Hybrid Reinforcement Learning Architecture with Geometric Curriculum Learning for Non-Holonomic Leader-Follower Navigation

Natasha Araújo Caxias , Abel Severo Rocha , José Reginaldo Hughes Carvalho

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

{natasha.caxias}{abel.severo}{reginaldo}@icomp.ufam.edu.br

Abstract. *This paper proposes a hybrid architecture for leader-follower non-holonomic navigation, combining a Dueling Deep Q-Network (DDQN), Prioritized Experience Replay (PER), Frame Stacking, and Golden Ratio action discretization. A three-phase Geometric Curriculum Learning strategy addresses sparse rewards and catastrophic forgetting, progressing from fixed geometries to full Domain Randomization. Ablation studies confirm that Frame Stacking and Golden Ratio discretization are strictly necessary, their removal collapses success to 0%, while PER acts as a convergence accelerator rather than a behavior enabler. The full system achieves 100% success on fixed configurations and 74% on procedural generalization.*

1. Introduction

Autonomous navigation of non-holonomic robots in cluttered environments remains challenging due to kinematic constraints [Siegwart et al. 2011]. Leader-follower architectures offer a modular solution, concentrating navigation intelligence in the leader while the follower applies deterministic tracking [Oh et al. 2015, Wang et al. 2022]. Applying DRL directly to a narrow-passage geometry, however, results in severe sample inefficiency: the agent rarely discovers the navigation sequence by random exploration, and successful trajectories are diluted by millions of collision experiences. Architectural extensions to the baseline DQN [Mnih et al. 2015] address several limitations: Double Q-learning [Hasselt et al.] mitigates overestimation bias, dueling networks [Wang et al. 2016] decompose state values from action advantages, and Prioritized Experience Replay [Schaul et al. 2016] improves sample efficiency. Complementing these, Curriculum Learning [Bengio et al. , Narvekar et al. 2020] and Domain Randomization [Tobin et al. 2017] jointly address sparse rewards and generalization. The specific utility of these components for narrow-passage leader-follower navigation, however, has not been systematically validated.

The main contributions are: i) A hybrid architecture combining Dueling DDQN, PER, Frame Stacking, and Golden Ratio action discretization; ii) A three-phase Geometric Curriculum Learning strategy that progressively exposes the agent to increasingly constrained and randomized environments, enabling robust policy generalization; iii) Quantitative ablation evidence that Frame Stacking and action space resolution are the most critical architectural components, while PER's contribution is primarily to convergence speed.

2. Proposed Architecture

Dueling DDQN network architecture is depicted in Figure 1. A shared 3-layer FC extractor (256 neurons, LayerNorm + ReLU) feeds into a Value stream ($128 \rightarrow 1$) and an Advantage stream ($128 \rightarrow 40$). Eq. 3 combines both streams into Q-values for all 40 discrete actions. This hybrid architecture decouples navigation complexity: the leader handles obstacle avoidance through a learned Dueling DQN policy, while the follower executes deterministic proportional-derivative tracking. This assigns computational and learning complexity to the leader, enabling lightweight follower implementations for resource-constrained platforms.

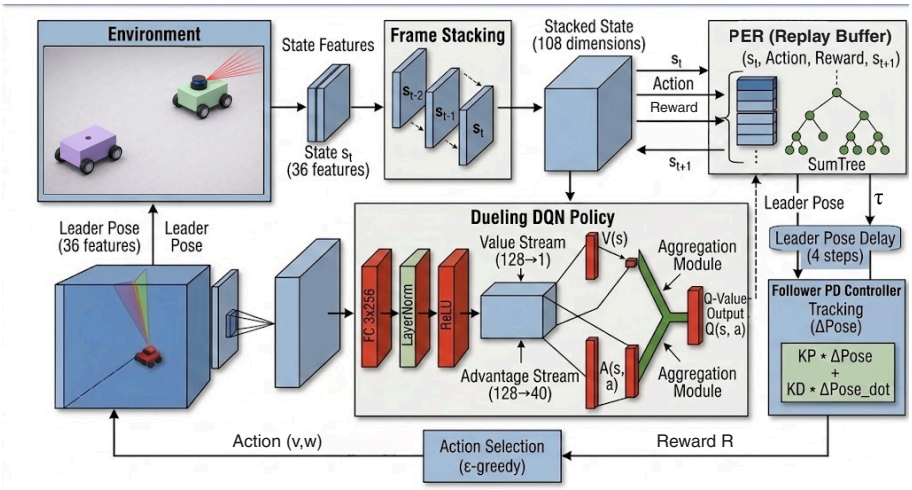

Figure 1. The proposed Dueling DDQN network architecture.

2.1. Observation Space and Frame Stacking

The single-frame observation $o_t \in \mathbb{R}^{36}$ encodes positional errors (e_x, e_y), goal and inter-robot distances, heading angles (sin/cos), door and gap breach indicators, previous velocity commands, follower bearing, and 21 normalized LiDAR readings spanning $[-60°, +60°]$ at 3.0 m range. Frame Stacking concatenates the $N{=}3$ most recent observations into a 108-dimensional state vector $s_t = [o_t;\, o_{t-1};\, o_{t-2}]$, providing implicit velocity and acceleration cues. This lightweight temporal mechanism avoids LSTM overhead while supplying sufficient inertial context for reactive braking near obstacles.

2.2. Golden Ratio Action Discretization

A principled discrete action space of $K = 40$ pairs (v_i, ω_j) is constructed using the Golden Ratio $\phi \approx 1.618$:

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

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

This parametrization provides exponentially denser resolution near zero velocities, critical for precise maneuvers near obstacles, while maintaining sufficient maximum velocities for open-area traversal.

2.3. Dueling DDQN Architecture

The action-value function is approximated by the Dueling DQN architecture [Wang et al. 2016], which decomposes the Q-value:

$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + \left(A(s,a;\theta,\alpha) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s,a';\theta,\alpha) \right) \tag{3}$$

The 108-dimensional input passes through three fully connected layers of 256 neurons with Layer Normalization and ReLU activations, bifurcating into a value stream $V(s) \in \mathbb{R}$ and an advantage stream $A(s,a) \in \mathbb{R}^{40}$ (Fig. 1). Double Q-learning [Hasselt et al.] is used for target computation. Parameter updates use Adam (α=5×10^{-4}), gradient clipping ($\|\nabla\| \leq 10$), and Soft Updates (τ=0.005).

2.4. Prioritized Experience Replay and Follower Controller

PER [Schaul et al. 2016] weights transitions proportionally to absolute TD-error (α_{PER}=0.6), with importance-sampling (β annealed $0.4\rightarrow1.0$) to correct distribution bias. High-TD-error transitions, corresponding to surprising successes or unexpected collisions, are replayed more frequently, accelerating convergence in early training phases. The follower tracks a virtual target at d_{des}=0.90 m behind the leader via PD control:

$$v_F = k_{v,p} e_x + k_{v,d} \dot{e}_x, \quad \omega_F = k_{\omega,p} \arctan\left(\frac{e_y}{e_x}\right) + k_{\omega,d} \dot{e}_\theta \tag{4}$$

with $(k_{v,p}, k_{\omega,p})$=$(0.8, 1.4)$, $(k_{v,d}, k_{\omega,d})$=$(0.1, 0.1)$, $v_{\max}$=0.45 m/s, and a simulated communication delay of τ_{delay}=4 steps.

2.5. Reward Function

The dense reward at each step t is:

$$r_t = r_{\text{prog}} + r_{\text{orient}} + r_{\text{vel}} + r_{\text{check}} - r_{\text{smooth}} - r_{\text{lidar}} - r_{\text{safety}} \tag{5}$$

Table 1 summarizes each component. The follower safety penalty r_{safety} is a novel contribution that discourages kinematically infeasible leader paths by penalizing small inter-robot distances near walls.

Table 1. Dense reward components (d_{safe}=0.65 m).

Component	Symbol	Role
Progress	r_{prog}	Proportional to Δd_{goal}; drives forward motion.
Orientation	r_{orient}	Cosine of heading–goal error; promotes alignment.
Velocity	r_{vel}	Positive for nonzero forward speed; penalizes stalling.
Checkpoints	r_{check}	Discrete bonuses ($+10$ gap, $+20$ door) for milestone passage.
Smoothness	r_{smooth}	Quadratic penalty on $\|\Delta\omega\|$; suppresses oscillatory steering.
LiDAR	r_{lidar}	Quartic repulsion for $d_{obs} < d_{safe}$; anticipatory avoidance.
Follower Safety	r_{safety}	Penalty for $d_F < d_{safe}$; prevents infeasible leader paths.

3. Geometric Curriculum Learning

The task geometry evolves across three training phases [Narvekar et al. 2020, Tobin et al. 2017]. **Phase 1** (0–10k episodes) uses a fixed wide geometry (3.0 m gap, 2.0 m door) with relaxed collision margin (d_{coll}=0.35 m) and no randomization. **Phase 2** (10k–30k episodes) linearly interpolates domain randomization strength from 0 to 1 while the gap shrinks from 3.0 to 2.0 m and the door from 2.0 to 1.0 m; collision margin tightens to 0.50 m. **Phase 3** (30k+ episodes) enables full domain randomization: wall position, gap, and door locations vary randomly while door remains at 1.0 m. This progression prevents premature exposure to hard geometries, which would yield near-zero reward signal with no learning gradient, and avoids catastrophic forgetting by gradually increasing difficulty rather than switching abruptly.

4. Experimental Setup

The simulation environment is an 8×8 m 2D workspace featuring an inner wall with a variable gap and a lateral exit door (Fig. 2). The full system was trained for 40,000 episodes across 12 parallel environments (replay buffer 100k, mini-batch 128, γ=0.99, ϵ decayed 1.0→0.02). Each ablation variant was trained from scratch and evaluated over 500 episodes on the Hard configuration and a procedural Generalization test. Variants: **Full Architecture** (Dueling DDQN + PER + Frame Stacking + Golden Ratio); **No Stacking** (single frame, 36-dim); **No PER** (uniform replay); **No Dueling** (standard DQN); **No Golden Ratio** (uniform linear spacing: 5 velocities $\times$ 8 angular rates).

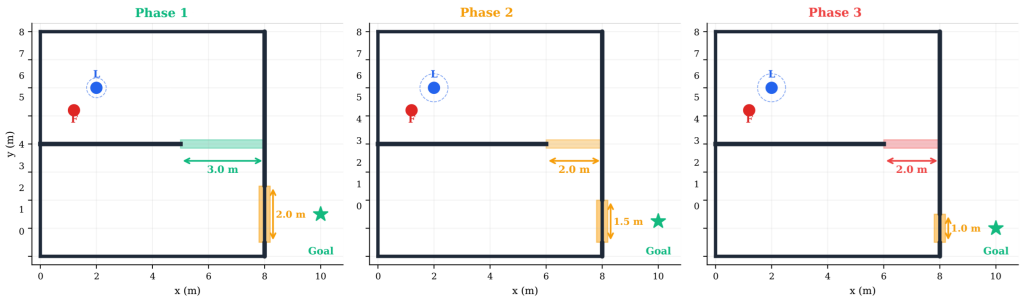

Figure 2. Top-down view of the 8×8 m workspace across the three curriculum phases. Follower (F) trails at d_{des}=0.90 m.

5. Results and Discussion

5.1. Ablation Study

Table 2 reports quantitative results on the Hard geometry; Fig. 3 shows global-average success rates across all difficulty configurations.

Frame Stacking is the most critical component: its removal collapses success to 0% with 100% collision rate across all difficulty levels. Without temporal context, the agent cannot anticipate follower kinematic inertia or compute implicit velocity cues for braking near obstacles. **Golden Ratio Discretization** is equally critical: uniform spacing collapses success to 0%, as coarse resolution near zero velocities precludes precise 1.0 m door navigation. **No Dueling** degrades success to 71.2%, confirming that value-advantage decomposition provides measurable benefit near the critical junction before

the door passage. **No PER** achieves 100% success, consistent with PER's primary role as a convergence accelerator: the No PER variant required ≈15% more episodes to reach the same reward plateau.

Table 2. Architecture Ablation Study (Hard Config., 500 Episodes per variant).

Configuration	Success (%)	Collision (%)	Avg Time (s)	Final Dist (m)
Full Architecture	**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 PER	100.0	0.0	23.8 ± 0.7	0.26 ± 0.03
No Stacking	0.0	100.0	–	7.58 ± 0.14
No Golden Ratio	0.0	100.0	–	5.63 ± 0.23

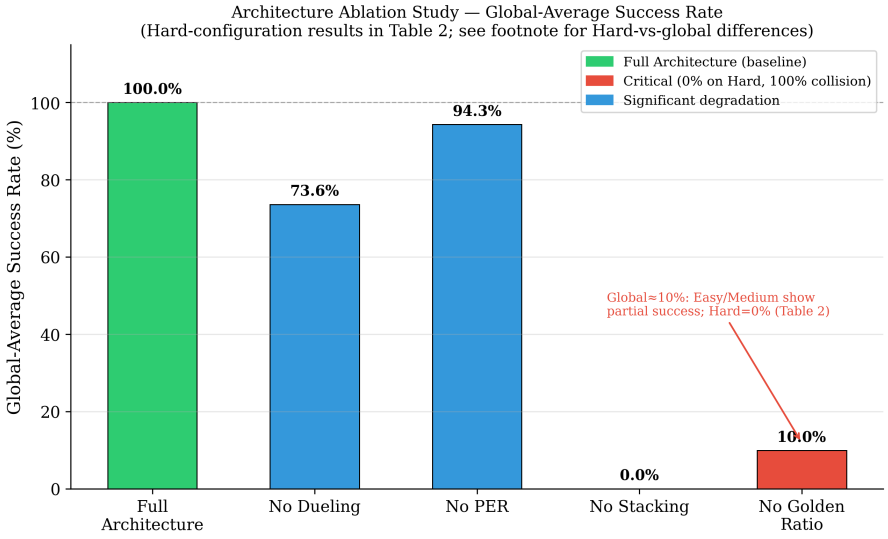

Figure 3. Global-average success rates across all evaluation configurations for each ablation variant. Red bars = Hard performance collapses to 0%.

5.2. Baseline Comparison and Curriculum Dynamics

Against a vanilla DQN and an Artificial Potential Field (APF) baseline: the APF controller achieves 0% success on all passage-constrained configurations due to local minima in non-convex geometry [Khatib 1986]; vanilla DQN reaches 68% on Easy but degrades to a global average of ≈34%, confirming that the proposed architectural extensions are necessary for realistic narrow-passage navigation. Regarding curriculum dynamics, Phase 1 enables rapid discovery of the basic navigation sequence (gap → door → goal); Phase 2 prevents catastrophic forgetting by maintaining partial environment familiarity while increasing difficulty; Phase 3 forces policy generalization beyond any single geometry instance, yielding 74% success under full domain randomization [Narvekar et al. 2020].

6. Conclusion

This paper presented a hybrid RL architecture integrating Dueling DDQN, PER, Frame Stacking, and Golden Ratio action discretization, trained via a three-phase Geometric

Curriculum Learning strategy for leader-follower non-holonomic navigation. Ablation studies identified Frame Stacking and Golden Ratio discretization as strictly necessary components, while PER contributes primarily to convergence speed. The full system achieves 100% success on fixed-geometry evaluations and 74% on procedural generalization under full Domain Randomization. Future work will address sim-to-real transfer to Pioneer 3-AT robots and extension to $N>2$ robot platoons [Zhang et al. 2024].

7. Acknowledgements

This study was partially financed CAPES – Finance Code 001, and CNPq through PIBIC (grants #PIB-E/0106/2025 and #PIB-E/0108/2025).

References

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*.

Hasselt, H. V., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence*.

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

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

Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M. E., and Stone, P. (2020). Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50.

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

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. (2016). 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.

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.

Zhang, K., Yang, Z., and Basar, T. (2024). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384.