Planner-Conditioned Diffusion for Coordinated Multi-Agent Exploration

Marcus Yu Siong Teo*, Jeric Lew*, Tanishq Duhan, and Guillaume Sartoretti

Department of Mechanical Engineering, National University of Singapore, Singapore
{e0958027,jericlew,e1280621}@u.nus.edu, mpegas@nus.edu.sg

Abstract. Coordinated multi-agent exploration requires not only efficient individual coverage but also non-redundant coverage across agents over extended planning horizons. Conventional approaches rely on short-horizon planning or hand-crafted coordination rules, while end-to-end multi-agent learning methods are difficult to scale and train. Diffusion-based planners such as DARE offer a promising alternative by generating long-horizon trajectories instead of single-step actions, but existing methods are trained on a narrow planner distribution, limiting behavioral diversity and inference-time controllability. We propose a Planner-Conditioned Diffusion Policy (PCDP) for graph-based multi-agent exploration. PCDP is trained on demonstrations from multiple planner styles with planner identity as an explicit one-hot conditioning input, enabling a single shared model to learn a multimodal trajectory distribution and produce diverse, controllable proposals from the same observation. Rather than learning coordination end-to-end, we reuse this multimodal single-agent policy across all agents and introduce coordination through local communication-aware reranking, selecting trajectory combinations with minimal predicted overlap. We evaluate PCDP against against both diffusion-based and classical multi-agent exploration baselines under a common 4-agent local-reranking framework over 100 paired test episodes. PCDP maintains a perfect success rate of the diffusion-based baselines while improving mean max-agent travel by 1.94%, mean total team travel by 1.58%, and mean agent imbalance by 5.66%. Crucially, reranking alone over a single-planner baseline yields only marginal gains, confirming that planner-conditioned multimodality is the primary driver of improved coordination. Qualitative simulation results and real-robot experiments with two agents further validate that diverse long-horizon trajectory generation produces emergent spatial separation between agents without any explicit repulsion mechanism.

Keywords: Multi-agent systems · Robot exploration · Diffusion Policy

1 Introduction

Autonomous exploration is a well research problem in robotics, with applications in search and rescue, infrastructure inspection, and mapping of unknown environments. The objective is to uncover unseen areas while maximizing efficiency

* The first two authors contributed equally to this work.

in terms of distance traveled or time taken. This requires reasoning over partial observations and making sequential decisions under uncertainty, which makes exploration inherently challenging even in the single-agent setting [4,16].

A natural extension is to deploy multiple agents to accelerate coverage in large or time-critical environments. However, multi-agent exploration introduces additional complexity: agents must not only explore efficiently individually, but also coordinate to avoid redundant coverage and ensure effective spatial distribution [2,17]. The quality of the team's performance depends on how well agents' behaviors complement each other over time.

Conventional exploration methods, including frontier-based and information-theoretic approaches [1,3,9,16], rely on hand-crafted heuristics and typically operate over short planning horizons. While efficient, they primarily optimize immediate utility and often fail to capture the long-term horizon of exploration. In multi-agent settings, this limitation is exacerbated, as coordination depends on how agents' future trajectories interact rather than just their next actions.

Learning-based methods offer a data-driven alternative, but many still focus on next-action prediction or local goal selection [4,6], limiting their ability to model long-horizon behavior. End-to-end multi-agent learning approaches attempt [18,20] to address coordination directly, but are often computationally expensive, difficult to scale, and sensitive to training instability. Diffusion-based [7,10,12] planners for exploration, such as DARE [5], provide a promising direction by generating long-horizon trajectories from data, enabling richer reasoning over future motion. However, existing diffusion-based exploration methods are typically trained on a single planner distribution, resulting in limited behavioral diversity and reduced controllability at inference time.

In this work, we aim to leverage long-horizon diffusion-based planning for multi-agent exploration without requiring end-to-end multi-agent training. We propose a Planner-Conditioned Diffusion Policy (PCDP) that is trained on trajectories from multiple planner types, with planner identity provided as an explicit conditioning variable. This enables the model to learn a multimodal distribution over exploration behaviors and to generate diverse trajectory proposals under different planner conditions. At inference time, we reuse this shared single-agent policy across all agents and introduce coordination through a lightweight, communication-aware reranking procedure that selects complementary trajectories with minimal overlap.

Our key insight is that adding structured diversity and to long-horizon trajectory proposals directly improves multi-agent coordination under a fixed selection mechanism. We evaluate the proposed method in a 4-agent simulation setting and show that planner-conditioned diffusion consistently improves travel efficiency and agent balance compared to single-planner baselines while maintaining perfect success rates. We further validate the approach on a real two-agent robot platform, demonstrating that the coordination behaviour transfers to physical deployment. The contributions of this work are: (1) a planner-conditioned diffusion policy that learns a controllable, multimodal trajectory distribution for exploration; (2) a modular multi-agent framework that achieves coordination

via a shared trajectory generator reused across agents with local inference-time reranking; and (3) empirical evidence that enhanced trajectory diversity improves coordination efficiency without end-to-end multi-agent training, validated in both simulation and real-robot experiments.

2 Related Work

Classical exploration has long been studied through frontier-based and information-driven formulations. Frontier-based exploration remains a foundational approach for directing an agent toward the boundary between known and unknown space [16]. Subsequent work has extended this line with stronger heuristics, information-gain objectives, and next-best-view style criteria [1,3,9], with TARE in particular representing a strong hierarchical classical planner that serves as a key demonstration source in our dataset. In multi-agent settings, these approaches are often combined with explicit coordination logic to reduce redundant exploration and improve coverage efficiency [2,17]. While effective in many practical scenarios, such methods typically rely on hand-crafted objectives and short-horizon decision rules.

Learning-based exploration methods [4,6,14,15,19] attempt to overcome these limitations by learning policies directly from data or interaction. End-to-end multi-agent learning approaches [20] address coordination directly by jointly training agents, but are often computationally expensive, difficult to scale, and sensitive to parameter tuning. More recent work has also explored topological and active mapping formulations for multi-agent exploration [13,18]. However, many of these methods still operate at the level of action prediction or local decision-making, rather than directly modeling a distribution over future trajectories. In contrast, our approach shares a single-agent policy across all agents and introduces coordination entirely at inference time, avoiding joint training while placing greater demand on the quality and diversity of trajectory proposals.

Diffusion models have recently emerged as a practical framework for trajectory generation in robotics. Diffuser and Diffusion Policy showed that diffusion models can represent multimodal trajectory distributions and support long-horizon planning through iterative denoising [7,12]. DARE extended this idea to autonomous exploration by learning long-term exploration trajectories from planner demonstrations [5]. Our work is most closely related to DARE in its use of diffusion for exploration, but differs in two key ways. First, we explicitly train on a planner-diverse dataset rather than a narrow planner set. Second, we treat planner identity as a controllable conditioning signal and leverage the resulting multimodal trajectory proposals directly for coordinated multi-agent planning through inference-time reranking.

3 Background

3.1 Multi-Agent Exploration

We consider a coordinated multi-agent exploration setting in which n agents operate in a bounded, initially unknown environment represented as a 2D occupancy grid map $\mathcal{M}$. The map is partitioned into free area $\mathcal{M}_f$, occupied area $\mathcal{M}_o$, and unknown area $\mathcal{M}_u$, such that $\mathcal{M}_f \cup \mathcal{M}_o \cup \mathcal{M}_u = \mathcal{M}$.

Agents operate under global communication: as agents navigate and observe the environment, all sensor readings are immediately reflected in a single shared map $\mathcal{M}_k = \mathcal{M}_f \cup \mathcal{M}_o$, which all agents plan from at every decision step. The environment is considered fully explored when $\mathcal{M}_u = \emptyset$, i.e., when no unknown area remains.

At each decision step t, each agent i receives a graph-based observation $o_t^{(i)}$ constructed from the full shared map following the graph representation of DARE [5]. The graph encodes the explored space as a set of nodes, where each node stores its position relative to the agent's current location (x, y), a utility value corresponding to the number of visible frontier cells reachable from that node, and a binary signal that activates for nodes lying on paths toward nodes with non-zero utility. This representation gives the agent a structured, global view of where unexplored area remains and how to navigate toward it.

Each agent selects and executes one action $a_t^{(i)}$ at each time step. The team objective is to minimize the maximum agent path length required to complete exploration:

$$\Psi^* = \arg\min_{\Psi} \max_{i \in [1,n]} L\Big(\psi^{(i)}\Big), \quad \text{s.t.} \quad \mathcal{M}_u = \emptyset, \tag{1}$$

where $L(\psi^{(i)})$ denotes the total path length of agent i and $\Psi = \{\psi^{(1)}, \ldots, \psi^{(n)}\}$ is the set of agent trajectories.

3.2 Diffusion-Based Exploration Planning

Diffusion models [10] learn to generate samples from a target distribution by reversing a forward noising process. Starting from a clean sample x_0, the forward process progressively corrupts it with Gaussian noise over S steps. The model is trained to reverse this process by recovering x_0 from a noisy observation, conditioned on any available context. At inference time, a clean sample is obtained by iteratively denoising from a Gaussian prior, which naturally produces diverse outputs across repeated samples.

DARE [5] applies this framework to single-agent exploration planning. A conditional diffusion policy $p_\theta(\tau_t \mid o_t)$ is trained on demonstrations from a ground-truth planner with full map access, learning to generate long-horizon trajectories over a planning horizon H from graph-based observations:

$$\tau_t = (a_t,\, a_{t+1},\, \ldots,\, a_{t+H-1})\,. \tag{2}$$

During execution, DARE follows a receding horizon strategy: a full trajectory is generated at each step, but only the first action is executed before replanning

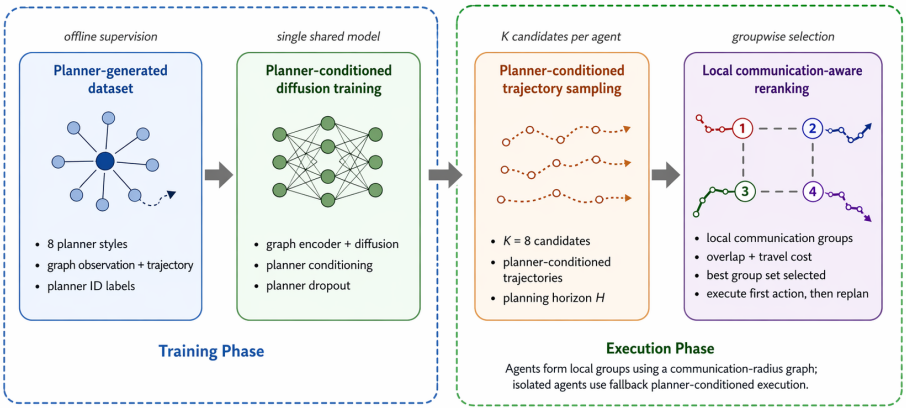


Fig. 1. Overview of the proposed PCDP framework. A planner-diverse dataset is constructed from multiple planner styles. A shared diffusion policy is trained conditioned on graph observation and planner identity. During execution, each agent samples multiple candidate trajectories, and a local reranking stage selects the most complementary combination for coordinated exploration.

with the updated map. This couples long-horizon trajectory generation with reactive, step-by-step execution, enabling the policy to reason over extended future motion while remaining responsive to newly observed information.

Because the denoising process is stochastic, repeated sampling from the same observation produces a diverse set of trajectory proposals, each representing a plausible future path. This property makes diffusion policies naturally suited for candidate-based planning, where multiple trajectories are generated and the best is selected. A key limitation of single-planner training, however, is that the learned distribution concentrates around one dominant behaviour mode, reducing candidate diversity and limiting how effectively a downstream selection mechanism can differentiate agent assignments.

4 Method

The proposed framework consists of three components: planner-diverse dataset construction, planner-conditioned diffusion policy training, and inference-time local reranking for multi-agent coordination. The learning problem is kept at the single-agent level throughout; coordinated behaviour emerges from trajectory selection rather than joint training.

4.1 Planner-Diverse Demonstration Dataset

To broaden the behavioral diversity of the trajectory generator, we construct a dataset from multiple planner families, including heuristic frontier planners,

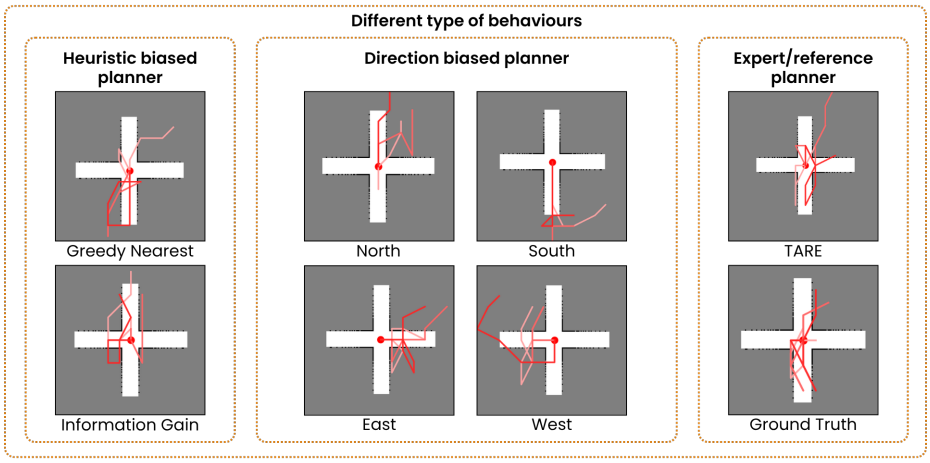


Fig. 2. Trajectory diversity under planner conditioning. From similar graph observations, different planner conditions produce distinct trajectory modes, illustrating that the learned policy captures a multimodal exploration distribution rather than a single deterministic behaviour.

directional-bias planners, information-gain planners, and the expert-style TARE planner [3]. In the final reported setup, demonstrations are collected from eight planner classes as shown in Figure 2.

Each training sample takes the form (o_t, τ_t, c_t), where o_t is the graph-based observation at decision time t, τ_t is the trajectory produced by the selected planner, and c_t is the associated planner label. The purpose of planner diversity is not merely to increase sample count, but to expose the model to structurally distinct exploration behaviors. Planner identity therefore acts as a structured source of behavioral variation rather than generic data augmentation. The baseline diffusion model follows the narrower DARE-style training setup [5], using demonstrations from a single ground-truth planner.

4.2 Planner-Conditioned Diffusion Policy

We extend the DARE baseline by introducing planner identity as an explicit conditioning variable. Rather than learning a single conditional policy $p_\theta(\tau_t \mid o_t)$, we define a planner-conditioned policy:

$$p_\theta(\tau_t \mid o_t, c_t), \tag{3}$$

where c_t is drawn from a finite set of planner identities. Here o_t describes the current exploration state from the shared map, while c_t specifies the intended exploration style. This enables a single shared model to represent a multimodal trajectory distribution, producing qualitatively different trajectories from the same observation under different planner conditions.

The graph-based observation o_t is encoded into a latent feature z_t via an attention-based graph encoder following the architecture of DARE [5] and Ariadne [4]. The planner identity c_t is represented as a one-hot vector over the set of M planner classes and projected to a learnable embedding before being fused with z_t as input to the diffusion model. The graph observation therefore provides state-dependent context capturing frontier structure and connectivity, while the planner embedding provides style-dependent context that steers the denoising process toward a particular trajectory mode. The same observation can yield qualitatively different trajectories under different planner conditions, which is the key mechanism through which the policy captures multimodality within a single shared model.

Diffusion training objective. Let x_0 denote the encoded clean trajectory, x_s its noisy version at diffusion step s, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ the injected noise. The model is trained to predict the noise residual:

$$\mathcal{L}_\epsilon = \mathbb{E}_{x_0,\, \epsilon,\, s}\left[\left\|\epsilon - \epsilon_\theta(x_s,\, s \mid z_t,\, c_t)\right\|^2\right]. \tag{4}$$

Planner dropout. To prevent over-reliance on the planner label and encourage robustness when planner identity is unavailable, planner conditioning is randomly removed for a fraction of training samples by substituting the true label with a null condition. This regularization encourages the model to learn both planner-aware specialization and planner-agnostic exploration behavior, mirroring the classifier-free guidance technique used in conditional diffusion models [11].

4.3 Multi-Agent Execution and Local Reranking

At inference time, the trained single-agent policy is shared across all n agents. All agents operate on the same globally shared map, and each agent independently forms its graph observation from this shared belief before sampling candidate trajectories.

Candidate trajectory generation. For **PCDP**, at each planning step every agent samples K trajectories per planner condition, yielding $K \times M$ candidates in total, where M is the number of planner classes. In the reported experiments, $K = 8$. Although the policy parameters are shared, trajectories differ across agents because each agent conditions on its own graph observation and its own stochastic diffusion samples. Each candidate trajectory is converted into a set of predicted traversed map cells and an estimated travel cost computed as the cumulative Euclidean length of the predicted path.

Communication groups. Although the map is shared globally, coordination during reranking is performed locally. At each planning step, robots are partitioned into local coordination groups via a proximity graph with communication radius $R_{\text{comm}} = 25\,\text{m}$. Two robots are connected if their positions lie within this radius; connected components of this graph define the coordination groups.

Local reranking. For a coordination group g, let $\mathcal{C}_g$ denote the set of all joint candidate combinations formed from the $K \times M$ trajectories sampled by each robot in the group. The selected combination minimises a local coordination score:

$$\begin{aligned} \hat{\mathcal{C}}_g &= \arg\min_{\mathcal{C} \in \mathcal{C}_g} S(\mathcal{C}), \\ S(\mathcal{C}) &= \lambda_{\text{overlap}}\, O(\mathcal{C}) + \lambda_{\text{travel}}\, D(\mathcal{C}), \end{aligned} \tag{5}$$

where $O(\mathcal{C})$ is the total predicted overlap in traversed map cells across the group's trajectories and $D(\mathcal{C})$ is the sum of their travel costs; in simple terms, the robots choose the combination that works best together by reducing overlap while keeping travel cost low. In the reported experiments, $\lambda_{\text{overlap}} = 1.0$ and $\lambda_{\text{travel}} = 0.02$.

Isolated robots. If a robot's communication group contains only itself, reranking is not applied. That robot instead selects the best single candidate from its own $K \times M$ samples, falling back to independent planner-guided behaviour.

Execution. Once a trajectory is selected for each robot, all agents follow the receding horizon strategy of DARE: only the first action of the chosen trajectory is executed before the shared map is updated and replanning occurs. This preserves long-horizon reasoning while keeping the system reactive to newly observed information.

The coordination layer is deliberately lightweight: the primary source of coordination benefit is the quality and diversity of the $K \times M$ candidate set generated by the planner-conditioned policy, while reranking is the mechanism that exploits this diversity at execution time.

5 Results and Discussion

We evaluate the proposed framework in a 4-agent simulation setting in which all agents share the same trained single-agent trajectory generator. Five completed methods are compared: two classical baselines, **Nearest** and **NBVP**, two diffusion-only baselines, **DARE + independent** and **DARE + local reranking**, and the proposed **PCDP + local reranking**. This comparison serves two purposes. First, the two DARE variants isolate the benefit of local reranking alone (DARE + local reranking vs. DARE + independent) and the benefit of planner conditioning under the same reranking rule (PCDP vs. DARE, both with local reranking). Second, the classical baselines provide broader context on whether the proposed method remains competitive beyond diffusion-only comparisons.The classical baselines are evaluated in their native multi-agent execution form on the same map instances, while the diffusion-based methods are evaluated under the shared local-reranking framework described in Section 4.3.

Evaluation is carried out over 100 paired test episodes using identical map instances across all completed methods. The reported experiments use an observation horizon of 2 steps and a trajectory horizon of 8 steps, with DDPM inference using 100 denoising steps.

5.1 Metrics

We report five metrics. **Explored rate** measures the fraction of the environment's free space that has been discovered by the team at the end of the episode. This metric is used to distinguish true efficiency from apparently low travel caused by under-exploration. **Success rate** is the fraction of episodes completed. An episode is counted as successful if the team fully explores the map within the allotted episode horizon. **Max-agent travel** $D_{\max} = \max_i d_T^{(i)}$ captures the largest individual burden within the team. **Total team travel** $D_{\text{sum}} = \sum_i d_T^{(i)}$ measures aggregate motion cost. **Agent imbalance** $D_{\text{imbalance}} = \max_i d_T^{(i)} - \min_i d_T^{(i)}$ reflects how evenly work is distributed; lower values indicate better-balanced coordination.

5.2 Real-Robot Experiments

To validate our approach beyond simulation, we conduct real-robot experiments using a two-agent platform (see Figure 3). We select mecanum-wheeled robots for the demonstration and use the OptiTrack motion capture system for accurate localization. Plan execution follows a P3GASUS [8]-based centralized planning, decentralized execution framework, with RViz used for visualization. The full video of the experiment is included in the supplementary material.

5.3 Quantitative Comparison

Table 1 summarises the 100-episode comparison across all completed methods. **PCDP + local reranking** is the strongest overall method. It preserves the same perfect success and exploration regime as the DARE baselines while also achieving the lowest mean max-agent travel, lowest mean total team travel, and lowest mean agent imbalance. Relative to the classical baselines, it also shows a substantially better efficiency–reliability trade-off: **Nearest** reaches reasonably high exploration but at much higher travel cost, while **NBVP** exhibits poor completion consistency and very high imbalance.

Table 1. Quantitative comparison over 100 paired episodes. Lower is better for $D_{\max}$, D_{sum}, and $D_{\text{imbalance}}$.

Method	Success	Explored	Mean $D_{\max}$	Mean D_{sum}	Mean $D_{\text{imbalance}}$
Nearest	88%	0.9618	982.04	3637.21	158.43
NBVP	51%	0.9554	1109.87	3188.08	685.94
DARE + independent	100%	1.0000	664.33	2559.38	49.78
DARE + local reranking	100%	1.0000	662.02	2542.91	51.37
PCDP + local reranking	**100%**	**1.0000**	**649.18**	**2502.85**	**48.46**

5.4 Planner Conditioning Is the Primary Source of Gain

Relative to the strongest classical baseline, **Nearest**, **PCDP + local reranking** achieves a substantially better efficiency–reliability trade-off. It improves success from 88% to 100%, increases mean explored rate from 0.9618 to 1.0000, and reduces mean max-agent travel by 33.9% (982.04 → 649.18), mean total team travel by 31.2% (3637.21 → 2502.85), and mean agent imbalance by 69.4% (158.43 → 48.46). This shows that the proposed method does not merely maintain coverage, but does so with substantially lower travel cost and better-balanced coordination.

Against the diffusion-based baselines, the gain can be isolated more precisely. Compared with **DARE + local reranking**, **PCDP + local reranking** reduces mean max-agent travel by 1.94% (662.02 → 649.18), total team travel by 1.58% (2542.91 → 2502.85), and agent imbalance by 5.66% (51.37 → 48.46). By contrast, the comparison between **DARE + local reranking** and **DARE + independent** shows that reranking alone yields only marginal gains: 0.35% on max-agent travel and 0.64% on total team travel, with imbalance slightly worsening. This shows that planner-conditioned multimodality, rather than reranking alone, is the primary source of improved coordination.

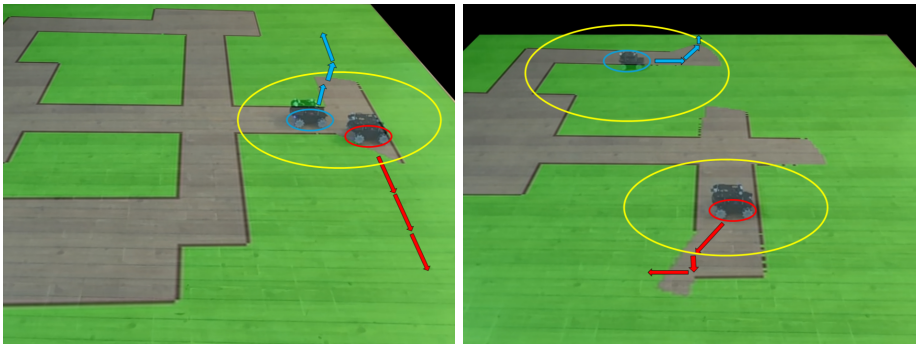
The gains are distributed across the team rather than concentrated in a single agent: mean travel distance is lower for all four agents under **PCDP + local reranking** than under both diffusion baselines. At the episode level, the proposed method achieves lower max-agent travel in 63 of 100 episodes against **DARE + local reranking** (59 against **DARE + independent**), and lower total team travel in 54 of 100 episodes (60 against independent). It ranks best on max-agent travel in 45 episodes and worst in only 23, confirming a consistent reduction in poor coordination outcomes rather than just an average shift.

5.5 Qualitative Results

Figure 2 shows that different planner conditions produce structurally distinct path proposals from similar observations, providing qualitative support for the multimodality claim and illustrating how diverse candidates arise from the same shared policy. Notably, trajectories generated under different conditions from the same graph observation diverge in both direction and extent, covering different regions of the frontier. This behavioural spread is what single-planner policies lack: they tend to cluster around the same dominant mode, leaving little room for the coordination mechanism to differentiate agent assignments.

5.6 Real-Robot Experiments

To further validate the approach, we deploy the framework on a two-agent real-robot platform. Two complementary behaviours are observed, as illustrated in Figure 3. When the agents are within communication range, the local reranking stage consistently selects trajectories that diverge toward different frontiers, actively reducing redundant coverage. When an agent is isolated, it falls back



(a) Within Reranking Range (b) Out of Reranking Range

Fig. 3. Real-robot two-agent exploration under the proposed framework.

to independently selecting its best planner-conditioned trajectory, continuing to explore efficiently on its own. Together, these observations confirm that planner-conditioned diversity is the key enabler: without it, agents operating from similar observations would concentrate probability on the same dominant trajectory mode and are likely to pursue redundant paths. The framework therefore produces emergent spatial separation when agents are nearby, and effective independent exploration when they are not, without requiring any explicit repulsion or role-assignment mechanism.

6 Conclusion

In this work we presented Planner-Conditioned Diffusion Policy (PCDP) for coordinated multi-agent exploration. By training a shared single-agent diffusion policy on demonstrations from multiple planner styles with planner identity as an explicit conditioning input, PCDP learns a multimodal trajectory distribution that produces diverse, controllable proposals from the same observation. Applied in a modular multi-agent framework with local communication-aware reranking, the policy achieves coordination without end-to-end multi-agent training.

The central finding is that planner-conditioned multimodality, rather than the reranking mechanism alone, is the primary driver of improved coordination. Reranking over a single-planner candidate set yields only marginal gains, whereas the same reranking rule applied to PCDP's richer candidate set consistently reduces max-agent travel, total team travel, and agent imbalance while preserving perfect task completion. Real-robot experiments with two agents further confirm that the framework transfers to physical deployment, producing emergent spatial separation between agents without any explicit repulsion or role-assignment mechanism.

More broadly, this work demonstrates that improving the quality and diversity of the trajectory proposal mechanism is a practical and scalable alterna-

tive to end-to-end multi-agent coordination learning. Future work could explore learned or adaptive reranking objectives, extend the framework to larger teams, and investigate whether richer planner conditioning can further reduce the gap between single-agent training and multi-agent execution.

References

1. Bircher, A., Kamel, M., Alexis, K., Oleynikova, H., Siegwart, R.: Receding horizon" next-best-view" planner for 3d exploration. In: 2016 IEEE international conference on robotics and automation (ICRA). pp. 1462–1468. IEEE (2016)
2. Burgard, W., Moors, M., Stachniss, C., Schneider, F.E.: Coordinated multi-robot exploration. IEEE Transactions on robotics **21**(3), 376–386 (2005)
3. Cao, C., Zhu, H., Choset, H., Zhang, J.: Tare: A hierarchical framework for efficiently exploring complex 3d environments. In: Robotics: Science and Systems. vol. 5, p. 2 (2021)
4. Cao, Y., Hou, T., Wang, Y., Yi, X., Sartoretti, G.: Ariadne: A reinforcement learning approach using attention-based deep networks for exploration. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 10219–10225. IEEE (2023)
5. Cao, Y., Lew, J., Liang, J., Cheng, J., Sartoretti, G.: Dare: Diffusion policy for autonomous robot exploration. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 11987–11993 (2025). https://doi.org/10.1109/ICRA55743.2025.11128196
6. Cao, Y., Zhao, R., Wang, Y., Xiang, B., Sartoretti, G.: Deep reinforcement learning-based large-scale robot exploration. IEEE Robotics and Automation Letters (2024)
7. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research (2024)
8. Duhan, T., He, C., Sartoretti, G.: P3gasus: Pre-planned path execution graphs for multi-agent systems at ultra-large scale. IEEE Robotics and Automation Letters **11**(2), 1274–1281 (2025)
9. Gao, W., Booker, M., Adiwahono, A., Yuan, M., Wang, J., Yun, Y.W.: An improved frontier-based approach for autonomous exploration. In: 2018 15th international conference on control, automation, robotics and vision (ICARCV). pp. 292–297. IEEE (2018)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021), https://openreview.net/forum?id=qw8AKxfYbI
12. Janner, M., Du, Y., Tenenbaum, J., Levine, S.: Planning with diffusion for flexible behavior synthesis. In: International Conference on Machine Learning (2022)
13. Lee, E.S., Kim, Y.M.: Multi-agent exploration with similarity score map and topological memory. IEEE Robotics and Automation Letters **9**(11), 10327–10334 (2024)
14. Li, H., Zhang, Q., Zhao, D.: Deep reinforcement learning-based automatic exploration for navigation in unknown environment. IEEE Transactions on Neural Networks and Learning Systems **31**(6), 2064–2076 (2019)
15. Xu, Y., Yu, J., Tang, J., Qiu, J., Wang, J., Shen, Y., Wang, Y., Yang, H.: Explore-bench: Data sets, metrics and evaluations for frontier-based and deep-reinforcement-learning-based autonomous exploration. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 6225–6231. IEEE (2022)

16. Yamauchi, B.: 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'. pp. 146–151. IEEE (1997)
17. Yamauchi, B.: Frontier-based exploration using multiple robots. In: Proceedings of the second international conference on Autonomous agents. pp. 47–53 (1998)
18. Yang, X., Yang, Y., Yu, C., Chen, J., Yu, J., Ren, H., Yang, H., Wang, Y.: Active neural topological mapping for multi-agent exploration. IEEE Robotics and Automation Letters **9**(1), 303–310 (2023)
19. Zhu, D., Li, T., Ho, D., Wang, C., Meng, M.Q.H.: Deep reinforcement learning supervised autonomous exploration in office environments. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 7548–7555. IEEE (2018)
20. Zhu, S., Zhou, J., Chen, A., Bai, M., Chen, J., Xu, J.: Maexp: A generic platform for rl-based multi-agent exploration. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5155–5161. IEEE (2024)