Revista Eletrônica de Iniciação Científica em Computação, 2026, 24:1
ISSN: 3085-8461 • doi: 10.5753/reic.2026.8466

Research Paper

Follower-Aware Reward Shaping for Leader–Follower Formation Control via Dueling Deep Q-Networks

Natasha Araújo Caxias [Universidade Federal do Amazonas (UFAM) | *natasha.caxias@icomp.ufam.edu.br*]
Abel Severo Rocha [Universidade Federal do Amazonas (UFAM) | *abel.severo@icomp.ufam.edu.br*]
José Reginaldo Hughes Carvalho [Universidade Federal do Amazonas (UFAM) | *reginaldo@icomp.ufam.edu.br*]

Instituto de Computação (IComp) – Universidade Federal do Amazonas (UFAM) Av. Gen. Rodrigo Octávio, 6200, Coroado I – 69080-900 – Manaus – AM – Brasil

Abstract. This paper proposes a follower-aware reward shaping strategy for leader–follower formation control of non-holonomic mobile robots, implemented within a hybrid Dueling Deep Q-Network (Dueling DQN) architecture. The leader employs a Dueling DQN enhanced with Prioritized Experience Replay, Frame Stacking, and a novel follower-aware safety penalty trained under a Curriculum Learning regime with progressive Domain Randomization. The follower tracks a virtual target located behind the leader using a proportional–derivative controller, thereby decoupling computational complexity between the agents. Validated in an 8×8 m environment featuring a 2.0 m inner gap and a 1.0 m lateral exit door ($d_{\text{collision}} = 0.50$ m), the proposed system achieves 100% success across three fixed-geometry configurations (500 episodes each) and 74.0% in a procedural generalization test with randomized obstacle placement. Comprehensive ablation studies confirm that the follower-aware reward shaping strategy is the single most impactful component, outweighing established techniques such as Prioritized Experience Replay and dueling decomposition. The findings provide empirical evidence that explicitly encoding cooperative safety constraints into the leader's policy yields emergent path-planning behaviors that account for the follower's kinematic limitations, offering a practical pathway toward robust multi-robot coordination in cluttered environments.

Keywords: Reinforcement Learning, Formation Control, Mobile Robotics, Dueling DQN, Reward Shaping

Received: 17 June 2026 • **Accepted:** 18 June 2026 • **Published:** 10 July 2026

1 Introduction

Multi-robot systems (MRS) have become an increasingly relevant paradigm in modern robotics, offering enhanced redundancy, parallel task execution, fault tolerance, and scalability when compared to single-agent solutions Rizk *et al*. [2019]. By distributing tasks among multiple heterogeneous or homogeneous agents, MRS enable applications that would otherwise be infeasible or excessively costly for a single robot, such as large-scale environmental monitoring, cooperative manipulation, and search and rescue operations in hazardous environments. Among the many coordination strategies investigated in the literature, leader–follower architectures have gained widespread adoption due to their intuitive structure, ease of deployment, and modest communication requirements Oh *et al*. [2015]. These architectures have demonstrated success in diverse application domains, including warehouse and intra-logistics operations Fragapane *et al*. [2021], autonomous convoy driving, agricultural fleets, and space exploration missions Huntsberger *et al*. [2003].

Typically, the leader dictates the global trajectory while the follower maintains a predefined spatial relationship relative to it, significantly mitigating the need for complex, heavy path-planning capabilities across all agents. However, coordinating non-holonomic robots—which are constrained by their inability to move laterally and thus require wider turning arcs—in cluttered environments remains a profound challenge due to these severe kinematic constraints Siegwart *et al*. [2011] and the strict demand for dynamic obstacle avoidance. Crucially, classical leader-follower formulations typically operate under a strict unidirectional assumption: the leader plans its trajectory independently, while the trailing units reactively track it. In real-world engineering applications, this lack of mutual awareness frequently triggers formation collapse, systemic resonance, or hardware collisions. For instance, in terrestrial autonomous truck platooning, a leader oblivious to the trailing vehicles' communication latencies or braking profiles risks causing severe rear-end accidents during sudden maneuvers. Similarly, in aerial search-and-rescue operations or aquatic oceanographic monitoring—where autonomous surface vehicles (ASVs) guide subaquatic autonomous underwater vehicles (AUVs) through stochastic currents—the leader must actively accommodate the follower's localized spatial bottlenecks and communication constraints. Therefore, deploying multi-agent systems in unstructured environments demands a transition toward follower-aware paradigms. In such frameworks, the leader's navigation policy intrinsically encodes the cooperative safety and physical limitations of the trailing unit, yielding emergent, proactive clearance behaviors that guarantee overall formation resilience.

Classical control approaches to formation maintenance and obstacle avoidance include Artificial Potential Fields (APF) Khatib [1986], which model navigation as the interaction between attractive forces toward the goal and repulsive forces from obstacles. While intuitive and computationally lightweight, APF methods are well known to suffer from local minima in non-convex environments, frequently trapping agents in unstable equilibria. Model Predictive Control (MPC) Nascimento *et al*. [2013], on the other hand, can produce optimal trajectories subject to explicit constraints, but

its computational demands may exceed the budget allowed by real-time embedded controllers, particularly when the prediction horizon is long and the dynamics are non-linear. More recently, Deep Reinforcement Learning (DRL) has emerged as a powerful alternative, enabling agents to learn navigation and coordination policies directly from experience without the need to hand-engineer behavior trees or control laws Tai *et al*. [2017]; Mnih *et al*. [2015]; Zhu and Zhang [2021]. DRL has shown remarkable success in tasks ranging from collision avoidance in crowded spaces to game-playing at superhuman levels, suggesting that it may also enable cooperative behaviors that are difficult to specify through analytical means.

This paper addresses the problem of leader–follower formation control through a hybrid architecture designed to combine the strengths of both learning-based and classical approaches. Specifically, the leader employs a Dueling DQN-based policy for navigation and obstacle avoidance, while the follower uses a proportional–derivative (PD) controller for formation tracking. This division of labor concentrates computational complexity in the leader, which is typically the better-equipped agent, while enabling lightweight followers through simple, well-understood control laws. The central contribution of the paper is a follower-aware reward shaping strategy that couples the leader's optimization objective to the follower's spatial safety, encouraging the leader to plan paths that remain kinematically feasible for the trailing unit. Concretely, the main contributions of this work can be summarized as follows:

- A hybrid DRL-classical control architecture for heterogeneous robot teams that integrates learned obstacle avoidance with deterministic formation tracking, balancing computational cost with safety and performance;
- A Dueling DQN agent enhanced with Prioritized Experience Replay (PER), a 21-ray simulated LiDAR observation space, Golden Ratio action discretization, and Frame Stacking to provide implicit temporal awareness of leader–follower dynamics;
- A follower-aware reward shaping strategy that explicitly encodes cooperative safety constraints into the leader's policy, producing emergent path-planning behaviors that proactively account for the follower's kinematic limitations;
- Quantitative validation through 500-episode evaluations across three difficulty levels (Easy, Medium, Hard) and a procedural generalization test with randomized geometry, complemented by an ablation study that isolates the contribution of each architectural component.

The remainder of this paper is organized as follows. Section 2 discusses the theoretical foundations and prior work most relevant to our approach. Section 3 formalizes the problem and notation. Section 4 presents the proposed hybrid architecture and its components in detail. Section 5 reports the simulation results, including the ablation study and a quantitative analysis of the impact of communication latency on follower stability. Section 6 addresses the main limitations of the results and how they will be addresses in future works. Finally, Section 7 concludes the paper and outlines directions for future work. Throughout the paper, an ablation-driven evaluation strategy is adopted, allowing each architectural component to be individually isolated and assessed.

2 Related Work

Formation control has been extensively studied through three principal paradigms: leader–follower architectures, virtual structures, and behavior-based approaches Oh *et al*. [2015]. In leader–follower schemes, one agent serves as the reference and the remaining agents follow predefined spatial offsets; virtual structures treat the entire formation as a single rigid body whose dynamics are controlled jointly; and behavior-based methods compose local rules such as cohesion, separation, and alignment to produce emergent flocking behaviors. While virtual-structure approaches typically demand tighter synchronization and behavior-based methods can be difficult to tune for specific geometric constraints, leader–follower architectures strike an attractive balance between simplicity, scalability, and predictability, motivating their selection in this work.

Classical control techniques for formation maintenance and navigation include Artificial Potential Fields (APF) Khatib [1986], which remain attractive due to their conceptual simplicity but are vulnerable to local minima in cluttered environments, and Model Predictive Control (MPC) Nascimento *et al*. [2013], which can yield near-optimal trajectories but at potentially prohibitive computational cost. Hybrid schemes that combine these classical methods with learning-based components have been investigated to leverage the strengths of each paradigm, though many such efforts focus on the homogeneous-agent case and do not explicitly address the asymmetric computational profile of leader–follower teams.

Deep Reinforcement Learning has more recently emerged as a powerful alternative for robot navigation Tai *et al*. [2017]; Zhu and Zhang [2021]. Everett *et al*. [2018] proposed a decentralized collision avoidance strategy based on DRL, demonstrating that policies trained in simulation can generalize to real-world pedestrian-rich environments. Hung and Regnier [2019] applied reinforcement learning to train follower agents directly, showing that learned followers can achieve performance comparable to hand-crafted controllers. More recent studies have explored DRL-based formation control under uncertainty Wang *et al*. [2022] and heterogeneous dynamics Li *et al*. [2023], often by training the entire team end-to-end in centralized critic frameworks. However, end-to-end multi-agent training typically incurs substantial sample complexity and may require careful credit assignment to avoid degenerate cooperative behaviors.

In contrast to these end-to-end approaches, our method assigns learning-based navigation exclusively to the leader while maintaining a lightweight proportional–derivative controller for the follower. This asymmetric design has three motivations: (i) it dramatically reduces the training burden, since only a single agent learns a policy; (ii) it enables a clear separation of concerns, with the learned policy handling the most cognitively demanding sub-task (obstacle avoidance) and the classical controller handling the more predictable subtask (formation tracking); and (iii) it allows the introduction of cooperative safety constraints through reward shaping rather

than through complex multi-agent learning algorithms. The integration of Dueling DQN Wang *et al.* [2016], Prioritized Experience Replay Schaul *et al.* [2015], Frame Stacking, a 21-ray simulated LiDAR observation space, and Golden Ratio action discretization further distinguishes the proposed method from prior work, providing a concrete recipe for cooperative navigation while preserving computational simplicity for follower agents.

3 Problem Formulation

Consider a two-robot system operating in a planar workspace $\mathcal{W} \subset \mathbb{R}^2$ populated with a finite set of static obstacles $\mathcal{O} = \{O_1, \ldots, O_k\}$. The team comprises a Leader robot (L) and a Follower robot (F), both modeled as differential-drive vehicles with non-holonomic kinematic constraints. At any time t, the pose of robot $i \in \{L, F\}$ is denoted $[x_i(t), y_i(t), \theta_i(t)]^T$, where (x_i, y_i) is the position in the global frame and θ_i is the heading angle. The control inputs of each robot are the linear velocity v_i and the angular velocity ω_i, subject to physical saturation bounds.

3.1 Formation Objective

The formation objective requires the follower to maintain a desired Euclidean distance $d_{des} = 0.90\,\text{m}$ and a relative bearing $\psi_{des} = \pi$ (directly behind) with respect to the leader at all times. To implement this objective, a virtual target pose $P_{des}(t)$ is generated from the leader's current state according to

$$P_{des}(t) = \begin{bmatrix} x_L(t) - d_{des}\cos(\theta_L(t)) \\ y_L(t) - d_{des}\sin(\theta_L(t)) \end{bmatrix}, \tag{1}$$

which places the virtual target at a fixed offset behind the leader along its heading direction.

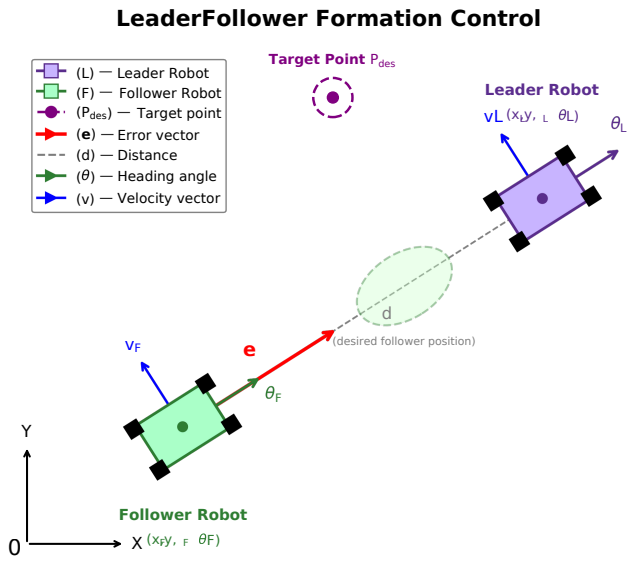

Figure 1. Leader–follower formation geometry. The follower tracks a virtual target P_{des} located at distance d_{des} behind the leader along its heading direction. The error vector $\mathbf{e}$ represents the deviation to be minimized.

The follower's control objective is to drive the tracking error $\mathbf{e} = P_{des} - [x_F, y_F]^T$ to zero. Simultaneously, the leader operates as a reinforcement learning agent navigating toward a goal $G = [x_g, y_g]^T$ while avoiding the obstacles in $\mathcal{O}$. The leader's success is conditioned not only on reaching the goal but also on ensuring that the follower does not collide with any obstacle during the traversal, which couples the optimization problem of the leader with the safety of the follower. The complete specification of the leader's state space, action space, and reward function is presented in Section 4.

4 Proposed Solution

The proposed hybrid architecture decouples navigation complexity between the two agents: the leader handles obstacle avoidance and high-level path planning through a learned Dueling DQN policy, while the follower executes deterministic proportional–derivative tracking of the virtual target defined in Equation (1). This division of labor balances expressiveness with tractability and provides a clean substrate for the introduction of cooperative reward shaping.

4.1 Leader Control Policy via Dueling DQN

4.1.1 State Space

To promote policy generalization across geometries and initial conditions, the state vector is defined entirely in the leader's local reference frame. At each time step t, the single-frame observation $o_t \in \mathbb{R}^{36}$ is structured as

$$\begin{aligned} o_t = [\, & e_x, e_y, d_{goal}, d_{pair}, \sin(\theta_L), \cos(\theta_L), \\ & \sin(\theta_F), \cos(\theta_F), \psi_{head}, I_{breach}, I_{door}, \\ & v_{t-1}, \omega_{t-1}, \sin(\alpha_F), \cos(\alpha_F), l_1, \ldots, l_{21}\,]^T, \end{aligned} \tag{2}$$

where e_x and e_y denote the waypoint positional errors, d_{goal} the distance to the goal, d_{pair} the inter-robot distance, ψ_{head} the heading error, I_{breach} and I_{door} binary checkpoint indicators signaling traversal of the gap and door respectively, v_{t-1} and ω_{t-1} the most recent actions, α_F the relative bearing to the follower, and $l_1, \ldots, l_{21}$ normalized LiDAR readings spanning $[-60°, +60°]$ at a 3.0 m range. The use of trigonometric encodings for angular variables avoids discontinuities at $\pm\pi$, and the use of binary checkpoint indicators provides an explicit signal regarding the leader's progress through the most geometrically constrained portions of the environment.

To further enhance temporal awareness, Frame Stacking concatenates the three most recent observations into a 108-dimensional state vector. This temporal context supplies the agent with implicit velocity and acceleration cues, enabling proactive response to leader–follower kinematic resonance that would be invisible from a single-frame snapshot.

4.1.2 Action Space

The Dueling DQN requires a discrete action space. While continuous-action algorithms (e.g., DDPG, SAC, PPO) offer in principle unlimited resolution, they often suffer from instability during the early exploration phase near obstacles, where a single ill-chosen action can result in a catastrophic collision. We therefore deliberately discretize the velocity space into $K = 40$ predefined pairs (v_i, ω_j), composed of 5 linear velocities and 8 angular velocities derived from the Golden Ratio ($\phi \approx 1.618$):

$$v_i \in \{0,\ 0.5\phi^{-3},\ 0.5\phi^{-2},\ 0.5\phi^{-1},\ 0.5\}\ \text{m/s}, \tag{3}$$

$$\omega_j \in \{\pm\phi^{-3},\ \pm\phi^{-2},\ \pm\phi^{-1},\ \pm 1.0\}\ \text{rad/s}. \tag{4}$$

This non-linear discretization concentrates resolution at low velocities, which is critical for fine-grained maneuvering in

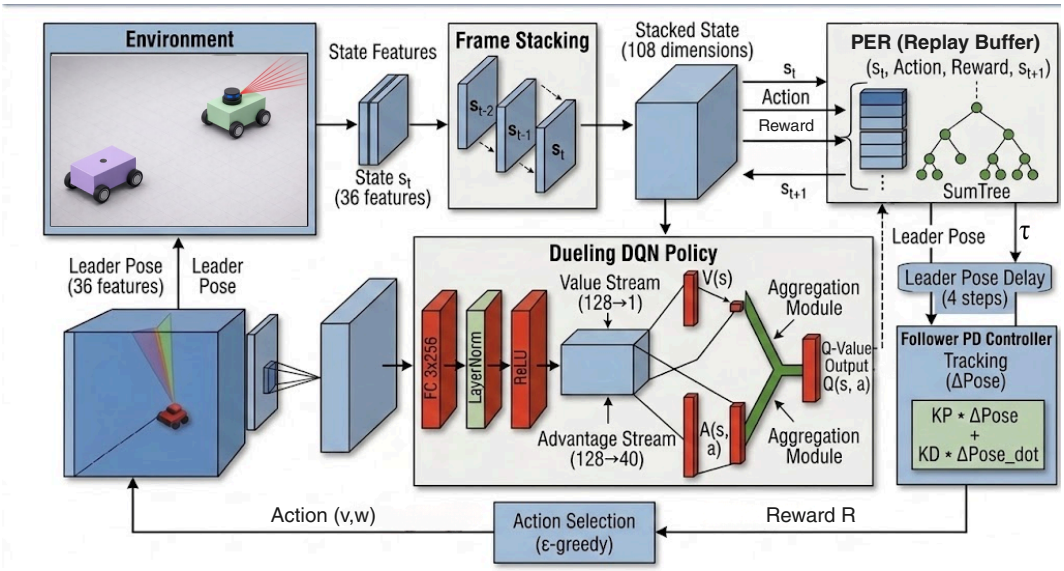

Figure 2. System architecture overview, depicting the data flow from sensors to the Dueling DQN policy, the follower's PD controller, and the environment.

narrow passages, while preserving full-speed capability for open-area traversal. The Golden Ratio is adopted because of its self-similar geometric properties, which empirically yield smoother transitions between adjacent action choices than uniform discretization.

4.1.3 Network Architecture

The action-value function is approximated by the Dueling DQN architecture Wang *et al.* [2016], which decomposes the Q-value into a state-value term and an advantage term, thereby allowing the network to evaluate global spatial risks without explicitly assessing every individual action:

$$Q(s,a) = V(s;\theta,\beta) + A(s,a;\theta,\alpha) - \frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}} A(s,a';\theta,\alpha). \tag{5}$$

The 108-dimensional input (36 features $\times$ 3 stacked frames) passes through a shared feature extractor consisting of three fully connected hidden layers with 256 neurons each, employing Layer Normalization Ba *et al.* [2016] and ReLU activations after every layer. The output bifurcates into a value stream $V(s) \in \mathbb{R}$ and an advantage stream $A(s,a) \in \mathbb{R}^{40}$, each implemented as Linear($256 \to 128$) $\to$ ReLU $\to$ Linear($128 \to \cdot$). The resulting network contains approximately 232k trainable parameters, making it small enough to be trained on commodity GPUs within reasonable time budgets.

Training uses the Smooth L_1 (Huber) loss optimized against target network predictions:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\Big[\big(r + \gamma \max_{a'} Q(s',a';\theta^-) - Q(s,a;\theta)\big)^2\Big]. \tag{6}$$

The Huber loss is preferred over plain mean-squared error because it attenuates the influence of outlier transitions, which is particularly beneficial in sparse-reward navigation tasks where occasional large TD errors are common.

To expedite convergence, minibatches are drawn using Prioritized Experience Replay (PER) Schaul *et al.* [2015]. The priority probability p_i of each transition is proportional to its temporal-difference error δ_i raised to the exponent α_{PER}, and importance-sampling weights w_i correct the resulting distribution bias:

$$p_i = \frac{|\delta_i|^{\alpha_{PER}}}{\sum_k |\delta_k|^{\alpha_{PER}}} \quad ; \quad w_i = \left(\frac{1}{N\, p_i}\right)^{\beta}. \tag{7}$$

The optimizer is Adam with learning rate $\alpha = 5 \times 10^{-4}$, accompanied by gradient clipping at $\|\nabla\| \le 10$ and Polyak (soft) updates with $\tau = 0.005$ for the target network. The combination of PER, soft updates, and gradient clipping is essential to maintain stable learning over the 40,000 training episodes.

4.1.4 Reward Function

Central to the proposed architecture is a follower-aware reward shaping strategy that couples the leader's optimization objective with the follower's spatial safety. Beyond standard navigation incentives, the reward encodes cooperative constraints via the follower-aware safety penalty, penalizing the leader whenever the follower approaches an obstacle and thereby incentivizing path choices that accommodate the follower's kinematic limitations. The full dense reward at each time step t is given by

$$r_t = r_{\text{prog}} + r_{\text{orient}} + r_{\text{vel}} - r_{\text{smooth}} - r_{\text{form}} - r_{\text{align}} - r_{\text{lidar}} - r_{\text{follower\_safety}}, \tag{8}$$

where each component is defined as follows:

- $r_{\text{prog}} = \text{clip}(5\,\Delta d_{goal}, -1, 1)$: progress toward the current waypoint, where $\Delta d_{goal} = d_{goal}^{t-1} - d_{goal}^{t}$;
- $r_{\text{orient}} = 0.1\cos(\psi)$: heading alignment bonus;
- $r_{\text{vel}} = 0.2\, v\cos(\psi)$: linear velocity projected onto the goal direction;
- $r_{\text{smooth}} = 0.15\,|\omega_t - \omega_{t-1}|$: anti-resonance penalty discouraging abrupt angular velocity oscillations;
- $r_{\text{form}} = 0.3\,\min(|d_{pair} - d_{des}|, 1)$: formation distance error, saturated at one meter to bound the contribution of large transient deviations;
- $r_{\text{align}} = 0.1\,|\theta_L - \theta_F|$: leader–follower heading misalignment;

- $r_{\text{lidar}} = 0.5 \cdot (1 - \min_i(l_i))^4$: continuous proximity penalty based on the 21-ray LiDAR, growing super-linearly as obstacles approach;
- $r_{\text{follower\_safety}} = \begin{cases} 0.5, & \text{if } d_{F,obs} < 0.40 \\ 0, & \text{otherwise} \end{cases}$, a shared-safety penalty applied to the leader whenever the follower's distance to the nearest obstacle $d_{F,obs}$ falls below a 0.40 m safety buffer, thereby encouraging the leader to select wider paths that respect the follower's kinematic envelope.

Terminal rewards are assigned upon episode termination: $+50$ if the goal is reached, -5 on timeout, and -10 on collision. Additionally, checkpoint bonuses ($+10$ for gap traversal, $+20$ for door traversal) provide sparse intermediate guidance that helps overcome the credit assignment difficulty inherent to long-horizon navigation tasks.

4.2 Follower Control Law

The follower tracks the virtual target located at distance d_{des} behind the leader. The tracking error is first transformed to the follower's local frame, and the proportional–derivative control law then computes

$$v_F = k_{v,p}\, e_x + k_{v,d}\, \dot{e}_x, \tag{9}$$

$$\omega_F = k_{\omega,p}\, \arctan\left(\frac{e_y}{e_x}\right) + k_{\omega,d}\, \dot{e}_\theta, \tag{10}$$

where $k_{v,p} = 0.8$ and $k_{\omega,p} = 1.4$ are the proportional gains, and $k_{v,d} = k_{\omega,d} = 0.1$ are the derivative gains for damping. Linear and angular velocities are saturated at $v_{max} = 0.45$ m/s and $\omega_{max} = 1.0$ rad/s respectively, ensuring kinematic feasibility. To realistically model wireless inter-robot communication, the follower receives the leader's pose with a simulated communication delay of $\tau_{delay} = 4$ time steps (0.8 s at $\Delta t = 0.05$ s $\times$ 4 sub-steps). This lightweight controller intentionally concentrates computational demand on the leader, in line with the asymmetric philosophy of the proposed architecture.

5 Simulation Results

To validate the architectural choices underlying the proposed system, we adopt an ablation-driven evaluation strategy. Rather than comparing against external baselines—which would introduce confounding variables such as differing reward structures, training budgets, network sizes, and environment assumptions—we systematically isolate the contribution of each architectural component by selectively disabling it and re-training the resulting variant under identical conditions. This methodology provides direct, interpretable evidence for each design decision and helps disentangle the impact of the proposed follower-aware reward shaping from the well-known benefits of standard DRL techniques such as Dueling decomposition and Prioritized Experience Replay.

5.1 Experimental Setup

The simulation environment, depicted in Figure 3, consists of an 8×8 m workspace with a 2.0 m inner gap and a 1.0 m lateral exit door. The leader is tasked with navigating from its initial pose to a goal located beyond the door, succeeding only if the final distance to the goal is within 0.3 m. The safety distances are set to $d_{collision} = 0.50$ m (which triggers episode termination) and $d_{safe} = 0.65$ m (which triggers a proximity penalty without termination).

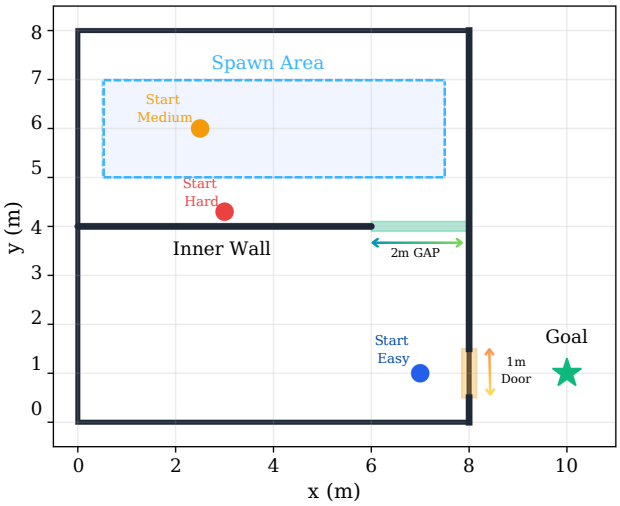

Figure 3. Simulation environment, illustrating the inner wall with the central gap, the lateral exit door, and the goal beyond.

The follower is initialized within 0.80 m of the leader, in a random relative pose. Training and simulation parameters are summarized in Table 1.

Table 1. Training and Simulation Parameters.

Symbol	Parameter	Value
γ	Discount factor	0.99
$\epsilon_{\min}$	Minimum exploration	0.02
α	Learning rate (Adam)	5×10^{-4}
N_{ep}	Total training episodes	40,000
τ	Soft update rate (Polyak)	0.005
N_{env}	Parallel environments	12
Cap	Replay buffer capacity	100,000
d_{des}	Desired formation distance	0.90 m
Batch	Mini-batch size	128
$d_{collision}$	Collision distance	0.50 m
α_{PER}	PER prioritization exponent	0.6
d_{safe}	Minimum safe distance	0.65 m
β_{start}	PER bias correction (initial)	0.4
$(k_{v,p}, k_{\omega,p})$	Follower P-gains	(0.8, 1.4)
ϵ_{start}	Initial exploration	1.0
$(k_{v,d}, k_{\omega,d})$	Follower D-gains	(0.1, 0.1)
ϵ_{decay}	Exploration decay (per episode)	0.9997
$\|\nabla\|_{max}$	Gradient clip norm	10.0

5.1.1 Experimental Protocol

We validate the system using three fixed-geometry configurations (Easy, Medium, Hard) and a procedural Generalization test in which the inner wall's y-position varies in $[3.0, 5.0]$ m, and both the gap and the door positions shift randomly within available bounds. Training employs a three-phase Curriculum Learning strategy combined with progressive Domain Randomization Tobin *et al.* [2017], structured as follows:

- **Phase 1** (0–10k episodes): a fixed Easy geometry (3.0 m gap, 2.0 m door) with a relaxed collision margin of 0.35 m allows the agent to learn the basic skills of progress, obstacle avoidance, and formation maintenance.
- **Phase 2** (10k–30k episodes): the domain randomization strength is linearly interpolated from 0 to 1, the gap

shrinks to 2.0 m, the door narrows to 1.0 m, and the collision margin is tightened to 0.50 m, forcing the agent to adapt to progressively harder conditions.

- **Phase 3** ($>$ 30k episodes): full randomization under Hard conditions, with the geometry drawn from the most challenging end of the configuration space at every reset.

5.2 Ablation Study

Table 2 reports the ablation results obtained from 500 evaluation episodes on the Hard configuration, isolating the impact of each component on the system's ability to solve the most geometrically severe environments.

Table 2. Ablation Study Results (Hard Configuration).

Agent	Success (%)	Collisions (%)	Avg Time (s)	Final Dist (m)
Full System	100.0	0.0	22.6 ± 0.5	0.25 ± 0.02
No dueling	71.2	28.8	30.1 ± 0.8	0.60 ± 0.54
No golden ratio	0.0	100.0	–	5.63 ± 0.23
No per	100.0	0.0	23.8 ± 0.7	0.26 ± 0.03
No stacking	0.0	100.0	–	7.58 ± 0.14
No follower safety	0.0	81.4	–	6.52 ± 0.22

Removing Frame Stacking collapses success to 0% with a 100.0% collision rate, confirming that temporal awareness is essential for the leader to anticipate follower inertia and plan accordingly. Removing the follower-aware safety penalty similarly yields 0% success and an 81.4% collision rate, demonstrating that without explicit follower-safety incentives the leader consistently selects paths that are kinematically infeasible for the trailing unit. Removing the Golden Ratio action discretization collapses success to 0% (100.0% collision rate), highlighting that fine-grained action resolution near obstacles is a critical bottleneck. Finally, removing PER yields 100.0% success, suggesting that PER primarily accelerates convergence rather than enabling qualitatively different final behavior.

For reference, an Artificial Potential Field controller, evaluated under identical environmental conditions, achieved 0% success across 500 episodes, consistent with the well-known local-minima failures of APF in non-convex geometries. This comparison is presented for illustrative purposes only, as APF does not share the same training infrastructure as the learned agents.

Collectively, the ablation results establish that the 100% success rate of the full system is not attributable to any single mechanism, but rather to the synergistic interaction of temporal awareness, follower-aware reward shaping, and fine-grained action resolution. The follower-aware reward shaping strategy emerges as the single most critical contribution: its removal produces the worst observed behavior across all experiments, outweighing in behavioral impact even established techniques such as PER and dueling decomposition. This criticality also extends to unseen environments. In the procedural generalization tests, removing the follower-aware safety penalty collapses the success rate to 0% (down from 74.0% in the full system), with a 58.6% collision rate, confirming that the proposed reward shaping is essential for robust adaptability under geometric variation.

5.3 Trajectory Analysis

Trajectory inspection (Fig. 4) reveals the emergence of clear cooperative behavior. In the Hard configuration in particular, rather than exploiting the mathematically shortest route, the leader deliberately aligns perpendicularly adjacent to the internal wall before approaching the gap. This maneuver preemptively creates geometric clearance favoring the follower's wider turning arc, providing strong qualitative evidence that the multi-component reward function effectively induces implicit follower-spatial awareness in the leader's policy. Such emergent behavior is non-trivial: it cannot be observed in the ablation variants lacking the follower-aware penalty, where the leader gravitates toward shorter but kinematically hostile trajectories.

5.4 Follower Analysis

The systemic resilience of the proportional–derivative tracker depends strongly on the operational communication latency between the two robots. To map the boundary conditions of tracking stability, we simulated varying observation delays $\tau_{delay} \in \{0, 2, 4, 6, 8\}$ time steps and recorded the resulting formation error $|d_{pair} - d_{des}|$ across the episode.

Although mild latencies pose no substantial threat to the overall formation, increasing periods of communication blackout trigger widening deviations in $|d_{pair} - d_{des}|$, especially following abrupt directional shifts by the leader. This observation ultimately confirms the necessity of the leader's proactive collision-clearance maneuvers: without them, even small delays could cause the follower to enter safety-critical margins. The interplay between leader-side prevention and follower-side latency tolerance thus emerges as a defining characteristic of the proposed system.

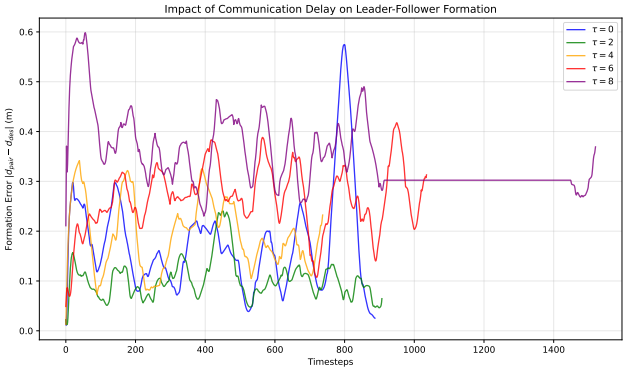

Figure 5. Impact of communication delay on the leader–follower formation error over time, for delays of $\tau_{delay} \in \{0, 2, 4, 6, 8\}$ steps.

6 Limitations and Improvements

Despite the high success rates achieved during the evaluations, a core limitation of this study lies in its strictly numerical and simulation-based nature. Virtual environments operate under idealized conditions that frequently fail to faithfully replicate the dynamic uncertainties of the physical world, such as sudden friction variations, imperfections in real LiDAR sensors, and stochastic fluctuations in communication latency. To bridge this gap, the next critical step for enhancing this system involves transitioning and validating the hybrid architecture on physical robotic platforms. This validation will be conducted using the ROS 2 framework, enabling the assessment of the shaped policy's behavior and robustness under real-world constraints and disturbances.

Additionally, the reliance on a discretized action space

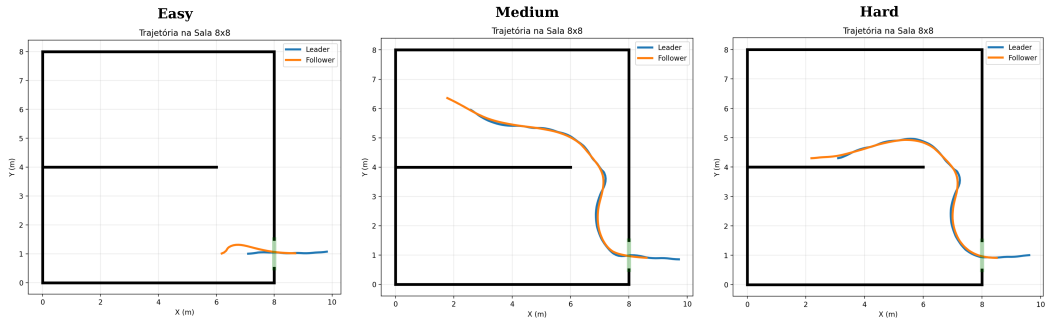

Figure 4. Leader–follower trajectories obtained in all evaluation cases. Note the deliberate widening of the leader's path adjacent to the inner wall in the Hard scenario, induced by the follower-aware reward shaping.

and an empirically weighted reward function imposes constraints on the system's optimal performance. Although the Golden Ratio-based discretizations provides refined control essential for complex maneuvers near obstacles, it restricts motion fluidity compared to continuous action spaces. Furthermore, the manual tuning of the reward components' weights demands substantial trial-and-error effort. As a future improvement, we plan to investigate Reinforcement Learning formulations in continuous action spaces to eliminate these discretizations barriers, as well as the application of Multi-Objective Reinforcement Learning (MORL) techniques to systematically optimize and explore the intrinsic trade-offs between follower safety and leader navigation efficiency.

Finally, the currently proposed architecture exempts the follower robot from any independent environmental perception, centralizing hazard detection exclusively within the leader and the reward penalty mechanism. Should the follower encounter external disturbances or severe communication failures that divert it from the ideal trajectory, its lack of local obstacle perception precludes autonomous emergency evasion maneuvers. Another theoretical limitation of the current model is the absence of a formal proof connecting the empirical safety function with stable asymptotic convergence. To address these deficiencies, future enhancements will focus on integrating low-computational, decentralized policies into the follower to provide it with basic obstacle awareness, alongside formally mapping the proposed reward shaping strategy to the consolidated theoretical principles of potential-based reward shaping.

7 Conclusion

This paper presented a follower-aware reward shaping strategy for leader–follower formation control, implemented within a hybrid architecture in which navigation and obstacle avoidance are handled by a Dueling DQN-based leader agent while a dedicated proportional–derivative controller governs follower tracking. The architecture combines a 21-ray simulated LiDAR observation space, Frame Stacking for implicit temporal awareness, Prioritized Experience Replay, Golden Ratio action discretization, and the proposed follower-aware safety penalty, which represents the central contribution of this work.

Ablation studies validated the follower-aware reward shaping strategy as the primary driver of safe cooperative navigation: the follower-aware safety penalty alone accounts for the difference between 0% and 100% success in the most challenging configuration. Quantitative results demonstrated 100% success across all fixed-geometry configurations and 74% generalization in procedural environments, verifying that the Dueling framework and Frame Stacking jointly support but do not substitute the contribution of the reward shaping mechanism. Trajectory analysis further revealed the emergence of cooperative behaviors, with the leader deliberately taking longer paths to accommodate the follower's kinematic envelope.

Future efforts will focus on transitioning this architecture from simulation to physical robot platforms using ROS 2, with experiments conducted in more diverse and realistic environments in order to better assess the sim-to-real transfer capabilities of the proposed system. Furthermore, while the current reward structure relies on empirically tuned weightings, future work will investigate Multi-Objective Reinforcement Learning (MORL) approaches that can systematically explore and optimize the trade-offs between the different reward components. We also intend to formally connect our empirical follower-aware reward shaping with the theoretical guarantees of potential-based reward shaping, and to evaluate continuous-action RL formulations (e.g., SAC, PPO) to eliminate the resolution boundaries imposed by discrete action spaces. Finally, we plan to address the follower's current lack of independent obstacle awareness by integrating decentralized, low-overhead learning policies, further diminishing collision rates in highly unstructured settings and paving the way to scalable multi-follower formations.

Declarations

Acknowledgements

The authors gratefully acknowledge the support of the **Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)**, the **Instituto de Computação (IComp) of the Universidade Federal do Amazonas (UFAM)**, and the **Programa Institucional de Bolsas de Iniciação Científica (PIBIC)**, whose funding, institutional infrastructure, and scholarship programs made this research possible. The authors also thank the colleagues and faculty of IComp/UFAM for their valuable feedback during the development of this work.

Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. The authors would like to thank the National Council for Scientific and Technological Development (CNPq) for the financial support through the Programa Institucional de Bolsas

de Iniciação Científica (PIBIC). Natasha Araujo Caxias acknowledges support from grant #PIB-E/0106/2025. Abel Severo Rocha acknowledges support from grant #PIB-E/0108/2025. This research was carried out at the Instituto de Computação (IComp) of the Universidade Federal do Amazonas (UFAM).

Authors' Contributions

Natasha Caxias: Formal analysis, Methodology, Investigation, Writing –review & editing; Abel Severo: Formal analysis, Investigation Writing – review& editing; Reginaldo Carvalho: Writing – review & editing, Supervision All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets and softwares generated and/or analysed during the current study will be made available upon request.

References

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization.

Everett, M., Chen, Y. F., and How, J. P. (2018). Motion planning among dynamic, decision-making agents with deep reinforcement learning. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3052–3059.

Fragapane, G., de Koster, D., Sgarbossa, F., and Strandhagen, J. O. (2021). Planning and control of autonomous mobile robots for intralogistics: Literature review and research agenda. *Journal of Manufacturing Systems*, 59:132–150.

Hung, H. Q. and Regnier, F. (2019). Reinforcement learning for leader-follower formation control of mobile robots. In *Proceedings of the International Conference on System Science and Engineering (ICSSE)*, pages 137–141.

Huntsberger, T., Pirjanian, P., Trebi-Ollennu, A., Nayar, H. D., Aghazarian, H., Ganino, A., Garrett, M., Joshi, S. S., and Schenker, P. S. (2003). Campout: A control architecture for tightly coupled coordination of multirobot systems for planetary surface exploration. *IEEE Transactions on Systems, Man, and Cybernetics—Part A*, 33(5):550–559.

Khatib, O. (1986). Real-time obstacle avoidance for manipulators and mobile robots. *International Journal of Robotics Research*, 5(1):90–98.

Li, J. *et al*. (2023). Deep reinforcement learning for collision avoidance in heterogeneous non-holonomic teams. *Robotics and Autonomous Systems*, 162:104368.

Mnih, V., Kavukcuoglu, K., Silver, D., *et al*. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Nascimento, T. P., Moreira, A. P., and Conceicao, A. G. S. (2013). Multi-robot nonlinear model predictive formation control: Moving target and target absence. *Robotics and Autonomous Systems*, 61(12):1502–1515.

Oh, K.-K., Park, M.-C., and Ahn, H.-S. (2015). A survey of multi-agent formation control. *Automatica*, 53:424–440.

Rizk, Y., Awad, M., and Tunstel, E. W. (2019). Cooperative heterogeneous multi-robot systems: A survey. *ACM Computing Surveys*, 52(2):1–31.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. (2015). Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.

Siegwart, R., Nourbakhsh, I. R., and Scaramuzza, D. (2011). *Introduction to Autonomous Mobile Robots*. MIT Press, Cambridge, MA, 2 edition.

Tai, L., Paolo, G., and Liu, M. (2017). Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 31–36.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30.

Wang, X. *et al*. (2022). Leader-follower formation control of multiple nonholonomic mobile robots with deep reinforcement learning. *IEEE Transactions on Industrial Informatics*, 18(10):6683–6692.

Wang, Z., Schaul, T., Hessel, M., van Hasselt, H., Lanctot, M., and de Freitas, N. (2016). Dueling network architectures for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1995–2003.

Zhu, K. and Zhang, T. (2021). Deep reinforcement learning based mobile robot navigation: A review. *Tsinghua Science and Technology*, 26(5):674–691.