

Article

PPO-GAT-Follow: Graph-Attention Reinforcement Learning for Robust Robot Person Following in Dense Crowds

Xinyu Zhou ^{1,*}, Yongliang Shi ², Songhao Piao ^{1,*}  and Chao Gao ^{2,*}
¹ Multi-Agent Robot Research Center, Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China; zhouxy@stu.hit.edu.cn

² Xincheng Qihang Inc., Beijing 100084, China; yongliang.shi@bdi.tech

* Correspondence: piaosh@hit.edu.cn (S.P.); super.gao@bdi.tech (C.G.)

Highlights

What are the main findings?

- PPO-GAT-Follow improves RPF in dense crowds by explicitly modeling interactions among the follower robot, target pedestrian, and surrounding pedestrians.
- The task-oriented reward mechanism guides the policy toward a better task-level balance among target maintenance, visibility preservation, collision avoidance, rear position maintenance, proximity-aware social compliance, and action stability.

What are the implications of the main findings?

- Structured interaction representation and task-oriented reward design are effective for dense-crowd RPF under state-based geometric visibility loss with available target-relative pose estimates.
- The proposed framework shows zero-shot generalization across different crowd densities and can be integrated with localization, point cloud-based surrounding pedestrian perception and tracking, and velocity control modules under UWB-like target-relative pose input.

Abstract

Robot person following (RPF) in dense crowds requires a mobile robot to maintain an appropriate relative position with respect to a moving target while avoiding surrounding pedestrians and satisfying rear-following and social constraints. This paper proposes PPO-GAT-Follow, an interaction-aware reinforcement learning framework for dense-crowd RPF under geometric visibility loss with available target-relative pose estimates. The follower, target pedestrian, and surrounding pedestrians are represented as graph nodes, and a graph attention encoder models their local interactions. A task-oriented reward mechanism jointly accounts for target maintenance, visibility preservation, collision avoidance, proximity-aware social compliance, rear position maintenance, post-arrival stabilization, and action stability. Experiments are conducted in IR-SIM under fixed-route and random-route settings, with comparisons against MPC, DWA, SFM, and an adapted SARL baseline. In the fixed-route setting with 12 background pedestrians, PPO-GAT-Follow achieves a task success rate of 98.8% and a collision rate of 1.1%, improving task success by 10.9 percentage points over MPC. In the random-route setting at the training density, it achieves 83.1% task success and an SPL of 0.815, outperforming MPC by 18.3 percentage points in task success; at this density, it also surpasses SARL in the main task-level metrics. Zero-shot evaluations across crowd densities, together with structural and reward ablations, reward weight sensitivity analysis, tolerance shift tests, multi-seed training, and stress testing under target pose noise and heterogeneous pedestrian dynamics, further demonstrate the effectiveness and reliability of the proposed framework. Gazebo-based validation also demonstrates



Academic Editor: Francesco Aggogeri

Received: 29 June 2026

Revised: 18 July 2026

Accepted: 21 July 2026

Published: 24 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

system integration feasibility with localization, point cloud-based surrounding pedestrian perception, tracking, and UWB-like target-relative pose input. Nevertheless, visual target identification, re-identification, and perception-level occlusion recovery remain outside the scope of the present validation.

Keywords: robot person following; crowd navigation; socially aware navigation; deep reinforcement learning; graph attention network; robot learning

1. Introduction

Mobile robots are increasingly expected to operate in human-populated environments, such as campuses, hospitals, shopping malls, transportation hubs, care facilities, and public service areas. In these scenarios, navigation is no longer a purely geometric path-planning problem; instead, robots must move safely and efficiently while respecting human comfort, personal space, motion legibility, politeness, and other implicit social norms [1–4]. This line of research is commonly referred to as socially aware navigation, and its applications in pedestrian-dense environments are often studied as crowd navigation. For simplicity, this paper uses the term “crowd navigation” throughout.

Conventional crowd navigation mainly focuses on guiding a robot toward a static or predefined goal. However, many service scenarios require a robot to accompany a designated person, motivating robot person following (RPF). In RPF, a mobile robot must maintain an appropriate spatial relationship with a target pedestrian while safely interacting with surrounding pedestrians [5,6]. Typical applications include shopping mall service robots, hospital and care-facility assistive robots, logistics robots following workers, and guide robots in transportation hubs [6,7]. Unlike point-to-point navigation, RPF requires continuous relative position maintenance, target observability, and safe crowd interaction.

Although RPF already differs from standard point-to-point navigation, the problem becomes substantially more challenging when the robot operates in dense crowds. Compared with sparse environments, where the follower can often maintain a clear line of sight to the target and approach the desired following position with limited interference, dense pedestrian environments introduce frequent occlusions, restricted free space, complex human–human and human–robot interactions, and conflicts between target maintenance and collision avoidance [1,6,8]. The target may be partially or fully occluded by surrounding pedestrians, violating the full-observation assumption adopted by many RPF systems [9]. After the target reappears, visually similar pedestrians, viewpoint changes, and illumination variations may further cause re-identification ambiguity [10,11]. Therefore, RPF in dense crowds requires a careful balance between safety and target maintenance: overly aggressive following may reduce tracking error but increase collisions and close proximity exposure, whereas overly conservative behavior may improve short-term safety but lead to target loss or inefficient detours.

As illustrated in Figure 1, RPF in dense crowds involves several coupled requirements. The robot must maintain an appropriate following distance and relative angle with respect to the target, rather than simply moving toward the target’s instantaneous position. Meanwhile, it must avoid dynamic pedestrians, respect proxemic constraints and personal spaces [3,4,12], handle temporary geometric visibility loss, and generate stable real-time actions. Abrupt or oscillatory motion may reduce motion predictability and social acceptability in human-populated environments [1,4]. These coupled requirements make RPF in dense crowds substantially more challenging than standard goal-reaching navigation.

Traditional local navigation methods, which generate short-horizon velocity commands or local trajectories based on the current robot state and nearby obstacles, have been widely used for robot navigation and collision avoidance. Representative examples include the Dynamic Window Approach (DWA) [13], the Social Force Model (SFM) [14], and Model Predictive Control (MPC)-based planners [15]. These methods are generally efficient and interpretable, and recent variants have incorporated social costs, pedestrian prediction, or risk-aware optimization for human-shared environments [16–18]. However, when applied to RPF in dense crowds, their performance is often limited by hand-crafted objective functions, simplified interaction assumptions, short-horizon decision making, or dependence on accurate pedestrian prediction. Specifically, sampling-based planners may overreact to immediate obstacles, force-based methods are sensitive to parameter tuning, and optimization-based planners depend heavily on prediction accuracy and cost-function design.

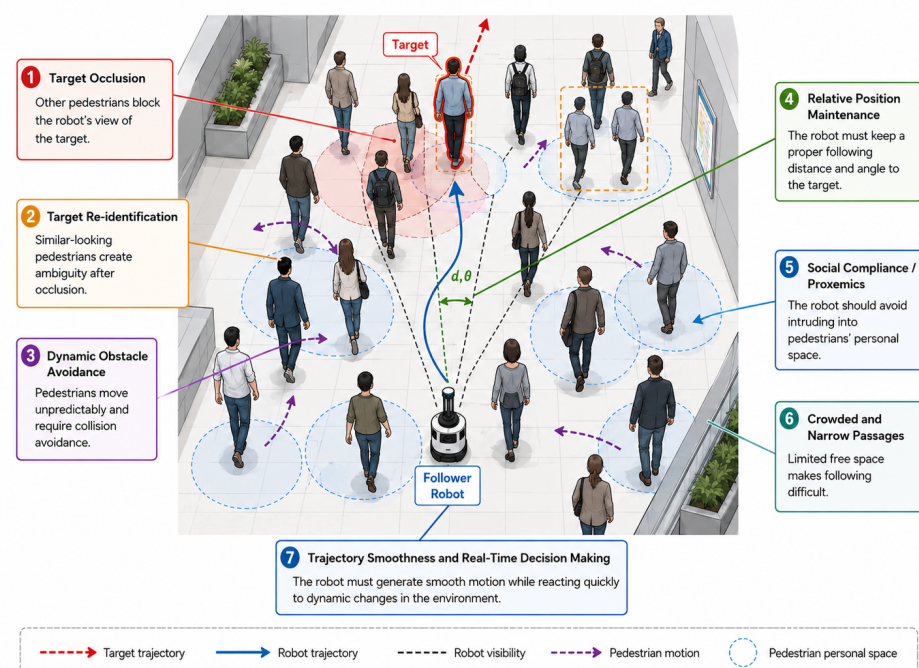


Figure 1. Key challenges of robot person following in dense pedestrian environments. The follower must track the target pedestrian while addressing occlusion, re-identification ambiguity, dynamic obstacle avoidance, relative position maintenance, social-compliance constraints, limited free space, and smooth real-time decision making. The red dashed arrow denotes the target trajectory, the blue arrow denotes the follower trajectory, black dashed lines indicate target visibility, purple arrows indicate pedestrian motion, and blue dashed ellipses represent pedestrian personal spaces.

Deep reinforcement learning (DRL), which learns decision-making policies through trial-and-error interactions with simulated environments, provides a promising alternative to hand-crafted local planning methods [19–21]. Early DRL-based crowd-navigation methods mainly focused on decentralized collision avoidance or goal-reaching navigation in dynamic crowds. Subsequent studies improved interaction reasoning by introducing attention mechanisms, graph neural networks, recurrent models, and spatial-temporal representations, enabling robots to better capture human–robot and human–human interactions [22–26].

Other studies have incorporated dynamic feasibility constraints, velocity-obstacle-guided rewards, memory-based reasoning, and reward-shaping strategies to improve safety, efficiency, and social compliance [27–32]. These advances suggest the potential of

DRL for RPF in dense crowds, where long-term interaction reasoning and multi-objective behavior balancing are both required.

Nevertheless, most existing DRL-based crowd-navigation methods are designed for goal-reaching tasks, where the robot moves toward a static destination while avoiding pedestrians. In contrast, RPF in dense crowds imposes additional task-specific constraints: the robot must continuously maintain a moving target relationship, jointly reason about the target and surrounding pedestrians, handle geometric visibility loss, and balance following accuracy with safety and proximity-aware social compliance. Although several studies have investigated robot following or guiding in crowded environments [6,7], DRL-based RPF under dense crowds with geometric visibility loss remains less explored. In particular, how to represent the interactions among the follower robot, target pedestrian, and surrounding pedestrians, and how to design rewards that jointly encourage target maintenance, visibility preservation, collision avoidance, proximity-aware social compliance, and stable motion remain open problems.

To address these challenges, this paper proposes PPO-GAT-Follow, an interaction-aware DRL framework for RPF in dense crowds. The proposed framework combines graph-based interaction modeling with task-oriented reward shaping to balance target maintenance, collision avoidance, proximity-aware social compliance, and stable following behavior. The main contributions are summarized as follows:

- We formulate dense-crowd RPF as an interaction-aware crowd-navigation problem that jointly considers target maintenance, local collision avoidance, and proximity-aware social compliance. A GAT-based representation is developed to model the relationships among the follower robot, target pedestrian, and surrounding pedestrians.
- We design a task-oriented reward mechanism that combines terminal rewards with step-wise dense shaping. The reward balances target maintenance, visibility preservation, collision avoidance, rear position maintenance, proximity-aware social compliance, post-arrival stabilization, and action stability.
- We conduct extensive IR-SIM experiments under fixed-route and random-route settings with different crowd densities. Comparisons with MPC, DWA, and an adapted SARL baseline, together with structural/reward ablations and reliability analyses, demonstrate the effectiveness and robustness of the proposed framework. SFM is retained as a force-based reference and analyzed separately due to its structural limitations under strict RPF completion conditions.
- We construct a Gazebo-based system integration validation pipeline combining localization, UWB-like target-relative pose input, point cloud-based surrounding pedestrian perception and tracking, and PPO-GAT-Follow control. This validation confirms the integration feasibility of the learned policy, while visual target identification, re-identification, and perception-level occlusion recovery remain outside the current scope.

The remainder of this paper is organized as follows. Section 2 reviews related work on DRL-based crowd navigation, traditional local planning methods, and RPF. Section 3 presents the proposed PPO-GAT-Follow framework, including the task formulation, graph-based state representation, GAT encoder, reward design, and PPO-based policy learning. Section 4 describes the experimental settings, including the simulation environment, episode randomization, visibility and occlusion handling, baseline methods, evaluation metrics, and ablation settings. Section 5 reports fixed-route and random-route performance, density generalization, following quality analysis, ablation studies, robustness and reliability analyses, qualitative trajectory comparison, and Gazebo-based system integration validation. Section 6 discusses the main findings, baseline interpretation, system integration implications, limitations, and future work. Section 7 concludes this paper.

2. Related Work

2.1. DRL-Based Crowd Navigation

Crowd navigation aims to enable mobile robots to move safely, efficiently, and acceptably in human-populated environments. Beyond geometric collision avoidance, it requires robots to consider human comfort, personal space, motion naturalness, and social acceptability [3,33]. Recent surveys show that this field has evolved from reactive obstacle avoidance and hand-crafted social rules toward human-centered motion generation, context-aware planning, and learning-based decision making [1,2,34].

Deep reinforcement learning (DRL) has been widely studied for crowd navigation because it can learn interaction-aware policies from simulated experience [35]. Early methods, such as CADRL and GA3C-CADRL, learned decentralized collision avoidance policies for multi-agent navigation [19,20]. Later studies improved crowd-level interaction reasoning through attention mechanisms, graph neural networks, recurrent models, and spatial-temporal representations. Representative examples include SARL for attention-based crowd reasoning [21], RGL for graph-based relation modeling [22], DS-RNN for spatial-temporal dependency modeling [23], and intention-aware methods for incorporating pedestrian intention prediction [26]. Other studies have explored velocity-obstacle-guided rewards, memory-based reasoning, and reward shaping to improve safety, efficiency, and social compliance [29–32].

However, most existing DRL-based crowd-navigation methods are still designed for goal-reaching tasks, where the robot moves toward a fixed destination while avoiding pedestrians. In contrast, RPF requires the robot to maintain a continuous spatial relationship with a moving target while avoiding surrounding pedestrians and respecting social constraints. Although recent work has begun to formulate socially aware RPF as a benchmark problem [6], RPF in dense crowds remains less explored than fixed-goal crowd navigation. Further investigation is therefore needed to balance target maintenance, occlusion handling, collision avoidance, rear position maintenance, and social compliance when robots follow a target pedestrian in dense crowds.

2.2. Traditional Local Planning Methods

Traditional local planning and obstacle-avoidance methods remain important baselines for mobile robot navigation because they are interpretable, computationally efficient, and easy to deploy. Representative categories include sampling-based reactive planners, force-based interaction models, and optimization-based planners. The Dynamic Window Approach (DWA) selects feasible velocity commands under robot kinematic and dynamic constraints using hand-crafted objectives such as heading alignment, obstacle clearance, and velocity preference [13,36]. Although socially aware variants have incorporated human-related costs for shared environments [16], DWA remains a short-horizon reactive method and tends to exhibit limited progress, local oscillations, or freezing-like behaviors in dense pedestrian interactions [37,38].

Force-based methods provide an intuitive formulation for navigation and collision avoidance. Artificial potential fields model goals and obstacles as attractive and repulsive forces [39], while the Social Force Model (SFM) describes pedestrian motion using human-inspired interaction forces [14]. These methods are lightweight and can encode interpersonal distance or social zone effects, but they are sensitive to parameter tuning and tend to suffer from local minima, oscillatory motion, or unrealistic interactions in dense crowds [40–42]. In RPF scenarios, force-based formulations may also overemphasize either attraction to the target or repulsion from surrounding pedestrians, making stable and safe following difficult.

Optimization-based planners, especially Model Predictive Control (MPC), formulate local navigation as a finite-horizon optimization problem and can explicitly incorporate robot dynamics, collision avoidance constraints, smoothness terms, and task-related costs [43]. Recent studies have combined MPC with control barrier functions, residual models, risk-aware optimization, or learning-based pedestrian prediction to improve safety and interaction awareness in crowded environments [15,17,18,44,45]. However, MPC-based methods often depend on accurate prediction, carefully designed cost functions, appropriate horizon selection, and tractable online optimization.

Overall, traditional local planning methods provide interpretable and deployable baselines, but they are mainly designed for fixed-goal navigation or local obstacle avoidance. RPF in dense crowds imposes additional requirements, including maintaining a desired relative position with a moving target, handling temporary occlusions, avoiding surrounding pedestrians, and respecting social spacing constraints. These coupled requirements are difficult to address using only hand-crafted rules or short-horizon optimization, motivating learning-based methods with interaction-aware representations.

2.3. Robot Person Following

Existing RPF studies have investigated different following configurations, including rear following, side-by-side following, frontal following, and adaptive following position selection [46–48]. Rear following is widely adopted because it is less intrusive and easier to maintain, whereas side-by-side following may provide more natural interaction but can introduce stronger conflicts with pedestrians and obstacles. To improve comfort and social acceptability, several studies have incorporated proxemic constraints and socially compliant behaviors into RPF systems [46,49,50].

From the motion-planning perspective, RPF methods have evolved from relative position tracking to local planning, optimization, and learning-based decision making. DWA-based methods generate feasible velocity commands under following and collision avoidance constraints [51], SFM-based methods model attractive and repulsive interactions among the robot, target, and surrounding pedestrians [52,53], and MPC-based methods optimize finite-horizon robot actions while considering target prediction, robot dynamics, and social space requirements [47,54]. Reinforcement learning-based methods have also been explored for human-following and follow-ahead tasks to address obstacle avoidance, occlusion avoidance, and frequent target direction changes [7,55,56].

Despite these advances, RPF in dense crowds remains insufficiently explored, as many existing studies are still evaluated in sparse or small-crowd scenarios [6]. In dense crowds, the robot must simultaneously maintain the target relationship, avoid surrounding pedestrians, handle target occlusion and re-identification, and respect social spacing constraints [9,11]. These coupled requirements motivate interaction-aware decision-making methods that balance target maintenance, safety, efficiency, and social compliance.

3. Materials and Methods

We formulate RPF as an interaction-aware sequential decision-making problem and propose PPO-GAT-Follow, a PPO-based DRL framework that uses GAT to model interactions among the follower, leader, and surrounding pedestrians. As shown in Figure 2, graph-structured observations and ego-centric task features are encoded by the GAT and ego-feature branches, respectively, and then fused into a 128-dimensional latent representation. This representation is fed into the PPO actor–critic network to generate continuous planar velocity commands, while rollout transitions and rewards are used for policy updates.

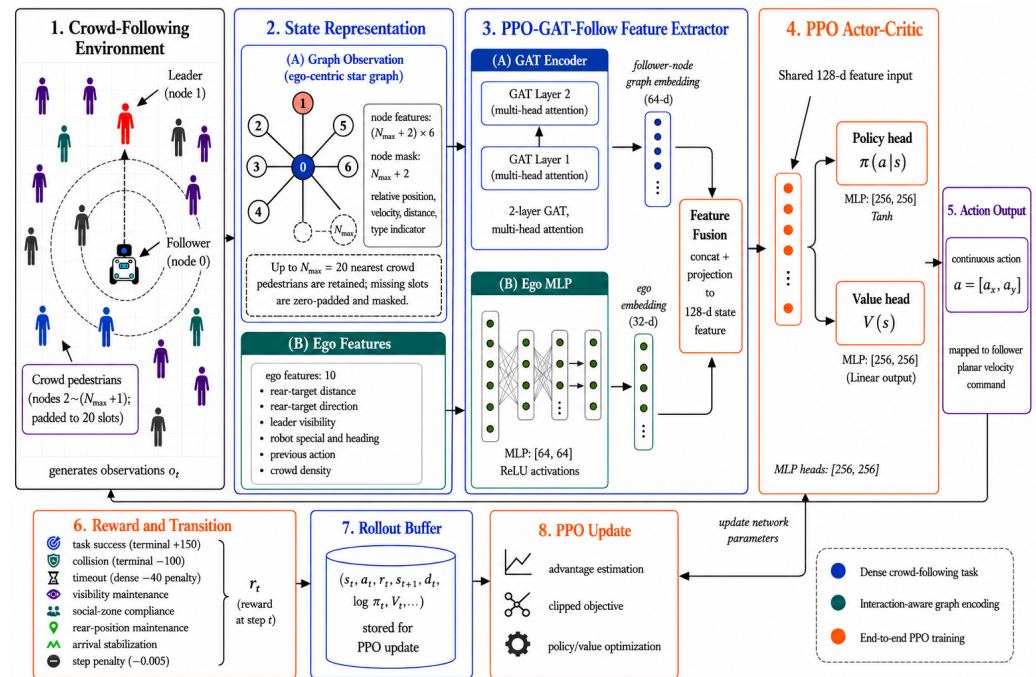


Figure 2. Overview of the proposed PPO-GAT-Follow framework. Graph-structured observations and ego-centric task features are encoded by separate branches and fused for PPO-based continuous control. The graph uses a follower-centered star topology with zero-padded background pedestrian slots and node masks to support variable crowd sizes.

The remainder of this section introduces the task formulation, graph-based state representation, GAT encoder, reward design, PPO learning process, and training configuration.

3.1. Crowd-Following Task and Environment

The RPF task is defined in a two-dimensional crowd environment containing a follower robot, a designated target pedestrian, referred to as the leader in this paper, and multiple surrounding pedestrians, referred to as background pedestrians, as shown in Figure 3. The follower is required to accompany the leader until goal arrival while maintaining a rear-following position and avoiding collisions. Unlike conventional goal-reaching navigation, this task imposes both collision avoidance constraints with respect to background pedestrians and relative position constraints with respect to a moving leader.

Let $p_t^r = [x_t^r, y_t^r]^\top$ and $p_t^l = [x_t^l, y_t^l]^\top$ denote the positions of the follower robot and leader at time step t , respectively. The leader heading is denoted by θ_t^l , and the desired rear-following distance is d_b . As shown in Figure 3a, the desired following reference is defined behind the leader along the opposite direction of its heading:

$$p_t^{\text{ref}} = p_t^l - d_b \begin{bmatrix} \cos \theta_t^l \\ \sin \theta_t^l \end{bmatrix}, \quad (1)$$

where $d_b = 1.0$ m in all experiments. The following error is defined as

$$e_t = \|p_t^r - p_t^{\text{ref}}\|_2. \quad (2)$$

This error evaluates whether the follower reaches the desired rear reference rather than merely staying close to the leader.

To distinguish rear following from simple distance keeping, the leader heading unit vector is defined as

$$u_t^l = \begin{bmatrix} \cos \theta_t^l \\ \sin \theta_t^l \end{bmatrix}. \quad (3)$$

The rear half-space indicator is then given by

$$\chi_t^{\text{rear}} = \mathbb{I}\left((p_t^l - p_t^r)^\top u_t^l > 0\right), \quad (4)$$

which is true when the follower lies behind the leader with respect to the leader heading. For reward shaping, we further define a continuous rear position score:

$$\tilde{b}_t = 0.6 \max\left(0, 1 - \frac{e_t}{2e_{\max}}\right) + 0.4 \text{clip}\left(\frac{(p_t^l - p_t^r)^\top u_t^l}{d_b + \epsilon}, 0, 1\right), \quad (5)$$

where $e_{\max} = 0.45$ m is the following position tolerance and ϵ is a small constant for numerical stability. The score $\tilde{b}_t \in [0, 1]$ jointly measures rear reference accuracy and rear alignment. It is used only for reward shaping during training, rather than as a hard success condition.

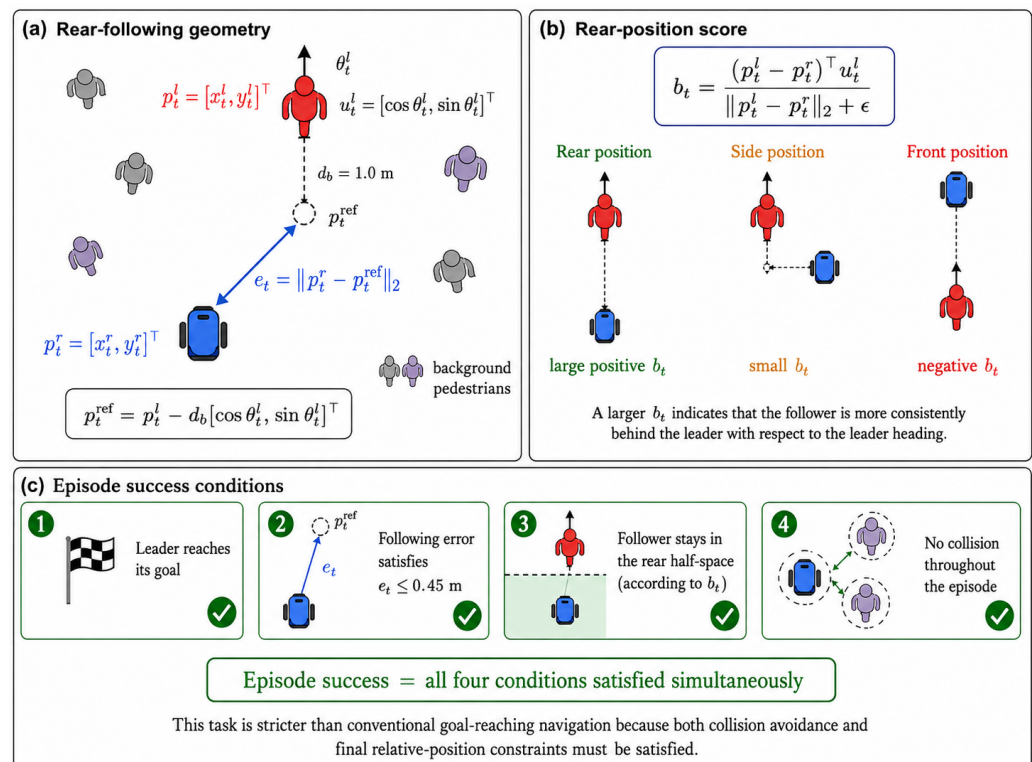


Figure 3. Visualization of the RPF task. (a) Rear-following geometry. (b) Rear position score $\tilde{b}_t$ for distinguishing rear following from side or frontal positioning. (c) Episode success conditions, including leader arrival, bounded following error, rear half-space satisfaction, low follower speed, and collision-free navigation.

As shown in Figure 3c, an episode is considered successful only if the leader reaches its goal, the follower satisfies $e_t \leq 0.45$ m and $\chi_t^{\text{rear}} = 1$, the follower speed is lower than $v_{\text{stop}} = 0.18$ m/s, and no collision occurs throughout the episode. This definition is stricter than conventional point-to-point navigation because the robot must stop at a geometrically meaningful rear-following position after the leader arrives.

3.2. MDP Formulation and Learning Objective

The RPF problem is formulated as a Markov decision process (MDP), denoted by $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. At each time step t , the follower observes a state $s_t \in \mathcal{S}$, selects an action $a_t \in \mathcal{A}$ according to the policy $\pi_\theta(a_t|s_t)$, receives a reward $r_t = \mathcal{R}(s_t, a_t)$, and transitions

to the next state s_{t+1} according to $\mathcal{P}(s_{t+1}|s_t, a_t)$. The learning objective is to maximize the expected discounted return:

$$J(\theta) = \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^T \gamma^t r_t \right], \quad (6)$$

where γ is the discount factor and T is the maximum episode length.

The follower is modeled as a holonomic mobile robot. The policy outputs a two-dimensional normalized action:

$$a_t = [a_t^x, a_t^y]^\top, \quad a_t^x, a_t^y \in [-1, 1]. \quad (7)$$

The action is then mapped to a planar velocity command in the world frame:

$$v_t^x = a_t^x v_{\max}, \quad v_t^y = a_t^y v_{\max}, \quad (8)$$

where $v_{\max} = 1.2$ m/s denotes the maximum velocity component used for policy execution and benchmarking. In implementation, the follower is controlled by an omni-directional robot model. The world-frame velocity command is converted into a body-frame command, and the robot heading is updated to align with the motion direction. This implementation preserves holonomic motion capability while remaining compatible with the simulator interface.

3.3. Graph-Based State Representation

RPF in dense crowds requires the robot to jointly reason about the leader, surrounding background pedestrians, and possible target occlusions. Therefore, PPO-GAT-Follow adopts a graph-structured observation instead of a flattened state vector. At time step t , the graph is defined as

$$\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}), \quad (9)$$

where $\mathcal{V}_t$ is a fixed-size node set and $\mathcal{E}$ denotes a bidirectional star topology centered on the follower.

The graph contains $N_{\max} + 2$ nodes. Node 0 represents the follower, node 1 represents the leader, and nodes $2, \dots, N_{\max} + 1$ represent background pedestrians. If more than $N_{\max}$ pedestrians are observed, only the nearest $N_{\max}$ are retained; otherwise, unused slots are zero-padded and masked out. This fixed-size design enables the same network architecture to handle different crowd densities. In our implementation, $N_{\max} = 20$, producing a 22×6 node feature matrix and a 22-dimensional node mask. For example, in the medium-density benchmark with 12 background pedestrians, only the corresponding pedestrian nodes are activated, while the remaining slots are padded and masked.

For each valid node j , the node feature is defined in the follower-centered coordinate frame:

$$x_t^j = \left[\frac{\Delta p_{t,x}^j}{s_c}, \frac{\Delta p_{t,y}^j}{s_c}, \frac{v_{t,x}^j}{v_{\max}}, \frac{v_{t,y}^j}{v_{\max}}, \frac{\|\Delta p_t^j\|_2}{s_c}, \tau_j \right], \quad (10)$$

where $\Delta p_t^j = p_t^j - p_t^r$, s_c is the scene normalization scale, v_t^j is the velocity of node j , and τ_j is the node-type indicator. The type indicator is set to 0.0 for the follower, 0.5 for the leader, and 1.0 for background pedestrians.

In addition to graph node features, an ego feature vector is used to encode follower-specific task information. Let p_t^{tgt} denote the active following target:

$$p_t^{\text{tgt}} = \begin{cases} p_t^{\text{ref}}, & \text{if the leader is visible,} \\ \hat{p}_t^{\text{ref}}, & \text{otherwise,} \end{cases} \quad (11)$$

where $\hat{p}_t^{\text{ref}}$ is the last known rear-following reference maintained during geometric target occlusion. The ego feature is defined as

$$g_t = \left[\frac{\|p_t^r - p_t^{\text{tgt}}\|_2}{s_c}, \cos \psi_t, \sin \psi_t, o_t, \frac{\|v_t^r\|_2}{v_{\max}}, \cos \theta_t^r, \sin \theta_t^r, a_{t-1}^x, a_{t-1}^y, \frac{n_t}{N_{\max}} \right], \quad (12)$$

where ψ_t is the direction from the follower to p_t^{tgt} , $o_t \in \{0, 1\}$ indicates leader visibility, θ_t^r is the follower heading, a_{t-1} is the previous action, and $n_t = \min(|\mathcal{P}_t|, N_{\max})$ is the number of valid background pedestrian nodes. Here, $\mathcal{P}_t$ denotes the set of observed background pedestrians at time step t . When the leader is visible, the first ego feature corresponds to the normalized following error. When the leader is occluded, it represents the distance to the last known rear-following reference, providing a simple target memory cue. Invalid background pedestrian slots are zero-padded and suppressed by the node mask M_t before GAT encoding.

In the main IR-SIM experiments, the leader-related state used in the observation is obtained from the simulator. When the leader is visible according to the geometric line-of-sight test, the current rear-following reference is used. When the leader is geometrically occluded, the last known rear-following reference is maintained as a simple memory cue. This design isolates the decision-making problem of following and collision avoidance, but it does not model visual detection uncertainty, re-identification failure, or tracking drift.

3.4. Graph Attention Encoder and Feature Fusion

The graph observation is encoded using a graph attention network (GAT). The edge set follows a follower-centered bidirectional star topology:

$$\mathcal{E} = \{(0, j), (j, 0) \mid j = 1, \dots, N_{\max} + 1\}, \quad (13)$$

which connects the follower with the leader and all valid background pedestrian nodes. This topology reflects the follower-centered decision process and enables the policy to assign different importance to neighboring agents.

For the l -th GAT layer, the attention coefficient from node j to node i is computed as

$$\alpha_{ij}^{(l)} = \frac{\exp\left(\text{LeakyReLU}\left(a_l^\top \left[W_l h_i^{(l)} \parallel W_l h_j^{(l)}\right]\right)\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(\text{LeakyReLU}\left(a_l^\top \left[W_l h_i^{(l)} \parallel W_l h_k^{(l)}\right]\right)\right)}, \quad (14)$$

where $h_i^{(l)}$ is the hidden feature of node i , W_l is a learnable linear projection, a_l is the attention vector, $\parallel$ denotes concatenation, and $\mathcal{N}(i)$ is the neighborhood of node i . The node feature is updated by

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} W_l h_j^{(l)}\right), \quad (15)$$

where $\sigma(\cdot)$ denotes a nonlinear activation function. Multi-head attention is used in the intermediate layer, and the final follower node embedding is taken as the graph-level interaction representation.

As shown in Figure 2, the GAT branch extracts follower-centered interaction features, while the ego MLP branch encodes task-specific ego features. The two embeddings are concatenated and projected into a latent representation:

$$z_t = \phi_{\text{proj}}\left(\left[\phi_{\text{GAT}}(X_t, M_t)_0 \parallel \phi_{\text{ego}}(g_t)\right]\right), \quad (16)$$

where X_t is the node feature matrix, M_t is the node mask, $\phi_{\text{GAT}}(\cdot)_0$ denotes the follower node embedding, $\phi_{\text{ego}}(\cdot)$ denotes the ego MLP, and $\phi_{\text{proj}}(\cdot)$ projects the fused feature into a 128-dimensional latent representation. The network configuration is summarized in Table 1.

Table 1. Architecture of the PPO-GAT-Follow feature extractor.

Component	Value	Description
Graph nodes	$N_{\max} + 2$	1 follower, 1 leader, and $N_{\max}$ background pedestrian slots
Node feature dimension	6	Relative position, velocity, distance, and type indicator
Ego feature dimension	10	Following error, target direction, visibility, velocity, heading, previous action, and density
GAT layers	2	Graph attention encoder
GAT heads	4	Multi-head attention in the intermediate layer
GAT hidden dimension	64	Hidden feature dimension
GAT output dimension	64	Follower node graph embedding
Ego hidden dimension	32	Output dimension of the ego MLP branch
Fused feature dimension	128	Input feature dimension for PPO policy and value networks
Policy/value MLP	[256, 256]	Two-layer MLP for actor and critic branches

3.5. Reward Design

The reward is designed to balance task completion, target maintenance, visibility preservation, collision avoidance, proximity-aware social compliance, rear position maintenance, and post-arrival stabilization. Regular steps use dense shaping rewards, whereas task success, collision, and timeout are handled by terminal conditions. The overall reward is defined as

$$r_t = \begin{cases} +150, & \text{if the task is successful,} \\ -100, & \text{if a collision occurs,} \\ r_t^{\text{dense}} - 40, & \text{if the episode times out,} \\ r_t^{\text{dense}}, & \text{otherwise.} \end{cases} \quad (17)$$

For success and collision steps, only the corresponding terminal reward is returned. For timeout steps, dense shaping is first computed and then penalized by -40 .

The dense reward is computed as

$$r_t^{\text{dense}} = -5.0 \frac{\|p_t^r - p_t^{\text{tgt}}\|_2}{s_c} + 4.0 \frac{\|p_{t-1}^r - p_{t-1}^{\text{tgt}}\|_2 - \|p_t^r - p_t^{\text{tgt}}\|_2}{s_c} + r_t^{\text{vis}} + r_t^{\text{zone}} + r_t^{\text{prog}} + r_t^{\text{behind}} + r_t^{\text{arr}} - 0.005. \quad (18)$$

The first two terms penalize the distance to the active following target and reward progress toward it, while the constant term acts as a small step penalty.

The visibility reward encourages maintaining geometric leader visibility:

$$r_t^{\text{vis}} = \begin{cases} +0.25, & \text{if the leader is visible,} \\ -0.10, & \text{otherwise.} \end{cases} \quad (19)$$

The proximity penalty discourages excessively close interaction with the leader and background pedestrians. To avoid ambiguity caused by different agent radii, we use signed surface clearance rather than raw center-to-center distance. Let ρ_r , ρ_l , and ρ_i denote the radii of the follower, leader, and background pedestrian i , respectively. The corresponding clearances are defined as

$$c_t^{r,l} = \|p_t^r - p_t^l\|_2 - (\rho_r + \rho_l), \quad (20)$$

and

$$c_t^{r,i} = \|p_t^r - p_t^i\|_2 - (\rho_r + \rho_i). \quad (21)$$

A negative clearance indicates geometric overlap, whereas a positive clearance represents the free space between the agent boundaries. The proximity penalty is then defined as

$$r_t^{\text{zone}} = -1.0\Delta t \mathbb{I}(c_t^{r,l} < 0.8) - 4.0\Delta t \mathbb{I}\left(\min_i c_t^{r,i} < 0.55\right). \quad (22)$$

The two terms penalize close-range interaction with the leader and surrounding pedestrians, respectively.

The leader progress reward is applied before the leader reaches its goal:

$$r_t^{\text{prog}} = \mathbb{I}(\neg \text{leader arrived}) \cdot 2.0 \frac{d_{t-1}^{l,g} - d_t^{l,g}}{s_c}, \quad (23)$$

where $d_t^{l,g}$ is the distance between the leader and its goal. The rear-following reward is based on the continuous rear position score $\tilde{b}_t$ defined in Equation (5):

$$r_t^{\text{behind}} = \begin{cases} 0.8 \tilde{b}_t, & \text{if the leader has arrived,} \\ 0.4 \tilde{b}_t, & \text{otherwise.} \end{cases} \quad (24)$$

After the leader arrives, an additional post-arrival reward encourages the follower to settle near the rear-following reference:

$$r_t^{\text{arr}} = \mathbb{I}(\text{leader arrived}) \left[1.0 \left(1 - \frac{\|p_t^r - p_t^{\text{tgt}}\|_2}{s_c} \right) - 4.0 \frac{\|v_t^r\|_2}{v_{\max}} - 1.5 \|a_t\|_2^2 \right]. \quad (25)$$

The velocity and action-magnitude penalties in Equation (25) are applied only after leader arrival to suppress oscillation near the final following position.

The reward weights were selected according to task priority, reward-scale normalization, and preliminary empirical tuning. First, terminal rewards are assigned larger magnitudes than dense shaping terms because task success, collision avoidance, and time-out prevention represent the primary episode-level outcomes. Specifically, successful completion is strongly rewarded, while collision and timeout are penalized to discourage unsafe or incomplete following behaviors. Second, dense shaping terms are scaled according to the scene scale and simulation time step when applicable, so that no single term dominates training only because of its numerical range. For example, distance-related terms are normalized by the scene scale s_c , progress terms are also scale-normalized, and proximity penalties are multiplied by the simulation time step Δt . Third, the relative weights of the dense terms were tuned through preliminary training runs to balance target maintenance, visibility preservation, proximity-aware social compliance, rear position maintenance, post-arrival stabilization, and action stability. These manually selected values are not claimed to be globally optimal; therefore, we further conduct reward weight sensitivity analysis in Section 5.5 to examine whether the learned policy remains stable under moderate reward weight perturbations.

Reward ablations are conducted by removing the visibility term, the proximity penalty term, or the rear-following term while keeping the evaluation metrics unchanged. This allows us to examine whether each reward component improves following performance rather than merely changing the evaluation criterion. The influence of this tolerance choice is further examined through tolerance shift evaluation without retraining in Section 5.5.

3.6. PPO-Based Policy Learning

The policy is trained using Proximal Policy Optimization (PPO), which stabilizes policy updates through a clipped surrogate objective. The probability ratio between the current and old policies is defined as

$$\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}. \quad (26)$$

The clipped policy objective is then given by

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t[\min(\rho_t(\theta)\hat{A}_t, \text{clip}(\rho_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)], \quad (27)$$

where $\hat{A}_t$ is the estimated advantage and ϵ is the clipping coefficient. The final loss combines the clipped policy loss, value loss, and entropy regularization:

$$L(\theta) = -L^{\text{CLIP}}(\theta) + c_v L^V(\theta) - c_e \mathcal{H}(\pi_\theta), \quad (28)$$

where c_v and c_e denote the value loss and entropy coefficients, respectively.

3.7. Training Configuration

PPO-GAT-Follow is implemented using Stable-Baselines3 (version 2.8.0) [57] with a multi-input policy. The graph observation is encoded by the custom GAT feature extractor, and the fused latent representation is shared by the policy and value networks. The main training parameters are summarized in Table 2.

Table 2. Training parameters of PPO-GAT-Follow.

Parameter	Value
Learning rate	3×10^{-4}
Rollout steps	512
Batch size	256
Optimization epochs per update	10
Discount factor γ	0.99
GAE parameter λ	0.95
Clip range	0.2
Entropy coefficient	0.01
Policy/value MLP	[256, 256]
Parallel environments	8
Total training steps	2,000,000
Main training seed	42
Evaluation frequency	50,000 steps
Validation episodes during training	10
Checkpoint frequency	50,000 steps

Additional independent training seeds are used in the training seed reliability analysis in Section 5.5.

4. Experimental Settings

4.1. Simulation Environment

The crowd-following benchmark is implemented in the IR-SIM [58] multi-agent simulator using a 20 m $\times$ 20 m obstacle-free workspace and a 0.1 s time step. Each episode is limited to 700 steps (70 s) and terminates upon follower collision; incomplete episodes are counted as timeouts.

As summarized in Table 3, the leader and background pedestrians are controlled by IR-SIM's RVO module, while the follower is controlled by the evaluated planner or policy. All methods are therefore tested under identical pedestrian dynamics.

Table 3. Agent configuration in the RPF environment.

Agent Type	ID	Radius	Behavior and Parameters
Follower	0	0.35 m	Controlled by the tested planner or learned policy; each velocity component capped at 1.2 m/s RVO controller; $v_{\max} = 1.0$ m/s; acceleration limit 1.0 m/s ² ; avoidance factor 2.0; goal threshold 0.3 m RVO loop motion between predefined waypoints, producing crossing and occlusion patterns
Leader	1	0.28 m	
Background pedestrian	$2 \sim (N_c + 1)$	0.25 m	

Three crowd-density settings are constructed by varying the number of background pedestrians while keeping the follower–leader pair unchanged, as shown in Table 4. The medium-density setting with 12 background pedestrians is used as the default benchmark, while the low- and high-density settings are used for density generalization analysis and stress testing.

Table 4. Crowd-density settings used in the simulation experiments.

Setting	Background Pedestrians	Total Agents	Purpose
Low density	6	8	Low-density evaluation
Medium density	12	14	Default benchmark
High density	20	22	Dense-crowd stress testing

4.2. Episode Randomization

To ensure reproducibility and reduce overfitting to a specific crowd configuration, each episode uses deterministic randomization controlled by a base seed and the episode index:

$$\text{seed}_{\text{ep}} = \text{base_seed} + 1,000,003 \times \text{episode_idx}. \quad (29)$$

Two evaluation regimes are considered. In the *fixed-route* setting, the leader and follower start positions remain unchanged. In the *random-route* setting, the leader start and goal positions are resampled per episode, with path lengths of 6–16 m and a 2 m boundary margin. The follower is initialized approximately 1.8 m behind the leader. In both settings, each background pedestrian receives an independent uniform offset in $[-0.25, 0.25]$ m along each axis.

The same randomization protocol is used for training and testing. During testing, each method is evaluated using 10 base seeds with 200 episodes per seed.

4.3. Visibility and Occlusion Handling

Leader visibility is determined by a geometric line-of-sight test between the follower and leader centers against background pedestrian disks enlarged by a 0.05 m margin. If blocked, the leader is marked invisible. Classical controllers search around the last known reference, and an occlusion is considered successfully handled if the leader is re-acquired within 5 s. Learning-based methods use no hand-crafted search rule and act from observations containing the last known leader state.

The IR-SIM occlusion model is state-based and geometric. The simulator provides the leader state, while occlusion affects only the visibility indicator and the use of the last-known rear-following reference. Thus, the benchmark evaluates decision making

under geometric line-of-sight loss with available target state information, but not visual detection, re-identification, tracking drift, or perception-pipeline failure.

The main visibility and task completion parameters are summarized in Table 5.

Table 5. Visibility and task completion parameters.

Parameter	Value
Visibility margin	0.05 m
Search time limit	5.0 s
Desired rear-following distance	1.0 m
Following position tolerance	0.45 m
Maximum episode length	700 steps

4.4. Baseline Methods

PPO-GAT-Follow is compared with three representative classical navigation baselines and one adapted learning-based baseline:

- *MPC*: A model-predictive local planner that optimizes finite-horizon commands using tracking and collision avoidance costs.
- *DWA*: A sampling-based local planner that selects feasible velocities based on heading, speed, and obstacle clearance.
- *SFM*: A social force-based method combining target attraction and pedestrian repulsion. Since standard SFM does not explicitly enforce precise rear reference positioning, rear half-space satisfaction, or post-arrival stopping, it is used as a force-based reference rather than the primary baseline for performance claims.
- *SARL*: A representative attention-based DRL crowd-navigation baseline adapted to RPF by replacing its original static navigation goal with the dynamically updated rear-following reference behind the leader. The adapted SARL baseline retains its original joint-state observation, discrete action formulation, DQN-based value learning, ϵ -greedy exploration, and discomfort-aware value term. For fair comparison, the RPF task definition, crowd configuration, random leader protocol, training budget, discount factor, learning rate, evaluation seeds, and major reward scales are aligned with those of PPO-GAT-Follow, while SARL-specific algorithmic components are kept unchanged.

MPC is treated as the strongest classical task-oriented baseline because it explicitly incorporates tracking and collision avoidance costs and achieves non-trivial RPF success. Accordingly, the main comparisons focus on MPC and SARL, while DWA and SFM serve as representative sampling-based and force-based references.

4.5. Evaluation Metrics

The evaluation metrics assess task completion, safety, target visibility, search ability, efficiency, and social compliance.

- *Task Success Rate*: The percentage of episodes in which the leader reaches the goal, the follower reaches the required rear-following position, and no follower collision occurs. Since task success requires collision-free navigation in our setting, this metric jointly reflects task completion, rear position satisfaction, and collision-free following.
- *Collision Rate*: The percentage of episodes in which the follower collides with the leader or any background pedestrian.
- *SPL*: Success weighted by path length. For episode i , it is computed as

$$\text{SPL}_i = S_i \frac{L_i^*}{\max(L_i, L_i^*)}, \quad (30)$$

where $S_i \in \{0, 1\}$ is the task success indicator, L_i is the actual follower path length, and L_i^* is the reference shortest path length.

- *Target Visibility Ratio*: The proportion of time steps in which the leader is visible to the follower.
- *Search Success Rate*: The proportion of occlusion events in which the leader is re-acquired within 5 s.
- *Path Length*: The total trajectory length traveled by the follower during an episode.
- *Navigation Time*: The elapsed time of an episode before task success, collision, or timeout.
- *Close Proximity Time*: The accumulated time for which the follower–leader surface clearance is below 0.8 m or the minimum follower–pedestrian clearance is below 0.55 m:

$$T_{\text{prox}} = \Delta t \sum_{t=1}^T \mathbb{I} \left(c_t^{r,l} < 0.8; \vee; \min_i c_t^{r,i} < 0.55 \right). \quad (31)$$

The logical OR avoids double counting when both conditions hold. This metric measures close-range exposure and should be interpreted jointly with task success and collision rate, since rear following inherently requires proximity to the leader.

Path length and navigation time are interpreted together with task success and collision rate, because failed or prematurely terminated episodes may also produce shorter paths or shorter durations. Close proximity time is used as the quantitative metric corresponding to the proximity-related social zone constraint.

4.6. Fixed-Route Experiment

The fixed-route experiment evaluates performance along a repeatable leader trajectory. The leader and follower start positions remain fixed, while background pedestrians receive small initial perturbations. As shown in Figure 4a, the follower starts at (2.7, 2.7, 0), and the leader moves from (4.0, 4.0, 0.785) to (17.3, 17.3, 0.785). The random leader option is disabled during training and testing. PPO-GAT-Follow is trained for 2 million steps and evaluated over 10 seeds with 200 episodes each. The detailed training and testing settings for the fixed-route experiment are summarized in Table 6.

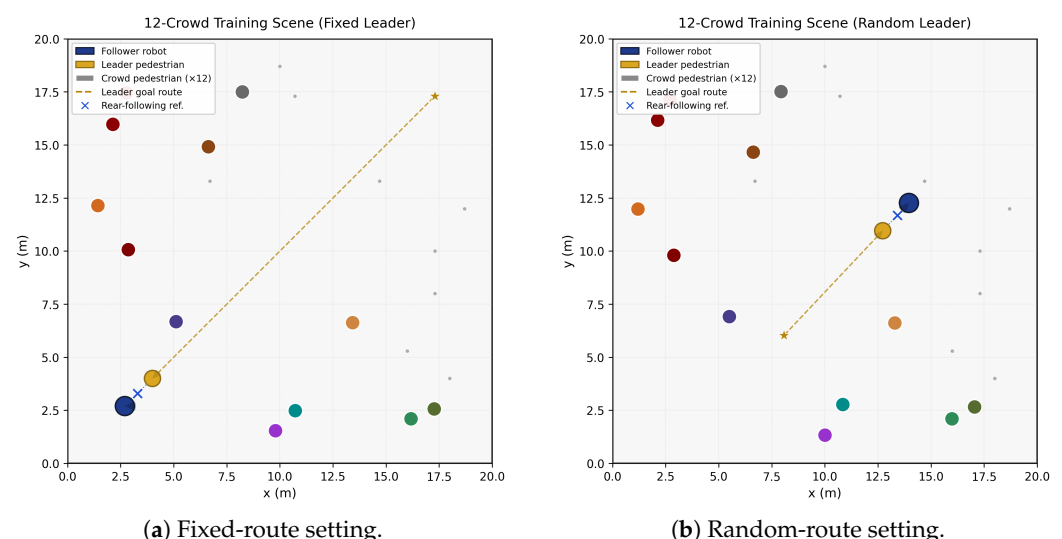


Figure 4. Top-down views of the training environments with 12 background pedestrians in IR-SIM. The follower, leader, and RVO background pedestrians share the same 20 m × 20 m scene template. In the fixed-route setting, the leader follows a predefined start–goal route; in the random-route setting, the leader start and goal positions are resampled across episodes. The dashed line and star denote the leader route and goal, and the blue cross marks the rear-following reference point.

Table 6. Training and testing settings for the fixed-route experiment.

Item	Value
Training steps	2,000,000
Background pedestrians	12
Testing seeds	0, 200, 400, . . . , 1800
Episodes per seed	200

4.7. Random-Route Experiment

The random-route experiment evaluates generalization across varied leader trajectories. As shown in Figure 4b, the leader start and goal are resampled per episode, and the follower is initialized behind the leader. The random leader option is enabled during training and testing. This setting is more challenging because the policy must adapt to varying target directions, path lengths, and crowd interactions. The detailed training and testing settings for the random-route experiment are summarized in Table 7.

Table 7. Training and testing settings for the random-route experiment.

Item	Value
Leader path length	6–16 m
Boundary margin	2 m
Follower initialization	Approximately 1.8 m behind the leader
Training steps	2,000,000
Testing seeds	0, 200, 400, . . . , 1800
Episodes per seed	200

4.8. Ablation Settings

Ablations are conducted under the random-route setting using the full PPO-GAT-Follow model as reference. PPO-MLP-Follow replaces the GAT encoder with a flat observation, while GAT-1Layer reduces the graph encoder to one layer. Reward ablations remove one component at a time. All PPO-based variants are trained for 2 million steps with identical seeds, hyperparameters, scenarios, and evaluation protocols. The structural and reward ablation variants are summarized in Table 8.

Table 8. Ablation variants of PPO-GAT-Follow.

Type	Variant	Modification
Full model	PPO-GAT-Follow	Two-layer GAT with the full reward design
Structure	PPO-MLP-Follow	Replace graph observation with flat vector observation and remove GAT
Structure	GAT-1Layer	Use a one-layer GAT encoder instead of the default two-layer encoder
Reward	w/o visibility	Remove the visibility reward term
Reward	w/o social zone	Remove the proximity penalty term
Reward	w/o behind	Remove the rear-following reward term

5. Results and Analysis

PPO-GAT-Follow is evaluated through six experiment groups: (i) fixed-route performance; (ii) random-route performance and density generalization; (iii) structural and reward ablations; (iv) reward weight, tolerance shift, training seed, and stress robustness analyses; (v) qualitative trajectory comparison; and (vi) Gazebo-based system integration validation. Unless stated otherwise, results are averaged over 10 evaluation seeds with 200 episodes each. Baselines include MPC, DWA, SFM, and the adapted SARL method.

5.1. Fixed-Route Performance

Table 9 reports fixed-route performance with 12 background pedestrians. The leader follows an identical start–goal route across episodes, while pedestrian initial positions are perturbed. PPO-GAT-Follow uses a separately trained fixed-route checkpoint; thus, these results are analyzed independently of the random-route experiments in Section 5.2.

Table 9. Fixed-route performance with 12 background pedestrians (10 seeds $\times$ 200 episodes).

Method	Task (%)	Coll. (%)	SPL	Vis.	Search	Path (m)	Time (s)	Close Proximity Time (s)
MPC	87.9 $\pm$ 1.8	11.9 $\pm$ 1.8	0.783 $\pm$ 0.021	0.888 $\pm$ 0.007	0.947 $\pm$ 0.018	21.3 $\pm$ 0.3	20.8 $\pm$ 0.3	6.88 $\pm$ 0.22
DWA	35.7 $\pm$ 2.4	64.2 $\pm$ 2.2	0.323 $\pm$ 0.024	0.848 $\pm$ 0.004	0.522 $\pm$ 0.020	14.4 $\pm$ 0.2	15.8 $\pm$ 0.3	5.98 $\pm$ 0.23
SFM	0.0 $\pm$ 0.0	8.1 $\pm$ 2.0	0.000 $\pm$ 0.000	0.997 $\pm$ 0.001	0.987 $\pm$ 0.009	77.6 $\pm$ 1.4	65.2 $\pm$ 1.2	54.42 $\pm$ 1.13
SARL	94.3 $\pm$ 1.8	5.4 $\pm$ 1.9	0.941 $\pm$ 0.018	0.997 $\pm$ 0.001	0.987 $\pm$ 0.006	19.9 $\pm$ 0.3	16.7 $\pm$ 0.3	5.32 $\pm$ 0.12
PPO-GAT-Follow (Ours)	98.8 $\pm$ 0.9	1.1 $\pm$ 0.9	0.987 $\pm$ 0.008	0.999 $\pm$ 0.001	0.998 $\pm$ 0.002	19.8 $\pm$ 0.0	16.1 $\pm$ 0.1	15.48 $\pm$ 0.09

Note: Bold values indicate the best numerical value in each metric column. Path length, navigation time, and close-proximity time should be interpreted jointly with the task success rate, since failed episodes may terminate prematurely.

PPO-GAT-Follow achieves the best task-level performance, with 98.8% task success, a 1.1% collision rate, and an SPL of 0.987. Relative to MPC, it improves success by 10.9 percentage points and reduces collisions by 10.8 percentage points. DWA yields shorter paths and times but substantially lower success and higher collision rates.

Compared with SARL, PPO-GAT-Follow improves task success by 4.5 percentage points, reduces collisions by 4.3 percentage points, and increases SPL by 0.046, supporting the benefit of the proposed graph-based representation. SFM preserves visibility but fails the strict rear position and post-arrival conditions and is therefore treated only as a force-based reference.

PPO-GAT-Follow has higher close proximity time because successful rear following keeps the robot near the leader; this metric should therefore be interpreted jointly with task success and collision rate rather than as a direct measure of social violations. Overall, PPO-GAT-Follow provides the strongest fixed-route RPF performance among the evaluated methods.

5.2. Random-Route Performance and Density Generalization

In the random-route experiments, the leader start and goal are resampled per episode. PPO-GAT-Follow is trained with 12 background pedestrians and evaluated without retraining in scenes containing 6, 12, and 20 pedestrians. This protocol tests generalization to sparser and denser crowds under identical route configurations and episode seeds. The complete random-route results across the three crowd-density settings are reported in Table 10.

At the training density, PPO-GAT-Follow achieves 83.1% task success, a 16.7% collision rate, and the highest SPL of 0.815. It outperforms MPC, DWA, and SARL in the main task-level metrics, including a 5.5 percentage-point success gain and a 0.125 SPL improvement over SARL.

Without retraining, it reaches 92.0% success in the six-pedestrian setting. Under 20 pedestrians, SARL attains slightly higher success (66.5% vs. 64.8%), while PPO-GAT-Follow provides higher SPL, visibility, and search success. These results indicate effective, though incomplete, density generalization.

PPO-GAT-Follow also records higher close proximity time than SARL and DWA, reflecting a trade-off between rear-following accuracy and conservative spacing. SFM preserves visibility but fails the strict rear position and stopping conditions and is therefore treated only as a force-based reference.

Table 10. Random-route performance under different numbers of background pedestrians (10 seeds $\times$ 200 episodes, random leader layout). PPO-GAT-Follow is trained with 12 background pedestrians and tested at all densities without retraining.

Density	Method	Task (%)	Coll. (%)	SPL	Vis.	Search	Path (m)	Time (s)	Close Proximity Time (s)
Low (6 background pedestrians)	MPC	77.6 $\pm$ 2.2	22.1 $\pm$ 2.3	0.712 $\pm$ 0.021	0.960 $\pm$ 0.013	0.985 $\pm$ 0.012	14.2 $\pm$ 0.5	14.2 $\pm$ 0.5	5.03 $\pm$ 0.48
	DWA	67.7 $\pm$ 3.0	31.9 $\pm$ 3.0	0.667 $\pm$ 0.031	0.892 $\pm$ 0.013	0.911 $\pm$ 0.019	10.7 $\pm$ 0.3	17.7 $\pm$ 0.6	4.25 $\pm$ 0.30
	SFM	0.0 $\pm$ 0.0	50.8 $\pm$ 3.5	0.000 $\pm$ 0.000	0.975 $\pm$ 0.013	0.954 $\pm$ 0.015	57.0 $\pm$ 1.9	48.4 $\pm$ 1.6	42.88 $\pm$ 1.49
	SARL	88.1 $\pm$ 1.4	10.5 $\pm$ 1.1	0.772 $\pm$ 0.018	0.958 $\pm$ 0.013	0.963 $\pm$ 0.013	14.5 $\pm$ 0.3	15.3 $\pm$ 0.5	3.85 $\pm$ 0.30
	PPO-GAT-Follow (Ours)	92.0 $\pm$ 1.7	7.8 $\pm$ 1.7	0.912 $\pm$ 0.016	0.978 $\pm$ 0.012	0.976 $\pm$ 0.011	11.2 $\pm$ 0.2	10.1 $\pm$ 0.2	7.89 $\pm$ 0.22
Medium (12 background pedestrians)	MPC	64.8 $\pm$ 3.5	34.9 $\pm$ 3.3	0.548 $\pm$ 0.030	0.885 $\pm$ 0.014	0.953 $\pm$ 0.020	15.5 $\pm$ 0.6	16.0 $\pm$ 0.6	6.70 $\pm$ 0.58
	DWA	51.2 $\pm$ 4.7	48.6 $\pm$ 4.8	0.495 $\pm$ 0.048	0.805 $\pm$ 0.013	0.836 $\pm$ 0.015	10.0 $\pm$ 0.2	15.9 $\pm$ 0.4	4.97 $\pm$ 0.29
	SFM	0.0 $\pm$ 0.0	68.0 $\pm$ 2.8	0.000 $\pm$ 0.000	0.939 $\pm$ 0.016	0.918 $\pm$ 0.014	43.9 $\pm$ 1.6	37.6 $\pm$ 1.4	32.82 $\pm$ 1.37
	SARL	77.6 $\pm$ 3.3	21.7 $\pm$ 3.4	0.690 $\pm$ 0.029	0.915 $\pm$ 0.015	0.930 $\pm$ 0.016	13.4 $\pm$ 0.5	13.6 $\pm$ 0.6	5.21 $\pm$ 0.39
	PPO-GAT-Follow (Ours)	83.1 $\pm$ 1.9	16.7 $\pm$ 1.9	0.815 $\pm$ 0.017	0.952 $\pm$ 0.014	0.954 $\pm$ 0.014	10.8 $\pm$ 0.3	10.1 $\pm$ 0.4	7.87 $\pm$ 0.22
High (20 background pedestrians)	MPC	47.2 $\pm$ 3.0	52.6 $\pm$ 2.9	0.369 $\pm$ 0.017	0.818 $\pm$ 0.020	0.936 $\pm$ 0.015	16.9 $\pm$ 0.7	18.2 $\pm$ 0.8	8.49 $\pm$ 0.49
	DWA	36.4 $\pm$ 4.7	63.5 $\pm$ 4.7	0.347 $\pm$ 0.046	0.734 $\pm$ 0.018	0.758 $\pm$ 0.026	9.0 $\pm$ 0.4	13.7 $\pm$ 0.6	5.57 $\pm$ 0.23
	SFM	0.0 $\pm$ 0.0	85.3 $\pm$ 1.6	0.000 $\pm$ 0.000	0.899 $\pm$ 0.022	0.883 $\pm$ 0.023	31.1 $\pm$ 1.3	27.1 $\pm$ 1.1	23.27 $\pm$ 1.05
	SARL	66.5 $\pm$ 3.4	33.5 $\pm$ 3.4	0.601 $\pm$ 0.030	0.870 $\pm$ 0.025	0.887 $\pm$ 0.025	11.9 $\pm$ 0.4	11.1 $\pm$ 0.4	5.98 $\pm$ 0.22
	PPO-GAT-Follow (Ours)	64.8 $\pm$ 4.4	33.7 $\pm$ 3.9	0.628 $\pm$ 0.044	0.904 $\pm$ 0.024	0.911 $\pm$ 0.019	10.5 $\pm$ 0.3	11.5 $\pm$ 0.7	7.75 $\pm$ 0.38

Note: Bold values indicate the best numerical value in each metric column. Path length, navigation time, and close-proximity time should be interpreted jointly with the task success rate, since failed episodes may terminate prematurely.

5.3. Following Quality Analysis

Task success reflects complete RPF completion, including leader arrival, rear position accuracy, and collision avoidance, but does not fully describe following quality. We therefore also evaluate SPL, target visibility, and search success.

PPO-GAT-Follow achieves the highest SPL at all densities: 0.912, 0.815, and 0.628 for 6, 12, and 20 pedestrians, respectively. Despite degradation in denser crowds, it consistently outperforms MPC and DWA in SPL and collision rate, indicating a better balance among completion, efficiency, and safety.

Its visibility ratio decreases from 97.8% to 90.4%, while search success remains above 91% with 20 pedestrians. The larger decline in task success shows that dense-crowd RPF is also limited by access to the rear-following region and stable collision-free completion.

SFM maintains high visibility and search success but achieves zero task success because it does not explicitly enforce rear position, tolerance, or stopping constraints. It is therefore used only as a force-based reference.

5.4. Ablation Study

All ablation experiments are conducted under the random-route setting with 12 background pedestrians. Each variant is trained for 2 million steps using the same training configuration and evaluated over 10 evaluation seeds with 200 episodes per seed. The structural ablation compares the full two-layer GAT model (A0) with PPO-MLP-Follow without graph encoding (A1) and a single-layer GAT variant (A2). The reward ablation individually removes the visibility reward (R1), social zone penalty (R2), and rear-following reward (R3), while keeping the task definition and evaluation metrics unchanged.

As shown in Table 11, removing the GAT encoder causes the largest structural degradation. Compared with the full model, A1 reduces task success from 83.1% to 74.0%, increases collisions from 16.7% to 25.8%, and decreases SPL from 0.815 to 0.665, indicating that flattened observations inadequately represent follower–leader–pedestrian interactions.

The single-layer GAT performs better than A1 but remains inferior to the full two-layer model, with a 4.3 percentage-point reduction in task success and a 3.5 percentage-point increase in collisions. Its occasionally higher visibility or search success does not improve overall task completion.

Table 12 shows that all reward components contribute to performance. Removing the proximity penalty causes the largest degradation, reducing task success by 5.1 percentage points and increasing both collisions and close proximity time. Removing the rear-following reward weakens position maintenance and increases navigation time, while removing the visibility reward consistently reduces visibility, search success, task success, and SPL.

Overall, the full model provides the best task success, collision rate, SPL, and close proximity time, demonstrating the complementary benefits of graph-based interaction modeling and task-oriented reward design.

Table 11. Structural ablation under the random-route setting with 12 background pedestrians (10 evaluation seeds $\times$ 200 episodes).

Type	Variant	Task (%)	Coll. (%)	SPL	Vis.	Search	Path (m)	Time (s)	Close Proximity Time (s)
Structure	A1: w/o GAT (MLP)	74.0 $\pm$ 3.6	25.8 $\pm$ 3.6	0.665 $\pm$ 0.038	0.954 $\pm$ 0.019	0.956 $\pm$ 0.019	13.6 $\pm$ 1.0	12.8 $\pm$ 1.1	12.16 $\pm$ 1.12
	A2: 1-layer GAT	78.8 $\pm$ 1.9	20.2 $\pm$ 1.5	0.769 $\pm$ 0.020	0.952 $\pm$ 0.017	0.955 $\pm$ 0.015	10.9 $\pm$ 0.3	10.0 $\pm$ 0.6	8.42 $\pm$ 0.30
	A0: Full (Ours)	83.1 $\pm$ 1.9	16.7 $\pm$ 1.9	0.815 $\pm$ 0.017	0.952 $\pm$ 0.014	0.954 $\pm$ 0.014	10.8 $\pm$ 0.3	10.1 $\pm$ 0.4	7.87 $\pm$ 0.22

Note: Bold values indicate the best numerical result among the structural variants.

Table 12. Reward ablation under the random-route setting with 12 background pedestrians (10 evaluation seeds $\times$ 200 episodes).

Type	Variant	Task (%)	Coll. (%)	SPL	Vis.	Search	Path (m)	Time (s)	Close Proximity Time (s)
Reward	R1: w/o visibility	80.5 $\pm$ 3.6	18.8 $\pm$ 3.6	0.788 $\pm$ 0.036	0.942 $\pm$ 0.018	0.950 $\pm$ 0.019	10.8 $\pm$ 0.3	10.7 $\pm$ 0.5	7.96 $\pm$ 0.31
	R2: w/o social zone	78.0 $\pm$ 2.8	20.5 $\pm$ 2.6	0.762 $\pm$ 0.027	0.949 $\pm$ 0.020	0.946 $\pm$ 0.021	10.9 $\pm$ 0.2	12.3 $\pm$ 0.6	8.44 $\pm$ 0.28
	R3: w/o behind reward	79.2 $\pm$ 2.3	19.6 $\pm$ 2.4	0.775 $\pm$ 0.022	0.947 $\pm$ 0.018	0.949 $\pm$ 0.022	11.0 $\pm$ 0.3	11.6 $\pm$ 0.5	8.17 $\pm$ 0.28
	R0: Full (Ours)	83.1 $\pm$ 1.9	16.7 $\pm$ 1.9	0.815 $\pm$ 0.017	0.952 $\pm$ 0.014	0.954 $\pm$ 0.014	10.8 $\pm$ 0.3	10.1 $\pm$ 0.4	7.87 $\pm$ 0.22

Note: Bold values indicate the best numerical result among the reward variants.

5.5. Robustness and Reliability Analysis

To assess policy reliability beyond the main benchmark, we conduct four analyses under the random-route setting. Reward-weight sensitivity evaluates dependence on the selected reward configuration. Tolerance-shift robustness tests whether the learned behavior generalizes to nearby success criteria without retraining. Training seed analysis examines reproducibility across independent initializations. Finally, stress testing introduces target pose noise, heterogeneous pedestrian speeds, perturbed RVO parameters, and mixed motion patterns, and evaluates the unchanged policy with 12 and 20 pedestrians.

Together, these experiments assess reward robustness, criterion generalization, training reproducibility, and performance under less idealized target state and crowd dynamics conditions.

5.5.1. Reward-Weight Sensitivity Analysis

To evaluate reward weight sensitivity, we independently retrain several PPO-GAT-Follow variants in the 12-pedestrian random-route setting. The Base model uses the original configuration; Target-heavy, Safety-heavy, and Visibility-heavy increase the corresponding dense reward group by 20%; and Random-perturb scales all dense reward weights by factors sampled from Uniform(0.8, 1.2). Each variant is trained for 2 million steps and evaluated over 10 seeds with 200 episodes each.

As shown in Table 13, Base achieves the best overall performance, with 83.1% task success and an SPL of 0.815. Visibility-heavy and Random-perturb remain close, reaching 81.0% and 80.2% success, respectively, with comparable visibility and search performance.

By contrast, Target-heavy and Safety-heavy reduce success to 73.2% and 65.0% and increase collisions and navigation time. Excessive emphasis on a single objective therefore disrupts the balance among tracking, safety, stabilization, and efficiency; stronger safety weighting does not necessarily improve episode-level safety.

Overall, the policy tolerates balanced perturbations but remains sensitive to large relative weight changes, supporting the original configuration as a more balanced design.

Table 13. Reward-weight sensitivity of PPO-GAT-Follow under the random-route setting with 12 background pedestrians (10 evaluation seeds $\times$ 200 episodes). Each variant is independently trained for 2 million steps, while Base reuses the main checkpoint. The directed variants increase the corresponding dense reward group by 20%. Random-perturb multiplies all dense reward weights by factors sampled from Uniform(0.8, 1.2) and fixed for each training run.

Setting	Task (%)	Coll. (%)	SPL	Vis.	Search	Path (m)	Time (s)	Close Proximity Time (s)
Base	83.1 $\pm$ 1.9	16.7 $\pm$ 1.9	0.815 $\pm$ 0.017	0.952 $\pm$ 0.014	0.954 $\pm$ 0.014	10.8 $\pm$ 0.3	10.1 $\pm$ 0.4	7.87 $\pm$ 0.22
Target-heavy (+20%)	73.2 $\pm$ 2.3	23.7 $\pm$ 2.6	0.712 $\pm$ 0.025	0.943 $\pm$ 0.019	0.945 $\pm$ 0.018	10.7 $\pm$ 0.3	12.2 $\pm$ 0.6	7.78 $\pm$ 0.27
Safety-heavy (+20%)	65.0 $\pm$ 3.6	25.4 $\pm$ 3.2	0.634 $\pm$ 0.036	0.926 $\pm$ 0.017	0.950 $\pm$ 0.016	10.8 $\pm$ 0.3	18.4 $\pm$ 0.9	8.72 $\pm$ 0.30
Visibility-heavy (+20%)	81.0 $\pm$ 2.6	16.5 $\pm$ 2.5	0.798 $\pm$ 0.023	0.955 $\pm$ 0.017	0.960 $\pm$ 0.017	10.7 $\pm$ 0.3	10.9 $\pm$ 0.5	7.74 $\pm$ 0.29
Random-perturb	80.2 $\pm$ 3.1	18.1 $\pm$ 3.0	0.790 $\pm$ 0.030	0.951 $\pm$ 0.020	0.957 $\pm$ 0.019	10.7 $\pm$ 0.3	10.8 $\pm$ 0.5	8.17 $\pm$ 0.25

Note: Bold values indicate the best numerical result among the evaluated reward-weight settings.

5.5.2. Tolerance-Shift Robustness Analysis

To assess dependence on the training tolerance, the same PPO-GAT-Follow checkpoint is evaluated without retraining using final position tolerances of 0.35, 0.45, 0.50, and 0.60 m, with all other settings unchanged.

As shown in Table 14, relaxing the tolerance to 0.50 or 0.60 m causes only minor changes: task success increases from 83.1% to 83.6% and 84.0%, while SPL, visibility, search success, path length, and navigation time remain nearly unchanged. Thus, the policy does not simply exploit the original 0.45 m threshold.

Under the stricter 0.35 m criterion, task success decreases to 63.1%, while collisions and navigation time increase to 32.2% and 18.1 s. Precise stabilization within a narrower rear-following region prolongs close interactions and increases conflict exposure. Overall, the policy is robust to moderate tolerance relaxation but more sensitive to stricter positioning requirements.

Table 14. Tolerance-shift robustness of PPO-GAT-Follow under the random-route setting with 12 background pedestrians (10 evaluation seeds $\times$ 200 episodes). The trained checkpoint is fixed, and only the final following position tolerance is varied. The original training and main-table tolerance is 0.45 m.

Tolerance	Task (%)	Coll. (%)	SPL	Vis.	Search	Path (m)	Time (s)	Close Proximity Time (s)
0.35 m (strict)	63.1 $\pm$ 2.7	32.2 $\pm$ 2.7	0.610 $\pm$ 0.028	0.953 $\pm$ 0.015	0.947 $\pm$ 0.013	11.6 $\pm$ 0.5	18.1 $\pm$ 1.2	15.84 $\pm$ 0.95
0.45 m (original)	83.1 $\pm$ 1.9	16.7 $\pm$ 1.9	0.815 $\pm$ 0.017	0.952 $\pm$ 0.014	0.954 $\pm$ 0.014	10.8 $\pm$ 0.3	10.1 $\pm$ 0.4	7.87 $\pm$ 0.22
0.50 m (relaxed)	83.6 $\pm$ 1.6	16.2 $\pm$ 1.6	0.820 $\pm$ 0.015	0.952 $\pm$ 0.015	0.954 $\pm$ 0.014	10.8 $\pm$ 0.3	10.0 $\pm$ 0.4	7.78 $\pm$ 0.24
0.60 m (relaxed)	84.0 $\pm$ 1.8	15.8 $\pm$ 1.7	0.825 $\pm$ 0.015	0.952 $\pm$ 0.015	0.954 $\pm$ 0.014	10.8 $\pm$ 0.3	10.0 $\pm$ 0.4	7.71 $\pm$ 0.25

Note: The 0.45 m setting is the original training and evaluation tolerance and is highlighted only for reference. The other rows represent criterion-shift evaluations under stricter or more relaxed tolerances and are not ranked as better or worse configurations.

5.5.3. Training Seed Reliability Analysis

To assess training seed reliability, five PPO-GAT-Follow policies are independently trained for 2 million steps under the same 12-pedestrian random-route setting and evaluated over 10 seeds with 200 episodes each.

As shown in Table 15, the policies achieve average task success of 79.7% $\pm$ 3.3%, collision rate of 19.0% $\pm$ 2.1%, and SPL of 0.777 $\pm$ 0.041. Task success ranges from 75.5% to 83.1%, indicating moderate variation without performance collapse. Visibility and search success remain particularly stable at 0.951 $\pm$ 0.002 and 0.952 $\pm$ 0.009, respectively.

Although seed 42 yields the highest task success, the remaining policies retain comparable performance. The results indicate that PPO-GAT-Follow is not dependent on a single favorable initialization and exhibits reasonable reproducibility across independent training runs.

Table 15. Training seed reliability of PPO-GAT-Follow under the random-route setting with 12 background pedestrians. Each policy is independently trained for 2 million steps and evaluated with 10 evaluation seeds $\times$ 200 episodes. Seed 42 corresponds to the main checkpoint. The final row reports the mean $\pm$ standard deviation across five independent training seeds.

Train Seed	Task (%)	Coll. (%)	SPL	Vis.	Search	Path (m)	Time (s)	Close Proximity Time (s)
42 (Main checkpoint)	83.1 $\pm$ 1.9	16.7 $\pm$ 1.9	0.815 $\pm$ 0.017	0.952 $\pm$ 0.014	0.954 $\pm$ 0.014	10.8 $\pm$ 0.3	10.1 $\pm$ 0.4	7.87 $\pm$ 0.22
142	75.5 $\pm$ 3.8	20.4 $\pm$ 2.9	0.720 $\pm$ 0.040	0.948 $\pm$ 0.018	0.939 $\pm$ 0.028	12.7 $\pm$ 0.5	13.3 $\pm$ 0.9	7.79 $\pm$ 0.34
242	80.0 $\pm$ 1.9	18.9 $\pm$ 1.9	0.785 $\pm$ 0.019	0.952 $\pm$ 0.019	0.959 $\pm$ 0.017	11.1 $\pm$ 0.4	10.9 $\pm$ 0.6	7.77 $\pm$ 0.29
342	77.2 $\pm$ 2.7	21.6 $\pm$ 3.0	0.751 $\pm$ 0.025	0.953 $\pm$ 0.019	0.961 $\pm$ 0.019	10.7 $\pm$ 0.5	10.2 $\pm$ 0.7	8.22 $\pm$ 0.48
442	82.5 $\pm$ 3.5	17.4 $\pm$ 3.5	0.811 $\pm$ 0.035	0.951 $\pm$ 0.019	0.950 $\pm$ 0.023	10.8 $\pm$ 0.3	9.6 $\pm$ 0.3	7.32 $\pm$ 0.23
Mean $\pm$ std (5 seeds)	79.7 $\pm$ 3.3	19.0 $\pm$ 2.1	0.777 $\pm$ 0.041	0.951 $\pm$ 0.002	0.952 $\pm$ 0.009	11.2 $\pm$ 0.8	10.8 $\pm$ 1.5	7.80 $\pm$ 0.32

Note: Seed 42 corresponds to the main checkpoint used in the primary experiments and is highlighted in bold only for identification. Bold formatting is not used to rank the individual training seeds because this experiment evaluates training reproducibility rather than model selection.

5.5.4. Stress Robustness Test

To evaluate PPO-GAT-Follow under less idealized yet still controlled dense-crowd conditions, we conduct a stress robustness test without retraining. The same checkpoint trained in the default 12-pedestrian random-route setting is directly evaluated under the same random leader protocol. This experiment is intended to examine policy robustness when several benchmark assumptions are relaxed, rather than to rank all methods under a new benchmark.

The stress setting jointly introduces three types of perturbations. First, zero-mean Gaussian noise with $\sigma = 0.10$ m is added to the controller-facing leader pose and rear-following reference, while the clean simulator state is used only for evaluation. Second, pedestrians are assigned heterogeneous preferred speeds and per-episode RVO parameter jitter, with $v_{\max} \sim U[0.6, 1.1]$ m/s, the avoidance factor sampled from $U[1.5, 3.0]$, and the neighbor range sampled from $U[4, 7]$ m, together with random crowd goals. Third, mixed motion models are used, including RVO-loop, SFM-loop, and wandering pedestrians, with ratios of 6/4/2 for 12 pedestrians and 10/7/3 for 20 pedestrians.

This experiment should therefore be interpreted as a stress test with noisier target input, heterogeneous crowd dynamics, and mixed motion patterns, rather than as a fully realistic human crowd benchmark. Evaluation is conducted at both 12- and 20-pedestrian densities using 10 seeds with 200 episodes per seed, and the original random-route results are included as references.

As shown in Table 16, performance decreases under the stress setting. With 12 pedestrians, task success drops from 83.1% to 75.0%, collision rate increases from 16.7% to 24.4%, and SPL decreases from 0.815 to 0.735. Visibility and search success decline only slightly, suggesting that the stress conditions mainly affect safe and precise completion rather than target observability.

With 20 pedestrians, task success decreases from 64.8% to 52.1%, collision rate rises from 33.7% to 46.5%, and SPL falls from 0.628 to 0.509. Shorter paths and times should not be interpreted as improved efficiency because unsuccessful or collision-terminated episodes may end earlier.

Despite these degradations, the unchanged policy retains 75.0% and 52.1% success in the 12- and 20-pedestrian settings, respectively. Thus, it remains functional under noisier target input and heterogeneous crowd motion, although these factors remain important limitations. Because other methods are not evaluated under the same stress conditions, this experiment assesses policy robustness rather than comparative superiority.

Table 16. Stress robustness of PPO-GAT-Follow under the original and stress settings.

Density	Setting	Task (%)	Coll. (%)	SPL	Vis.	Search	Path (m)	Time (s)	Close Proximity Time (s)
12 background pedestrians	Original random route	83.1 ± 1.9	16.7 ± 1.9	0.815 ± 0.017	0.952 ± 0.014	0.954 ± 0.014	10.8 ± 0.3	10.1 ± 0.4	7.87 ± 0.22
	Stress setting	75.0 ± 2.6	24.4 ± 2.6	0.735 ± 0.026	0.948 ± 0.014	0.940 ± 0.016	10.8 ± 0.3	10.1 ± 0.4	7.67 ± 0.29
20 background pedestrians	Original random route	64.8 ± 4.4	33.7 ± 3.9	0.628 ± 0.044	0.904 ± 0.024	0.911 ± 0.019	10.5 ± 0.3	11.5 ± 0.7	7.75 ± 0.38
	Stress setting	52.1 ± 4.0	46.5 ± 3.5	0.509 ± 0.039	0.889 ± 0.016	0.887 ± 0.015	9.7 ± 0.2	10.9 ± 0.4	6.93 ± 0.32

5.6. Qualitative Trajectory Analysis

To qualitatively illustrate the behavioral differences among the evaluated methods, Figure 5 presents representative trajectories from the random-route setting with 12 background pedestrians. All methods are evaluated using the same random seed and episode index, ensuring matched leader motion, pedestrian initialization, and crowd interaction conditions. The thick blue curve denotes the follower trajectory, the yellow curve denotes the leader trajectory, and the translucent curves show the historical trajectories of background pedestrians. The simulation time and current episode status are displayed above each panel. Additional dynamic demonstrations are provided in Supplementary Video S1.

The representative episode reveals distinct interaction patterns. MPC remains collision-free at the displayed instant but deviates broadly from the leader trajectory in the dense interaction region, reflecting conservative short-horizon optimization and delayed recovery to the desired rear-following position.

DWA repeatedly reacts to local crowd changes but fails to resolve the congestion and eventually collides. SFM also terminates in collision; although its attraction–repulsion mechanism maintains proximity to the leader, it does not explicitly enforce rear position accuracy or post-arrival stabilization.

Both learning-based methods show more coordinated behavior. SARL adapts to pedestrian flow but exhibits a larger trajectory deviation before returning toward the leader path. PPO-GAT-Follow follows a more compact route, avoids the densest pedestrian flow, and maintains the rear-following relationship with limited lateral deviation.

This episode is qualitative rather than statistical evidence, but it is consistent with Table 10: learning-based methods achieve more reliable completion than the classical baselines, while PPO-GAT-Follow provides a favorable balance among rear position maintenance, collision avoidance, and path efficiency.

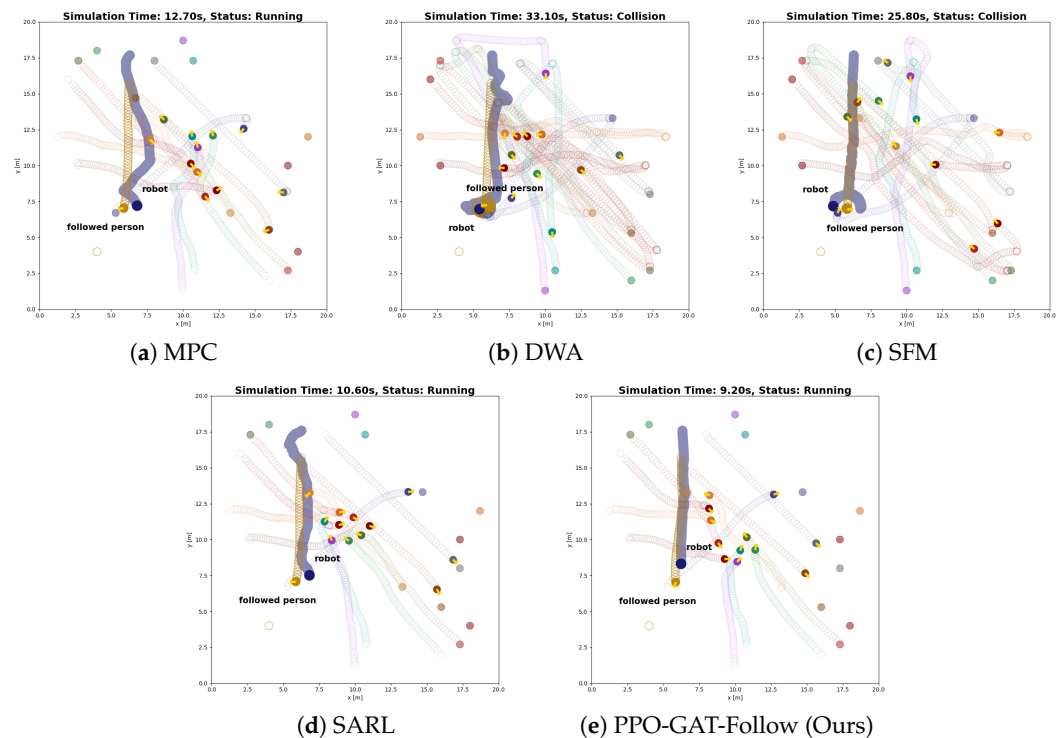


Figure 5. Representative trajectory comparison under the random-route setting with 12 background pedestrians. All methods are evaluated using the same random seed and episode index. The thick blue curve denotes the follower trajectory, the yellow curve denotes the leader trajectory, and the translucent curves denote the historical trajectories of the background pedestrians. The titles above the panels indicate the simulation time and episode status at the displayed instant.

5.7. Gazebo-Based System-Integration Validation

To evaluate system integration feasibility, PPO-GAT-Follow is deployed in a Gazebo Classic 11 simulation pipeline combining robot localization, UWB-like target-relative pose input, surrounding pedestrian perception and tracking, and policy-based control. This experiment assesses integration with robotic sensing and control modules rather than a complete visual RPF system.

The UWB-like interface directly provides the target pose relative to the robot and therefore excludes visual detection, re-identification, tracking drift, and perception-level occlusion recovery. The evaluation consequently focuses on policy execution, pedestrian perception, and collision avoidance with available target-relative pose estimates.

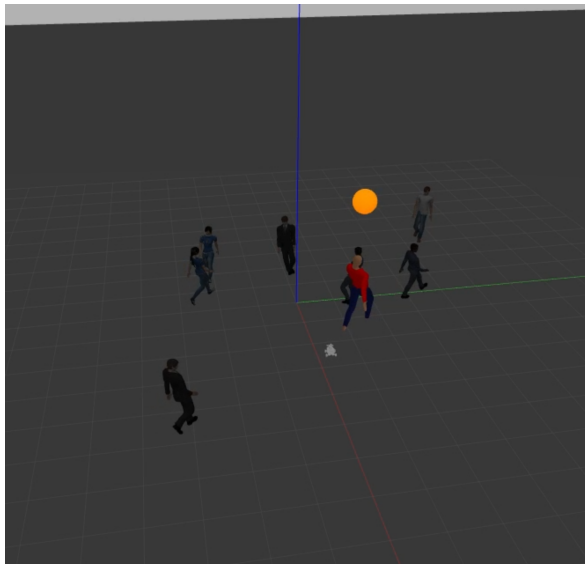
As shown in Figure 6, data from a simulated Livox Mid-360 LiDAR (Livox Technology Co., Ltd., Shenzhen, China) are processed by Super-LIO [59] for localization. The robot, target, and pedestrian states are converted into the graph observation, from which PPO-GAT-Follow generates continuous velocity commands. Dynamic demonstrations are provided in Supplementary Video S1.

Surrounding pedestrians are estimated from Livox point clouds using ROI cropping, voxel downsampling, RANSAC ground removal, adaptive Euclidean clustering, and 3D size filtering. Detected candidates are tracked by a two-dimensional constant-velocity Kalman filter, and their positions and velocities are published through `/pedestrian_state` for policy inference. As shown in Figure 6b, this module converts raw point clouds into structured crowd state inputs.

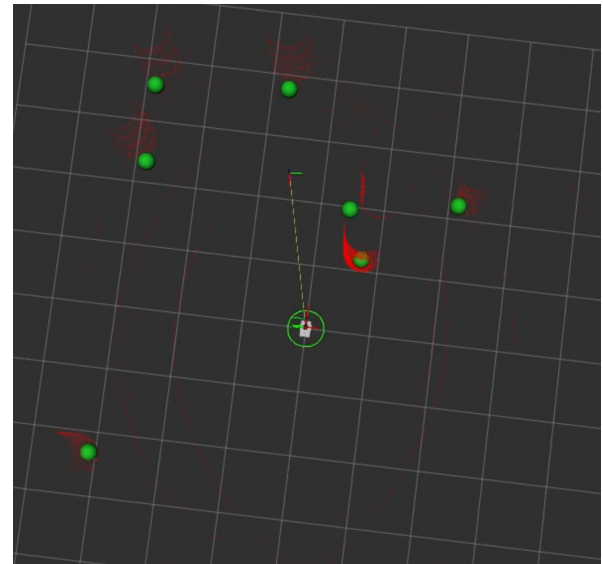
The complete pipeline integrates Super-LIO localization, UWB-like target-relative pose input, pedestrian perception, graph-based inference, and velocity control. Gazebo Classic 11 and RViz2 (ROS 2 Humble) demonstrations confirm that PPO-GAT-Follow

can execute closed-loop following and avoidance within an integrated three-dimensional simulation system.

These results demonstrate system integration feasibility under simulated sensing and available target-relative pose estimates. They do not validate a complete visual RPF system or robustness to real-world target occlusion, since visual identification, re-identification, tracking failure, and perception-level recovery are not evaluated.



(a) Gazebo deployment scenario.



(b) RViz visualization of pedestrian perception and tracking.

Figure 6. Gazebo-based system integration validation. (a) Simulation environment containing the robot, target pedestrian, and surrounding pedestrians. (b) RViz visualization of point cloud-based surrounding pedestrian detection and tracking. The red points denote the filtered Livox point cloud, the green circles denote detected pedestrian candidates.

6. Discussion

6.1. Influence of Crowd Density and Generalization

Crowd density strongly affects RPF performance. Increasing pedestrian numbers causes more frequent geometric occlusions, narrower collision-free passages, and stronger conflicts between rear position maintenance and avoidance, leading to degradation across all methods.

PPO-GAT-Follow is trained with 12 pedestrians and evaluated without retraining with 6 and 20 pedestrians. It performs better in sparse crowds but degrades under high density because the desired rear-following region is more often occupied or blocked. Nevertheless, it retains non-trivial success and competitive efficiency and observability.

Compared with SARL, PPO-GAT-Follow performs better at the training density. With 20 pedestrians, SARL achieves slightly higher success and marginally fewer collisions, whereas PPO-GAT-Follow maintains higher SPL, visibility, and search success. The two learning-based methods therefore show comparable robustness but different trade-offs.

The fixed-size masked graph supports moderate crowd-size variation without architectural changes, while follower-centered attention adapts to changing agent relevance. However, the high-density performance drop confirms that severe congestion remains challenging.

6.2. Interaction Modeling and Reward Shaping

Structural ablations confirm the importance of interaction-aware representation. Replacing the graph encoder with a flattened observation reduces task success and SPL

while increasing collisions, navigation time, and close proximity time, indicating weaker modeling of follower–leader–pedestrian interactions.

The single-layer GAT outperforms the MLP variant but remains inferior to the two-layer model, suggesting that deeper attention aggregation improves the follower-centered representation. However, the star topology mainly captures direct follower–agent relations and does not explicitly model all pedestrian–pedestrian interactions.

Reward ablations show that RPF cannot rely on target attraction or visibility alone. Removing the visibility reward weakens observability, removing the rear-following reward degrades position maintenance and increases navigation time, and removing the proximity penalty causes the largest degradation in collisions and close proximity time.

Sensitivity analysis further shows that balanced perturbations are tolerated, whereas excessive emphasis on target- or safety-related rewards substantially degrades performance. Thus, the original coefficients provide a practical multi-objective balance rather than a globally optimal configuration.

6.3. Robustness, Reliability, and Remaining Trade-Offs

The robustness analyses examine sensitivity to reward weights, final position tolerance, training initialization, and less idealized operating conditions. Relaxing the tolerance from 0.45 m to 0.50 m or 0.60 m causes only minor metric changes, indicating that the policy does not exploit the exact original threshold. By contrast, the stricter 0.35 m criterion substantially reduces success and increases collisions and navigation time. Because this tolerance determines online success termination, stricter episodes continue longer and require stabilization within a narrower rear-following region, increasing exposure to later conflicts.

Independent training runs exhibit moderate variation without severe performance collapse, while visibility and search metrics remain stable, supporting reasonable reproducibility. Under target pose noise and heterogeneous crowd dynamics, performance declines, particularly with 20 pedestrians, but remains non-trivial. These results demonstrate policy-level robustness rather than robustness to realistic perception failures or human behavior.

Close proximity time also reveals a remaining trade-off. Because it combines task-required leader proximity with undesired pedestrian proximity, a high value does not directly imply unsafe behavior. PPO-GAT-Follow does not minimize this metric in every setting, as stronger rear position maintenance may increase close-range exposure. Future work will separate leader and pedestrian proximity components to distinguish accurate following from personal space intrusion.

6.4. Interpretation of Baseline Comparisons

The baselines should be interpreted according to their original objectives. MPC is the strongest classical task-oriented reference because it explicitly optimizes tracking and collision avoidance. DWA represents reactive short-horizon planning, while SARL provides an attention-based learning baseline.

The SARL comparison is particularly informative because both methods use aligned RPF task definitions, training budgets, reward scales, and evaluation protocols, while retaining their respective algorithm-specific action formulations. PPO-GAT-Follow performs better in the fixed-route and training-density settings, supporting the value of its graph representation and task-specific features. However, their comparable high-density results show that PPO-GAT-Follow does not dominate under every condition. Its advantage is therefore best characterized as a favorable overall balance among completion, safety, efficiency, and observability.

SFM requires separate interpretation because it does not explicitly enforce the rear reference tolerance, rear half-space constraint, or post-arrival stopping condition. Its zero task success reflects this structural mismatch rather than an inability to maintain target proximity or visibility. Accordingly, SFM is retained as a force-based reference, while the main quantitative comparisons focus on MPC and SARL.

6.5. Implications for Dense-Crowd RPF

The experiments show that target visibility and geometric reacquisition are necessary but insufficient for successful RPF. A method may preserve line of sight while still failing to reach the required rear-following position, avoid collisions, or stabilize after leader arrival.

Unlike conventional target tracking, RPF requires both target observability and appropriate relative positioning under dynamic crowd constraints. In dense scenes, the rear reference may be temporarily blocked, and direct pursuit can conflict with collision avoidance and social space preservation.

The proposed graph representation and task-oriented reward jointly address these requirements. Accordingly, dense-crowd RPF should be evaluated using complementary metrics, including task success, collision rate, SPL, target visibility, reacquisition success, navigation efficiency, and close proximity time, since no single metric fully characterizes following quality.

6.6. System-Integration Implications

The Gazebo validation demonstrates system integration beyond isolated 2D policy evaluation. The closed-loop pipeline combines simulated Livox Mid-360 LiDAR sensing, Super-LIO localization, UWB-like target-relative pose input, point cloud-based surrounding pedestrian perception, Kalman tracking, graph-based inference, and velocity control, showing that PPO-GAT-Follow can operate within a three-dimensional robotic simulation framework.

However, this experiment assumes directly available target-relative pose estimates and is not a complete visual RPF deployment. The designated target is not detected or identified through the point cloud perception pipeline, excluding visual detection, identity association, re-identification, and perception-level occlusion recovery.

Thus, the results support integration feasibility for perception, tracking, policy execution, and control, but do not validate robustness to target loss, identity switches, tracking drift, communication failure, or delayed state estimates.

6.7. Limitations and Future Work

Several limitations remain. First, the quantitative experiments rely mainly on IR-SIM with predefined RVO-based pedestrian dynamics. Although reproducible, this setting cannot fully capture abrupt motion changes, group behavior, cooperation, yielding, or socially dependent human responses. The stress tests add target pose noise and more diverse pedestrian dynamics, but remain controlled simulations rather than substitutes for real-crowd validation.

Second, the IR-SIM experiments use simulator-provided leader states, while the Gazebo system assumes a UWB-like target-relative pose interface. Thus, visual detection uncertainty, identity switches, re-identification failure, tracking drift, latency, and intermittent target loss are not evaluated. The Gaussian pose perturbations provide only preliminary robustness evidence and do not represent temporally correlated, biased, delayed, or discontinuous perception errors.

Third, PPO-GAT-Follow does not always minimize close proximity time. Stronger rear position maintenance and task completion may increase close-range interactions, revealing a remaining trade-off between strict following accuracy and social space preservation.

Fourth, the graph retains only a fixed number of nearest pedestrians and uses a follower-centered star topology. This supports efficient inference but may discard relevant long-range interactions and does not explicitly model pedestrian–pedestrian relations or group structures.

Future work will investigate real robot evaluation, uncertainty- and latency-aware target estimation, visual re-identification, sim-to-real adaptation, socially adaptive rewards, and scalable graph representations incorporating group behavior and broader crowd interactions.

7. Conclusions

This paper proposed PPO-GAT-Follow, an interaction-aware reinforcement learning framework for robot person following in dense crowds. The method combines graph attention-based interaction modeling with task-oriented reward shaping to jointly address target maintenance, collision avoidance, rear position accuracy, visibility preservation, close proximity regulation, and motion stability.

In the fixed-route setting with 12 background pedestrians, PPO-GAT-Follow achieves 98.8% task success, a 1.1% collision rate, and an SPL of 0.987. At the random-route training density, it achieves 83.1% task success and an SPL of 0.815, while generalizing without retraining to 6- and 20-pedestrian settings. It consistently outperforms MPC in task success, collision rate, and SPL, and surpasses SARL in the fixed-route and training-density experiments while remaining competitive under high density. The SFM results further show that target attraction and repulsion alone are insufficient for strict rear-following completion.

Ablation studies confirm the contributions of graph-based interaction modeling and balanced reward design. Reward-weight, tolerance shift, multi-seed, and stress analyses demonstrate reasonable robustness to moderate configuration changes, stochastic initialization, target pose noise, and heterogeneous crowd dynamics, although performance degrades under less idealized high-density conditions.

The Gazebo validation further demonstrates integration with localization, UWB-like target-relative pose input, point cloud pedestrian perception, tracking, and velocity control. However, visual target identification, re-identification, tracking failure, and perception-level occlusion recovery remain outside the current scope. Future work will focus on real robot validation, uncertainty-aware perception, human-response modeling, and sim-to-real transfer.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/s26154711/s1>, Video S1: Dynamic demonstrations of PPO-GAT-Follow in the IR-SIM and Gazebo environments.

Author Contributions: Conceptualization, C.G. and X.Z.; methodology, X.Z.; software, X.Z.; validation, X.Z. and Y.S.; formal analysis, X.Z. and Y.S.; investigation, X.Z. and Y.S.; resources, S.P. and C.G.; data curation, X.Z.; writing—original draft preparation, X.Z.; writing—review and editing, S.P., C.G., and Y.S.; visualization, X.Z.; supervision, Y.S. and C.G.; project administration, Y.S. and C.G.; funding acquisition, C.G. and S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Xincheng Qihang Inc. No specific grant number was assigned.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-5.5, OpenAI, San Francisco, CA, USA) for language polishing and writing assistance. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: Authors Yongliang Shi and Chao Gao were employed by Xincheng Qihang Inc. The remaining 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.

References

1. Mavrogiannis, C.; Baldini, F.; Wang, A.; Zhao, D.; Trautman, P.; Steinfeld, A.; Oh, J. Core Challenges of Social Robot Navigation: A Survey. *ACM Trans. Hum.-Robot. Interact.* **2023**, *12*, 1–39. <https://doi.org/10.1145/3583741>.
2. Singamaneni, P.T.; Bachiller-Burgos, P.; Manso, L.J.; Garrell, A.; Sanfeliu, A.; Spalanzani, A.; Alami, R. A survey on socially aware robot navigation: Taxonomy and future challenges. *Int. J. Robot. Res.* **2024**, *43*, 1533–1572.
3. Gao, Y.; Huang, C.M. Evaluation of Socially-Aware Robot Navigation. *Front. Robot. AI* **2022**, *8*, 721317. <https://doi.org/10.3389/frobt.2021.721317>.
4. Francis, A.; Pérez-d’Arpino, C.; Li, C.; Xia, F.; Alahi, A.; Alami, R.; Bera, A.; Biswas, A.; Biswas, J.; Chandra, R.; et al. Principles and guidelines for evaluating social robot navigation algorithms. *ACM Trans. Hum.-Robot. Interact.* **2025**, *14*, 1–65.
5. Islam, M.J.; Hong, J.; Sattar, J. Person-Following by Autonomous Robots: A Categorical Overview. *Int. J. Robot. Res.* **2019**, *38*, 1581–1618. <https://doi.org/10.1177/0278364919881683>.
6. Ye, H.; Situ, W.; Peng, J.; Zhan, Y.; Xia, B.; Cai, K.; Zhang, H. Follow-Bench: A Unified Motion Planning Benchmark for Socially-Aware Robot Person Following. *arXiv* **2025**, arXiv:2509.10796. Available online: <http://arxiv.org/abs/2509.10796> (accessed on 20 July 2026).
7. Kästner, L.; Fatloun, B.; Shen, Z.; Gawrisch, D.; Lambrecht, J. Human-Following and -Guiding in Crowded Environments Using Semantic Deep-Reinforcement-Learning for Mobile Service Robots. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 833–839. <https://doi.org/10.1109/ICRA46639.2022.9812111>.
8. Stratton, A.; Hauser, K.; Mavrogiannis, C. Characterizing the Complexity of Social Robot Navigation Scenarios. *IEEE Robot. Autom. Lett.* **2025**, *10*, 184–191. <https://doi.org/10.1109/LRA.2024.3502060>.
9. Ye, H.; Zhao, J.; Pan, Y.; Chen, W.; He, L.; Zhang, H. Robot Person Following Under Partial Occlusion. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 7591–7597. <https://doi.org/10.1109/ICRA48891.2023.10160738>.
10. Ghimire, A.; Zhang, X.; Javed, S.; Dias, J.; Werghi, N. Robot Person Following in Uniform Crowd Environment. *arXiv* **2022**, arXiv:2205.10553. Available online: <http://arxiv.org/abs/2205.10553> (accessed on 20 July 2026).
11. Ye, H.; Zhao, J.; Zhan, Y.; Chen, W.; He, L.; Zhang, H. Person Re-Identification for Robot Person Following with Online Continual Learning. *IEEE Robot. Autom. Lett.* **2024**, *9*, 9151–9158. <https://doi.org/10.1109/LRA.2024.3438042>.
12. Kang, S.; Kim, J.; Yang, S.; Kim, B.; Yura, J.; Kim, D. Social Type-Aware Navigation Framework for Mobile Robots in Human-Shared Environments. *Sensors* **2024**, *24*, 4862. <https://doi.org/10.3390/s24154862>.
13. Fox, D.; Burgard, W.; Thrun, S. The Dynamic Window Approach to Collision Avoidance. *IEEE Robot. Autom. Mag.* **1997**, *4*, 23–33. <https://doi.org/10.1109/100.580977>.
14. Helbing, D.; Molnár, P. Social Force Model for Pedestrian Dynamics. *Phys. Rev. E* **1995**, *51*, 4282–4286. <https://doi.org/10.1103/PhysRevE.51.4282>.
15. Samavi, S.; Khan, M.O.U.; Warnell, G.; Stone, P. SICNav: Safe and Interactive Crowd Navigation Using Model Predictive Control and Bilevel Optimization. *IEEE Trans. Robot.* **2025**, *41*, 801–818. <https://doi.org/10.1109/TRO.2024.3484634>.
16. Mozzarelli, L.; Corno, M.; Savaresi, S.M. Socially Aware Local Planning: A Dynamic Window-Based Approach. In Proceedings of the 2024 European Control Conference (ECC), Stockholm, Sweden, 25–28 June 2024; IEEE: Piscataway, NJ, USA, 2024.
17. Han, J.R.; Thomas, H.; Zhang, J.; Rhinehart, N.; Barfoot, T.D. DR-MPC: Deep Residual Model Predictive Control for Real-World Social Navigation. *arXiv* **2024**, arXiv:2410.10646. Available online: <http://arxiv.org/abs/2410.10646> (accessed on 20 July 2026).
18. Aslam, M.P.; Derajic, B.; Bouzidi, M.-K.; Bernhard, S.; Ringert, J.O. Model Predictive Control for Crowd Navigation via Learning-Based Trajectory Prediction. In Proceedings of the 22nd International Conference on Informatics in Control, Automation and Robotics (ICINCO), Marbella, Spain, 20–22 October 2025; pp. 251–258. <https://doi.org/10.5220/0013710400003982>.
19. Chen, Y.F.; Liu, M.; Everett, M.; How, J.P. Decentralized Non-Communicating Multiagent Collision Avoidance with Deep Reinforcement Learning. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 285–292. <https://doi.org/10.1109/ICRA.2017.7989037>.

20. Everett, M.; Chen, Y.F.; How, J.P. Motion Planning Among Dynamic, Decision-Making Agents with Deep Reinforcement Learning. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 3052–3059. <https://doi.org/10.1109/IROS.2018.8593871>.
21. Chen, C.; Liu, Y.; Kreiss, S.; Alahi, A. Crowd-Robot Interaction: Crowd-Aware Robot Navigation with Attention-Based Deep Reinforcement Learning. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 6015–6022. <https://doi.org/10.1109/ICRA.2019.8794134>.
22. Chen, C.; Hu, S.; Nikdel, P.; Mori, G.; Savva, M. Relational Graph Learning for Crowd Navigation. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Virtually, 24 October–21 January 2021; IEEE: Piscataway, NJ, USA, 2020; pp. 10007–10013. <https://doi.org/10.1109/IROS45743.2020.9341699>.
23. Liu, S.; Chang, P.; Liang, W.; Chakraborty, N.; Driggs-Campbell, K. Decentralized Structural-RNN for Robot Crowd Navigation with Deep Reinforcement Learning. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 3517–3524. <https://doi.org/10.1109/ICRA48506.2021.9561595>.
24. Yang, C.; Li, Y.; Zhang, J.; Zhao, H.; Sun, Q. ST²: Spatial-Temporal State Transformer for Crowd-Aware Autonomous Navigation. *IEEE Robot. Autom. Lett.* **2023**, *8*, 912–919. <https://doi.org/10.1109/LRA.2023.3234815>.
25. Wang, W.; Wang, R.; Mao, L.; Min, B.C. NaviSTAR: Socially Aware Robot Navigation with Hybrid Spatio-Temporal Graph Transformer and Preference Learning. *arXiv* **2023**, arXiv:2304.05979. Available online: <http://arxiv.org/abs/2304.05979> (accessed on 20 July 2026).
26. Liu, S.; Chang, P.; Huang, Z.; Chakraborty, N.; Hong, K.; Liang, W.; McPherson, D.L.; Geng, J.; Driggs-Campbell, K. Intention Aware Robot Crowd Navigation with Attention-Based Interaction Graph. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 12015–12021. <https://doi.org/10.1109/ICRA48891.2023.10160632>.
27. Patel, U.; Kumar, N.; Sathiamoorthy, A.J.; Manocha, D. Dynamically Feasible Deep Reinforcement Learning Policy for Robot Navigation in Dense Mobile Crowds. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 6052–6059. <https://doi.org/10.1109/ICRA48506.2021.9561317>.
28. Zhou, Z.; Zhu, P.; Zeng, Z.; Xiao, J.; Lu, H.; Zhou, Z. Robot Navigation in a Crowd by Integrating Deep Reinforcement Learning and Online Planning. *Appl. Intell.* **2022**, *52*, 15600–15616. <https://doi.org/10.1007/s10489-022-03191-2>.
29. Xie, Z.; Dames, P. DRL-VO: Learning to Navigate Through Crowded Dynamic Scenes Using Velocity Obstacles. *IEEE Trans. Robot.* **2023**, *39*, 2700–2719. <https://doi.org/10.1109/TRO.2023.3257549>.
30. Samsani, S.S.S.; Muhammad, H. Memory-Based Crowd-Aware Robot Navigation Using Deep Reinforcement Learning. *Complex Intell. Syst.* **2023**, *9*, 2147–2163. <https://doi.org/10.1007/s40747-022-00906-3>.
31. Kim, J.; Kang, S.; Yang, S.; Kim, B.; Yura, J.; Kim, D. Transformable Gaussian Reward Function for Socially-Aware Navigation with Deep Reinforcement Learning. *Sensors* **2024**, *24*, 4540. <https://doi.org/10.3390/s24144540>.
32. Maddumage, V.; Kodagoda, S.; Carmichael, M.G.; Gunatilake, A.; Thiagarajan, K.; Martin, J. Relative Velocity-Based Reward Model for Socially-Aware Navigation with Deep Reinforcement Learning. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA), Atlanta, Georgia, 19–23 May 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 10448–10454.
33. Möller, R.; Furnari, A.; Battiato, S.; Härmä, A.; Farinella, G.M. A survey on human-aware robot navigation. *Robot. Auton. Syst.* **2021**, *145*, 103837.
34. Guillén-Ruiz, S.; Bandera, J.P.; Hidalgo-Paniagua, A.; Bandera, A. Evolution of Socially-Aware Robot Navigation. *Electronics* **2023**, *12*, 1570. <https://doi.org/10.3390/electronics12071570>.
35. Kabir, I.K.; Mysorewala, M.F. Socially aware navigation for mobile robots: a survey on deep reinforcement learning approaches. *Appl. Intell.* **2026**, *56*, 38.
36. Ögren, P.; Leonard, N.E. A convergent dynamic window approach to obstacle avoidance. *IEEE Trans. Robot.* **2005**, *21*, 188–195. <https://doi.org/10.1109/TRO.2004.838008>.
37. Trautman, P.; Ma, J.; Murray, R.M.; Krause, A. Robot Navigation in Dense Human Crowds: Statistical Models and Experimental Studies of Human–Robot Cooperation. *Int. J. Robot. Res.* **2015**, *34*, 335–356. <https://doi.org/10.1177/0278364914557874>.
38. Fan, T.; Cheng, X.; Pan, J.; Long, P.; Liu, W.; Yang, R.; Manocha, D. Getting Robots Unfrozen and Unlost in Dense Pedestrian Crowds. *IEEE Robot. Autom. Lett.* **2019**, *4*, 1178–1185. <https://doi.org/10.1109/LRA.2019.2891491>.
39. Khatib, O. Real-Time Obstacle Avoidance for Manipulators and Mobile Robots. *Int. J. Robot. Res.* **1986**, *5*, 90–98. <https://doi.org/10.1177/027836498600500106>.
40. Koren, Y.; Borenstein, J. Potential Field Methods and Their Inherent Limitations for Mobile Robot Navigation. In Proceedings of the IEEE International Conference on Robotics and Automation, Sacramento, CA, USA, 9–11 April 1991; IEEE: Piscataway, NJ, USA, 1991; pp. 1398–1404. <https://doi.org/10.1109/ROBOT.1991.131810>.

41. Kretz, T. On Oscillations in the Social Force Model. *Phys. A Stat. Mech. Its Appl.* **2015**, *438*, 272–285. <https://doi.org/10.1016/j.physa.2015.07.002>.
42. Chraïbi, M. Oscillating Behavior within the Social Force Model. *arXiv* **2014**, arXiv:1412.1133. <https://doi.org/10.48550/arXiv.1412.1133>.
43. Kiss, D.; Tevesz, G. A Model Predictive Navigation Approach Considering Mobile Robot Shape and Dynamics. *Period. Polytech. Electr. Eng. Comput. Sci.* **2012**, *56*, 43–50. <https://doi.org/10.3311/PPee.7162>.
44. Jian, Z.; Yan, Z.; Lei, X.; Lu, Z.; Lan, B.; Wang, X.; Liang, B. Dynamic Control Barrier Function-Based Model Predictive Control to Safety-Critical Obstacle-Avoidance of Mobile Robot. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 3679–3685. <https://doi.org/10.1109/ICRA48891.2023.10160857>.
45. Le, V.A.; Chalaki, B.; Tadiparthi, V.; Mahjoub, H.N.; D'Sa, J.; Moradi-Pari, E. Social Navigation in Crowded Environments with Model Predictive Control and Deep Learning-Based Human Trajectory Prediction. *arXiv* **2023**, arXiv:2309.16838. Available online: <http://arxiv.org/abs/2309.16838> (accessed on 20 July 2026).
46. Repiso, E.; Garrell, A.; Sanfeliu, A. Adaptive Side-by-Side Social Robot Navigation to Approach and Interact with People. *Int. J. Soc. Robot.* **2020**, *12*, 909–930. <https://doi.org/10.1007/s12369-019-00559-2>.
47. Sekiguchi, S.; Yorozu, A.; Kuno, K.; Okada, M.; Watanabe, Y.; Takahashi, M. Uncertainty-Aware Non-linear Model Predictive Control for Human-Following Companion Robot. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 8316–8322. <https://doi.org/10.1109/ICRA48506.2021.9561974>.
48. Peng, J.; Liao, Z.; Su, Z.; Yao, H.; Zeng, Y.; Dai, H. A Dual Closed-Loop Control Strategy for Human-Following Robots Respecting Social Space. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 11252–11258.
49. Kuderer, M.; Burgard, W. An Approach to Socially Compliant Leader Following for Mobile Robots. In *Social Robotics*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 239–248. https://doi.org/10.1007/978-3-319-11973-1_24.
50. Montesdeoca, J.; Toibero, J.M.; Jordan, J.; Zell, A.; Carelli, R. Person-Following Controller with Socially Acceptable Robot Motion. *Robot. Auton. Syst.* **2022**, *153*, 104075. <https://doi.org/10.1016/j.robot.2022.104075>.
51. Dang, C.V.; Ahn, H.; Kim, J.W.; Lee, S.C. Collision-Free Navigation in Human-Following Task Using a Cognitive Robotic System on Differential Drive Vehicles. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *15*, 78–87.
52. Ferrer, G.; Zulueta, A.G.; Cotarelo, F.H.; Sanfeliu, A. Robot Social-Aware Navigation Framework to Accompany People Walking Side-by-Side. *Auton. Robot.* **2017**, *41*, 775–793.
53. Repiso, E.; Garrell, A.; Sanfeliu, A. Adaptive Social Planner to Accompany People in Real-Life Dynamic Environments. *Int. J. Soc. Robot.* **2024**, *16*, 1189–1221.
54. Wang, Z.; Huang, S.; Li, M.; Ouyang, J.; Wang, Y.; Chen, H. Continuous Robotic Tracking of Dynamic Targets in Complex Environments Based on Detectability. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 16338–16344. <https://doi.org/10.1109/ICRA57147.2024.10610993>.
55. Leisiazar, S.; Park, E.J.; Lim, A.; Chen, M. An MCTS-DRL Based Obstacle and Occlusion Avoidance Methodology in Robotic Follow-Ahead Applications. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1–5 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 221–228. <https://doi.org/10.1109/IROS55552.2023.10342150>.
56. Leisiazar, S.; Rohani, S.R.R.; Park, E.J.; Lim, A.; Chen, M. Adapting to Frequent Human Direction Changes in Autonomous Frontal Following Robots. *IEEE Robot. Autom. Lett.* **2025**, *10*, 2934–2941. <https://doi.org/10.1109/LRA.2025.3537860>.
57. Stable-Baselines3 Contributors. Stable-Baselines3: Reliable Reinforcement Learning Implementations in PyTorch. Version 2.8.0. Available online: <https://github.com/DLR-RM/stable-baselines3> (accessed on 20 July 2026).
58. Han, R. IR-SIM: An Open-Source Lightweight Simulator for Robot Navigation, Control, and Learning. Version 2.9.0. 2026. Available online: <https://github.com/hanruihua/ir-sim> (accessed on 20 June 2026).
59. Wang, L.; Zhang, X.; Li, C.; He, D.; Pan, Y.; Yi, J. Super-LIO: Official Software Implementation. Available online: <https://github.com/Liansheng-Wang/Super-LIO> (accessed on 20 July 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.