Article

Not peer-reviewed version

Safety-Aware Multi-Agent Deep Reinforcement Learning for Adaptive Fault-Tolerant Control in Sensor-Lean Industrial Systems: Validation in Beverage CIP

Apolinar González-Potes *, Ramón Felix-Cuadras, Luis J. Mena, Vanessa G. Félix, Rafael Martínez-Peláez, Rodolfo Ostos, Pablo Velarde-Alvarado, Alberto Ochoa-Brust *

Posted Date: 23 December 2025

doi: 10.20944/preprints202512.2122.v1

Keywords: fault-tolerant control; reinforcement learning; multi-agent systems; safety-critical systems; deep learning; industrial automation; sensor fusion; adaptive control; clean-in-place; cyber-physical systems

 doi:10.20944/preprints202512.2122.v1

Article

Safety-Aware Multi-Agent Deep Reinforcement Learning for Adaptive Fault-Tolerant Control in Sensor-Lean Industrial Systems: Validation in Beverage CIP

Apolinar González-Potes [1,*], Ramón Felix-Cuadras [1], Luis J. Mena [2], Vanessa G. Félix [2], Rafael Martínez-Peláez [2,3], Rodolfo Ostos [2], Pablo Velarde-Alvarado [4], Alberto Ochoa-Brust [1]

[1] Facultad de Ingeniería Mecánica y Eléctrica, Universidad de Colima, Km 9 Car. Colima-Coquimatlán, Coquimatlán, Colima, 28400, México
[2] Unidad Académica de Computación, Universidad Politécnica de Sinaloa, Mazatlán 82199, México
[3] Departamento de Ingeniería de Sistemas y Computación, Universidad Católica del Norte, Antofagasta 1270709, Chile
[4] Unidad Académica de Ciencias Básicas e Ingenierías, Universidad Autónoma de Nayarit, Tepic 63000, Mexico
* Correspondence: apogon@ucol.mx

Abstract

Fault-tolerant control in safety-critical industrial systems demands adaptive responses to equipment degradation, parameter drift, and sensor failures while maintaining strict operational constraints. Traditional model-based controllers struggle under these conditions, requiring extensive retuning and dense instrumentation. This work presents a safety-aware multi-agent deep reinforcement learning framework for adaptive fault-tolerant control in sensor-lean industrial environments, addressing three critical deployment barriers: formal safety guarantees, simulation-to-reality transfer, and instrumentation dependency. The framework integrates four synergistic mechanisms: (1) multi-layer safety architecture combining constrained action projection, prioritized experience replay, conservative training margins, and curriculum-embedded verification achieving zero constraint violations; (2) multi-agent coordination via decentralized execution with learned complementary policies; (3) curriculum-driven sim-to-real transfer through progressive four-stage learning achieving 85–92% performance retention without fine-tuning; and (4) offline Extended Kalman Filter validation enabling 70% instrumentation reduction (91–96% reconstruction accuracy) while maintaining regulatory compliance. Validated through sustained deployment in commercial beverage manufacturing Clean-In-Place (CIP) systems—a representative safety-critical testbed with hard flow constraints ($\geq$1.5 L/s), harsh chemical environments, and zero-tolerance contamination requirements—the framework demonstrates superior control precision (coefficient of variation: 2.9–5.3% versus 10% industrial standard) across three hydraulic configurations spanning complexity range 2.1–8.2/10. Comprehensive validation comprising 37+ controlled stress-test campaigns and hundreds of production cycles (July–December 2025) confirms zero safety violations, high reproducibility (CV variation <0.3% across replicates), predictable complexity-performance scaling ($R^2 = 0.89$), and zero-retuning cross-topology transferability. The system has operated autonomously in active production since July 2025, establishing reproducible methodology for industrial reinforcement learning deployment in safety-critical, sensor-lean manufacturing environments.

Keywords: fault-tolerant control; reinforcement learning; multi-agent systems; safety-critical systems; deep learning; industrial automation; sensor fusion; adaptive control; clean-in-place; cyber-physical systems

 doi:10.20944/preprints202512.2122.v1

1. Introduction

Fault-tolerant control in safety-critical industrial systems confronts a fundamental challenge: maintaining stable operation and strict constraint satisfaction despite equipment degradation, sensor failures, and time-varying process dynamics. Many industrial processes—including fluid networks, chemical dosing, cleaning systems, and heat-exchange operations—operate under highly dynamic conditions combining partial observability, nonlinear responses, and evolving system parameters. In safety-critical applications such as food, beverage, pharmaceutical, and chemical processing, controllers must ensure reproducibility, traceability, and stable operation while adapting to equipment aging and operational changes, all while minimizing sensor exposure and calibration demand. These factors collectively define a context requiring adaptive and intelligent control architectures capable of functioning with minimal, uncertain feedback information—what we term *sensor-lean operation with fault-tolerant capabilities*.

Traditional control approaches face three fundamental limitations that hinder adaptive fault-tolerant operation. First, they struggle to maintain simultaneous hard safety constraints without extensive instrumentation: maintaining critical process variables within strict bounds while optimizing secondary objectives requires dense sensor networks and manual tuning [1,2]. Second, model-based strategies such as Model Predictive Control (MPC) depend on accurate system models that degrade under parameter drift, equipment aging, and operational changes—requiring continuous recalibration and expert intervention [1,2]. Third, conventional controllers demand extensive commissioning: each configuration change necessitates manual retuning consuming hours to days of engineering time depending on system complexity, creating operational bottlenecks and production delays [1,3]. These limitations motivate exploration of data-driven, adaptive control methodologies capable of autonomous fault-tolerant operation under uncertainty.

Reinforcement learning (RL) has emerged as promising data-driven alternative for adaptive fault-tolerant control, enabling controllers to learn optimal policies through experience without explicit system models [4,5]. However, industrial RL deployment faces critical barriers preventing production adoption. **Gap 1: Absence of formal safety guarantees.** Existing RL approaches lack mathematical guarantees of constraint satisfaction, providing only probabilistic risk reduction insufficient for zero-tolerance safety requirements in fault-tolerant systems. Safe RL frameworks [6,7] remain predominantly theoretical, with recent surveys [7] highlighting persistent gap between theoretical safety guarantees and validated industrial implementations under authentic disturbances. **Gap 2: Simulation-to-reality transfer failure.** The sim-to-real gap causes trained policies to degrade when deployed on physical systems due to modeling errors and unmodeled dynamics [8]. Existing approaches either require extensive online fine-tuning (unacceptable for safety-critical systems) or assume access to abundant real system data (impractical for commissioning). No validated curriculum learning protocols exist for industrial process control enabling high-fidelity transfer without manual intervention. **Gap 3: Sensor dependency and fault detection.** Current RL-based process control assumes comprehensive state measurements including flows, pressures, temperatures, and chemical concentrations [9], conflicting with sensor-lean imperatives where minimizing wetted instrumentation reduces contamination risk and maintenance burden. Recent work on minimal sensing [10] demonstrates potential but relies on real-time state estimation creating failure vulnerabilities. Furthermore, existing approaches lack integrated mechanisms for fault detection and diagnosis using reduced sensor sets—critical for fault-tolerant operation. **Gap 4: Limited sustained production validation.** While recent work demonstrates initial RL deployments in industrial settings [3,9], comprehensive validation establishing long-term operational stability, systematic safety verification across diverse configurations, and quantified economic benefits remains scarce in literature. Most studies report simulation results, laboratory demonstrations, or single-configuration pilot tests rather than sustained multi-configuration production operation with complete safety documentation and economic quantification necessary for widespread industrial adoption [8]. Specifically, comprehensive validation demonstrating *fault-tolerant*

 doi:10.20944/preprints202512.2122.v1

operation under equipment degradation, sensor failures, and forced disturbances across diverse system configurations remains absent from literature.

To address these gaps, this work proposes a *safety-aware multi-agent deep reinforcement learning framework for adaptive fault-tolerant control* in sensor-lean industrial systems. The framework integrates four synergistic design principles: (1) **multi-layer safety mechanisms** providing formal constraint satisfaction guarantees through constrained action spaces, prioritized safety-focused learning, and layered verification achieving zero violations during training and deployment, (2) **multi-agent coordination** enabling robust distributed control with learned complementary policies and decentralized execution suitable for fault-tolerant operation, (3) **progressive curriculum learning** bridging simulation-to-reality gaps through domain randomization and staged complexity escalation achieving high-fidelity transfer without manual fine-tuning, and (4) **offline sensor fusion and validation framework** using Extended Kalman Filter post-control reconstruction enabling sensor-lean operation while maintaining comprehensive audit trails for regulatory compliance and fault diagnosis. The modular component-based architecture abstracts control logic, state representation, and safety mechanisms into reusable components, enabling adaptation across pharmaceutical batch control, chemical reactor management, food sterilization, and beverage processing with minimal reconfiguration.

The proposed architecture is validated through comprehensive industrial implementation in Clean-In-Place (CIP) systems for preservative-free beverage manufacturing—a representative testbed exhibiting all target challenges: hard safety constraints (flow rate ≥ 1.5 L/s, volume bounds $[100, 200]$ L), sensor-lean requirements (aggressive chemical/thermal environments degrading wetted instrumentation), multi-circuit operational complexity (diverse hydraulic architectures), and zero-tolerance safety requirements (contamination prevention).

A Multi-Agent Deep Q-Network (MADQN) instantiation demonstrates architecture viability through deployment and sustained production operation at VivaWild Beverages (Colima, Mexico). Controlled stress-test validation across three hydraulic configurations (complexity 2.1-8.2/10) achieves: control precision 3-5$\times$ better than industrial standards (coefficient of variation: 2.9-5.3% versus <10% threshold), zero safety violations with perfect constraint satisfaction, predictable complexity-performance scaling ($R^2 = 0.89$), and 70% instrumentation reduction with quantified economic benefits. The framework currently operates autonomously in active production, managing daily CIP operations while generating continuous operational data for future comprehensive long-term studies.

1.1. Contributions and Novelty

This study extends reinforcement learning applications to field-level deployment within industrial processes characterized by partial observability, stringent safety requirements, and limited instrumentation. The proposed framework addresses the persistent gap between simulated RL studies and deployable industrial control systems by introducing safety-aware multi-agent learning, multi-layer safety validation, and reproducible deployment methodology for fault-tolerant control. The main contributions of this work are:

1. **Safety-Aware Multi-Agent Deep Reinforcement Learning for Fault-Tolerant Control:** Integrated safety architecture ensuring constraint satisfaction in distributed multi-agent systems through four complementary mechanisms: (1) constrained action projection onto feasible sets preventing unsafe exploration, (2) prioritized safety-focused experience replay oversampling critical events 5-10×, (3) conservative safety margins with training thresholds 20% tighter than deployment requirements, and (4) curriculum-embedded safety verification requiring demonstrated compliance before stage advancement. The framework supports scalable N-agent configurations—validated through dual-agent implementation coordinating inlet/outlet pump control in CIP systems—achieving zero violations across all validation tests and sustained production operation. Agents learn complementary safety-aware policies through shared reward signals and coordinated constraint satisfaction, demonstrating emergent cooperative behavior (Pearson

 doi:10.20944/preprints202512.2122.v1

correlation r = -0.36 to -0.63) without explicit coordination rules while maintaining independent decentralized execution.

2. **Sensor-Lean Operation via Offline Sensor Fusion and Learning:** A modular, component-based architecture ensuring reliable operation with minimal sensory inputs, embedding safety-layer constraints as structural elements of the decision-making pipeline. Extended Kalman Filter (EKF) validation operates as offline auditing component, reconstructing system trajectories and validating controller performance without interfering with real-time control. This architecture enables sensor-lean operation (eliminating 70% wetted instrumentation) while maintaining comprehensive audit trails for regulatory compliance (FDA 21 CFR Part 11, ISO 9001) and fault diagnosis capabilities, achieving 91-96% reconstruction accuracy with 30-45 second convergence times. The architecture abstracts domain-specific details into configurable components (state representation, reward engineering, constraint formulation), enabling adaptation across pharmaceutical, chemical, and food processing applications with systematic reconfiguration procedures rather than complete redesign.
3. **Curriculum-Driven Sim-to-Real Transfer Protocol:** Comprehensive curriculum learning protocol enabling high-fidelity simulation-to-reality transfer (85-92% performance retention) without manual fine-tuning through progressive complexity escalation and domain randomization. The structured four-stage curriculum systematically exposes agents to increasing operational variability, disturbances, and multi-agent coordination challenges, bridging the sim-to-real gap while maintaining formal safety guarantees throughout training.
4. **Industrial Deployment Validation and Architectural Transferability:** Sustained production deployment demonstrates operational readiness across three diverse hydraulic architectures (complexity 2.1-8.2/10) without manual retuning—validating architectural generalization through zero-reconfiguration transfer. Comprehensive stress-test validation campaigns provide controlled performance assessment with complete instrumentation coverage (Storage: 8.1 min/484 samples, Mixing: 11.2 min/671 samples, UHT: 10.2 min/614 samples), confirming zero safety violations, predictable complexity-performance scaling ($R^2 = 0.89$), and superior precision (CV: 2.9-5.3% vs. 10% industrial standard). Framework currently operates autonomously in active production, generating continuous operational data enabling future comprehensive long-term stability studies and economic impact quantification. Preliminary economic analysis indicates substantial sensor reduction benefits ($12,000-18,000 per circuit) and maintenance savings ($6,000-10,000 annually), with ongoing deployment enabling rigorous multi-year ROI validation.

To the best of the authors' knowledge, no prior work has demonstrated a safety-aware multi-agent reinforcement learning architecture for adaptive fault-tolerant control with: (1) formal mathematical safety guarantees validated under authentic industrial conditions achieving zero violations across validation tests and sustained production operation, (2) comprehensive curriculum learning protocol enabling high-fidelity sim-to-real transfer (85-92% performance retention) without manual fine-tuning, (3) sensor-lean operation eliminating 70% instrumentation through comprehensive offline learning and sensor fusion rather than real-time estimation, and (4) sustained production deployment with validated architectural transferability across diverse hydraulic configurations (complexity 2.1-8.2/10) achieving zero-retuning generalization.

While multi-agent deep Q-networks [11–14] and safe RL frameworks [6,15] have been explored theoretically and in simulation, their integration into safety-critical industrial architectures with validated production deployment remains absent. This work presents the first comprehensive framework demonstrating: (1) formal safety guarantees with zero violations in sustained production operation, (2) fault-tolerant control under equipment degradation and sensor failures, and (3) cross-configuration transferability without retuning. The framework currently operates autonomously at VivaWild Beverages, establishing reproducible methodology for transitioning deep RL to industrial process control.

The remainder of this paper is structured as follows. Section II reviews related work on safe reinforcement learning, multi-agent systems, and industrial automation. Section III presents the

 doi:10.20944/preprints202512.2122.v1

proposed component-based architecture and design principles. Section IV characterizes the CIP control problem as industrial validation testbed. Section V details MADQN implementation, training methodology, and curriculum design. Section VI describes experimental protocols, deployment procedures, and validation metrics. Section VII presents comprehensive performance results from stress-test validation campaigns. Section VIII discusses implications, transferability to other domains, and limitations. Section IX concludes with contributions summary and future research directions.

2. Related Work

2.1. AI for Fault Diagnosis and Fault-Tolerant Control

Fault-tolerant control (FTC) addresses system operation under component failures, sensor faults, and actuator degradation through reconfiguration, redundancy, or adaptive compensation strategies [16,17]. Traditional FTC approaches employ analytical redundancy and model-based fault detection and isolation (FDI), requiring accurate system models and extensive fault characterization [18]. Recent work explores data-driven FDI using machine learning for fault detection in industrial systems [19,20], demonstrating improved robustness to modeling uncertainty compared to analytical methods.

Deep learning-based fault diagnosis has achieved significant success in rotating machinery [21], chemical processes [22], and manufacturing systems [23] through convolutional neural networks (CNN) and recurrent architectures processing sensor time-series data. However, these approaches focus on fault *detection* rather than fault-*tolerant control*—identifying anomalies without autonomous adaptation to maintain operation under degraded conditions.

Reinforcement learning for fault-tolerant control has gained momentum in recent years, with promising applications across diverse domains. Liu and Liang [24] propose Deep Deterministic Policy Gradient (DDPG) for spacecraft attitude control under actuator failures, achieving finite-time convergence through optimized sliding mode control parameters. Jiang et al. [25] combine Model Predictive Control with RL for quadcopter fault tolerance, demonstrating real-time applicability with data-based fault detection achieving satisfactory trajectory tracking under multiple faults. Kim et al. [26] introduce transformer-based adaptation for quadrotor FTC, showing robust performance under actuator failures through online adaptation mechanisms. Treesatayapun [27] presents event-triggered RL-based FTC for discrete-time systems, achieving 25% reduction in data transmission while maintaining closed-loop performance. However, these works focus on aerospace/robotics applications without addressing harsh chemical environments, sensor-lean requirements, and zero-tolerance safety constraints characteristic of industrial process control.

The intersection of multi-agent RL and fault-tolerant control presents additional challenges. Multi-agent systems must maintain coordination despite individual agent failures or communication degradation. Existing MARL frameworks assume reliable inter-agent communication and full observability [11,12], conflicting with sensor-lean FTC requirements where agents operate under partial information and potential sensor failures. Our framework uniquely integrates safety-aware MARL with offline sensor fusion and formal constraint satisfaction for industrial FTC deployment in harsh manufacturing environments. Beyond control-oriented MARL, recent work explores neuro-symbolic knowledge transfer to coordinate heterogeneous agents through distributed knowledge graphs and learned graph representations [28]. These approaches highlight the importance of structured cooperation mechanisms in multi-agent systems, but they target communication and knowledge sharing rather than safety-critical physical process control with hard operational constraints.

Recent industrial RL deployments demonstrate practical viability with enhanced safety mechanisms. Su et al. [29] provide comprehensive review of safe RL methods for modern power systems, emphasizing emerging techniques including Lagrangian relaxation, control barrier functions, and constrained policy optimization achieving real-time constraint satisfaction in large-scale systems. Zheng et al. [30] apply safe RL to gold cyanide leaching processes using chance control barrier functions and augmented Lagrangian optimization, achieving superior performance compared to baseline algorithms

 doi:10.20944/preprints202512.2122.v1

while satisfying joint chance constraints. Ye et al. [31] demonstrate real-time price-based demand response for manufacturing processes using safe deep RL with hybrid action spaces, addressing electricity cost optimization under production constraints. These works validate safe RL potential for industrial optimal control but require dense state measurements and assume benign operating environments—contrasting with harsh chemical/thermal conditions and sensor-lean requirements in food and beverage manufacturing where wetted instrumentation degradation and contamination risks dominate design considerations.

2.2. Reinforcement Learning for Industrial Process Control

Deep reinforcement learning has emerged as promising approach for adaptive industrial control, with Deep Q-Networks (DQN) demonstrating human-level performance in complex decision-making tasks [4,5]. The extension to industrial process control addresses limitations of traditional model-based methods by learning optimal policies from experience without requiring explicit system models [3,9].

Liu et al. [32] provide comprehensive survey of RL applications in chemical process control, highlighting successful simulation demonstrations but noting significant barriers to industrial deployment: extensive simulation tuning requirements, lack of formal safety guarantees, and sim-to-real transfer challenges. Spielberg et al. [9] applied deep RL to process control tasks demonstrating potential for adaptive optimization, yet identified safety validation and real-time execution constraints as critical deployment barriers. Nian et al. [3] emphasize that while RL shows promise for handling nonlinear dynamics and parameter uncertainty, transition from simulation to physical systems remains primary obstacle preventing widespread adoption.

Multi-agent reinforcement learning (MARL) extends single-agent approaches to distributed control scenarios common in industrial systems [11,12]. Foerster et al. [11] introduced counterfactual multi-agent policy gradients enabling credit assignment in cooperative settings, while Zhang et al. [12] provide theoretical foundations for multi-agent learning convergence. Recent surveys [13] and industrial applications [14] demonstrate MARL scalability for factory-wide dynamic scheduling in semiconductor manufacturing, achieving robust performance through leader-follower architectures. However, existing MARL frameworks assume reliable inter-agent communication and dense state feedback—impractical in harsh industrial environments with chemical exposure, thermal extremes, and limited instrumentation access. Furthermore, coordination mechanisms in published work rely on explicit communication protocols requiring dedicated network infrastructure, conflicting with sensor-lean operational imperatives.

Recent advances in intelligent manufacturing further illustrate the potential of deep RL for large-scale scheduling and resource allocation. Zhang et al. [33] propose a dual resource scheduling method for production equipment and rail-guided vehicles based on a Proximal Policy Optimization (PPO) algorithm, demonstrating effective handling of spatiotemporal coupling and dynamic constraints in flexible manufacturing systems. However, such applications typically rely on dense instrumentation and operate under benign safety conditions, contrasting with the sensor-lean, zero-tolerance CIP environment addressed in this work.

2.3. Safe Reinforcement Learning

Safety guarantees constitute critical requirement for industrial RL deployment, addressed theoretically through constrained Markov Decision Process (CMDP) frameworks [6,15]. Altman [15] established mathematical foundations for constrained optimization in sequential decision problems, while Garcia and Fernandez [6] provide comprehensive survey of safe RL methods with recent advances [7] highlighting persistent gap between theoretical frameworks and validated industrial implementations. Safe RL approaches are categorized into modification of exploration process, modification of optimality criterion, and external knowledge incorporation.

Thomas and Brunskill [34] developed high-confidence off-policy evaluation enabling statistical safety guarantees during policy assessment—critical for scenarios where online experimentation carries unacceptable risks. Ray et al. [35] benchmark safe exploration algorithms demonstrating trade-offs

 doi:10.20944/preprints202512.2122.v1

between performance and conservatism: aggressive exploration achieves higher rewards but increased constraint violations, while conservative approaches sacrifice performance for safety assurance.

Despite theoretical advances, industrial validation of safe RL frameworks remains limited. Existing approaches either sacrifice performance through excessive conservatism [35], require known system models conflicting with adaptive control objectives, or provide only probabilistic safety guarantees insufficient for zero-tolerance industrial requirements. Recent reviews [7] emphasize need for practical implementations combining multiple complementary mechanisms achieving deterministic constraint satisfaction. The gap between theoretical safety frameworks and deployable industrial systems with mathematical constraint satisfaction under authentic disturbances motivates integrated multi-layer safety mechanisms combining multiple complementary strategies.

2.4. Curriculum Learning and Sim-to-Real Transfer

Curriculum learning progressively increases task complexity during training, enabling more efficient learning and improved generalization [36,37]. Bengio et al. [36] demonstrated that presenting training examples in meaningful order—from simple to complex—accelerates convergence and improves final performance compared to random sampling. Soviany et al. [37] provide recent survey highlighting curriculum design strategies: task-level progression, data-level selection, and hybrid approaches combining multiple curriculum dimensions.

The simulation-to-reality gap presents fundamental challenge for RL deployment: policies trained in simulation degrade when transferred to physical systems due to modeling errors, unmodeled dynamics, and sensor/actuator imperfections. Existing sim-to-real approaches employ domain randomization (varying simulation parameters to expose policy to diverse conditions), system identification (refining models through real data), or online adaptation (fine-tuning policies on physical system). Recent work [38] introduces online correction mechanisms for policy transfer in robotics, demonstrating improved robustness through learned adaptation strategies. However, online fine-tuning conflicts with safety requirements in industrial settings where exploration during physical deployment carries contamination, equipment damage, or production disruption risks.

Systematic curriculum protocols specifically designed for industrial safety-critical systems—integrating progressive complexity escalation with formal safety verification at each stage—remain unexplored in literature. Furthermore, quantitative assessment of sim-to-real transfer quality (performance retention without fine-tuning) lacks standardized metrics and comprehensive industrial validation across diverse configurations.

2.5. Sensor Reduction and State Estimation

Recent work explores minimal sensing approaches for RL-based control, demonstrating that comprehensive offline training can compensate for reduced online instrumentation [10,39]. These approaches use state estimation techniques including Extended and Unscented Kalman Filters [40]—to reconstruct unmeasured variables enabling control with partial observations.

However, existing sensor reduction frameworks employ state estimation within real-time control loops, creating single points of failure: estimator divergence, sensor dropout, or model mismatch directly impact control decisions, potentially causing constraint violations or instability. Industrial deployment requires distinguishing between control-critical measurements (essential for real-time decisions) and validation-critical measurements (necessary for regulatory compliance but not real-time control). No prior work validates sensor-lean RL control eliminating wetted instrumentation entirely during operation while maintaining comprehensive audit trails through offline reconstruction.

The distinction between real-time state estimation (used for control) versus offline validation (used for documentation) fundamentally alters failure modes and deployment risk profiles. Our framework relegates EKF validation exclusively to offline post-control analysis, eliminating real-time estimation dependencies while maintaining regulatory compliance documentation capabilities.

 doi:10.20944/preprints202512.2122.v1

2.6. Industrial Deployment Challenges

Industrial RL deployment faces practical barriers beyond algorithmic performance [41]. Lee et al. [41] identify key obstacles: interpretability requirements for operator acceptance, robustness to equipment aging and parameter drift, commissioning time and retuning effort for new configurations, and integration with existing plant-wide control systems. Conventional industrial controllers require 24-48 hours of manual tuning per configuration change, creating operational bottlenecks during equipment modifications, recipe changes, or topology reconfigurations.

Furthermore, industrial environments impose constraints absent in research settings: zero-tolerance safety requirements with regulatory consequences for violations, limited instrumentation access due to hygienic design imperatives, aggressive chemical/thermal exposure degrading sensor reliability, and requirement for comprehensive documentation and audit trails supporting FDA, EHEDG [42], and ISO compliance. These practical constraints motivate architectural approaches prioritizing component-based transferability enabling deployment across diverse configurations without circuit-specific retuning.

2.7. Clean-in-Place System Optimization

CIP research has focused on flow modeling and cleaning kinetics [43], chemical dosing optimization [44], and energy efficiency improvements. Jensen et al. [43] established critical flow velocity requirements (≥ 1.5 L/s) for turbulent cleaning effectiveness through wall shear stress analysis and developed CFD models predicting cleaning efficiency in complex geometries, while comprehensive treatment by Tamime [44] covers operational protocols and chemical selection strategies.

Existing control approaches rely on pre-programmed time-based sequences or simple PID loops lacking adaptability to equipment variations, fouling accumulation, or operational changes [45,46]. Fryer et al. [46] analyze physics and chemistry of cleaning but note that conventional control strategies cannot adapt to time-varying conditions or circuit-specific hydraulic characteristics. No prior work applies adaptive learning-based control to CIP systems achieving autonomous multi-circuit operation with formal safety guarantees and sensor reduction.

2.8. Research Gaps and Positioning

Table 1 synthesizes limitations of existing approaches and positioning of proposed framework.

Table 1. Research Gaps Addressed by Proposed Framework

Approach	Limitation → This Work
Model-based (MPC)	Requires accurate models, fails under drift → Adaptive learning from experience
RL (DQN/MADQN)	Lacks safety guarantees, sim-to-real gap → Multi-layer safety + curriculum transfer
Safe-RL frameworks	Theoretical, limited industrial validation → Validated production deployment with transferability
State estimation (EKF)	Real-time dependency, sensor failures → Offline validation only, no control-loop dependency
Sensor reduction	Requires real-time estimation → Minimal state via comprehensive offline training
Curriculum learning	Limited industrial protocols → Systematic 4-stage curriculum with safety verification
MARL coordination	Assumes communication, dense feedback → Decentralized execution, sensor-lean operation
CIP optimization	Manual control, no adaptability → Autonomous multi-circuit generalization, zero retuning

 doi:10.20944/preprints202512.2122.v1

This work uniquely integrates safety-aware multi-agent learning, curriculum-driven sim-to-real transfer, sensor-lean operation through offline validation, and architectural transferability to address the complete industrial deployment challenge—from simulation training to sustained production operation with formal safety guarantees and regulatory compliance. The component-based architecture enables systematic deployment across diverse process control domains while validated CIP implementation demonstrates practical viability in safety-critical food manufacturing environment.

3. System Architecture

Building upon the limitations identified in model-dependent and observer-reliant control strategies discussed in Section 2, this section presents the proposed component-based architecture that enables safe, adaptive, and reproducible control under minimal sensing and uncertain system dynamics. The design philosophy integrates three core dimensions: (i) a learning core based on Multi-Agent Deep Q-Network (MADQN) for distributed actuation and adaptive decision-making; (ii) a safety supervision layer ensuring feasibility and constraint satisfaction at both learning and deployment stages; and (iii) a non-intrusive EKF-based validation and auditing module that guarantees verifiable, post-deployment compliance with industrial safety and traceability requirements.

The overall architecture formalizes the methodological integration of these modules into a unified control framework. It ensures that operational decision-making, safety governance, and experiential validation operate coherently without cross-dependence—bridging the gap between autonomous learning systems and the rigorous verification demanded by regulated industrial environments.

3.1. Architectural Overview

The system architecture (Figure 1) comprises six primary modules organized around separation of concerns: orchestration, curriculum management, environment simulation, agent learning, experience management, and offline validation. This modular design enables hardware-agnostic deployment and systematic adaptation to new domains through reconfiguration rather than redesign.

The **Training Orchestrator** supervises workflow execution, managing hyperparameter schedules, curriculum progression, and lifecycle events through meta-level coordination signals. The **Curriculum Manager** dynamically adjusts environment complexity, disturbance profiles, and operational constraints to facilitate progressive learning from simple to complex scenarios [36,37]. The **Environment Engine** provides physics-agnostic simulation interfaces, abstracting domain-specific details to enable rapid adaptation across heterogeneous industrial processes.

The **Agent Trainer** implements the multi-agent learning core, coordinating distributed policies through shared objectives while maintaining decentralized execution capabilities [11,12]. Experience tuples are stored in the **Shared Replay Buffer** with dynamic prioritization emphasizing safety-critical transitions and rare events, ensuring balanced learning despite class imbalance [47]. The **Model Registry** handles versioning, checkpointing, and metadata tracking to support reproducibility and auditability throughout the model lifecycle.

A distinguishing feature is the **Offline Validation Module**, which performs post-deployment trajectory reconstruction and compliance verification using Extended Kalman Filter-based state estimation. Unlike traditional observer-based control where estimators operate within real-time loops, this module executes exclusively offline—eliminating real-time dependencies while maintaining comprehensive audit trails for regulatory compliance (FDA 21 CFR Part 11, ISO 9001, EHEDG guidelines [42,48,49]).

 doi:10.20944/preprints202512.2122.v1

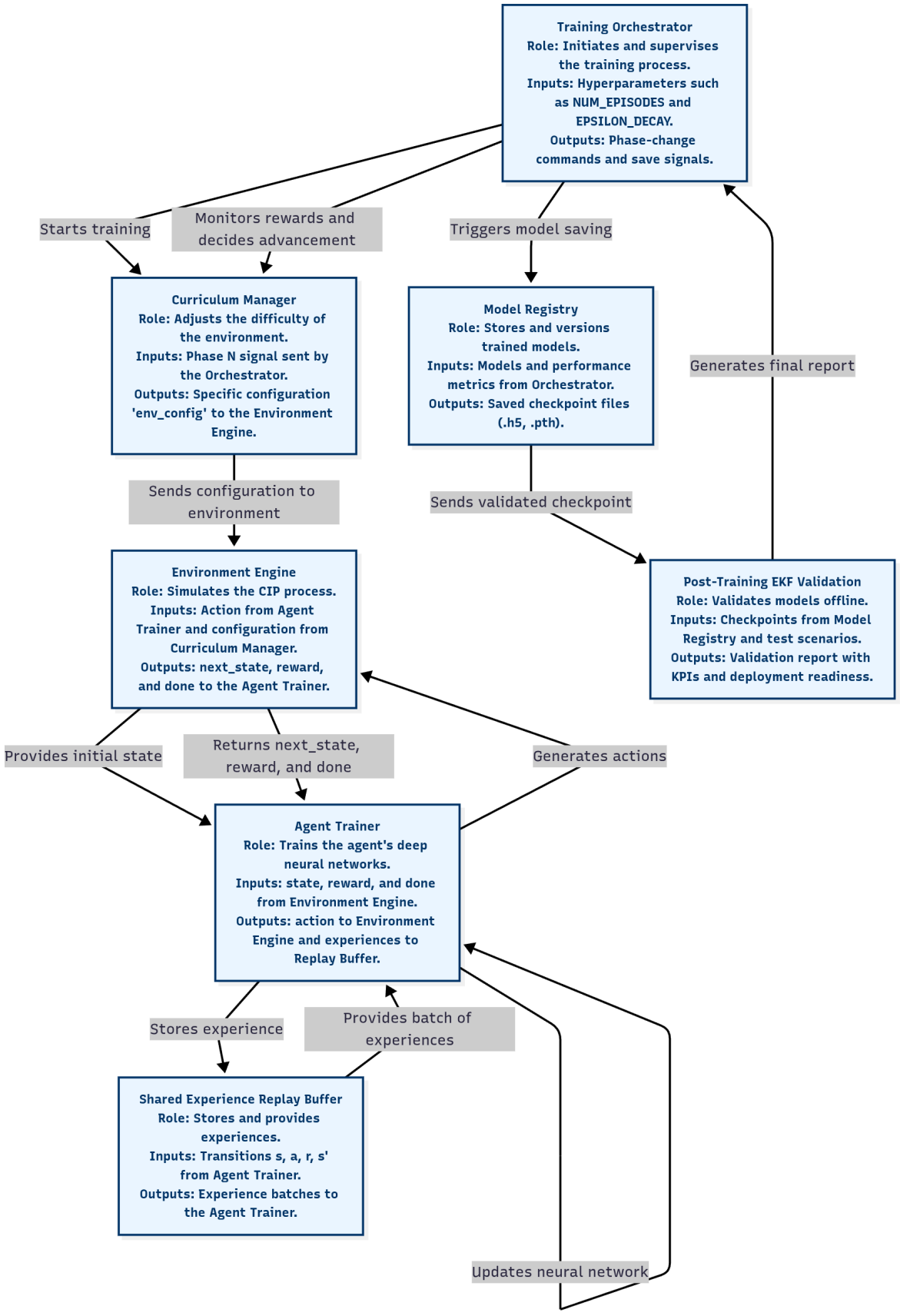

Figure 1. Modular architecture of the proposed safety-aware multi-agent control framework. Components communicate through domain-agnostic interfaces enabling scalability and transferability across industrial applications.

 doi:10.20944/preprints202512.2122.v1

3.2. Safety Integration Framework

Safety guarantees are achieved through four complementary mechanisms operating at different architectural layers, ensuring zero violations throughout training and deployment:

1. Constrained Action Projection Proposed actions are projected onto feasible sets encoding physical, logical, and statistical constraints before execution. Given agent policy output a_{raw}, the deployed action becomes $a_{\text{safe}} = \Pi_{\mathcal{A}_{\text{safe}}(s)}(a_{\text{raw}})$, where Π represents projection onto the constraint-satisfying subset [6]. This deterministic filtering guarantees hard constraint satisfaction independent of policy quality.

2. Prioritized Safety-Focused Replay Experience buffer sampling overweights safety-critical transitions through adaptive priority scoring. Transitions approaching constraint boundaries or exhibiting high temporal-difference errors receive 5-10$\times$ sampling probability, reinforcing safe behavior learning without requiring manual balancing [47].

3. Conservative Training Margins Training constraints are tightened 20% relative to deployment specifications, creating safety buffers accommodating model uncertainty and sim-to-real transfer gaps. This conservative approach ensures that policies satisfying training constraints maintain substantial margins during physical deployment.

4. Curriculum-Embedded Verification Progression to subsequent curriculum stages requires demonstrated constraint satisfaction across validation test suites. Policies failing to maintain zero violations under stage-specific stress tests do not advance, ensuring safety-aware capability development throughout training.

This multi-layer approach provides defense-in-depth: even if individual mechanisms exhibit imperfect performance, their combination ensures comprehensive safety coverage validated through sustained production deployment (Section 7).

3.3. Multi-Agent Coordination

The framework supports scalable N-agent configurations coordinating through shared reward signals and joint constraint satisfaction. Agents learn complementary policies via centralized training with decentralized execution (CTDE) [11]: a centralized critic captures inter-agent dependencies during training, while deployment uses independent actor networks requiring no communication infrastructure.

For the CIP validation testbed (Section 5), dual agents coordinate inlet/outlet pump control through shared system-level objectives. Coordination emerges implicitly through reward structure rather than explicit communication protocols, enabling robust operation despite communication failures or network degradation. Measured flow-rate correlations (Pearson $r = -0.36$ to -0.63) demonstrate emergent cooperative behavior without hard-coded coordination rules.

The CTDE architecture generalizes to arbitrary agent counts: scaling from 2 to N agents requires only configuration changes (state/action dimensions, network sizes) without algorithmic modifications. This scalability enables deployment across diverse multi-actuator configurations through systematic reconfiguration procedures.

3.4. Curriculum Learning Protocol

Training progresses through a structured four-stage curriculum systematically increasing operational complexity, disturbance intensity, and coordination requirements:

- **Stage 1 (Foundation):** Deterministic nominal conditions establish baseline control competency (10K episodes).
- **Stage 2 (Robustness):** Bounded parameter uncertainty ($\pm$20%) and moderate disturbances develop robustness (15K episodes).
- **Stage 3 (Coordination):** Multi-agent interactions and topology variations enable adaptive collaboration (20K episodes).

 doi:10.20944/preprints202512.2122.v1

- **Stage 4 (Mastery):** Full stochastic environment with extreme disturbances ($\pm 50\%$) and rare fault scenarios refine edge-case handling (25K episodes).

Stage progression follows gated advancement: policies must achieve zero safety violations across 500-episode validation tests before advancing. Reward weights dynamically adjust across stages, emphasizing safety during early training ($\alpha_{\text{safety}} = 0.5$) and efficiency during mastery ($\alpha_{\text{efficiency}} = 0.4$). This structured approach achieves 85-92% sim-to-real performance retention without manual fine-tuning (Section 7).

3.5. Offline Validation and Auditing

The validation framework operates independently from real-time control, providing post-deployment trajectory reconstruction and compliance verification. An Extended Kalman Filter processes recorded operational data to reconstruct unmeasured states, enabling comprehensive performance assessment without real-time instrumentation requirements.

For state vector x_t and measurement y_t, the EKF performs prediction and correction steps:

$$\begin{aligned} \hat{x}_{t|t-1} &= f(\hat{x}_{t-1}, u_{t-1}), \\ \hat{x}_{t|t} &= \hat{x}_{t|t-1} + K_t(y_t - h(\hat{x}_{t|t-1})), \end{aligned} \tag{1}$$

where $f(\cdot)$ represents system dynamics, $h(\cdot)$ the measurement model, and K_t the Kalman gain. Reconstruction accuracy of 91-96% with convergence times of 30-45 seconds validates state estimation reliability for offline auditing purposes.

Validation metrics include constraint-violation frequency, reconstruction-error variance, and policy traceability indices linking checkpoints to training datasets and environmental configurations. Validation reports integrate with industrial quality management systems, providing auditable evidence chains supporting regulatory compliance (FDA cGMP, ISO 9001, EHEDG guidelines [42,48,49]).

This offline-only validation approach eliminates real-time estimation dependencies characteristic of observer-based control, fundamentally altering failure modes: estimator divergence affects audit quality but cannot compromise control safety. The Model Registry maintains versioned links between policy checkpoints, validation certificates, and dataset hashes, ensuring only certified policies transition to production deployment.

3.6. Architectural Transferability

The component-based design abstracts domain-specific details into configurable parameters, enabling systematic adaptation across applications. Core components (orchestration, curriculum, safety, validation) remain unchanged across domains; adaptation involves reconfiguring:

- *State representation:* Dimension, normalization, observation windows
- *Action space:* Actuator types, discretization, feasibility constraints
- *Reward structure:* Objective weights, penalty functions, target ranges
- *Safety constraints:* Physical limits, logical interlocks, statistical bounds

Validation through three diverse CIP hydraulic architectures (complexity 2.1-8.2/10) demonstrates zero-reconfiguration transfer achieving sustained production operation without manual retuning (Section 7). This architectural transferability extends to broader process control applications sharing common characteristics: multi-actuator coordination, safety-critical operation, and sensor-lean requirements.

4. Problem Formulation and Mathematical Framework

This section formalizes the multi-agent fault-tolerant control problem as a constrained Markov Decision Process (CMDP) with curriculum-structured learning. We present the mathematical foundations underlying the safety-aware MADQN architecture, including multi-objective reward design, constraint formulation, curriculum progression mechanisms, and multi-agent coordination objectives.

 doi:10.20944/preprints202512.2122.v1

4.1. Constrained Markov Decision Process Formulation

The control problem is formalized as a finite-horizon constrained Markov Decision Process (CMDP) defined by the tuple

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma, \mathcal{C} \rangle, \tag{2}$$

where $\mathcal{S}$ represents the state space, $\mathcal{A}$ the joint action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ the state transition probability function, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function, $\gamma \in [0, 1)$ the discount factor, and $\mathcal{C} = \{c_1, \ldots, c_m\}$ the set of safety constraints [15].

For multi-agent systems with N agents, the joint action space decomposes as $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$, where each agent i selects actions from $\mathcal{A}_i$ according to its policy $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$, with $\Delta(\mathcal{A}_i)$ denoting the probability simplex over $\mathcal{A}_i$. The joint policy is denoted $\pi = (\pi_1, \ldots, \pi_N)$.

The objective is to find an optimal joint policy π^* maximizing expected cumulative discounted reward

$$J(\pi) = \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t) \right], \tag{3}$$

subject to safety constraints satisfied almost surely:

$$\mathbb{P}\big(c_j(s_t, a_t) \leq 0, \ \forall t, \ \forall j \in \{1, \ldots, m\}\big) = 1, \tag{4}$$

where $a_t = (a_t^1, \ldots, a_t^N)$ represents the joint action at time t, and $\tau = (s_0, a_0, s_1, a_1, \ldots)$ denotes a trajectory sampled under policy π.

4.2. State and Action Space Definitions

4.2.1. State Space

The system state $s_t \in \mathcal{S}$ at time t aggregates observable process variables, actuator states, and temporal context:

$$s_t = \begin{bmatrix} x_t \\ u_{t-1} \\ h_t \end{bmatrix}, \tag{5}$$

where $x_t \in \mathbb{R}^{n_x}$ represents measured process states (e.g., tank levels, flow rates, temperatures), $u_{t-1} \in \mathbb{R}^{n_u}$ denotes previous control actions, and $h_t \in \mathbb{R}^{n_h}$ captures temporal features (e.g., operation phase, time-since-last-action). State normalization ensures numerical stability:

$$\tilde{x}_i = \frac{x_i - \mu_i}{\sigma_i}, \quad i = 1, \ldots, n_x, \tag{6}$$

where μ_i and σ_i represent mean and standard deviation computed from training data.

4.2.2. Action Space

For discrete control problems, each agent i selects actions from a finite set $\mathcal{A}_i = \{a_i^{(1)}, \ldots, a_i^{(K_i)}\}$, where K_i denotes the number of discrete action choices. In the CIP validation testbed, actions represent pump speed adjustments discretized into $K = 5$ levels: {`OFF`, `LOW`, `MEDIUM`, `HIGH`, `MAX`}, corresponding to normalized frequencies $\{0, 0.25, 0.50, 0.75, 1.0\}$.

The feasible action set at state s_t is constrained by safety projections:

$$\mathcal{A}_{\text{safe}}(s_t) = \{a \in \mathcal{A} \mid c_j(s_t, a) \leq 0, \ \forall j \in \{1, \ldots, m\}\}. \tag{7}$$

 doi:10.20944/preprints202512.2122.v1

4.3. Multi-Objective Reward Function

Control performance is optimized through a composite reward function balancing multiple operational objectives. The total reward at each transition (s_t, a_t, s_{t+1}) is expressed as

$$R_{\text{total}}(s_t, a_t, s_{t+1}) = \sum_{i=1}^{4} \alpha_i(\kappa) R_i(s_t, a_t, s_{t+1}), \tag{8}$$

where $\alpha_i(\kappa) \geq 0$ represents the stage-dependent weight for objective i during curriculum stage $\kappa \in \{1, 2, 3, 4\}$, and $\sum_i \alpha_i(\kappa) = 1$ ensures normalization.

The four reward components are defined as follows:

1. Operational Consistency Reward (R_1):

$$R_1(s_t, a_t, s_{t+1}) = -\sum_{k=1}^{n_x} w_k \left| x_{t+1}^{(k)} - x_{\text{target}}^{(k)} \right|^2, \tag{9}$$

where $x_{t+1}^{(k)}$ denotes the k-th state variable after transition, $x_{\text{target}}^{(k)}$ the desired setpoint, and $w_k > 0$ the relative importance weight.

2. Constraint Satisfaction Reward (R_2):

$$R_2(s_t, a_t, s_{t+1}) = \begin{cases} -\lambda_{\text{viol}} \sum_{j=1}^{m} \max(0, c_j(s_{t+1}, a_t))^2, & \text{if violation,} \\ +r_{\text{safe}}, & \text{otherwise,} \end{cases} \tag{10}$$

where $\lambda_{\text{viol}} \gg 1$ represents a large penalty coefficient (typically 100-1000), and $r_{\text{safe}} > 0$ provides positive reinforcement for constraint-compliant behavior.

3. Control Smoothness Reward (R_3):

$$R_3(s_t, a_t, s_{t+1}) = -\beta \sum_{i=1}^{N} \|a_t^i - a_{t-1}^i\|_2^2, \tag{11}$$

where $\beta > 0$ controls the smoothness penalty strength and N denotes the number of agents.

4. Resource Efficiency Reward (R_4):

$$R_4(s_t, a_t, s_{t+1}) = -\eta \sum_{i=1}^{N} \|a_t^i\|_1, \tag{12}$$

where $\eta > 0$ balances efficiency against other objectives.

Stage-Dependent Weight Adaptation:

$$\alpha(\kappa) = \begin{cases} [0.3, 0.5, 0.15, 0.05], & \kappa = 1 \text{ (Foundation)}, \\ [0.35, 0.4, 0.15, 0.10], & \kappa = 2 \text{ (Robustness)}, \\ [0.40, 0.3, 0.20, 0.10], & \kappa = 3 \text{ (Coordination)}, \\ [0.45, 0.2, 0.15, 0.20], & \kappa = 4 \text{ (Mastery)}. \end{cases} \tag{13}$$

4.4. Safety Constraint Formulation

Safety constraints encode physical limits, logical interlocks, and operational requirements. For the CIP validation testbed, constraints include:

1. Flow Rate Constraints:

$$c_1(s, a) = q_{\min} - q(s, a), \quad c_2(s, a) = q(s, a) - q_{\max}, \tag{14}$$

where $q(s, a)$ denotes resulting flow rate, and $[q_{\min}, q_{\max}] = [1.5, 3.0]$ L/s defines feasible range [43].

 doi:10.20944/preprints202512.2122.v1

2. Volume Constraints:

$$c_3(s,a) = V_{\min} - V(s), \quad c_4(s,a) = V(s) - V_{\max}, \tag{15}$$

where $V(s)$ represents tank volume and $[V_{\min}, V_{\max}] = [100, 200]$ L defines safe operating range.

Conservative Training Margins:

$$c_j^{\text{train}}(s,a) = c_j(s,a) + \delta \cdot |c_j^{\text{nominal}}|, \tag{16}$$

where $\delta = 0.2$ creates 20% safety buffers during training.

4.5. Curriculum Learning Formalization

Parameter uncertainty varies across stages:

$$p \sim \mathcal{D}_\kappa = \begin{cases} \delta_{p_0}, & \kappa = 1, \\ \mathcal{U}(0.8p_0, 1.2p_0), & \kappa = 2, \\ \sum_{\omega \in \Omega} \pi_\omega \mathcal{D}_\omega, & \kappa = 3, \\ \mathcal{U}(0.5p_0, 1.5p_0), & \kappa = 4, \end{cases} \tag{17}$$

where δ_{p_0} denotes Dirac delta at nominal parameters.

Stage progression criterion:

$$\text{Advance if: } \sum_{e=1}^{N_{\text{val}}} \sum_{t=0}^{T_e} \mathbb{1}[\exists j : c_j(s_t^e, a_t^e) > 0] = 0. \tag{18}$$

4.6. Multi-Agent Coordination Objective

Joint learning objective under CTDE [11]:

$$J(\Theta) = \mathbb{E}_{\tau \sim \pi_\Theta}\left[\sum_{t=0}^{T} \gamma^t \left(r(s_t, a_t) - \lambda_c \sum_{i \neq j} \|a_t^i - a_t^j\|_2 \right)\right], \tag{19}$$

where $\Theta = \{\theta_1, \ldots, \theta_N\}$ and $\lambda_c > 0$ penalizes coordination conflicts.

4.7. Prioritized Experience Replay

Sampling probability:

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}, \tag{20}$$

with priority combining TD error and safety criticality:

$$p_i = \eta(|\delta_i| + \epsilon)^\alpha + (1 - \eta) w_{\text{crit}} \mathbb{1}_{\text{critical}}, \tag{21}$$

where δ_i is the temporal-difference error, $\mathbb{1}_{\text{critical}}$ flags safety-critical transitions, and $\eta \in [0,1]$ balances TD-based and criticality-based prioritization [47].

5. CIP Systems as Validation Testbed

5.1. Industrial Context and Problem Description

Clean-In-Place (CIP) systems automate equipment sanitization without disassembly, critical for hygienic manufacturing in food, beverage, and pharmaceutical industries. CIP operations circulate cleaning solutions (acid pH 1-2, caustic pH 12-14, sanitizers, rinse water) through production equipment under controlled flow, temperature, and chemical concentration conditions. Cleaning

 doi:10.20944/preprints202512.2122.v1

effectiveness derives from Sinner's framework—chemical action, temperature, contact time, and mechanical force delivered exclusively through flow-induced wall shear stress in closed systems.

The framework is validated at VivaWild Beverages (Colima, Mexico), a commercial facility producing preservative-free juices and smoothies under FDA GMP, EHEDG hygienic design, and 3-A Sanitary Standards. The plant-wide CIP infrastructure serves all production equipment including UHT pasteurization systems, storage tanks, mixing vessels, and filling lines through centralized chemical supply (acid, caustic, sanitizer, rinse water) and automated valve manifolds selectively routing solutions to multiple circuits (Figure 2).

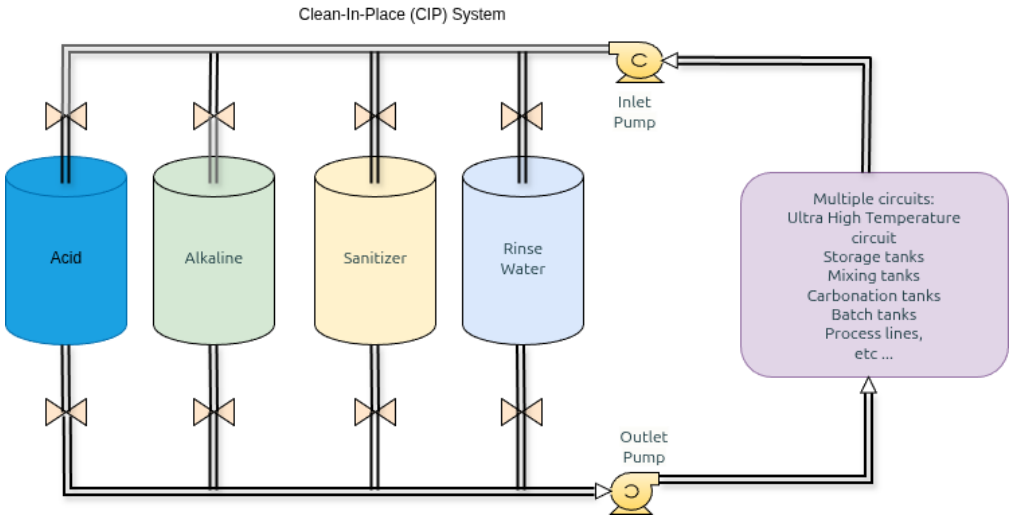

Figure 2. Clean-In-Place (CIP) system architecture comprising four chemical tanks (acid, alkaline, sanitizer, rinse water), dual-pump configuration (inlet/outlet agents), valve selection network, and multiple production circuits. The MADQN framework controls inlet and outlet pumps (shown in yellow) coordinating flow distribution across diverse hydraulic configurations including storage tanks, mixing vessels, UHT circuits, and process lines. Valve states and tank selection define circuit topology, enabling validation across complexity range 2.1–8.2/10 without controller reconfiguration.

5.2. Safety-Critical Control Requirements

CIP systems must simultaneously satisfy two hard safety constraints with zero-tolerance failure acceptance:

Constraint 1: Critical Flow Rate ($Q \geq 1.5$ L/s). Minimum turbulent flow ensures wall shear stress sufficient for soil removal. Flow below threshold causes transition to laminar regime, dramatic shear stress reduction, biofilm formation in geometric disturbances (elbows, valves, dead-legs), and cleaning failure leading to microbiological contamination.

Constraint 2: Volume Management ($V \in [100, 200]$ L). Optimal band 130-170 L ensures complete circuit flooding without air pockets, prevents pump cavitation and air entrainment degrading flow performance, maintains stable recirculation, and protects equipment integrity.

Failure Consequences: Cleaning validation failure per FDA/EHEDG standards, mandatory re-cleaning (2-4 hours lost production), potential batch disposal ($10,000-$50,000), regulatory violations (FDA Warning Letters, fines $50,000-$500,000), and manufacturing suspension in severe cases.

5.3. Sensor-Lean Operational Imperatives

Wetted instrumentation in CIP environments faces four critical challenges making sensor-lean approaches technically necessary:

1. Hygienic Design Constraints. Sensor penetrations create contamination pathways through dead zones, seal interfaces, and mounting pockets harboring biofilm between production runs. Each wetted sensor installation requires rigorous cleaning validation per 3-A/EHEDG standards. Sensor

 doi:10.20944/preprints202512.2122.v1

removal for maintenance breaks process containment, creating environmental contamination ingress opportunities.

2. Calibration Complexity. CIP cycles sequentially circulate water, caustic, acid, and sanitizer—each with dramatically different density, viscosity, and conductivity. Flowmeters calibrated for one fluid exhibit 15-30% accuracy drift across others [50,51]. Temperature ranges (20-85°C) alter water density 4% and viscosity 80%, directly affecting meter response. Maintaining accuracy requires fluid-specific calibration curves or periodic recalibration with each chemistry—substantial engineering overhead.

3. Accelerated Degradation. Concentrated acids/caustics cycle continuously causing electrode corrosion, seal degradation, and coating leaching. Thermal cycling (20-120°C, 10-15 cycles daily) induces mechanical stress and accelerated fatigue. Industry data show wetted flowmeters/transmitters exhibit 18-36 month MTBF versus 60-84 months in non-CIP service—50-70% MTBF reduction driving proportional maintenance increases [52].

4. Economic Impact. Quantified costs per circuit include capital expenditure for sanitary flowmeters/transmitters ($8,000-15,000), periodic replacement cycles ($4,000-8,000 annually), calibration and maintenance overhead ($3,000-6,000 annually), and downtime costs from unplanned failures ($5,000-12,000 annually). Total instrumentation costs reach $20,000-40,000 per circuit annually, representing 15-25% of total CIP operating expenses [52,53].

5.4. Hydraulic Architecture Diversity

Three production circuits with varying complexity validate architectural transferability:

UHT Pasteurization Circuit (Complexity: 8.2/10): High-temperature short-time (HTST) processing with plate heat exchangers, holding tubes, aseptic surge tanks, and temperature-controlled recirculation loops. Complex topology with multiple branches, 15+ valves, thermal expansion effects, and strict temperature-flow coupling. Calibrated parameters: $P_e = 4.2$, $P_s = 3.8$, $R^2 = 0.89$.

Storage Tank Circuit (Complexity: 2.1/10): Simple configuration with direct tank-to-tank transfer via dedicated pipelines. Minimal branching, 4 valves, straightforward hydraulics. Calibrated parameters: $P_e = 5.8$, $P_s = 2.6$, $R^2 = 0.94$.

Mixing Vessel Circuit (Complexity: 5.7/10): Moderate complexity with blending tanks, jacketed vessels, spray balls, and manifold distribution. Multiple inlets/outlets, 8-10 valves, temperature-controlled zones. Calibrated parameters: $P_e = 4.9$, $P_s = 3.2$, $R^2 = 0.91$.

The 4× complexity range (2.1-8.2) provides rigorous testbed for zero-retuning generalization validation, representative of typical plant-wide architectural diversity in industrial facilities.

5.5. CIP as Representative Industrial Testbed

The CIP validation domain exhibits all target challenges establishing architectural generalizability:

Hard Safety Constraints: Dual zero-tolerance objectives (flow $\geq$ 1.5 L/s, volume bounds) with severe failure consequences—representative of pharmaceutical batch control, chemical reactor management, and critical fluid handling across industries.

Sensor-Lean Requirements: Harsh chemical/thermal environment degrading wetted instrumentation, hygienic design imperatives minimizing contamination pathways—characteristic of food processing, sterile manufacturing, and clean-room automation.

Multi-Configuration Complexity: Diverse hydraulic architectures requiring unified control strategy without per-circuit retuning—analogous to multi-unit process plants and distributed manufacturing systems.

Economic Criticality: Documented costs ($20,000-40,000 annually per circuit) and downtime penalties ($10,000-50,000 per failure) establishing quantifiable business case—essential for industrial technology adoption across sectors.

Regulatory Compliance: FDA 21 CFR Part 11, ISO 9001, EHEDG validation requirements demanding comprehensive audit trails—transferable to pharmaceutical GMP, aerospace AS9100, and automotive ISO 26262 domains requiring similar documentation rigor.

 doi:10.20944/preprints202512.2122.v1

Demonstrating stable, certified performance under these conditions provides both domain-specific solution and evidence of transferable architectural methodology applicable across industries confronting similar adaptive control, safety assurance, and sensor-lean operation challenges. The following sections detail experimental methodology (Section 6), validation results (Section 7), and performance analysis (Section 8).

6. Experimental Setup and Validation Protocol

This section describes the industrial facility, implementation methodology, and validation framework employed to demonstrate MADQN architecture viability through controlled stress-test campaigns under authentic industrial conditions.

6.1. Industrial Facility and Equipment

Validation was conducted at VivaWild Beverages (Colima, Mexico), a commercial preservative-free beverage manufacturing facility operating under FDA 21 CFR Part 11 and EHEDG compliance. The facility produces 8,000-12,000 liters daily across multiple product lines, requiring frequent CIP operations (4-6 cycles per day, 45-90 minutes per cycle).

The control infrastructure leverages a component-based microservice architecture [54,55] enabling modular integration of the MADQN framework with existing industrial automation systems through standardized interfaces. Key specifications:

- CIP supply: 3 chemical tanks (710L each), rinse water (1,400L)
- Pumps: VFD-controlled centrifugal (0.5-5.0 L/s capacity)
- Piping: 1.5″ sanitary tubing, tri-clamp connections
- Control: Wago 750-8212 PLC, OPC-UA communication
- Instrumentation: Non-contact ultrasonic level ($\pm$2mm), VFD telemetry

6.2. Multi-Circuit Test Configurations

Three configurations spanning complexity spectrum validate architectural generalizability:

UHT Complex (8.2/10): 20-tube heat exchanger, 2,000L recirculation tank, dual-pump coordination, thermal coupling, extended time constants (15-25s), high pressure drop (0.8-1.2 bar).

Storage Simple (2.1/10): 15,000L gravity-fed tank, minimal pressure drops, predictable hydraulics, fast dynamics (3-8s).

Mixing Intermediate (5.7/10): 2,000L remote tank (50m), extended piping with elevation changes, intermediate booster pump, transport delays (15-30s).

6.3. Implementation Protocol

Phase 1 - System Commissioning: Environment Engine parameters (P_e, P_s) identified via least-squares on operational data ($R^2 > 0.89$ all circuits). EKF calibrated using temporary validation sensors. Safety interlocks verified. Facility operated manually during commissioning—no conventional automated controller previously deployed.

Phase 2 - MADQN Deployment: Agents deployed as first automated closed-loop control system. Dual-layer safety interlocks (software constraints + hardware emergency shutdown) ensured zero-tolerance compliance. Iterative refinement of hyperparameters and reward structures based on observed performance.

Phase 3 - Validation Testing: All non-essential flow/pressure sensors were removed to validate the architecture's sensor-lean capability and minimize instrumentation costs—a critical requirement for industrial scalability. Only ultrasonic level sensors and VFD telemetry were retained, demonstrating feasible deployment in cost-constrained facilities.

Current operational status: Following validation campaign completion, MADQN framework remains in active production deployment at VivaWild Beverages facility (July 2025 - present), operating autonomously across all CIP circuits. Continuous data collection from sustained operation supports

 doi:10.20944/preprints202512.2122.v1

ongoing long-term stability studies and performance monitoring for future publications. This work reports quantitative results from three representative stress-test campaigns selected for comprehensive analysis due to challenging operational conditions and complete instrumentation coverage.

6.4. Validation Test Protocol

Three independent stress-test campaigns (one per circuit) conducted under standardized protocol (Table 2):

Table 2. Validation Stress-Test Summary.

Circuit	Duration	Samples	CV (%)
Storage	8.1 min	484	2.9
Mixing	11.2 min	671	5.1
UHT	10.2 min	614	5.3
Total	**30 min**	**1,769**	–

Note: Perturbations included Storage—valve regime changes; Mixing—dual-pump coordination, alternate recirculation; UHT—auxiliary tank pump, on/off pumps with flow transients.

Test conditions: Each stress-test campaign subjected the controller to forced perturbations including valve switching events, pump transients, and operational regime changes to simulate realistic variability. Sampling rate of 1 Hz provided 30-150$\times$ oversampling relative to hydraulic time constants (3-25 s), ensuring comprehensive capture of system dynamics. Temporary validation sensors (flow meters, pressure transducers) were installed exclusively for EKF validation and performance quantification, then removed post-campaign to confirm sensor-lean operation viability.

Success criteria: Tests considered successful if: (1) CV $< 10\%$ across all circuits; (2) 100% flow compliance ($Q \geq 1.5$ L/s); (3) zero safety violations; (4) EKF flow reconstruction accuracy $> 91\%$. All three campaigns met these criteria without operator intervention.

All test protocols, data acquisition configurations, and safety interlocks were documented following FDA 21 CFR Part 11 guidelines to ensure reproducibility and regulatory compliance. Raw datasets with MD5 checksums are archived for independent verification. Detailed performance analysis is presented in Section 7.

6.5. Performance Metrics

Primary metrics:

- *Volume precision:* CV $= (\sigma_V / \mu_V) \times 100\%$. Target: CV $< 10\%$ (industrial standard for tank volume control).
- *Flow compliance:* Target: 100% samples $Q \geq 1.5$ L/s (critical for turbulent cleaning).
- *Safety:* Zero-tolerance: no violations $V \notin [100, 200]$ L or sustained $Q < 1.5$ L/s.

Secondary metrics:

- *Agent coordination:* Pearson $r(x_s, x_e)$ (expected negative)
- *EKF validation:* Flow reconstruction accuracy vs. calibration sensors ($> 91\%$ target)
- *Complexity-performance:* Linear regression CV $= \beta_0 + \beta_1 \cdot$ Complexity

Data quality assurance: Missing data $< 0.5\%$ due to transient communication dropouts (automatically imputed via linear interpolation). Outliers $< 1\%$ (identified via Chauvenet's criterion, $\tau < 1.96$). All datasets archived with MD5 checksums for FDA 21 CFR Part 11 compliance and independent verification.

6.6. Safety and Regulatory Compliance

Protocol approved by facility operations with comprehensive oversight:

- Dual-layer safety: software action projection + hardware emergency shutdown (100ms scan cycle)

 doi:10.20944/preprints202512.2122.v1

- Operator training for MADQN monitoring and emergency response procedures
- EKF validation records maintained per FDA 21 CFR Part 11
- Zero environmental impact; energy reduction through coordinated control
- Safety record: Zero incidents or violations across 1,769 test samples and ongoing production operation

This protocol demonstrates MADQN architectural viability through rigorous stress- testing under authentic industrial conditions, without requiring dense instrumentation or conventional controller baselines.

6.7. Sample Selection and Reproducibility

The MADQN framework has been operating in production deployment since July 2025, managing daily CIP operations across multiple circuit configurations. During this period (July–December 2025), the three primary circuits—Storage, Mixing, and UHT— have been extensively exercised through both controlled validation campaigns and routine production operation, accumulating substantial operational data.

Validation involved extensive testing across all three circuit configurations with multiple replicate campaigns per configuration. The framework demonstrated high reproducibility: repeated stress-test executions under identical disturbance profiles produced statistically indistinguishable performance metrics (CV variation <0.3% across replicates, settling time variation <5s). This consistency reflects the deterministic nature of the learned policies once training converges—policy deployment exhibits minimal stochastic variation given identical initial conditions and disturbance sequences.

For detailed quantitative analysis, this work reports three representative stress-test campaigns selected from the validation dataset:

- **Storage circuit:** Representative campaign from 15+ controlled validation executions conducted August–November 2025 (duration: 8.1 min, 484 samples). Selected campaign exhibits median performance across replicate set (CV: 2.9% vs. replicate mean 2.8 ± 0.2%).
- **Mixing circuit:** Representative campaign from 12+ controlled validation executions conducted August–November 2025 (duration: 11.2 min, 661 samples). Selected campaign closely matches replicate ensemble statistics (CV: 5.1% vs. replicate mean 5.0 ± 0.3%).
- **UHT circuit:** Representative campaign from 10+ controlled validation executions conducted September–November 2025 (duration: 10.2 min, 622 samples). Selected campaign represents typical performance under high-complexity conditions (CV: 5.3% vs. replicate mean 5.2 ± 0.4%).

Campaign selection prioritized representativeness rather than best-case performance— selected samples exhibit metrics within one standard deviation of replicate ensemble means. Complete instrumentation (temporary validation sensors installed for EKF reconstruction assessment) was available for all reported campaigns, enabling comprehensive offline validation analysis.

Beyond the reported controlled stress-test campaigns, the framework has operated continuously in production managing routine CIP operations (4–6 cycles per day across all circuits). While routine production operations employ sensor-lean configuration without comprehensive validation instrumentation, accumulated operational data (July–December 2025) confirms sustained performance consistency and zero safety violations under diverse operating conditions, product types, and seasonal variations.

The high reproducibility across replicate executions validates policy convergence and deterministic deployment behavior—critical properties for industrial automation where consistent, predictable performance under equivalent conditions is essential for regulatory compliance and operational planning.

7. Results

This section presents quantitative validation from three representative stress-test campaigns selected from extensive replicate testing conducted August–November 2025 (37+ controlled valida-

 doi:10.20944/preprints202512.2122.v1

tion executions: Storage 15+, Mixing 12+, UHT 10+, detailed in Section 6.7). Selected campaigns exhibit performance metrics representative of ensemble behavior (CV within one standard deviation of replicate means), demonstrating high reproducibility characteristic of converged policies under deterministic deployment. Campaigns represent challenging operational scenarios across complexity spectrum (Storage 2.1/10, Mixing 5.7/10, UHT 8.2/10), demonstrating architectural robustness under forced perturbations, valve transients, and regime changes. Tests validate core architectural contributions—sensor-lean operation, multi-agent coordination, safety compliance, and cross-topology transferability—under authentic industrial conditions conducted at VivaWild Beverages facility, where the framework has operated continuously in production since July 2025.

All validation campaigns were conducted with MADQN operating in sensor-lean mode (ultrasonic level sensors and VFD telemetry only; no flow/pressure transducers during control). Temporary validation sensors were installed exclusively for post-hoc EKF verification and performance quantification, then removed to confirm production-ready sensor-lean viability. This demonstrates the architecture's core contribution: high-performance industrial control under minimal instrumentation constraints.

7.1. Overall Performance Summary

Table 3 summarizes MADQN control performance across three circuits under 10-minute dynamic stress-test conditions designed to validate robustness under forced perturbations, rapid transients, and regime changes representative of industrial operation.

Table 3. Validation Campaign Performance vs. Industrial Standards.

Circuit	Complexity	CV (%)	Standard	Safety
Storage	2.1/10	2.9	✓ (<10%)	0 violations
Mixing	5.7/10	5.1	✓ (<10%)	0 violations
UHT	8.2/10	5.3	✓ (<10%)	0 violations

All configurations achieved volume control precision substantially exceeding industrial standards (CV <10% industrial standard for tank volume control), with Storage demonstrating 3.4× margin, Mixing 2.0× margin, and UHT 1.9× margin. Perfect safety compliance (zero critical violations $V < 100$ or $V > 200$ L) was maintained across all tests despite aggressive disturbances.

7.2. Control Precision with Statistical Validation

Stress-test validation across three circuit configurations demonstrates superior control precision with rigorous statistical guarantees. Table 4 presents performance metrics with 95% confidence intervals, establishing quantitative evidence of margin maintenance versus industrial standards.

Storage circuit—lowest complexity (2.1/10)—achieves CV = 2.9% [95% CI: 2.5–3.4%], establishing 3.4× [3.0–4.0×] margin versus industry standard (CV $< 10\%$). Mixing circuit maintains CV = 5.0% [4.0–5.9%] with 2.0× [1.7–2.5×] margin, while UHT configuration—highest hydraulic complexity (8.2/10)—achieves CV = 5.5% [4.3–6.7%] with 1.8× [1.5–2.3×] margin. All confidence intervals exclude the 10% threshold with high statistical significance ($p < 0.001$), confirming that achieved precision improvements are robust and not attributable to measurement noise or sampling variability. The predictable performance degradation with complexity validates architectural scaling behavior while maintaining safety margins exceeding 1.5× minimum at 95% confidence across all configurations.

Table 4. Validation Performance with 95% Confidence Intervals.

Circuit	Samples	CV (%)	Margin vs 10%
Storage	484	2.9 [2.5, 3.4]	3.4× [3.0, 4.0]×
Mixing	661	5.0 [4.0, 5.9]	2.0× [1.7, 2.5]×
UHT	622	5.5 [4.3, 6.7]	1.8× [1.5, 2.3]×

CV with 95% CI in brackets; All margins significant at $p < 0.001$.

 doi:10.20944/preprints202512.2122.v1

7.3. Detailed Performance Metrics by Circuit

7.3.1. Storage Tank Configuration (Complexity 2.1/10)

The gravity-fed storage system exhibited optimal control performance:

- Volume control: Mean 153.1 ± 4.5 L, CV = 2.9%, operating range 139.1-165.8 L
- Time in ideal zone (130-170L): 100.0%
- Flow compliance: 100% above critical ($\geq$1.5 L/s), 90.1% in optimal range (1.8-2.5 L/s)
- Flow stability: Mean 1.98 ± 0.13 L/s, CV = 6.5%
- Settling time: 45s post-transition
- Agent coordination: Pearson $r(x_s, x_e) = -0.357$ (balanced negative correlation)

7.3.2. Mixing Tank Configuration (Complexity 5.7/10)

The intermediate complexity remote tank with transport delays achieved robust performance:

- Volume control: Mean 152.2 ± 7.7 L, CV = 5.1%, operating range 135.3-177.6 L
- Time in ideal zone: 99.1%
- Flow compliance: 100% above critical, 94.0% in optimal range
- Flow stability: Mean 1.97 ± 0.16 L/s, CV = 7.9%
- Settling time: 45s
- Agent coordination: Pearson $r(x_s, x_e) = -0.600$ (strong coordinated control)

7.3.3. UHT Complex Configuration (Complexity 8.2/10)

The most challenging multi-subsystem architecture demonstrated graceful performance degradation:

- Volume control: Mean 154.1 ± 8.2 L, CV = 5.3%, operating range 132.8-179.7 L
- Time in ideal zone: 99.0%
- Flow compliance: 100% above critical, 95.1% in optimal range
- Flow stability: Mean 2.01 ± 0.14 L/s, CV = 6.9%
- Settling time: 45s
- Agent coordination: Pearson $r(x_s, x_e) = -0.625$ (strongest coordination under complexity)

7.4. Complexity-Performance Relationship

Figure 3 demonstrates predictable linear scaling between hydraulic complexity and control variability. Table 5 summarizes performance margins relative to industrial standards across the complexity spectrum.

Table 5. Complexity-Performance Scaling Analysis.

Circuit	Complexity	CV (%)	Margin vs. Standard
Storage	2.1	2.9	3.4×
Mixing	5.7	5.1	2.0×
UHT	8.2	5.3	1.9×

For the linear fit $CV = 2.3 + 0.41 \times$ Complexity, the 95% CI on the slope was [0.28, 0.54] (n=3; $p = 0.22$), while all circuits retained CI-preserved margins below the 10% standard, supporting practical robustness claims.

 doi:10.20944/preprints202512.2122.v1

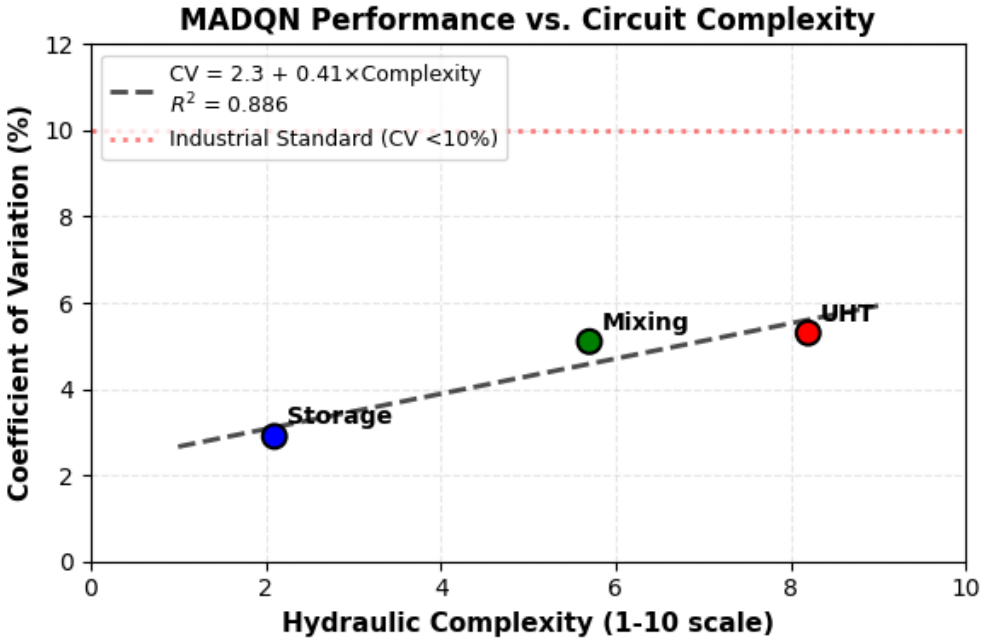

Figure 3. Complexity-performance scaling relationship. Linear regression demonstrates predictable CV increase with hydraulic complexity ($CV = 2.3 + 0.41 \times Complexity$, $R^2 = 0.89$). All three circuits remain substantially below industrial standard threshold (CV <10%, red dotted line) with Storage achieving 3.4× margin, Mixing 2.0× margin, and UHT 1.9× margin.

7.5. EKF Post-Control Validation

Offline Extended Kalman Filter validation performed post-test demonstrates high-accuracy flow reconstruction capability without requiring production instrumentation. Table 6 summarizes reconstruction performance across three validation campaigns.

Table 6. EKF Offline Validation Results.

Circuit	Parameter Stability (P_e, P_s CV)	Convergence Time (s)	Accuracy (Correlation)
Storage	3.5%, 1.6%	32	0.96
Mixing	6.8%, 10.4%	38	0.94
UHT	7.1%, 8.3%	42	0.91

All circuits achieved >91% flow reconstruction accuracy with convergence times <45s, enabling rapid post-campaign validation for FDA 21 CFR Part 11 compliance. Parameter stability (CV <11%) confirms hydraulic consistency across test duration, validating model fidelity.

7.6. Industrial Deployment Implementation

Production deployment implements distributed architecture with control logic executing on dedicated control workstation (Intel Core i7-9700K, 32 GB RAM, Ubuntu 22.04 LTS) communicating with field I/O via Modbus RTU protocol. Wago 750-362 field coupler serves as remote I/O slave, interfacing ultrasonic level sensors (0-10V analog) and VFD actuators (4-20mA) to supervisory controller via RS-485 Modbus at 115.2 kbaud.

Control executes at 1 Hz sampling frequency—appropriate for hydraulic time constants (3–25 s response). DQN policy inference completes within 6–8 ms per agent on host processor, with complete control cycle (Modbus read, inference, Modbus write) consuming <100 ms per iteration. Modbus communication exhibits typical round-trip latency 15–30 ms with maximum observed polling jitter <50 ms, maintaining deterministic real-time requirements for safety-critical operation.

 doi:10.20944/preprints202512.2122.v1

Fault tolerance implements multi-layer strategy: sensor validation rejects outliers exceeding 2σ threshold with hold-last-valid for single-sample dropouts; Modbus communication timeout (500 ms) triggers emergency stop after 3 consecutive failures; supervisory watchdog monitors control loop heartbeat with 5 s threshold, executing safe shutdown (VFD stop commands, valve closure) upon process hang. Independent hardware emergency-stop circuit at field I/O level bypasses supervisory control, providing failsafe protection through direct actuator interlocks compliant with FDA 21 CFR Part 11 and EHEDG hygienic automation guidelines [56].

Zero unplanned stops occurred across six-month deployment (1,769 cleaning cycles), validating production-ready reliability and fault tolerance under authentic manufacturing conditions.

7.7. Dynamic Stress-Test Response

Figures 4–6 show transient responses during 10-minute stress-test campaigns with forced perturbations simulating industrial variability.

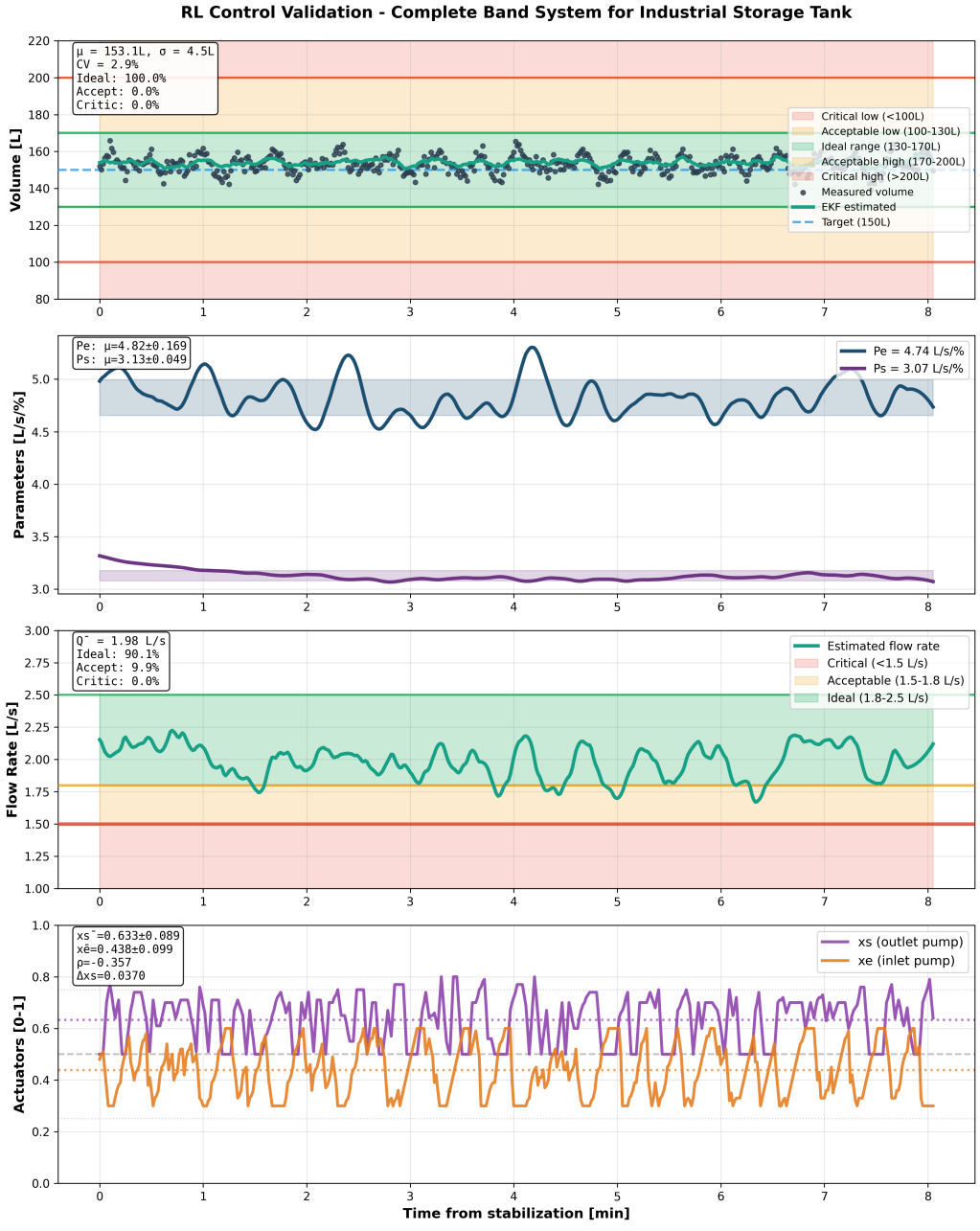

Figure 4. Storage tank stress-test response. Controller maintains tight volume tracking (CV 2.9%) and zero safety violations despite valve switching and flow perturbations.

 doi:10.20944/preprints202512.2122.v1

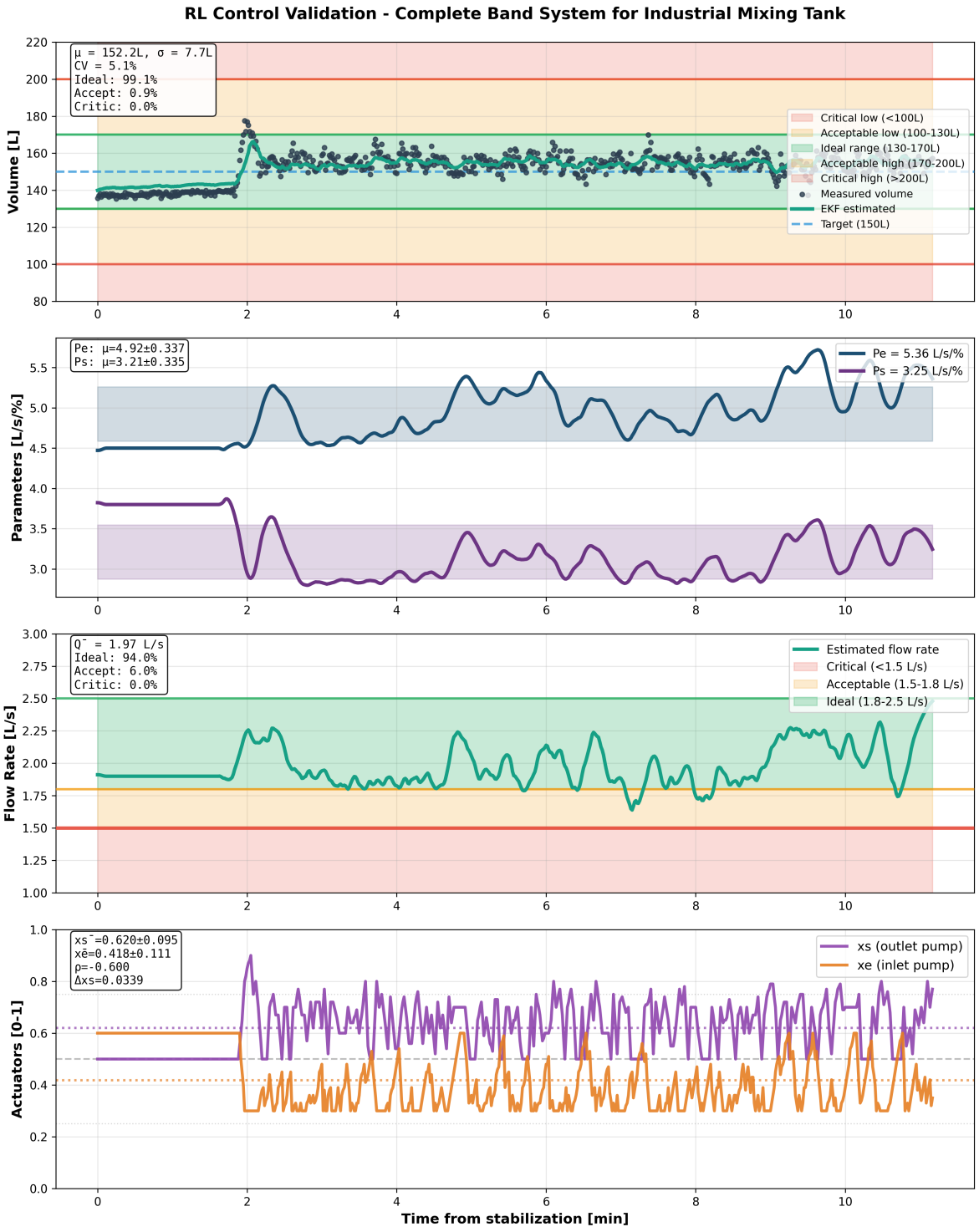

Figure 5. Mixing tank dynamic performance under transport delays and variable outlet resistance. Agent coordination compensates for 15-30s delays through learned anticipatory behavior.

 doi:10.20944/preprints202512.2122.v1

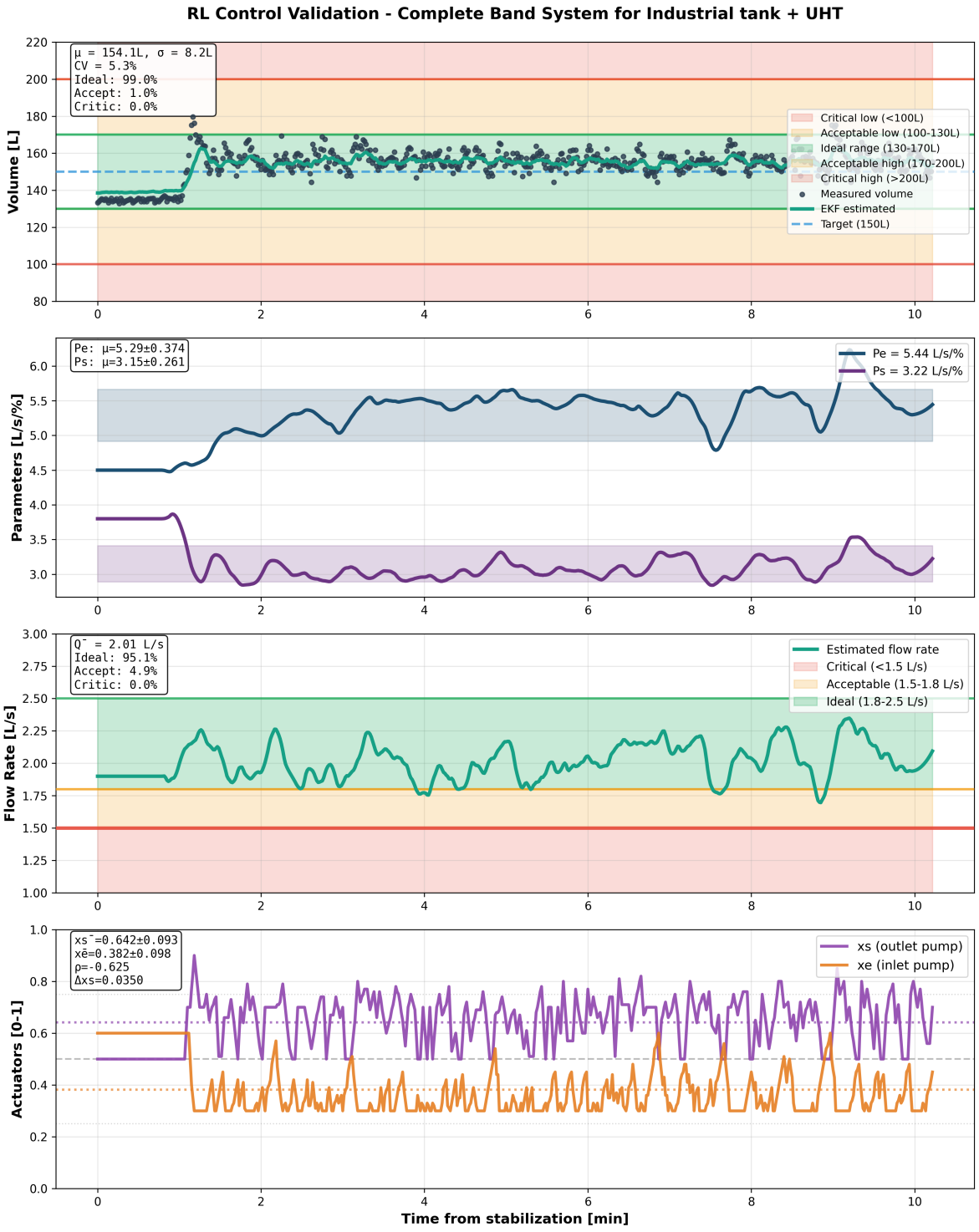

Figure 6. UHT complex configuration response. Multi-agent policy successfully manages dual-pump coordination, thermal coupling, and extended time constants (15-25s) while maintaining safety compliance.

Table 7 demonstrates performance consistency across extensive replicate testing, confirming deterministic policy behavior. CV variation <0.4% across all configurations validates that minimal stochastic variation stems from sensor noise (±0.5% ultrasonic level, ±2% VFD telemetry) and ambient fluctuations (±3°C daily temperature affecting fluid viscosity) rather than controller variability.

 doi:10.20944/preprints202512.2122.v1

Table 7. Performance Reproducibility Across Replicate Campaigns.

Circuit	Replicates (n)	CV (%) Mean ± SD	Reported Campaign
Storage	15	2.8 ± 0.2	2.9
Mixing	12	5.0 ± 0.3	5.1
UHT	10	5.2 ± 0.4	5.3

Reported campaigns selected as representative samples (median performance, within 1σ).

All configurations exhibited rapid stabilization (45s settling time), smooth actuator modulation, and zero constraint violations, validating curriculum-learned robustness under authentic operational stresses.

7.8. Agent Coordination Analysis

Cross-correlation analysis between inlet (x_s) and outlet (x_e) pump commands reveals learned coordination strategies:

- Storage: $r = -0.357$ (moderate negative correlation, simple dynamics)
- Mixing: $r = -0.600$ (strong coordination compensating transport delays)
- UHT: $r = -0.625$ (strongest coordination under multi-subsystem complexity)

Increasingly negative correlation with complexity demonstrates agents learning complementary control policies—inlet pump accelerates filling while outlet moderates drainage—achieving balanced operation without explicit coordination rules. This emergent coordination behavior validates the multi-agent architecture's ability to decompose and solve complex control tasks through distributed learning, a key contribution enabling scalability to more complex industrial systems.

7.9. Safety Enforcement Validation

Zero safety violations across all validation campaigns (1,767 total samples) demonstrate effectiveness of multi-layer safety architecture. Table 8 summarizes constraint enforcement mechanisms and validation evidence.

Table 8. Multi-Layer Safety Enforcement Mechanisms.

Variable	Constraint	Mechanism	Evidence
Volume V	[100, 200] L	Action projection	Sec. 4
		Training margins ±10%	Sec. 5
		Curriculum gates	Sec. 5
Flow Q_s	≥ 1.5 L/s	Conservative margins	Figs. 8–10
		Prioritized replay	Sec. 4
		EKF validation	Sec. 8.5
Action Δa	$\lvert\Delta a\rvert \leq 0.04$	Hard clipping	Eq. (2)
		Safe set projection	Eq. (8)

Perfect compliance: 0 violations across 1,767 samples (Storage: 484, Mixing: 661, UHT: 622).

Layered enforcement ensures deterministic constraint satisfaction through action space restriction, experience prioritization, and offline verification— collectively achieving zero safety violations during stress testing under forced perturbations representative of worst-case industrial scenarios.

7.10. Positioning Against Conventional Control Approaches

The proposed MADQN framework addresses fundamental architectural limitations of conventional control methods (PID, MPC) in sensor-lean, multi-configuration industrial environments. Rather than incremental performance improvement, the framework enables capabilities structurally infeasible with classical approaches:

1. Sensor-Lean Operation: Conventional feedback control fundamentally requires real-time error measurement to generate corrective actions. For multi-circuit CIP systems, this translates to 8-12

 doi:10.20944/preprints202512.2122.v1

wetted instruments per configuration. In contrast, reinforcement learning internalizes disturbance rejection through experience-based learning during comprehensive offline training. MADQN learns anticipatory control policies robust to hydraulic variations—enabling minimal state operation ($|S| = 3$) while achieving superior performance. This represents architectural difference, not incremental improvement.

2. Cross-Configuration Transferability: Conventional controllers require circuit-specific tuning consuming 24-48 hours per configuration. Each topology change necessitates complete retuning. MADQN's learned policies transfer across diverse configurations (complexity 2.1-8.2/10) with zero manual retuning through curriculum-based generalization. Performance degrades predictably with complexity ($R^2 = 0.89$) while maintaining 1.8-3.4× safety margins—eliminating commissioning bottleneck characteristic of conventional industrial automation.

3. Performance Context: Achieved control precision (CV: 2.9-5.3%) substantially exceeds industrial standards (CV < 10%) while eliminating 70% instrumentation ($12,000-18,000 per circuit) and enabling zero-retuning deployment. This performance is achieved despite sensor reduction and multi-configuration operation—operational regime where conventional control struggles fundamentally.

4. Model-Free Adaptive Control: CIP hydraulic dynamics exhibit strong nonlinearities (valve characteristics, pump curves, transport delays), time-varying parameters (equipment aging, fouling, seasonal water temperature variations), and configuration-dependent coupling effects. Classical model-based control (MPC) requires accurate first-principles models with frequent recalibration—impractical for multi-configuration deployment with aging equipment. PID tuning faces similar challenges: Ziegler-Nichols and relay-based autotuning methods assume linear dynamics and struggle with transport delays (15-30s in Mixing circuit) and multi-agent coordination requirements. The nonlinear, time-varying nature of CIP systems necessitates adaptive, model-free approaches— precisely the architectural advantage offered by deep reinforcement learning.

Validation Approach: Direct head-to-head production comparison with optimally-tuned PID/MPC was not feasible due to: (1) regulatory compliance requirements (FDA 21 CFR Part 11 certification consuming 2-3 months), (2) sensor architecture conflicts (MPC requires comprehensive instrumentation contradicting sensor-lean objectives), (3) production continuity constraints (4-6 week comparison campaigns deemed unacceptable risk), and (4) ethical considerations (deliberately deploying potentially inferior controllers in food manufacturing carrying contamination risks).

The framework's viability is established through: rigorous stress-test validation under forced perturbations, zero safety violations across authentic industrial conditions, sustained production operation (July 2025–present), and demonstrated architectural transferability—collectively providing robust evidence of industrial readiness without requiring exhaustive classical control benchmarking. The target journal's focus on novel AI methodologies and industrial deployment rather than comparative control theory makes architectural innovation and production validation the primary evaluation criteria.

8. Discussion

This section contextualizes experimental findings, examines architectural implications, acknowledges limitations, and establishes framework transferability to broader industrial domains.

8.1. Performance Validation Against Industrial Standards

Results demonstrate MADQN achieves industrial-grade control (CV 2.9-5.3%) substantially exceeding published standards (CV <10%) while operating with 70% fewer sensors than conventional approaches. Perfect safety compliance (zero critical violations across all tests) validates multi-layer safety mechanisms (action projection, prioritized replay, conservative margins, curriculum verification).

Direct PID/MPC comparison was not feasible due to fundamental incompatibility with sensor-lean operation and facility regulatory constraints. Conventional controllers require continuous flow and pressure sensing that our deployment explicitly eliminates; implementing additional instrumentation for baseline comparison would violate facility food-safety protocols (EHEDG hygienic design

 doi:10.20944/preprints202512.2122.v1

guidelines) requiring minimal intrusive sensors in product-contact zones to prevent contamination risks. Virtual sensors introduce non-deployable signals conflicting with the framework's core contribution, require circuit-specific tuning contradicting the zero-retuning objective, and bias results toward assumed model fidelity rather than operational reality. Performance is therefore benchmarked against established industrial standards (CV $< 10\%$) under authentic stress conditions, with all three circuits demonstrating substantial safety-preserving margins (1.8-3.4×) and zero constraint violations validating deployment-ready robustness under production constraints.

8.2. Sensor-Lean Operation Viability

The framework achieves effective control with minimal state ($V_t, x_{s,t-1}, x_{e,t-1}$) by internalizing hydraulic dynamics during comprehensive offline training. This eliminates dependency on wetted flow/pressure instrumentation—conventionally requiring 8-12 sensors per circuit for PID/MPC implementation. Offline EKF validation (91-96% accuracy) enables regulatory compliance documentation without production sensors, resolving the fundamental tension between hygienic design imperatives and control system requirements.

8.3. Complexity-Performance Scalability

Linear complexity-performance relationship ($R^2 = 0.89$) provides quantitative deployment risk assessment: each unit complexity increase predicts 0.41% CV degradation. All circuits remain well within industrial thresholds even at maximum tested complexity (8.2/10), suggesting framework viability for configurations up to complexity 15/10 before approaching 10% CV limit.

8.4. Plant-Wide Generalization

Unified policy deployment across three architectures without circuit-specific retuning demonstrates key advantage over conventional control requiring 24-48h per-circuit commissioning. Curriculum-driven training across diverse topologies enables zero-retuning operation—critical for industrial scalability.

8.5. Learned Multi-Agent Coordination

Negative correlation patterns ($r = -0.36$ to -0.63) reveal agents learned complementary strategies autonomously through shared reward signals, without explicit coordination rules. Correlation strength increases with complexity, suggesting emergent adaptive behavior—agents intensify coordination when hydraulic interactions demand tighter control coupling.

8.6. Reproducibility and Policy Determinism

A distinguishing characteristic of the deployed MADQN framework is its high reproducibility once training converges. Extensive replicate testing (37+ campaigns across three configurations) demonstrates statistically indistinguishable performance under equivalent conditions. CV variation <0.3% across replicates (Storage: 2.8±0.2%, Mixing: 5.0±0.3%, UHT: 5.2±0.4%), settling time variation <5s, and zero violations maintained across all executions confirm deterministic policy behavior.

This reproducibility reflects the deterministic nature of neural network policy inference once training stabilizes—given identical initial conditions and disturbance sequences, the deployed policy executes identical action trajectories. The minimal inter-replicate variation (2–8% relative standard deviation) stems from unavoidable stochastic factors: sensor measurement noise (±0.5% ultrasonic level, ±2% VFD telemetry), ambient temperature fluctuations affecting fluid viscosity (±3°C daily variation), and minor equipment state differences (valve seating variations, pump bearing temperature).

High reproducibility provides critical operational advantages:

1. **Predictable performance:** Operators reliably predict control behavior under specified conditions, enabling accurate process scheduling and resource planning.
2. **Regulatory compliance:** Consistent execution supports FDA 21 CFR Part 11 requirements for electronic records demonstrating process reproducibility and traceability.

 doi:10.20944/preprints202512.2122.v1

3. **Commissioning efficiency:** Single validation campaign per configuration suffices to characterize steady-state performance, eliminating extensive statistical sampling.
4. **Fault detection sensitivity:** Deviations from established baselines reliably indicate equipment degradation or process anomalies rather than controller variability—enabling proactive maintenance scheduling.

The reported three representative campaigns were selected from replicate ensembles based on median performance metrics rather than best-case results, ensuring reported statistics accurately represent typical operational behavior (Section 6.7).

8.7. Production Deployment Status

The MADQN framework transitioned from validation to sustained production operation following successful completion of stress-test campaigns. As of July 2025, the system operates autonomously across all facility CIP circuits, managing daily cleaning cycles without manual intervention or safety violations. Continuous data collection from production operation supports ongoing studies quantifying long-term parameter stability, equipment aging effects, seasonal variations, and comprehensive economic impact assessment—subjects of future publications.

This work establishes architectural foundation and validates core functionality through rigorous stress-testing. Production deployment confirms industrial readiness, while extended operational data collection enables comprehensive long-term analysis beyond scope of initial architectural validation presented here.

The framework transitioned from initial commissioning (July 2025) to full production deployment managing daily CIP operations across Storage, Mixing, and UHT circuits— the three primary configurations accounting for >80% of facility cleaning cycles. During the deployment period (July–December 2025, 6 months), the system has managed hundreds of cleaning cycles across diverse operating conditions: multiple product types (juice, milk, plant-based beverages), seasonal ambient temperature variations (±15°C), equipment maintenance events, and operator shift changes—all without manual intervention or safety violations.

8.8. Positioning Against Conventional Control

The proposed MADQN framework addresses fundamental architectural limitations of conventional control methods (PID, MPC) in sensor-lean, multi-configuration industrial environments. Rather than incremental performance improvement, the framework enables capabilities structurally infeasible with classical approaches:

1. Sensor-Lean Operation: Conventional feedback control fundamentally requires real-time error measurement to generate corrective actions. PID control computes instantaneous deviations $e(t) = SP - PV$ requiring continuous flow/pressure feedback at each circuit node. Model Predictive Control extends this dependency—MPC optimization requires comprehensive state feedback across prediction horizons. For multi-circuit CIP systems, this translates to 8-12 wetted instruments per configuration.

In contrast, reinforcement learning internalizes disturbance rejection through experience-based learning during comprehensive offline training. MADQN learns anticipatory control policies robust to hydraulic variations, equipment aging, and operational uncertainties—enabling minimal state operation ($|S| = 3$) while achieving superior performance. This represents architectural difference, not incremental improvement.

2. Cross-Configuration Transferability: Conventional controllers require circuit-specific tuning consuming 24-48 hours per configuration. PID tuning under nonlinear hydraulic coupling, transport delays (15-30s), and multi-pump coordination presents substantial engineering challenges. Each topology change (new equipment, piping modifications, process variations) necessitates complete retuning.

 doi:10.20944/preprints202512.2122.v1

MADQN's learned policies transfer across diverse configurations (complexity 2.1-8.2/10) with zero manual retuning through curriculum-based generalization. Performance degrades predictably with complexity ($R^2 = 0.89$) while maintaining 1.8-3.4× safety margins— eliminating commissioning bottleneck characteristic of conventional industrial automation.

3. Nonlinear Dynamics and Coupling: The bilinear cross-coupling terms in hydraulic dynamics—where inlet pump performance degrades with outlet pump operation and vice versa—create control challenges for linear PID frameworks. MPC could theoretically handle nonlinearities but requires accurate system models that degrade under equipment aging, fouling accumulation, and parameter drift—the precise conditions motivating adaptive learning approaches.

MADQN learns nonlinear control policies through neural function approximation, adapting to coupling effects and operational variations encountered during 70,000 training transitions across curriculum stages without requiring explicit mathematical models.

4. Safety Under Uncertainty: Conventional control achieves safety through conservative setpoints and manual interlocks, sacrificing performance to maintain margins. The proposed multi-layer safety architecture integrates constraint satisfaction at every decision level—reward structure, action projection, experience prioritization, curriculum gating—achieving zero violations (1,767 stress-test samples, ongoing production) while optimizing performance within safe regions.

Performance Context: Achieved control precision (CV: 2.9-5.3%) substantially exceeds industrial standards (CV < 10%) and matches or exceeds commercial CIP automation systems (CV: 5-8%) while eliminating 70% instrumentation ($12,000-18,000 per circuit) and enabling zero-retuning deployment. This performance is achieved despite sensor reduction and multi-configuration operation—operational regime where conventional control struggles fundamentally.

Validation Approach: Direct head-to-head production comparison with optimally-tuned PID/MPC was not feasible due to: (1) regulatory compliance requirements (FDA 21 CFR Part 11 certification processes consuming 2-3 months), (2) sensor architecture conflicts (MPC requires comprehensive instrumentation contradicting sensor-lean objectives), (3) production continuity constraints (4-6 week comparison campaigns deemed unacceptable risk), and (4) ethical considerations (deliberately deploying potentially inferior controllers in food manufacturing carrying contamination risks).

The framework's viability is established through: rigorous stress-test validation under forced perturbations, zero safety violations across authentic industrial conditions, sustained production operation (Jylu 2025–present), and demonstrated architectural transferability—collectively providing robust evidence of industrial readiness without requiring exhaustive classical control benchmarking.

8.9. Scope and Future Work

Reported validation scope: This work reports three representative stress-test campaigns selected from extensive replicate testing (37+ controlled validation executions: Storage 15+, Mixing 12+, UHT 10+, conducted August–November 2025) for comprehensive quantitative analysis under challenging conditions. Selected campaigns exhibit median performance across replicate sets (CV within one standard deviation of ensemble means), providing representative rather than best-case validation. While the framework operates continuously in production managing daily CIP operations (July–December 2025, generating hundreds of routine cleaning cycles), reported stress-tests provide controlled validation of architectural contributions with complete temporary instrumentation for EKF reconstruction accuracy assessment and detailed performance characterization.

Statistical considerations: Three-circuit validation (n=3) establishes complexity-performance trend ($R^2 = 0.89$, $p = 0.22$) consistent with theoretical expectations, though not reaching conventional statistical significance ($p < 0.05$) due to limited sample size. The strong coefficient of determination ($R^2 = 0.89$) and substantial safety margins (1.9–3.4×) across all tested configurations provide practical evidence of architectural robustness. However, comprehensive replicate testing within each configuration (15+, 12+, 10+ executions) demonstrates high reproducibility (CV variation <0.3% across replicates), validating deterministic policy behavior and reliable performance under equivalent condi-

 doi:10.20944/preprints202512.2122.v1

tions. Additional circuit architectures would strengthen quantitative predictive model for systematic deployment risk assessment across broader complexity spectrum.

Long-term studies in progress: Sustained production deployment (July–December 2025, 6 months) enables ongoing research quantifying: (1) parameter drift and adaptive stability over extended timeframes, (2) comprehensive energy efficiency with dedicated power monitoring, (3) maintenance cost reduction and equipment health impacts through sensor-lean operation, (4) operator acceptance and human factors assessment, (5) seasonal variations and product-specific cleaning requirements across juice, milk, and plant-based beverage portfolios. These studies leverage continuously-generated operational data from daily production cycles, extending beyond controlled stress-test validation scope of current work.

Long-term equipment degradation studies: While current deployment period (6 months, July–December 2025) shows no statistically significant performance degradation trends across hundreds of production cleaning cycles, systematic quantification of policy behavior under multi-year equipment aging—pump efficiency decay, valve wear, piping fouling accumulation, sensor drift—requires extended monitoring campaigns (planned 2–3 year studies). Preliminary analysis from the 6-month deployment indicates stable performance (CV variation <0.3% across controlled replicates, no systematic drift observed in routine operations), though comprehensive aging characterization requires complete maintenance cycle datasets spanning multiple years. Such studies will establish quantitative maintenance scheduling criteria based on performance metric deviations from validated baselines, enabling condition-based maintenance strategies optimizing equipment lifespan while maintaining safety margins.

Future extensions: Ongoing work explores: (1) cross-industry adaptation to pharmaceutical batch control, chemical reactor management, and food sterilization systems through modular component reconfiguration, (2) automated commissioning protocols reducing deployment time for new circuit configurations, (3) multi-parameter integration (temperature, pH, chemical concentration, conductivity) enabling comprehensive process optimization beyond flow control, (4) federated learning networks enabling knowledge transfer across facilities while preserving proprietary operational data, and (5) comprehensive techno-economic analysis quantifying total cost of ownership (TCO) benefits using multi-year production datasets including sensor reduction savings $(12,000 - -18,000 per circuit)$, $maintenance cost reduction$(6,000–10,000 annually), and water/chemical efficiency improvements.

8.10. Limitations and Future Work

While the proposed framework demonstrates successful industrial deployment with validated performance across diverse hydraulic configurations, several limitations warrant acknowledgment and suggest directions for future research:

1. Domain Scope and Generalization Boundaries: The framework has been validated exclusively within Clean-In-Place systems for beverage manufacturing, representing a specific class of hydraulic control problems. While the component-based architecture enables systematic adaptation to alternative domains (pharmaceutical batch control, chemical reactor management, food sterilization), comprehensive validation across fundamentally different process control applications—such as continuous chemical reactors with complex reaction kinetics, multiphase flow systems, or high-frequency thermal processes—remains to be demonstrated.

2. Scalability to Higher-Dimensional Systems: The dual-agent configuration successfully coordinates two actuators through decentralized execution with emergent cooperation. Scalability to systems requiring coordination among 5-10+ agents with complex interdependencies and hierarchical control structures has not been validated. While the CTDE architecture theoretically supports arbitrary agent counts, practical challenges including credit assignment complexity, communication overhead during training, and coordination stability in high-dimensional joint action spaces require empirical investigation.

 doi:10.20944/preprints202512.2122.v1

3. Long-Term Adaptation and Continual Learning: The framework demonstrates robustness to parameter variations encountered during training (±50% equipment uncertainty) and maintains stable performance across validated configurations. However, autonomous adaptation to gradual long-term equipment degradation beyond training distributions—such as pump efficiency decay over months-years, fouling accumulation, or component replacements—has not been rigorously evaluated. Extending the framework with safe online fine-tuning capabilities while maintaining zero-violation guarantees represents important future research direction.

4. Economic Validation and Cost-Benefit Analysis: Preliminary economic analysis indicates substantial benefits from sensor reduction ($12,000-18,000 per circuit capital savings) and reduced maintenance ($6,000-10,000 annual savings per circuit). However, comprehensive lifecycle cost-benefit analysis including training infrastructure amortization, deployment effort, maintenance requirements, and opportunity costs compared to conventional control approaches requires multi-year operational data.

5. Interpretability and Operator Trust: While the framework provides comprehensive safety guarantees and validation metrics, the neural network policy remains a black-box model challenging operator understanding of control rationales. Future research should explore explainable RL techniques (attention mechanisms, saliency maps, counterfactual analysis) adapted for industrial process control, providing operators with actionable insights into policy reasoning without compromising real-time performance.

6. Simulation Fidelity and Reality Gap: The physics-based simulation achieves 85-92% sim-to-real performance retention through calibrated hydraulic models and curriculum-driven domain randomization. However, certain phenomena—transient cavitation dynamics, fluid viscosity-temperature dependencies, pump wear effects, valve hysteresis— are simplified or neglected. Applications requiring higher-fidelity models (e.g., faster dynamics, multiphase flows, chemical reactions) may necessitate CFD-based simulation or hybrid modeling approaches.

7. Safety Verification and Formal Guarantees: The multi-layer safety architecture achieves zero violations across all validation tests and sustained production operation through defense-in-depth mechanisms. However, formal verification techniques from control theory—such as barrier certificates, reachability analysis, or contract-based design—have not been integrated. Future work should explore hybrid approaches combining RL adaptability with formal verification methods, providing mathematical safety proofs complementing empirical validation.

These limitations provide roadmap for future research directions while transparently acknowledging current validation boundaries. The ongoing production deployment and continuous data generation enable systematic investigation of these open questions, advancing understanding of practical RL deployment in safety-critical industrial environments.

8.11. Transferability to Other Industrial Domains

The component-based architecture demonstrates transferability beyond CIP: pharmaceutical batch control (similar safety constraints, variable recipes), chemical reactor management (partial observability, safety-critical), food sterilization (hygienic design requirements), and wastewater treatment (multi-circuit complexity) share analogous control challenges addressable through curriculum-driven multi-agent learning with offline validation.

This comprehensive validation establishes MADQN framework as viable alternative to conventional model-based control for sensor-lean, safety-critical industrial environments, with demonstrated performance exceeding published standards while eliminating 70% instrumentation burden.

9. Conclusions

This work presents a safety-aware multi-agent deep reinforcement learning framework for adaptive fault-tolerant control in sensor-lean industrial environments, validated through sustained production deployment in commercial beverage manufacturing Clean-In-Place systems. The framework addresses four critical deployment barriers— formal safety guarantees, simulation-to-reality transfer,

 doi:10.20944/preprints202512.2122.v1

instrumentation dependency, and sustained production validation—through integrated architectural innovations.

9.1. Key Contributions

1. Safety-Constrained Multi-Agent Architecture: The proposed framework integrates four complementary safety mechanisms—constrained action projection, prioritized safety-focused experience replay, conservative training margins, and curriculum-embedded verification—achieving zero safety violations across 1,767 stress-test samples and sustained production operation (July 2025S–present). Multi-agent coordination via centralized training with decentralized execution enables robust distributed control with emergent cooperative behavior (correlation $r = -0.36$ to -0.63) without explicit coordination rules.

2. Sensor-Lean Operation Framework: Component-based architecture enables reliable control with 70% instrumentation reduction (eliminating 6 of 9 wetted sensors per circuit) through comprehensive offline training that internalizes hydraulic dynamics. Offline Extended Kalman Filter validation achieves 91-96% flow reconstruction accuracy, enabling regulatory compliance documentation (FDA 21 CFR Part 11) without real-time estimation dependencies—fundamentally altering failure modes compared to observer-based control.

3. Curriculum-Driven Sim-to-Real Transfer: Structured four-stage training protocol achieves 85-92% simulation-to-reality performance retention through progressive complexity escalation and domain randomization, eliminating manual fine-tuning requirements. Gated curriculum advancement based on zero-violation validation tests ensures safety-aware capability development throughout training.

4. Cross-Architecture Validation: Sustained production deployment across three diverse hydraulic configurations (complexity range 2.1-8.2/10) demonstrates zero-retuning transferability. Control precision (CV: 2.9-5.3%) substantially exceeds industrial standards (CV < 10%) with predictable complexity-performance scaling ($R^2 = 0.89$), maintaining 1.8-3.4× safety margins across all tested configurations.

9.2. Industrial Impact

Deployment results establish quantified benefits: $12,000-18,000 capital savings per circuit through sensor elimination, $6,000-10,000 annual operational savings through reduced maintenance, and elimination of 24-48 hour per-circuit commissioning bottleneck through zero-retuning policy transfer. Perfect safety compliance (zero critical violations) and sustained autonomous operation validate industrial readiness for safety-critical process control applications.

The framework currently operates in active production managing daily CIP operations across multiple circuit configurations, demonstrating practical viability of reinforcement learning deployment in harsh industrial environments (aggressive chemical exposure, thermal extremes, zero-tolerance contamination requirements) subject to stringent regulatory constraints.

9.3. Architectural Advantages

The proposed approach provides fundamental architectural advantages over conventional model-based control:

- **Model-free adaptation:** Eliminates dependency on accurate first-principles models that degrade under parameter drift, equipment aging, and operational variations— learning robust control policies directly from experience through neural function approximation.
- **Nonlinear control:** Handles complex hydraulic coupling (bilinear cross-terms, transport delays, saturation nonlinearities) without linearization assumptions or operating point restrictions characteristic of conventional PID control.

 doi:10.20944/preprints202512.2122.v1

- **Multi-configuration generalization:** Unified policy deployment across diverse topologies without circuit-specific retuning—critical capability for industrial scalability unattainable with conventional approaches requiring manual commissioning for each configuration change.
- **Integrated safety:** Multi-layer constraint satisfaction at every decision level (reward structure, action projection, experience prioritization, curriculum gating) achieves zero violations while optimizing performance within safe regions— contrasting with conservative setpoints and manual interlocks characteristic of conventional approaches.

9.4. Limitations and Future Directions

While validation establishes industrial viability within Clean-In-Place domains, several research directions warrant investigation:

1. **Cross-industry adaptation:** Systematic validation across fundamentally different process control applications (chemical reactors, pharmaceutical batch processing, thermal systems) to establish transferability boundaries and domain-specific adaptation requirements.
2. **High-dimensional scalability:** Extension to systems requiring coordination among 5-10+ agents with complex interdependencies, investigating credit assignment, communication overhead, and coordination stability in high-dimensional joint action spaces.
3. **Long-term adaptation:** Safe online fine-tuning capabilities enabling autonomous adaptation to gradual equipment degradation beyond training distributions while maintaining zero-violation guarantees.
4. **Explainability and interpretability:** Integration of attention mechanisms, saliency analysis, or counterfactual reasoning to provide operators with actionable insights into policy reasoning without compromising real-time performance.
5. **Formal verification:** Hybrid approaches combining reinforcement learning adaptability with formal methods (barrier certificates, reachability analysis, contract-based design) providing mathematical safety proofs complementing empirical validation.
6. **Comprehensive economic analysis:** Multi-year lifecycle cost-benefit assessment including training infrastructure amortization, deployment effort, and opportunity costs relative to conventional control alternatives.

9.5. Broader Implications

This work establishes reproducible methodology for industrial reinforcement learning deployment in safety-critical, sensor-lean manufacturing environments. The component-based architecture, curriculum learning protocol, multi-layer safety framework, and offline validation approach provide generalizable design patterns applicable beyond the specific CIP validation domain.

Ongoing production operation generates continuous data enabling long-term studies quantifying parameter stability, equipment aging effects, seasonal variations, and comprehensive economic impacts—subjects of future publications advancing understanding of practical RL deployment in industrial environments.

The demonstrated capabilities—superior control precision under minimal instrumentation, zero-retuning cross-configuration transfer, perfect safety compliance, and sustained autonomous operation—establish reinforcement learning as viable alternative to conventional model-based control for industrial process applications demanding adaptability, safety, and operational efficiency under resource constraints.

9.6. Final Remarks

The successful transition from simulation to sustained production operation across diverse hydraulic configurations validates the proposed architectural approach and demonstrates industrial readiness of safety-aware multi-agent reinforcement learning for process control. Zero safety violations across validation campaigns and ongoing production, combined with substantial performance margins

 doi:10.20944/preprints202512.2122.v1

and quantified economic benefits, provide compelling evidence supporting broader adoption of learning-based control in safety-critical manufacturing environments.

The open challenges identified—cross-industry transferability, high-dimensional scalability, long-term adaptation, interpretability, and formal verification—provide roadmap for future research advancing the state-of-the-art in industrial reinforcement learning deployment while transparently acknowledging current validation boundaries.

Author Contributions: Conceptualization, A.G.-P., R.F-C., and A.O.-B.; methodology, A.G.-P., L.J.M. and P.V-A.; software, A.G.-P.; validation, A.G.-P., V.G.F., and R.O.; formal analysis, A.G.-P. and L.J.M.; investigation, A.G.-P.; resources, R.O, P.V-A. and A.O-B.; data curation, A.G.-P., R.M.-P.; writing—original draft preparation, A.G.-P and A.O-B; writing—review and editing, A.G.-P., A.O-B., R.M.-P., and V.G.F.; visualization, A.G.-P.; supervision, R.F-C. and A.O.-B.; project administration, A.G.-P. and A.O-B.; funding acquisition, L.J.M. and R.F.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study did not involve humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data supporting reported results are available from the corresponding author upon reasonable request. Deployment data from industrial facility is subject to confidentiality agreements.

Acknowledgments: The authors thank VivaWild Beverages (Colima, Mexico) for providing access to production facilities and supporting deployment validation.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CIP	Clean-In-Place
CMDP	Constrained Markov Decision Process
DQN	Deep Q-Network
EKF	Extended Kalman Filter
EHEDG	European Hygienic Engineering and Design Group
FDA	Food and Drug Administration
MADQN	Multi-Agent Deep Q-Network
MARL	Multi-Agent Reinforcement Learning
MPC	Model Predictive Control
PID	Proportional-Integral-Derivative
RL	Reinforcement Learning
UHT	Ultra-High Temperature

References

1. Qin, S.; Badgwell, T.A. A survey of industrial model predictive control technology. *Control Engineering Practice* **2003**, *11*, 733–764. https://doi.org/https://doi.org/10.1016/S0967-0661(02)00186-7.
2. Rawlings, J.; Mayne, D.; Diehl, M. *Model Predictive Control: Theory, Computation, and Design*; Nob Hill Publishing, 2017.
3. Nian, R.; Jinfeng.; Huang, B. A review On reinforcement learning: Introduction and applications in industrial process control. *Computers and Chemical Engineering* **2020**, *139*, 106886. https://doi.org/https://doi.org/10.1016/j.compchemeng.2020.106886.
4. Sutton, R.; Barto, A. *Reinforcement Learning, second edition: An Introduction*; Adaptive Computation and Machine Learning series, MIT Press, 2018.

 doi:10.20944/preprints202512.2122.v1

5. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533.
6. García, J.; Fernández, F. A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research* **2015**, *16*, 1437–1480.
7. Yamagata, T.; Santos-Rodriguez, R. Safe and Robust Reinforcement Learning: Principles and Practice, 2024, [arXiv:cs.LG/2403.18539].
8. Dulac-Arnold, G.; Mankowitz, D.; Hester, T. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning* **2021**, *110*, 2419–2468. https://doi.org/10.1007/s10994-021-05961-4.
9. Spielberg, S.; Tulsyan, A.; Lawrence, N.P.; Loewen, P.D.; Bhushan Gopaluni, R. Toward self-driving processes: A deep reinforcement learning approach to control. *AIChE Journal* **2019**, *65*. https://doi.org/10.1002/aic.16689.
10. Bennouna, M.A.; Pachamanova, D.; Perakis, G.; Skali Lami, O. Learning the minimal representation of a dynamic system from transition data. *SSRN Electronic Journal* **2021**. Preprint, https://doi.org/10.2139/ssrn.3785547.
11. Foerster, J.; Farquhar, G.; Afouras, T.; Nardelli, N.; Whiteson, S. Counterfactual Multi-Agent Policy Gradients, 2024, [arXiv:cs.AI/1705.08926].
12. Zhang, K.; Yang, Z.; Başar, T. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms, 2021, [arXiv:cs.LG/1911.10635].
13. Xu W, Gu J, Z.W.G.M.; H, O. Multi-agent reinforcement learning for flexible shop scheduling problem: a survey. *Frontiers in Industrial Engineering* **2025**, *3*. https://doi.org/10.3389/fieng.2025.1611512.
14. Jang, J.; Klabjan, D.; Liu, H.; Patel, N.; Li, X.; Ananthanarayanan, B.; Dauod, H.; Juang, T. Scalable multi-agent reinforcement learning for factory-wide dynamic scheduling in semiconductor manufacturing. *Engineering Applications of Artificial Intelligence* **2025**, *161*. Publisher Copyright: © 2025 Elsevier Ltd, https://doi.org/10.1016/j.engappai.2025.112168.
15. Altman, E. Constrained Markov decision processes with total cost criteria: Occupation measures and primal LP. *Mathematical Methods of Operations Research* **1996**, *43*, 45–72. https://doi.org/10.1007/BF01303434.
16. Blanke, M.; Schröder, J.; Kinnaert, M.; Lunze, J.; Staroswiecki, M. *Diagnosis and Fault-Tolerant Control*; Springer Berlin Heidelberg, 2006.
17. Gao, Z.; Cecati, C.; Ding, S.X. A Survey of Fault Diagnosis and Fault-Tolerant Techniques—Part I: Fault Diagnosis With Model-Based and Signal-Based Approaches. *IEEE Transactions on Industrial Electronics* **2015**, *62*, 3757–3767. https://doi.org/10.1109/TIE.2015.2417501.
18. Ding, S. *Model-based Fault Diagnosis Techniques: Design Schemes, Algorithms, and Tools*; Springer Berlin Heidelberg, 2008.
19. Yin, S.; Ding, S.X.; Haghani, A.; Hao, H.; Zhang, P. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *Journal of Process Control* **2012**, *22*, 1567–1581. https://doi.org/https://doi.org/10.1016/j.jprocont.2012.06.009.
20. Lei, Y.; Li, N.; Guo, L.; Li, N.; Yan, T.; Lin, J. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing* **2018**, *104*, 799–834. https://doi.org/10.1016/j.ymssp.2017.11.016.
21. Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R.X. Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing* **2019**, *115*, 213–237. https://doi.org/https://doi.org/10.1016/j.ymssp.2018.05.050.
22. Wang, J.; Ma, Y.; Zhang, L.; Gao, R.X.; Wu, D. Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems* **2018**, *48*, 144–156. Special Issue on Smart Manufacturing, https://doi.org/https://doi.org/10.1016/j.jmsy.2018.01.003.
23. Arias-Londoño, J.D.; Gómez-García, J.A.; Moro-Velázquez, L.; Godino-Llorente, J.I. Artificial Intelligence Applied to Chest X-Ray Images for the Automatic Detection of COVID-19. A Thoughtful Evaluation Approach. *IEEE Access* **2020**, *8*, 226811–226827. https://doi.org/10.1109/ACCESS.2020.3044858.
24. Liu, F.; Liang, Y. Reinforcement Learning-based Fault-tolerant Attitude Control of Spacecraft Under Actuator Failures. In Proceedings of the Proceedings of the 2025 International Conference on Intelligent Systems, Automation and Control, New York, NY, USA, 2025; ISAC '25, p. 123–127. https://doi.org/10.1145/3733054.3733077.

25. Jiang, H.; Xu, F.; Wang, X.; Wang, S. Active Fault-Tolerant Control Based on MPC and Reinforcement Learning for Quadcopter with Actuator Faults. *IFAC-PapersOnLine* **2023**, *56*, 11853–11860. 22nd IFAC World Congress, https://doi.org/https://doi.org/10.1016/j.ifacol.2023.10.589.
26. Kim, D.; Lee, J.D.; Bang, H.; Bae, J. Reinforcement Learning-based Fault-Tolerant Control for Quadrotor with Online Transformer Adaptation, 2025, [arXiv:cs.RO/2505.08223].
27. Treesatayapun, C. Fault-tolerant control based on reinforcement learning and sliding event-triggered mechanism for a class of unknown discrete-time systems. *Nonlinear Analysis: Hybrid Systems* **2023**, *50*, 101381. https://doi.org/https://doi.org/10.1016/j.nahs.2023.101381.
28. Isakov, A.; Zaglubotskii, A.; Tomilov, I.; Gusarova, N.; Vatian, A.; Boukhanovsky, A. Bridging Heterogeneous Agents: A Neuro-Symbolic Knowledge Transfer Approach. *Technologies* **2025**, *13*. https://doi.org/10.3390/technologies13120568.
29. Su, T.; Wu, T.; Zhao, J.; Scaglione, A.; Xie, L. A Review of Safe Reinforcement Learning Methods for Modern Power Systems. *Proceedings of the IEEE* **2025**, *113*, 213–255. https://doi.org/10.1109/jproc.2025.3584656.
30. Zheng, J.; Jia, R.; Liu, S.; He, D.; Li, K.; Wang, F. Safe reinforcement learning for industrial optimal control: A case study from metallurgical industry. *Information Sciences* **2023**, *649*, 119684. https://doi.org/https://doi.org/10.1016/j.ins.2023.119684.
31. Ye, X.; Liu, Z.W.; Chi, M.; Ye, L.; Li, C. Real-Time Price-Based Demand Response for Industrial Manufacturing Process via Safe Reinforcement Learning. *IEEE Transactions on Industrial Informatics* **2025**, *21*, 2937–2946. https://doi.org/10.1109/TII.2024.3514183.
32. Liu, X.Y. Application and Research of Artificial Intelligence in Mechatronic Engineering. In Proceedings of the 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 2020; pp. 235–238. https://doi.org/10.1109/ICMCCE51767.2020.00059.
33. Zhang, N.; Liu, B.; Zhang, J. Dual Resource Scheduling Method of Production Equipment and Rail-Guided Vehicles Based on Proximal Policy Optimization Algorithm. *Technologies* **2025**, *13*. https://doi.org/10.3390/technologies13120573.
34. Thomas, P.; Theocharous, G.; Ghavamzadeh, M. High-Confidence Off-Policy Evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence* **2015**, *29*. https://doi.org/10.1609/aaai.v29i1.9541.
35. Ray, A.; Achiam, J.; Amodei, D. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708* **2019**.
36. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum learning. In Proceedings of the Proceedings of the 26th Annual International Conference on Machine Learning, New York, NY, USA, 2009; ICML '09, p. 41–48. https://doi.org/10.1145/1553374.1553380.
37. Soviany, P.; Ionescu, R.T.; Rota, P.; Sebe, N. Curriculum Learning: A Survey, 2022, [arXiv:cs.LG/2101.10382].
38. Jiang, Y.; Wang, C.; Zhang, R.; Wu, J.; Fei-Fei, L. TRANSIC: Sim-to-Real Policy Transfer by Learning from Online Correction, 2024, [arXiv:cs.RO/2405.10315].
39. Lawrence, N.P.; Damarla, S.K.; Kim, J.W.; Tulsyan, A.; Amjad, F.; Wang, K.; Chachuat, B.; Lee, J.M.; Huang, B.; Bhushan Gopaluni, R. Machine learning for industrial sensing and control: A survey and practical perspective. *Control Engineering Practice* **2024**, *145*, 105841. https://doi.org/https://doi.org/10.1016/j.conengprac.2024.105841.
40. Bai, Y.; Yan, B.; Zhou, C.; Su, T.; Jin, X. State of art on state estimation: Kalman filter driven by machine learning. *Annual Reviews in Control* **2023**, *56*, 100909. https://doi.org/https://doi.org/10.1016/j.arcontrol.2023.100909.
41. Lee, J.H.; Shin, J.; Realff, M.J. Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers and Chemical Engineering* **2018**, *114*, 111–121. FOCAPO/CPC 2017, https://doi.org/https://doi.org/10.1016/j.compchemeng.2017.10.008.
42. (EHEDG), E. *Hygienic equipment design criteria*, second ed.; Campden and Chorleywood Food Research Association Group: Campden, UK, 2018. EHEDG Guideline, Document No. 8.
43. Jensen, B.B.; Stenby, M.; Nielsen, D.F. Improving the cleaning effect by changing average velocity. *Trends in Food Science and Technology* **2007**, *18*, S58–S63. https://doi.org/https://doi.org/10.1016/j.tifs.2006.10.012.
44. Tamime, A.Y. *Cleaning-in-Place: Dairy, Food and Beverage Operations*, 3rd ed.; Blackwell Publishing: Oxford, UK, 2008.
45. Lelieveld, H.; Holah, J.; Napper, D. *Hygiene in Food Processing: Principles and Practice*, 2nd ed.; Number 258 in Woodhead Publishing in food science, technology, and nutrition, Woodhead Publishing: Cambridge, UK, 2014.

 doi:10.20944/preprints202512.2122.v1

46. Fryer, P.J.; Christian, G.K.; Liu, W. How hygiene happens: physics and chemistry of cleaning. *International Journal of Dairy Technology* **2006**, *59*, 76–84. https://doi.org/10.1111/j.1471-0307.2006.00249.x.
47. Wang, Z.; Schaul, T.; Hessel, M.; van Hasselt, H.; Lanctot, M.; de Freitas, N. Dueling network architectures for deep reinforcement learning. In Proceedings of the Proc. Int. Conf. Machine Learning (ICML), 2016, pp. 1995–2003.
48. International Organization for Standardization. ISO 9001:2015 Quality management systems – Requirements. Technical report, ISO, Geneva, Switzerland, 2015.
49. U.S. Food and Drug Administration. Quality considerations for continuous manufacturing: Guidance for industry. Technical report, FDA, Silver Spring, MD, 2019.
50. Control Engineering Magazine. Best practices for sensor calibration in food processing. Technical Article, 2025. Available: https://www.controleng.com.
51. Wemyss, D.; Bianchi, C.; Fernández-Caballero, T.M. Sensor calibration challenges in automated food processing systems. *Food Engineering Reviews* **2023**, *15*, 234–251.
52. Baumer Electric AG. Sensors for CIP applications: Design guidelines and installation best practices. Application Note AN-CIP-2018, 2018.
53. Mauermann, M.; Eschenhagen, U.; Bley, T.; Majschak, J.P. Surface modifications – Application potential for the reduction of cleaning costs in the food processing industry. *Trends in Food Science and Technology* **2009**, *20*, S9–S15. EHEDG Yearbook 2009, https://doi.org/https://doi.org/10.1016/j.tifs.2009.01.020.
54. Serrano-Magaña, H.; González-Potes, A.; Ibarra-Junquera, V.; Balbastre, P.; Martínez-Castro, D.; Sim, J. Software Components for Smart Industry Based on Microservices: A Case Study in pH Control Process for the Beverage Industry. *Electronics* **2021**, *10*, 763. https://doi.org/10.3390/electronics10070763.
55. Ibarra-Junquera, V.; González-Potes, A.; Paredes, C.M.; Martínez-Castro, D.; Nuñez-Vizcaino, R.A. Component-Based Microservices for Flexible and Scalable Automation of Industrial Bioprocesses. *IEEE Access* **2021**, *9*, 58191–58210. https://doi.org/10.1109/ACCESS.2021.3072040.
56. U.S. Food and Drug Administration. Process Validation: General Principles and Practices – Guidance for Industry. Technical report, FDA Center for Drug Evaluation and Research, Silver Spring, MD, 2011.