frontiers | Frontiers in Robotics and AI

TYPE Original Research
PUBLISHED 25 March 2025
DOI 10.3389/frobt.2025.1492526

OPEN ACCESS

EDITED BY
Shude He,
Guangzhou University, China

REVIEWED BY
Ke Lu,
Guangxi University, China
Lin Chen,
Hunan University, China

*CORRESPONDENCE
Wencen Wu,
wencen.wu@sjsu.edu

RECEIVED 07 September 2024
ACCEPTED 17 February 2025
PUBLISHED 25 March 2025

CITATION
Lu T, Sobti D, Talwar D and Wu W (2025) Reinforcement learning-based dynamic field exploration and reconstruction using multi-robot systems for environmental monitoring.
Front. Robot. AI 12:1492526.
doi: 10.3389/frobt.2025.1492526

Reinforcement learning-based dynamic field exploration and reconstruction using multi-robot systems for environmental monitoring

Thinh Lu, Divyam Sobti, Deepak Talwar and Wencen Wu*

Computer Engineering Department, San Jose State University, San Jose, CA, United States

In the realm of real-time environmental monitoring and hazard detection, multi-robot systems present a promising solution for exploring and mapping dynamic fields, particularly in scenarios where human intervention poses safety risks. This research introduces a strategy for path planning and control of a group of mobile sensing robots to efficiently explore and reconstruct a dynamic field consisting of multiple non-overlapping diffusion sources. Our approach integrates a reinforcement learning-based path planning algorithm to guide the multi-robot formation in identifying diffusion sources, with a clustering-based method for destination selection once a new source is detected, to enhance coverage and accelerate exploration in unknown environments. Simulation results and real-world laboratory experiments demonstrate the effectiveness of our approach in exploring and reconstructing dynamic fields. This study advances the field of multi-robot systems in environmental monitoring and has practical implications for rescue missions and field explorations.

KEYWORDS

multi-robot systems, mobile sensor networks, reinforcement learning, dynamic field reconstruction, source seeking, environmental monitoring

1 Introduction

Environmental monitoring, including the identification and tracing of areas impacted by environmental hazards, is paramount for safeguarding human life and property. Early warning systems allow for swift responses to potential threats. Effective environmental monitoring relies on a deep understanding of key processes like wildfire propagation and pollutant dispersion. These phenomena often involve spatial and temporal changes, making them suitable for modeling using partial differential equations (PDEs). For instance, the advection-diffusion equation can be used to simulate the movement of smoke plumes from wildfires, providing crucial insights for predicting their evolution over time (Khaled et al., 2004; Reisch et al., 2024). This information is essential for effective environmental hazard monitoring and mitigation.

For environmental monitoring tasks, multi-robot systems offer significant advantages over single-robot setups by enabling faster coverage of larger areas and providing

redundancy against individual failures. These systems excel in complex missions across diverse environments, including search and rescue (Niroui et al., 2019; Shuvo et al., 2023; Cao et al., 2024), underwater surveillance (Martins et al., 2018; Luvisutto et al., 2022), and space exploration (Gautam et al., 2019; Bi et al., 2024; Long and Zhang, 2024). These works place additional emphasis on developing robust coordination strategies and efficient path-planning algorithms. Coordination may be centralized, with a leader directing actions, or decentralized, with robots making their own decisions based on local observations. Reinforcement learning has advanced these strategies, with actor–critic models enhancing control stability of the whole unit under dynamic disturbances (Hu et al., 2023) and graph-based methods enabling scalable, distributed decision-making across large robot teams (Chen et al., 2024). Depending on the mission, formation control may also play an essential role, where the robot system can be operated in organized patterns for high-quality data collection, or independently for greater flexibility. Beyond coordination, reinforcement learning-based approaches have also been increasingly adopted for path planning, further enhancing adaptability and performance of multi-robot systems in unknown environments (Zhu et al., 2023). These approaches require careful design of both the simulation environment and reward functions, which should closely model real-world conditions, to ensure effective learning and reliable performance in deployment. For applications in environmental monitoring, multi-robot systems can be equipped with specialized sensors to enable real-time data collection and reconstruction of environmental processes (Kinaneva et al., 2019; Dunbabin and Marques, 2012; Queralta et al., 2020; Rossi and Brunelli, 2016).

To reconstruct dynamic processes through limited measurements from multi-robot systems, it is necessary to identify unknown parameters in the PDEs that describe these processes, such as the diffusion coefficient in a diffusion equation. A common approach is to deploy static sensor networks (Mourikis and Roumeliotis, 2006; Burgard et al., 2005). Although effective, this approach is both costly and impractical for large-scale regions due to the need for extensive sensor installations. Mobile sensor networks, with collaborative mobile sensing robots, present a more practical alternative, offering great flexibility and broad coverage while using fewer sensors. In mobile sensor networks, parameter identification can be performed in two primary ways: offline and online (Zhang et al., 2023). Offline parameter identification requires mobile sensor networks to explore the entire spatial domain before any parameter estimation begins (Ucinski, 2005; Ucinski and Chen, 2005; Tricaud and Chen, 2010). This approach often uses techniques like least squares optimization to minimize the error between the observed and estimated states, typically requiring complex computations to solve PDEs. While this approach generally yields more accurate results, it is time-consuming, as full data collection must be completed before any estimation can take place. Due to the limitations of offline methods, increasing attention has shifted toward online parameter identification approaches (Wu et al., 2020; Zhang et al., 2023; Christopoulos and Roumeliotis, 2005). Online identification continuously updates parameter estimates as mobile sensors collect data in real time. While this approach may not provide the most accurate solution to PDEs compared to offline methods, it is far more efficient for time-sensitive applications like environmental hazard management (Zhang et al., 2023).

A key challenge of online parameter identification is determining an information-rich trajectory for the mobile sensing network, as this directly impacts the speed and accuracy of field reconstruction. However, since online methods operate in real-time, predicting the optimal path in advance is challenging, making efficient trajectory planning a complex problem. As a result, recent works in this field often provide additional strategies for effective trajectory planning and navigation for mobile sensor networks. In (You et al., 2016; Zhang et al., 2023), the authors employ a cooperative Kalman filter (CKF) combined with recursive least squares (RLS) to identify advection-diffusion field parameters in real-time using live sensor readings from a formation of mobile robots. To ensure that the robot formation follows information-rich trajectories, several studies, including (You and Wu, 2018; You et al., 2022), have integrated robot dynamics into the field dynamics and focused on minimizing mapping errors. However, these approaches may converge to local optima and may not adequately address the complexity of field reconstruction in environments involving multiple diffusion fields with varying characteristics. To address this issue, The author in (Talwar, 2020) proposes an exploration strategy that samples nearby candidate destinations based on custom weights calculated from cosine similarity to the centroid of unvisited regions and distance from explored diffusion fields. However, this approach may result in inefficient backtracking and revisits, which are undesirable in time-critical missions.

To tackle the problem of exploring complex dynamic fields, this research introduces a strategy for path planning and control of mobile sensing robots designed to effectively explore and reconstruct a dynamic field consisting of multiple non-overlapping diffusion fields while offering a good balance between speed and accuracy. In our proposed algorithm, the robot formation alternates between two primary operational modes: Field Exploration and Source Mapping. In Source Mapping mode, the formation makes use of reinforcement learning (RL), specifically, proximal policy optimization (PPO) to direct the robot formation to the center of a newly discovered diffusion field, while attempting to estimate its diffusion and advection coefficients through the CKF and RLS developed in (You et al., 2022). When dealing with the challenging problem that multiple sources exist in the field and the path planned in Source Mapping mode only leads to one source (local maximum) in the field, we develop a novel K-means clustering algorithm in the Field Exploration mode, to allow the robot formation advances toward unexplored regions to identify traces of potential new diffusion fields. The K-means clustering algorithm is used to partition the unexplored regions and facilitate faster scanning of the whole map. We validate our proposed strategy through both computer simulations and controlled laboratory experiments. In these scenarios, the robot formation is randomly placed within a spatially and temporally varying field, and we compare the field reconstruction errors to baseline strategies that employ random or lawn-mowing trajectories. Our research demonstrates the potential of multi-robot formations for accurate field reconstruction in complex environments characterized by multiple spatial-temporal diffusion fields.

To summarize, the main contributions of the paper are twofold: (1) it introduces a novel two-mode strategy for path planning and control of mobile sensing robots in dynamic environments, specifically for exploring and reconstructing fields with multiple

non-overlapping diffusion sources. The strategy integrates RL-based path planning with a CKF and RLS for estimating unknown parameters of the field. A key innovation is the use of K-means clustering algorithm to facilitate efficient exploration of unexplored regions, ensuring a balance between speed and accuracy. (2) Through both simulations and controlled experiments, the research demonstrates the effectiveness of the proposed strategy in improving field reconstruction accuracy using only a limited number of mobile sensing robots.

The remainder of this paper is structured as follows. In Section 2, we formally define the problem. We present some preliminary information in Section 3. The proposed algorithm is presented in Section 4, followed by a detailed analysis of the simulation and experimental results in Section 5 and Section 6 respectively. Finally, Section 7 summarizes our findings and outlines future research directions.

2 Problem formulation

In this section, we formulate the problem of reconstructing an unknown spatial-temporal varying field represented by a linear combination of several advection-diffusion equations, using a team of mobile sensing robots.

2.1 Spatial-temporal varying fields

Various processes that exhibit spatial and temporal variations, such as the dispersion of pollutants in the atmosphere or water bodies, are often represented by two-dimensional (2D) PDEs over a domain R. A typical example is the 2D advection-diffusion equation, which models the transfer of substances via advection (the movement of substances through a fluid) and diffusion (the spreading of substances from areas of higher to lower concentration). This can be mathematically expressed as:

$$\frac{\partial z}{\partial t}(r,t) = \theta \nabla^2 z(r,t) + \mathbf{v}^T \nabla z(r,t), \quad r \in R, \tag{1}$$

where $z(r,t)$ denotes the concentration function of the field at position r at time instance t, ∇ and ∇^2 are the gradient and Laplacian operators, respectively. The coefficient $\theta > 0$ represents the rate of diffusion, and $\mathbf{v}$ is the two-dimensional advection coefficient, representing the speed at which a quantity such as heat, concentration, or pollutant is transported by the bulk movement of a fluid. Both θ and $\mathbf{v}$ are considered constant but possibly unknown over a fixed interval.

In this work, we consider the field as a linear superposition of multiple non-overlapping advection diffusion phenomena, each governing a spatial-temporal region $R_i, \quad i = 1, \ldots, M$ where M is the number of the advection-diffusion processes. The concentration $z_{i(r,t)}$ in each region satisfies the advection-diffusion equation: $\frac{\partial z_i}{\partial t}(r,t) = \theta \nabla^2 z_i(r,t) + \mathbf{v}^T \nabla z_i(r,t), r \in R_i$. The global concentration field is then represented as

$$z(r,t) = \sum_i \chi_i(r) z_i(r,t), \quad r \in \Omega, \tag{2}$$

where $\chi_i(r)$ is an indicator function defined as $\chi_i(r) = 1$ if $r \in R_i$ and 0 otherwise, and $\Omega = \bigcup R_i, \quad i = 1, \ldots, M$. Moreover, in various real-world environmental monitoring scenarios, the domain Ω is significantly larger than the dimensions of the operational robots, enabling the approximation of the boundary as essentially flat. Under these conditions, we apply the initial and Dirichlet boundary conditions as shown in Equation 3 at the boundary $\partial\Omega$ (Demetriou et al., 2013):

$$\begin{aligned} z(r,0) &= z_0(r), \\ z(r,t) &= 0, \quad r \in \partial\Omega. \end{aligned} \tag{3}$$

2.2 Mobile sensing robots

In this work, we consider a group of N mobile sensing robots moving in a coordinated formation in the field Ω represented in Equation 2. The algorithm proposed in this work commands the formation to travel on planned paths to efficiently reconstruct the unknown field. We make the following assumption regarding these mobile sensing robots.

Assumption 2.1: *Each sensing robot is equipped with sensors to localize itself and to measure the field concentration value at its current location at each discrete time step k.*

The measurement of the ith sensing robot at time step k is modeled as follows:

$$p\left(r_i^k, k\right) = z\left(r_i^k, k\right) + n_i, \quad i = 1, \ldots N, \tag{4}$$

where r_i^k represents the location of the ith robot at the discrete time step k and n_i is assumed to be i.i.d Gaussian noise. Additionally, using the locations of all the robots in the formation at time step k, we can determine the location of the formation center r_c^k at time step k as $r_c^k = \frac{1}{N}\sum_i^N r_i^k$.

When the robots move in a desired formation, it covers a time-varying view-scope $\Gamma(k)$, which is the area of the field domain Ω that lies inside the polygon formed by sensing robot locations. As shown in Figure 1, the shaded region illustrates the time-varying view-scope $\Gamma(k)$ at discrete time step k, the blue circles represent the four mobile robots, and the red circle represents the formation center. At any given time, the mobile sensing robots can measure and exchange concentration values at their specific locations as shown in Equation 4 and the field values $z(r^k,k), \quad r^k \in \Gamma(k)$ can be estimated by interpolating the measured values from the robots. Consequently, it is reasonable to assume that the estimated field values, $z(r^k,k), r^k \in \Gamma(k)$, are available to us at all times.

In this work, to facilitate the implementation of the PPO algorithm for source mapping in Section 4.1, we discretize the global field Ω to a $E \times F$ grid, where each grid cell represents a single location r^k in the map and associates with a concentration value $z(r^k,k)$. The following assumption holds for the formation center.

Assumption 2.2: *Robots travel in a coordinate formation and the formation center moves along the eight possible directions "up", "down", "left", "right", "up-left", "up-right", "down-left", "down-right" in the discretized domain.*

With the robots moving in a formation, a CKF developed in (You et al., 2016; Wu et al., 2020) is employed to output estimates of concentration $z(r_c^k,k)$ and gradients $\nabla z(r_c^k,k)$ at the formation center r_c at time step k. These estimated values will play a major

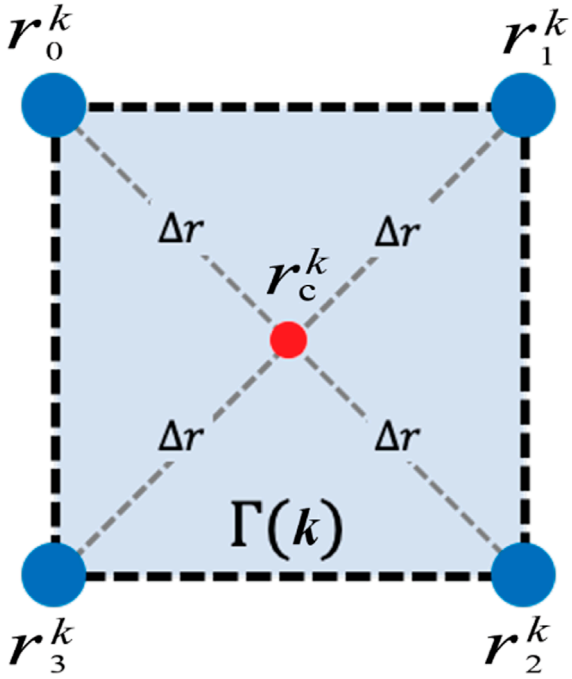

FIGURE 1
A symmetric formation composed of four mobile robots r_i^k, $i = 0,1,2,3$ shown in blue. The formation center r_c^k is shown in red. The distance between each robot and the formation center is Δr. The shaded region is the time-varying view-scope $\Gamma(k)$.

role in the developed algorithm in Section 4. Furthermore, we apply the parameter identification algorithms developed in (Wu et al., 2020) to estimate the unknown diffusion coefficient θ in real-time using the output from the CKF and the RLS algorithm. Thus, in the following discussions, we consider θ as a known value for field reconstruction.

Remark 2.1: *Multi-robot formation control is a well-studied topic and researchers have developed numerous formation control algorithms* (Zhang and Leonard, 2010; Ren and Beard, 2008; Wu and Zhang, 2012). *In this work, we employ the formation control strategy developed in* (Zhang and Leonard, 2010) *and applied in* (Wu et al., 2020). *The strategy uses the Jacobi transform to decouple the formation control from the motion control of the multi-robot formation, which enables us to only plan the path and design the controller for the formation center* r_c. *The individual robot controllers are then derived using the formation controller.*

2.3 The multi-robot source seeking and field reconstruction problem

In real-world scenarios, the task of mapping complex dynamic fields for cases like gas-leaking and wildfires is important and is often time-critical. It is essential for the robot formation to explore and detect diffusion sources in unknown areas and generate a map as quickly as possible. With the field defined in Section 2.1 and the multi-robot formation defined in Section 2.2, the goal of this study is to design a path for the multi-robot formation so that the formation can identify the multiple non-overlapping diffusion sources in the dynamic field and reconstruct the field in real-time with the limited concentration measurements collected by the multi-robot formation along its trajectory. To achieve the goal, we will introduce a two-mode strategy in Section 4, which consists of a Source Mapping mode and a Field Exploration mode. In the Source Mapping mode, we employ the RL-based algorithm and train a PPO model to guide the multi-robot formation toward a diffusion source in the field and reach a stationary state, where the formation arrives at the source and moves with the field at the same speed as the advection flow. In the Field Exploration mode, we develop a K-means clustering-based exploration strategy to enable efficient exploration of unknown areas.

3 Preliminaries

3.1 Proximal policy optimization

PPO (Schulman et al., 2017) is a significant development in reinforcement learning, introduced as a more efficient and simpler alternative to Trust Region Policy Optimization (TRPO) (Schulman et al., 2015). PPO is based on the policy gradient approach, a class of algorithms that optimize policies by directly computing gradients of expected rewards for policy parameters. This approach allows the learning agent to improve its policy iteratively by following the gradient of expected rewards. PPO enhances this process by addressing the complexities of earlier methods while retaining their benefits, particularly in maintaining stable and reliable policy updates.

PPO operates as an on-policy method. Unlike traditional policy gradient methods that apply a single update after each interaction with the environment, PPO refines the policy by using multiple updates on the same batch of data. The core of PPO is its surrogate objective function, designed to prevent large, potentially destructive policy updates. This is achieved through a probability ratio r_k between the new and old policies, which is clipped to keep updates within a safe range. The surrogate objective function $L^{CLIP}(\theta)$ is expressed as:

$$L^{CLIP}(\theta) = \widehat{E}_k\left[\min\left(r_k(\theta)\hat{A}_k, clip\left(r_k(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_k\right)\right]. \quad (5)$$

In Equation 5, $\widehat{E}_k$ denotes the expectation over timestep k, θ represents the policy parameters, $\widehat{A}_k$ is an estimate of the advantage function at time step k, and ε is a small hyperparameter that controls the clipping range. It is important to note that while r_k and θ follow the conventional notations used in literature, they differ from the notations in other sections of this paper, where r refers to the locations in the field and θ refers to the diffusion coefficient. The clipping mechanism ensures that if the probability ratio deviates outside the predefined range $[1-\varepsilon, 1+\varepsilon]$, the function applies the clipped values to prevent excessively large updates, thereby maintaining the stability of the learning process. By constraining the probability ratio, PPO effectively controls the size of policy updates, balancing stability and performance. PPO is particularly well-suited for discrete action spaces, which makes it an ideal choice for our environment setup. The PPO algorithm is summarized in Algorithm 1.

Input: Initial policy parameters θ_0, initial value function parameters ϕ_0
for $k = 0,1,2,\ldots$ **do**
Generate a set of trajectories $D_k = \{\tau_i\}$ by running the policy $\pi_k = \pi(\theta_k)$ in the environment
Calculate the rewards-to-go $\hat{R}_k$
Estimate advantages $\hat{A}_k$ using a suitable method based on the current value function V_{ϕ_k}
Update the policy by maximizing the PPO-Clip objective

$$\theta_{k+1} = \arg\max_{\theta} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in D_k} \sum_{k=0}^{T} \min\left(r_k(\theta)\hat{A}_k, clip\left(r_k(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_k\right),$$

where $r_k(\theta) = \frac{\pi_\theta(a_k|s_k)}{\pi_{\theta_k}(a_k|s_k)}$.
Update the value function by minimizing the mean-squared error:

$$\phi_{k+1} = \arg\min_{\phi} \frac{1}{|D_k|T} \sum_{\tau \in D_k} \sum_{k=0}^{T} \left(V_\phi(s_k) - \hat{R}_k\right)^2.$$

Algorithm 1. PPO-Clip Algorithm.

3.2 Cooperative Kalman Filter

Cooperative Kalman Filter (CKF) is a collaborative state estimation scheme first developed in (Zhang and Leonard, 2010), then used in later studies (Wu and Zhang, 2012; You et al., 2016; Wu et al., 2020; You et al., 2022), by combining live sensor data collected by the network of multiple mobile robots to collaboratively improve the accuracy of the state estimation process. In particular, when applied to the state estimation in dynamic fields, the authors incorporated the dynamics of the mobile robot formation and the diffusion equation into the formulation of the state equation of the CKF. This integration facilitates reliable and accurate state estimation, taking into account how changes in diffusion fields and the formation trajectory over time affect sensor data measurements. More specifically, the state vector $X(k)$ at each time step k is defined as:

$$X(k) = \left[z\left(r_c^k, k\right), \nabla z\left(r_c^k, k\right), z\left(r_c^k, k-1\right), \nabla z\left(r_c^k, k-1\right)\right]^T,$$

where $z\left(r_c^k, k\right)$ and $z\left(r_c^k, k-1\right)$ denote the field concentration values at location r_c^k at two consecutive time steps k and $k-1$, respectively, and $\nabla z\left(r_c^k, k\right)$ and $\nabla z\left(r_c^k, k-1\right)$ denote the field gradient at location r_c^k at two consecutive time steps k and $k-1$, respectively. Given that the mobile robots maintain a symmetrical formation while traversing the environment, CKF estimates the state vector $X(k)$ along the trajectory of the formation center. Note that since the field is spatial-temporal varying, the field concentration values and gradients are different at time steps k and $k-1$ even at the same location r_c^k. This fact is critical in the construction of the CKF to provide reliable estimates of the state vector. The estimated state vector is subsequently employed to iteratively identify the diffusion coefficients of the field over time, based on the RLS algorithm. The estimated diffusion coefficients are vital for the task of identifying and reconstructing spatial diffusion fields in our paper. To save time and avoid excessive length in this paper, we will not provide the complete derivation of the CKF. For additional details, interested readers may refer to the original papers.

4 Methodology

In this section, we introduce the proposed path-planning algorithm for guiding the mobile sensing robot formation to quickly explore an open field while reliably mapping and reconstructing all detected diffusion sources along its trajectory. The algorithm aims to find a balance between speed and reliability for the dynamic field reconstruction. To achieve this, the solution alternates the robot formation between two operational modes: Map Exploration and Source Mapping. In Map Exploration, the robots systematically advance toward unexplored regions to detect new diffusion fields. Upon detecting a new diffusion field, the system transitions to Source Mapping, where the formation converges on the field's center to achieve a stationary state, necessary for estimating advection coefficients.

Throughout both modes, the robots continuously collect data, using the CKF for real-time concentration and gradient estimation and the RLS algorithm for identifying diffusion coefficients. In the discrete simulation environment, concentration estimates are interpolated across the formation's view-scope. Mode transitions are based on the formation's state and the concentration estimates at its center. Figure 2 provides an overview of all major components in our algorithm and their interaction within the two operation modes. In the following sections, we will provide details of the algorithms developed for the two modes.

4.1 Source mapping mode

As discussed in the high-level overview, the goal of the formation in Source Mapping mode is to move toward the source of a diffusion field and facilitate diffusion field reconstruction by estimating advection and diffusion parameters. For this purpose, we train a PPO model that takes the field information vector state to predict the optimal action. In this section, we describe the setup of our training environment and the architecture of our PPO model.

We define the observation input state $S(k)$ for our PPO model as follows:

$$S(k) = \left[z\left(r_c^k, k\right), \nabla z_x\left(r_c^k, k\right), \nabla z_y\left(r_c^k, k\right)\right], \quad (6)$$

where $\nabla z_x\left(r_c^k, k\right)$ and $\nabla z_y\left(r_c^k, k\right)$ represent the estimated concentration gradients at the formation center at time step k, in the x and y directions, respectively. As discussed in the previous section, we rely on the CKF to provide estimates of the field concentration and gradients at the formation center, which form the complete input state of our model. By incorporating concentration gradients in the observation state, we provide the model with the direction of the largest concentration value change at the current location, which can be useful for heading toward the source center. Additionally, our definition of the state vector $S(k)$ in Equation 6 limits the model to

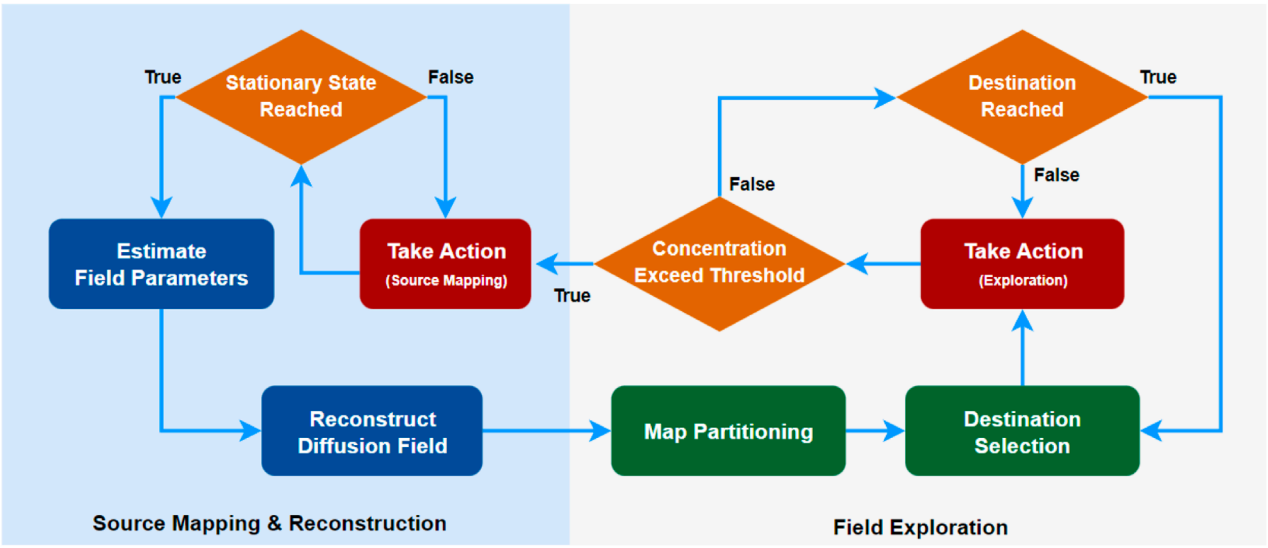

FIGURE 2
Flow chart showing the key components of our algorithm and two operation modes of the robot formation.

learning state-action values solely based on the characteristics of the field. With the formation controller, the mobile sensor robots can maintain a constant desired formation while traversing the environment; thus; we only plan the path for the formation center instead of individual robots.

For every time step, our robot formation can move to any adjacent cells (including diagonal) or stay at the current location. Thus, we can define the action space $A(k)$ consisting of nine actions as follows:

$$A(k) = \{\text{“up”}, \text{“down”}, \text{“left”}, \text{“right”}, \text{“up – left”}, \text{“up – right”}, \text{“down – left”}, \text{“down –right”}, \text{“stay”}\}. \quad (7)$$

Since our goal is to train a PPO model that can guide the formation toward the center of a diffusion field and maintain stationary state as long as possible, it is crucial to develop a reward function that incentivizes this behavior. For this reason, we model the reward function based on the concentration values inside the formation view-scope as follows:

$$R(k) = \alpha \sum_{r \in \Gamma(k)} z\left(r^k, k\right), \quad (8)$$

where α is a rescale constant. As regions with high concentration values play a significant role in field reconstruction, a reward proportional to the total concentration values within the view-scope as shown in Equation 8 motivates the model to learn to navigate towards areas with higher concentration values. This, in turn, greatly reduces the error in reconstruction and prioritizes information-rich trajectory. We chose PPO as our model due to its long-standing role as a crucial component in various state-of-the-art solutions in reinforcement learning. Furthermore, PPO can be used for environments with discrete action space, as in our case. Figure 3 shows the architecture of our PPO model. Since PPO is a type of actor-critic algorithm, it has two neural networks - the actor network and critic network. In our case, we employed the same architecture

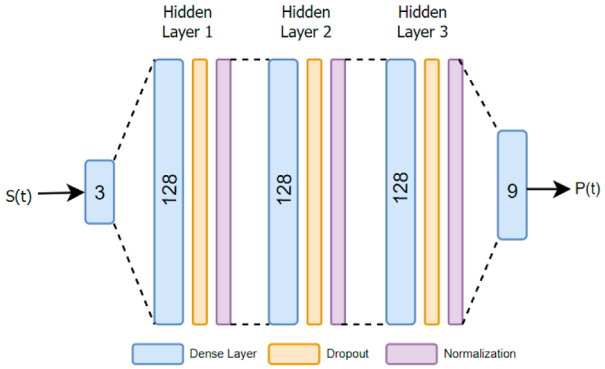

FIGURE 3
The PPO neural network architecture used in the source mapping mode.

for both. This network consists of three dense layers of size 128 each, an input layer of size 3 and an output layer of size 9. We added random dropout layers and regularization between the hidden layers, with ReLu as the non-linear activation function (Agarap, 2018). The network is trained using Adams optimizer (Kingma and Ba, 2015).

Using the trained PPO model, the robot formation takes actions chosen from the action space in Equation 7, and is guided toward the source of the diffusion field until it reaches the stationary state. This stationary state is achieved when the estimated field concentration reaches a local maximum and the estimated field gradient approaches zero, i.e., $z\left(r_c^k, k\right) > z\left(r_c^{k-1}, k-1\right)$, $z\left(r_c^k, k\right) > z\left(r_c^{k+1}, k+1\right)$, and $\nabla z\left(r_c^k, k\right) \approx 0$. At this point, the formation moves at the same speed as the advection flow for a designated period of T steps before switching to the Field Exploration mode to search for other diffusion fields in unexplored areas.

4.2 Field exploration mode

Whenever the robot formation reaches a stationary state, indicating the detection of the source of a diffusion field, the robot formation switches back to Field Exploration mode and moves toward unexplored area in the map to look for the remaining diffusion fields. During this phase, it is essential for the robot formation to come up with a new destination that is away from the already explored locations to avoid revisiting the same source, but also not too far to cause the formation to go back and forth when scanning the whole map. In short, our main objective is to generate a path that allows our robot formation to scan the whole field as quickly as possible without leaving any diffusion field undetected. Lawn mowing is a good example of generating a deterministic trajectory for scanning an unknown map. However, since our mobile sensing robots are often deployed in time-critical missions. It is necessary to opt for a more aggressive exploration strategy that allows the formation to discover all sources as quickly as possible. In our algorithm, we partition the unvisited cells in the entire map into multiple clusters using the K-Means clustering algorithm (Na et al., 2010). The selection of K directly affects how aggressive or cautious the scanning behavior of our robot formation will be. The initial value of K is selected based on the initial estimate of the minimum size of a diffusion field. In this work, we use the term "size" to refer to the bounded area of a diffusion field where the concentration value exceeds a certain threshold. For different scenarios (such as wildfires and gas leaks), we are often able to come up with a rough estimation of the average size of a diffusion field. Let us denote this as $\tilde{S}_{field}$ and the global map size as $S_{map} = E \times F$. Then, K is estimated as:

$$K = \max\left(2, \left\lfloor \frac{S_{map}}{\tilde{S}_{field}} \right\rfloor\right). \tag{9}$$

After partitioning the map, the robot formation selects the centroid of the nearest cluster as the new destination and moves toward that destination to explore the field. It continues to visit the centroids of other clusters, prioritizing nearby clusters, as long as it is in the Field Exploration mode. The formation switches to Source Mapping mode when the estimated field concentration value exceeds a chosen threshold, i.e., $z(r_c^k, k) > \delta$. At all times, the formation maintains a record of visited clusters and diffusion fields, effectively creating a mask that distinguishes between explored and unexplored regions.

Whenever a switch occurs from Field Exploration mode to Source Mapping mode and the formation reaches the stationary state in the Source Mapping mode (indicating a new diffusion field is detected), the robot formation evaluates the field and computes a new estimated K value to repartition the remaining unexplored areas in the map. The "size" of the newly discovered diffusion field can be roughly estimated based on the circular area with radius R_{dist} extending from the source center to the location where the concentration first surpasses the threshold. This is also where the formation switches from Field Exploration to Source Mapping mode. Denote the size as S_{new} represented as:

$$S_{new} \approx \pi * R_{dist}^2. \tag{10}$$

With the size of the latest detected diffusion field calculated based on Equation 10, we can update the estimated average size of the diffusion field $\tilde{S}_{field}$ as follows:

$$\tilde{S}'_{field} = (1-\beta) \cdot \tilde{S}_{field} + \beta \cdot S_{new}, \tag{11}$$

where $\beta \in [0, 1]$ is the weighting factor that determines the influence of the newly discovered field on the updated estimate $\tilde{S}'_{field}$. This approach allows for increasing accurate field size estimation with new discovery. Note that K will only decrease or maintain unchanged with each new source found. When the size of an unexplored area in the map falls below a certain threshold, we consider the whole map has been adequately covered. At this point, The robot formation can either be set to idle mode or be directed to transition to a new map (potentially a neighboring global field) to initiate its operations from a different starting location. The exploration strategy can be summarized in Algorithm 2.

```
Input: Unexplored region M_0 as an N×N grid map.
Initialize:
Compute initial estimate K_0 based on diffusion field radius using Equation 9.
Apply K-Means Clustering with K = K_0 to partition M_0 into K_0 clusters.
Let A_centroids be the set of centroids for these clusters.
Set concentration threshold δ for mode switching.
Set target destination Target = (x_k, y_k) = r_c^k.
Let DONE ← False.
while z(r_c^k, k) < δ do
  if formation center r_c^k = Target then
    if A_centroids ≠ ∅ then
      Set Target as the nearest centroid C_nearest in A_centroids.
      Remove C_nearest from A_centroids.
    else
      DONE ← True
      break
  Move formation towards Target using A*path-finding algorithm.
if DONE then
  Transition to Source Mapping Mode.
else
  Move to new region M_i and restart exploration.
```

Algorithm 2. K-Means Clustering Based Exploration Mode.

4.3 Parameter estimation and field reconstruction

The field reconstruction begins when the formation detects the source of a diffusion field within the global domain. Given that our environment is modeled as a 2D grid, we discretize the diffusion Equation 1 to enable field reconstruction. Assuming the domain of interest Ω is divided into square cells of size $\Delta r_x =$

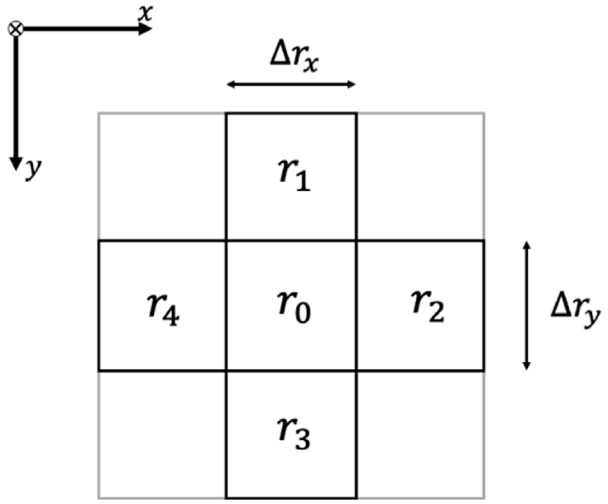

FIGURE 4
A 3 × 3 section of the discretized advection-diffusion field.

Δr_y, as illustrated in Figure 4, where a 3×3 grid is demonstrated. Here, r_1, r_2, r_3, r_4 are the neighboring cells of grid cell r_0. Let $z(r_0,k), z(r_1,k), z(r_2,k), z(r_3,k)$ and $z(r_4,k)$ denote the concentration values at grid cells r_0 to r_4 at discrete time step k. Using the finite difference method, the discretized advection-diffusion equation can be expressed as:

$$\begin{aligned}\frac{z(r_0,k+1)-z(r_0,k)}{t_s} = \theta\Bigg[&\frac{z(r_2,k)+z(r_4,k)-2z(r_0,k)}{\Delta r_x^2}\\ &+\frac{z(r_1,k)+z(r_3,k)-2z(r_0,k)}{\Delta r_y^2}\Bigg]\\ &+\mathbf{v}^T\nabla z(r_0,k)+e(r_0,k),\end{aligned} \tag{12}$$

where t_s denotes the sampling interval and $e(r_0,k)$ represents the approximation error. With the symmetric property, Equation 12 can be further simplified to:

$$\frac{z(r_0,k+1)-z(r_0,k)}{t_s} = \theta\frac{\sum_{i=1}^{4} z(r_i,k)-4z(r_0,k)}{\Delta r_x^2}+\mathbf{v}^T\nabla z(r_0,k)+e(r_0,k). \tag{13}$$

To reconstruct a diffusion field using the measurements taken by the robot formation with Equation 13, we need the estimated field concentration values $\hat{z}(r,k)$ within the view-scope of the robot formation at each time step k, as well as the estimated diffusion coefficient $\widehat{\theta}$ and the advection vector $\hat{\mathbf{v}}$. The former can be obtained through interpolation at each time step using the measurements taken by the robots. As mentioned previously, we employ the strategy developed in (Wu et al., 2020) to identify the diffusion coefficient θ. In many scenarios, the advection coefficient $\mathbf{v}$ is assumed to be a known constant. When the advection coefficient is unknown, we estimate it when the robot formation reaches a stationary state, where the robot's velocity matches that of the advection flow. Let $\mathbf{v} = (v_x, v_y)^T$ and $k_S > 0$ represent the time step when a stationary state is detected. Assuming the formation stays at the stationary state for T steps, (v_x, v_y) can be estimated as

$$\begin{aligned}\hat{v}_x &= \frac{r_{c,x}^{k_S+T}-r_{c,x}^{k_S}}{T},\\ \hat{v}_y &= \frac{r_{c,y}^{k_S+T}-r_{c,y}^{k_S}}{T}.\end{aligned} \tag{14}$$

With these values determined, the field values across the diffusion field can be propagated through Equation 13. This approach enables field reconstruction using only the sparse measurements gathered along the robots' paths.

5 Simulation results

In this section, we provide a comprehensive analysis of the proposed multi-robot field reconstruction strategy, which encompasses source mapping and field exploration modes in simulations. We begin by outlining the implementation details, followed by a discussion on the PPO training specifics. Finally, we present the results derived from these simulations.

5.1 Simulation environment

To assess the overall solution, we developed a low-fidelity simulation environment within a discrete space. This environment is structured as a 100×100 square grid, incorporating between one to four non-overlapping spatial diffusion fields of varying sizes and configurations, as depicted in Figure 5. We generated a total of 15 different environments, with increasing level of complexity, to investigate and assess the model's efficiency in mapping unknown environments. The color of each cell in the grid denotes the concentration value of the diffusion field. Each diffusion field possesses distinct and independent advection and diffusion coefficients. Figure 6 shows the evolution of a sample diffusion field over time. The environment was simulated for up to 400 time steps with the discrete interval $\Delta t = 0.1$, $\Delta x = 0.1$, and $\Delta y = 0.1$. Table 1 lists the configurations of the 15 diffusion field environments with advection terms (v_x, v_y) and diffusion term θ. The center of the field is denoted as $pos(x,y)$. Note that *size* is only used internally by the generator as a scale factor to control how large a diffusion field appears on the map.

Since the map is discretized in our approach, selecting an appropriate grid cell size also plays a role in both the accuracy of data collection and computational efficiency. Each grid cell should be small enough to capture meaningful concentration gradients but large enough to reduce computational demands. Ideally, the grid cell size should reflect both the overall map dimensions and the characteristics of the environment being monitored. For example, when studying gas leaks, where subtle concentration changes are significant, a finer grid may be required. On the other hand, wildfire propagation fields, which tend to cover larger areas, can accommodate slightly larger cells. Adjusting the grid cell size based on the specific characteristics of the environment allows us to find the right balance between resolution and computational cost, facilitating effective and efficient exploration and reconstruction.

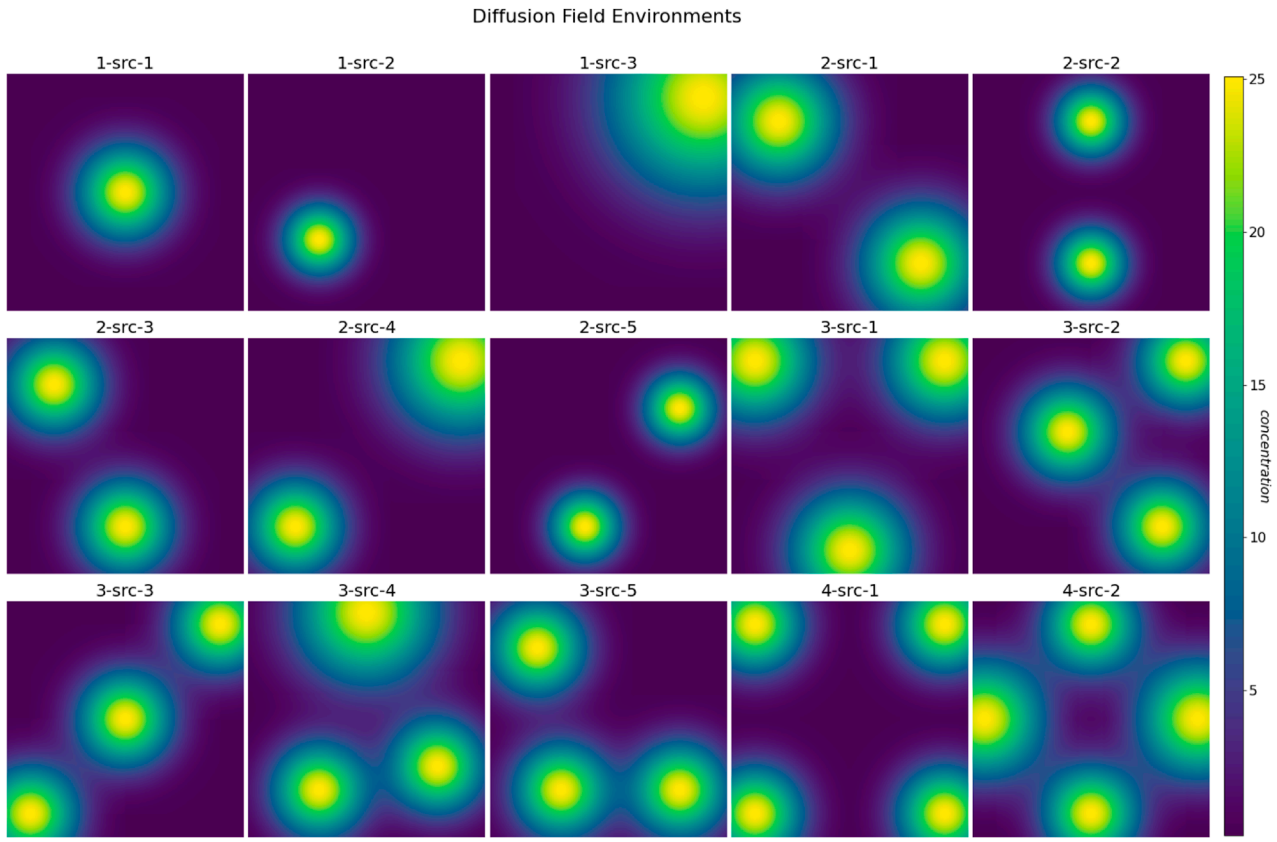

FIGURE 5
A set of 15 simulation environments, each consisting of one to four diffusion fields centered at various locations with distinct diffusion and advection coefficients.

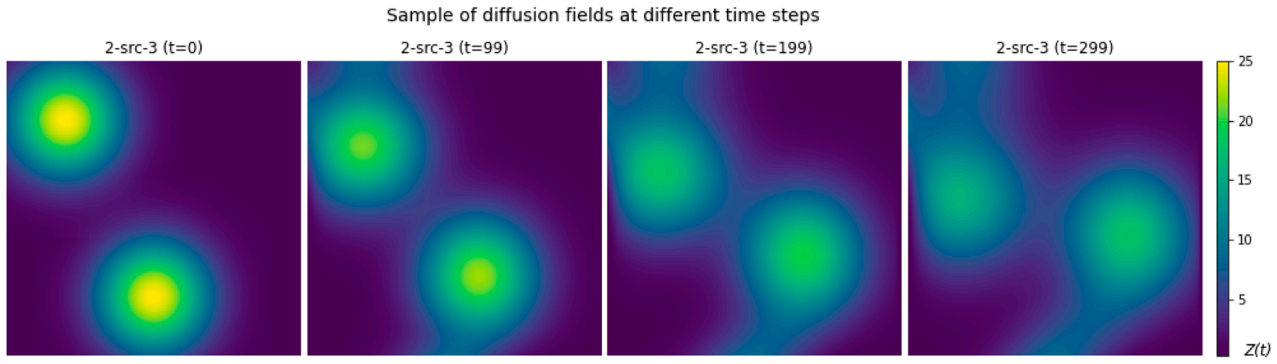

FIGURE 6
Sample of a diffusion field environment over different time steps.

5.2 PPO training

For the training of the PPO model, we follow a curriculum learning approach (Wang et al., 2023; Wang et al., 2022) that involves gradually increasing the complexity of the training environment. We created a 100×100 training environment featuring a single diffusion field at the center of the map $p = (50, 50)$ with a constant diffusion coefficient $\theta = 1$. The spread of this diffusion field at the beginning of an episode is drawn randomly based on the parameter *size*, which controls how large the area with non-zero concentration is due to the presence of the diffusion field. This environment has two variants: a "static" environment with the advection term set to zero, and a "dynamic" environment with a fixed non-zero advection term.

We simulate a group of four mobile robots in a symmetric formation as shown in Figure 1 to move in the environments, with the formation controller running to maintain the desired formation.

TABLE 1 Configurations of the 15 diffusion field environments with advection terms (v_x,v_y) and diffusion term θ. The center of the field is denoted as *pos*(*x*,*y*). Note that *size* is only used internally by the generator as a scale factor to control how large a diffusion field appears on the map.

Source	pos (x,y)	Size	v_x	v_y	θ
1-src-1	(50, 50)	80	0.73	−0.44	0.76
1-src-2	(30, 70)	60	0.14	−0.65	1.02
1-src-3	(90, 10)	160	0.74	0.09	1.24
2-src-1	(20, 20)	100	0.62	−0.01	1.07
	(80, 80)	100	−0.04	−0.14	0.96
2-src-2	(50, 20)	60	−0.22	−0.47	1.04
	(50, 80)	60	−0.59	−0.45	1.08
2-src-3	(20, 20)	80	−0.71	0.09	1.17
	(50, 80)	80	0.56	−0.67	0.78
2-src-4	(20, 80)	80	−0.20	0.37	0.75
	(90, 10)	120	0.71	0.78	0.99
2-src-5	(40, 80)	60	0.10	0.38	1.15
	(80, 30)	60	−0.09	0.25	1.07
3-src-1	(10, 10)	100	0.54	0.19	1.21
	(50, 90)	100	0.03	0.58	0.94
	(90, 10)	100	−0.64	0.67	0.79
3-src-2	(40, 40)	80	0.39	−0.50	0.81
	(80, 80)	80	0.60	0.25	0.93
	(90, 10)	80	−0.12	0.11	1.18
3-src-3	(50, 50)	80	−0.37	0.24	1.13
	(10, 90)	80	−0.32	−0.75	1.21
	(90, 10)	80	0.76	0.71	1.02
3-src-4	(50, 5)	120	−0.49	0.17	0.91
	(30, 80)	80	−0.50	0.56	0.84
	(80, 70)	80	−0.44	0.42	0.98
3-src-5	(20, 20)	80	0.58	0.45	0.97
	(80, 80)	80	0.66	−0.18	0.99
	(30, 80)	80	0.73	−0.23	0.99
4-src-1	(10, 10)	80	−0.08	0.22	1.14
	(90, 10)	80	0.37	0.35	0.77
	(90, 90)	80	−0.72	−0.38	0.76
	(10, 90)	80	0.37	0.63	0.86
4-src-2	(50, 10)	80	0.72	−0.73	1.05
	(50, 90)	80	−0.75	0.74	1.06
	(5, 50)	100	0.35	−0.05	1.03
	(95, 50)	100	0.12	0.73	1.25

(Continued on the following page)

With the CKF providing the state $S(k)$ which includes the estimated field value $z(r_c^k, k)$ and gradient along the trajectory of the formation center $\nabla z_x(r_c^k, k)$ and $\nabla z_y(r_c^k, k)$, the PPO model underwent training on static maps first before advancing to training on dynamic maps. After training, our PPO model outputs the policy to direct the robot formation to move toward the source of a diffusion field and maintain the stationary state, which is required for the estimation of advection coefficients.

We provide the training results in Figure 7. The PPO model was initially trained for 400,000 time steps on the static environment where advection terms are set to zero, as shown in Figure 7A. After that, we proceeded to train the PPO model on the dynamic environment for an additional 2,000,000 time steps, as shown in Figure 7B The model was trained multiple times with orthogonal random weight initialization using the Stable-Baselines3 framework (Raffin et al., 2021), and the best-performing model was selected for use in our experiments. In both environments, the formation is initially placed in a low-concentration region at the start of each episode to avoid starting too close to the source. Additionally, in our dynamic environment, the source is assigned random, non-zero advection and diffusion coefficients, causing it to move in a different direction in each episode. This setup encourages the model to adapt to various scenarios but also introduces some fluctuations in performance early in training, as the formation may take suboptimal actions initially and struggle to catch up to the moving source. In both training phases, however, the average reward per episode increases steadily, indicating that the model successfully improves its given task over time. Additionally, Figure 7C provide some samples of source-heading operation performed by our PPO model post-training. In these samples, the 4-robot formation, shown as the red square in the map, are tasked with locating the source while maintaining its formation. The yellow region denotes the

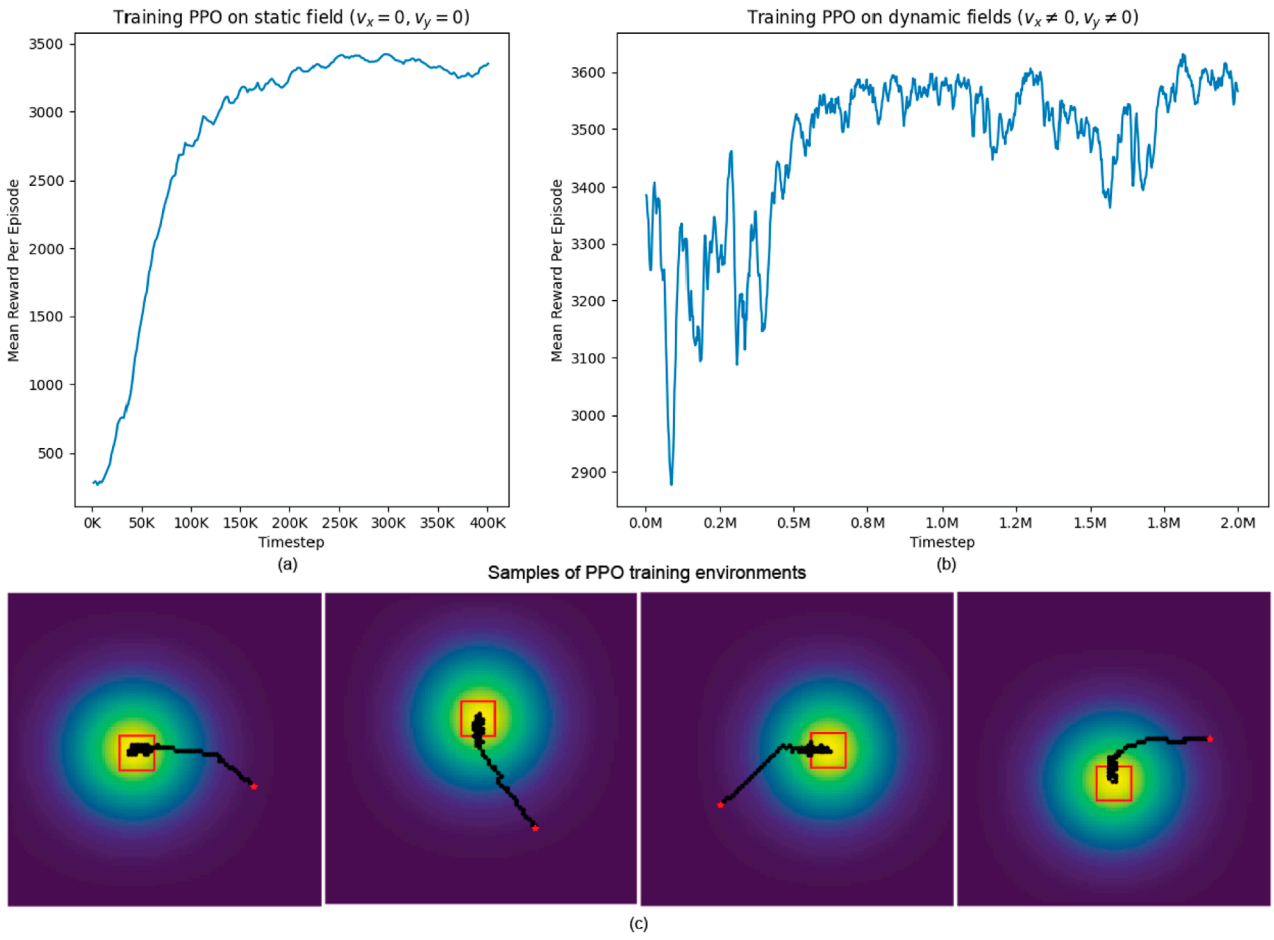

FIGURE 7
(A) Stage 1: Pre-training PPO on static fields. **(B)** Stage 2: Training PPO on dynamic fields. **(C)** Samples of training environments and generated trajectories post-training.

area with high concentration - where the source is located, the red square represents the robot formation, and the red dot is the starting position of the formation center. The robot formation is directed towards the center of the source, which has maximum concentration, per the PPO's objective of reward maximization. PPO's role in this task is to update the policy per iteration to make an informed decision on where to go next. The simulation results show the PPO algorithm's efficacy in source mapping.

The PPO algorithm's training process effectively learns a policy that guides the robot formation to explore the field, detect diffusion sources, and assist in reconstruction while adapting to dynamic changes in the environment. The algorithm's capability in handling complex and dynamic environments indicates its potential in real-life scenarios where rapid environmental change happens, and accurate detection of diffusion sources is crucial. This capability is precious in pollution tracking or gas leak detection scenarios, where time-sensitive and precise localization is essential.

5.3 K-means clustering-based exploration

When the PPO model leads the robot formation to move toward a diffusion source and the formation center reaches the stationary state, the advection coefficient can be estimated based on Equation 14, and the field reconstruction process can start using Equation 13. Along with the field reconstruction process, the robot formation switches to the field exploration mode, where the K-means clustering-based exploration algorithm 2 plays a role.

Figure 8 provides an example of how the K-means clustering-based exploration works in a 100×100 grid map with two diffusion fields. The formation started in Field Exploration mode with the starting position shown as the red dot. Based on our initial assumption about the average size of diffusion fields, the map is partitioned into six clusters with $K = 6$ as illustrated in Figure 8B. The formation moves toward the centroid of the nearest cluster and continues to other neighboring centroids until an area

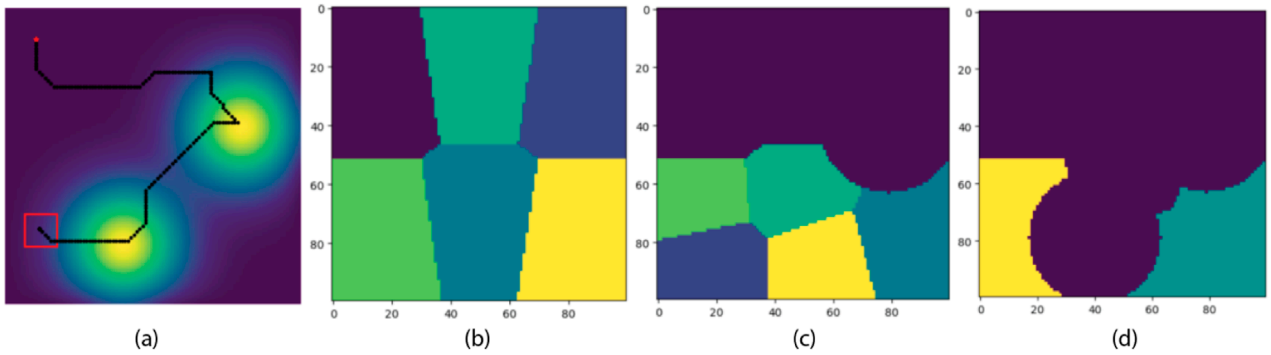

FIGURE 8
Field Re-partitioning Behavior of the Exploration Module using K-Mean Clustering. The red dot is the starting position of the formation center. Partitioning happens at the beginning of an episode and whenever a new source is detected. **(A)** The trajectory of the formation center. **(B)** The initial partition of the field. **(C)** The updated partition after the first source on the upper right corner is detected. **(D)** The updated partition after the second source on the lower left corner is detected.

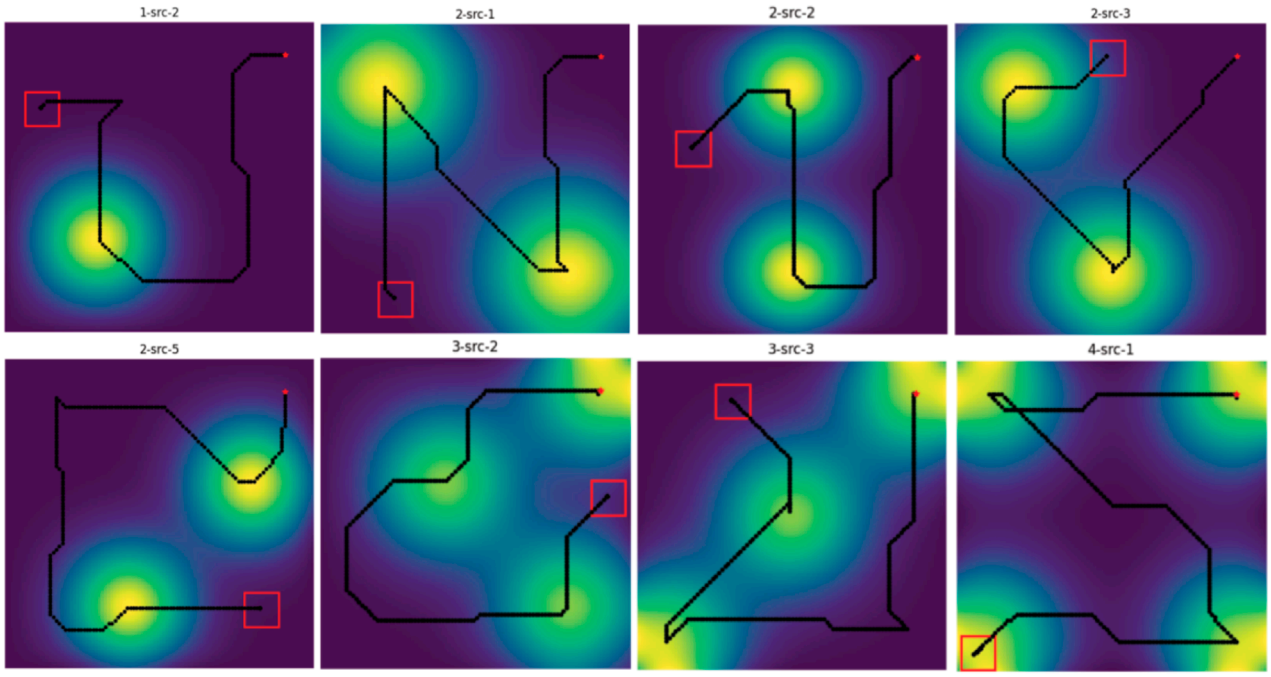

FIGURE 9
Generated trajectories obtained from simulations across different spatial-diffusion environments. The robot formation center is initially located at r_c = (90,10) - shown as the red marker, and the red square shows the position of the formation the end of the episode.

with a high concentration is detected. When that happens, the formation transitions to Source Mapping mode and attempts to move toward the source center to reconstruct the field. When the new diffusion field is fully reconstructed, the formation switches back to Field Exploration mode, re-calculates a new K as described in Equations 9-11 and re-partitions the unexplored area, as shown in Figure 8C. Note that the unexplored area has excluded visited clusters and any detected diffusion fields. This process continues to repeat until the map is fully covered.

5.4 Multi-robot source seeking and field exploration results

With the trained PPO model and the K-means clustering-based exploration algorithm, we now ready to implement the overall multi-robot Source Mapping and Field Exploration strategy to reconstruct a dynamic field. Figure 9 shows different trajectories of the robot formation obtained from the simulation across multiple spatial-diffusion environments. In this setup, the robot formation started

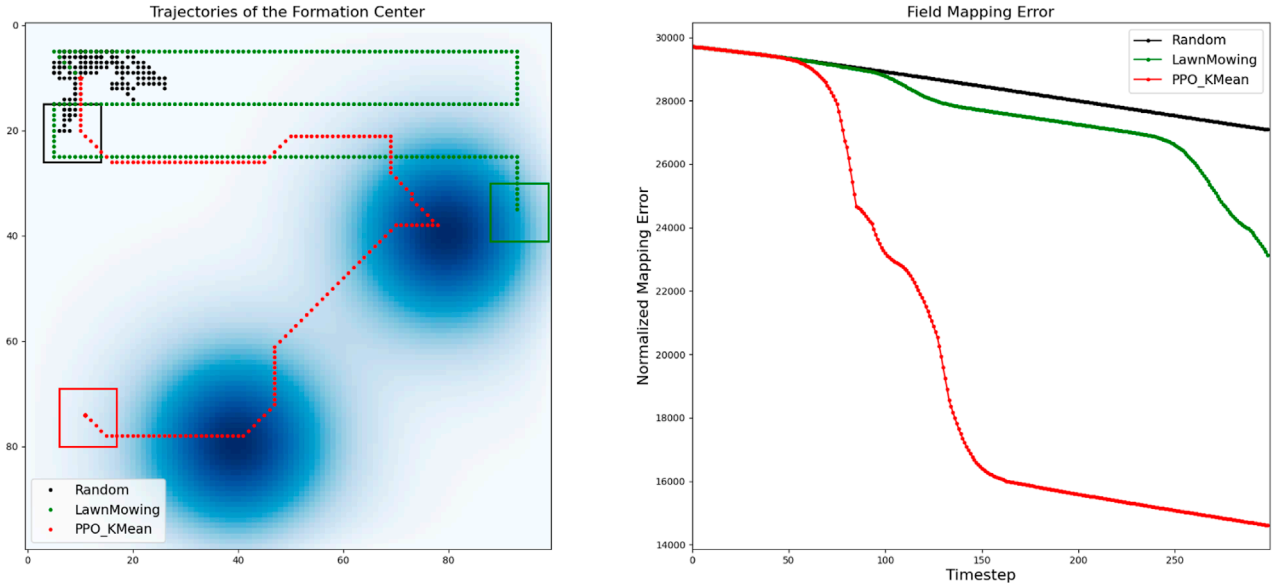

FIGURE 10
Analysis of the generated trajectory and mapping errors of our solution in comparison with the Lawn Mowing and Random Walking approaches.

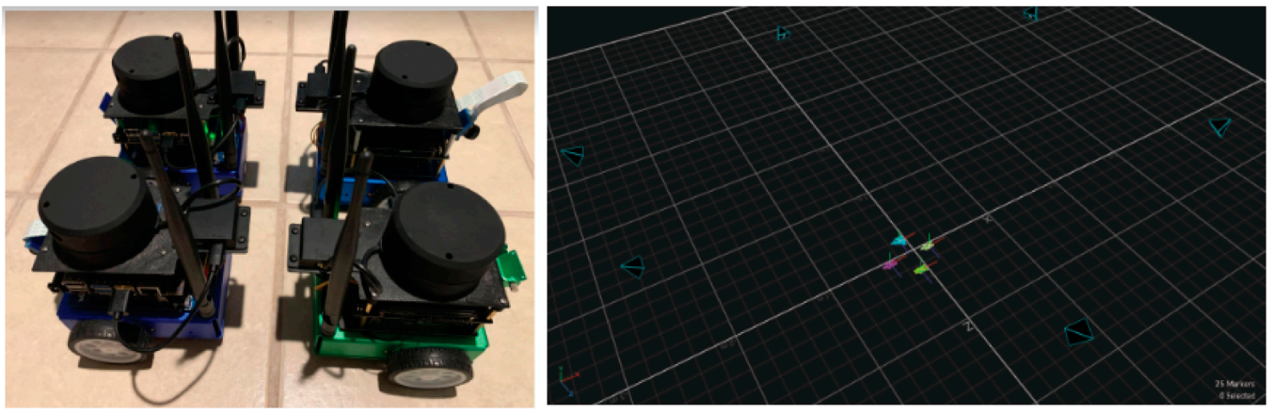

FIGURE 11
Mobile robot formation setup for real-world testing and evaluation.

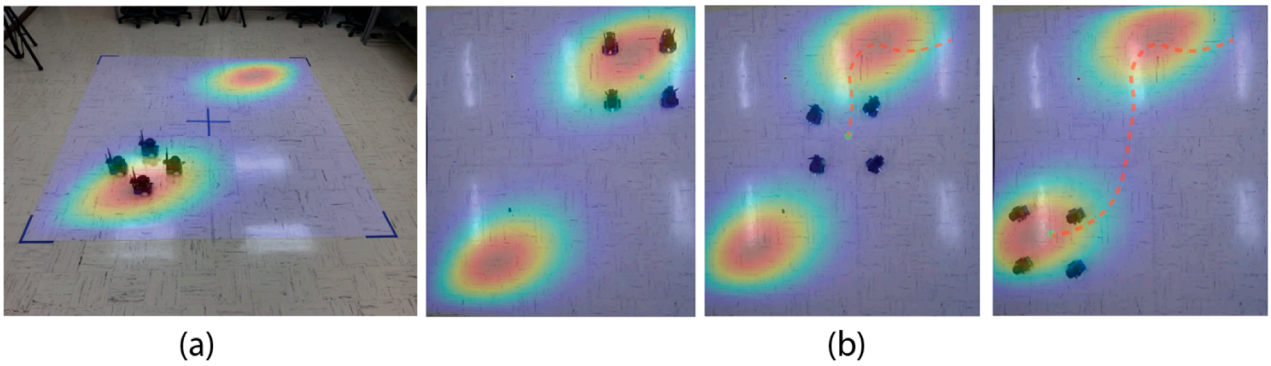

FIGURE 12
(a) Lab experiment setup with a projected dynamic field and four mobile robots. **(b)** Snapshots of the trajectories of the robot formation at three different time steps in an experiment. The red dashed lines represent the trajectories.

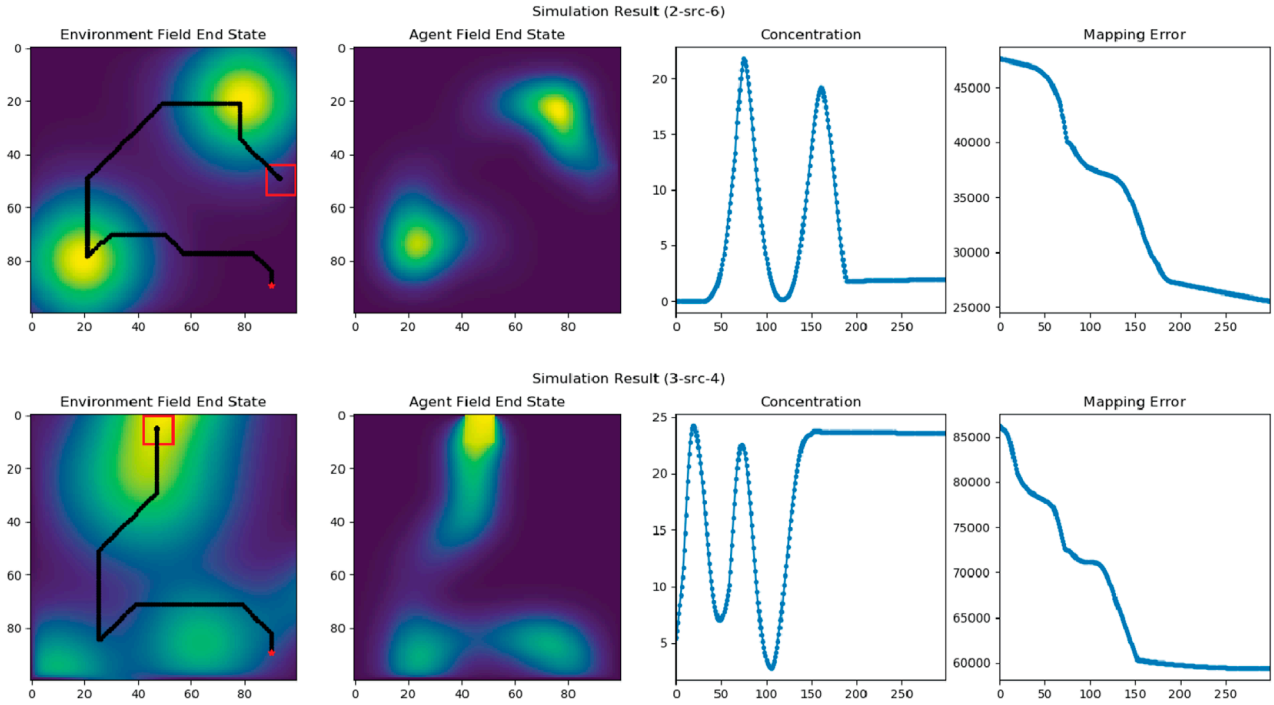

FIGURE 13
The field exploration and reconstruction results in two experiments with two and three diffusion sources. "Environment Field End State" figures illustrate the end states of the two experiments with corresponding trajectories of the robot formation. The red dots indicate the starting locations of the formation center and the red squares are the ending locations of the formation. "Agent Field End State" figures show the end states of the reconstructed fields in the two experiments. "Concentration" figures illustrate the estimated field concentration along the trajectories of the formation center, and "Mapping Error" figures show the mapping errors while reconstructing the fields.

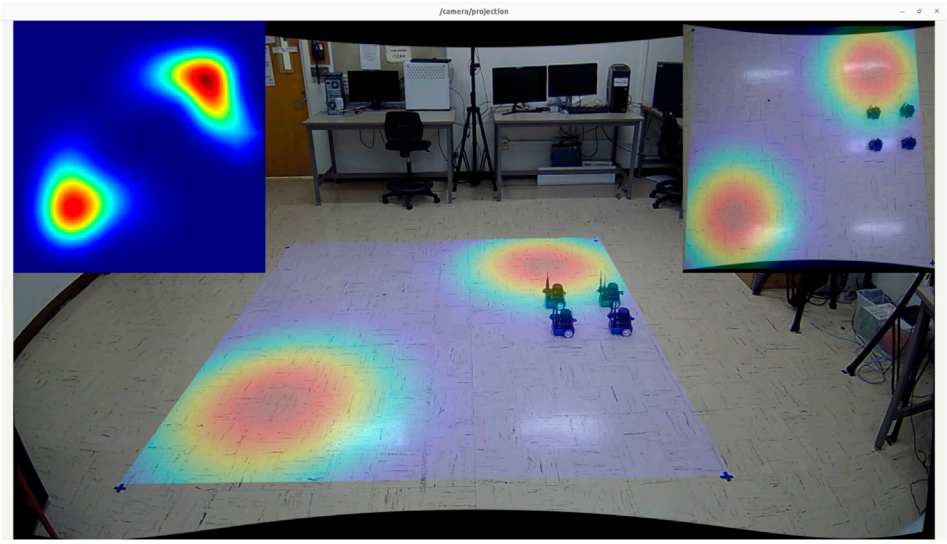

FIGURE 14
Screenshot captured from the experiment on the first environment with two diffusion fields.

at the same initial position $r_c^0 = (90, 10)$ (top-right corner, shown as a red marker). The red square displays the final position of the robot formation at the end of the episode ($k = 300$). Despite variations in spatial-diffusion field distributions and characteristics, we can see that the robot formation managed to detect all sources in the map while efficiently covering the entire map before the episode ends.

In Figure 10, we compare the trajectory and mapping errors of our solution against other alternative path-planning algorithms,

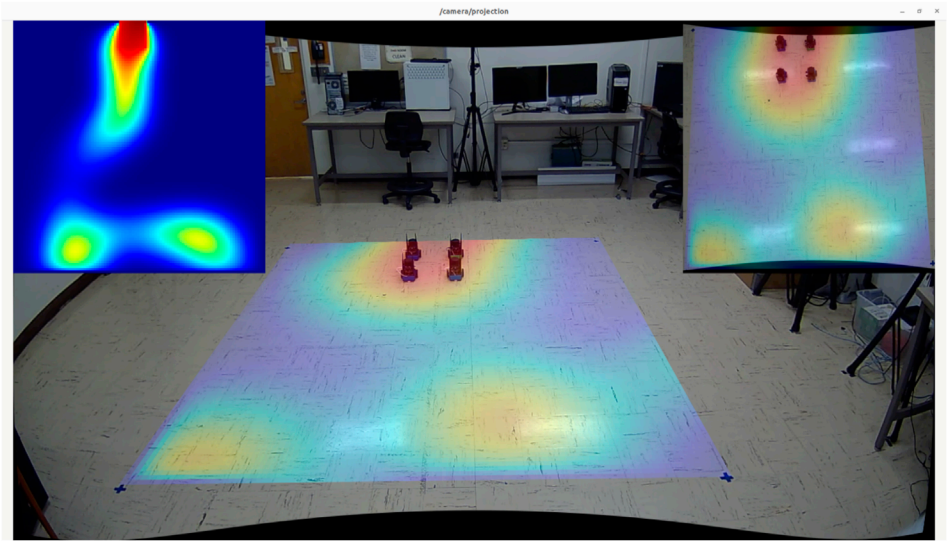

FIGURE 15
Screenshot captured from the experiment on the second environment with three diffusion fields.

including Random-Walking (shown in black) and Lawn-Mowing (shown in green). With the Random-Walking strategy, the formation takes a random action in every time-step, resulting in an arbitrary trajectory. On the other hand, with Lawn-Mowing strategy, the formation attempts to scan the field row-by-row until the entire field is fully covered. Compared to these approaches, the trajectory generated by our solution is faster at detecting and mapping diffusion fields, resulting in a much lower mapping error. In this map, our K-Mean-based approach is the only one that detects all diffusion fields before the episode ends ($k = 300$).

6 Experimental results

To validate our solution in a real-world setting, we developed a high-fidelity testing environment in our lab. Our setup includes four mobile robots operating in a 12 × 12 square foot open field that simulates an actual advection-diffusion environment. Figure 11 shows our laboratory setup of the mobile sensor network consisting of 4 mobile robots with motion tracking enabled, allowing for accurate collection of real-time trajectory data. The robots are two-wheel differential drive and ROS-based, running on the Jetson Nano (Developer Nvidia, 2024) computing platform. Each robot is equipped with a 2D Lidar scanner [YDLidar-G4 (YDLIDAR-G4-Datasheet, 2024)], a speed encoder, and an IMU (BNO-055 (Industries, 2024)). To enable low-latency sensor fusion, a Teensy 4.0 (PJRC, 2024) collects and preprocesses sensor readings from the speed encoder and IMU before streaming the results to the main board via rosserial. Lidar is installed to enable basic obstacle avoidance behaviors, allowing the formation to adapt to various scenarios when navigating in outdoor environments.

For localization, we rely on an indoor motion capture system to provide absolute positional tracking, analogous to GPS in outdoor scenarios. An Extended Kalman Filter (Ribeiro, 2004) fuses data from both the IMU and motion capture system to improve real-time position estimation of the robots. Our software stack uses ROS Noetic (Quigley et al., 2009) and its ecosystem to facilitate sensor fusion for localization and obstacle avoidance, as well as to simulate and visualize the behaviors of the advection-diffusion field. In our stack, each robot has its own action server (based on ROS Action), which is responsible for moving the robot to a target destination. A master node running on a stand-alone computer is responsible for broadcasting the concentration values of the simulated field as well as performing formation control during the experiment.

Figure 12 presents various views of the simulated spatial-temporal diffusion field in our laboratory as well as an example of the trajectory generated by our robot formation. Given the difficulties of installing physical diffusion field sources indoors, we utilized computational models to simulate the environment. The simulated field is projected onto the floor in real-time footage captured by side and top-down cameras. Sensor measurements are generated based on the robots' locations, which are tracked using the motion capture system.

To validate our solution in real-world settings, we selected two spatial-diffusion environments from our list and ran simulation experiments using actual mobile robots. While the spatial-temporal diffusion field is generated by computer simulation, the mobile sensor network is still designed to function exactly like how they should behave in the real-world. This involves having individual mobile robots take raw measurements and combine the results to estimate the concentration and gradients at the formation center, using CKF. Figure 13 provides a summary of our experimental results. The first column "Environment Field End State" shows the final states of our spatial diffusion environments and the trajectories of the robot formation. In both experiments, the formation enters the map from the bottom right corner with $r_c^0 = (90, 90)$ - shown as a red marker, and the red square again shows the formation's final location when the episode ends. The "Agent Field End State" column

shows the final reconstructed field computed by our mobile sensor network. As we can see, the formation center managed to explore the entire map while following an information-rich trajectory, which resulted in consistently high concentration readings and low mapping errors. Additionally, the results from experiments in the high-fidelity simulation environment closely resemble those from the low-fidelity testing environment, demonstrating that our solution can be adapted to more realistic scenarios. Additional screenshots from our laboratory experiments are available in Figures 14, 15. The plots in the upper left corners in both figures illustrate the reconstructed fields.

7 Conclusions and future work

In this paper, we developed a strategy to map and reconstruct dynamic fields with multiple diffusion sources using a multi-robot formation. This strategy proved effective on various maps with different configurations. Our approach efficiently explores unknown maps while ensuring that potential diffusion sources are detected. The results from our experiments show that the robot formation can effectively utilize environment data from all robots to navigate toward the source center and accurately reconstruct the advection and diffusion coefficients. While we did encounter some challenges with overlapping diffusion fields, these complexities only underscore the need for further research and detailed experiments. Our system holds potential for practical use in scenarios like rescue missions and field explorations, where robots can assess hazards before sending humans into these environments. This research shows the capability and versatility of our multi-robot system in environmental monitoring and could be important in enhancing safety measures during high-risk missions.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

TL: Writing–original draft, Formal Analysis, Methodology, Software, Validation, Visualization, Data curation. DS: Writing–original draft, Formal Analysis, Software, Validation, Visualization. DT: Writing–review and editing, Methodology, Software. WW: Writing–original draft, Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The research work is supported by NSF grants CMMI-1917300 and RINGS-2148353.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Agarap, A. F. (2018). *Deep learning using rectified linear units (ReLU)*. CoRR abs/1803 08375. doi:10.48550/arXiv.1803.08375

Bi, Q., Zhang, X., Wen, J., Pan, Z., Zhang, S., Wang, R., et al. (2024). Cure: a hierarchical framework for multi-robot autonomous exploration inspired by centroids of unknown regions. *IEEE Trans. Automation Sci. Eng.* 21 (3), 3773–3786. doi:10.1109/tase.2023.3285300

Burgard, W., Moors, M., Stachniss, C., and Schneider, F. (2005). Coordinated multi-robot exploration. *Robotics, IEEE Trans.*21, 376–386. doi:10.1109/tro.2004.839232

Cao, X., Li, M., Tao, Y., and Lu, P. (2024). Hma-sar: multi-agent search and rescue for unknown located dynamic targets in completely unknown environments. *IEEE Robotics Automation Lett.* 9 (6), 5567–5574. doi:10.1109/lra.2024.3396097

Chen, L., Dai, S.-L., and Dong, C. (2024). Adaptive optimal tracking control of an underactuated surface vessel using actor–critic reinforcement learning. *IEEE Trans. Neural Netw. Learn. Syst.* 35 (6), 7520–7533. doi:10.1109/tnnls.2022.3214681

Christopoulos, V., and Roumeliotis, S. (2005). *Adaptive sensing for instantaneous gas release parameter estimation*, 4450–4456.

Demetriou, M. A., Gatsonis, N. A., and Court, J. R. (2013). Coupled controls-computational fluids approach for the estimation of the concentration from a moving gaseous source in a 2-D domain with a Lyapunov-guided sensing aerial vehicle. *IEEE Trans. Control Syst. Technol.* 22 (3), 853–867. doi:10.1109/tcst.2013.2267623

Developer Nvidia (2024). Developer Nvidia. Available at: https://developer.nvidia.com/embedded/jetson-nano (Accessed January 9, 2024).

Dunbabin, M., and Marques, L. (2012). Robots for environmental monitoring: significant advancements and applications. *IEEE Robot. Autom. Mag.* 19 (1), 24–39. doi:10.1109/mra.2011.2181683

Gautam, A., Shekhawat, V. S., and Mohan, S. (2019). "A graph partitioning approach for fast exploration with multi-robot coordination," in *2019 IEEE international conference on systems, man and cybernetics (SMC)*, 459–465.

Hu, Y., Fu, J., and Wen, G. (2023). Graph soft actor–critic reinforcement learning for large-scale distributed multirobot coordination. *IEEE Trans. Neural Netw. Learn. Syst.*, 1–12. doi:10.1109/TNNLS.2023.3329530

Industries, A. (2024). Adafruit 9-dof absolute orientation imu fusion breakout - bno055. Available at: https://www.adafruit.com/product/4646 (Accessed January 9, 2024).

Khaled, C., Mustapha, E.-R., Olivier, (2004). On the rate of spread for some reaction-diffusion models of forest fire propagation. *Numer. Heat. Transf. Part A Appl.* 46 (8), 765–784. doi:10.1080/104077890504456

Kinaneva, D., Hristov, G., Raychev, J., and Zahariev, P. (2019). "Early forest fire detection using drones and artificial intelligence," in *2019 42nd international convention on information and communication technology, electronics and microelectronics* MIPRO Opatija, Croatia, 1060–1065.

Kingma, D., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *International conference on learning representations (ICLR)* (San Diega, CA, USA).

Long, J., and Zhang, B. (2024). "Multi-robots path planning and mapping for exploring unknown environment," in *2024 7th international conference on intelligent robotics and control engineering (IRCE)*, 72–76.

Luvisutto, A., Shehhi, A. A., Mankovskii, N., Renda, F., Stefanini, C., and De Masi, G. (2022). "Robotic swarm for marine and submarine missions: challenges and perspectives," in *2022 IEEE/OES autonomous underwater vehicles symposium (AUV)*, 1–8.

Martins, A., Almeida, J., Almeida, C., Dias, A., Dias, N., Aaltonen, J., et al. (2018). "Ux 1 system design - a robotic system for underwater mining exploration," in *IEEE/RSJ international conference on intelligent robots and systems IROS*, 1494–1500.

Mourikis, A., and Roumeliotis, S. (2006). Performance analysis of multirobot cooperative localization. *IEEE Trans. Robotics* 22 (4), 666–681. doi:10.1109/tro.2006.878957

Na, S., Xumin, L., and Yong, G. (2010). "Research on k-means clustering algorithm: an improved k-means clustering algorithm," in *2010 third international symposium on intelligent information technology and security informatics*, 63–67.

Niroui, F., Zhang, K., Kashino, Z., and Nejat, G. (2019). Deep reinforcement learning robot for search and rescue applications: exploration in unknown cluttered environments. *IEEE Robotics Automation Lett.* 4 (2), 610–617. doi:10.1109/lra.2019.2891991

PJRC (2024). Teensy® 4.0 development board. Available at: https://www.pjrc.com/store/teensy40.html (Accessed on January 9, 2024).

Queralta, J. P., Taipalmaa, J., Can Pullinen, B., Sarker, V. K., Nguyen Gia, T., Tenhunen, H., et al. (2020). Collaborative multi-robot search and rescue: planning, coordination, perception, and active vision. *IEEE Access* 8, 191617–191643. doi:10.1109/access.2020.3030190

Quigley, M., Conley, K., Gerkey, B. P., Faust, J., Foote, T., Leibs, J., et al. (2009). "ROS: an open-source robot operating system," in *ICRA workshop on open source software*.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. (2021). Stable-baselines3: reliable reinforcement learning implementations. *J. Mach. Learn. Res.* 22 (268), 1–8. doi:10.5555/3546258.3546526

Reisch, C., Navas-Montilla, A., and Özgen Xian, I. (2024). Analytical and numerical insights into wildfire dynamics: exploring the advection–diffusion–reaction model. *Comput. Math. Appl.* 158, 179–198. doi:10.1016/j.camwa.2024.01.024

Ren, W., and Beard, R. W. (2008). *Distributed consensus in multi-vehicle cooperative control*, 27. Springer.

Ribeiro, M. I. (2004). Kalman and extended kalman filters: concept, derivation and properties. *Inst. Syst. Robotics* 43 (46), 3736–3741.

Rossi, M., and Brunelli, D. (2016). Autonomous gas detection and mapping with unmanned aerial vehicles. *IEEE Trans. Instrum. Meas.* 65 (4), 765–775. doi:10.1109/tim.2015.2506319

Schulman, J., Levine, S., Moritz, P., Jordan, M., and Abbeel, P. (2015). "Trust region policy optimization," in *Proceedings of the 32nd international Conference on international Conference on machine learning - volume 37, Lille, France ICML'15* (JMLR.org), 1889–1897.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *CoRR* abs/1707, 06347. doi:10.48550/arXiv.1707.06347

Shuvo, M. I. R., Wimer, B., Mahmud, S., and Kim, J.-H. (2023). "A novel collaborative knowledge sharing and self-learning framework for robotic systems in search and rescue operations," in *IECON 2023- 49th annual conference of the IEEE industrial electronics society*, 1–6.

Talwar, D. (2020). *Deep reinforcement learning based path-planning for multi-agent systems in advection-diffusion field reconstruction tasks*.

Tricaud, C., and Chen, Y. Q. (2010). "Optimal trajectories of mobile remote sensors for parameter estimation in distributed cyber-physical systems," in *Proceedings of the 2010 American control conference*, 3211–3216.

Ucinski, D. (2005). "Optimal measurement methods for distributed parameter system identification," in *Taylor and Francis series in systems and control*. Boca Raton, Fla: CRC Press.

Ucinski, D., and Chen, Y. (2005). "Time-optimal path planning of moving sensors for parameter estimation of distributed systems," in *Proceedings of the 44th IEEE conference on decision and control*, 5257–5262.

Wang, H.-C., Huang, S.-C., Huang, P.-J., Wang, K.-L., Teng, Y.-C., Ko, Y.-T., et al. (2023). Curriculum reinforcement learning from avoiding collisions to navigating among movable obstacles in diverse environments. *IEEE Robotics Automation Lett.* 8 (5), 2740–2747. doi:10.1109/lra.2023.3251193

Wang, X., Chen, Y., and Zhu, W. (2022). A survey on curriculum learning. *IEEE Trans. Pattern Analysis Mach. Intell.* 44 (9), 4555–4576. doi:10.1109/TPAMI.2021.3069908

Wu, W., You, J., Zhang, Y., Li, M., and Su, K. (2020). Parameter identification of spatial–temporal varying processes by a multi-robot system in realistic diffusion fields. *Robotica* 39, 842–861. doi:10.1017/s0263574720000788

Wu, W., and Zhang, F. (2012). Robust cooperative exploration with a switching strategy. *IEEE Trans. Robotics* 28 (4), 828–839. doi:10.1109/tro.2012.2190182

YDLIDAR-G4-Datasheet (2024). YDLIDAR-G4-Datasheet. Available at: http://www.ydlidar.com/Public/upload/files/2020-04-13/YDLIDAR%20G4%20Datasheet.pdf (Accessed on January 9, 2024).

You, J., and Wu, W. (2018). "Geometric reinforcement learning based path planning for mobile sensor networks in advection-diffusion field reconstruction," in *2018 IEEE conference on decision and control (CDC)* (IEEE), 1949–1954.

You, J., Zhang, F., and Wu, W. (2016). "Cooperative filtering for parameter identification of diffusion processes," in *2016 IEEE 55th conference on decision and control* IEEE: CDC, 4327–4333.

You, J., Zhang, Z., Zhang, F., and Wu, W. (2022). Cooperative filtering and parameter identification for advection–diffusion processes using a mobile sensor network. *IEEE Trans. Control Syst. Technol.* 31 (2), 527–542. doi:10.1109/tcst.2022.3183585

Zhang, F., and Leonard, N. E. (2010). Cooperative filters and control for cooperative exploration. *IEEE Trans. Automatic Control* 55 (3), 650–663. doi:10.1109/tac.2009.2039240

Zhang, Z., Mayberry, S. T., Wu, W., and Zhang, F. (2023). Distributed cooperative kalman filter constrained by advection–diffusion equation for mobile sensor networks. *Front. Robotics AI* 10, 1175418. doi:10.3389/frobt.2023.1175418

Zhu, L., Cheng, J., and Liu, Y. (2023). "Multi-robot autonomous exploration in unknown environment: a review," in *2023 China automation congress (CAC)*, 7639–7644.