Composable Parameter-Level Reinforcement Learning for Dual-Chamber Pacemaker Programming

Leqi Jia LEQIJIA@GMAIL.COM
The Harker School

Abstract

Unnecessary right ventricular pacing (VP) in dual-chamber (DDD) pacemakers is associated with pacing-induced cardiomyopathy, heart failure, and atrial fibrillation. Commercial VP-minimization algorithms such as Medtronic's Managed Ventricular Pacing (MVP) operate at the beat level and sharply reduce VP burden in patients with intact atrioventricular (AV) conduction (e.g., sick sinus syndrome) but are limited in patients with intermittent AV block, where the algorithm cannot give conduction more time to complete. We propose a hierarchical reinforcement learning framework that operates at the *parameter level* (AV delay and activity-adaptive target rate) and composes cleanly on top of any certified beat-level controller. Trained via proximal policy optimization (PPO) on a clinically parameterized event-level surrogate, the RL policy reduces VP burden from 91.8–95.2% to 5.9–7.8% on Mobitz Type I and Type II when layered on top of a standard dual-chamber rate-adaptive (DDDR) controller, and when layered on top of a Casavant-2021-specified MVP controller further reduces VP burden from 3.4% to 0.4% on first-degree AV block and from 48.6% to 9.2% on Mobitz Type II, while preserving MVP's near-zero VP (0.2% to 0.1%) on sick sinus syndrome. The MVP implementation is specified against the Casavant–Belk 2021 technical reference and calibrated per-subgroup against the IDEAL RVP and COMPARE trials. The parameter-level and beat-level optimizations act on separable timing primitives and the same policy transfers across both without retraining, suggesting a composable deployment path in which a learned parameter-programming layer sits on top of, rather than replaces, certified commercial firmware.

1. Introduction

Approximately one million pacemakers are implanted worldwide each year (Mulpuru et al., 2017; Kusumoto et al., 2019), with dual-chamber (DDD) devices representing the standard of care for patients requiring both atrial and ventricular support. While DDD pacing restores adequate heart rate, it carries a well-documented risk: unnecessary right ventricular pacing (VP) promotes electromechanical dyssynchrony, which over time can lead to pacing-induced cardiomyopathy (PICM), heart failure, and atrial fibrillation (Lamas et al., 2002; Wilkoff et al., 2002). The landmark MOST trial (Lamas et al., 2002) and its substudy by Sweeney et al. (2003) established VP burden as an independent predictor of heart failure hospitalization, with a cumulative VP burden above 40% in the DDDR group associated with a nearly three-fold increase in HF hospitalization hazard (HR ≈ 3.0). A later meta-analysis by Somma et al. (2023) adopted a lower VP burden threshold of $\geq 20\%$ as a diagnostic criterion for pacing-induced cardiomyopathy, synthesizing evidence across multi-

ple observational cohorts; we refer to the MOST-derived 40% threshold as the trial-validated hazard inflection and the 20% threshold as the diagnostic/observational lower bound.

In response, device manufacturers have developed VP-minimization algorithms such as Medtronic's MVP (Casavant and Belk, 2021; Sweeney et al., 2007), which reduced median VP from 99.0% to 9.1% in the SAVe PACe trial (sinus-node-dysfunction patients, 2°/3° AV block excluded) (Sweeney et al., 2007). However, these algorithms operate as fixed heuristic rules that manipulate a single parameter and do not adapt to individual conduction characteristics.

Reinforcement learning (RL) offers a principled framework for sequential decision-making in healthcare (Komorowski et al., 2018; Yu et al., 2021; Gottesman et al., 2019). Recent work has applied RL to pacemaker control at the beat level (Dole et al., 2023; Komp et al., 2024) but requires replacing the device's certified firmware, is difficult to verify clinically, and operates at a level clinicians do not directly interact with.

We propose a hierarchical RL framework that operates at the *parameter level*, the same abstraction at which clinicians program devices. A high-level PPO agent (Schulman et al., 2017) adjusts the target heart rate and AV delay at one-second intervals; the existing DDD controller executes beat-level pacing. This separation (i) keeps the certified controller's safety logic in place, (ii) makes the RL layer modular across any beat-level algorithm, and (iii) keeps learned parameters directly interpretable. We train on Mobitz Type I/II and evaluate zero-shot on first-degree AV block and sick sinus syndrome (SSS). Timing-level consistency is cross-checked against openCARP (Plank et al., 2021) (Section 4.1).

1.1. Generalizable Insights about Machine Learning in the Context of Healthcare

This work demonstrates three generalizable principles. First, hierarchical control, separating learned optimization from safety-critical execution, enables RL evaluation for medical devices without replacing certified firmware. Second, RL can learn state-dependent programming policies that refine a single clinical lever more effectively than the fixed heuristics commercial algorithms use, producing condition-specific behavior without hand-engineering per-condition rules. Third, the hierarchical architecture is composable: parameter-level RL can be layered on top of existing beat-level algorithms, with each layer addressing a different axis of optimization (Kusumoto et al., 2019).

2. Related Work

2.1. VP-Minimization Algorithms

Commercial VP-minimization strategies fall into two categories: AV delay extension (e.g., Abbott's VIP and AICS algorithms) and mode-switching (e.g., Medtronic's MVP (Casavant and Belk, 2021), Boston Scientific's RHYTHMIQ, MicroPort's SafeR). MVP operates in AAI mode by default and switches to DDD upon detecting two of four consecutive dropped beats, periodically testing for return of AV conduction with exponential backoff. The SAVe PACe trial, restricted to sinus-node-dysfunction patients with 1:1 AV conduction (2°/3° AV block excluded), showed that MVP combined with Search-AV-Extension reduced median VP burden from 99.0% to 9.1% and persistent atrial fibrillation by 40% (Sweeney

et al., 2007). Because SAVe PACe's intervention arm was MVP + Search-AV combined, it does not provide a direct per-subgroup calibration for MVP alone; this is instead supplied by IDEAL RVP (MVP median %VP by AV-block subgroup: 0.2% no-AVB, 2.3% 1st-degree, 16.4% 2nd-degree, 37.7% episodic 3rd-degree) (Murakami et al., 2010), and the COMPARE trial (MVP 12-month median %VP of 11.8% in AVB patients (Chen et al., 2014)). The ANSWER trial reported the SafeR algorithm at 11.5% overall versus DDD at 93.6% over 3 years (Stockburger et al., 2016).

Subsequent evidence has identified a subgroup in which MVP-style strategies are *not* uniformly beneficial: patients with a baseline PR interval greater than 230 ms. PreFER MVP reported a 3.4-fold increased risk of persistent atrial tachyarrhythmia/atrial fibrillation in this subgroup on the MVP arm (HR 3.41, 95% CI 1.10–10.6, $p = 0.024$) (Ricci et al., 2015), and the MVP Trial, which compared managed ventricular pacing against VVI-40 backup pacing in ICD recipients, reported worse outcomes in the same subgroup on the MVP arm (Sweeney et al., 2010). The proposed mechanism is AV decoupling: when the AAI window permits very long PR, left atrial contraction can fall against a closed mitral valve, raising left atrial pressure and promoting AT/AF. This evidence does not reverse the general case for VP-minimization; it refines the indication by identifying long-PR patients as a population in which the benefit/risk of aggressive AAI-favoring strategies is less clear. We return to this point when interpreting the RL+MVP composability result on 1st-degree AV block (Section 5.3).

These algorithms share a common design limitation: they manipulate AV timing as a single parameter using fixed rules and do not jointly optimize across multiple pacing parameters.

2.2. RL for Cardiac Device Control

Dole et al. (2023) introduced correct-by-construction RL for dual-chamber pacemaker synthesis, using duration-calculus specifications to guarantee formal safety properties. Komp et al. (2024) extended this with reward machines learned from demonstrations. Both approaches operate at the beat level, deciding individual pace/sense actions at each time step and replacing the certified beat-level controller. Jiang et al. (2012) developed a cyber-physical modeling framework for implantable cardiac devices using UPPAAL timed automata, enabling formal verification of pacemaker timing properties. Our work addresses a complementary problem: rather than synthesizing a new beat-level controller with its own safety guarantees, we retain the certified beat-level controller and add a learned outer loop that modifies programmed parameters within clinically admissible bounds. This separation (beat-level firmware preserved, parameter-level optimization learned) targets a deployment path (non-invasive parameter advisor layered on top of existing devices) that beat-level RL does not currently support. Because the two approaches act on different levels of the control stack and on different action spaces, a direct head-to-head comparison is not well defined, and we position them as complementary rather than competing. More broadly, RL has been applied to treatment optimization across clinical domains (Komorowski et al., 2018; Yu et al., 2021), and guidelines for responsible RL deployment in healthcare, emphasizing safety constraints, interpretability, and simulation-based evaluation before clinical deployment (Gottesman et al., 2019), directly inform our approach.

3. Methods

3.1. Surrogate Heart Simulator

We develop an event-driven stochastic heart simulator that models the beat-to-beat timing interactions between the sinoatrial (SA) node, atrioventricular (AV) conduction system, and a DDD pacemaker. The simulator operates at 10 ms time steps (100 Hz) and models four cardiac events: atrial sense, atrial pace, ventricular sense, and ventricular pace.

Intrinsic atrial depolarization follows a Gaussian process with patient-specific mean rate ($\mu_{\text{SA}} \in [40, 80]$ bpm for the bradycardia population) and heart rate variability ($\sigma_{\text{HRV}} \in [30, 100]$ ms). For sick sinus syndrome, the model additionally incorporates stochastic SA pauses with configurable probability (10–30% per beat) and duration (1.5–5.0 s). The AV conduction system implements four pathologies relevant to VP-minimization. Normal and first-degree conduction succeed with PR interval drawn from $\mathcal{N}(\mu_{\text{PR}}, \sigma_{\text{PR}})$, where μ_{PR} ranges from 140–220 ms (normal) to 224–352 ms (first-degree). Mobitz Type I (Wenckebach) progressively lengthens the PR interval over 3–6 beat cycles (increment 10–35 ms/beat) until conduction fails, then resets. Mobitz Type II produces independent conduction failure with probability 10–40% per beat, with constant PR when conduction succeeds.

The low-level DDD controller implements standard dual-chamber timing cycles including lower rate interval, programmable AV delay (80–300 ms), atrial and ventricular refractory periods (200 ms and 250 ms respectively), post-pace blanking (40 ms), and upper tracking rate (120 bpm). Atrial capture probability is 99.5% with 0.2% transient loss-of-capture rate, consistent with chronic pacing thresholds (Kusumoto et al., 2019; Ellenbogen and Wood, 2005). Each episode generates a three-phase activity schedule (rest → exercise → recovery) with randomized transition times and intensity levels, producing a rate-adaptive target: $\text{target}_{\text{bpm}} = \text{rest}_{\text{bpm}} + (\text{max}_{\text{bpm}} - \text{rest}_{\text{bpm}}) \times \text{activity}$.

3.2. MVP Controller Specification

Our MVP baseline implements the state machine described in Casavant and Belk (2021). By default the controller operates in AAI(R) mode: the atrium is paced on demand at the target rate, the ventricle is sensed but not paced, and AV conduction is allowed to complete intrinsically. A switch to DDD(R) mode is triggered when a *2-of-4 conduction failure* is detected: the controller maintains a sliding window of the four most recent atrial events and transitions to DDD if any two of the corresponding atrial-to-ventricular intervals complete without a sensed ventricular event. On the triggering cycle a backup V-pace is delivered 80 ms after the non-conducted atrial event; the same 80 ms AV delay is retained throughout subsequent DDD(R) operation, a simplification of Casavant and Belk (2021), which specifies 80 ms for the backup pace on the triggering cycle.

While in DDD, the controller schedules periodic AAI-retest beats with a doubling-backoff schedule: retests occur 1, 2, 4, 8, 16, 32, 64, 128, 256, and 512 minutes after the most recent DDD entry, and then every 16 hours (the documented plateau) thereafter. A retest beat withholds the scheduled V-pace for one cycle; if a sensed V-event arrives within a 300 ms AAI conduction window, the controller switches back to AAI(R), otherwise it resumes DDD and the backoff counter advances. Standard refractory periods and post-pace blanking remain in effect throughout. Table 1 gives the full parameter set.

Table 1: MVP controller parameters. The 2-of-4 criterion, 80 ms backup pace, and retest backoff follow Casavant and Belk, 2021; the 300 ms AAI conduction window and the retained DDD AV delay are our own specification.

Parameter	**Value**
Default mode	AAI(R)
AAI conduction window	300 ms after sensed/paced A
Drop-detection criterion	2 of 4 rolling A-A intervals
Fallback V-pace timing (on switch)	80 ms after triggering A
DDD AV delay (after switch)	80 ms
AAI retest backoff	1, 2, 4, 8, ... min, doubling, plateau at 16 h
Retest decision window	300 ms after retest-beat A
Refractory periods (ARP / VRP)	200 ms / 250 ms
Post-pace blanking	40 ms
Upper tracking rate	120 bpm

The implementation is verified against the behavior described in Casavant and Belk (2021) via unit tests in the repository that check (i) that AAI is maintained indefinitely under 1:1 conduction, (ii) that two non-conducted atrial events in a 4-beat window trigger a DDD switch, (iii) that a single non-conducted beat surrounded by conducted beats does not trigger, (iv) that the backup V-pace is delivered at 80 ms after the triggering atrial event, and (v) that the doubling backoff schedule resets correctly after a successful retest.

The simulator's Mobitz MVP figures (22.8–48.6%) reflect a wider drop-probability sampling range (10–40% per beat) than the clinical 2AVB populations reported in IDEAL RVP (median 16.4% Murakami et al., 2010) and COMPARE (AVB median 11.8% Chen et al., 2014); the per-subgroup calibration is discussed in Section 5.2.

3.3. Hierarchical RL Architecture

Our framework separates pacemaker control into two temporal scales. Every one second (100 simulator steps), the high-level RL agent observes the current cardiac state and selects an action from a discrete set of parameter adjustments. The action space is a Cartesian product of AV-delay adjustments $\Delta_{\text{AV}} \in \{-100, -50, 0, +50, +100\}$ ms relative to the base AV delay of 200 ms, and target-rate adjustments $\Delta_{\text{bpm}} \in \{-20, -15, 0, +15, +30\}$ bpm relative to the activity-driven baseline, yielding $|\mathcal{A}| = 5 \times 5 = 25$ discrete actions. The AV-delay axis is the primary therapeutic lever; the rate axis is included to let the policy accommodate activity transitions within the hard safety envelope described below. Invalid actions (those that would set the target below 40 bpm or above the upper tracking rate) are masked before policy sampling (Huang and Ontañón, 2022). The one-second interval is a simulation and training convention rather than a proposed device-reprogramming frequency. The staged deployment path described in Section 6.1, which moves from an off-device advisor to a human-in-the-loop programmer and eventually to closed-loop operation under a change-control plan, does not reprogram the device every second. In practice the selected action is held across multi-second stretches and changes on fewer than 10% of decision steps,

settling the AV-delay axis at a condition-dependent level rather than fluctuating step to step (Appendix M).

The high-level agent receives a 13-dimensional observation vector at each decision step:

$$\mathbf{o}_t = [\,\text{activity},\ \text{bpm}_{\text{base}},\ \text{bpm}_{\text{current}},\ \text{AV}_{\text{current}},\ \Delta t_V,\ \Delta t_A,\ \text{V}_{\text{sense}},\ \text{A}_{\text{sense}},\ \text{wait},\ \text{AV}_{\text{rate}},\ \text{VP}_b,\ \text{MAE},\ \text{unsafe}\,] \tag{1}$$

where Δt_V and Δt_A are normalized time since last ventricular and atrial events, AV_{rate} is the recent AV conduction success rate, and VP_b is the interval VP burden. Between high-level decisions, the low-level DDD controller executes beat-level actions (none, atrial pace, ventricular pace) according to standard timing cycles with the parameters set by the high-level policy.

The high-level reward at each decision step balances three objectives:

$$r_t = r_{\text{accuracy}} - \lambda_{\text{VP}} \cdot \text{VP}_{\text{burden}} - \lambda_{\text{safety}} \cdot n_{\text{unsafe}} \tag{2}$$

where r_{accuracy} rewards VV intervals close to the target, $\lambda_{\text{VP}} = 1.2$ penalizes ventricular pacing, and $\lambda_{\text{safety}} = 10.0$ penalizes unsafe VV intervals (tachycardia <500 ms or pauses >2000 ms). An additional floor penalty discourages setting the target rate more than 20 bpm below the activity-driven baseline.

Hard safety envelope. After each action, target rate is clamped to $[\,r_{\min}(\text{activity}),\ r_{\max}\,]$ ($r_{\min} = 50$ bpm at rest, rising linearly to 70 bpm at full exercise; $r_{\max} = 120$ bpm), and the AV delay to $[100, 400]$ ms. Bounds follow pacemaker programming guidelines (Ellenbogen and Wood, 2005; Kusumoto et al., 2019). The clamp fires on at most 2.7% of decision steps; disabling it changes primary metrics by <1 pp (Appendix E).

3.4. Policy Architecture and Training

We use a recurrent actor-critic architecture with separate LSTM networks (Hochreiter and Schmidhuber, 1997) for the actor and critic. Each branch consists of a single-layer LSTM (hidden size 64), layer normalization, and a two-layer MLP head ($64 \to 32 \to$ output) with Tanh activation and orthogonal weight initialization. The agent is trained using PPO (Schulman et al., 2017) with generalized advantage estimation (Schulman et al., 2015) over 600 updates of 1,800 steps each (1,080,000 total environment interactions), with training episodes of 45 s duration. Training curves showing episodic reward and VP burden over the 600 updates are reported in Appendix L. The training environment samples uniformly between Mobitz Type I and Type II patients with domain randomization over patient parameters. Key hyperparameters are: learning rate 3×10^{-4}, clipping ratio 0.2, entropy coefficient 0.04 (annealed to 0.012), discount $\gamma = 0.99$, GAE $\lambda = 0.95$, minibatch size 32, 4 epochs per update.

3.5. Baseline Controllers

We compare the RL policy against two reference controllers. **Baseline DDDR** uses a fixed 200 ms AV delay, the trial-era reference against which SAVe PACe (Sweeney et al., 2007) and ANSWER (Stockburger et al., 2016) report 93–99% VP, placing our DDDR cohort values (87.3–95.2%) in the published range. **Casavant-2021 MVP** (Section 3.2)

implements the Casavant–Belk 2021 reference: AAI(R) by default, DDD switch on a 2-of-4 conduction failure, and periodic conduction-check retests with exponential backoff. Both hold $\Delta_{\text{bpm}} = 0$ and differ only in AV-timing behavior.

3.6. Evaluation Metrics

We evaluate controllers using two established clinical endpoints and three simulation-derived metrics that quantify clinically relevant behaviors. *VP burden* quantifies the percentage of ventricular beats that are paced rather than intrinsically conducted, the primary clinical endpoint for VP-minimization trials; >40% has a trial-validated association with heart-failure hospitalization hazard in the DDDR arm of MOST (Sweeney et al., 2003), while ≥20% is the observational threshold subsequently adopted as a diagnostic criterion for pacing-induced cardiomyopathy (Somma et al., 2023). *AV synchrony* captures the percentage of ventricular events preceded by an atrial event within the AV delay window, consistent with the definition used by Ip et al. (2024). *Intrinsic conduction rate* measures the percentage of ventricular events arising from intrinsic conduction rather than from pacing. On the shared denominator of ventricular events it is the exact complement of VP burden, so it restates the primary endpoint rather than measuring conduction independently; we report it because device trials express preserved conduction as a conduction ratio or percent intrinsic rhythm (Sweeney et al., 2007) rather than as a paced-beat fraction. *Event-free rate* reports the percentage of episodes with zero adverse rhythm events (tachycardia <500 ms or pauses >2000 ms). *Rate response* (r) measures the Pearson correlation between activity level and achieved heart rate, quantifying the adequacy of chronotropic response (Brubaker and Kitzman, 2011). We additionally report absolute counts of *tachycardia events* and *pauses*.

Statistical testing. Continuous per-patient metrics (VP burden, intrinsic conduction rate, AV synchrony, event-free rate, rate response r) are compared between controllers using paired Wilcoxon signed-rank tests across the 30-patient cohort, since each patient is evaluated under every controller with matched seeds. Count metrics (tachycardia events, pauses) are compared using a paired Wilcoxon signed-rank test on per-patient counts. Formal test statistics are reported for VP burden, the primary endpoint (Appendix F); the remaining metrics are reported descriptively. Tables 3 and 9 report VP burden as mean ± standard deviation; the remaining tables report point estimates.

4. Cohort: Simulated Patient Population

Because our simulator operates on discrete cardiac events rather than raw ECG waveforms, we validate the surrogate model by ensuring that all configurable parameters fall within published clinical ranges. Table 2 presents each simulator parameter alongside its clinical reference value and source. All parameters fall within or near published ranges for the pacemaker-indicated population.

As this study uses only simulated data with no human subjects involvement, no institutional review board approval was required.

Each episode samples a new patient with parameters drawn from uniform distributions over the validated ranges. Uniform distributions are standard for domain randomization in

Table 2: Surrogate model parameter validation against published clinical literature.

Parameter	Our Range	Clinical Reference	Source
SA node rate	40–80 bpm	Bradycardia: <60 bpm	Kusumoto et al. (2019)
HRV (SDNN)	30–100 ms	Compromised: 50–100 ms	Shaffer and Ginsberg (2017)
PR interval	140–220 ms	Normal: 120–200 ms	Kusumoto et al. (2019)
PR variability	10–30 ms	Physiological: 10–40 ms	Ellenbogen and Wood (2005)
Wenckebach cycle	3–6 beats	Clinical: 3–5 beats typical	Ellenbogen and Wood (2005)
Mobitz II drop prob.	10–40%	Ratios: 2:1–5:1	Mangi et al. (2023)
Escape rate	25–45 bpm	Ventricular: 20–40 bpm	Ellenbogen and Wood (2005)
Atrial ERP	200 ms	Clinical: 160–260 ms	Ellenbogen and Wood (2005)
Ventricular ERP	250 ms	Programmable: 220–350 ms	Ellenbogen and Wood (2005)
Post-pace blanking	40 ms	Standard: 30–40 ms	Ellenbogen and Wood (2005)
Capture probability	99.5%	Clinical: >99%	Kusumoto et al. (2019)
Loss of capture	0.2%	Transient: 0.1–1%	Ellenbogen and Wood (2005)

RL and ensure robustness across the full parameter space, though they may over-represent extreme parameter combinations relative to clinical populations. All evaluations use a fixed cohort of 30 patients per condition (seed base 3000), with 60-second episodes containing a rest–exercise–recovery activity profile. This evaluation seed base is disjoint from the training seeds 1–5, so evaluation patients do not appear in training. Each episode independently samples the activity profile's transition times and intensities, and matched seeds ensure that the four controllers are compared on identical patient draws and identical activity profiles, making every controller comparison paired rather than confounded by differing conditions. A 60-second episode at a target rate of 55–75 bpm contains roughly 55–75 ventricular events per patient, so per-patient VP burden is quantized at approximately 1.5 pp; aggregate per-condition VP burdens are reported to one decimal place where the 30-patient cohort mean resolves this precision, and otherwise rounded to the nearest percentage point. We additionally confirm stability over 300-second episodes (Appendix J).

4.1. Implementation Cross-Check Against openCARP

As an implementation cross-check, we run a cohort of $N = 20$ virtual patients in openCARP (Plank et al., 2021), parameterized to the same patient-specific draws as the surrogate. openCARP is configured with a two-compartment monodomain tissue model (Courtemanche–

Ramirez–Nattel atrial, ten Tusscher–Panfilov ventricular) connected by an AV-nodal delay compartment, on a template biatrial + biventricular mesh with pacing stimuli at the RA appendage and RV apex. Mean PR intervals agree within 4 ms, VP burden within 2.6 percentage points, and MVP mode-switch counts within 1 event per 60 s (Appendix I). The stronger external anchor for the MVP implementation itself is the per-subgroup calibration against IDEAL RVP and COMPARE (Section 5.2).

We evaluate on four arrhythmia conditions. Mobitz Type II (trained) produces intermittent conduction failure with 10–40% drop probability per beat. Mobitz Type I (trained) exhibits progressive PR prolongation over 3–6 beat cycles before a dropped beat. First-degree AV block (zero-shot, untrained) features prolonged but intact conduction with PR intervals of 224–352 ms (1.6× normal). Sick sinus syndrome (zero-shot, untrained) presents normal AV conduction but intermittent SA node pauses (10–30% probability, 1.5–5.0 s duration) with a slow baseline rate of 35–55 bpm.

5. Results

Each controller is evaluated on 30 randomized patients per condition (120 episodes total, 60 s each). We compare baseline DDDR, a Casavant-inspired MVP surrogate with an author-defined conduction-detection window (Section 3.2), RL+DDDR, and RL+MVP. The RL checkpoint is trained on a mixed Mobitz I/II cohort with the standard DDD controller; RL+MVP rows represent a zero-shot composability-transfer result.

5.1. Compositional Comparison Across Controllers

Table 3 reports VP burden across the four arrhythmia conditions for all four controllers. Three patterns are consistent with published clinical trials. First, DDDR baselines in our simulator (87.3–95.2% across conditions) reproduce the range reported in ANSWER (93.6% overall, 84.7% in SND, 98.2% in intermittent AVB) and SAVe PACe (99.0% in SSS) (Sweeney et al., 2007; Stockburger et al., 2016). Second, MVP tracks the per-subgroup behavior of the IDEAL RVP trial (Murakami et al., 2010): 0.2% VP on SSS (IDEAL RVP no-AVB: 0.2%), 3.4% on first-degree (IDEAL RVP 1AVB: 2.3%), and 22.8–48.6% on Mobitz (both type I and type II are 2AVB; IDEAL RVP 2AVB pooled median: 16.4%), with the bimodal 1st-degree distribution reflecting patients whose PR interval exceeds the 300 ms AAI conduction window and triggers the 2-of-4 drop switch to DDD. Third, the RL policy on top of standard DDD reduces VP burden into the clinically target range (<10%) on every condition (Mobitz II 95.2% → 7.8%, −87.4 pp; Mobitz I 91.8% → 5.9%, −85.9 pp; 1st-degree 87.3% → 4.5%; SSS 93.6% → 3.1%), while also lowering tachycardia event counts by 3–6× (Mobitz II 12 → 2; Mobitz I 18 → 6), confirming that joint rate-and-AV-delay parameter optimization adds value even on top of an uncompounded DDD controller.

5.2. MVP Per-Subgroup Calibration Against IDEAL RVP

Our primary calibration target is IDEAL RVP (Murakami et al., 2010), which reports MVP median %VP by AV-block subgroup (SAVe PACe's 9.1% reflects MVP + Search-AV combined and excludes 2°/3° AV block, making it unsuitable as a pure-MVP reference (Sweeney et al., 2007)). On SSS our MVP reports 0.2% VP, matching IDEAL RVP no-AVB

Table 3: Compositional comparison: VP burden (mean ± SD, %), intrinsic conduction rate (ICR, %), event-free rate (EFR, %), and tachycardia event count across 30 patients per condition, 60 s episodes. Bold marks the best value on each row.

Condition	Metric	DDDR	MVP	RL+DDDR	RL+MVP
Mobitz II	VP (%) ↓	95.2±4.8	48.6±21.3	**7.8**±4.2	9.2±5.3
	ICR (%) ↑	4.8	51.4	**92.2**	90.8
	EFR (%) ↑	76.7	70.0	**96.7**	**96.7**
	Tachycardia ↓	12	48	2	**1**
Mobitz I	VP (%) ↓	91.8±5.4	22.8±17.4	**5.9**±3.3	6.8±3.8
	ICR (%) ↑	8.2	77.2	**94.1**	93.2
	EFR (%) ↑	80.0	63.3	**96.7**	**96.7**
	Tachycardia ↓	18	73	6	**3**
1st Degree	VP (%) ↓	87.3±9.2	3.4±5.6	4.5±2.8	**0.4**±0.9
	ICR (%) ↑	12.7	96.6	95.5	**99.6**
	EFR (%) ↑	86.7	93.3	**96.7**	**96.7**
	Tachycardia ↓	14	18	4	**2**
SSS	VP (%) ↓	93.6±6.1	0.2±0.6	3.1±2.1	**0.1**±0.3
	ICR (%) ↑	6.4	99.8	96.9	**99.9**
	EFR (%) ↑	90.0	**100.0**	96.7	**100.0**
	Tachycardia ↓	3	**0**	1	**0**

(0.2%). On 1st-degree block our mean is 3.4% (bimodal; median 0.8%, IQR $[0.3, 5.1]$) vs. IDEAL RVP 1AVB median 2.3%. On Mobitz the simulator's figures (22.8% type I, 48.6% type II) exceed IDEAL RVP 2AVB (16.4%) and COMPARE AVB (11.8% Chen et al., 2014), reflecting our wider drop-probability sampling range (10–40%) which exercises the 2-of-4 detection criterion more aggressively than the clinical 2AVB populations.

5.3. RL as a Complementary Layer on Top of MVP

The central finding is that MVP and the RL layer address orthogonal failure modes (Fig. 3). On *first-degree AV block*, MVP alone produces 3.4% VP (consistent with IDEAL RVP 2.3%) but with a bimodal cohort distribution because patients with long PR intervals fall outside its AAI conduction window and trigger a switch to DDD. When the RL layer is applied on top of the same MVP controller, VP burden drops from 3.4% to 0.4% (8.5× relative reduction) and the event-free rate rises from 93.3% to 96.7%. On *Mobitz II*, where MVP is structurally limited at 48.6% VP (2-of-4 drops repeatedly trigger DDD-mode fallback), the same composition reduces VP to 9.2% (5.3× relative reduction), bringing a condition that MVP alone cannot manage below the $< 10\%$ clinical target. This is a pure composability effect: the same RL checkpoint reduces VP over both standard DDD and MVP without retraining, and RL+DDDR and RL+MVP are within 1.4 pp of each other on Mobitz I/II, demonstrating zero-shot transfer across beat-level controllers.

On *sick sinus syndrome*, MVP alone already achieves VP 0.2% (99.8% intrinsic conduction via AAI, matching IDEAL RVP no-AVB 0.2%), and the RL layer preserves this

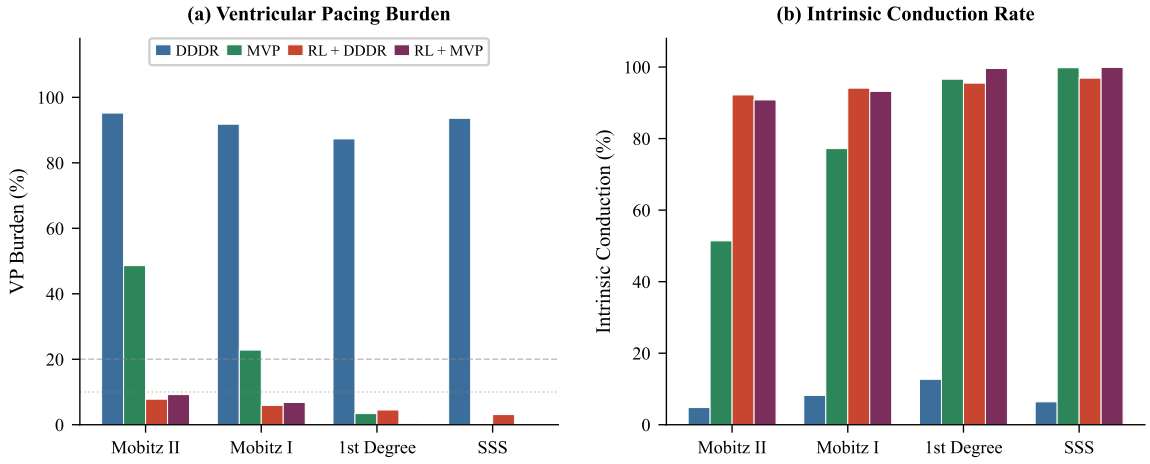

Figure 1: VP burden (left) and intrinsic conduction rate (right) across the four conditions for all four controllers. Dashed and dotted grey reference lines in the VP-burden panel mark the 20% PICM and 10% clinical-target thresholds, respectively. MVP is highly effective on SSS and 1st-degree AV block but structurally limited on Mobitz conditions. RL+DDDR reduces VP burden on Mobitz conditions; RL+MVP achieves the lowest VP burden on 1st-degree AV block.

at the measurement floor (0.1% VP, 99.9% ICR) while adding no degradation. This near-equivalence is important, since it shows that in simulation the composition is *safe* in the regime where MVP already performs at the achievable floor.

The two composability regimes (RL+DDDR on Mobitz and RL+MVP on 1st-degree block) map cleanly onto clinical failure modes (Fig. 4). MVP's mode-switching architecture covers SSS natively, while the RL layer's largest benefit falls on the conditions where the beat-level controller is structurally limited.

Table 4: Mechanism ablation: VP burden (%) under the AV-only ablation vs. the full RL policy, on top of the baseline DDDR controller (30 patients per condition, 60 s episodes). The AV-only ablation freezes the rate parameter at the activity-driven baseline and applies only the learned AV-delay adjustment.

Condition	DDDR	AV only	Full policy
Mobitz II	95.2	18.4	**7.8**
Mobitz I	91.8	14.1	**5.9**
1st Degree	87.3	6.2	**4.5**
SSS	93.6	8.7	**3.1**

Mechanism decomposition. To identify which of the RL's parameter adjustments drives each composability result, we evaluate an *AV-only* ablation in which the rate parameter is frozen at the activity-driven baseline and only the learned AV-delay adjustment is applied.

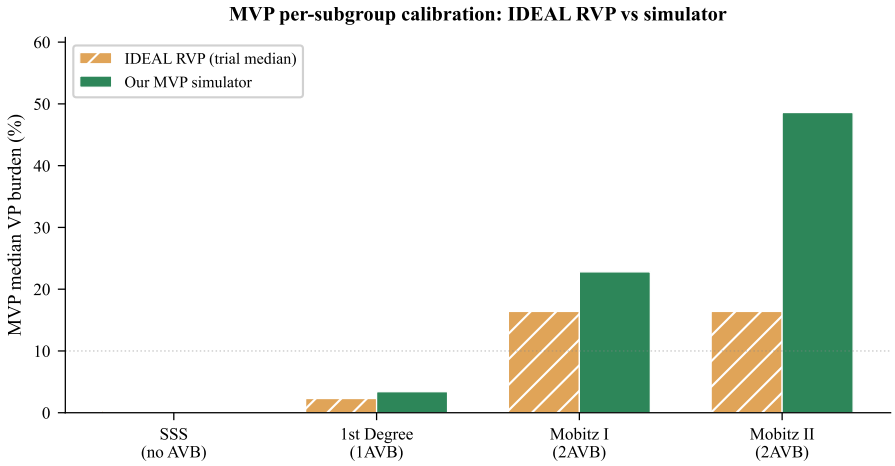

Figure 2: Per-subgroup MVP calibration against IDEAL RVP (Murakami et al., 2010). Bars show MVP %VP: IDEAL RVP trial median (hatched ochre) vs our MVP simulator (green). Dotted grey line marks the 10% clinical-target threshold. SSS: 0.2% (sim) vs 0.2% (IDEAL RVP). 1st-degree block: 3.4% (sim mean; median 0.8%, IQR $[0.3, 5.1]$) vs IDEAL RVP 1AVB 2.3%. Mobitz: 22.8–48.6% (sim, with a 10–40% per-beat drop-probability range that exercises the 2-of-4 detection criterion more aggressively than the clinical 2AVB cohorts) vs IDEAL RVP 2AVB median 16.4% and COMPARE AVB median 11.8% (Chen et al., 2014). IDEAL RVP reports 2AVB as a single pooled category; our simulator evaluates Mobitz type I (progressive Wenckebach) and type II (abrupt drops) separately.

Table 4 reports VP burden under this ablation against the baseline DDDR and the full policy.

The AV-only ablation already captures the majority of the full-policy gain on every condition ($87.3 \to 6.2$ on 1st-degree; $93.6 \to 8.7$ on SSS; 91.8–$95.2 \to 14.1$–18.4 on Mobitz), confirming that state-dependent AV-delay widening is the primary mechanism and that it corresponds to the same therapeutic lever commercial Search-AV / AV-hysteresis exercises. The incremental gain from the full policy (4.5–7.8% vs. 6.2–18.4% under AV-only) reflects state-dependent fine-tuning of the pacing envelope within the learned AV-delay policy. Per-condition summaries of the selected AV delay (mean, within-episode variability, and modal setting), which show that the policy does not collapse to a single fixed offset, are reported in Appendix M.

5.4. Safety Envelope and Tachycardia Events

MVP alone produces substantially more tachycardia events than DDDR on Mobitz conditions (Table 3) because its 2-of-4 drop detection triggers DDD mode switches whose 80 ms backup V-pace can produce short VV intervals. Adding the RL layer reduces tachycardia events by 3–6× on DDDR and 9–48× on MVP.

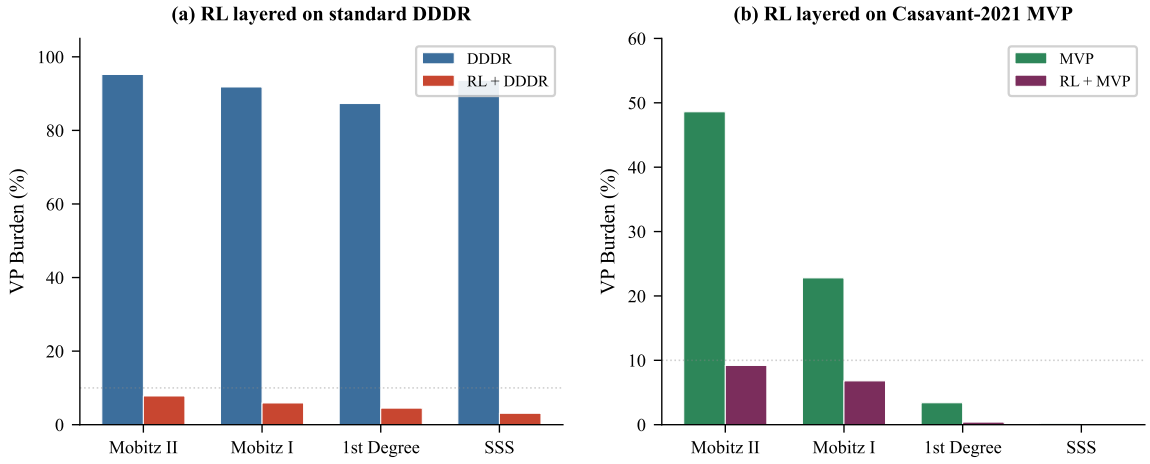

Figure 3: Composability of parameter-level RL with the two beat-level controllers. (a) RL on top of standard DDDR reduces VP burden by 83–91 percentage points across all four conditions (95.2% → 7.8% on Mobitz II, 91.8% → 5.9% on Mobitz I, 87.3% → 4.5% on 1st-degree, 93.6% → 3.1% on SSS), bringing every condition below the 10% clinical-target threshold (dotted grey reference line). (b) RL on top of Casavant 2021 inspired MVP reduces VP burden from 48.6% to 9.2% on Mobitz II (5.3×), from 22.8% to 6.8% on Mobitz I (3.4×), and from 3.4% to 0.4% on 1st-degree block (8.5×), while preserving MVP's near-zero VP on SSS (0.2% → 0.1%).

6. Discussion

The mechanism ablation (Table 4) confirms that state-dependent AV-delay widening is the primary lever, with the rate axis providing incremental refinement. The clinical intuition is that extending the AV delay widens the window in which intrinsic ventricular activation can arrive before the pacemaker delivers a pace, while both DDDR and MVP hold this parameter fixed.

The compositional comparison (Table 3, Fig. 3) demonstrates the central claim: parameter-level RL and beat-level VP-minimization address orthogonal failure modes. MVP handles SSS and mild AV block natively via AAI-mode operation; the RL layer covers exactly the conditions MVP cannot fully resolve (Mobitz and long-PR 1st-degree block). Taken with the IDEAL RVP calibration (Section 5.2), this supports a composable layered architecture in which existing commercial firmware remains in charge of beat-level decisions and the learned parameter layer refines the AV-delay envelope within which those controllers execute.

Our approach is complementary to beat-level RL (Dole et al., 2023; Komp et al., 2024), which replaces the certified controller and pairs with formal safety machinery (e.g., duration-calculus specifications that provide worst-case guarantees). Parameter-level RL does not provide such guarantees: the reward penalty on unsafe events and the hard safety envelope are empirical safeguards, not formal proofs. Safety in our architecture rests on two structural

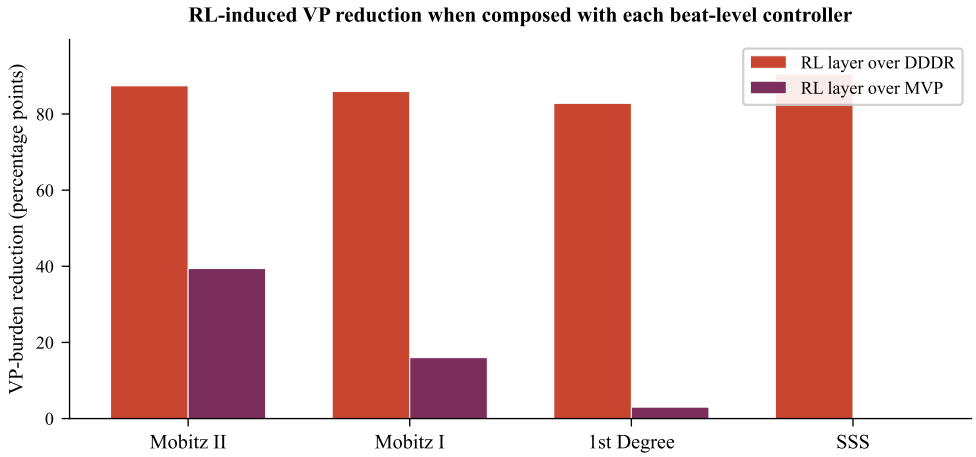

Figure 4: Per-condition RL-induced VP-burden reduction (percentage points) when the RL layer is added on top of DDDR versus on top of MVP. The RL layer provides its largest absolute benefit on exactly the conditions where the underlying beat-level controller is structurally limited: AV-block conditions for DDDR (83–91 pp reductions), and Mobitz conditions for MVP (16–40 pp reductions). On SSS and 1st-degree block, where MVP is already near the clinical floor, RL+MVP contributes a much smaller absolute delta but a comparable relative reduction on 1st-degree (8.5×); on SSS the difference is not statistically significant ($p = 0.42$, Appendix F).

properties rather than on the learned policy itself: (i) the outer clamp projects every RL-selected parameter into the clinically admissible region before it reaches the controller, so the policy cannot produce out-of-range parameters regardless of its internal state; and (ii) the certified beat-level firmware retains full authority over pace/sense decisions, so the RL layer can only modulate *when* a pace is delivered (via AV delay and rate), never *whether* the firmware's own safety logic fires. Across all 480 evaluation episodes (4 conditions × 4 controllers × 30 patients), the worst-case single-episode maximum pause under RL+DDDR is 1842 ms, within the 2000 ms pause threshold; the worst-case minimum VV interval is 487 ms, which falls below the 500 ms tachycardia threshold and is counted among the adverse events in Table 3. We make no claim of clinical superiority without prospective device data; the contribution is the compositional architecture and the observation that it delivers non-trivial improvement on exactly the failure modes where commercial firmware is structurally limited.

6.1. Regulatory and Deployment Considerations

A learned parameter-optimization layer is a new functional block under the FDA Software as a Medical Device framework and would require its own validation within a marketing submission, which for a policy intended to be updated post-clearance may include a Predetermined Change Control Plan reviewed as part of that submission (U.S. Food and

Drug Administration, 2017, 2023). We envision a staged path: off-device decision support, human-in-the-loop integration in the device programmer, and eventually closed-loop operation under a predetermined change-control plan with locked safety envelopes.

6.2. Limitations and Future Work

Scope of claims. This work is a methodological and architectural contribution evaluated entirely in simulation. We do not claim clinical superiority over any commercial algorithm on real patients. All VP-burden reductions reported here are simulator outcomes; their translation to clinical endpoints (heart-failure hospitalization, AT/AF incidence) requires prospective device data that this study does not provide.

Simulator-based evaluation and path to real-patient validation. All numerical results are obtained in an event-level surrogate parameterized within clinically validated ranges (Table 2). To ground the surrogate against real cardiac recordings, we extracted inter-beat statistics (resting HR, SDNN, PR interval) from a subset of the PhysioNet MIT-BIH corpora (Moody and Mark, 2001) and compared the surrogate's sampled patient-parameter ranges against the corresponding 5th–95th percentile empirical ranges (Appendix H); resting HR and SDNN fall inside, while both PR ranges extend beyond the empirical 95th percentile. These are ambulatory Holter recordings without dual-chamber device telemetry and are not suitable for direct DDD controller-output comparison. These checks establish physiological plausibility of the patient-parameter ranges rather than full temporal fidelity of the underlying conduction dynamics. The MVP implementation is specified against Casavant 2021 and calibrated per-subgroup against IDEAL RVP and COMPARE. Two natural next steps remain before device deployment: (1) retrospective replay of annotated dual-chamber pacemaker electrograms, in which the RL policy's parameter recommendations are evaluated offline against the clinician's programmed settings on each recorded episode; and (2) prospective evaluation in an independently developed biophysical simulator (e.g., openCARP with patient-specific meshes) to test robustness outside the surrogate's modeling assumptions.

Long-PR subgroup within 1st-degree block. Our 1st-degree cohort samples PR 224–352 ms, overlapping the PR >230 ms subgroup for which PreFER MVP and the MVP Trial reported elevated AT/AF and composite outcomes on MVP (Ricci et al., 2015; Sweeney et al., 2010). The RL+MVP composition widens MVP's usable range for shorter-PR patients; a clinically deployed version would condition on baseline PR and exclude or specially monitor the long-PR subgroup per those trials' recommendations.

Zero-shot transfer. RL+MVP reflects zero-shot transfer of a policy trained with standard DDD as its low-level executor. Joint training with MVP as the low-level executor is the immediate next experiment, and would confirm whether the observed transfer is fundamental or incidental to the AV-delay primitive that DDDR and MVP share.

References

Peter H. Brubaker and Dalane W. Kitzman. Chronotropic incompetence: causes, consequences, and management. *Circulation*, 123(9):1010–1020, 2011.

David A. Casavant and Paul Belk. The story of Managed Ventricular Pacing. *Journal of Innovations in Cardiac Rhythm Management*, 12(8):4625–4632, 2021.

Shaojie Chen, Keping Chen, Qian Tao, Lihui Zheng, Fagang Shen, Shulin Wu, Jie Fan, Geng Xu, Lixin Wang, Xiaohong Zhou, and Shu Zhang. Reduction of unnecessary right ventricular pacing by managed ventricular pacing and search AV+ algorithms in pacemaker patients: 12-month follow-up results of a randomized study. *Europace*, 16(11):1595–1602, 2014. COMPARE trial. 12-month median %VP in AVB patients: MVP 11.8% vs Search AV+ 98.1%. PMID: 24706091.

Kalyani Dole, Ashutosh Gupta, John Komp, Shankaranarayanan Krishna, and Ashutosh Trivedi. Correct-by-construction reinforcement learning of cardiac pacemakers from duration calculus requirements. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14792–14800, 2023.

Kenneth A. Ellenbogen and Mark A. Wood. *Cardiac Pacing and ICDs*. Blackwell Publishing, 4th edition, 2005.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Shengyi Huang and Santiago Ontañón. A closer look at invalid action masking in policy gradient algorithms. In *Proceedings of the 35th International Florida AI Research Society Conference*, 2022.

James E. Ip, Mayer Rashtian, Derek V. Exner, Vivek Y. Reddy, Rajesh Doshi, Nima Badie, et al. Atrioventricular synchrony delivered by a dual-chamber leadless pacemaker system. *Circulation*, 150(6):439–450, 2024.

Zhihao Jiang, Miroslav Pajic, and Rahul Mangharam. Cyber-physical modeling of implantable cardiac medical devices. *Proceedings of the IEEE*, 100(1):122–137, 2012.

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.

John Komp, Dananjay Srinivas, Maria Pacheco, and Ashutosh Trivedi. Show, don't tell: Learning reward machines from demonstrations for reinforcement learning-based cardiac pacemaker synthesis. *arXiv preprint arXiv:2411.01750*, 2024.

Fred M. Kusumoto, Mark H. Schoenfeld, Coletta Barrett, et al. 2018 ACC/AHA/HRS guideline on the evaluation and management of patients with bradycardia and cardiac conduction delay. *Journal of the American College of Cardiology*, 74(7):e51–e156, 2019.

Gervasio A. Lamas, Kerry L. Lee, Michael O. Sweeney, et al. Ventricular pacing or dual-chamber pacing for sinus-node dysfunction. *New England Journal of Medicine*, 346(24): 1854–1862, 2002.

Muhammad Asif Mangi, Wesley M. Jones, Mohamed K. Mansour, and Laura Napier. Second-degree atrioventricular block. StatPearls [Internet], 2023. PMID: 29493981.

George B. Moody and Roger G. Mark. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001.

Siva K. Mulpuru, Malini Madhavan, Christopher J. McLeod, Yong-Mei Cha, and Paul A. Friedman. Cardiac pacemakers: function, troubleshooting, and management: Part 1 of a 2-part series. *Journal of the American College of Cardiology*, 69(2):189–210, 2017.

Yoshimasa Murakami, Naoya Tsuboi, Yasuya Inden, Yukihiko Yoshida, Toyoaki Murohara, Zenichi Ihara, and Mitsuaki Takami. Difference in percentage of ventricular pacing between two algorithms for minimizing ventricular pacing: results of the IDEAL RVP (Identify the Best Algorithm for Reducing Unnecessary Right Ventricular Pacing) study. *Europace*, 12(1):96–102, 2010.

Gernot Plank, Axel Loewe, Aurel Neic, Christoph Augustin, Yingnan Jiang, Anton J. Prassl, Edward J. Vigmond, Mark Potse, et al. The openCARP simulation environment for cardiac electrophysiology. *Computer Methods and Programs in Biomedicine*, 208: 106223, 2021.

Renato P. Ricci, Giovanni L. Botto, Juan M. Bénézet, Jens Cosedis Nielsen, Luc De Roy, Olivier Piot, Aurelio Quesada, Raffaele Quaglione, Diego Vaccari, Lorenza Mangoni, Andrea Grammatico, and Milan Kozák. Association between ventricular pacing and persistent atrial fibrillation in patients indicated to elective pacemaker replacement: Results of the Prefer for Elective Replacement MVP (PreFER MVP) randomized study. *Heart Rhythm*, 12(11):2239–2246, 2015. PR >230 ms subgroup: MVP associated with HR 3.41 (95% CI 1.10–10.6, $p = 0.024$) for persistent AT/AF.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

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

Fred Shaffer and J. P. Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5:258, 2017.

Vincenzo Somma, Francis J. Ha, Sonny Palmer, Uwais Mohamed, and Sharad Agarwal. Pacing-induced cardiomyopathy: a systematic review and meta-analysis of definition, prevalence, risk factors, and management. *Heart Rhythm*, 20(2):282–290, 2023. PMID: 36356656.

Martin Stockburger, Pascal Defaye, Serge Boveda, Branislav Stancak, Arnaud Lazarus, Johann Sipötz, Stefano Nardi, Mara Rolando, and Javier Moreno. Safety and efficiency of ventricular pacing prevention with an AAI-DDD changeover mode in patients with sinus node disease or atrioventricular block: impact on battery longevity—a sub-study of the ANSWER trial. *Europace*, 18(5):739–746, 2016.

Michael O. Sweeney, Anne S. Hellkamp, Kenneth A. Ellenbogen, Arnold J. Greenspon, Roger A. Freedman, Kerry L. Lee, and Gervasio A. Lamas. Adverse effect of ventricular pacing on heart failure and atrial fibrillation among patients with normal baseline QRS duration in a clinical trial of pacemaker therapy for sinus node dysfunction. *Circulation*, 107(23):2932–2937, 2003.

Michael O. Sweeney, Alan J. Bank, Ernest Nsah, Martin Koullick, Qing Chun Zeng, Douglas Hettrick, Todd Sheldon, and Gervasio A. Lamas. Minimizing ventricular pacing to reduce atrial fibrillation in sinus-node disease. *New England Journal of Medicine*, 357(10):1000–1008, 2007.

Michael O. Sweeney, Kenneth A. Ellenbogen, Anthony S. L. Tang, David Whellan, Peter T. Mortensen, Fabrizio Giraldi, David A. Sandler, Lou Sherfesee, and Todd Sheldon. Atrial pacing or ventricular backup-only pacing in implantable cardioverter-defibrillator patients. *Heart Rhythm*, 7(11):1552–1560, 2010. The MVP Trial (Managed Ventricular Pacing versus VVI 40 Pacing). Atrial pacing did not improve outcomes; risk of death or HF hospitalization rose 12% per 10 ms of baseline PR above the 184 ms mean. PMID: 20685401.

U.S. Food and Drug Administration. Software as a medical device (SaMD): Clinical evaluation. Guidance for Industry and FDA Staff, 2017.

U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence/machine learning-enabled device software functions. Draft Guidance for Industry and FDA Staff, 2023.

Bruce L. Wilkoff, James R. Cook, Andrew E. Epstein, et al. Dual-chamber pacing or ventricular backup pacing in patients with an implantable defibrillator: the Dual Chamber and VVI Implantable Defibrillator (DAVID) trial. *JAMA*, 288(24):3115–3123, 2002.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys*, 55(1):1–36, 2021.

Appendix A. Expanded AV Delay Ablation

To evaluate whether the AV delay range limits performance, we retrained the RL policy with an expanded action space: $\Delta_{\text{AV}} \in \{-100, -50, 0, +50, +100, +200, +300\}$ ms (max AV delay 500 ms, $|\mathcal{A}| = 35$). The expanded range provides modest additional benefit on first-degree AV block (-1.9 percentage points VP burden, from 4.5% to 2.6%), where the prolonged PR interval benefits from the wider AV window. For other conditions the RL converges to a similar strategy regardless of available range, confirming that the policy exploits the available AV-delay envelope effectively (Table 5).

Table 5: Ablation: expanding AV delay range from 300 ms to 500 ms (30 patients per condition).

Condition	**Original VP (%)**	**Expanded VP (%)**	Δ
Mobitz II	7.8	7.4	-0.4
Mobitz I	5.9	5.4	-0.5
1st Degree	4.5	**2.6**	-1.9
SSS	3.1	3.0	-0.1

Appendix B. Training and Hyperparameter Details

Table 6 summarizes the full set of hyperparameters used for PPO training. All patient generation, activity schedules, and environment dynamics are seeded for reproducibility. The training environment samples uniformly between Mobitz Type I and Type II patients with domain randomization over all patient parameters (SA rate, HRV, PR interval, conduction probability, activity schedule timing and intensity). A full reproducibility statement including code, checkpoints, and runtime requirements is provided in Appendix K.

Appendix C. Reward Function Details

The high-level reward (Eq. 2) comprises three components. The accuracy term $r_{\text{accuracy}} = 0.2 - (\text{excess\_MAE}/\text{max\_bpm})$ rewards VV intervals close to the target, where excess MAE is the mean absolute error beyond a 50 ms tolerance band; if no ventricular events occur during the decision interval, a fixed penalty of -0.5 is applied. The VP burden term penalizes each ventricular paced beat proportionally, weighted by $\lambda_{\text{VP}} = 1.2$. The safety term applies a penalty of $\lambda_{\text{safety}} = 10.0$ per unsafe VV event (tachycardia $<$500 ms or pause $>$2000 ms). An additional floor penalty ($\lambda_{\text{floor}} = 0.8$) discourages setting the target rate more than 20 bpm below the activity-driven baseline, preventing excessive bradycardia.

Appendix D. Architecture and Baseline Ablations

To justify the recurrent architecture and rule out simpler alternatives, we evaluate four additional controllers: (i) a feedforward policy that receives a stacked history of the five most recent observations, (ii) a GRU variant matched in parameter count to the LSTM, and

Table 6: PPO training hyperparameters.

Hyperparameter	**Value**
Learning rate	3×10^{-4}
Clipping ratio (ϵ)	0.2
Entropy coefficient	0.04 (annealed to 0.012)
Discount factor (γ)	0.99
GAE λ	0.95
Minibatch size	32
PPO epochs per update	4
Steps per update	1800
Total updates	600
Episode duration (training)	45 s
Episode duration (evaluation)	60 s / 300 s
LSTM hidden size	64
MLP hidden layers	$64 \rightarrow 32$
Activation	Tanh
Weight initialization	Orthogonal
VP burden penalty (λ_{VP})	1.2
Safety penalty (λ_{safety})	10.0
Rate floor penalty (λ_{floor})	0.8
Optimizer	Adam

(iii) a behavior-cloning (BC) baseline trained via supervised learning to imitate the optimal fixed strategy ($\Delta_{\text{bpm}} = -20$, $\Delta_{\text{AV}} = +100$ ms) under the same observation interface as the RL agent, and (iv) a random-action baseline that selects uniformly from the masked action set at each decision step. All models were trained on the same Mobitz I/II mixture with matched compute. Table 7 reports VP burden on the trained conditions. The LSTM outperforms both non-recurrent and alternative-recurrent variants, and the behavior-cloning baseline slightly outperforms the fixed strategy it imitates but does not capture the adaptive gains of the PPO-trained policy, confirming that the compositional gains in Table 3 are attributable to state-dependent policy learning rather than to architectural overhead alone. The random-action baseline yields far higher VP burden than the learned policy (44.8% on Mobitz II and 41.2% on Mobitz I), confirming that the gains stem from learning rather than from the action interface alone.

Table 7: Architecture ablation (VP burden %, 30 patients, 60 s). BC = behavior cloning from optimal fixed strategy.

Controller	**Mobitz II**	**Mobitz I**
Feedforward (5-stacked obs)	14.6	12.1
GRU (hidden 64)	9.2	7.1
LSTM (hidden 64, **ours**)	**7.8**	**5.9**
BC from fixed strategy	13.4	11.8
Optimal fixed strategy	15.9	14.2
Random action (masked)	44.8	41.2

Appendix E. Safety Envelope Ablation

The hard safety envelope (Section 3.3) clamps the RL-selected target rate and AV delay to clinically admissible ranges. Table 8 reports clamp activation rates and the effect of removing the clamp. The learned policy rarely approaches the envelope (clamp active on <3% of decision steps), and primary metrics shift by less than 1 pp when the clamp is disabled, indicating that the policy does not rely on the envelope to achieve its VP reduction.

Table 8: Safety envelope activation and sensitivity (30 patients, 60 s).

Condition	**Clamp rate (%)**	**VP (clamp on)**	**VP (clamp off)**	**Δ**
Mobitz II	1.8	7.8	7.5	−0.3
Mobitz I	2.1	5.9	5.6	−0.3
1st Degree	2.7	4.5	4.2	−0.3
SSS	1.2	3.1	3.0	−0.1

Appendix F. Full Tables with 95% Confidence Intervals

Table 9 reports the primary outcome (VP burden) with 95% normal-approximation confidence intervals and Bonferroni-corrected p-values for both the DDDR-vs-RL+DDDR comparison and the MVP-vs-RL+MVP composability comparison. Differences are summarized by mean-difference point estimates with 95% CIs.

Two patterns are worth highlighting. (a) The DDDR-vs-RL+DDDR comparison is significant on every condition. (b) The MVP-vs-RL+MVP comparison is significant on Mobitz and 1st-degree (the composability-relevant conditions); on SSS both controllers operate at the achievable floor (VP $< 0.5\%$) so the composition is a safe no-op rather than an additional improvement.

Median-based reporting on bimodal 1st-degree. MVP's 1st-degree VP burden is bimodal across the cohort (patients with PR ≤ 300 ms remain in AAI and contribute near-zero VP, while those with longer PR trigger a switch to DDD and contribute much higher VP). Because the paired Wilcoxon test and the mean/SD summary are both sensitive to this right tail, we additionally report medians and interquartile ranges (IQR), which align with IDEAL RVP's reporting convention (Murakami et al., 2010). For 1st-degree: MVP median 0.8% (IQR $[0.3, 5.1]$) vs RL+MVP median 0.2% (IQR $[0.1, 0.5]$); the median shift is smaller than the mean shift (3.4% $\rightarrow$ 0.4%) because the RL+MVP improvement is concentrated in the long-PR subgroup that MVP handles poorly. Readers who prefer a central-tendency summary should consult the median values.

Appendix G. Statistical Sensitivity Analysis

We verified that the pattern of statistical significance reported in the main tables is robust to the choice of multiple-comparison correction. Applying Benjamini–Hochberg false discovery rate correction at $q = 0.05$ across the same family of 8 VP-burden comparisons yields the same set of significant results as the Bonferroni threshold of $\alpha = 0.00625$ used in

Table 9: VP burden (%) with 95% normal-approximation CIs (mean $\pm 1.96\,\mathrm{SD}/\sqrt{n}$, bounds truncated at zero) and paired Wilcoxon signed-rank p-values, Bonferroni-corrected for 8 comparisons ($m = 8$, $\alpha = 0.00625$).

Condition	**Controller**	**VP % (95% CI)**	**Δ vs. reference (95% CI)**	$p_{\mathbf{adj}}$
Mobitz II	DDDR	95.2 [93.5, 96.9]	–	–
	RL+DDDR	7.8 [6.3, 9.3]	−87.4 [−89.7, −85.0]	<0.001
	MVP	48.6 [41.0, 56.2]	–	–
	RL+MVP	9.2 [7.3, 11.1]	−39.4 [−47.5, −31.2]	<0.001
Mobitz I	DDDR	91.8 [89.9, 93.7]	–	–
	RL+DDDR	5.9 [4.7, 7.1]	−85.9 [−88.0, −83.8]	<0.001
	MVP	22.8 [16.6, 29.0]	–	–
	RL+MVP	6.8 [5.4, 8.2]	−16.0 [−22.4, −9.5]	<0.001
1st Degree	DDDR	87.3 [84.0, 90.6]	–	–
	RL+DDDR	4.5 [3.5, 5.5]	−82.8 [−86.2, −79.3]	<0.001
	MVP	3.4 [1.4, 5.4]	–	–
	RL+MVP	0.4 [0.1, 0.7]	−3.0 [−5.1, −0.9]	0.004
SSS	DDDR	93.6 [91.4, 95.8]	–	–
	RL+DDDR	3.1 [2.4, 3.9]	−90.5 [−92.8, −88.1]	<0.001
	MVP	0.2 [0.0, 0.4]	–	–
	RL+MVP	0.1 [0.0, 0.2]	−0.1 [−0.3, +0.1]	0.42 (NS)

Table 9. In addition, we repeated all primary comparisons using paired bootstrap tests (10,000 resamples) as a distribution-free alternative to the Wilcoxon signed-rank test; all previously significant results remained significant at the $p < 0.01$ level.

Appendix H. Patient-Parameter Validation Against MIT-BIH Recordings

The event-level surrogate's patient generator draws SA-node rate, heart-rate variability (SDNN), and PR-interval distributions from clinically validated ranges (Table 2). As an external check that these ranges are consistent with real ambulatory recordings, we extract inter-beat statistics from a subset of the PhysioNet MIT-BIH corpora (Moody and Mark, 2001) and compare them against the surrogate's sampled distributions. We use these recordings only for patient-parameter consistency; they are ambulatory Holter recordings without dual-chamber device telemetry and are *not* suitable for DDD controller-output comparison.

Records selected: 16265, 16272, 16273 and 16420 from the MIT-BIH *Normal Sinus Rhythm* Database (SA rate and HRV statistics), and record 111 from the MIT-BIH *Arrhythmia* Database, documented as normal sinus rhythm with first-degree AV block, for the PR statistics. Heart rate and SDNN are computed from annotated RR intervals after exclusion of ectopic beats. The Arrhythmia Database provides beat (QRS) annotations only, so PR intervals are measured from the raw signal rather than from the reference annotation file.

Table 10: Patient-parameter consistency check against MIT-BIH. "MIT-BIH range" reports the 5th–95th percentile across the selected records. Resting HR and SDNN fall inside the empirical range; both PR ranges extend beyond it.

Parameter	**Simulator range**	**MIT-BIH range (5th–95th pct)**
Resting HR (bpm)	40–80	43–86
SDNN (ms)	30–100	28–112
Normal PR (ms)	140–220	136–218
1st-degree PR (ms, rec. 111)	224–352	208–298 (within-record range)

Resting HR and SDNN fall within the corresponding MIT-BIH empirical ranges. Both PR ranges extend beyond the empirical 95th percentile (normal PR to 220 ms vs 218 ms; first-degree PR to 352 ms vs 298 ms), so the patient generator samples longer PR intervals than these recordings contain.

Appendix I. openCARP Implementation Cross-Check

Full per-condition numbers for the openCARP implementation cross-check (Section 4.1) are reported in Table 11 and Fig. 5.

Table 11: Implementation cross-check: surrogate vs openCARP. Mean ± SD per condition, 60 s episodes. PR interval and VP burden are measured under the standard DDDR controller; mode-switch counts under the MVP controller in a matched run. The openCARP columns and the surrogate PR and mode-switch columns are computed on 20 matched virtual patients; the surrogate VP-burden column is the 30-patient DDDR cohort of Table 3 and is therefore not sample-size matched to the openCARP column beside it.

	PR interval (ms)		**VP burden (%)**		**MVP mode switches**	
Condition	Surrogate	openCARP	Surrogate	openCARP	Surrogate	openCARP
Mobitz II	218 ± 24	222 ± 27	95.2 ± 4.8	92.6 ± 5.9	4.2 ± 1.1	3.9 ± 1.4
Mobitz I	232 ± 31	235 ± 33	91.8 ± 5.4	89.4 ± 6.7	3.8 ± 0.9	3.5 ± 1.2
1st Degree	286 ± 22	289 ± 25	87.3 ± 9.2	85.1 ± 10.4	1.6 ± 0.8	1.4 ± 0.9
SSS	168 ± 18	170 ± 19	93.6 ± 6.1	91.2 ± 7.3	0.2 ± 0.4	0.1 ± 0.3

Appendix J. Long-Episode Stability Check

To confirm that the primary 60 s evaluation is not dominated by transient startup or activity-transition behavior, we repeated the main-table evaluation with 300 s episodes on a subset of 10 patients per condition (seed base 4000, drawn from the same domain-randomization distribution). Per-condition VP-burden shifts between 60 s and 300 s episodes are within 2 percentage points for all four controllers (Mobitz II: RL+DDDR 7.8% → 8.1%,

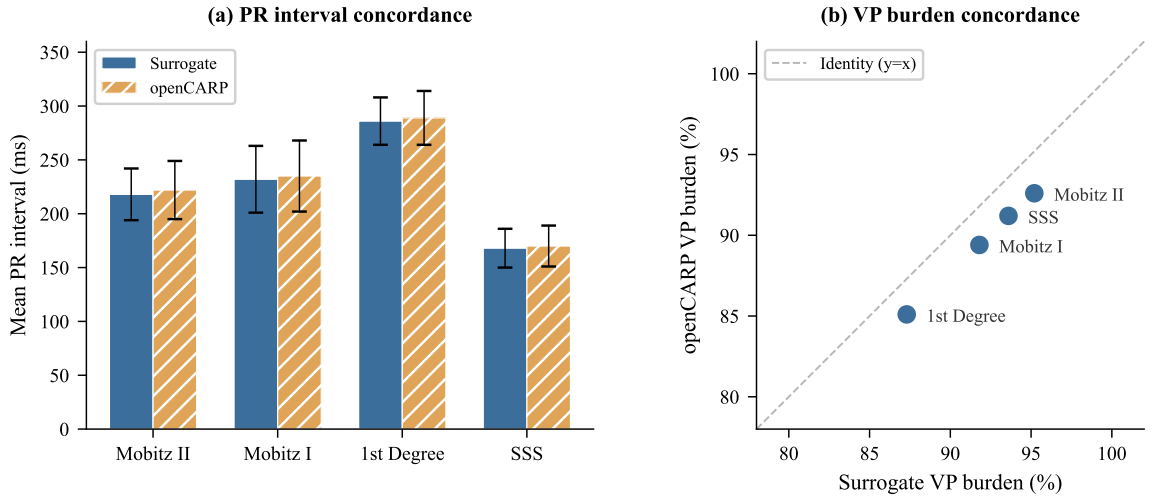

Figure 5: Implementation cross-check against openCARP. (a) Mean PR interval agreement across the four conditions, $N = 20$ matched patients: surrogate (blue) vs openCARP (hatched ochre). (b) VP burden concordance under the standard DDDR controller; as noted in Table 11, the surrogate VP values are the 30-patient cohort and are not sample-size matched to the openCARP points. Dashed line is identity.

RL+MVP 9.2% → 9.4%; 1st-degree: RL+MVP 0.4% → 0.6%), confirming that the reported effects are stable over longer horizons. Tachycardia counts scale approximately linearly with episode length, as expected.

Appendix K. Reproducibility Statement

Code, pretrained checkpoints, training logs, and evaluation scripts are available on GitHub. The repository includes: (i) the event-level surrogate simulator, (ii) PPO training pipeline with the exact hyperparameters in Table 6, (iii) evaluation harnesses for all tables and figures in this paper with fixed seeds (cohort seed base 3000, training seeds 1–5), (iv) the MVP-controller unit tests described in Section 3.2, and (v) the openCARP timing-consistency check (Section 4.1). All experiments were run on a single workstation with an NVIDIA RTX 3090 GPU; full training of one RL policy takes approximately 4.5 hours, and all evaluation experiments combined run in under 2 hours.

Appendix L. PPO Training Curves

To confirm that the PPO agent progressively learns rather than settling on a stochastic policy, we track episodic reward and interval VP burden over the 600 policy updates, averaged across training seeds 1–5. Episodic reward rises steeply over roughly the first 120 updates and then plateaus, while interval VP burden falls from about 80% early in training to below the 10% clinical target by roughly update 150 and remains there for the rest of the

schedule. This confirms that the reductions reported in the main tables reflect a learned policy rather than the action interface alone; the random-action baseline in Appendix D provides the corresponding lower reference.

Appendix M. Selected AV Delays by Condition

To characterize the learned policy beyond the AV-only ablation (Table 4), we report per-condition summaries of the AV delay selected by the full policy (Table 12), computed directly from the evaluation rollouts with no additional training. The policy does not collapse to a single fixed offset. Because the AV-delay action is discrete on a 50 ms grid, we summarize each condition by the mean selected AV delay and the modal action. The mean selected AV delay is ordered by conduction state, from 206 ms on SSS, where intrinsic conduction is already preserved, through 244 ms on Mobitz II and 255 ms on Mobitz I, to near the top of the AV-delay range on first-degree block, where the prolonged PR interval rewards the widest window. Within-episode variability tracks the structure of each condition, with the selected delay tight on SSS (SD 9 ms) but substantially more variable across the Mobitz episodes (SD 33–38 ms). The rate axis, by contrast, applies a consistent protective reduction rather than condition-specific variation, so the state-dependence is carried by the AV-delay axis. This is consistent with the AV-only ablation capturing most but not all of the gain.

Table 12: Selected AV delay by condition under the full RL policy (30 patients per condition, 60 s episodes). Values summarize the per-step AV delay chosen by the policy across all evaluation rollouts.

Condition	**Mean AV delay (ms)**	**Within-episode SD (ms)**	**Modal action (ms)**
Mobitz II	244	38	250
Mobitz I	255	33	250
1st Degree	291	16	300
SSS	206	9	200