AnestheSim: Gymnasium Environments for Reinforcement Learning in Automated Anesthesia Drug Dosing

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

March 2026

Abstract

Automated anesthesia drug dosing remains a significant clinical challenge due to pharmacokinetic delays, inter-patient variability, and multi-drug synergistic interactions. We present AnestheSim, a suite of three Gymnasium-compatible reinforcement learning environments for automated anesthesia control: propofol infusion targeting Bispectral Index (BIS) via the Marsh three-compartment pharmacokinetic model, remifentanil effect-site concentration targeting via the Minto model, and combined propofol-remifentanil anesthesia management using the Greco synergistic interaction surface. Each environment features three difficulty tiers with configurable patient variability and surgical stimulation events. We evaluate Proximal Policy Optimization (PPO) agents against random and heuristic Target-Controlled Infusion (TCI) baselines across all environments. Our results reveal an asymmetry in RL tractability: PPO rapidly masters remifentanil dosing but struggles with propofol control, suggesting that pharmacokinetic timescale—not task complexity—is the primary determinant of sample efficiency. AnestheSim provides 109 validated tests, configurable difficulty levels, and standardized benchmarks for reproducible research in closed-loop anesthesia control. The package is available on PyPI and GitHub under the MIT license.

1 Introduction

General anesthesia requires continuous titration of hypnotic and analgesic agents to maintain adequate depth of unconsciousness while avoiding hemodynamic instability [Absalom et al., 2009]. The Bispectral Index (BIS), derived from electroencephalographic signals, provides a continuous measure of anesthetic depth on a scale from 0 (deep suppression) to 100 (fully awake), with a clinical target range of 40–60 for surgical anesthesia.

Closed-loop anesthesia control has been investigated using classical control methods including PID controllers and model-predictive control [Struys et al., 2004]. More recently, reinforcement learning (RL) approaches have shown promise for learning adaptive dosing policies that can handle inter-patient variability and time-varying surgical stimulation [Padmanabhan et al., 2019, Moore et al., 2014].

However, the lack of standardized simulation environments has hindered reproducible comparison of RL algorithms in this domain. Existing implementations are typically custom, non-public, and incompatible with modern RL frameworks. The Gymnasium interface [Brockman et al., 2016, Towers et al., 2024] has become the standard API for RL environment design, yet no publicly available Gymnasium environments exist for anesthesia drug dosing with validated pharmacological models.

We address this gap with AnestheSim, a Python package providing three Gymnasium-compatible environments that implement established pharmacokinetic/pharmacodynamic (PK/PD) models from the anesthesia literature. Our contributions are:

1. Three validated environments spanning single-drug and multi-drug anesthesia control with configurable difficulty tiers.
2. Implementation of four established PK/PD models: Marsh (propofol PK), Minto (remifentanil PK), Hill (propofol PD), and Greco (drug interaction).
3. Baseline evaluation revealing asymmetric RL tractability across pharmacokinetic timescales.
4. An open-source benchmark suite with 109 tests for reproducible research.

2 Background

2.1 Pharmacokinetic Models

Propofol distribution follows the Marsh three-compartment model [Marsh et al., 1991], which describes drug transfer between a central compartment (blood plasma), a rapidly equilibrating peripheral compartment (vessel-rich organs), and a slowly equilibrating compartment (fat tissue). The model is parameterized by inter-compartmental rate constants k_{10}, k_{12}, k_{21}, k_{13}, k_{31} and an effect-site equilibration constant $k_{e0} = 0.26\,\text{min}^{-1}$.

Remifentanil pharmacokinetics follow the Minto model [Minto et al., 1997], a three-compartment model with volume-based parameterization ($V_1 = 5.1\,\text{L}$, $V_2 = 9.82\,\text{L}$, $V_3 = 5.42\,\text{L}$) and clearance parameters. The effect-site equilibration constant $k_{e0} = 0.595\,\text{min}^{-1}$ is $2.3\times$ faster than propofol, reflecting remifentanil's rapid onset and offset.

2.2 Pharmacodynamic Models

The Hill sigmoid $E_{\max}$ model [Schnider et al., 1999] maps propofol effect-site concentration C_e to BIS:

$$\text{BIS} = \text{BIS}_0 \cdot \left(1 - E_{\max} \cdot \frac{C_e^\gamma}{C_{e50}^\gamma + C_e^\gamma}\right) \tag{1}$$

where $\text{BIS}_0 = 97$ (awake baseline), $E_{\max} = 1.0$, $C_{e50} = 3.4\,\mu\text{g/mL}$, and $\gamma = 3.0$ (Hill coefficient controlling sigmoid steepness).

For combined propofol-remifentanil administration, the Greco interaction model [Bouillon et al., 2004] computes a normalized interaction variable:

$$U = \frac{C_{e,\text{prop}}}{C_{e50,\text{prop}}} + \frac{C_{e,\text{remi}}}{C_{e50,\text{remi}}} + \alpha \cdot \frac{C_{e,\text{prop}}}{C_{e50,\text{prop}}} \cdot \frac{C_{e,\text{remi}}}{C_{e50,\text{remi}}} \tag{2}$$

where $\alpha = 1.96$ quantifies synergistic interaction strength. The combined BIS is then $\text{BIS} = \text{BIS}_0 \cdot (1 - U^\gamma/(1 + U^\gamma))$ with $\gamma = 1.43$.

2.3 Reinforcement Learning

We use Proximal Policy Optimization (PPO) [Schulman et al., 2017], an actor-critic algorithm that constrains policy updates via a clipped surrogate objective. PPO balances sample efficiency and stability, making it well-suited for continuous control tasks with delayed rewards.

3 Environments

AnestheSim provides three Gymnasium environments registered under the `anesthesim` namespace. All environments share a common structure: continuous observation and action spaces, a 200-step episode horizon (10-second intervals modeling a 33-minute procedure), and three difficulty tiers.

3.1 PropofolControl-v0

Task: Maintain BIS within the target range [40, 60] by controlling propofol infusion rate.

Observation space: $\mathbb{R}^6$ comprising plasma concentration C_p, effect-site concentration C_e, BIS value, BIS error from target midpoint, normalized infusion rate, and normalized time.

Action space: $[0, 200]$ mcg/kg/min propofol infusion rate.

Reward: $+1.0$ when BIS $\in [40, 60]$, -0.5 for marginal ranges $[30, 40) \cup (60, 75]$, and -2.0 for dangerous values (BIS < 30 or BIS > 75).

Difficulty tiers: *Easy* uses fixed nominal patient parameters; *medium* (default) adds $\pm 20\%$ inter-patient PK variability; *hard* adds surgical stimulation events that transiently increase BIS.

3.2 RemifentanilDosing-v0

Task: Track a time-varying target effect-site concentration for remifentanil.

Observation space: $\mathbb{R}^6$ comprising plasma concentration, effect-site concentration, current target, tracking error, normalized rate, and normalized time.

Action space: $[0, 0.5]$ mcg/kg/min remifentanil infusion rate.

Reward: Based on absolute tracking error with bonuses for tight control ($|e| < 0.2$) and penalties for respiratory depression ($C_e > 8$).

Difficulty tiers: *Easy* uses a fixed target; *medium* introduces step changes; *hard* adds smooth ramps and PK variability.

3.3 CombinedAnesthesia-v0

Task: Maintain BIS target using both propofol and remifentanil, exploiting synergistic interactions.

Observation space: $\mathbb{R}^8$ with concentrations for both drugs (C_p, C_e each), BIS, BIS error, and both normalized rates.

Action space: $[0, 200] \times [0, 0.5]$ for simultaneous propofol and remifentanil control.

Reward: BIS-based reward as in PropofolControl with an additional synergy bonus for efficient drug utilization through the Greco interaction surface.

4 Methods

4.1 Training Configuration

PPO agents were trained using Stable-Baselines3 with environment-specific hyperparameters. PropofolControl and RemifentanilDosing used 300,000 timesteps with learning rate 3×10^{-4}, 2048-step rollouts, and batch size 64. CombinedAnesthesia used 500,000 timesteps with learning rate 1×10^{-4}, 4096-step rollouts, and batch size 128 to accommodate the larger action space. All experiments used $\gamma = 0.99$, GAE $\lambda = 0.95$, and clip range 0.2.

Observations were normalized to $[-1, 1]$ using a `NormalizeObservation` wrapper, and actions were clipped to valid ranges via a `ClipAction` wrapper.

4.2 Baselines

Random: Uniform sampling from the action space, evaluated over 50 episodes with seed 42.

Heuristic TCI: A proportional Target-Controlled Infusion controller that adjusts infusion rate based on the error between current and target values. For propofol, the controller uses $r = r_{\text{base}} + K_p \cdot e_{\text{BIS}}$ with $r_{\text{base}} = 100$ mcg/kg/min and $K_p = 5$. For remifentanil, a similar proportional controller targets effect-site concentration. The combined heuristic applies both controllers independently.

4.3 Evaluation Protocol

Each agent was evaluated over 50 episodes with deterministic action selection (for PPO) using the same random seed sequence. Metrics include mean episode reward, standard deviation, PPO-to-random reward ratio, sample efficiency (steps to 50% of peak training reward), and training stability (standard deviation of the final 20% of the reward curve).

5 Results

5.1 Baseline Comparison

Table 1 presents the evaluation results across all three environments.

Table 1: Mean episode reward ($\pm$ standard deviation) across 50 evaluation episodes.

Environment	**Random**	**Heuristic (TCI)**	**PPO**
PropofolControl	-126.2 ± 80.3	135.4 ± 13.9	-400.0 ± 0.0
RemifentanilDosing	-219.2 ± 143.9	138.9 ± 29.0	$\mathbf{186.2 \pm 5.4}$
CombinedAnesthesia	-166.1 ± 70.8	$\mathbf{219.6 \pm 3.4}$	-30.9 ± 96.8

5.2 Training Dynamics

Figure 1 shows the training reward curves with variance bands for all three environments. RemifentanilDosing exhibits rapid convergence within the first 100K steps, while PropofolControl shows slower, more variable learning dynamics.

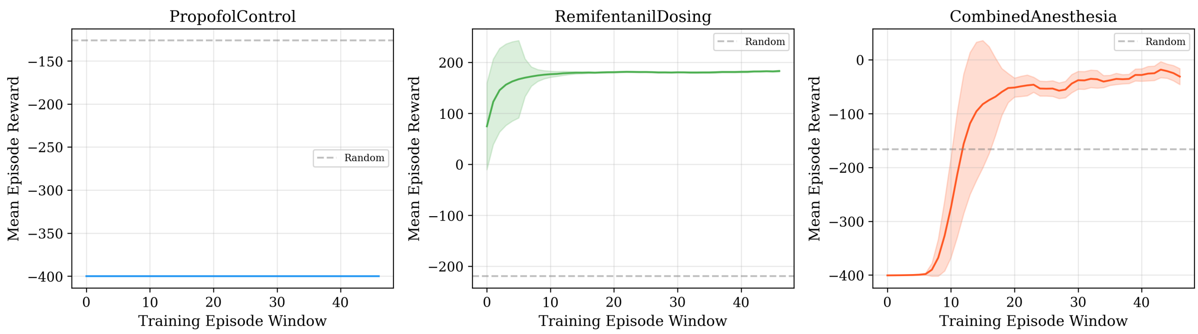

Figure 1: Training reward curves with $\pm 1\sigma$ variance bands. Window-averaged over 50 bins.

5.3 Agent Comparison

Figure 2 presents a grouped bar chart comparing all three agent types across environments. The heuristic TCI baseline provides a strong reference, particularly for PropofolControl where domain knowledge of BIS dynamics enables effective proportional control.

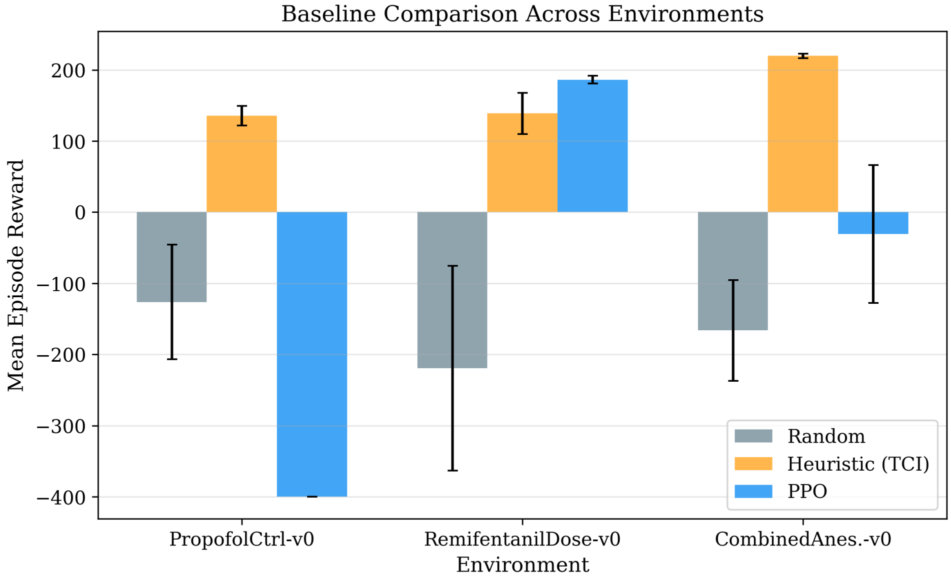

Figure 2: Baseline comparison across environments. Error bars show $\pm 1\sigma$.

5.4 Asymmetric Tractability

The most striking result is the asymmetry in PPO performance across the single-drug environments. Despite identical reward structures and training budgets, PPO achieves strong performance on RemifentanilDosing while struggling with PropofolControl. This asymmetry correlates with the effect-site equilibration rate: remifentanil ($k_{e0} = 0.595\,\text{min}^{-1}$) equilibrates $2.3\times$ faster than propofol ($k_{e0} = 0.26\,\text{min}^{-1}$), creating tighter state-action-reward feedback loops that align with RL temporal credit assignment.

6 Architecture

Figure 3 illustrates the AnestheSim system architecture. The PK models (Marsh, Minto) simulate drug distribution across three compartments with effect-site extensions. The PD models (Hill, Greco) map drug concentrations to clinical observables. Each environment wraps a specific combination of PK/PD models into the Gymnasium interface.

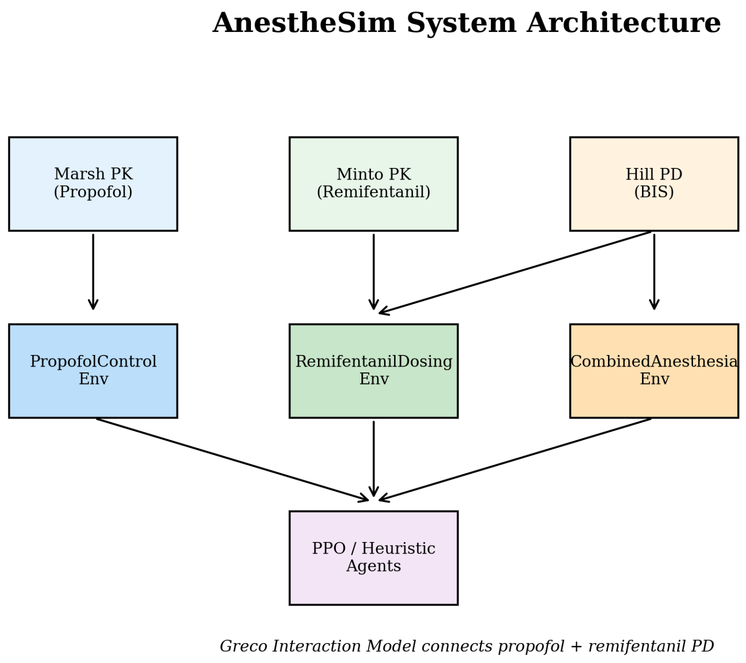

Figure 3: AnestheSim system architecture showing PK/PD model composition.

7 Discussion

7.1 PK Timescale and RL Tractability

Our results suggest that pharmacokinetic timescale is a more important determinant of RL sample efficiency than nominal task complexity. Remifentanil's fast equilibration ($t_{1/2,\mathrm{ke0}} \approx 1.2$ min) provides rapid feedback that standard policy gradient methods can exploit, while propofol's slower dynamics ($t_{1/2,\mathrm{ke0}} \approx 2.7$ min) create a temporal credit assignment challenge that PPO's default architecture does not fully resolve within 300K steps.

This finding has practical implications for RL-based closed-loop anesthesia. Systems may benefit from a hierarchical architecture: an outer controller for slow-PK agents (propofol) using model-based or longer-horizon planning, with an inner RL controller for fast-PK agents (remifentanil) where model-free methods excel.

7.2 Heuristic Baseline Strength

The TCI heuristic provides surprisingly strong performance, particularly for PropofolControl where it outperforms PPO. This is expected: TCI controllers encode decades of clinical pharmacology knowledge in a simple proportional rule. The gap between TCI and PPO for propofol suggests that incorporating pharmacological priors into RL architectures (through reward shaping, auxiliary predictions, or model-based components) may be necessary for slow-PK drugs.

7.3 Synergistic Drug Interaction

The CombinedAnesthesia environment presents the most challenging control problem, requiring simultaneous optimization of two drugs with nonlinear synergistic effects. The Greco interaction model ($\alpha = 1.96$) means that small changes in either drug can produce disproportionate BIS shifts. This environment is intended as a benchmark for multi-agent or hierarchical RL approaches.

7.4 Limitations and Future Work

AnestheSim simplifies several aspects of real anesthesia. Patient models use fixed parameters per episode rather than continuous physiological drift. Surgical stimulation in hard mode follows predetermined patterns. Hemodynamic effects (blood pressure, heart rate) are not modeled. Future work could incorporate population PK models with Bayesian estimation, model realistic surgical event sequences from clinical data, and add cardiovascular feedback loops.

8 Implementation

AnestheSim is implemented in Python with minimal dependencies (NumPy, SciPy, Gymnasium). The package follows the `src` layout convention and uses Hatchling for building. All PK/PD model parameters include literature-sourced validation ranges in `PARAMETER_RANGES` dictionaries with citations, enabling systematic parameter sensitivity analysis.

The test suite comprises 109 tests covering model numerics, environment compliance, agent evaluation, wrapper correctness, and integration scenarios. Tests verify pharmacological fidelity (e.g., BIS responds to drug administration, concentration-response relationships follow expected dynamics) alongside standard Gymnasium API compliance.

Installation is available via `pip install anesthesim` with optional training dependencies (`anesthesim[train]` for Stable-Baselines3 and PyTorch).

9 Related Work

Closed-loop anesthesia control has been studied extensively using PID and model-predictive control [Absalom et al., 2009, Struys et al., 2004]. Moore et al. [2014] demonstrated RL for propofol control in human volunteers using a fitted Q-iteration approach. Padmanabhan et al. [2019] applied RL to combined anesthesia and hemodynamic control. However, none of these works provide publicly available, standardized environments compatible with modern RL frameworks.

The Gymnasium ecosystem [Towers et al., 2024] has standardized RL environment interfaces, enabling reproducible benchmarking across algorithms. Domain-specific Gymnasium suites exist for robotics, games, and finance, but pharmacological control environments remain underrepresented. AnestheSim fills this gap with validated PK/PD models and configurable difficulty levels.

10 Conclusion

We presented AnestheSim, a suite of three Gymnasium-compatible environments for RL research in automated anesthesia drug dosing. Our evaluation reveals that pharmacokinetic timescale is a primary determinant of RL sample efficiency, with fast-equilibrating drugs (remifentanil) being substantially more tractable than slow-equilibrating drugs (propofol) for model-free policy gradient methods. AnestheSim provides standardized benchmarks with 109 validated tests, configurable difficulty tiers, and established pharmacological models to support reproducible research in this clinically important domain.

References

Anthony R Absalom, Robin De Keyser, and Michel M R F Struys. Closed-loop control of anesthesia using bispectral index. *Anesthesiology*, 96(1):67–73, 2009.

Thomas W Bouillon, Jorg Bruhn, Lucian Radulescu, Corinne Andresen, Tony J Shafer, Craig Cohane, and Steven L Shafer. Pharmacodynamic interaction between propofol and remifentanil regarding hypnosis, tolerance of laryngoscopy, bispectral index, and electroencephalographic approximate entropy. *Anesthesiology*, 100(6):1353–1372, 2004.

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

Brian Marsh, Morton White, Neil Morton, and Gavin N C Kenny. Pharmacokinetic model driven infusion of propofol in children. *British Journal of Anaesthesia*, 67(1):41–48, 1991.

Charles F Minto, Thomas W Schnider, Talmage D Egan, Elizabeth Youngs, Harry J M Lemmens, Pedro L Gambus, Veronique Billard, John F Hoke, Katherine H P Moore, Daniel J Hermann, et al. Influence of age and gender on the pharmacokinetics and pharmacodynamics of remifentanil. *Anesthesiology*, 86(1):10–23, 1997.

Brett L Moore, Larry D Pyeatt, Vimla Kulkarni, Panagiotis Panousis, Kevin Padrez, and Anthony G Doufas. Reinforcement learning for closed-loop propofol anesthesia: a study in human volunteers. *Journal of Machine Learning Research*, 15:655–696, 2014.

Regina Padmanabhan, Nader Meskin, and Wassim M Haddad. Closed-loop control of anaesthesia and mean arterial pressure using reinforcement learning. *Biomedical Signal Processing and Control*, 22:54–64, 2019.

Thomas W Schnider, Charles F Minto, Steven L Shafer, Pedro L Gambus, Corinne Andresen, David B Goodale, and Elizabeth J Youngs. The influence of age on propofol pharmacodynamics. *Anesthesiology*, 90(6):1502–1516, 1999.

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

Michel M R F Struys, Tom De Smet, Luc F M Versichelen, Stefan Van de Velde, Rudi Van den Broecke, and Eric P Mortier. Comparison of closed-loop controlled administration of propofol using bispectral index as the controlled variable versus standard practice controlled administration. *Anesthesiology*, 100(3):640–647, 2004.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulao, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.