2026 American Control Conference (ACC)
May 27-29, 2026. New Orleans, Louisiana, USA

Optimal Control of Information Diffusion in Robot Swarms via a Policy-Driven SIR Model

Yanjing Huang*, Hao Gao*, Xianzhe Cheng, Yiding Ji, and Yun Gao

Abstract— **Efficient information diffusion is critical for coordination in large-scale robot swarms operating under dynamic communication topologies and limited energy resources. This paper studies diffusion control by modeling information propagation as a node-level susceptible–infected–recovered (SIR) process and formulating a finite-horizon optimal control problem that jointly regulates transmission intensity and robot mobility policies. We first derive a spectral diffusion condition with a time window that characterizes when information spreading is sustained in time-varying networks, and establish input-to-state stability (ISS) bounds that reveal a fundamental trade-off between diffusion speed and robustness. These theoretical insights motivate a theory-guided multi-robot reinforcement learning framework termed SIR-informed multi-agent proximal policy optimization (SIR-MAPPO). The framework adopts centralized training with decentralized execution and embeds feasibility projections, Lyapunov-informed critics, and hybrid model-based/model-free gradients to enforce stability and bandwidth constraints during learning. Convergence and stability of the learning dynamics are established analytically. Simulations of large-scale robot swarms demonstrate that the proposed approach achieves faster information diffusion and lower communication energy than representative baselines.**

Index Terms— **Robot swarms, information diffusion, epidemic dynamics, optimal control, reinforcement learning.**

I. Introduction

Robot swarms (RSs) consist of large groups of decentralized mobile robots that collaboratively accomplish tasks beyond the capability of any single robot [1]. They are widely applied in disaster response [2] and large-scale exploration [3], where timely and reliable information sharing is critical. Effective coordination relies on information dissemination, referred to here as diffusion, since its speed and coverage directly influence collective awareness and decision-making [4]–[6]. An effective diffusion strategy must achieve wide coverage with limited energy consumption, minimize redundant transmissions, and remain robust under mobility-induced topology changes and strict bandwidth constraints.

Classical diffusion protocols such as broadcasting [7], gossiping, and flooding [8] attempt to balance coverage and energy efficiency, yet they often suffer from excessive redundancy and vulnerability to network disruptions in large-scale mobile swarms. Epidemic models, particularly the Susceptible–Infected–Recovered (SIR) framework, provide a structured perspective for analyzing diffusion dynamics. By regulating transmission intensity over time, the SIR model enables a principled trade-off between coverage speed and energy expenditure and supports systematic intervention mechanisms. However, most existing epidemic-based approaches rely on mean-field approximations or centralized coordination, which are difficult to apply to decentralized robotic networks with time-varying communication graphs. Although the co-design problem can be formulated as an optimal control task [9], classical approaches such as dynamic programming and model predictive control become computationally prohibitive for large swarms with nonlinear epidemic dynamics and evolving communication topologies [10]. Moreover, gradient-based distributed optimization methods typically rely on linearization and fail to capture the strongly nonlinear and stochastic nature of node-level diffusion processes [11]. These limitations motivate the integration of control-theoretic analysis with scalable learning-based coordination mechanisms.

To address these challenges, we integrate control theory with reinforcement learning to optimize information diffusion in robot swarms. Unlike existing mean-field adaptations [12], we model the diffusion process at the node level. By formulating it as a discrete-time optimal control problem, we jointly regulate transmission intensity and robot mobility to balance coverage, energy, redundancy, and mobility costs. Building upon these analytical insights, we design a constrained, decentralized learning approach that effectively embeds system stability into policy optimization. The main contributions of this paper are summarized as follows:

- We establish a control-theoretic characterization of information diffusion in robot swarms by deriving a windowed spectral diffusion condition and ISS bounds for node-level SIR dynamics over time-varying communication graphs.
- We propose a theory-guided multi-robot reinforcement learning framework (SIR-MAPPO) that integrates feasibility projection, Lyapunov-informed critics, and hybrid model-based/model-free gradients to enable stable and scalable decentralized diffusion control.

Corresponding author: Yun Gao.

Yanjing Huang, Hao Gao, and Yiding Ji are with the Robotics and Autonomous Systems Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China (e-mail: yhuang403@connect.hkust-gz.edu.cn; hgao906@connect.hkust-gz.edu.cn).

Xianzhe Cheng is with the Information Support Force Engineering University, Wuhan, China (e-mail: chengxianzhe11@nudt.edu.cn).

Yun Gao is with the Robotics and Autonomous Systems Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China, and also with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China (e-mail: y.gao@gaoyunailab.com).

*The authors contributed equally to this work.

National Natural Science Foundation of China grants 62303389 and 62373289; Guangdong Basic and Applied Basic Research Funding grant 2024A1515012586; Guangdong Scientific Research Platform and Project Scheme grant 2024KTSCX039; Youth Talent Support Program of Guangdong Association for Science and Technology grant SKXRC2025463 and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007).

979-8-3315-9381-0//$31.00 ©2026 AACC

- We provide convergence analysis and stability performance guarantees for the closed-loop system, showing the proposed framework achieves efficient diffusion while maintaining bounded stability.

The remainder of this paper is organized as follows. Section II details the system model and formally defines the problem. Section III presents the control-theoretic analysis, the design of the SIR-MAPPO framework, and the associated theoretical guarantees. Section IV presents the simulation results, and Section V concludes the paper.

II. Preliminaries and Problem Formulation

A. Robot Swarm Communication Network

Consider N robots in $\mathbb{R}^d$ with discrete-time dynamics:

$$x_i(k+1) = x_i(k) + u_i(k), \tag{1}$$

where $x_i(k) \in \mathbb{R}^d$ denotes the position of robot i at time k, and $u_i(k) \in \mathbb{R}^d$ is the control input satisfying $\|u_i(k)\| \leq u_{\max}$.

At each time step k, the robots form a time-varying directed graph $G(k) = (V, E(k))$, which represents the underlying communication network. Here $V = \{1, \ldots, N\}$ denotes the set of robots and $E(k)$ represents the set of links. A directed edge $(i, j) \in E(k)$ indicates that robot i can successfully transmit information to robot j, which occurs if j is within i's communication range. The connectivity of the graph is described by the adjacency matrix $A(k) = [a_{ij}(k)]$, where $a_{ij}(k) \in [0, 1]$ quantifies the reliability of the link from robot i to robot j. To facilitate the analysis of information diffusion over the time-varying network, the following assumptions are introduced.

Assumption 1: There exists an integer $T > 0$ such that for every $k \geq 0$ the union graph $\bigcup_{t=k}^{k+T} G(t)$ is connected.

Assumption 2: There exists a fixed positive constant $L_A > 0$ such that for all time steps $k \geq 0$, the time-varying adjacency matrix $A(k)$ satisfies the 2-norm bound:

$$\|A(k)\|_2 \leq L_A. \tag{2}$$

B. Node-Level SIR Dynamics

To model the evolution of information states over the network, we associate each robot with a node-level discrete-time SIR diffusion process. At time step k, robot i has three normalized state probabilities $S_i(k), I_i(k), R_i(k) \in [0, 1]$, corresponding to the susceptible (uninformed), infected (informed and actively transmitting), and recovered (informed but inactive) states, respectively, and satisfying:

$$S_i(k) + I_i(k) + R_i(k) = 1. \tag{3}$$

The discrete-time state transition dynamics are given by:

$$S_i(k+1) = S_i(k)p_i(k), \tag{4}$$

$$I_i(k+1) = (1-\mu)I_i(k) + S_i(k)(1-p_i(k)), \tag{5}$$

$$R_i(k+1) = R_i(k) + \mu I_i(k), \tag{6}$$

where $p_i(k) = \prod_{j=1}^{N}(1-\beta_j(k)a_{ji}(k)I_j(k))$, $\mu \in (0, 1)$ is the recovery probability, $\beta_j(k) \in [0, \beta_{\max}]$ is the transmission rate assigned to robot j at time k, and $\beta_{\max} > 0$ is the maximum per-robot transmission intensity.

C. Problem Statement

The objective of the robot network is to disseminate information efficiently while avoiding excessive communication overhead. To quantify this objective, we introduce a cost function composed of four terms: coverage deficiency (7), communication energy (8), redundancy (9), and mobility penalty (10).

$$C(k) = 1 - \frac{1}{N}\sum_{i=1}^{N}(I_i(k) + R_i(k)), \tag{7}$$

$$E_n(k) = \sum_{j=1}^{N} I_j(k)(\alpha_1\beta_j(k) + \alpha_2\beta_j(k)^2), \tag{8}$$

$$R_d(k) = \sum_{j=1}^{N}\sum_{i=1}^{N} \beta_j(k)a_{ji}(k)I_j(k)(I_i(k) + R_i(k)), \tag{9}$$

$$M_o(k) = \sum_{i=1}^{N} \alpha_m \|u_i(k)\|^2. \tag{10}$$

In the above, α_1 and α_2 weight the linear and quadratic communication power terms, capturing circuit and transmission power consumption, respectively, while α_m reflects the mobility–energy conversion coefficient. The overall cost is:

$$\ell(k) = w_c C(k) + w_e E_n(k) + w_r R_d(k) + w_m M_o(k), \tag{11}$$

where $w_c, w_e, w_r, w_m > 0$ are design weights balancing coverage, energy, redundancy, and mobility. Based on the above diffusion dynamics and cost formulation, the control objective of this paper is to design control policies for the robots that minimize the cumulative cost over the network.

Problem 1 (Optimal diffusion control): Given the robot motion dynamics and the SIR diffusion model, design transmission intensity policies $\{\beta_j(k)\}$ and mobility control inputs $\{u_i(k)\}$ that minimize the expected finite-horizon cost:

$$\begin{aligned} \min_{\{\beta_j(k)\},\{u_i(k)\}} \quad & J = \mathbb{E}\left[\sum_{k=0}^{T}\gamma^k \ell(k)\right], \\ \text{s.t.} \quad & \|u_i(k)\| \leq u_{\max}, \quad \forall i, k \\ & 0 \leq \beta_j(k) \leq \beta_{\max}, \quad \forall j, k \end{aligned} \tag{12}$$

where $\gamma \in (0, 1)$ is the discount factor.

III. Theory-Guided SIR-MAPPO Framework

This section presents the proposed control and learning framework for multi-robot information diffusion. We first analyze the control-theoretic properties of the node-level SIR dynamics and derive conditions that characterize stable diffusion behaviors over the communication network. Based on these insights, we develop a centralized-training decentralized-execution (CTDE) reinforcement learning architecture that enables distributed policy learning for individual robots. Finally, learning-theoretic guarantees are established to justify the stability and convergence properties of the proposed framework.

A. Control-Theoretic Analysis

1) Diffusion Threshold and Speed Characterization: For small infection probabilities $I_i(k)$, expand (5) to first order.

Let $\mathbf{I}(k) = [I_1(k), \ldots, I_N(k)]^\top$, and let $\mathcal{I}^N$ denote the $N \times N$ identity matrix. Using $S_i(k) \approx 1$ since $I_i(k), R_i(k) \ll 1$ in the initial infection stage, we obtain:

$$\mathbf{I}(k+1) \approx (1-\mu)\mathbf{I}(k) + A(k)^\top B(k)\mathbf{I}(k) \tag{13}$$

where $B(k) = \mathrm{diag}(\beta_1(k), \ldots, \beta_N(k))$. Let $M(k) = (1-\mu)\mathcal{I}^N + A(k)^\top B(k)$, under the small-infection approximation, the dynamics can be written as the switched linear system $\mathbf{I}(k+1) = M(k)\mathbf{I}(k)$.

To analyze the evolution of information propagation over multiple time steps in the time-varying network, we introduce the multi-step transition matrix. Let the T-step transition matrix be:

$$\Phi(k+T, k) = M(k+T-1)\cdots M(k). \tag{14}$$

The matrix $\Phi(k+T,k)$ characterizes the cumulative effect of the time-varying diffusion dynamics over the interval $[k, k+T)$. To establish conditions under which information diffusion can be effectively sustained or suppressed, we introduce the following assumptions on the communication network and diffusion parameters.

Assumption 3: Since $a_{ij}(k) \geq 0$ and $\beta_j(k) \geq 0$, it follows that $M(k)$ and $\Phi(k+T,k)$ are nonnegative, with diagonal entries $1-\mu > 0$.

Theorem 1: Suppose Assumptions 1–3 hold. If spectral radius $\rho(\Phi(k+T,k)) > 1$, then any sufficiently small nonzero $\mathbf{I}(k) \succeq 0$ is amplified over the horizon T and spreads through the swarm due to joint connectivity. Moreover, the controllable diffusion speed over horizon T can be characterized as:

$$v(k) = \frac{1}{T}\log \rho(\Phi(k+T,k)). \tag{15}$$

Proof: Nonnegativity of $M(k)$ implies that $\Phi(k+T,k)$ is a nonnegative matrix. Since $M(k)$ has positive diagonal entries, joint connectivity of the communication graph over the interval $[k, k+T)$ ensures that the product matrix $\Phi(k+T,k)$ is irreducible.

By the Perron–Frobenius theorem for nonnegative irreducible matrices, $\Phi(k+T,k)$ has a simple dominant eigenvalue $\rho(\Phi(k+T,k)) > 0$ with an associated strictly positive eigenvector $v \succ 0$. Using the Perron decomposition, the powers of the matrix satisfy:

$$\Phi(k+T,k)^\ell \approx \rho(\Phi(k+T,k))^\ell v b^\top, \tag{16}$$

where b is the corresponding left eigenvector. Thus:

$$\mathbf{I}(k+\ell T) \approx \rho(\Phi(k+T,k))^\ell (\langle b, \mathbf{I}(k)\rangle) v. \tag{17}$$

If $\rho(\Phi(k+T,k)) > 1$, the dominant mode grows exponentially, and since $v \succ 0$, the infection spreads to all nodes. The speed characterization follows directly from (15). ■

Theorem 1 provides a windowed spectral condition for guaranteed spreading: diffusion is sustained whenever $\rho(\Phi(k+T,k)) > 1$ over some horizon T. Practically, this condition can be promoted by (i) increasing selected transmission intensities $\beta_j(k)$ during favorable connectivity periods, and (ii) exploiting mobility patterns enhancing temporal connectivity. Since $\rho(\Phi(k+T,k))$ increases monotonically with each $\beta_j(k)$ for nonnegative $M(k)$, the spreading condition is equivalent to the existence of a critical threshold $\beta_{\min}$ such that diffusion occurs whenever $\beta_j(k) \geq \beta_{\min}$.

2) Lyapunov Stability and ISS Robustness: To analyze the stability of the diffusion dynamics under nonlinear infection effects and topology perturbations, we consider the quadratic Lyapunov function:

$$V(\mathbf{I}(k)) = \mathbf{I}(k)^\top P \mathbf{I}(k), \tag{18}$$

where $P \succ 0$. And let $\lambda_{\min}(P)$ and $\lambda_{\max}(P)$ denote the minimum and maximum eigenvalues of P, respectively.

If the system evolved according to the nominal linear model, the Lyapunov function would satisfy $V(\mathbf{I}(k+1)) = \mathbf{I}(k)^\top M(k)^\top P M(k)\mathbf{I}(k)$. However, the true dynamics deviate from this ideal evolution due to nonlinear infection effects and perturbations in the interaction matrix. To quantify these deviations, we define the residual terms:

$$\Delta_{\mathrm{nl}}(k) = V(\mathbf{I}(k+1)) - \mathbf{I}(k)^\top M(k)^\top P M(k)\mathbf{I}(k), \tag{19}$$

$$\Delta_A(k) = \mathbf{I}(k)^\top (M(k)^\top P M(k) - M_0(k)^\top P M_0(k))\mathbf{I}(k), \tag{20}$$

with $M_0(k)$ being the unperturbed version of $M(k)$. Therefore, the one-step drift of the full nonlinear dynamics can be decomposed as:

$$\begin{aligned}\Delta V(k) = \mathbf{I}(k)^\top (M(k)^\top P M(k) - P)\mathbf{I}(k) \\ + \Delta_{\mathrm{nl}}(k) + \Delta_A(k).\end{aligned} \tag{21}$$

To establish stability guarantees, we impose boundedness conditions on the residual terms introduced above.

Assumption 4: Assume there exist constants $L_1, L_2, \delta \geq 0$ such that $|\Delta_{\mathrm{nl}}(k)| \leq L_1 \|\mathbf{I}(k)\|^2$ and $|\Delta_A(k)| \leq L_2\delta$.

Theorem 2: Suppose Assumptions 2–4 hold. If there exist $P \succ 0$ and some constant $\chi > 0$ such that:

$$M(k)^\top P M(k) - P \preceq -\chi \mathcal{I}^N, \tag{22}$$

with $\chi > L_1$, then:

$$\|\mathbf{I}(k)\| \leq h e^{-\lambda k}\|\mathbf{I}(0)\| + \frac{L_2\delta}{\chi - L_1}, \tag{23}$$

where $h = \sqrt{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}}$ and $\lambda = \frac{1}{2}\log(1 + \frac{\chi - L_1}{\lambda_{\max}(P)})$.

Proof: Combining (21) and (22) yields:

$$\Delta V(k) \leq -(\chi - L_1)\|\mathbf{I}(k)\|^2 + L_2\delta. \tag{24}$$

Using the eigenvalue bounds:

$$\lambda_{\min}(P)\|\mathbf{I}(k)\|^2 \leq V(\mathbf{I}(k)) \leq \lambda_{\max}(P)\|\mathbf{I}(k)\|^2, \tag{25}$$

we obtain:

$$V(\mathbf{I}(k+1)) \leq (1 - \frac{\chi - L_1}{\lambda_{\max}(P)})V(\mathbf{I}(k)) + L_2\delta. \tag{26}$$

Applying the discrete Grönwall inequality yields (23). ■

The steady-state term $L_2\delta/(\chi - L_1)$ quantifies robustness degradation under disturbances. A larger χ accelerates convergence but restricts the admissible transmission intensity. Since χ decreases monotonically as $\beta_j(k)$ increases, there exists a maximum admissible bound $\beta_{\max}$ such that stability is preserved if $\beta_j(k) \leq \beta_{\max}$ for all robots.

Theorem 1 and Theorem 2 establish a fundamental design region: transmission intensities must exceed the diffusion threshold while remaining within the ISS-stable regime.

B. SIR-MAPPO Framework

1) Constraint-Embedded Actor–Critic Architecture: The actor operates under a CTDE paradigm. During training, global information can be used to update the policy parameters. During execution, however, each robot has access only to local observations (27).

$$o_i(k) = [I_i(k), S_i(k), R_i(k), u_i(k), N_i(k), r_i(k)], \quad (27)$$

where $\mathcal{N}_i(k) = \{j \mid a_{ji}(k) > 0\}$ denotes the neighbor set of robot i from the communication graph $G(k)$, and $r_i(k) = |N_i(k)|^{-1} \sum_{j \in N_i(k)} a_{ji}(k) I_j(k)$ denotes the average received infection rate.

The actor is implemented as a function approximator that maps $o_i(k)$ to an unconstrained base output $\beta_{i,\mathrm{base}}(k) = f_\theta(o_i(k)) \in [0,1]$, where $f_\theta(\cdot)$ denotes the neural policy parameterized by θ. To enforce diffusion and stability constraints, the base output is first rescaled to the admissible interval $[\beta_{\min}, \beta_{\max}]$:

$$\beta_{i,\text{scaled}}(k) = \beta_{\min} + (\beta_{\max} - \beta_{\min})\beta_{i,\mathrm{base}}(k). \quad (28)$$

The transmission rates are then normalized to satisfy the global bandwidth constraint:

$$\beta_i(k) = \beta_{i,\text{scaled}}(k) \frac{W_{\max}}{\max(\hat{W}(k), \zeta)}, \quad (29)$$

where $W_{\max}$ denotes maximum total bandwidth, $\hat{W}(k)$ denotes an estimate of the aggregate bandwidth $\sum_{j=1}^{N} \beta_{j,\text{scaled}}(k)$, and $\zeta > 0$ avoids division by zero. This projection ensures that all executed actions remain physically admissible and consistent with the ISS condition.

Under centralized training, a critic value network $V_\phi(\mathbf{s}(k))$ parameterized by ϕ that evaluates an augmented global state $\mathbf{s}(k) = [o_1(k), \ldots, o_N(k)]$ constructed by concatenating the observations of all robots. To improve value estimation, the critic observes global diffusion metrics, including $\Delta C(k) = C(k-1) - C(k)$, which reflects the instantaneous reduction of infection level. These signals are used during training to compute value estimates and advantages. These global metrics are not available to the actors, ensuring fully decentralized execution. The critic is initialized using a Lyapunov function (18). This aligns early value estimates with stability requirements and accelerates convergence.

2) Training Loss Functions: The critic minimizes a temporal-difference (TD) loss augmented with a Lyapunov-based penalty:

$$\begin{aligned}\mathcal{L}_\phi = \mathbb{E}_{\mathcal{D}}[&(V_\phi(\mathbf{s}(k)) - (\ell(k) + \gamma V_\phi(\mathbf{s}(k+1))))^2 \\ &+ \lambda_V \max(0, V_\phi(\mathbf{s}(k)) - \mathbf{I}(k)^\top P \mathbf{I}(k))] \end{aligned} \quad (30)$$

where $\mathcal{D}$ is the replay buffer, γ denotes the discount factor, and λ_V weights the stability penalty.

The actor is trained with a clipped PPO objective combined with a bandwidth violation penalty:

$$\begin{aligned}\mathcal{L}_\theta = \mathbb{E}_{\mathcal{D}}[&\min(r_{\mathrm{trans}}(\theta)\mathcal{A}^\pi, \mathrm{clip}(r_{\mathrm{trans}}(\theta), 1-\epsilon, 1+\epsilon)\mathcal{A}^\pi) \\ &- \lambda_W \max(0, \sum_{j=1}^{N} \beta_j(k) - W_{\max})] \end{aligned} \quad (31)$$

where $r_{\mathrm{trans}}(\theta)$ is the policy ratio, $\mathcal{A}^\pi$ is the advantage estimate under the current policy π, clip$(\cdot)$ denotes the standard clipping operator used in PPO, ϵ the PPO clipping parameter, and λ_W weights the bandwidth penalty.

To reduce the variance of model-free (MF) gradients, we fuse MF and model-based (MB) gradients derived from the SIR dynamics:

$$\nabla_{\theta_i} J_{\text{hybrid}} = (1-\eta)\nabla_{\theta_i} J_{\mathrm{MF}} + \eta \nabla_{\theta_i} J_{\mathrm{MB}}, \quad (32)$$

where θ_i denotes the policy parameters associated with robot i, J_{hybrid} is the hybrid objective optimized during training, J_{MF} denotes the model-free reinforcement learning objective, J_{MB} represents the model-based objective derived from the diffusion dynamics, and $\eta \in [0,1]$ balances the contributions of the two terms. This hybrid estimator reduces variance relative to pure MF updates while preserving unbiasedness, thereby improving training stability.

3) Dynamic Weight Adaptation for Speed–Stability Trade-off: To incorporate the stability–speed tradeoff, the stage-cost weights in (11) are adapted based on the instantaneous diffusion rate $\Delta C(k)$. The weights are adjusted according to the diffusion phase, with a high w_c in the early stage to accelerate infection reduction, balanced weighting in the middle stage, and a reduced w_c in the late stage to emphasize energy efficiency and stability. The overall training procedure is summarized in Algorithm 1.

Algorithm 1 SIR-MAPPO

Initialize: Actor θ, Critic ϕ (with Riccati/Lyapunov matrix $P \succ 0$), Buffer $\mathcal{D}$.
for episode $= 1$ to MaxEpisodes **do**
  Reset environment; initialize epidemic states $S(0), I(0), R(0)$.
  for step $k = 0$ to $T-1$ **do**
    [Decentralized Execution]
    Each robot computes $\beta_{i,\mathrm{base}}(k)$ via θ, projects to $[\beta_{\min}, \beta_{\max}]$, and scales to satisfy $W_{\max}$.
    Execute $\{\beta_i(k), u_i(k)\}$; observe $\ell(k)$, $C(k)$, and next state.
    Compute $\Delta C(k) = C(k-1) - C(k)$; adapt weights (w_c, w_e, w_r, w_m) via $\Delta C(k)$.
    Store transition in $\mathcal{D}$.
  end for
  [Centralized Training]
  Compute advantages $\mathcal{A}^\pi$ using Critic V_ϕ.
  Update Critic ϕ by minimizing TD loss + Lyapunov penalty $\mathcal{L}_\phi$.
  Estimate hybrid gradient $\nabla J_{\mathrm{hybrid}} = (1-\eta)\nabla J_{\mathrm{MF}} + \eta \nabla J_{\mathrm{MB}}$.
  Update Actor θ using PPO objective $\mathcal{L}_\theta$ with the hybrid gradient estimate.
end for

C. Learning-Theoretic Guarantees for Closed-Loop Control

1) Convergence of MAPPO for Diffusion Control: We analyze convergence under standard two-timescale stochastic approximation. Let $\{\sigma_\theta^t\}$ and $\{\sigma_\phi^t\}$ denote the actor and critic learning rates. To ensure convergence of the coupled updates, we impose the following assumptions.

Assumption 5: The learning rates satisfy:

$$\sum_{t=0}^{\infty} \sigma_\theta^t = \infty, \quad \sum_{t=0}^{\infty} (\sigma_\theta^t)^2 < \infty, \quad \frac{\sigma_\phi^t}{\sigma_\theta^t} \to 0, \quad (33)$$

ensuring faster critic convergence and sufficient exploration.

Assumption 6: For each time step k, for any two feasible transmission-rate vectors $\boldsymbol{\beta}_1(k), \boldsymbol{\beta}_2(k) \in [0, \beta_{\max}]^N$, the infection state $\mathbf{I}(k+1)$ satisfies:

$$\|\mathbf{I}(k+1;\boldsymbol{\beta}_1(k)) - \mathbf{I}(k+1;\boldsymbol{\beta}_2(k))\| \le L_\beta \|\boldsymbol{\beta}_1(k) - \boldsymbol{\beta}_2(k)\|, \quad (34)$$

Here $\mathbf{I}(k+1;\boldsymbol{\beta}(k))$ denotes the infection state generated by the diffusion dynamics under the transmission-rate vector $\boldsymbol{\beta}(k)$, where $\boldsymbol{\beta}(k) = [\beta_1(k), \ldots, \beta_N(k)]^\top$. $L_\beta > 0$ denotes the Lipschitz constant with respect to $\hat{\boldsymbol{\beta}} = [\boldsymbol{\beta}(0), \ldots, \boldsymbol{\beta}(T-1)]^\top$, which depends on $\beta_{\max}$ and the bounded adjacency matrix $A(k)$.

Let $J(\theta)$ denote the expected discounted cumulative cost induced by the policy π_θ under the closed-loop diffusion dynamics, as defined in Problem 1.

Theorem 3: Under Assumptions 1, 5, and 6, SIR-MAPPO ensures that $\theta^t \to \theta^\star$ almost surely as $t \to \infty$, where $\theta^\star$ is a stationary point of $J(\theta)$. Moreover, under the limiting policy $\pi_{\theta^\star}$, the closed-loop diffusion metric satisfies $|C^{\theta^\star}(k) - C^\star| \to 0$ almost surely as $k \to \infty$, where $C^{\theta^\star}(k)$ denotes the diffusion metric at time k under the policy $\pi_{\theta^\star}$, and $C^\star$ denotes its limiting equilibrium value.

Proof: a) Critic convergence: Lyapunov-informed initialization and bounded TD error ensure $V_{\phi^t}(\mathbf{s}) \to V^{\pi_{\theta^\star}}(\mathbf{s})$ almost surely. b) Actor convergence: Actor gradients are unbiased under the converged critic; Robbins–Monro stochastic approximation ensures $\theta^t \to \theta^\star$ almost surely. c) Diffusion metric convergence: By Assumption 6, $\boldsymbol{\beta}^{\theta^t}(k) \to \boldsymbol{\beta}^{\theta^\star}(k)$ implies $\mathbf{I}^{\theta^t}(k) \to \mathbf{I}^{\theta^\star}(k)$. Since $C(k)$ is continuous in $\mathbf{I}(k)$, it follows that $C^{\theta^t}(k) \to C^\star$ almost surely. ■

2) Gradient Variance and ISS Robustness Enlargement: SIR-MAPPO employs a hybrid gradient (32). To analyze the stability and robustness of the learning dynamics, we first characterize the variance of the hybrid gradient and then study its effect on the closed-loop diffusion system. The following lemma and theorem formalize these properties.

Lemma 1: If the model-based estimator is unbiased and has lower variance, then $\mathrm{Var}(\nabla J_{\mathrm{hybrid}}) \leq \tau\,\mathrm{Var}(\nabla J_{\mathrm{MF}})$ with $\tau \in (0, 1)$.

To establish the impact of gradient variance on the closed-loop ISS robustness, we impose the following assumption.

Assumption 7: The transmission policy $\boldsymbol{\beta}(k) = \pi_\theta(\cdot)$ is Lipschitz continuous in θ with constant L_π. Moreover, under stochastic approximation, the asymptotic parameter variance is proportional to the gradient variance.

Theorem 4: Suppose Assumptions 2-4, 6, and 7 hold, and let the SIR dynamics satisfy the ISS bound in Theorem 2. If the variance reduction condition in Lemma 1 holds, then the steady-state disturbance term in Theorem 2 satisfies:

$$\frac{L_2 \delta_{\mathrm{hybrid}}}{\chi - L_1} \leq \sqrt{\tau}\,\frac{L_2 \delta_{\mathrm{MF}}}{\chi - L_1}, \tag{35}$$

where δ_{MF} and δ_{hybrid} denote the disturbance standard deviations entering the SIR dynamics under the MF policy gradient and the hybrid gradient estimator, respectively. Consequently, the admissible initial infection norm preserving stability enlarges by a factor $1/\sqrt{\tau}$.

Proof: Let θ_{MF} and θ_{hybrid} denote the asymptotic policy parameters obtained using the MF policy gradient and the proposed hybrid gradient estimator, respectively. The corresponding transmission policies produce transmission rates $\beta_{\mathrm{MF}} = \pi_{\theta_{\mathrm{MF}}}(\cdot)$ and $\beta_{\mathrm{hybrid}} = \pi_{\theta_{\mathrm{hybrid}}}(\cdot)$.

Under Assumption 7 and standard stochastic approximation theory, the asymptotic parameter variance scales proportionally with the gradient variance. Hence, $\mathrm{Var}(\theta_{\mathrm{hybrid}}) \leq \tau_\theta\,\mathrm{Var}(\theta_{\mathrm{MF}})$ for some constant $\tau_\theta \propto \tau$.

Since the policy is Lipschitz continuous in θ, we obtain $\mathrm{Var}(\beta_{\mathrm{hybrid}}) \leq L_\pi^2\,\mathrm{Var}(\theta_{\mathrm{hybrid}})$. Because the SIR dynamics are Lipschitz with respect to the transmission rate $\hat{\boldsymbol{\beta}}$, the disturbance magnitude entering the SIR system scales proportionally with $\sqrt{\mathrm{Var}(\hat{\boldsymbol{\beta}})}$. Consequently, the disturbance standard deviations under the hybrid and MF policies, denoted by δ_{hybrid} and δ_{MF}, satisfy $\delta_{\mathrm{hybrid}} \leq \sqrt{\tau}\,\delta_{\mathrm{MF}}$. ■

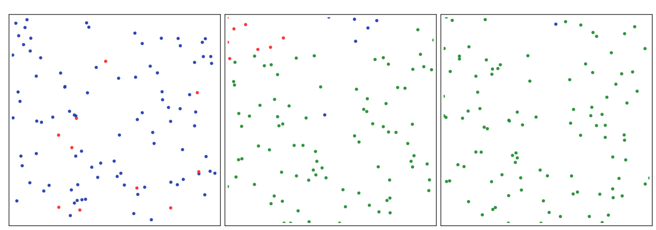
IV. Experimental Results

We construct a simulation environment with $N = 200$ robots operating in a $100 \times 100\,\mathrm{m}$ workspace over 300 discrete-time steps with a sampling period of $1\,\mathrm{s}$. Each robot follows the dynamics in (1) with maximum velocity $u_{\max} = 0.5\,\mathrm{m/s}$. Communication links are established within a radius of $20\,\mathrm{m}$ and are subject to stochastic fading modeled as a multiplicative factor $U[0.8, 1.0]$.

The information diffusion process follows the SIR model and is initialized with 15 randomly selected informed robots. The transmission rate satisfies $\boldsymbol{\beta}(k) \in [0, 0.8]$ under a global bandwidth budget $W_{\max} = 9.0$, and the recovery probability is set to $\mu = 0.1$.

The optimization objective uses cost weights $(w_c, w_e, w_r, w_m) = (2.0, 0.5, 0.3, 0.2)$, while the energy model adopts coefficients $(\alpha_1, \alpha_2, \alpha_m) = (0.1, 0.05, 0.02)$. For training, PPO is configured with clipping parameter $\epsilon = 0.2$, discount factor $\gamma = 0.95$, and hybrid gradient ratio $\eta = 0.3$. The penalty coefficients are $\lambda_V = 0.1$ and $\lambda_W = 0.5$, and the learning rates for the actor and critic are 3×10^{-4} and 10^{-3}, respectively.

Fig. 1 illustrates the initial robot deployment and the evolution of the communication topology. Robots are randomly distributed in the workspace, and communication links are formed according to the communication radius. As robots move during the coverage and information diffusion process, the interaction graph evolves dynamically while preserving network connectivity.



(a) k = 0 (b) k = 25 (c) k = 50

Fig. 1. Simulation environment comprising 200 mobile robots. At $k = 0$, 15 robots are initialized in the informed state. Blue represents susceptible (S), red represents infected (I), and green represents recovered (R).

We benchmark the SIR-MAPPO against five baselines:

- *Adaptive Random*: Random transmission rates in different ranges are selected based on the current coverage defects.
- *Greedy Broadcasting*: Prioritize bandwidth for infected nodes and allocate the remaining bandwidth to susceptible nodes.
- *Random Broadcasting*: each robot randomly selects a transmission rate from $[0, \beta_{\max}]$ at every step.
- *Periodic Gossiping*: robots transmit with a fixed probability at regular intervals.

- *Fixed-SIR*: each robot maintains a constant transmission rate throughout the mission.

To evaluate the effectiveness of the proposed framework in information dissemination, we compare the convergence behavior of the coverage defect. Fig. 2 shows the proposed policy achieves the fastest reduction of the coverage defect and reaches the smallest steady-state value, indicating rapid and complete information diffusion. In particular, the curve exhibits a steep initial decrease, demonstrating aggressive early-stage dissemination, followed by a smooth convergence phase that avoids excessive oscillations. This behavior is consistent with the theoretical analysis, where the diffusion rate is regulated to balance fast initial spreading with long-term stability of the information propagation process.

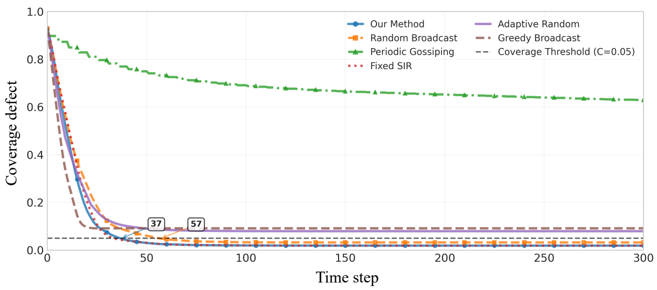


Fig. 2. The comparison of coverage defect demonstrates that our method achieves the fastest convergence to the minimum level, proving its superior diffusion efficiency.

Furthermore, to verify the learned decentralized policy adheres to the stability constraints derived from our theoretical analysis, we examine the mean transmission rate dynamics across all robots. Fig. 3 shows that SIR-MAPPO starts with an aggressive rate near $0.3 - 0.5$ in the early diffusion stage, then rapidly decreases to near zero after $k \approx 100$ as diffusion completes, remaining within the theoretical bounds.

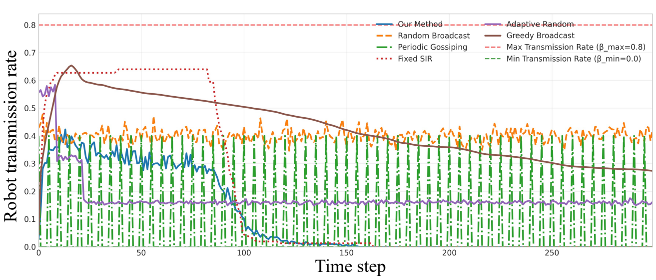


Fig. 3. The mean transmission rate across all robots remains within the theoretical bounds while adaptively decreasing as diffusion completes.

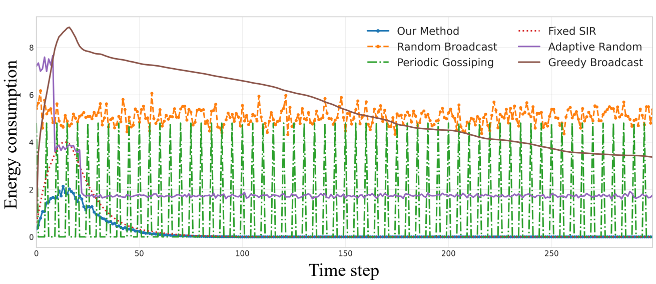


Fig. 4. The total energy consumption comparison indicates that SIR-MAPPO intelligently manages energy, achieving lower overall consumption while ensuring rapid diffusion.

Finally, Fig. 4 demonstrates the resource efficiency, which relates to the energy term in the cost function. As shown in the figure, the energy consumption is concentrated in the early diffusion stage, where a temporary increase in transmission activity accelerates information spread. Once the majority of robots become informed, the energy usage gradually decreases and stabilizes at a low maintenance level. This adaptive allocation of communication energy enables rapid initial dissemination while avoiding unnecessary long-term consumption.

V. Conclusion

In this paper, we developed SIR-MAPPO, a control-informed MARL framework for optimal information diffusion in swarm robotics. The proposed approach enables rapid and energy-efficient diffusion while mitigating redundancy and maintaining decentralized scalability. Relying only on local observations, the framework naturally scales to large swarms without communication bottlenecks, demonstrating strong potential for real-world large-scale robotic coordination in infrastructure-denied environments such as disaster response and wide-area monitoring. Future work will extend the framework to heterogeneous swarms, realistic communication conditions, and physical robotic experiments.

References

[1] W. Zhu, S. Oğuz, M. K. Heinrich, M. Allwright, M. Wahby, A. L. Christensen, E. Garone, and M. Dorigo, "Self-organizing nervous systems for robot swarms," *Science Robotics*, vol. 9, no. 96, p. eadl5161, 2024.

[2] G. Vásárhelyi, C. Virágh, G. Somorjai, T. Nepusz, A. E. Eiben, and T. Vicsek, "Optimized flocking of autonomous drones in confined environments," *Science Robotics*, vol. 3, no. 20, p. eaat3536, 2018.

[3] Y. Gao, Z. Zhou, H. Gao, S. Zhang, and Y. Ji, "Hybrid control for robot swarms to detect critical nodes in heterogeneous sensor networks," in *2025 IEEE 21st International Conference on Automation Science and Engineering*, pp. 3462–3467, 2025.

[4] B. Zhou, H. Xu, and S. Shen, "RACER: Rapid collaborative exploration with a decentralized multi-UAV system," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 1816–1835, 2023.

[5] N. Bai, Z. Duan, and Q. Wang, "Distributed optimal consensus of multi-agent systems: A randomized parallel approach," *Automatica*, vol. 159, p. 111339, 2024.

[6] H. Wei, L. Pan, H. Shao, and P. Wang, "Consensus problem of asymmetric higher-order interaction multi-agent networks," in *2025 American Control Conference*, pp. 3590–3595, 2025.

[7] S. Li, X. Zheng, and T. Jiang, "Stably global broadcasting for multi-robot flocking under multi-hop ad-hoc networks," *IEEE Internet of Things Journal*, 2025.

[8] Y. Gao, H. Gao, Y. Ji, J. Zhou, and Y. Shi, "A dual calibration framework for exploring environments using heterogeneous robot swarms," in *51st Annual Conference of the IEEE Industrial Electronics Society*, pp. 1–7, 2025.

[9] J. Zhou, D. Tian, Z. Sheng, X. Duan, and X. Shen, "Joint mobility, communication and computation optimization for UAVs in air-ground cooperative networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 3, pp. 2493–2507, 2021.

[10] Y. Lian, Y. Jiang, D. F. Opila, and C. N. Jones, "A proximal-point lagrangian-based parallelizable nonconvex solver for bilinear model predictive control," *IEEE Transactions on Control Systems Technology*, vol. 31, no. 5, pp. 2245–2260, 2023.

[11] C. De Persis, M. Rotulo, and P. Tesi, "Learning controllers from data via approximate nonlinearity cancellation," *IEEE Transactions on Automatic Control*, vol. 68, no. 10, pp. 6082–6097, 2023.

[12] Y. Wang, H. Ishii, F. Bonnet, and X. Défago, "Resilient consensus for multi-agent systems under adversarial spreading processes," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 5, pp. 3316–3331, 2022.