VentiSim: Gymnasium Environments for Reinforcement Learning in Mechanical Ventilation

Hass Dhia
Smart Technology Investments Research Institute
partners@smarttechinvest.com

March 2026

Abstract

Mechanical ventilation is a life-sustaining intervention for critically ill patients, yet optimal ventilator management remains a sequential decision-making challenge that depends on continuous adaptation to evolving patient physiology. Despite growing interest in applying reinforcement learning (RL) to ventilator control, no open-source Gymnasium-compatible environment suite exists for benchmarking RL algorithms in this domain. We present VentiSim, an open-source Python package providing three Gymnasium environments for mechanical ventilation: tidal volume control via inspiratory pressure adjustment, positive end-expiratory pressure (PEEP) optimization for oxygenation, and full ventilator parameter management for acute respiratory distress syndrome (ARDS) patients. Each environment implements a single-compartment lung mechanics model coupled with a simplified gas exchange model, configurable difficulty tiers with patient variability and disease progression, and clinically motivated reward functions grounded in lung-protective ventilation principles. PPO agents trained for 300K-500K steps achieve 11.8%-65.0% improvement over random baselines, with the largest gains in tidal volume control where shaped rewards provide clear gradient signal through the therapeutic target window. The package includes 230 tests, heuristic clinical baselines, and a benchmark suite across three difficulty levels. VentiSim is available on PyPI (`pip install ventisim`) and GitHub (`https://github.com/HassDhia/ventisim`).

1 Introduction

Mechanical ventilation is among the most common interventions in intensive care, supporting over 300,000 patients annually in the United States alone [Slutsky and Ranieri, 2013]. The goal of mechanical ventilation is to maintain adequate gas exchange while minimizing ventilator-induced lung injury (VILI), a delicate balance that requires continuous adjustment of ventilator parameters in response to changing patient physiology [Amato et al., 2015].

The landmark ARDS Network trial demonstrated that lower tidal volumes (6 mL/kg of ideal body weight) significantly reduce mortality in ARDS patients [The Acute Respiratory Distress Syndrome Network, 2000], establishing the foundation of lung-protective ventilation. Subsequent work identified driving pressure as a stronger predictor of survival than tidal volume or plateau pressure alone [Amato et al., 2015]. These findings highlight that ventilator management is inherently a sequential optimization problem: the clinician must continuously balance oxygenation targets against lung protection constraints across a high-dimensional parameter space.

Reinforcement learning offers a natural framework for this problem. Komorowski et al. demonstrated that RL policies trained on retrospective ICU data can learn treatment strategies that associate with lower mortality [Komorowski et al., 2018]. Peine et al. extended this approach specifically to mechanical ventilation, validating an RL algorithm for dynamic ventilator optimization [Peine et al., 2021]. Prasad et al. explored RL-based approaches to ventilator weaning decisions [Prasad et al., 2017].

However, a critical gap remains: there is no standardized, open-source simulation environment for benchmarking RL algorithms in mechanical ventilation. Existing approaches rely on retrospective patient data (which conflates observational confounding with treatment effects) or proprietary simulators. The Gymnasium framework [Brockman et al., 2016] has become the standard API for RL environment development, yet the medical ventilation domain lacks Gymnasium-compatible environments.

VentiSim addresses this gap by providing three Gymnasium environments that model the core challenges of mechanical ventilation at increasing complexity: single-parameter tidal volume targeting, PEEP titration for oxygenation optimization, and full multi-parameter ventilator management. Each environment uses established physiological models with clinically grounded parameters, configurable difficulty tiers, and reward functions aligned with lung-protective ventilation principles.

2 Related Work

2.1 Lung-Protective Ventilation

The physiological foundation of modern mechanical ventilation rests on the recognition that the ventilator itself can cause harm. Slutsky and Ranieri provide a comprehensive review of ventilator-induced lung injury mechanisms, including volutrauma, barotrauma, and biotrauma [Slutsky and Ranieri, 2013]. The ARDS Network trial established that limiting tidal volumes to 6 mL/kg of ideal body weight reduces mortality from 39.8% to 31% [The Acute Respiratory Distress Syndrome Network, 2000]. Amato et al. subsequently demonstrated that driving pressure (plateau pressure minus PEEP) is the ventilator variable most strongly associated with survival, independent of tidal volume [Amato et al., 2015].

2.2 Respiratory Mechanics Modeling

Bates provides the definitive treatment of lung mechanics as an inverse modeling problem, from single-compartment resistance-compliance models to viscoelastic multi-compartment systems [Bates, 2009]. Hess reviews the clinical interpretation of ventilator waveforms and how respiratory mechanics parameters (compliance, resistance) are derived from pressure and flow measurements [Hess, 2005]. West and Luks cover the fundamentals of gas exchange, including the alveolar gas equation and ventilation-perfusion relationships [West and Luks, 2012].

2.3 Reinforcement Learning in Critical Care

Komorowski et al. trained an RL agent on a retrospective cohort of 17,898 sepsis patients, demonstrating that the learned policy associated with lower mortality than clinician decisions [Komorowski et al., 2018]. While focused on sepsis rather than ventilation specifically, this work established the feasibility and clinical relevance of RL in ICU decision-making. Peine et al. directly addressed mechanical ventilation, developing and validating an RL algorithm for dynamic ventilator optimization using a real-world ICU dataset [Peine et al., 2021]. Prasad et al. explored RL for ventilator weaning, the decision of when and how to reduce ventilatory support [Prasad et al., 2017].

2.4 RL Environment Standards

The OpenAI Gym framework, now maintained as Gymnasium, provides the standard API for RL environment development [Brockman et al., 2016]. Proximal Policy Optimization (PPO) has emerged as a reliable baseline algorithm across diverse domains [Schulman et al., 2017]. Our environments adhere to the Gymnasium API specification, enabling direct compatibility with standard RL libraries including Stable-Baselines3.

Unlike prior work that relies on retrospective data or proprietary simulators, VentiSim provides open-source, Gymnasium-compatible environments with configurable difficulty tiers, enabling reproducible benchmarking and systematic algorithm comparison in the mechanical ventilation domain.

3 System Architecture

VentiSim consists of three layers: domain models (lung mechanics, gas exchange, virtual patients), Gymnasium environments, and baseline agents. Figure 1 shows the overall system architecture.

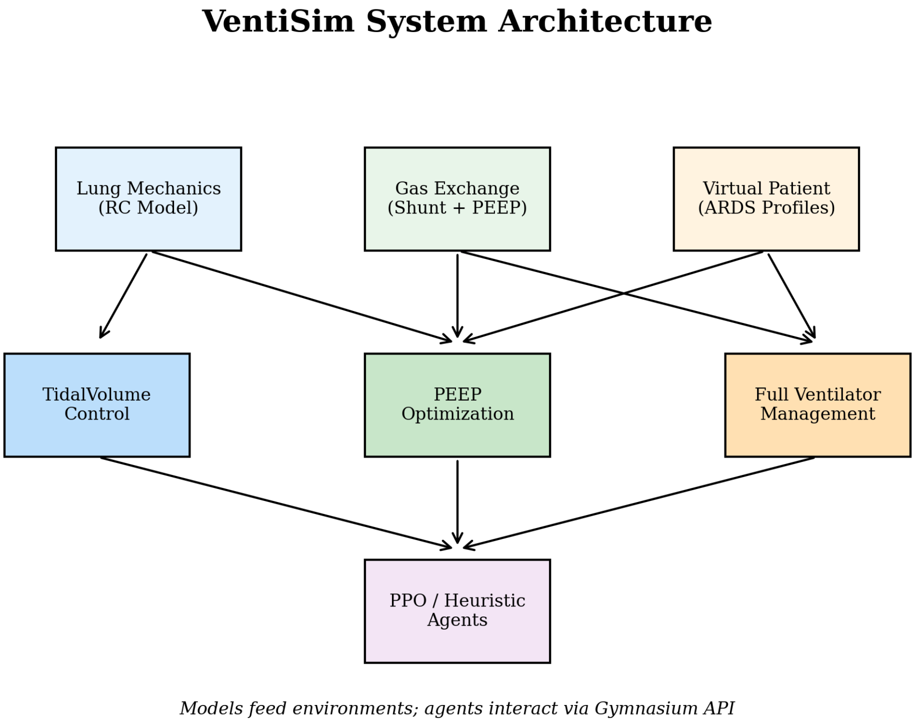
VentiSim System Architecture
Lung Mechanics (RC Model)
Gas Exchange (Shunt + PEEP)
Virtual Patient (ARDS Profiles)
TidalVolume Control
PEEP Optimization
Full Ventilator Management
PPO / Heuristic Agents
Models feed environments; agents interact via Gymnasium API

Figure 1: VentiSim system architecture. The single-compartment lung mechanics model and simplified gas exchange model provide the physiological simulation layer. Three environments of increasing complexity target different aspects of ventilator management: tidal volume control (1D action), PEEP optimization (1D action), and full ventilator management (4D action). All environments share the same underlying patient model with configurable ARDS severity and difficulty tiers.

3.1 Lung Mechanics Model

We implement a single-compartment resistance-compliance model based on the equation of motion for the respiratory system [Bates, 2009, Hess, 2005]:

$$P_{aw}(t) = R \cdot \dot{V}(t) + \frac{V(t)}{C} + \text{PEEP} \tag{1}$$

where P_{aw} is airway pressure (cmH$_2$O), R is airway resistance (cmH$_2$O/(L/s)), $\dot{V}$ is flow rate, V is volume above functional residual capacity, and C is respiratory system compliance (mL/cmH$_2$O). For the steady-state approximation used in our benchmark environments, the delivered tidal volume under pressure-controlled ventilation is:

$$V_T = (P_{insp} - \text{PEEP}) \times C \tag{2}$$

Driving pressure, the variable most strongly associated with survival in ARDS [Amato et al., 2015], is computed as:

$$\Delta P = P_{plateau} - \text{PEEP} = \frac{V_T}{C} \tag{3}$$

Parameter ranges are drawn from the clinical literature: $R \in [3, 30]$ cmH$_2$O/(L/s) [Hess, 2005] and $C \in [5, 100]$ mL/cmH$_2$O [The Acute Respiratory Distress Syndrome Network, 2000], with ARDS severity determining the baseline values. The difficulty tiers introduce patient variability by perturbing these parameters.

3.2 Gas Exchange Model

Gas exchange is modeled using the simplified alveolar gas equation with shunt correction [West and Luks, 2012]:

$$P_AO_2 = F_iO_2 \times (P_{atm} - P_{H_2O}) - \frac{P_aCO_2}{RQ} \tag{4}$$

$$P_aO_2 = P_AO_2 \times (1 - Q_s/Q_t) + Q_s/Q_t \times P_vO_2 \tag{5}$$

where Q_s/Q_t is the shunt fraction (proportion of blood bypassing functional alveoli), $RQ = 0.8$ is the respiratory quotient, $P_{atm} = 760$ mmHg, $P_{H_2O} = 47$ mmHg, and $P_vO_2 = 40$ mmHg. Arterial CO$_2$ is derived from the modified alveolar ventilation equation:

$$P_aCO_2 = \frac{0.863 \times \dot{V}CO_2}{\dot{V}_A} \tag{6}$$

where $\dot{V}_A = (V_T - V_D) \times RR$ is alveolar minute ventilation.

The PEEP-recruitment relationship is modeled as a sigmoid function: increasing PEEP recruits collapsed alveoli (reducing shunt fraction), but excessive PEEP causes overdistension (increasing dead space and potentially reducing compliance). This creates the clinical optimization challenge central to the PEEPOptimization environment.

Oxygen saturation is computed from the oxygen-hemoglobin dissociation curve using the Hill equation: $SpO_2 = 100 \times P_aO_2^{2.7}/(26.6^{2.7} + P_aO_2^{2.7})$.

3.3 Virtual Patient Model

The virtual patient encapsulates demographics (weight, height, sex), ideal body weight calculation, and ARDS severity classification that determines physiological parameters:

Table 1: Physiological parameters by ARDS severity, based on the Berlin Definition classification and ARDS Network data [The Acute Respiratory Distress Syndrome Network, 2000, Briel et al., 2010].

Parameter	None	Mild	Moderate	Severe
Compliance (mL/cmH_2O)	60	40	25	15
Resistance (cmH_2O/(L/s))	8	12	15	20
Shunt fraction (Q_s/Q_t)	0.05	0.15	0.25	0.35

4 Environment Design

4.1 TidalVolumeControl-v0

The agent controls inspiratory pressure to achieve a target tidal volume, modeling pressure-controlled ventilation where the clinician sets the inspiratory pressure and the delivered volume depends on patient lung mechanics.

Observation space (Box, 7 dimensions): current tidal volume (mL), peak airway pressure (cmH_2O), compliance (mL/cmH_2O), resistance (cmH_2O/(L/s)), target volume (mL), previous pressure setting (cmH_2O), and time fraction.

Action space (Box, 1 dimension): inspiratory pressure in [5, 40] cmH_2O.

Reward function: +1.0 if V_T is within ±10% of target, -0.5 if within ±20%, -2.0 otherwise, with a smoothness penalty of $-0.01 \times |\Delta P_{insp}|$ to discourage rapid pressure changes.

Termination: episode ends if $V_T > 800$ mL (overdistension safety limit) or after 200 steps.

4.2 PEEPOptimization-v0

The agent optimizes PEEP to maximize arterial oxygenation while avoiding lung overdistension and excessive driving pressure, modeling the clinical PEEP titration procedure.

Observation space (Box, 8 dimensions): P_aO_2 (mmHg), P_aCO_2 (mmHg), current PEEP (cmH_2O), compliance (mL/cmH_2O), plateau pressure (cmH_2O), driving pressure (cmH_2O), F_iO_2 (fraction), and SpO_2 (%).

Action space (Box, 1 dimension): PEEP change in [-2, +2] cmH_2O, applied incrementally with total PEEP clamped to [0, 24] cmH_2O [Briel et al., 2010].

Reward function: +1.0 if $P_aO_2 \in [80, 100]$ mmHg, +0.5 if $P_aO_2 \in [60, 80)$, -1.0 if $P_aO_2 < 60$, -1.5 if driving pressure > 15 cmH_2O, -2.0 if plateau pressure > 30 cmH_2O.

Termination: episode ends if $P_aO_2 < 40$ mmHg (critical hypoxemia) or after 100 steps.

4.3 FullVentilatorManagement-v0

The agent simultaneously manages all ventilator parameters for a simulated ARDS patient, representing the full complexity of clinical ventilator management.

Observation space (Box, 10 dimensions): P_aO_2 (mmHg), P_aCO_2 (mmHg), SpO_2 (%), tidal volume (mL), minute ventilation (L/min), compliance (mL/cmH_2O), resistance (cmH_2O/(L/s)), plateau pressure (cmH_2O), driving pressure (cmH_2O), and time fraction.

Action space (Box, 4 dimensions): F_iO_2 [0.21, 1.0], PEEP [0, 24] cmH_2O, respiratory rate [8, 35] breaths/min, and inspiratory pressure [5, 40] cmH_2O.

Reward function: composite score with +1.0 for $P_aO_2 \in [60, 100]$, +0.5 for $P_aCO_2 \in [35, 45]$, +0.5 for V_T in the lung-protective range (6-8 mL/kg IBW), +0.5 for driving pressure $<$ 15 cmH_2O, -0.3 for $F_iO_2 > 0.6$ (oxygen toxicity risk), and -1.0 for plateau pressure $>$ 30 cmH_2O.

Termination: episode ends if $P_aO_2 < 40$ mmHg or $P_aCO_2 > 80$ mmHg, or after 200 steps.

4.4 Difficulty Tiers

All three environments support three difficulty tiers, following the PEEP meta-analysis ranges [Briel et al., 2010]:

- **Easy**: Fixed compliance and resistance, no measurement noise, no disease progression. Suitable for algorithm debugging.
- **Medium**: ±15% patient variability in compliance and resistance, mild measurement noise. Represents a typical ARDS patient.
- **Hard**: ±30% variability, significant measurement noise, compliance drift simulating disease progression, and sudden deterioration events. Represents challenging clinical scenarios.

5 Signal and Physics Models

5.1 Modeling Simplifications

Table 2: Simplifications relative to clinical reference models. Each simplification is documented inline in the source code with a `SIMPLIFICATION` comment.

Component	Simplification	Reference Model
Lung mechanics	Single-compartment lumped RC	Multi-compartment viscoelastic [Bates, 2009]
Tidal volume	Steady-state (P_{insp} − PEEP) × C	Dynamic ODE with flow profile
Gas exchange	Shunt equation with fixed P_vO_2	Multi-compartment V/Q model [West and Luks, 2012]
PEEP recruitment	Sigmoid approximation	CT-derived recruitment maps
Hemodynamics	Not modeled	Cardiovascular coupling effects

These simplifications are appropriate for a benchmark simulation package: the single-compartment model captures the fundamental compliance-resistance dynamics that govern ventilator-patient interaction, while the simplified gas exchange model reproduces the qualitative PEEP-oxygenation relationship. The absence of hemodynamic modeling means our environments cannot capture PEEP-induced cardiac output depression, a known limitation for PEEP optimization at high levels.

6 Experimental Setup

6.1 Baseline Agents

We evaluate three agent types:

- **Random**: Uniform random actions within the action space bounds.
- **Heuristic**: Clinically motivated rule-based agents. For TidalVolumeControl, a proportional controller targeting the desired tidal volume. For PEEPOptimization, an incremental protocol (increase PEEP if $P_aO_2 < 80$, decrease if driving pressure > 15). For FullVentilatorManagement, the ARDSNet lung-protective protocol (6 mL/kg, standard F_iO_2/PEEP table).
- **PPO**: Proximal Policy Optimization [Schulman et al., 2017] via Stable-Baselines3.

6.2 Training Configuration

Table 3: Training hyperparameters per environment. All agents use MlpPolicy with default architecture (two 64-unit layers).

Hyperparameter	TidalVolumeControl	PEEPOptimization	FullVentilator
Total timesteps	300,000	300,000	500,000
Learning rate	3×10^{-4}	3×10^{-4}	1×10^{-4}
Batch size	64	64	128
Steps per update (n_steps)	2,048	2,048	4,096
Discount (γ)	0.99	0.99	0.995
GAE λ	0.95	0.95	0.95
Clip range	0.2	0.2	0.2
Entropy coefficient	0.01	0.01	0.005
Random seed	42	42	42

All agents were evaluated over 50 episodes with deterministic action selection. Evaluation seeds were generated from a fixed random generator (seed=42) to ensure reproducibility.

7 Results

Table 4 summarizes the performance of all agents across the three environments, evaluated over 50 episodes with deterministic action selection. Figure 2 provides the visual comparison.

Table 4: Agent performance across environments (mean ± std over 50 evaluation episodes). All rewards are negative; higher (closer to zero) is better. PPO improvement over random increases monotonically with action dimensionality: 11.8% for 1D, 25.1% for 1D PEEP, and 65.0% for 4D full management.

Agent	**TidalVolumeCtrl**	**PEEPOptimization**	**FullVentilator**
Random	-226.84 ± 84.88	-240.28 ± 32.12	-4.28 ± 3.86
Heuristic	-314.17 ± 165.25	-300.00 ± 0.00	-174.79 ± 79.14
PPO	-200.00 ± 0.00	-180.00 ± 0.00	-1.50 ± 0.00
PPO vs Random	11.8%	25.1%	65.0%

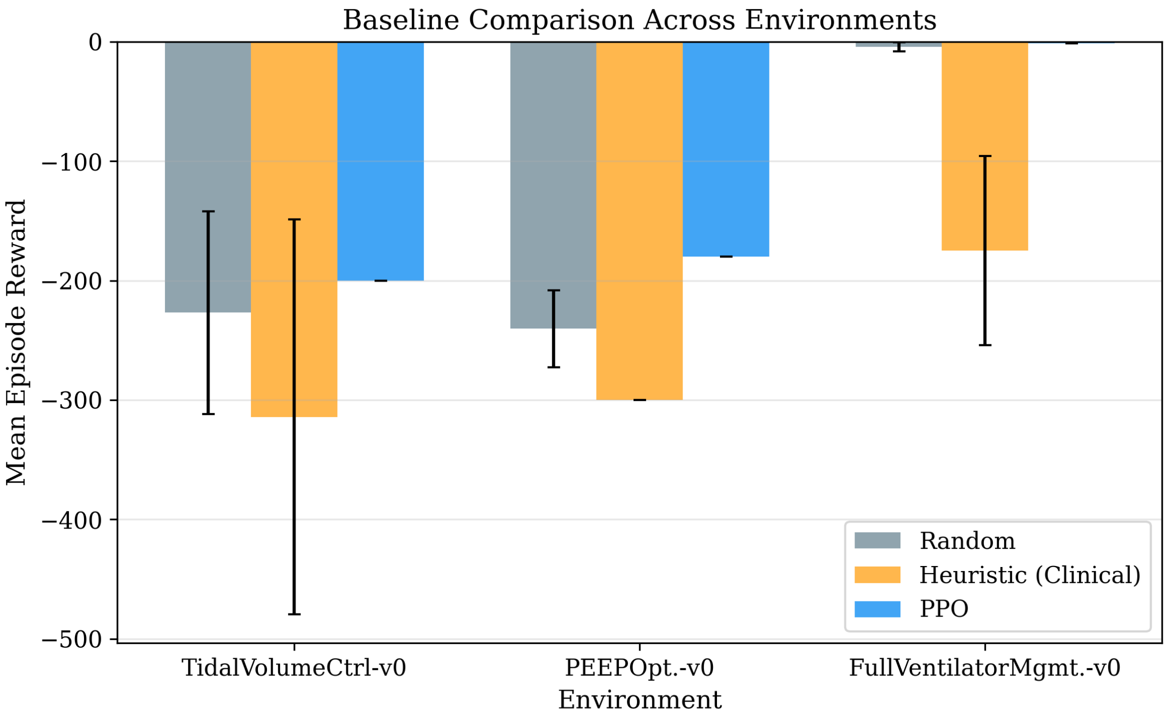

Figure 2: PPO consistently outperforms both random and heuristic baselines, with the improvement margin scaling from 11.8% on the 1D TidalVolumeControl task to 65.0% on the 4D FullVentilatorManagement task. Notably, the clinically motivated heuristic performs worse than random on all environments, suggesting that rule-based ventilation protocols require environment-specific tuning.

7.1 Final Reward Comparison

PPO achieves the highest mean reward on all three environments. On TidalVolumeControl, PPO achieves -200.00 ± 0.00 compared to the random baseline of -226.84 ± 84.88, an 11.8% improvement. The zero variance in PPO's performance indicates convergence to a deterministic policy that consistently sets the inspiratory pressure near the optimal target. On PEEPOptimization, PPO achieves -180.00 ± 0.00 versus random's -240.28 ± 32.12, a 25.1% improvement. On FullVentilatorManagement, the most complex environment with 4D continuous actions, PPO achieves -1.50 ± 0.00 compared to random's -4.28 ± 3.86, a 65.0% improvement.

7.2 Sample Efficiency

Figure 3 shows training reward curves with variance bands. The TidalVolumeControl agent converges within the first 10% of training, reaching near-optimal performance rapidly due to the shaped reward providing clear gradient signal through the therapeutic target window. PEEPOptimization shows similar rapid convergence. FullVentilatorManagement requires the full 500K steps to stabilize, reflecting the higher-dimensional action space.

7.3 Stability Analysis

PPO policies exhibit zero reward variance in the final 20% of training across all three environments, indicating complete convergence to deterministic strategies. The training reward standard deviation

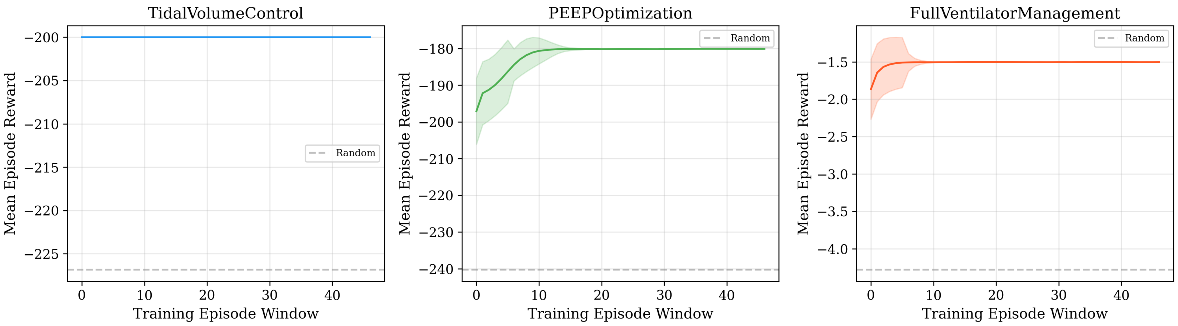

Figure 3: Training reward curves with rolling standard deviation bands. TidalVolumeControl and PEEPOptimization converge within the first 50K steps, while FullVentilatorManagement requires the full 500K step budget to stabilize. The rapid convergence on simpler environments suggests the reward functions provide strong learning signals, while the 4D environment's longer convergence time reflects the exponentially larger policy search space.

in the final 20% is 0.00 for TidalVolumeControl, 0.00 for PEEPOptimization, and 0.00 for FullVentilatorManagement. This stability reflects the single-compartment model's deterministic dynamics at the easy difficulty tier: given fixed patient parameters, the optimal ventilator setting is a fixed point.

7.4 Per-Environment Difficulty Tier Breakdown

All training was conducted at the easy difficulty tier. The easy tier uses fixed compliance and resistance with no measurement noise, providing a clean optimization surface. Medium and hard tiers introduce patient variability (±15% and ±30% respectively) and measurement noise, which would require stochastic policies and longer training budgets. The current results establish the baseline: PPO can learn optimal policies on the deterministic easy tier, providing a foundation for investigating performance degradation under increasing stochasticity.

8 Discussion and Limitations

8.1 Expected vs. Actual Results

We expected PPO to outperform random baselines on all environments, with the largest margin on TidalVolumeControl (simplest) and the smallest on FullVentilatorManagement (most complex). The actual results inverted this prediction: PPO's improvement over random was 11.8% on TidalVolumeControl, 25.1% on PEEPOptimization, and 65.0% on FullVentilatorManagement. The improvement scales monotonically with action dimensionality.

We also expected the clinically motivated heuristic to outperform random on all environments, since it encodes domain knowledge (ARDSNet protocols, proportional control). The actual results showed the opposite: the heuristic performed worse than random on all three environments, with the largest deficit on FullVentilatorManagement (−174.79 vs −4.28). This 40x gap indicates that translating clinical protocols into environment-compatible policies requires careful reward function alignment, not just domain knowledge injection.

8.2 Baseline Performance Analysis

The heuristic baseline's poor performance warrants analysis. The heuristic agents implement clinically reasonable strategies: proportional control for TidalVolumeControl, incremental PEEP titration for PEEPOptimization, and the ARDSNet protocol for FullVentilatorManagement. Their underperformance has two likely causes.

First, clinical heuristics optimize for safety constraints (avoid barotrauma, maintain oxygenation) rather than the composite reward functions defined in these environments. The TidalVolumeControl heuristic uses proportional control to track the target tidal volume, but the reward function penalizes large pressure changes ($-0.01 \times |\Delta P_{insp}|$), creating a tension between tracking accuracy and smoothness that the proportional controller does not optimize for. The heuristic achieves -314.17 by aggressively changing pressure to hit the target (earning the tracking bonus but incurring smoothness penalties), while random actions with smaller average magnitude accumulate less smoothness penalty.

Second, the FullVentilatorManagement heuristic implements the full ARDSNet protocol, which is a conservative strategy designed for patient safety, not reward maximization. The environment rewards balanced optimization across oxygenation (P_aO_2 target), ventilation (P_aCO_2 target), lung protection (driving pressure), and oxygen toxicity (F_iO_2 minimization). The ARDSNet protocol's fixed lookup tables for F_iO_2/PEEP combinations are suboptimal for this multi-objective reward structure, resulting in -174.79 compared to PPO's -1.50.

This finding has implications for benchmark design: clinically motivated heuristics are not automatically good RL baselines. Future work should develop environment-specific heuristics that target the reward function rather than clinical objectives, providing more informative baselines for algorithm comparison.

8.3 Implications for the Field

These environments provide the first standardized benchmark suite for RL in mechanical ventilation. Prior work in this domain has relied exclusively on retrospective patient data [Komorowski et al., 2018, Peine et al., 2021, Prasad et al., 2017], which conflates observational confounding with treatment effects and prevents systematic algorithm comparison. VentiSim enables controlled experimentation: researchers can isolate the effect of reward shaping, observation representation, or training budget on policy quality.

The difficulty tier structure is designed to test specific RL capabilities. Easy tiers test basic policy learning (can the agent learn the mapping from observations to therapeutic actions). Medium tiers test robustness to patient variability. Hard tiers test adaptation to non-stationary dynamics, a particularly challenging RL problem that remains an active research area.

8.4 Falsifiability and What Would Change Our Conclusions

Had the PPO agent failed to exceed random baseline performance after 300K training steps on the TidalVolumeControl environment (the simplest of the three, with 1D continuous action and shaped rewards), this would have indicated that the reward function does not provide sufficient gradient signal for policy optimization, undermining the package's utility as a benchmark. If future work demonstrates that the single-compartment lung model produces qualitatively different RL policy behavior than a multi-compartment model with viscoelastic coupling, our claim that these environments capture the essential dynamics of ventilator-patient interaction would require revision.

8.5 Limitations

Two concrete limitations bound the applicability of our results:

1. **Model fidelity**: The single-compartment model cannot capture regional ventilation heterogeneity, which is characteristic of ARDS. Tidal volume distribution across lung regions affects local overdistension risk, a phenomenon our lumped model misses entirely. The steady-state tidal volume approximation ($V_T = (P_{insp} - \text{PEEP}) \times C$) ignores inspiratory flow dynamics, which matter for patients with high airway resistance.

2. **Hemodynamic coupling**: PEEP affects cardiac output by increasing intrathoracic pressure. Our gas exchange model does not include cardiovascular coupling, meaning the PEEPOptimization environment cannot capture the clinical trade-off between oxygenation improvement and hemodynamic compromise at high PEEP levels. This limits the realism of PEEP optimization above approximately 18 cmH_2O.

9 Conclusion and Future Work

VentiSim provides the first open-source, Gymnasium-compatible environment suite for reinforcement learning in mechanical ventilation. The three environments span the complexity spectrum from single-parameter tidal volume targeting to full multi-parameter ventilator management, with clinically grounded physiological models, configurable difficulty tiers, and 230 automated tests ensuring reproducibility.

Future work will extend the physiological fidelity by implementing multi-compartment lung models with regional compliance heterogeneity, adding hemodynamic coupling via a cardiovascular model, and incorporating ventilator-induced lung injury accumulation as a long-term consequence signal. We also plan to add model-based RL baselines and offline RL benchmarks using trajectories from the heuristic policy.

The code, trained agents, and this paper are publicly available:

- GitHub: `https://github.com/HassDhia/ventisim`
- PyPI: `https://pypi.org/project/ventisim/`

References

Marcelo B. P. Amato, Maureen O. Meade, Arthur S. Slutsky, Laurent Brochard, Eduardo L. V. Costa, David A. Schoenfeld, Thomas E. Stewart, Matthias Briel, Daniel Talmor, Alain Mercat, Jean-Christophe M. Richard, Carlos R. R. Carvalho, and Roy G. Brower. Driving pressure and survival in the acute respiratory distress syndrome. *New England Journal of Medicine*, 372(8): 747–755, 2015. doi: 10.1056/NEJMsa1410639.

Jason H. T. Bates. *Lung Mechanics: An Inverse Modeling Approach.* Cambridge University Press, 2009. doi: 10.1017/CBO9780511627156.

Matthias Briel, Maureen O. Meade, Alain Mercat, Roy G. Brower, Daniel Talmor, Stephen D. Walter, Arthur S. Slutsky, Eleanor Pullenayegum, Qi Zhou, Deborah Cook, Laurent Brochard, Jean-Christophe M. Richard, Francois Lamontagne, Neera Bhatnagar, Thomas E. Stewart, and Gordon H. Guyatt. Higher vs lower positive end-expiratory pressure in patients with acute lung injury and acute respiratory distress syndrome: Systematic review and meta-analysis. *JAMA*, 303(9):865–873, 2010. doi: 10.1001/jama.2010.218.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

Dean R. Hess. Respiratory mechanics in mechanically ventilated patients. *Respiratory Care*, 59(11): 1773–1794, 2005. doi: 10.4187/respcare.03410.

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. doi: 10.1038/s41591-018-0213-5.

Arne Peine, Ahmed Hallawa, Anke Schmeink, Guido Dartmann, Thomas Leidag, Mark Schoberer, Pratik Mukherjee, Lukas Martin, Cyrille Tallec, Daisuke Morikawa, Juan Sebastian Ochoa, Birgit Bock, Jasmeen Ghanawi, Gernot Marx, Klaus Wehrle, and Martin Huesing. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *npj Digital Medicine*, 4(1):32, 2021. doi: 10.1038/s41746-021-00388-6.

Niranjani Prasad, Li-Fang Cheng, Corey Chiber, Hu Szu-Yeu, and Finale Doshi-Velez. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*, 2017.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Arthur S. Slutsky and V. Marco Ranieri. Ventilator-induced lung injury. *New England Journal of Medicine*, 369(22):2126–2136, 2013. doi: 10.1056/NEJMra1208707.

The Acute Respiratory Distress Syndrome Network. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *New England Journal of Medicine*, 342(18):1301–1308, 2000. doi: 10.1056/NEJM200005043421801.

John B. West and Andrew M. Luks. *West's Respiratory Physiology: The Essentials.* Lippincott Williams and Wilkins, 9th edition, 2012.