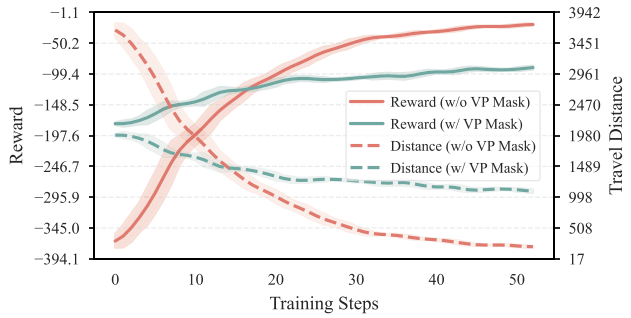



Fig. 7. Ablation study of viewpoint-based action mask. With viewpoint masking, the policy achieves stable rewards; without masking, training converges to a degenerate strategy with high rewards but poor coverage.

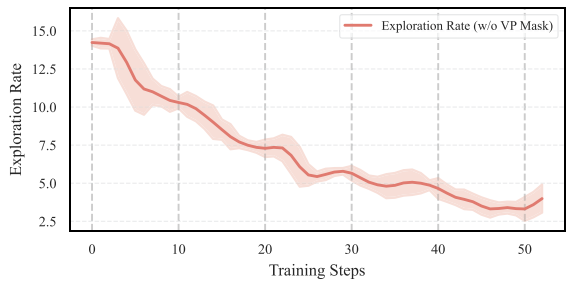



Fig. 8. Ablation study of viewpoint-based action mask. The exploration rate rapidly declines without masking, indicating that the agent learns to remain idle instead of exploring the environment.

degenerate because the exploration rate steadily decreases and eventually approaches zero, meaning that the agent prefers to stay idle or oscillate locally rather than covering new areas. As a result, the environment is not sufficiently explored, and task completion cannot be guaranteed. In contrast, the masked formulation maintains stable reward, reduced travel distance, and a consistently positive exploration rate, showing that it balances reward optimization with coverage. These results indicate that viewpoint-based action masking is not a minor heuristic but a key mechanism that prevents collapse into trivial strategies and ensures robust exploration performance.

### G. Real-World Experiments

To validate the effectiveness and robustness of our proposed method in realistic scenarios, we conducted extensive field experiments in both indoor (Fig. 5(a)) and outdoor (Fig. 5(c)) environments using the wheeled bipedal robot DIABLO, equipped with a Mid-360 LiDAR and an Orin NX onboard computer (Fig. 5(e)). Remarkably, the policy trained in simulation was directly deployed in these real-world scenarios in a zero-shot manner, without any fine-tuning or adaptation. All calculations were performed onboard, with dynamics constraints set to $v_{max} = 1.0\,\mathrm{m/s}$ and $a_{max} = 0.8\,\mathrm{m/s}$.

In the indoor scenario, we explored a large and complex floor that spans three interconnected buildings, covering approximately $2412.75\,\mathrm{m}^2$. The exploration terminated once no frontiers or candidate viewpoints remained in the constructed topological graph, which naturally occurred since the environment is a closed structure. The robot started from the edge of the unexplored area without prior environmental knowledge, and the exploration concluded in $629\,\mathrm{s}$, traversing a trajectory of $514.72\,\mathrm{m}$. Fig. 5(b) shows the final generated map and exploration paths, demonstrating the effectiveness of our method in comprehensively covering corridors, rooms, and cluttered areas.

The outdoor scenario covered a significantly larger test area of approximately $9846.50\,\mathrm{m}^2$, featuring typical campus elements such as open areas, walkways, and structured buildings. Since the environment is open and unbounded, we defined a bounding

box in the robot's initial coordinate frame by specifying fixed ranges along the $x$ and $y$ axes (highlighted by the red dashed outline in Fig. 5(c)) to ensure a well-posed stopping condition. The robot continued exploration within this bounded region until no new viewpoints were generated, which resulted in completion after $629\,\mathrm{s}$ and a trajectory of $351.93\,\mathrm{m}$. Fig. 5(d) illustrates accurate and comprehensive maps generated by our method, validating reliable performance under challenging outdoor conditions.

Overall, these real-world experiments highlight the scalability, efficiency, and zero-shot performance of our proposed method under realistic and complex environments.

## VI. Conclusion

We proposed a scalable graph-based exploration framework that integrates heterogeneous topological representations, lightweight global-local reasoning, and reinforcement learning for large-scale autonomous exploration. The heterogeneous graph encodes key spatial semantics in a compact yet expressive form, while the combination of linear global attention and local graph convolutions enhances inference efficiency and scalability. A minimalist reward with curriculum learning mitigates sparse-reward challenges and improves policy generalization, and viewpoint-based action masking ensures robust and efficient decision-making. Extensive simulations and real-world experiments demonstrate that our method consistently outperforms state-of-the-art approaches in efficiency, stability, and generalization. Nevertheless, several limitations remain and should be addressed in future work. First, the framework has not yet been evaluated in highly dynamic environments, where viewpoint estimation may fail; addressing this issue could further enhance the framework's uncertainty-aware capabilities. Furthermore, the current graph representation is restricted to 2D grids and ESDF skeletons. Extending it to 3D structures with richer features will broaden the framework's applicability. Finally, the existing policy selects only the next waypoint, whereas autoregressive prediction of waypoint sequences would facilitate a tighter integration between global and local planning.

## References


[1] A. Romero, C. Delgado, L. Zanzi, R. Suárez, and X. Costa-Pérez, "Cellular-enabled collaborative robots planning and operations for search-and-rescue scenarios," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2024, pp. 5942–5948.

[2] S. Kim et al., "Gas source localization in unknown indoor environments using dual-mode information-theoretic search," *IEEE Robot. Automat. Lett.*, vol. 10, no. 1, pp. 588–595, Jan. 2025.

[3] B. Yamauchi, "Frontier-based exploration using multiple robots," *ACM SIGGRAPH*, vol. 97, pp. 47–53, 1997.

[4] C. Cao, H. Zhu, H. Choset, and J. Zhang, "TARE: A hierarchical framework for efficiently exploring complex 3D environments," in *Proc. Robot.: Sci. Syst.*, vol. 5, 2021, p. 2.

[5] J. Huang et al., "FAEL: Fast autonomous exploration for large-scale environments with a mobile robot," *IEEE Robot. Automat. Lett.*, vol. 8, no. 3, pp. 1667–1674, Mar. 2023.

[6] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural SLAM," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 2193–2210.

[7] Y. Cao, R. Zhao, Y. Wang, B. Xiang, and G. Sartoretti, "Deep reinforcement learning-based large-scale robot exploration," *IEEE Robot. Automat. Lett.*, vol. 9, no. 5, pp. 4631–4638, May 2024.

[8] Z. Li, Y. Yang, and H. Cheng, "Efficient multi-agent cooperation: Scalable reinforcement learning with heterogeneous graph networks and limited communication," *Knowl.-Based Syst.*, vol. 300, 2024, Art. no. 112124.

[9] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *IEEE Comput.*, vol. 22, no. 6, pp. 46–57, Jun. 1989.

[10] H. Oleynikova, Z. Taylor, M. Fehr, R. Siegwart, and J. Nieto, "Voxblox: Incremental 3D euclidean signed distance fields for on-board MAV planning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 1366–1373.

[11] T. Noël, S. Kabbour, A. Lehuger, E. Marchand, and F. Chaumette, "Disk-graph probabilistic roadmap: Biased distance sampling for path planning in a partially unknown environment," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 5707–5714.

[12] T. Noël, A. Lehuger, E. Marchand, and F. Chaumette, "Skeleton disk-graph roadmap: A sparse deterministic roadmap for safe 2D navigation and exploration," *IEEE Robot. Automat. Lett.*, vol. 9, no. 1, pp. 555–562, Jan. 2024.

[13] B. Zhou, Y. Zhang, X. Chen, and S. Shen, "FUEL: Fast UAV exploration using incremental frontier structure and hierarchical planning," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 779–786, Apr. 2021.

[14] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 2713–2726.

[15] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 2920–2931.

[16] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim, "Graph transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 11983–11993.

[17] U. Alon and E. Yahav, "On the bottleneck of graph neural networks and its practical implications," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 14048–14063.

[18] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5156–5165.

[19] Q. Wu et al., "SGFormer: Simplifying and empowering transformers for large-graph representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 64753–64773.

[20] J. Yu, H. Shen, J. Xu, and T. Zhang, "ECHO: An efficient heuristic viewpoint determination method on frontier-based autonomous exploration for quadrotors," *IEEE Robot. Automat. Lett.*, vol. 8, no. 8, pp. 5047–5054, Aug. 2023.

[21] R. S. Sutton et al., *Reinforcement Learning: An Introduction*, vol. 1, no. 1. Cambridge, MA, USA: MIT Press, 1998.

[22] A. Ecoffet, J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune, "Go-explore: A new approach for hard-exploration problems," 2019, *arXiv:1901.10995*.

[23] A. Jonnarth, J. Zhao, and M. Felsberg, "Learning coverage paths in unknown environments with deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 22491–22508.

[24] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 16–17.

[25] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Sci. Robot.*, vol. 5, no. 47, 2020, Art. no. eabc5986.

[26] L. Weng, "Reward hacking in reinforcement learning," 2024. [Online]. Available: https://lilianweng.github.io/posts/2024-11-28-reward-hacking/

[27] S. Narvekar, B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor, and P. Stone, "Curriculum learning for reinforcement learning domains: A framework and survey," *J. Mach. Learn. Res.*, vol. 21, no. 181, pp. 1–50, 2020.

[28] O. Devillers, S. Hornus, and C. Jamin, "dD triangulations," in *CGAL User and Reference Manual*, CGAL Editorial Board, 2024. [Online]. Available: https://doc.cgal.org/6.0.1/Manual/packages.html#PkgTriangulations

[29] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.

[30] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[31] S. Huang et al., "CleanRL: High-quality single-file implementations of deep reinforcement learning algorithms," *J. Mach. Learn. Res.*, vol. 23, no. 274, pp. 1–18, 2022.

[32] F. Kong et al., "MARSIM: A light-weight point-realistic simulator for LiDAR-based UAVs," *IEEE Robot. Automat. Lett.*, vol. 8, no. 5, pp. 2954–2961, May 2023.