

Multi-Agent Reinforcement Learning for Cooperative Warehouse Automation: QMIX Value Decomposition for Sparse-Reward Coordination

Price Allman, Lian Thang, Dre Simmons, Salmon Riaz

Department of Computer Science, Oral Roberts University, Tulsa, OK, USA

December 2025

Abstract

We present a comparative study of multi-agent reinforcement learning (MARL) algorithms for cooperative warehouse robotics. We evaluate QMIX and IPPO on the Robotic Warehouse (RWARE) environment and a custom Unity 3D simulation. Our experiments reveal that QMIX’s value decomposition significantly outperforms independent learning approaches (achieving 3.25 mean return vs. 0.38 for advanced IPPO), but requires extensive hyperparameter tuning—particularly extended epsilon annealing (5M+ steps) for sparse reward discovery. We demonstrate successful deployment in Unity ML-Agents, achieving consistent package delivery after 1M training steps. While MARL shows promise for small-scale deployments (2-4 robots), significant scaling challenges remain. Code and analyses: [project documentation](#).

1 Introduction

Warehouse automation represents a critical domain for multi-agent systems, where robots must coordinate to move goods without collisions or deadlocks. Traditional approaches rely on centralized planning systems, but these face scalability limitations as warehouse complexity increases. Multi-agent reinforcement learning (MARL) offers a promising alternative by enabling decentralized decision-making through learned policies.

The challenge of warehouse robotics extends beyond single-agent navigation. Robots must learn to coordinate pickup and delivery tasks, avoid collisions, navigate narrow corridors, and handle dynamic environments—all while optimizing global throughput. This requires algorithms that can effectively solve the credit assignment problem: determining how each agent’s actions contribute to team success.

We investigate MARL algorithms representing different paradigms: QMIX [Rashid et al., 2018], which uses value decomposition with a monotonic mixing network; IPPO, an independent learning approach; and MASAC, which employs a centralized critic with maximum entropy objectives. Our study progresses from the Multi-Agent Particle Environment (MPE) for initial algorithm validation to RWARE for warehouse-specific evaluation, with QMIX emerging as the primary focus after demonstrating superior performance. The project culminates in Unity 3D deployment for visual verification.

Our key contributions include: (1) systematic hyperparameter analysis revealing that default configurations fail on sparse-reward warehouse tasks, (2) comparative performance evaluation across algorithms and environment scales, (3) successful Unity ML-Agents integration demonstrating sim-to-sim transfer, and (4) identification of scaling challenges for production deployment. Complete experimental details are available in our [Quarto documentation book](#).

2 Related Work

2.1 Value Decomposition Methods

Value decomposition methods address credit assignment by factorizing the joint action-value function into individual agent utilities. VDN [Sunehag et al., 2018] assumes additive decomposition: $Q_{tot} = \sum_i Q_i$, while QMIX [Rashid et al., 2018] relaxes this to monotonic relationships through a hypernetwork-based mixing network. These methods follow the Centralized Training with Decentralized Execution (CTDE) paradigm, where agents access global information during training but act on local observations at execution time. This paradigm is particularly well-suited for warehouse robotics, where centralized coordination during training can leverage full environment state while deployed robots must act on local sensor readings.

2.2 Policy Gradient and Actor-Critic Approaches

Independent PPO (IPPO) trains separate policies for each agent, treating other agents as part of the environment. While theoretically limited by non-stationarity, IPPO has shown surprising effectiveness in cooperative settings [Yu et al., 2022]. Adding LSTM memory enables handling of partial observability, crucial for warehouse environments where agents have limited sensing range. MASAC extends soft actor-critic to multi-agent settings with centralized critics and entropy regularization, promoting exploration in sparse reward environments.

2.3 Warehouse Robotics Environments

The RWARE environment [Papoudakis et al., 2021] provides a standardized benchmark for multi-robot warehouse coordination. Agents must navigate grid-based warehouses, pick up shelves, and deliver them to designated locations. The environment features sparse rewards (only granted upon successful delivery), partial observability (agents see limited local regions), and configurable difficulty levels based on grid size and agent count.

3 Methods

3.1 Problem Formulation

We model the warehouse coordination task as a decentralized partially observable Markov decision process (Dec-POMDP). Each agent i receives local observation o_i , selects action a_i from discrete action space $\mathcal{A} = \{\text{left, right, forward, load/unload, no-op}\}$, and the team receives shared reward r based on successful deliveries. The sparse reward structure—where agents only receive positive signal upon completing multi-step delivery sequences—makes credit assignment particularly challenging.

3.2 QMIX Architecture

QMIX factorizes the joint action-value function as:

$$Q_{tot}(\tau, \mathbf{a}) = f_{mix}(Q_1(\tau_1, a_1), \dots, Q_n(\tau_n, a_n); s) \quad (1)$$

where f_{mix} is a mixing network with non-negative weights generated by hypernetworks conditioned on global state s . This ensures monotonicity: $\frac{\partial Q_{tot}}{\partial Q_i} \geq 0$, enabling consistent greedy action selection across individual and joint Q-values.

We use RNN-based agent networks (GRU with 64 hidden units) to handle partial observability, allowing agents to maintain beliefs about unobserved state. The mixing network uses 2-layer hypernetworks with 64-dimensional embeddings, providing sufficient capacity for warehouse coordination.

3.3 Training Configuration

Our optimized QMIX configuration differs substantially from defaults ([detailed analysis](#)):

Table 1: Hyperparameter Configuration: Default vs Optimized

Parameter	Default	Optimized
Batch Size	32	256
Buffer Size	5,000	200,000
Epsilon Anneal Time	50,000	5,000,000
Training Steps	2,000,000	20,000,000
Learning Rate	0.0005	0.0005

The extended epsilon annealing schedule proved critical—rapid decay to greedy behavior prevents agents from discovering sparse reward signals characteristic of warehouse logistics tasks. The larger replay buffer enables learning from diverse experiences across the extended training horizon.

3.4 Unity ML-Agents Integration

We developed a custom Unity ML-Agents environment mirroring RWARE dynamics with 3D visualization ([implementation details](#)). The environment features 3 agents with 6 discrete actions, 36-dimensional observation vectors including LIDAR-style sensing, and 200-step episodes. Training uses no-graphics mode with 50x time scaling for efficiency, enabling rapid iteration while maintaining physics fidelity.

4 Experiments and Results

4.1 Environment Progression

We validated our implementations on MPE Simple Spread before transitioning to RWARE. The transition revealed significant differences: MASAC on MPE converged in approximately 30,000 steps (200 episodes) achieving -55 mean reward, representing 63% improvement over random baselines. In contrast, RWARE required 20M+ steps for QMIX to achieve comparable task completion, with default hyperparameters producing zero learning even after 2M steps. This 600x difference in sample complexity underscores how dense-reward benchmarks like MPE can mislead practitioners about algorithm readiness for sparse warehouse tasks ([transition analysis](#)).

4.2 Algorithm Comparison

Table 2: Performance Comparison on RWARE Environments (Test Return)

Algorithm	tiny-2ag-v2	small-4ag-v1	Training Steps
QMIX	3.25	2.10	20-30M
IPPO (Advanced)	0.38	—	5M
IPPO (Vanilla)	0.13	—	20M
Random Baseline	0.05	0.02	N/A

QMIX achieved significantly higher returns than independent learning approaches. On tiny-2ag-v2, QMIX (3.25) outperformed advanced IPPO (0.38) by approximately 755%, representing an 8.5x improvement in task completion. Even vanilla IPPO with 4x longer training (20M vs 5M steps) achieved

only 0.13 mean return—25x lower than QMIX. These results demonstrate that value decomposition provides substantial advantages for sparse-reward coordination where credit assignment is challenging. Note: MASAC was evaluated on MPE (achieving 63% improvement over random baseline) but not on RWARE ([MPE results](#)). Detailed learning curves are available in our [training analysis](#).

4.3 Scaling Analysis

Performance degrades with increased agent count, while training requirements grow super-linearly ([scaling discussion](#)):

Table 3: QMIX Scaling Results Across Environment Sizes

Configuration	Agents	Test Return	Required Steps
tiny-2ag-v2	2	3.25	20M
small-4ag-v1	4	2.10	30M
medium-6ag-v1	6	1.45	40M

Scaling from 2 to 6 agents (3x increase) requires 2x more training steps (20M to 40M) while performance drops 55% (3.25 to 1.45 mean return). The effective cost—training time per unit performance—increases approximately 4–5x. The joint action space grows as $|A|^n$: with 5 discrete actions, 6 agents yield 15,625 joint actions versus 25 for 2 agents, making exhaustive exploration infeasible and suggesting the need for hierarchical or factored approaches at scale.

4.4 Unity Deployment Results

We validated trained QMIX policies in Unity on a Windows Server 2022 environment (Intel Xeon E5-2680, 196GB RAM). Two training runs demonstrated successful learning ([deployment details](#)):

Table 4: Unity Training Results on Windows Server

Metric	500K Steps	1M Steps
Test Return Mean	95.2	238.6
Peak Return	~150	443
Episode Length	200	200
Training Time	3h 2min	6h 18min

Console logs confirmed active package delivery behavior, with agents consistently navigating to packages, picking them up, and delivering to goal zones. At 500K steps, agents exhibited functional but suboptimal behavior: occasional hesitation near obstacles, inefficient path selection, and sporadic coordination failures during simultaneous pickups. By 1M steps, these issues largely resolved—policies became nearly deterministic (return standard deviation <0.01) with smooth navigation and effective implicit coordination. Compared to grid-based RWARE, Unity’s continuous physics introduced additional challenges: agents occasionally collided when pathways narrowed, and LIDAR-based sensing required learning spatial relationships that discrete grid observations provided directly. Despite these differences, the core coordination behaviors transferred successfully, validating the sim-to-sim pipeline.

5 Discussion

5.1 Key Findings

Our experiments reveal several critical insights for applying MARL to warehouse robotics:

Hyperparameter Sensitivity: Default configurations consistently fail on RWARE. Extended exploration through slow epsilon decay (5M+ steps) is essential for sparse reward discovery. This finding has significant implications for practitioners—off-the-shelf algorithms require substantial tuning for warehouse domains.

Value Decomposition Advantages: QMIX’s monotonic mixing network substantially outperforms independent learning approaches (3.25 vs. 0.38 mean return on RWARE tiny-2ag), demonstrating the value of centralized coordination mechanisms. The ability to decompose credit assignment while maintaining decentralized execution makes value-based methods well-suited for warehouse coordination.

Scaling Challenges: Performance degrades sub-linearly with agent count, but training requirements increase super-linearly. Industrial warehouses requiring 50+ robots will need hierarchical decomposition or domain-specific constraints to maintain tractability.

5.2 Practitioner Notes

For researchers attempting QMIX on RWARE or similar sparse-reward warehouse tasks, we distill our key lessons. **Most critical:** extend epsilon annealing to at least 5M steps (100x default)—premature greedy behavior prevents agents from ever discovering sparse rewards. **Second:** increase replay buffer size substantially (we used 200K vs 5K default) to maintain experience diversity across long training horizons. **Third:** plan for 10–20M training steps minimum; 2M steps consistently produced zero learning regardless of other hyperparameters.

Recommended starting configuration for RWARE+QMIX: batch size 256, buffer size 200K, epsilon annealing 5M steps, learning rate 0.0005, GRU hidden dimension 64, total training 20M steps. Our experiments ran on Mac M3 (16GB RAM) for RWARE and Windows Server 2022 (Intel Xeon E5-2680, 196GB RAM) for Unity; the latter completed 1M Unity steps in approximately 6 hours using no-graphics mode with 50x time scaling.

5.3 Limitations

Several limitations should be considered: (1) our experiments focus on simulation—transfer to physical systems requires additional domain adaptation and robust perception; (2) testing was limited to 2-6 agents due to computational constraints; (3) RWARE’s simplified task structure does not capture all warehouse operations such as inventory management and dynamic obstacle avoidance; (4) extended training requirements (20M+ steps) may be impractical for rapid deployment scenarios.

5.4 Practical Implications

Our findings suggest QMIX-based coordination is viable for small-scale warehouse deployments (2-4 robots). For larger fleets, hierarchical decomposition or domain-specific constraints may be necessary. The significant hyperparameter tuning required indicates that production deployments should incorporate automated hyperparameter optimization and robust monitoring systems.

6 Conclusion

We presented a systematic comparison of MARL algorithms for warehouse robotics, demonstrating that QMIX achieves superior performance through value decomposition while highlighting the critical importance of hyperparameter tuning for sparse-reward environments. Our Unity integration validates that learned policies transfer effectively to 3D simulation environments with realistic physics.

Future work should explore hierarchical approaches for scaling beyond small agent teams, curriculum learning strategies to accelerate training, and sim-to-real transfer for physical robot deployment. The gap between simulation success and real-world deployment remains a significant challenge for warehouse MARL systems.

Code and Data Availability

All code, trained models, and experimental results are available at:

- GitHub: <https://github.com/pallman14/MARL-QMIX-Warehouse-Robots>
- Documentation: <https://pallman14.github.io/MARL-QMIX-Warehouse-Robots/>

References

- Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., and Whiteson, S. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *International Conference on Machine Learning*, 2018.
- Sunehag, P., Lever, G., Gruslys, A., et al. Value-Decomposition Networks for Cooperative Multi-Agent Learning. *International Conference on Autonomous Agents and Multiagent Systems*, 2018.
- Papoudakis, G., Christianos, F., Schäfer, L., and Albrecht, S.V. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.
- Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Baez, A., and Wu, Y. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. *Neural Information Processing Systems*, 2022.