CHOICE HETEROGENEITY AS STRUCTURED DEVIATIONS FROM A SHARED VALUE ANCHOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Action preferences can differ systematically across trajectories and over time. In clinical care, patients with similar observed physiological conditions may receive different treatments even when clinicians share the goal of improving patient outcomes. Behavioral prediction alone leaves these persistent differences unexplained against a common reference. We introduce a decomposition of sequential behavior that combines a shared action-value anchor with switching regimes characterized by interpretable action-cost profiles. A continuous-time latent-state model learns interpretable physiological phenotypes and their dynamics from irregular, partially observed records. It supports forecasting, constructs the anchor under a prespecified reward and continuation rule, and provides physiological context for clinical audit. With the anchor fixed, the regime layer learns action-dependent costs within a common feature dictionary, making treatment preferences comparable across regimes. Controlled synthetic experiments support regime and action-cost contrast recovery. In CKD–MBD and MIMIC–IV, the framework improves physiological forecasting and action prediction, while inferred regimes capture persistent treatment patterns. Clinical cases link these patterns to physiological trajectories and decompose anchor-relative departures into treatment-feature contributions, supporting longitudinal clinical audit.

1 INTRODUCTION

Sequential decisions in clinical care, dynamic pricing, and operations are commonly formalized as Markov decision processes (BELLMAN, 1957), with offline reinforcement learning used to learn from logged trajectories despite distribution shift (Levine et al., 2020; Fu et al., 2020; Komorowski et al., 2018). A recurring empirical pattern, however, is *apparent behavioral heterogeneity*: under similar observed conditions, agents take different actions across time and contexts (Widmer & Kubat, 1996; Jung & Shah, 2015).

In ICU sepsis management, for example, patients with similar observed physiological conditions may receive different treatments even when clinicians share the goal of improving patient outcomes. Drug shortages, staffing changes across shifts, evolving hospital protocols, and temporary restrictions on interventions can shape the decision context over several decisions. Observed choices may therefore reflect both clinical reasoning and persistent features of the surrounding decision context.

Standard modeling approaches capture such variation through drifting rewards, flexible policy classes, or stochastic behavior (Cheung et al., 2020; Lecarpentier & Rachelson, 2019). Approaches in offline and inverse reinforcement learning explain heterogeneity through inferred rewards (Ng & Russell, 2000; Ziebart et al., 2008), latent tasks (Bloem & Bambos, 2015; Yu et al., 2019; Chen et al., 2024), or bounded decision-making and interaction (Hadfield-Menell et al., 2016; Jacob et al., 2023). Meanwhile, switching state-space models represent persistent behavior through latent regime dynamics (Ghahramani & Hinton, 2000; Hamilton, 1989). These formulations describe heterogeneity through variation in rewards, policies, tasks, or latent states. Accurate behavioral prediction alone does not establish how persistent differences should be interpreted against a common value reference. We seek to separate that reference from the systematic adjustments to action preferences expressed by different behavioral regimes. An extended discussion of related work is provided in Appendix A.

We propose a different perspective: behavioral heterogeneity can be understood as persistent, structured adjustments around a shared decision anchor. Motivated by persistent decision pressures, we

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

introduce *shared-anchor regime decomposition*, separating behavior into (*i*) a shared action-value anchor providing a stable, outcome-based reference and (*ii*) a switching regime process with interpretable, action-dependent cost profiles. The anchor remains fixed during fitting, while each regime learns costs over a common feature dictionary. Treatment identity, regimen changes, and state–treatment interactions describe how each regime adjusts action rankings. This structure connects recurring treatment patterns to explicit, comparable action-cost contrasts.

A central difficulty is that clinical records are irregularly sampled and contain missing measurements. Stage I therefore uses a continuous-time latent-state model to propagate state beliefs between observations and update them using available measurements (Rubanova et al., 2019; Habiba & Pearlmutter, 2020; Ansari et al., 2023). This provides a common basis for decision-time state inference and action valuation. Physiological phenotypes are interpretable through measurement profiles, persistence, and transitions. Under a prespecified reward and continuation rule, the fitted dynamics define the shared anchor. Conditioned on pre-action beliefs and contexts, Stage II captures persistent behavioral regimes through action-dependent costs while keeping this reference fixed. Together, the two stages link treatment patterns and cost contributions to the patient's physiological course, supporting clinical audit.

Controlled synthetic experiments support regime and cost recovery and isolate the roles of the anchor and persistence. In CKD–MBD and MIMIC–IV, Stage I achieves the lowest mean forecasting NLL among the compared methods and provides interpretable physiological trajectories and anchor-reference actions for clinical audit. Stage II improves held-out action prediction and characterizes persistent treatment patterns through comparable action-cost profiles, with comparisons against fixed trajectory classes supporting the contribution of switching. Clinical cases connect these patterns to physiological changes and explain individual departures through treatment-feature contributions. Sensitivity analyses assess the conditional reproducibility of these patterns and explanations.

Finally, rather than assigning different rewards or tasks to each regime, we keep the outcome objective and shared value reference fixed and explain behavioral variation through regime-dependent action costs. This yields a factored interpretation of sequential behavior: stability lies in the anchor, and heterogeneity lies in structured deviations. Regimes remain the main object of analysis, with the shared anchor and common cost features making their action preferences comparable and interpretable. By linking these regimes and their inferred cost contributions to the patient's physiological course, we aim to help hospitals understand persistent departures from a specified clinical value reference, identify behavioral regimes and their transitions, and obtain leads for clinical review and management investigation, supporting longitudinal clinical audit.

2 METHOD

We model structured behavioral heterogeneity in sequential decision logs. The goal is not to learn a separate reward, policy, or MDP for each latent group. Instead, we ask whether heterogeneous actions can be explained as persistent, structured deviations from a shared value-based anchor.

The method has two main stages (Figure 1). **Stage I** constructs decision-time latent states and a shared anchor value function. **Stage II** learns behavioral regimes that modulate action selection relative to this anchor. At a high level, the action score under regime u is

$$S_u(x, \xi, a) = \underbrace{Q_{\mathrm{anc}}(x, a)}_{\text{shared anchor value}} - \underbrace{C_u(\xi, a)}_{\text{regime-specific anchor-relative cost}}, \qquad C_u(\xi, a) = w_u^\top g(\xi, a). \tag{1}$$

Here x is the inferred decision-time state and ξ contains pre-action variables available to the regime layer, including x, the previous action, and optional side information. The anchor Q_{anc} is shared across trajectories and regimes. Every regime uses the same action-dependent feature dictionary g, but has its own complete coefficient vector w_u. Thus changing the candidate action can change both $Q_{\mathrm{anc}}(x, a)$ and $C_u(\xi, a)$, allowing the learned regime cost to change action rankings and probabilities. Since candidate-normalized softmax probabilities are invariant to action-independent offsets, we center C_u over the candidate set and interpret its signed values through within-decision action contrasts.

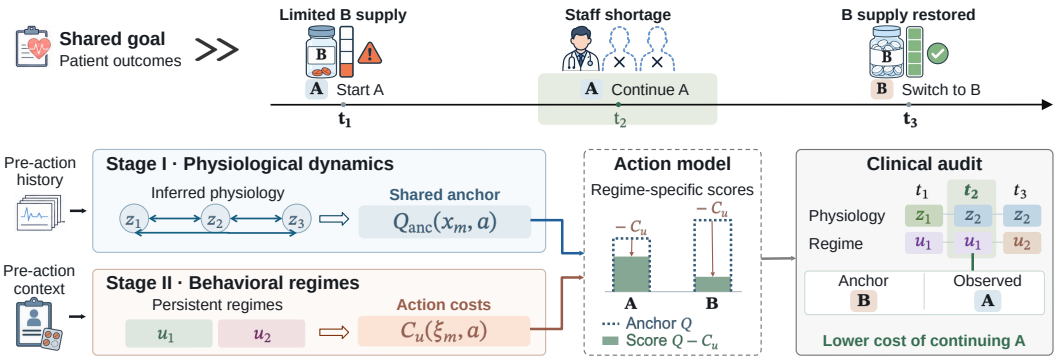


Figure 1: Illustrative treatment course. Stage I infers physiological changes from irregular, incomplete histories and constructs a shared value anchor. Stage II uses inferred states and pre-action context to learn persistent regimes and action costs. Scores $Q_{\text{anc}} - C_u$ (bars) yield choice probabilities through within-regime softmax and regime mixing. Clinical audit links both trajectories to the cost contrast favoring continued A over anchor-preferred B at t_2.

2.1 Sequential decision data

We observe N trajectories. In trajectory i, decisions occur at irregular times

$$t_1^i < t_2^i < \cdots < t_{M_i}^i.$$

At decision epoch m, o_m^i denotes newly available pre-action observations, with their timestamps retained, and $a_m^i \in \mathcal{A}$ is the chosen action, where $\mathcal{A}$ is finite:

$$\mathcal{D}^i = \{(t_m^i, o_m^i, a_m^i)\}_{m=1}^{M_i}.$$

We write $\mathcal{A}_m \subseteq \mathcal{A}$ when specifying the candidates at a particular decision; the clinical experiments use a common candidate vocabulary within each split.

The observed covariates need not be Markovian or fully informative. We therefore introduce a latent system state $z(t) \in \{1, \ldots, K_z\}$, evolving in continuous time, and use the posterior belief over this state as the decision-time representation:

$$b_m^i(k) = \Pr\big(z(t_m^i) = k \mid o_{1:m}^i, a_{1:m-1}^i\big).$$

We also maintain continuous outcome-proxy variables $\lambda_m^i = \lambda(t_m^i)$. The state used by the action model is

$$x_m^i = (b_m^i, \lambda_m^i).$$

Thus, x_m summarizes the model's belief when the action is chosen. We write $\mathcal{H}_m^-$ for the available pre-action history, including observation times, past actions, and current context. In the clinical analyses, observations recorded in the decision's own month or hour enter subsequent histories.

2.2 Stage I: constructing the shared anchor

Stage I produces two objects for every decision epoch:

$$x_m \quad \text{and} \quad Q_{\text{anc}}(x_m, a),\ a \in \mathcal{A}.$$

The first object is the decision-time belief state. The second is a shared action-value anchor used to compare actions across trajectories.

In our implementation, the latent system state follows an action-modulated continuous-time Markov chain. If action a_m is held fixed over the interval $[t_m, t_{m+1})$, then

$$\Pr(z(t_{m+1}) = \cdot \mid z(t_m), a_m) = \exp\{A(a_m)\Delta_m\}, \qquad \Delta_m = t_{m+1} - t_m,$$

where $A(a)$ is the generator under action a. This directly handles irregular decision times without forcing trajectories onto a fixed grid.

The outcome proxies $\lambda(t)$ evolve continuously between observations according to state-dependent dynamics; the explicit ODE form is deferred to Appendix C. At observation time s, an available measurement $o(s)$ has likelihood

$$p_\theta(o(s) \mid z(s), \lambda(s)).$$

Over an interval $[s_0, s]$ with fixed action a, belief updating takes the prediction-correction form

$$b^-(s) = b(s_0)\exp\{A(a)(s - s_0)\}, \qquad b(s;k) \propto b^-(s;k)\, p_\theta(o(s) \mid z(s) = k, \lambda(s)).$$

Propagation is split at action changes and observation times; decision-time beliefs use only observations already available. Only available measurement channels contribute to the emission likelihood; between observations, the CTMC and ODE propagate the state under recorded treatments.

Given the fitted dynamics, a reference return over horizon T is

$$Q^T_{\mathrm{raw}}(x, a) = \mathbb{E}_{\pi_0}\left[\int_0^T e^{-\rho t} r_{\mathrm{anc}}(z(t), \lambda(t), a(t))\, dt \;\middle|\; x(0) = x,\ a(0) = a\right],$$

where r_{anc} is a reference reward, $\rho \geq 0$ is the discount rate, and π_0 specifies continuation after the initial action interval. The clinical experiments use six-step returns over six months or six hours, repeating the candidate action with per-step discount 0.99. We denote the resulting action-centered, training-scaled values by Q_{anc}; Appendix F specifies the reward and normalization. This anchor is fixed before Stage II and shared by all regimes.

Stage I tracks and predicts physiological evolution, constructs shared action values, and supplies patient-state context for clinical audit. Stage II explains persistent treatment patterns through action-dependent costs. The specified reward, continuation assumption, and value scale give the costs their reference meaning. Comparing anchor and fitted action rankings locates departures for review, while action-dependent costs describe the fitted adjustments. Action likelihood measures behavioral fit; it does not by itself establish the correctness of the reference or the resulting interpretation.

2.3 Stage II: switching behavioral regimes

Stage II models structured behavioral heterogeneity through a latent switching regime layer. A *behavioral regime* is a persistent mode of action adjustment relative to the shared anchor Q_{anc}. Resource availability, policy requirements, and management burden motivate this representation; in the clinical analyses, each regime is characterized by its treatment-cost profile.

We introduce a latent regime

$$u_m \in \{1, \ldots, K_u\},$$

which evolves along the decision sequence according to

$$\Pr(u_m = v \mid u_{m-1} = u, \Delta_{m-1}) = \Pi_{uv}(\Delta_{m-1}).$$

In the clinical model, $\Pi(\Delta) = e^{-\Delta/\tau} I + (1 - e^{-\Delta/\tau})\mathbf{1}\pi^\top$, where π is the stationary regime distribution and $\tau > 0$ controls persistence in physical time. The previous regime's influence decays toward π as the time between decisions increases. Regime persistence lets successive actions provide evidence about a common behavioral pattern. Each regime defines one complete action-dependent cost $C_u(\xi, a)$ over the common feature dictionary. Here ξ_m denotes the pre-action context used to explain behavioral variation, including the latent decision state x_m, action-history features, and available protocol or operational indicators.

Conditional on regime u, actions are generated according to

$$p(a_m = a \mid x_m, \xi_m, u_m = u) = \frac{\exp\{\beta[Q_{\mathrm{anc}}(x_m, a) - C_u(\xi_m, a)]\}}{\sum_{a' \in \mathcal{A}} \exp\{\beta[Q_{\mathrm{anc}}(x_m, a') - C_u(\xi_m, a')]\}}. \tag{2}$$

At prediction time the current regime is unknown. Let $\rho^-_{m,u} = p(u_m = u \mid a_{<m}, x_{\leq m}, \xi_{\leq m})$ be the pre-action regime distribution. The model probability used for held-out evaluation is therefore

$$p(a_m = a \mid \mathcal{H}^-_m) = \sum_{u=1}^{K_u} \rho^-_{m,u}\, p(a_m = a \mid x_m, \xi_m, u_m = u), \tag{3}$$

which uses only information available before observing a_m.

Each cost feature in equation 1 is an action descriptor or a context–action interaction. Features with zero candidate-action contrast remain in the dictionary but do not affect equation 2 or receive an action-level interpretation.

With Q_{anc} fixed before fitting, the dictionary constrains how behavior can depart from the reference. An unanchored linear score can also be rewritten relative to Q_{anc} after fitting, but its implied cost generally leaves the prescribed dictionary unless the anchor's action contrasts lie in that feature span.

For fixed (x, ξ, u), the softmax distribution in equation 2 uniquely maximizes expected anchor-minus-cost score plus entropy weighted by $1/\beta$ (Appendix B.2).

After fitting, Stage II produces

$$\gamma_m^i(u) = \Pr(u_m^i = u \mid \mathcal{D}^i), \qquad \Pi(\cdot), \qquad \{C_u\}_{u=1}^{K_u}.$$

3 Learning and Model Analysis

Learning follows the decomposition above. Stage I estimates latent system dynamics and constructs (x_m, Q_{anc}). Stage II learns the switching regime layer conditional on these fixed anchor objects. After fitting, the model can also be used for anchor-based diagnostics, but these diagnostics are separate from the estimator.

3.1 Stage I: fitting latent dynamics and computing anchor values

Stage I estimates the continuous-time latent dynamics from the observation sequence, treating actions as inputs to the system dynamics. Let Θ_{dyn} denote the parameters of the CTMC generators, outcome-proxy dynamics, emission model, and initial latent-state distribution. We fit Θ_{dyn} by maximum likelihood:

$$\widehat{\Theta}_{\text{dyn}} \in \arg\max_{\Theta_{\text{dyn}}} \sum_{i=1}^{N} \log p_{\Theta_{\text{dyn}}}\big(o_{1:M_i}^i \mid a_{1:M_i-1}^i, t_{1:M_i}^i\big)\,.$$

This objective models the non-choice observations. It is not an action-prediction objective; action prediction is handled by Stage II.

In practice, we use an EM-style routine. The E-step filters and smooths the latent trajectory $z(t)$. The M-step updates the transition generators, outcome dynamics, and emission parameters. After fitting, we compute pre-action states $\widehat{x}_m^i = (\widehat{b}_m^i, \widehat{\lambda}_m^i)$ at each decision epoch and evaluate $\widehat{Q}_{\text{anc}}(\widehat{x}_m^i, a)$ for all candidates. Anchor evaluation uses CTMC belief propagation and ODE updates over the fixed clinical horizon. Both outputs remain fixed in Stage II.

3.2 Stage II: fitting switching behavioral regimes

Stage II estimates the latent regime process and one action-dependent cost for each regime. With a fixed normalized dictionary shared across regimes,

$$C_u(\xi, a) = w_u^\top g(\xi, a), \qquad u = 1, \ldots, K_u. \tag{4}$$

For the clinical transition kernel, the regime-layer parameters are

$$\Theta_{\text{reg}} = \{\mu, \pi, \tau, w_1, \ldots, w_{K_u}\},$$

where μ is the initial regime distribution, π is the stationary distribution, and τ is selected on validation data. Given the fixed Stage I outputs, the action log-likelihood is

$$\mathcal{L}(\Theta_{\text{reg}}) = \sum_{i=1}^{N} \log \sum_{u_{1:M_i}^i} \mu_{u_1^i} \prod_{m=2}^{M_i} \Pi_{u_{m-1}^i, u_m^i}(\Delta_{m-1}^i) \prod_{m=1}^{M_i} p_{\Theta_{\text{reg}}}(a_m^i \mid \widehat{x}_m^i, \widehat{\xi}_m^i, u_m^i),$$

evaluated sequentially at pre-action histories using the emissions in equation 2.

The clinical fitting routine alternates forward–backward inference and regularized parameter updates. For each candidate τ, forward–backward computes $\gamma_m^i(u) = \Pr(u_m^i = u \mid \mathcal{D}^i)$. We update μ and π from smoothed initial and overall posterior occupancies, respectively, and take Adam steps on the posterior-weighted cost objective

$$\max_{\{w_u\}} \frac{1}{n} \sum_{i,m} \sum_{u=1}^{K_u} \gamma_m^i(u) \log p_{\Theta_{\text{reg}}}(a_m^i \mid \widehat{x}_m^i, \widehat{\xi}_m^i, u) - \sum_{u=1}^{K_u} \left(\frac{\ell_2}{2} \|w_u\|_2^2 + \ell_1 \|w_u\|_1 \right), \qquad n = \sum_i M_i.$$

Here $\ell_2, \ell_1 \geq 0$ control regularization; sparsity is not required. The occupancy updates are moment updates for the gap-dependent kernel, so this is an approximate EM-style routine. Appendix K specifies these updates and the selected settings.

The single-regime $K_u = 1$ model provides one global cost profile. Comparing it with $K_u > 1$ assesses the value of heterogeneous costs; comparison with an IID mixture isolates the contribution of regime persistence.

3.3 Scope of the theoretical results

Conditional identification and cost recovery. With fixed Q_{anc}, β, and g, regime kernels identifiable up to a common label permutation determine within-regime cost contrasts:

$$\log \frac{p(a \mid x, \xi, u)}{p(a' \mid x, \xi, u)} = \beta \left[Q_{\mathrm{anc}}(x, a) - Q_{\mathrm{anc}}(x, a') - \{C_u(\xi, a) - C_u(\xi, a')\} \right].$$

Coefficients are identified only along feature-difference directions; grouped allocations require additional conditions. Appendix B.3 also characterizes the anchor's feature-space role and gives a sufficient condition for improved cost recovery over post-hoc anchoring. Section 4.2 examines conditional reproducibility under the prescribed fitting rule.

Stability of longitudinal inference. Appendix B.4 extends the fixed-parameter Stage I plug-in bound to regime posteriors and full-mixture action probabilities, accounting for perturbations in fitted costs and transitions. Margin conditions then ensure stable action rankings, cost-contribution signs, and review priorities (Appendix B.7).

Persistence and predictive information. Switching yields a positive population likelihood gap when a single-regime model cannot represent the true action law within the switching class. Appendix B.6 quantifies how persistence retains predictive information across decision gaps and how action-cost separation supports regime distinction within persistent intervals, explaining their joint role in recovery.

3.4 Anchor-based diagnostics after fitting

After Stages I and II are fitted, the model can be used for descriptive diagnostics. These diagnostics are not part of training.

One diagnostic is the anchor-reference action

$$a_m^Q \in \arg\max_{a \in \mathcal{A}} Q_{\mathrm{anc}}(x_m, a),$$

which provides a reference for comparing the observed action and the fitted top action $\hat{a}_m$. Another diagnostic modifies selected components of C_u, such as setting a feature group to zero or scaling a subset of deviation weights, and recomputes the implied action probabilities. These analyses describe how learned regimes affect action selection under the fitted model.

All such diagnostics are conditional on the learned dynamics, anchor reward, and feature map. They are useful for summarizing regimes and inspecting trajectories.

4 Experiments

We evaluate regime and action-cost recovery in synthetic data and persistent treatment patterns in long-horizon CKD–MBD care and high-frequency MIMIC–IV treatment (Johnson et al., 2024; 2023). Clinical analyses link these patterns to physiological phenotypes and anchor-relative cost contrasts; held-out prediction checks behavioral fit. Appendices E–G detail the clinical datasets, anchor, and cost features.

4.1 Controlled Recovery of the Q–C–Action Mechanism

We test recovery under resource-pressure, inertia, and clinical-guardrail regimes, which penalize intensity, treatment changes, and under- or overtreatment across four treatment levels (Appendices D.1–D.3). Actions follow equation 2; across ten seeds, costs change the anchor-preferred action

in $45.5\% \pm 6.5\%$ of held-out decisions. The known anchor improves cost recovery over no-anchor switching, while persistence improves regime recovery over the independent mixture (Table 1).

Perturbing the anchor barely changes NLL (0.809 versus 0.811) but raises cost RMSE to 0.808, separating behavioral fit from cost recovery. Appendix D.4 also examines cost separation and persistence. With a learned CTMC anchor, during-fit anchoring improves cost RMSE from 0.547 to 0.125 and cost-direction accuracy from 85.1% to 94.9% over post-hoc anchoring of an unanchored GLM-HMM (Appendix D.5). Both use the same dictionary, unrestricted signed coefficients, and fitting budget; gains depend on anchor quality and the specified feature class.

Table 1: Controlled synthetic recovery on held-out trajectories (mean $\pm$ SD across ten seeds). Cost RMSE measures action-centered cost recovery; regime accuracy uses aligned labels and ARI is label invariant.

Model	NLL ↓	Cost RMSE ↓	Action acc. ↑	Regime acc. ↑	ARI ↑
Anchor only ($C \equiv 0$)	1.120 ± 0.054	–	0.528 ± 0.033	–	–
Global cost ($K_u = 1$)	0.885 ± 0.038	–	0.637 ± 0.027	–	–
No anchor ($Q \equiv 0$, $K_u = 3$)	0.826 ± 0.039	0.470 ± 0.060	0.646 ± 0.027	0.785 ± 0.073	0.498 ± 0.109
Independent mixture ($K_u = 3$)	0.880 ± 0.036	0.471 ± 0.185	0.636 ± 0.029	0.475 ± 0.037	0.078 ± 0.026
Switching cost ($K_u = 3$)	$\mathbf{0.809 \pm 0.039}$	$\mathbf{0.068 \pm 0.018}$	$\mathbf{0.659 \pm 0.026}$	$\mathbf{0.824 \pm 0.039}$	$\mathbf{0.555 \pm 0.083}$

4.2 Interpreting treatment choices through anchor-relative costs

Figure 2 decomposes action-ranking reversals in two held-out decisions. Anchor probabilities satisfy $p_Q(a \mid x_m) \propto \exp\{\beta Q_{\text{anc}}(x_m, a)\}$, and displayed costs average over pre-action regimes: $\bar{C}_m(a) = \sum_u \rho^-_{m,u} C_u(\xi_m, a)$. Final probabilities use the regime-specific softmax mixture in equation 3.

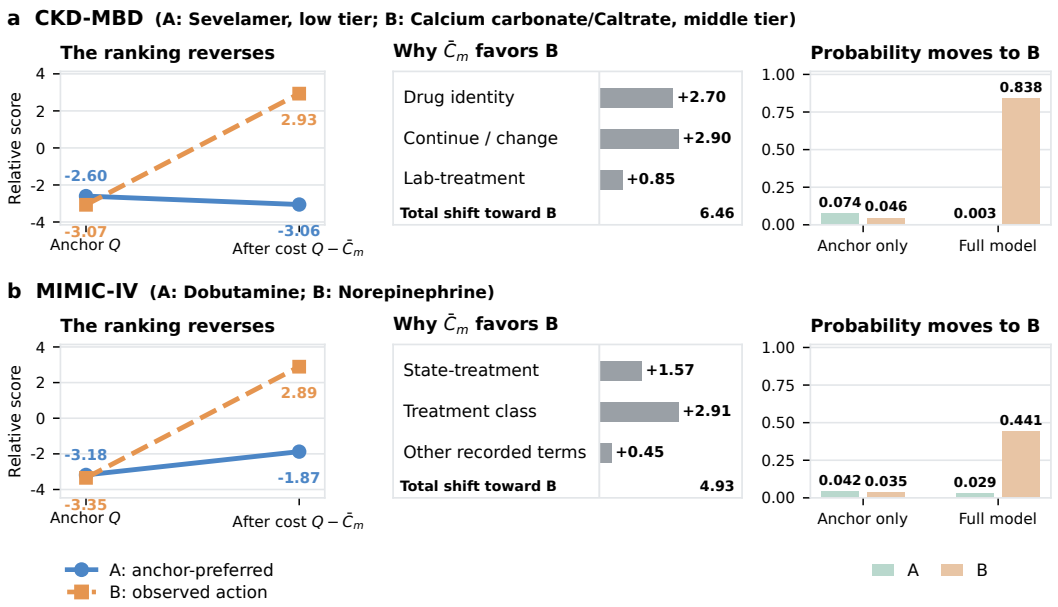


Figure 2: Two held-out decisions where the anchor favors A over observed B. Left: scores before and after subtracting regime-averaged costs $\bar{C}_m$. Middle: contributions to $\bar{C}_m(A) - \bar{C}_m(B)$; positive values favor B. Right: anchor and full pre-action mixture probabilities.

CKD–MBD: continuing an established regimen. With low calcium, high phosphorus, and low PTH, the anchor favors low-tier Sevelamer (A) over observed middle-tier calcium carbonate (B). Treatment identity, continuity, and lab–treatment interactions lower B's inferred cost, favoring continuation.

MIMIC–IV: choosing vasopressor support. With elevated lactate and low urine output, the six-hour anchor marginally favors Dobutamine (A) over observed Norepinephrine (B). Acute-state–treatment and treatment-class effects lower B's inferred cost; the full mixture ranks it first.

Feature contributions depend on the anchor, representation, scaling, and regularization; dependent features can redistribute them without changing total cost contrasts (Appendix B.3). With Stage I, the anchor, and representation fixed, principal contribution directions persist in all three clinical cases across 40 fits per case (20 patient/subject bootstrap resamples, two initializations), as well as regularization and fitting-budget checks. Appendix G.1 gives these checks and tests alternative representations.

4.3 Interpreting Physiological Phenotypes and Behavioral Regimes

Stage I physiological phenotypes. Figure 3 shows three phenotypes per dataset. In CKD–MBD, z_1 is predominantly PTH-normal; z_2 has low PTH and the longest dwell time; z_3 has lower calcium, higher phosphorus, shorter duration, and lower prevalence. In MIMIC–IV, z_1 is more acute, with higher lactate and lower blood pressure; z_2 and z_3 are intermediate and less acute.

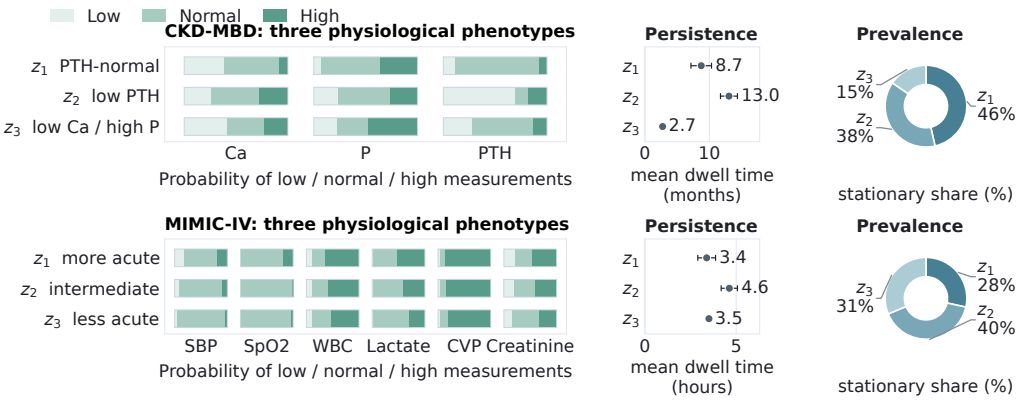


Figure 3: Stage I phenotypes. Cells: low/normal/high probabilities at each state's ODE equilibrium. Points and donuts: mean dwell time and stationary share from the CTMC at the training mean action-feature vector. Means over five splits aligned to the first by Euclidean distance between full emission profiles; dwell-time bars show SD.

Stage II behavioral regimes. Figure 4 compares regimes through the relative costs of treatment continuation, expansion, and drug choice. All regimes use the same feature dictionary within each dataset. In CKD–MBD, u_1's lower relative continuation cost and higher regimen-size cost favor maintaining a compact regimen; u_2's lowest costs for retaining and adding drugs favor extending it. In MIMIC–IV, u_1's strongest penalties on additions and treatment count, together with a lower continuation cost, favor maintaining existing support with limited expansion.

Appendix H examines joint and single-layer changes in the inferred phenotype and regime labels.

4.4 Predictive Performance and Model Robustness

Predictive performance of Stages I and II. Stage I models physiological evolution under observed treatments for action valuation and longitudinal clinical audit. Across four forecasting settings, it achieves the lowest mean NLL (Figure 5) and Brier score and the highest mean category accuracy, with competitive classwise calibration (Appendix J.2). Stage II achieves the lowest mean NLL (0.934 on CKD–MBD; 0.654 on MIMIC–IV) and changed-action NLL among predictors (Appendix J.1), and the lowest mean NLL among anchor and switching controls. Relative to fixed trajectory classes that also update class beliefs after each action, switching lowers NLL by 0.081 and 0.048, respectively. It also outperforms fixed classes on all five CKD–MBD splits when Stage II uses only recorded prescriptions (Appendix J.3).

Robustness to regime count and anchor specification. Among $K_u = 2, \ldots, 5$, the selected four-regime CKD–MBD and three-regime MIMIC–IV models combine near-best validation likelihood with favorable occupancy, separation, and parsimony (Appendix J.4). Varying rewards, continuation, and Stage I values changes mean held-out NLL by at most 0.006, preserves at least 99.1% of fitted top actions, and yields highly similar aligned regime profiles (Appendix J.5).

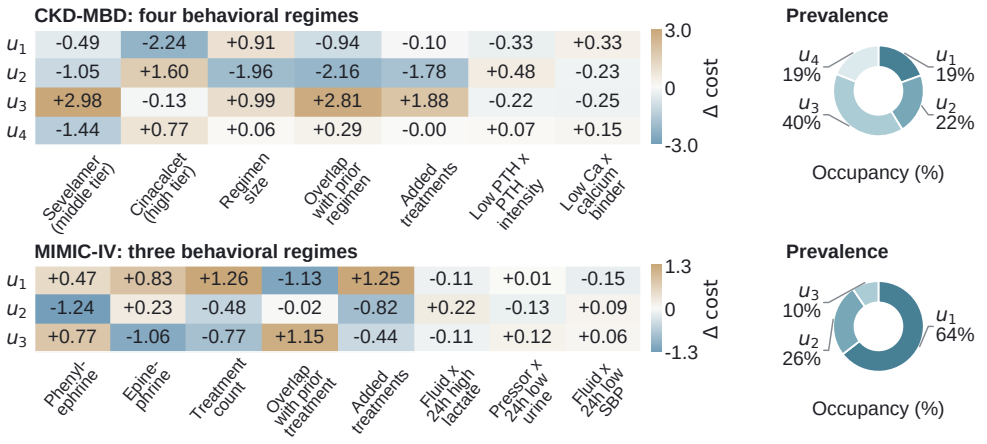


Figure 4: Stage II regimes. Heatmaps: selected effects $R_{u,j} = w_{u,j}\bar{\Delta}_j$ in model-score units, centered across regimes; $\bar{\Delta}_j$ is the mean training candidate-action feature range. Means over five splits aligned to the first by cosine similarity of full coefficient vectors. Donuts: mean held-out pre-action regime probabilities. Transition-memory scale τ: 18 or 36 months (CKD–MBD), 96 hours (MIMIC–IV).

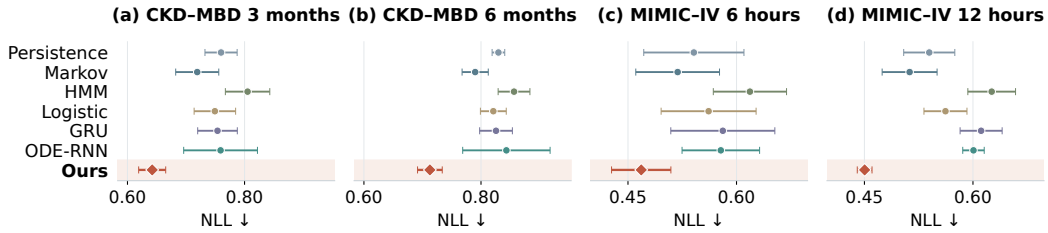


Figure 5: Held-out Stage I physiological forecasting: NLL. Points show means and whiskers show ± 1 SD; axes are shared across horizons within each dataset.

4.5 Anchor-Relative Decision Diagnostics and Clinical Auditing

Stage I reconstructs physiological evolution from irregular observations and provides a shared action-value reference. Stage II describes persistent treatment patterns through anchor-relative action costs. For longitudinal clinical audit, we compare the observed action a_m, anchor reference $a_m^Q = \arg\max_a Q_{\text{anc}}(x_m, a)$, and fitted top action $\hat{a}_m = \arg\max_a p(a \mid \mathcal{H}_m^-)$, alongside pre-action phenotypes and regimes. The cost contrast $\bar{C}_m(a_m^Q) - \bar{C}_m(\hat{a}_m)$ identifies treatment features favoring the fitted choice over the reference.

In the held-out CKD–MBD case (Figure 7, Appendix I.1), the inferred phenotype remains stable from month 2, while treatments and regimes change. At month 5, phosphorus remains high: the anchor favors combination treatment, while observed and fitted actions retain calcium-carbonate monotherapy. Continuation and treatment identity dominate the cost contrast. Under the two reward reweightings, retention of the primary 20% of decisions with the largest anchor-value gaps is 92.9–94.4% in CKD–MBD and 96.4–99.6% in MIMIC–IV. Month-5 contribution directions persist under all four anchor alternatives (Appendix I.2). Appendix I.3 extends the analysis to SCM-based alternative trajectories.

5 Conclusion

We introduced a framework for representing persistent behavioral heterogeneity through switching, action-dependent cost profiles. A shared value reference and common feature dictionary make regime differences comparable through explicit action-cost contrasts. Synthetic experiments support recovery of known costs and regimes, while fixed-class comparisons show the additional predictive value of switching in both clinical settings. By linking physiological trajectories to regime-specific action-cost contrasts, the framework supports longitudinal clinical audit.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript by improving the clarity and readability of the writing. We also used these tools to help identify relevant literature and to support the planning and refinement of experimental designs. In addition, we used them for code debugging, including identifying potential errors and exploring possible fixes. We have reviewed all AI-assisted code, analyses, and manuscript text for accuracy and consistency. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Abdul Fatir Ansari, Alvin Heng, Andre Lim, and Harold Soh. Neural continuous-discrete state space models for irregularly-sampled time series. In *International Conference on Machine Learning*, pp. 926–951. PMLR, 2023.

Zoe C Ashwood, Nicholas A Roy, Iris R Stone, International Brain Laboratory, Anne E Urai, Anne K Churchland, Alexandre Pouget, and Jonathan W Pillow. Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience*, 25(2):201–212, 2022.

RICHARD BELLMAN. A markovian decision process. *Journal of Mathematics and Mechanics*, 6 (5):679–684, 1957. ISSN 00959057, 19435274. URL http://www.jstor.org/stable/24900506.

Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela Van Der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083*, 2020.

Michael Bloem and Nicholas Bambos. Ground delay program analytics with behavioral cloning and inverse reinforcement learning. *Journal of aerospace information systems*, 12(3):299–313, 2015.

Elynn Y Chen, Rui Song, and Michael I Jordan. Reinforcement learning in latent heterogeneous environments. *Journal of the American Statistical Association*, 119(548):3113–3126, 2024.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *International conference on machine learning*, pp. 1843–1854. PMLR, 2020.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M Coopersmith, Craig French, Flávia R Machado, Lauralyn Mcintyre, Marlies Ostermann, Hallie C Prescott, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Critical care medicine*, 49(11):e1063–e1143, 2021.

Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 2017.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Zoubin Ghahramani and Geoffrey E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):831–864, 2000. doi: 10.1162/089976600300015619.

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.

Mansura Habiba and Barak A Pearlmutter. Neural ordinary differential equation based recurrent neural network model. In *2020 31st Irish signals and systems conference (ISSC)*, pp. 1–6. IEEE, 2020.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

James D Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, pp. 357–384, 1989.

Athul Paul Jacob, Abhishek Gupta, and Jacob Andreas. Modeling boundedly rational agents with latent inference budgets, 2023. URL https://arxiv.org/abs/2312.04030.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pp. 652–661. PMLR, 2016.

A Johnson, L Bulgarelli, T Pollard, B Gow, B Moody, S Horng, LA Celi, and R Mark. Mimic-iv (version 3.1). physionet, 2024.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Kenneth Jung and Nigam H. Shah. Implications of non-stationarity on predictive modeling using ehrs. *Journal of Biomedical Informatics*, 58:168–174, 2015. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2015.10.006. URL https://www.sciencedirect.com/science/article/pii/S1532046415002282.

Markus Ketteler, Geoffrey A Block, Pieter Evenepoel, Masafumi Fukagawa, Charles A Herzog, Linda McCann, Sharon M Moe, Rukshana Shroff, Marcello A Tonelli, Nigel D Toussaint, et al. Executive summary of the 2017 kdigo chronic kidney disease–mineral and bone disorder (ckd-mbd) guideline update: what's changed and why it matters. *Kidney international*, 92(1):26–36, 2017.

Taylor W Killian, Samuel Daulton, George Konidaris, and Finale Doshi-Velez. Robust and efficient transfer learning with hidden parameter markov decision processes. *Advances in neural information processing systems*, 30, 2017.

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.

Erwan Lecarpentier and Emmanuel Rachelson. Non-stationary markov decision processes, a worst-case approach using model-based reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Sharon M Moe, Tilman Drüeke, Norbert Lameire, and Garabed Eknoyan. Chronic kidney disease–mineral-bone disorder: a new paradigm. *Advances in chronic kidney disease*, 14(1):3–12, 2007.

Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In Pat Langley (ed.), *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pp. 663–670. Morgan Kaufmann, 2000.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146, 2009a. doi: 10.1214/09-SS057. URL https://doi.org/10.1214/09-SS057.

Judea Pearl. *Causality*. Cambridge university press, 2009b.

Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment, 2017a. URL https://arxiv.org/abs/1711.09602.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, pp. 147–163. PMLR, 2017b.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, and Mihaela van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. *arXiv preprint arXiv:2206.08311*, 2022.

Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23(1):69–101, 1996.

Lantao Yu, Tianhe Yu, Chelsea Finn, and Stefano Ermon. Meta-inverse reinforcement learning with probabilistic context variables. *Advances in neural information processing systems*, 32, 2019.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A Related Work

Our framework is closely related to offline reinforcement learning for high-stakes sequential decision-making, where policies must be learned from logged behavior with limited exploration (Levine et al., 2020; Fu et al., 2020). In healthcare, MDP and deep RL models have been used to derive treatment policies from MIMIC sepsis trajectories (Komorowski et al., 2018; Raghu et al., 2017b;a), but subsequent work emphasizes that clinician actions are heterogeneous, confounded, and not necessarily an optimal gold standard (Gottesman et al., 2019). Off-policy evaluation provides tools for estimating policy value from observational logs (Dudík et al., 2011; Jiang & Li, 2016), and hidden-parameter MDPs model patient-level variation during transfer (Killian et al., 2017). Our framework differs in goal: rather than learning a single deployable treatment policy, it treats the behavior policy itself as data to be decomposed into a shared anchor and switching, anchor-relative deviations.

Our regime layer also connects to inverse reinforcement learning and heterogeneous behavior modeling. Classical IRL and maximum-entropy IRL infer rewards that rationalize demonstrations (Ng & Russell, 2000; Ziebart et al., 2008), while later work handles diverse demonstrations through latent contexts, mixtures, or heterogeneous environments (Bloem & Bambos, 2015; Yu et al., 2019; Chen et al., 2024). Cooperative IRL and bounded-rational-agent models further recognize that observed behavior can reflect uncertainty, limited inference, or interaction between human and algorithmic objectives (Hadfield-Menell et al., 2016; Jacob et al., 2023). GLM-HMMs model persistent, switching decision strategies through state-specific choice coefficients (Ashwood et al., 2022). Stage II is a switching choice model with a value offset constructed from learned physiological dynamics under a prespecified reward and continuation rule. We hold this offset fixed during fitting and restrict anchor-relative adjustments to a prespecified feature class. Costs derived after fitting an unanchored GLM-HMM need not lie in this feature class, so the two approaches impose different constraints on learned behavior. This exposes how treatment features adjust action rankings and supports comparisons of observed and fitted decisions with the same reference, aligning with calls for interpretable models in high-stakes settings (Rudin, 2019; Frosst & Hinton, 2017).

Our framework builds on work on nonstationarity, regime switching, continuous-time latent dynamics, and counterfactual evaluation. Concept drift and nonstationary MDPs show why fixed decision rules can fail when data-generating processes evolve (Widmer & Kubat, 1996; Jung & Shah, 2015; Lecarpentier & Rachelson, 2019; Cheung et al., 2020); switching state-space models provide a classical latent-regime mechanism for piecewise-stationary structure (Hamilton, 1989; Ghahramani & Hinton, 2000). For irregularly sampled clinical records, latent ODEs and neural continuous-discrete state-space models learn dynamics without forcing observations onto an artificial grid (Rubanova et al., 2019; Habiba & Pearlmutter, 2020; Ansari et al., 2023). Counterfactual time-series methods and structural causal models provide semantics for asking how outcomes or values would change under alternative interventions (Pearl, 2009b;a; Bica et al., 2020; Seedat et al., 2022). The proposed framework couples continuous-time physiological dynamics with switching behavioral regimes. The dynamics support physiological forecasting and shared action valuation, while inferred states contextualize action-cost decompositions for longitudinal clinical audit.

B Theoretical Properties and Proofs

The analysis focuses on the Stage-II regime layer conditional on the Stage-I outputs. Throughout, the decision-time state is $x_m = (b_m, \lambda_m)$, the regime-layer context ξ_m collects the variables available to the regime layer (e.g., x_m itself and any optional side information), and the shared anchor value is $Q_{\mathrm{anc}}(x_m, a)$. For the proofs, every regime uses the same fixed action-dependent dictionary and has one complete linear cost, $C_u(\xi, a) = w_u^\top g(\xi, a)$, without a sparsity assumption.

We treat Q_{anc}, β, and g as fixed when analyzing identification and the population likelihood. Value scaling and β follow the experimental specifications.

B.1 Regime-layer setup

Let $\mathcal{A}$ be a finite action set with $|\mathcal{A}| \geq 2$, and let $K = K_u$ be the number of behavioral regimes. A regime $u_m \in \{1, \ldots, K\}$ evolves as a Markov chain with initial distribution μ and transition

kernel $\Pi(\Delta_{m-1})$. A constant kernel gives the homogeneous case; the clinical kernel is specified in equation 16. Conditional on the pre-action context (x_m, ξ_m) and $u_m = u$, the action emission model is

$$e_u(a \mid x, \xi) := p_\Theta(a \mid x, \xi, u) = \frac{\exp\{\beta[Q_{\mathrm{anc}}(x, a) - C_u(\xi, a)]\}}{\sum_{a' \in \mathcal{A}} \exp\{\beta[Q_{\mathrm{anc}}(x, a') - C_u(\xi, a')]\}}. \tag{5}$$

For a trajectory of length M, the sequential action likelihood is

$$L_\Theta(a_{1:M}; x_{1:M}, \xi_{1:M}, t_{1:M}) = \sum_{u_{1:M}} \mu_{u_1} \prod_{m=2}^{M} \Pi_{u_{m-1}, u_m}(\Delta_{m-1}) \prod_{m=1}^{M} e_{u_m}(a_m \mid x_m, \xi_m). \tag{6}$$

The contexts are evaluated at the available histories $\mathcal{H}_m^-$; equation 6 equals $\prod_m p_\Theta(a_m \mid \mathcal{H}_m^-)$, with the regime filter updated after each action. In the linear model, $C_u(\xi, a) = w_u^\top g(\xi, a)$, where $g : \Xi \times \mathcal{A} \to \mathbb{R}^d$ is the fixed action-dependent feature map and $w_u \in \mathbb{R}^d$ is the complete coefficient vector for regime u.

Two parameter values are considered equivalent if they induce the same conditional action laws on the support of pre-action histories, up to a permutation of regime labels and action-independent offsets in the deviations. Such offsets are unavoidable: replacing $C_u(\xi, a)$ by $C_u(\xi, a) + \kappa_u(\xi)$, where κ_u does not depend on a, does not change the softmax probabilities.

B.2 Entropy-regularized interpretation

Proposition 1 (Entropy-regularized anchor-minus-cost)**.** *Fix (x, ξ, u) and define*

$$S_u(x, \xi, a) = Q_{\mathrm{anc}}(x, a) - C_u(\xi, a).$$

Then the softmax distribution in equation 5 is the unique maximizer of

$$\max_{\pi \in \Delta(\mathcal{A})} \left\{ \sum_{a \in \mathcal{A}} \pi(a) S_u(x, \xi, a) + \frac{1}{\beta} H(\pi) \right\}, \qquad H(\pi) = -\sum_{a \in \mathcal{A}} \pi(a) \log \pi(a). \tag{7}$$

Proof. The objective is strictly concave because $\beta > 0$. At an interior optimum, the first-order condition with simplex multiplier ν is

$$S_u(x, \xi, a) - \frac{1}{\beta}\{1 + \log \pi(a)\} + \nu = 0.$$

Thus $\pi(a) \propto \exp\{\beta S_u(x, \xi, a)\}$. Normalizing gives equation 5, a positive distribution satisfying the first-order condition; strict concavity makes it the unique maximizer. □

B.3 Conditional identification and cost recovery

We characterize which cost quantities are determined by identifiable regime kernels. Establishing kernel identifiability for general history-dependent contexts is outside this result.

Assumption 1 (Regime-kernel identifiability)**.** On the support of pre-action histories and decision gaps, the observable action law identifies the initial distribution μ, transition kernel $\Pi(\cdot)$, and emission kernels $\{e_u\}_{u=1}^K$ up to one common permutation of regime labels.

Assumption 2 (Positivity and boundedness)**.** The parameter space is compact and the induced emissions satisfy $e_u(a \mid x, \xi) \geq \underline{p} > 0$ for all u, a and all contexts in the support.

Assumption 3 (Full rank for coefficient identification)**.** For the linear model, define

$$\mathcal{S} = \mathrm{span}\left\{ g(\xi, a) - g(\xi, a') : (\xi, a, a') \text{ in the support} \right\},$$

and let $\mathcal{N} = \mathcal{S}^\perp$. Full coefficient identification additionally assumes $\mathcal{S} = \mathbb{R}^d$; the contrast and projection results below do not require this condition.

Theorem 2 (Conditional identification of cost contrasts)**.** *Under Assumptions 1–2, with fixed Q_{anc}, $\beta > 0$, and g, each cost contrast $C_u(\xi, a) - C_u(\xi, a')$ is identified on the support, up to the common regime-label permutation. In the linear model, two coefficient vectors induce identical emissions if and only if their difference lies in $\mathcal{N}$. Thus the projection of w_u onto $\mathcal{S}$ is identified; Assumption 3 identifies the complete vector.*

For a fixed action pair, let d_J equal $g(\xi, a) - g(\xi, a')$ on feature group J and zero elsewhere. The grouped contribution $w_u^\top d_J$ is invariant under all coefficient shifts in $\mathcal{N}$ if and only if $d_J \in \mathcal{S}$.

Proof. Align regime labels using Assumption 1. For fixed u, x, ξ, the log-odds in equation 5 satisfy

$$\log \frac{e_u(a \mid x, \xi)}{e_u(a' \mid x, \xi)} = \beta \left[Q_{\mathrm{anc}}(x, a) - Q_{\mathrm{anc}}(x, a') - \{C_u(\xi, a) - C_u(\xi, a')\} \right]. \tag{8}$$

Since the emission law is identified and Q_{anc}, β are fixed, the cost contrast is identified. For linear costs, equality of all such contrasts is equivalent to

$$(w_u - w'_u)^\top \{g(\xi, a) - g(\xi, a')\} = 0$$

throughout the support, hence $w_u - w'_u \in \mathcal{N}$. Conversely, any shift in $\mathcal{N}$ adds only an action-independent cost offset and preserves the emissions. Finally, a shift $v \in \mathcal{N}$ changes the grouped contribution by $v^\top d_J$, which vanishes for every such v exactly when $d_J \in \mathcal{N}^\perp = \mathcal{S}$. □

For fixed posterior weights, positive L2 makes the linear coefficient objective strictly concave, yielding a unique exact update even with dependent features and optional L1. The full latent-model objective remains nonconvex. Grouped allocations depend on the dictionary, scaling, and fitting rule; the case checks assess conditional reproducibility rather than mechanism identification.

Shared-anchor geometry and cost recovery. Write $z = (x, \xi)$ and center functions across the $A = |\mathcal{A}|$ candidate actions: $\mathsf{C}h(z, a) = h(z, a) - A^{-1} \sum_b h(z, b)$. For a context distribution ν, use

$$\|h\|_\nu^2 = \mathbb{E}_\nu \left[A^{-1} \sum_a h(z, a)^2 \right], \qquad \mathcal{G} = \{w^\top \mathsf{C} g : w \in \mathbb{R}^d\}.$$

Assume these features are square integrable. The finite-dimensional function space $\mathcal{G}$ is closed even with dependent features; let $P_{\mathcal{G}}$ be its orthogonal projection. All functions below are centered, and regime labels are aligned.

Theorem 3 (Anchor geometry and cost recovery)**.** *For unrestricted signed coefficients, the anchored score class $q - \mathcal{G}$ and the unanchored class $\mathcal{G}$ induce the same emission class if and only if $q \in \mathcal{G}$. For a target cost $c_u^\star \in \mathcal{G}$, any learned anchor $\widehat{q}$, and any unanchored score $f_u \in \mathcal{G}$, the post-hoc cost $c_u^{\mathrm{post}} = \widehat{q} - f_u$ satisfies*

$$\|c_u^{\mathrm{post}} - c_u^\star\|_\nu^2 = \|(I - P_{\mathcal{G}})\widehat{q}\|_\nu^2 + \|P_{\mathcal{G}}\widehat{q} - f_u - c_u^\star\|_\nu^2, \tag{9}$$

$$\|c_u^{\mathrm{post}} - c_u^\star\|_\nu \geq \mathrm{dist}(\widehat{q}, \mathcal{G}). \tag{10}$$

The lower bound is sharp over $f_u \in \mathcal{G}$.

Let $e_u^\star$ and $\widehat{e}_u$ have scores $q^\star - c_u^\star$ and $\widehat{q} - \widehat{c}_u$, where $\widehat{c}_u \in \mathcal{G}$, and suppose both emission distributions have coordinates at least $p_0 > 0$. With

$$R_u = \mathbb{E}_\nu \, \mathrm{KL}(e_u^\star \| \widehat{e}_u), \qquad \varepsilon_Q = \|\widehat{q} - q^\star\|_\nu,$$

the during-fit cost obeys

$$\|\widehat{c}_u - c_u^\star\|_\nu \leq \varepsilon_Q + \sqrt{\frac{2R_u}{\beta^2 p_0 A}}.$$

If this upper bound is smaller than $\mathrm{dist}(\widehat{q}, \mathcal{G})$, during-fit anchoring has smaller cost error than every post-hoc cost above.

Proof. Softmax with fixed positive β is injective on centered score vectors, since equal log odds imply an action-independent score difference. Thus emission-class equality is equivalent to $q - \mathcal{G} = \mathcal{G}$, or $q \in \mathcal{G}$. Orthogonal decomposition gives equation 9; choosing $f_u = P_{\mathcal{G}}\widehat{q} - c_u^\star$ attains the floor.

For $F(v) = \log \sum_a e^{v_a}$, the KL divergence between softmax distributions is the corresponding Bregman divergence. Probabilities along the line between their logits are normalized geometric means of the endpoint probabilities. Their normalizer is at most one by Hölder's inequality, so they remain at least p_0. For a centered vector v,

$$v^\top \nabla^2 F \, v = \min_b \sum_a p_a (v_a - b)^2 \geq p_0 \min_b \sum_a (v_a - b)^2 = p_0 \|v\|_2^2.$$

Integration along the logit segment gives

$$R_u \geq \frac{\beta^2 p_0 A}{2} \|(\widehat{q} - \widehat{c}_u) - (q^\star - c_u^\star)\|_\nu^2.$$

The triangle inequality separates the anchor and score errors, proving the cost bound. □

This comparison concerns centered costs under a common dictionary and scale. The floor remains valid for bounded coefficient sets, while class equivalence additionally requires those sets to correspond under the anchor shift. Identical coefficient penalties need not correspond under this shift. The error R_u is an aligned component divergence, not the mixture NLL. An anchor error $\delta q \in \mathcal{G}$ can be absorbed into every cost as $c_u^\star + \delta q$, leaving emissions unchanged; this explains why similar behavioral fit can accompany different recovered costs.

B.4 Stability of regime inference

The Stage-II likelihood uses plug-in estimates from Stage I. We show that moderate error in these plug-in quantities induces only a controlled perturbation of the action likelihood.

Lemma 4 (Log-sum-exp Lipschitzness). *For any $r, s \in \mathbb{R}^{|\mathcal{A}|}$,*

$$\left| \log \sum_{a \in \mathcal{A}} e^{r_a} - \log \sum_{a \in \mathcal{A}} e^{s_a} \right| \le \|r - s\|_\infty.$$

Proof. Let $m = \|r - s\|_\infty$. Then $s_a - m \le r_a \le s_a + m$ for all a. Hence $\sum_a e^{s_a - m} \le \sum_a e^{r_a} \le \sum_a e^{s_a + m}$. Taking logs gives the result. □

Let $q_m(a) = Q_{\mathrm{anc}}(x_m, a)$ and $\widehat{q}_m(a) = \widehat{Q}_{\mathrm{anc}}(\widehat{x}_m, a)$. Suppose

$$\epsilon_Q = \sup_{m,a} |\widehat{q}_m(a) - q_m(a)|, \qquad \epsilon_\xi = \sup_m \|\widehat{\xi}_m - \xi_m\|_1.$$

Theorem 5 (Plug-in stability of the Stage-II likelihood). *Assume $C_u(\xi, a)$ is L_C-Lipschitz in ξ, uniformly over u and a. Then for any fixed regime-layer parameters,*

$$\sup_{m,u} \left| \log \widehat{e}_u(a_m \mid \widehat{x}_m, \widehat{\xi}_m) - \log e_u(a_m \mid x_m, \xi_m) \right| \le 2\beta(\epsilon_Q + L_C \epsilon_\xi),$$

where $\widehat{e}_u$ is the emission model evaluated with the plug-in anchor values and plug-in contexts. For the linear model, if $\|w_u\|_1 \le W$ and $\|g(\xi, a) - g(\xi', a)\|_\infty \le L_g \|\xi - \xi'\|_1$, then one may take $L_C \le W L_g$.

Proof. Fix m, u and define the scaled action scores

$$r_a = \beta\{\widehat{q}_m(a) - C_u(\widehat{\xi}_m, a)\}, \qquad s_a = \beta\{q_m(a) - C_u(\xi_m, a)\}.$$

The anchor-error and Lipschitz assumptions give $\|r - s\|_\infty \le \beta(\epsilon_Q + L_C \epsilon_\xi)$. Each log emission is its observed-action score minus the log-sum-exp over candidates. Bounding these two terms separately with Lemma 4 yields

$$\left| \log \widehat{e}_u(a_m \mid \widehat{x}_m, \widehat{\xi}_m) - \log e_u(a_m \mid x_m, \xi_m) \right| \le 2\|r - s\|_\infty \le 2\beta(\epsilon_Q + L_C \epsilon_\xi).$$

Taking the supremum over m, u proves the claim. For linear costs,

$$|C_u(\widehat{\xi}, a) - C_u(\xi, a)| \le \|w_u\|_1 \|g(\widehat{\xi}, a) - g(\xi, a)\|_\infty \le W L_g \|\widehat{\xi} - \xi\|_1,$$

so $L_C \le W L_g$. □

If one instead separates anchor-function error from state-estimation error, the same theorem applies with $\epsilon_Q \le \epsilon_{\mathrm{val}} + L_Q \epsilon_x$, where $\epsilon_{\mathrm{val}} = \sup_{m,a} |\widehat{Q}_{\mathrm{anc}}(\widehat{x}_m, a) - Q_{\mathrm{anc}}(\widehat{x}_m, a)|$, Q_{anc} is L_Q-Lipschitz in x, and $\epsilon_x = \sup_m \|\widehat{x}_m - x_m\|_1$.

Longitudinal filtering stability. Consider two fitted models on the same record, with aligned regime labels and candidate actions. Let $P_m = \Pi(\Delta_{m-1})$, $\rho_1^- = \mu$, $\rho_m^- = \rho_{m-1}^+ P_m$, and $\rho_m^+ \propto \rho_m^- \odot e_m(a_m)$; hats denote the second model. Assume positive initial distributions, transitions, and emissions. For positive vectors define

$$d_H(v,w) = \operatorname{osc}_i \log(v_i/w_i), \qquad D(P) = \max_{i,j,k,l} \log \frac{P_{ik}P_{jl}}{P_{il}P_{jk}}, \qquad \kappa(P) = \tanh\{D(P)/4\},$$

where osc is maximum minus minimum. Put

$$\eta_m = \operatorname{osc}_u \log \frac{\widehat{e}_{m,u}(a_m)}{e_{m,u}(a_m)}, \qquad \zeta_m = \operatorname{osc}_{i,j} \log \frac{\widehat{P}_{m,ij}}{P_{m,ij}}, \qquad \kappa_m = \kappa(P_m),$$

and define

$$B_1^- = d_H(\widehat{\mu}, \mu), \qquad B_1^+ = B_1^- + \eta_1,$$
$$B_m^- = \kappa_m B_{m-1}^+ + \zeta_m, \qquad B_m^+ = B_m^- + \eta_m \quad (m \geq 2).$$

Theorem 6 (Stability of regime filtering and action prediction)**.** *For every decision, $d_H(\widehat{\rho}_m^{\pm}, \rho_m^{\pm}) \leq B_m^{\pm}$. If $\kappa_m \leq \bar{\kappa} < 1$, $\eta_m \leq \bar{\eta}$, and $\zeta_m \leq \bar{\zeta}$, then*

$$B_m^+ \leq \bar{\kappa}^{m-1}(B_1^- + \bar{\eta}) + (\bar{\zeta} + \bar{\eta})\frac{1 - \bar{\kappa}^{m-1}}{1 - \bar{\kappa}}.$$

Let $s_{m,u}(a) = q_m(a) - C_u(\xi_m, a)$, $\delta_m = \max_{u,a} |\widehat{s}_{m,u}(a) - s_{m,u}(a)|$, and

$$V_m = \tanh(B_m^-/4), \qquad T_m = \min\{1, V_m + \tanh(\beta\delta_m/2)\}.$$

The pre-action regime distributions and full action mixtures satisfy

$$\mathrm{TV}(\widehat{\rho_m^-}, \rho_m^-) \leq V_m, \qquad \mathrm{TV}(\widehat{p}_m, p_m) \leq T_m, \qquad \eta_m \leq 4\beta\delta_m.$$

Proof. For positive probability vectors, $\mathrm{TV}(v,w) \leq \tanh\{d_H(v,w)/4\}$. To see this, the ratio $Z_i = v_i/w_i$, weighted by w_i, has mean one. If $Z \in [a,b]$, the chord bound for $(Z-1)_+$ yields $\mathrm{TV} \leq (b-1)(1-a)/(b-a)$. Maximizing at fixed b/a gives the stated bound; equal vectors are immediate.

Also $d_H(vP, wP) \leq \kappa(P) d_H(v,w)$. Interpolate $v_i(t) = w_i \exp(t h_i)$, $h_i = \log(v_i/w_i)$. The derivative of $\log[(v(t)P)_j]$ is the mean of h_i under weights proportional to $v_i(t)P_{ij}$. Two columns' weighting distributions have log-ratio range at most $D(P)$; their means differ by at most $\kappa(P)\operatorname{osc}(h)$ by the preceding TV bound. Integration proves contraction.

Multiplication by a common positive emission vector and normalization preserve d_H; different emissions add at most η_m. Each coordinate ratio of $r\widehat{P}_m$ to rP_m is a weighted average of entrywise transition ratios, giving an additional distance at most ζ_m. Induction proves the recursion, and summation gives the uniform bound. Component softmax distributions have log-ratio range at most $2\beta\delta_m$, hence TV at most $\tanh(\beta\delta_m/2)$. Mixing adds at most the TV error in regime weights. Finally, Lemma 4 gives log-emission error at most $2\beta\delta_m$, and its range across regimes is at most $4\beta\delta_m$. □

For the clinical kernel, with $\alpha_m = e^{-\Delta_{m-1}/\tau}$, $\Delta_{m-1} > 0$, and $\pi_{\min} > 0$,

$$\kappa_m \leq \frac{\alpha_m}{\alpha_m + 2(1-\alpha_m)\pi_{\min}} < 1.$$

Indeed, row-ratio ranges bound every cross ratio by $[1 + \alpha_m/\{(1-\alpha_m)\pi_{\min}\}]^2$. Substitution into $\tanh(D/4)$ yields the inequality. Small gaps or long persistence can weaken contraction; the finite-record recursion does not require a uniform gap bound. At a zero gap, a common identity transition is nonexpansive with $\kappa_m = 1$.

Conditional stability of regularized cost updates. For one regime, let

$$F(w) = n^{-1}\sum_{i=1}^{n} \gamma_i[-\log e_w(a_i \mid q_i, g_i)] + \lambda_2\|w\|_2^2/2 + \lambda_1\|w\|_1,$$

where $0 \leq \gamma_i \leq 1$, $\lambda_2 > 0$, $\lambda_1 \geq 0$, and e_w uses scores $q_i - w^\top g_i$. Compare with $\widehat{F}$ under changed anchors, features, and responsibility weights, but the same coefficient coordinates and penalties. Suppose $\|g_i(a)\|_2 \leq G$, $\|\widehat{g}_i(a) - g_i(a)\|_2 \leq \varepsilon_g$, $\|\widehat{q}_i - q_i\|_\infty \leq \varepsilon_q$, and $\|w\|_2 \leq W_2$.

Proposition 7 (Regularized update perturbation). *Let $r, \widehat{r}$ be norms of valid subgradient residuals of F at w and $\widehat{F}$ at $\widehat{w}$, and let $D_\gamma = n^{-1}\sum_i |\widehat{\gamma}_i - \gamma_i|$. Then*

$$\|\widehat{w} - w\|_2 \le \frac{r + \widehat{r} + 2\beta G D_\gamma + 2\beta\varepsilon_g + 2\beta G \tanh[\beta(\varepsilon_q + W_2\varepsilon_g)/2]}{\lambda_2}.$$

Proof. The row-loss gradient is $\beta\{g_i(a_i) - \sum_a e_w(a) g_i(a)\}$, with norm at most $2\beta G$. At fixed w, the score perturbation is at most $\varepsilon_q + W_2\varepsilon_g$; the preceding TV bound gives gradient change at most $2\beta\varepsilon_g + 2\beta G \tanh[\beta(\varepsilon_q + W_2\varepsilon_g)/2]$. Splitting the weighted-gradient difference gives the numerator's perturbation terms. Both subdifferentials are λ_2-strongly monotone. Use the same L1 subgradient at w in the two objectives and compare with a subgradient at $\widehat{w}$; Cauchy–Schwarz yields the result. □

This bound retains responsibility changes and optimization residuals, rather than assuming global convergence of the latent-model fit. At an evaluation context, $\|g_m(a)\|_2 \le G_m$, feature error $\varepsilon_{g,m}$, coefficient error $\varepsilon_w = \max_u \|\widehat{w}_u - w_u\|_2$, and $\widehat{W}_2 = \max_u \|\widehat{w}_u\|_2$ give

$$\delta_m \le \max_a |\widehat{q}_m(a) - q_m(a)| + G_m\varepsilon_w + \widehat{W}_2\varepsilon_{g,m},$$

connecting the cost-update comparison to Theorem 6.

B.5 Consistency of the regime-layer MLE

Let

$$\ell_i(\Theta) = \log L_\Theta(a^i_{1:M_i}; x^i_{1:M_i}, \xi^i_{1:M_i}, t^i_{1:M_i})$$

denote the sequential action log-likelihood contribution for trajectory i. The following results concern global maximizers of the unpenalized likelihood, not the finite-budget regularized fits. For comparisons, $\mathrm{KL}_a(P^\star \| P_\Theta)$ denotes the expected sum of conditional action KL divergences over pre-action histories generated under $P^\star$.

Assumption 4 (Sampling and trajectory lengths). Trajectories $\{(x^i_{1:M_i}, \xi^i_{1:M_i}, a^i_{1:M_i})\}_{i=1}^N$ are independent across i and identically distributed, including their observation times and decision gaps, with $\mathbb{E}[M_i] < \infty$.

Assumption 5 (Compactness and continuity). The parameter space for the initial distribution, transition-kernel parameters, and complete linear coefficient vectors $\{w_u\}_{u=1}^K$, with fixed g, is compact. The map $\Theta \mapsto \ell_i(\Theta)$ is almost surely continuous and dominated by an integrable envelope.

Theorem 8 (Consistency of the regime-layer MLE). *Assume the Stage-II model is well specified, with true parameter $\Theta^\star$, and Assumptions 1, 4, and 5 hold. Then any global maximizer*

$$\widehat{\Theta}_N \in \arg\max_\Theta \sum_{i=1}^N \ell_i(\Theta)$$

converges to $\Theta^\star$ in the quotient space induced by common label permutations and feature-difference null directions $\mathcal{N}$.

Proof. By Assumptions 4 and 5, a uniform law of large numbers gives

$$\sup_\Theta \left| \frac{1}{N}\sum_{i=1}^N \ell_i(\Theta) - \mathbb{E}[\ell_1(\Theta)] \right| \to 0$$

almost surely. Under well specification, the population criterion is uniquely maximized at $\Theta^\star$ modulo the equivalence relation defined above. This uniqueness follows from Assumption 1 and Theorem 2. Standard argmax consistency for M-estimators then implies convergence of any global maximizer in the quotient space. □

Theorem 9 (Quasi-MLE under misspecification). *Under Assumptions 4 and 5, if the true conditional law is not contained in the Stage-II model class, then any global maximizer converges to the set of maximizers of $\mathbb{E}[\ell_1(\Theta)]$, equivalently the minimizers of $\mathrm{KL}_a(P^\star \| P_\Theta)$ over the model class.*

Proof. The same uniform convergence argument applies. The population objective can be written as a constant minus the expected cumulative conditional action divergence $\mathrm{KL}_a(P^\star \| P_\Theta)$. Hence its maximizers are exactly the KL projections. Argmax consistency gives convergence to this set. □

B.6 WHY SWITCHING REGIMES CAN IMPROVE FIT

Let $\mathcal{M}_1$ denote the single-regime model and $\mathcal{M}_K$ the switching-regime model with $K \geq 2$. The single-regime model is nested in the switching model.

Theorem 10 (Strict population improvement over one global deviation)**.** *Suppose the true conditional law $P^\star$ belongs to $\mathcal{M}_K$ and is not contained in the closure of $\mathcal{M}_1$. Then*

$$\sup_{\Theta \in \mathcal{M}_K} \mathbb{E}_{P^\star}[\ell_1(\Theta)] > \sup_{\theta \in \mathcal{M}_1} \mathbb{E}_{P^\star}[\ell_1(\theta)].$$

The gap equals the minimum KL_a divergence from $P^\star$ to $\mathcal{M}_1$, and is strictly positive.

Proof. Since $\mathcal{M}_1 \subset \mathcal{M}_K$, the left-hand side is at least the right-hand side. Because $P^\star \in \mathcal{M}_K$, the best value over $\mathcal{M}_K$ is achieved by the true conditional law. For any $\theta \in \mathcal{M}_1$,

$$\mathbb{E}_{P^\star}[\ell_1(\Theta^\star)] - \mathbb{E}_{P^\star}[\ell_1(\theta)] = \mathrm{KL}_a(P^\star \| P_\theta).$$

Taking the infimum over $\theta \in \mathcal{M}_1$ gives the stated gap. Since $P^\star$ is not in the closure of $\mathcal{M}_1$ and the parameter space is compact, this infimum is strictly positive. □

With the uniform convergence conditions from Assumptions 4 and 5, the empirical likelihood gap converges to the positive population gap. This justifies the empirical comparison between $K_u = 1$ and $K_u > 1$.

Predictive information retained by persistence. For $m \geq 2$, let $Z_m = (x_m, \xi_m, \Delta_{m-1})$ be the current emission context, E_m the matrix of regime-specific action probabilities, $R_m = \rho^+_{m-1}$, and $\overline{R}_m = \mathbb{E}[R_m \mid Z_m]$. Contexts may depend on past actions. Consider a law obeying the sequential action model with the clinical transition, so

$$p_m = \Pr(a_m = \cdot \mid \mathcal{H}^-_m) = \{\alpha_m R_m + (1 - \alpha_m)\pi\}E_m, \qquad \alpha_m = e^{-\Delta_{m-1}/\tau}.$$

Theorem 11 (History information and regime separation)**.** *The optimal predictor using only Z_m is*

$$p^0_m = \{\alpha_m \overline{R}_m + (1 - \alpha_m)\pi\}E_m, \qquad p_m - p^0_m = \alpha_m(R_m - \overline{R}_m)E_m.$$

Its excess expected log loss relative to the full-history predictor is

$$\begin{aligned} \mathcal{I}_m &= \mathbb{E}\,\mathrm{KL}(p_m \| p^0_m) = I(a_m; \mathcal{H}^-_m \mid Z_m) \\ &\geq \frac{1}{2}\mathbb{E}\left[e^{-2\Delta_{m-1}/\tau}\|(R_m - \overline{R}_m)E_m\|_1^2\right]. \end{aligned}$$

If component probabilities are at least $p_0 > 0$, then

$$\mathcal{I}_m \leq p_0^{-1}\mathbb{E}\left[e^{-2\Delta_{m-1}/\tau}\|(R_m - \overline{R}_m)E_m\|_2^2\right].$$

For two regimes, writing $v_m = \mathrm{TV}(e_{m,1}, e_{m,2})$, the lower bound becomes

$$\mathcal{I}_m \geq 2\mathbb{E}\left[e^{-2\Delta_{m-1}/\tau} v_m^2 \,\mathrm{Var}(R_{m,1} \mid Z_m)\right].$$

Proof. Condition on Z_m, under which α_m and E_m are measurable, and apply the tower property. The expected log-score difference between the two conditional laws is their expected KL, or conditional mutual information. Pinsker's inequality gives $\mathrm{KL}(p\|r) \geq \|p - r\|_1^2/2$. It follows by grouping coordinates where $p_a \geq r_a$, applying the log-sum inequality, and using the Bernoulli KL curvature $1/[x(1-x)] \geq 4$. For the upper bound, $\log t \leq t - 1$ gives $\mathrm{KL}(p\|r) \leq \sum_a (p_a - r_a)^2/r_a$, and $p^0_m(a) \geq p_0$. With two regimes, $(R - \overline{R})E = (R_1 - \overline{R}_1)(e_1 - e_2)$; conditioning on Z_m gives the variance expression. □

The gain is positive when history changes regime weights in a direction distinguishable through the current emissions. Long gaps attenuate this information. Persistence alone is insufficient if regimes have identical emissions or the current context already determines the regime posterior. The comparator is the best predictor using Z_m, not a predictor that also retains the full history; no assumption that π gives the context-conditional mixing weights is needed.

Regime distinction within persistent intervals. Consider L actions under a fixed-regime hypothesis u, with kernels $e_{t,u}(\cdot \mid Z_t)$. Contexts may adapt to previous actions; additional context observations have a common conditional mechanism under the competing hypotheses. Kernels are fixed, conditional on any independent training data. Let a classifier maximize $\omega_u L_u$, where $\omega_u > 0$ and $L_u = \prod_{t=1}^{L} e_{t,u}(a_t \mid Z_t)$. Define the predictable quantities

$$h_{t,uv} = -\log \sum_a \sqrt{e_{t,u}(a) e_{t,v}(a)}, \qquad H_{uv} = \sum_{t=1}^{L} h_{t,uv}.$$

Proposition 12 (Accumulation of cost-separation evidence)**.** *For any deterministic $H_0 \geq 0$,*

$$\Pr_u\{\omega_v L_v \geq \omega_u L_u,\ H_{uv} \geq H_0\} \leq \sqrt{\omega_v/\omega_u}\, e^{-H_0}.$$

If $H_{uv} \geq H^0_{uv}$ almost surely, the classification error is at most

$$\min\left\{1, \sum_{v \neq u} \sqrt{\omega_v/\omega_u}\, e^{-H^0_{uv}}\right\}.$$

For shared-anchor softmax emissions with centered costs and probabilities at least $p_0 > 0$,

$$h_{t,uv} \geq \frac{\beta^2 p_0}{8} \|c_{t,u} - c_{t,v}\|_2^2.$$

Proof. Under hypothesis u,

$$M_l = \exp\left(\sum_{t=1}^{l} h_{t,uv}\right) \sqrt{\prod_{t=1}^{l} \frac{e_{t,v}(a_t)}{e_{t,u}(a_t)}}$$

is a nonnegative mean-one martingale: conditioning before the next action makes the expected multiplier equal to one. Common context kernels cancel. On the stated event, $M_L \geq e^{H_0}\sqrt{\omega_u/\omega_v}$; Markov's inequality and then a union bound prove the error bounds. To connect separation to costs, set $z_u = \beta(q - c_u)$ and $F(z) = \log \sum_a e^{z_a}$. Direct calculation gives

$$h_{t,uv} = \frac{F(z_u) + F(z_v)}{2} - F\left(\frac{z_u + z_v}{2}\right) \geq \frac{p_0}{8}\|z_u - z_v\|_2^2,$$

using the Hessian bound in the proof of Theorem 3. Since the anchor is common, $z_u - z_v = -\beta(c_u - c_v)$. □

With estimated kernels satisfying uniform log-emission errors $d_{t,u}$, put $D_{uv} = \sum_t (d_{t,u} + d_{t,v})$. An incorrect fitted comparison implies $\log(L_v/L_u) \geq \log(\omega_u/\omega_v) - D_{uv}$, so each summand in the error bound becomes $\sqrt{\omega_v/\omega_u}\exp(-H^0_{uv} + D_{uv}/2)$. Use deterministic error bounds, or add the probability that such bounds fail.

For a fixed observation schedule independent of actions and latent transitions, starting in u, the clinical kernel gives probability

$$S_u = \prod_{t=2}^{L} [\pi_u + (1 - \pi_u) e^{-\Delta_{t-1}/\tau}]$$

that all sampled regime labels remain u. Under the common context mechanism above, if $0 \leq \mathcal{E}_u \leq 1$ bounds classification error on this event, endpoint regime error is at most

$$(1 - S_u) + S_u \mathcal{E}_u \leq \min\{1, (1 - \pi_u)T/\tau + \mathcal{E}_u\}, \qquad T = \sum_{t=2}^{L} \Delta_{t-1}.$$

This follows from $1 - S_u \leq \sum_t (1 - P_{t,uu}) \leq (1 - \pi_u)T/\tau$. Longer intervals accumulate separation evidence but can span regime changes. The fixed-schedule condition applies to this interval probability calculation; the preceding history-information result allows history-dependent contexts. These are conditional interval results, not guarantees of full latent-path recovery or exact switch-time localization.

B.7 Diagnostics after fitting

The anchor-based diagnostics used after fitting are descriptive consequences of the fitted probabilistic model. For example, one may compare the observed action a_m to $a_m^Q \in \arg\max_a Q_{\mathrm{anc}}(x_m, a)$, or modify selected components of C_u and recompute the implied action probabilities. These diagnostics are conditional on the learned dynamics, anchor reward, feature map, and regime layer. They are not used in training and are not claimed here as standalone causal or decision-theoretic guarantees.

Stability of audit quantities. The bounds V_m, T_m in Theorem 6 give decision-specific sufficient conditions for preserving an interpretation. Define the pointwise anchor error $\varepsilon_{q,m} = \max_a |\widehat{q}_m(a) - q_m(a)|$.

Proposition 13 (Margins for stable audit conclusions)**.** *The fitted top action is unchanged if its probability gap over the runner-up exceeds $2T_m$; the pre-action regime label is unchanged if its posterior gap exceeds $2V_m$. The anchor top action is unchanged if its score gap exceeds $2\varepsilon_{q,m}$.*

For a fixed action pair (a, b), let $d_u = C_u(\xi_m, a) - C_u(\xi_m, b)$, $D = \sum_u \rho^-_{m,u} d_u$, and $\max_u |\widehat{d}_u - d_u| \le \varepsilon_d$. Then

$$|\widehat{D} - D| \le \varepsilon_d + (\max_u d_u - \min_u d_u) V_m =: J_m.$$

Thus $|D| > J_m$ preserves its sign. The same bound applies to a fixed feature-group contribution. If every group has error bound at most J, a gap exceeding $2J$ between the largest and second-largest absolute contributions preserves the dominant group.

The review score $s_m = \max_a q_m(a) - q_m(a_m)$ satisfies

$$|\widehat{s}_m - s_m| \le 2\varepsilon_{q,m}.$$

If $\varepsilon_{q,m} \le \varepsilon_q$ across the reviewed decisions, a top-k set with a strict boundary gap greater than $4\varepsilon_q$ is unchanged.

Proof. Individual probabilities change by at most TV, so pairwise gaps change by at most twice the corresponding bound. Apply the same argument to anchor scores. For contributions,

$$\widehat{\rho^-_m}\widehat{d} - \rho^-_m d = \widehat{\rho^-_m}(\widehat{d} - d) + (\widehat{\rho^-_m} - \rho^-_m) d;$$

the terms are bounded by ε_d and the range of d times V_m. Absolute values are 1-Lipschitz, giving the dominant-group condition. The maximum of a score vector is 1-Lipschitz in sup norm, proving the review-score and set-boundary claims. □

Group comparisons use the same representation and coefficient allocation rule, preserving the identification conditions in Appendix B.3. If each model uses its own anchor-preferred action in the contrast, the anchor-margin condition first ensures a common action pair. Tie-inclusive review sets require a strict boundary gap for full set invariance. Stable inferred labels do not establish the occurrence times of latent transitions.

C Stage I: Latent Dynamics Specification

Stage I learns three coupled components: an action-modulated continuous-time Markov chain (CTMC) for the latent state $z(t)$, a state-dependent ODE for the outcome proxies $\lambda(t)$, and an emission model linking (z, λ) to observations o.

Action-modulated CTMC for $z(t)$. Between decision epochs $[t_m, t_{m+1})$, action a_m is held fixed and $z(t) \in \{1, \dots, K_z\}$ evolves as a CTMC with action-dependent generator $A(a_m) \in \mathbb{R}^{K_z \times K_z}$. The off-diagonal entries are parameterized as

$$[A(a)]_{jk} = \exp\big(\theta_{jk}^\top \phi(a)\big), \qquad j \ne k,$$

where $\phi(a)$ is an action-feature embedding; diagonal entries satisfy $[A(a)]_{jj} = -\sum_{k \ne j} [A(a)]_{jk}$. Over an interval of length $\Delta_m = t_{m+1} - t_m$, the transition matrix is $\exp\{A(a_m)\Delta_m\}$, which directly handles irregular decision times.

Outcome-proxy ODE for $\lambda(t)$. Each outcome proxy $\lambda_\ell(t) \in (0,1)$ ($\ell = 1, \dots, L$) evolves between latent-state jumps according to a state-dependent first-order ODE

$$\frac{d}{dt}\lambda_\ell(t) = \alpha_\ell^{(z(t))}\big(1 - \lambda_\ell(t)\big) - \kappa_\ell^{(z(t))}\lambda_\ell(t), \tag{11}$$

with state-conditioned rates $\alpha_\ell^{(k)}, \kappa_\ell^{(k)} \geq 0$ for each $k \in \{1, \dots, K_z\}$. The ODE admits a closed-form solution conditional on $z(t) = k$ being constant on a sub-interval, which we use for efficient propagation between decision epochs.

Emission model. At an observation time s, an available measurement $o(s)$ has likelihood $p_\theta(o(s) \mid z(s) = k, \lambda(s))$. The clinical filter maintains state-conditional proxies $\lambda_{k\ell}$; Appendix F specifies their low/normal/high emission mapping and treatment of missing observations.

Belief filtering. Beliefs follow the prediction–correction update in Section 2, with propagation split at action changes and observation times. Clinical filtering and anchor rollouts use a state-conditional proxy approximation, mixing proxies across CTMC transitions before applying the destination-state ODE update. The choice model uses the pre-action state $x_m = (b_m, \lambda_m)$, before observations from the decision's own month or hour are incorporated.

D Controlled Synthetic Q–C–Action Experiment

This appendix specifies the controlled experiment in Section 4.1: its data-generating process, estimators, recovery metrics, and action-level mechanism. It then examines recovery under changes in specification, signal strength, and persistence.

D.1 Data-generating process

Each candidate action has intensity $\ell_a \in \{0, 1/3, 2/3, 1\}$, corresponding to no, low, medium, and high treatment. The anchor state is $x_m = s_m \in [0,1]$, a continuous severity, and the cost context is $\xi_m = (s_m, r_m, \ell_{a_{m-1}})$, where $r_m \in [0,1]$ is resource pressure. For every candidate action, the fixed patient-benefit anchor is

$$Q_{\text{anc}}(s_m, a) = 2.5 - 4(\ell_a - s_m)^2 + 0.8 s_m \ell_a,$$

and the action-dependent cost features are

$$g(\xi_m, a) = \big[\ell_a(0.5 + r_m), \quad \ell_a^2, \quad |\ell_a - \ell_{a_{m-1}}|, \quad (s_m - \ell_a)_+, \quad (\ell_a - s_m)_+\big]^\top.$$

The features represent resource intensity, quadratic complexity, treatment-change burden, undertreatment, and overtreatment. The three rows of the true nonnegative weight matrix correspond to resource, inertia, and clinical-guardrail regimes:

$$W^\star = \begin{bmatrix} 2.40 & 0.35 & 0.10 & 0 & 0 \\ 0.10 & 0.20 & 3.00 & 0 & 0 \\ 0 & 0.10 & 0.10 & 2.80 & 2.00 \end{bmatrix}, \qquad \Pi^\star = \begin{bmatrix} .94 & .03 & .03 \\ .03 & .94 & .03 \\ .03 & .03 & .94 \end{bmatrix}.$$

For this controlled experiment, $C_u(\xi_m, a) = w_u^{\star\top} g(\xi_m, a)$. The observed action is sampled from the candidate-normalized softmax in Equation (2) with $\beta = 2$. Severity and pressure evolve as clipped autoregressions,

$$s_m = [.78 s_{m-1} + .22 U_m + \epsilon_m^s]_{[0,1]}, \qquad r_m = [.82 r_{m-1} + .18 V_m + \epsilon_m^r]_{[0,1]},$$

where $U_m, V_m \sim \text{Unif}(0,1)$, $\epsilon_m^s \sim \mathcal{N}(0, .055^2)$, and $\epsilon_m^r \sim \mathcal{N}(0, .060^2)$. Initially, $s_0 \sim \text{Unif}(.1, .9)$, $r_0 \sim \text{Unif}(.05, .95)$, the previous action is uniform over the four intensities, and the regime is uniform over the three states.

D.2 Protocol and estimators

Each of ten independent repetitions contains 120 length-50 trajectories, split into 80 training, 20 validation, and 20 test trajectories. The core comparisons use three random restarts and at most 22 generalized-EM iterations. Each M-step takes 35 projected Adam updates with learning rate 0.045

and an ℓ_1 penalty of 0.002 on nonnegative weights. We initialize weights uniformly on $[0.05, 2.20]$, regime probabilities uniformly, and transition diagonals at 0.85. Transition updates add pseudocounts of 2 on the diagonal and 0.15 off the diagonal; initial-state updates add 0.5 per regime. The fitted anchor scale and $\beta = 2$ remain fixed. Restarts are selected by validation NLL.

We compare anchor only ($C \equiv 0$), a global cost ($K_u = 1$), no-anchor switching ($Q_{\mathrm{anc}} = 0$), an independent mixture, and switching costs ($K_u = 3$). The no-anchor control retains the same feature map and switching estimator; the independent mixture constrains all transition rows to equal its mixture weights. The fitted controls use the same data and core optimization budget. For the separation–persistence grid, we use two restarts, at most 16 EM iterations, and 25 weight updates per iteration, with the same learning rate and penalty.

NLL is the mean negative log probability of the held-out action, and action accuracy is the fraction ranked first. Both use pre-action regime probabilities, updated only after scoring each action. Regime recovery uses Viterbi decoding of complete held-out trajectories. Regime accuracy maximizes agreement with the true path over label permutations; ARI measures chance-adjusted agreement between the two regime partitions. The same maximizing permutation aligns fitted cost vectors. For each test context and regime, let $C_u^{\circ}(\xi, a) = C_u(\xi, a) - |\mathcal{A}|^{-1} \sum_{b \in \mathcal{A}} C_u(\xi, b)$. Cost RMSE is

$$\left[\frac{1}{N_{\mathrm{test}} K_u |\mathcal{A}|} \sum_{m,u,a} \left(\widehat{C}_u^{\circ}(\xi_m, a) - C_u^{\star\circ}(\xi_m, a) \right)^2 \right]^{1/2},$$

where N_{test} counts test decisions. Tables report the mean and sample standard deviation across repetitions.

D.3 Action-level mechanism

Table 2: Candidate-wise ground-truth decomposition for one held-out clinical-guardrail decision. The anchor favors medium treatment, but action-dependent costs make low treatment most probable.

Action	Q_{anc}	C_u	$Q_{\mathrm{anc}} - C_u$	$p(a)$
None	1.669	1.343	0.325	0.013
Low	2.561	0.388	**2.174**	**0.521**
Medium	**2.565**	0.466	2.099	0.449
High	1.681	1.222	0.459	0.017

Holding x_m, ξ_m, and u_m fixed in Table 2, changing only the candidate action changes both Q_{anc} and C_u; subtracting the cost reverses the low-versus-medium ranking.

D.4 Robustness and recovery conditions

Specification checks. The perturbed-anchor check replaces the true anchor by $\widetilde{Q}_{\mathrm{anc}} = 0.65 Q_{\mathrm{anc}} + 1.15(-1, -1/3, 1/3, 1)$. The missing-feature test removes treatment-change burden from the fitted feature map while leaving the DGP unchanged.

Table 3: Specification checks for switching-cost recovery (mean ± SD across ten seeds).

Specification	NLL ↓	Regime acc. ↑	ARI ↑	Cost RMSE ↓
Correct	0.809 ± 0.039	0.824 ± 0.039	0.555 ± 0.083	$\mathbf{0.068 \pm 0.018}$
Perturbed anchor	0.811 ± 0.038	0.826 ± 0.036	0.558 ± 0.077	0.808 ± 0.031
Missing change feature	0.857 ± 0.034	0.644 ± 0.062	0.342 ± 0.098	0.969 ± 0.157

Table 3 confirms that recovery is conditional on the specified anchor and feature map. Perturbing the anchor leaves NLL and ARI nearly unchanged but increases cost RMSE from 0.068 to 0.808, showing that the recovered costs are anchor-relative. Removing the change feature worsens both regime and cost recovery, as expected when a relevant action-cost direction is omitted.

Recovery across signal strength and persistence. We multiply all rows of $W^{\star}$ by a common separation factor and vary the diagonal persistence p of the symmetric transition matrix, whose

off-diagonal entries are $(1-p)/2$. These changes control how distinctly the regimes rank actions and how long their segments persist.

Cost sep. \ persistence	.50	.80	.94	.99
.15	.000	.000	.001	.036
.35	.001	.004	.052	.278
.60	.003	.035	.265	**.823**
1.00	.032	.192	**.558**	**.897**
1.50	.204	.352	**.660**	**.933**

Table 4: Regime recovery across signal strength and persistence (mean ARI across ten seeds). Larger cost separation makes regime-specific action preferences more distinct; larger persistence yields longer within-regime segments. Green cells mark ARI ≥ 0.50.

Table 4 shows a clear interaction: recovery improves when cost profiles are more distinct and regimes persist longer. In the lower-right region, ARI exceeds 0.50 and reaches 0.933 at the strongest setting, whereas weak separation or persistence yields limited regime information in the observed actions.

D.5 Anchor estimation and cost recovery

Estimated-anchor recovery. We generate partially observed trajectories from a three-state action-dependent CTMC with three categorical channels, gaps in $\{0.5, 1, 2\}$, and 25% missing measurements. This setting omits the clinical ODE layer and retains the preceding regimes, costs, and nonnegative Stage II estimator, with cost features from observed context. Stage I uses forward likelihood, validation selection, and five-fold cross-fitting. True-parameter and estimated anchors share pre-action histories and a six-step return ($\delta = 0.99$), averaging rewards of $+1$ for normal and -1 for abnormal channels. The estimated anchor retains most regime recovery (ARI 0.451 versus 0.473), with cost RMSE increasing from 0.068 to 0.257; both outperform no-anchor recovery, while switch timing remains partially recovered (Table 5).

Table 5: End-to-end recovery in the CTMC setting (mean $\pm$ sample SD across ten seeds; 80/20/20 training/validation/test trajectories of length 50). Both RMSEs use action-centered contrasts; boundary F1 allows ± 1 decision. All fits retain the same nonnegative cost dictionary. The true-parameter anchor is a reference condition, not a learned method.

Anchor	Q RMSE $\downarrow$	ARI $\uparrow$	Boundary F1 $\uparrow$	Cost RMSE $\downarrow$
True-parameter anchor	0.000 ± 0.000	0.473 ± 0.106	0.265 ± 0.082	0.068 ± 0.021
Estimated Stage I anchor	0.255 ± 0.094	0.451 ± 0.102	0.237 ± 0.085	0.257 ± 0.114
No anchor	0.980 ± 0.004	0.158 ± 0.083	0.094 ± 0.074	0.711 ± 0.172

During-fit versus post-hoc anchoring. We compare fitting $Q_{\text{anc}} - C_u$ with fitting an unanchored linear GLM-HMM score f_u and then computing $C_u^{\text{post}} = Q_{\text{anc}} - f_u$. Both use the same dictionary, unrestricted signed coefficients, fixed scale, fitting budget, and validation selection; regime labels are aligned using training trajectories. Table 6 evaluates the resulting costs against the same truth. With a correct anchor, during-fit anchoring lowers cost RMSE from 0.319 to 0.086 despite similar NLL. With the learned CTMC anchor, it improves prediction, regime recovery, and cost recovery. Post-hoc NLL and ARI are unchanged when only the reference changes. Under a misspecified anchor, both cost RMSEs are approximately 0.78. The recovery benefit therefore depends on an informative reference and the specified feature class.

E Clinical Datasets

The two datasets capture complementary clinical decision processes: long-horizon management of mineral balance in CKD–MBD and high-frequency treatment during acute ICU care. We organize laboratory measurements and treatment records into patient-specific event sequences, using months for CKD–MBD and hours for MIMIC–IV. For both datasets, Appendix F specifies the shared value anchor, and Appendix G defines the action-dependent cost features.

Table 6: During-fit versus post-hoc anchoring (mean ± sample SD over ten seeds). RMSE measures action-centered costs; direction accuracy measures the sign of nonzero true cost differences between candidate actions.

Anchor	Method	NLL ↓	ARI ↑	Cost RMSE ↓	Direction (%) ↑
Known correct	During-fit	$\mathbf{0.762 \pm 0.019}$	$\mathbf{0.638 \pm 0.042}$	$\mathbf{0.086 \pm 0.012}$	$\mathbf{98.4 \pm 0.2}$
	Post-hoc	0.779 ± 0.022	0.610 ± 0.091	0.319 ± 0.008	92.8 ± 1.4
CTMC true	During-fit	$\mathbf{0.628 \pm 0.070}$	$\mathbf{0.505 \pm 0.116}$	$\mathbf{0.055 \pm 0.018}$	$\mathbf{100.0 \pm 0.0}$
	Post-hoc	0.715 ± 0.062	0.289 ± 0.108	0.594 ± 0.026	79.3 ± 2.0
CTMC learned	During-fit	$\mathbf{0.644 \pm 0.067}$	$\mathbf{0.489 \pm 0.056}$	$\mathbf{0.125 \pm 0.081}$	$\mathbf{94.9 \pm 4.4}$
	Post-hoc	0.715 ± 0.062	0.289 ± 0.108	0.547 ± 0.024	85.1 ± 2.9
Misspecified	During-fit	$\mathbf{0.763 \pm 0.018}$	$\mathbf{0.641 \pm 0.065}$	0.781 ± 0.028	$\mathbf{87.5 \pm 1.3}$
	Post-hoc	0.779 ± 0.022	0.610 ± 0.091	$\mathbf{0.780 \pm 0.028}$	83.5 ± 2.2

E.1 CKD–MBD: LONG-HORIZON TREATMENT MANAGEMENT

Clinical setting and data source. We use a private clinical cohort assembled through a longstanding collaboration with a leading hospital. The records link repeated laboratory assessments to medication regimens prescribed during routine CKD–MBD care. Management involves balancing calcium, phosphorus, and PTH through adjustments to concurrent medications (Ketteler et al., 2017; Moe et al., 2007). The processed cohort contains 150 patients with nonempty event sequences and $2{,}934$ monthly treatment decisions, enabling us to study how treatment continuity and changes relate to evolving biochemical profiles.

Laboratory measurements. We harmonize laboratory fields across source files and align records by patient and month. The three indicators are corrected calcium (cCa), phosphorus (P), and intact PTH. When cCa is not supplied, it is calculated as $\mathrm{cCa} = \mathrm{Ca} + 0.02(40 - \mathrm{albumin})$, with calcium in mmol/L and albumin in g/L. The processed event dictionary distinguishes four cCa categories (very low, low, target-range, and high) and three categories each for P and PTH (low, target-range, and high), giving ten laboratory markers. For each indicator, the monthly representation retains one available value from the processed records. The two low-calcium categories are combined when constructing the three-category physiological observation channel.

Medication processing and decision sequences. We standardize drug-name variants and convert recorded dose frequencies into daily-equivalent amounts, including division by seven for weekly prescriptions. These amounts are grouped into 13 drug–dose markers: four calcium-carbonate tiers, three sevelamer tiers, one lanthanum-carbonate marker, three calcitriol tiers, and two cinacalcet tiers. A monthly regimen combines the markers for concurrently used drugs, retaining the highest recorded tier when multiple tiers of the same drug occur within a month. When a month has no encoded treatment, the preceding regimen is carried forward within the patient's prescription-record period; the sequence is not extended beyond the final prescription month. Each patient-month with a nonempty regimen contributes one treatment decision. Of the 2,934 decision months, 730 (24.9%) carry forward a regimen without a contemporaneous encodable prescription; the remaining 2,204 months have recorded, encodable prescriptions.

E.2 MIMIC–IV: HIGH-FREQUENCY ICU TREATMENT

Clinical setting and cohort construction. We derive an ICU cohort from MIMIC–IV using sepsis-related diagnosis and diagnosis-related-group codes (Johnson et al., 2024). The treatment setting involves vasopressor support, inotropic support, and fluid administration during acute physiological changes (Evans et al., 2021). Extraction identifies low-urine-output events using a six-hour mean below $1\,\mathrm{mL\,kg^{-1}\,h^{-1}}$, with no such event in the preceding observation window. Each sequence covers 48 hours before the event and up to 24 hours afterward, with overlapping windows excluded. The processed extract contains $2{,}475$ sequences. After linking sequences to subjects and retaining sequences with treatment decisions, the behavioral cohort comprises $1{,}011$ sequences from 818 subjects and $9{,}174$ treatment-event decisions.

Physiological observations. We construct hourly low/normal/high observations for systolic blood pressure, SpO_2, white blood cell count, arterial lactate, central venous pressure, and creatinine. When an indicator is measured multiple times within an hour, its last measurement is retained; hours without a measurement remain explicitly unobserved. The decision context also records recent physiological events and low urine output.

Treatment processing and decision sequences. Treatment events are encoded by nine markers spanning vasopressors, inotropes, and fluids. At each hour with a treatment event, the distinct treatment markers form one action; the hourly clock retains the intervals between successive decisions.

F Clinical Anchor Construction

Candidate actions. Using the treatment encodings in Appendix E, we retain the 20 most frequent training combinations in CKD–MBD and up to 50 in MIMIC–IV. Remaining combinations map to OTHER, with the same mapping applied across training, validation, and test sets. Stage I represents each combination by its individual treatment indicators and an intercept.

Pre-action states. The state x_m combines the pre-action filtered belief over three CTMC states with the current ODE proxies for the dataset's observation channels. Cross-fitting and optimization settings are given in Appendix K.

Primary reward. Let b_z be the current phenotype belief and $\lambda_{z\ell}$ the current ODE proxy for channel ℓ, conditional on phenotype z. The emission probabilities are $p^N_{z\ell} = \lambda_{z\ell}$, $p^L_{z\ell} = \varphi_{z\ell}(1-\lambda_{z\ell})$, and $p^H_{z\ell} = (1-\varphi_{z\ell})(1-\lambda_{z\ell})$, where the fitted $\varphi_{z\ell}$ allocates off-target mass to the low category. Missing measurements contribute no emission likelihood, so filtering propagates the fitted dynamics between available observations. The one-step reward combines target alignment, abnormality penalties, joint-risk terms, and treatment burden:

$$r(x,a) = \sum_z b_z \left[\sum_\ell \left(v_\ell p^N_{z\ell} - d^L_\ell p^L_{z\ell} - d^H_\ell p^H_{z\ell}\right) - \sum_{(\ell,c,j,d)\in\mathcal{P}} \eta_{\ell c,jd} p^c_{z\ell} p^d_{zj} \right] - 0.05\, n_{\mathrm{trt}}(a), \tag{12}$$

where $n_{\mathrm{trt}}(a)$ counts active treatment markers. For CKD–MBD, channel order is Ca, P, PTH, with $v = (1,1,1)$, $d^L = (1.2, 1.2, 2)$, and $d^H = (2, 1.5, 1.5)$. Joint penalties are 0.15 for high Ca with high P, 0.15 for high P with low PTH, and 0.10 for high Ca with low PTH. For MIMIC–IV, channel order is SBP, SpO_2, WBC, lactate, CVP, creatinine, with

$$v = (1, 1, 0.8, 1.2, 0.5, 1),$$
$$d^L = (1.5, 1.5, 1, 0.2, 0.5, 0.2), \qquad d^H = (0.5, 0.2, 1, 1.5, 0.5, 1.2).$$

Joint penalties are 0.20 for low SBP with high lactate, 0.15 for low SpO_2 with high lactate, and 0.10 for high lactate with high creatinine. Products are evaluated within each phenotype and then averaged using b_z. The anchor sums this reward over six monthly transitions in CKD–MBD or six hourly transitions in MIMIC–IV, with discount 0.99, repeating the candidate action throughout the rollout.

Value scale and choice probabilities. Let $Q^{\mathrm{raw}}_m(a)$ denote the six-step return. For each outer split, let s_Q be the median positive candidate-action range among the cross-fitted training values. The clinical choice model uses

$$Q_{\mathrm{anc}}(x_m, a) = \frac{Q^{\mathrm{raw}}_m(a) - |\mathcal{A}_m|^{-1} \sum_{b\in\mathcal{A}_m} Q^{\mathrm{raw}}_m(b)}{s_Q}, \qquad \beta = 1. \tag{13}$$

The same s_Q scales training, validation, and test values. The OTHER action receives the mean raw value of the named candidates before this normalization. Thus Q_{anc}, inferred costs, and the displayed action-cost contrasts share one score scale.

G Action-Dependent Cost Features

The cost context ξ_m combines x_m, the previous treatment, decision time, and clinical history. CKD–MBD uses the latest strictly pre-action Ca/P/PTH categories and the month index, excluding same-month laboratories. MIMIC–IV uses the hour index and strictly prior event histories summarized by

recency and 6- and 24-hour counts, excluding same-hour events. Every regime assigns its own cost $C_u(\xi_m, a) = w_u^\top g(\xi_m, a)$ using the same feature dictionary.

For a fixed pre-action context ξ_m, Stage II evaluates every candidate action with the same feature function $g(\xi_m, a)$. A retained coordinate has one of two forms:

$$g_j(\xi, a) = h_j(a) \qquad \text{or} \qquad g_j(\xi, a) = q_j(\xi) h_j(a),$$

where $h_j(a)$ is an action identity, action-group membership, intensity, or change relative to the previous action. Features are standardized on training decisions crossed with their candidate actions and centered over candidates at each context. Hence action-independent offsets vanish before fitting.

Complete grouped dictionaries. Table 7 partitions all 25 CKD–MBD and 78 MIMIC–IV coordinates.

Table 7: Action-dependent feature dictionaries used in the two clinical experiments. Counts include all implemented coordinates, and each feature family is available to every behavioral regime.

Dataset	Feature family (count)	Candidate-action construction
CKD–MBD	Drug/dose identity + OTHER (14)	Indicators for 13 dose-tier markers: calcium carbonate/Caltrate (4), sevelamer (3), lanthanum (1), calcitriol family (3), cinacalcet (2), plus OTHER.
	Regimen size (1)	Number of active drug/dose markers in candidate a.
	Relation to previous regimen (4)	Indicators/counts for changed, added, removed, and overlap with a_{m-1}.
	Lab–treatment interactions (6)	Calcium-binder intensity × Ca low/high; phosphate-binder intensity × P low/high; PTH-treatment intensity × PTH low/high.
MIMIC–IV	Treatment identity + OTHER (10)	Indicators for nine treatment markers (epinephrine, phenylephrine, norepinephrine, dopamine, dobutamine, milrinone, crystalloid, colloid, water) plus OTHER; dopamine and water are inactive in the current derived cohort.
	Combination size (1)	Number of active treatment markers in candidate a.
	Previous-action relation + group counts (7)	Changed, added, removed, overlap; counts of inotropes, fluids, and pressors in a.
	Core event–treatment interactions (5)	Pressor × recent low SBP/high lactate; fluid × recent high lactate/low urine; inotrope × recent low SBP.
	Rich group–history interactions (36)	Three candidate groups (pressor, inotrope, fluid) × four pre-action events (low SBP, low SpO_2, high lactate, low urine) × recent/6-hour/24-hour summaries.
	Change–history interactions (12)	Candidate treatment change × the same four events and three history windows.
	Clinical-rule features (7)	Pre-action event summaries multiplied by candidate treatment presence or absence, as defined below.

For example, the raw CKD–MBD calcium-binder interaction is

$$g_{\text{CaBinder}\times\text{CaHigh}}(\xi, a) = \mathbf{1}\{\text{last Ca} \in \{\text{high}, \text{very high}\}\}\, d_{\text{Ca}}(a).$$

Here $d_{\text{Ca}}(a)$ is the maximum normalized dose tier among calcium-binder markers in the candidate: the four ordered tiers have intensities $1/4, 1/2, 3/4, 1$, with zero for absence. Phosphate-treatment intensity is the maximum across calcium binder, sevelamer, and lanthanum; PTH-treatment intensity is the maximum across the vitamin-D and calcimimetic groups, each normalized by its own number of tiers.

The corresponding MIMIC–IV interaction is

$$g_{\text{Pressor}\times\text{LowSBP}}(\xi, a) = r_{\text{lowSBP}}(m)\, \mathbf{1}\{a \text{ contains a vasopressor}\}.$$

For each event type e, the recent value is $r_e(m) = \exp[-(t_m - t_e^{\text{last}})/6]$, where $t_e^{\text{last}} < t_m$ is its most recent event hour; it is zero if no such event exists. The 6- and 24-hour summaries use $\log(1 + n_{e,w}(m))$, with $n_{e,w}(m)$ counting recorded events in $[t_m - w, t_m)$. The candidate's group membership or change indicator multiplies these summaries, giving separate coordinates for recency and repeated event burden.

The seven clinical-rule coordinates are $r_{\text{lowSBP}} I_{\text{water}}$, $r_{\text{highLactate}} I_{\text{water}}$, $r_{\text{lowUrine}} I_{\text{water}\vee\text{colloid}}$, $(1-r_{\text{normalCreatinine}}) I_{\text{colloid}}$, $r_{\text{lowSBP}}(1-I_{\text{pressor}})$, $r_{\text{highLactate}}(1-I_{\text{fluid}\vee\text{pressor}})$, and $r_{\text{lowUrine}}(1-I_{\text{fluid}\vee\text{pressor}})$. Here I indicates whether candidate a contains the named treatment or group, and all r_e use the recency definition above. In particular, $1 - r_{\text{normalCreatinine}}$ measures the absence or age of prior normal-creatinine evidence. Coordinates with zero candidate-action contrast remain in the dictionary but are omitted from the displayed profiles.

G.1 Sensitivity to feature representation

With Stage I, the anchor, and representation fixed, the principal contribution directions in all three clinical cases persisted across all 40 fits per case (20 patient/subject bootstrap resamples, two initializations), at half or double the selected L2 strength, and at two or four times the EM budget.

We compare keeping the previous regimen with adding, removing, or replacing one treatment marker. Within each split, Stage I and the anchor are fixed; Stage II is refitted using the full dictionary or after removing regimen-size or added-treatment counts, selecting among three initializations by validation likelihood. Regimes are aligned using training-set emission similarity. We compare complete cost contrasts after centering across regimes, weighting eligible decision times equally and averaging over their eligible alternatives. Ordinary L2 is retained, so removing coordinates also changes regularization geometry. Table 8 shows generally similar action-cost contrasts, with greater variability for removal and replacement; MIMIC–IV removal contrasts cover only 5.1% of decision times. This evaluates functional contrasts, not invariance of individual group allocations.

With fitted models fixed, the checked equivalent dictionaries preserve the displayed group contributions in Figure 2. In the longitudinal CKD–MBD case, jointly eliminating redundant size and addition counts changes continuation and treatment-identity contributions from $(4.52, 2.46)$ to $(3.02, 4.34)$; these remain the two largest positive contributions, with total contrast unchanged at 7.21.

Table 8: Between-regime action-cost contrast cosine (↑), mean ± sample SD over ten patient/subject splits. Coverage is the percentage of held-out decision times supporting each comparison. Refits are compared with the reference fit within each split.

Dataset	Keep vs.	Coverage	Full dictionary refit	Drop regimen size	Drop added count
CKD–MBD	Add one	45.1%	0.881 ± 0.095	0.857 ± 0.092	0.870 ± 0.111
	Remove one	32.2%	0.832 ± 0.169	0.812 ± 0.172	0.762 ± 0.180
	Replace one	74.8%	0.814 ± 0.197	0.781 ± 0.183	0.764 ± 0.216
MIMIC–IV	Add one	88.1%	0.865 ± 0.093	0.935 ± 0.053	0.872 ± 0.088
	Remove one	5.1%	0.716 ± 0.191	0.892 ± 0.118	0.696 ± 0.140
	Replace one	88.6%	0.816 ± 0.140	0.946 ± 0.070	0.838 ± 0.146

H Phenotype–Regime Dynamics

Figure 6 shows joint and single-layer label changes. Joint label changes account for 8% and 9% of adjacent decision pairs with at least one label change in CKD–MBD and MIMIC–IV, respectively, but occur $1.6\times$ and $2.1\times$ as often as expected under independence. These label changes may also reflect posterior updates from new information.

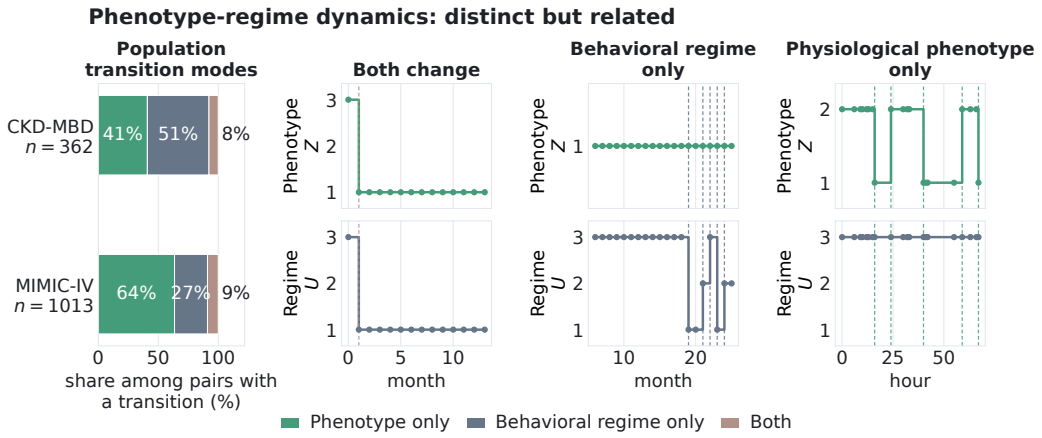


Figure 6: Phenotype–regime label changes, using the most probable pre-action phenotype and regime. Left: label-change proportions pooled across five held-out splits, conditional on at least one label change between adjacent decisions. Joint-change enrichment divides the joint frequency by the product of marginal change frequencies, using all adjacent pairs. Right: CKD–MBD examples of joint and regime-only label changes, and a MIMIC–IV example of phenotype-only label changes. Dashed lines mark the relevant changes.

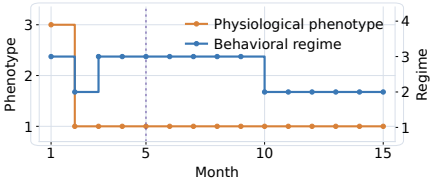
A longitudinal anchor-relative decision audit in CKD-MBD

(a) Patient course and three action views

Mo.	Ca	P	PTH	Observed	Anchor	Fitted
1	Low	High	—	Ca-M+Sev-L	Ca-H+Sev-M	Ca-H
2	Low	High	Normal	Ca-M	Ca-H+Sev-M	Ca-M+Sev-L
3	Low	High	Normal	Ca-M	Ca-H+Sev-M	Ca-M
4	Low	High	Normal	Ca-M	Ca-H+Sev-M	Ca-M
5	**Low**	**High**	**Normal**	**Ca-M**	**Ca-H+Sev-M**	**Ca-M**
6	Low	High	Normal	Ca-M	Ca-H+Sev-M	Ca-M
7	Low	High	Normal	Ca-M+Sev-L	Ca-H+Sev-M	Ca-M
8	Low	High	Normal	Ca-M+Sev-L	Ca-H+Sev-M	Ca-M+Sev-L
9	Normal	High	Normal	Ca-H+Sev-L	Ca-H+Sev-M	Ca-M+Sev-L
10	Normal	High	Normal	Ca-H+Sev-L	Ca-H+Sev-M	Ca-H+Sev-L
11	High	Normal	Normal	Ca-H+Sev-M	Ca-H+Sev-M	Ca-H+Sev-L
12	High	High	Normal	Ca-H+Sev-M	Ca-H+Sev-M	Ca-H+Sev-M
13	Low	Normal	Normal	Ca-H+Sev-H	Ca-H+Sev-M	Ca-H+Sev-M
14	Normal	High	Normal	Ca-H+Sev-L	Ca-H+Sev-M	Ca-H+Sev-H
15	Normal	Normal	Normal	Ca-H+Sev-L	Ca-H+Sev-M	Ca-H+Sev-L

Ca = calcium carbonate/Caltrate; Sev = sevelamer
L/M/H = low/medium/high dose tier
Highlighted month 5 is decomposed in (c).

(b) Pre-action latent dynamics



(c) Month 5: action-cost decomposition

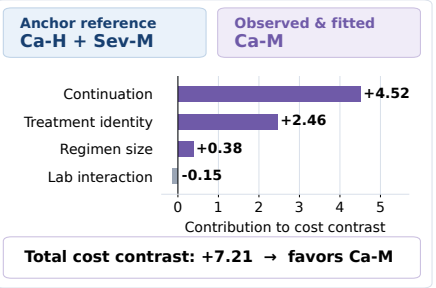


Figure 7: Clinical review of a held-out CKD–MBD trajectory over 15 decisions. **(a)** Most recent pre-action Ca/P/PTH categories and observed, anchor-reference, and fitted top actions. Ca, P, and PTH denote calcium, phosphorus, and parathyroid hormone; in action labels, Ca denotes calcium carbonate/Caltrate, Sev sevelamer, and L/M/H low/medium/high dose tiers. **(b)** Aligned labels of the most likely pre-action physiological phenotype and behavioral regime ($K_z = 3$, $K_u = 4$). The highlighted month 5 has phenotype z_1 and regime u_3. **(c)** Month-5 cost decomposition. Positive contributions to $\bar{C}_m(a_m^Q) - \bar{C}_m(\hat{a}_m)$ favor the fitted Ca-M over the anchor-reference Ca-H+Sev-M.

I Clinical Auditing and Counterfactual Extensions

I.1 Case-Level Anchor-Relative Decision Audit

Figure 7 applies the audit in Section 4.5 to an anonymous held-out CKD–MBD trajectory, using the models evaluated in Figure 8. We compare the observed action a_m, the anchor reference $a_m^Q := \arg\max_{a \in \mathcal{A}_m} Q_{\text{anc}}(x_m, a)$, and the fitted top action $\hat{a}_m := \arg\max_{a \in \mathcal{A}_m} p(a \mid \mathcal{H}_m^-)$, alongside the pre-action phenotype and behavioral regime. The cost contrast $\bar{C}_m(a_m^Q) - \bar{C}_m(\hat{a}_m)$

identifies which terms favor the fitted choice over the reference. Fitted probabilities use the regime mixture in equation 3.

Over the 15-decision course, the physiological phenotype moves from z_3 to z_1 at month 2 and then remains stable, while the behavioral regime follows $u_3 \to u_2 \to u_3 \to u_2$. The observed plan continues calcium alone from months 2 to 6, adds sevelamer at month 7, and changes dose tiers later.

At month 5, calcium is low, phosphorus is high, and PTH is normal. The anchor probabilities p_Q of Ca-H+Sev-M and Ca-M are 0.071 and 0.035; the fitted mixture assigns 0.003 and 0.786, respectively. The cost contrast $\bar{C}_m(\text{Ca-H+Sev-M}) - \bar{C}_m(\text{Ca-M}) = 7.21$ comprises continuation/change (4.52), treatment identity (2.46), regimen size (0.38), and lab–treatment interaction (-0.15). The higher relative overlap cost for u_3 in Figure 4 is centered across regimes coordinate by coordinate. Here, full candidate-regimen costs are summed and averaged over $\rho^-_{m,u}$, with continuation contributing most under the specified dictionary.

I.2 Stability of anchor-relative audits

Using the four anchor variants in Table 13, we assess stability of review priorities and case explanations on the same held-out decisions with each fit's pre-action posterior. For variant v (primary $v = 0$), define anchor gap $s^v_m = \max_a Q_v(x_m, a) - Q_v(x_m, a_m)$ and cost contrast $c^v_m = \bar{C}^v_m(a^{Q,v}_m) - \bar{C}^v_m(a_m)$, where $a^{Q,v}_m$ is its anchor preference. Spearman ρ_s compares gap ranks; retention measures overlap with the primary highest-gap 20% set at the same budget, including boundary ties. This budget is not a clinical threshold.

Cost metrics use decisions with $a_m \neq a^{Q,v}_m$ under both fits. Group agreement uses the largest absolute contribution, combining MIMIC identity and size as treatment class. Normalized changes are $E_Q = \sum_m |s^v_m - s^0_m| / \sum_m s^0_m$ over all decisions and $E_C = \sum_m |c^v_m - c^0_m| / \sum_m |c^0_m|$ over the cost subset.

Table 9: Audit stability, averaged over five splits. Retained: highest-gap-set retention; Group: dominant-contribution agreement; Sign: total-cost-sign agreement. E_Q, E_C: normalized absolute changes. All entries except Spearman ρ_s are percentages.

Dataset	Anchor variant	ρ_s	Retained	Group	Sign	E_Q	E_C
CKD–MBD	Equal-channel reward	0.982	92.9	98.8	99.4	8.2	1.5
	Amplified-risk reward	0.992	94.4	96.6	99.4	4.0	2.9
	One-step continuation	0.916	90.1	96.2	98.4	20.3	5.8
	Stage I value noise (20%)	0.636	54.4	88.1	97.4	51.1	19.6
MIMIC–IV	Equal-channel reward	0.999	99.6	98.5	99.4	1.5	1.8
	Amplified-risk reward	0.990	96.4	97.1	98.8	17.4	4.0
	One-step continuation	0.990	97.0	98.3	99.3	49.7	2.9
	Stage I value noise (20%)	0.526	68.3	80.1	96.1	63.4	49.9

Table 9 shows that reward reweighting and one-step continuation largely preserve review priorities and cost structure, whereas value noise produces larger changes. MIMIC one-step rankings remain stable despite changes in gap magnitude. Continuation changes redefine value; noise is a synthetic stress test, not calibrated Q uncertainty.

For the three cases in Figures 2 and 7, principal contribution directions persist in 12/12 checks with the original action pair and 11/12 with variant-specific anchor comparators. Only the CKD lab–treatment contribution under noise reverses ($+0.853$ to -0.429); all 12 updated cost contrasts remain positive. Month-5 continuation and identity directions persist: cost contrasts range 6.37–8.20 (primary 7.21), and anchor gaps 0.45–0.76 (primary 0.72).

I.3 SCM-Based Counterfactual Trajectories and Action Synthesis

This extension synthesizes alternative trajectories under the fitted structural causal model (SCM). Its retrospective return $Q_{\text{cf}}(a \mid \mathcal{H}^i, m)$ conditions on the complete observed trajectory $\mathcal{H}^i$, whereas the shared anchor $Q_{\text{anc}}(x_m, a)$ used for behavior prediction and Figure 7 uses only pre-action information.

The former evaluates alternatives given the recorded course; the latter compares candidate values before a decision.

I.3.1 Retrospective Action Values and Reward

For an intervention at decision epoch t_m^i, let $R_m(a) = \sum_{k=m}^{T} \delta^{k-m} r_k^{(a)}$, where T is the terminal decision index and $\delta \in (0, 1]$. The retrospective value Q_{cf} averages this return over posterior samples of latent states and exogenous noise given $\mathcal{H}^i$. The reward can use equation 12, evaluated from the simulated state and ODE proxies. Candidates share the same reward, horizon, discount, and subsequent-action convention. For regularly spaced epochs, $\delta = e^{-\rho\Delta}$; with irregular timing, the epoch-indexed discount approximates the continuous-time discount in Section 2.

I.3.2 Abduction-Action-Prediction Framework

Step 1: Abduction. Given the full observed history $\mathcal{H}^i = (\{o_m^i, a_m^i\}_{m=1}^{M_i})$, we use the fitted model to infer the posterior over latent states and exogenous noises $p(z_{1:T}, \varepsilon^{\mathrm{tr}}, \varepsilon^{\mathrm{em}} \mid \mathcal{H}^i)$ via a backward-smoothing particle-EM procedure. Here, $\varepsilon^{\mathrm{tr}}$ includes the latent Poisson arrival times and Gumbel noise used at uniformization events; $\varepsilon^{\mathrm{em}}$ denotes the Gumbel noise associated with categorical emissions o_m^i.

Step 2: Action Intervention. For each candidate action $a \in \mathcal{A}$ at decision epoch t_m^i, we replace the factual action with the intervention $\tilde{a}_{t_m}^i = a$. Subsequent actions can either follow the factual trajectory for a single-decision intervention or be recursively replaced by anchor-optimal actions for sequence synthesis, evaluating Q_{anc} at the counterfactual pre-action state.

Step 3: Prediction (Forward Rollout). We roll forward under $\mathrm{do}(a_t := \tilde{a}_t)$, reusing each posterior sample's Poisson arrival times and Gumbel noises across candidates to couple their alternative trajectories:

1. **Latent State** $z(t)$: Transitions at each uniformization jump time s_k are computed using the identical Gumbel realizations $g_{k,j} \in \varepsilon^{\mathrm{tr}}$ from the abduction step:
$$z_{k+1} = \arg\max_j \{\log \tilde{P}_{z_k j}(\tilde{a}_{s_k}) + g_{k,j}\}, \quad \tilde{P}(\tilde{a}_{s_k}) = I + \frac{1}{\Omega} A(\tilde{a}_{s_k}). \tag{14}$$
where $\Omega \geq \max_{j,a} |A_{jj}(a)|$ is the uniformization rate.
2. **Outcome Dynamics** λ_τ: Between latent-state jumps, λ_ℓ evolves deterministically under the fitted state-dependent ODE. For an interval of length Δ_τ in latent state z_τ,
$$\lambda_\ell(\tau + \Delta_\tau) = \bar{\lambda}_\ell^{(z_\tau)} + \left(\lambda_\ell(\tau) - \bar{\lambda}_\ell^{(z_\tau)}\right) \exp\left[-\left(\alpha_\ell^{(z_\tau)} + \kappa_\ell^{(z_\tau)}\right)\Delta_\tau\right],$$
$$\bar{\lambda}_\ell^{(z)} = \frac{\alpha_\ell^{(z)}}{\alpha_\ell^{(z)} + \kappa_\ell^{(z)}}. \tag{15}$$
3. **Emissions at Observations**: Counterfactual signals $\tilde{o}_{m,\ell}^i$ are reconstructed using the fixed emission Gumbels $\varepsilon^{\mathrm{em}}$ and the newly generated latent trajectory.

Retrospective action synthesis. Using S posterior rollouts per candidate, we estimate

$$\widehat{Q}_{\mathrm{cf}}(a \mid \mathcal{H}^i, m) = \frac{1}{S} \sum_{s=1}^{S} \sum_{k=m}^{T} \delta^{k-m} r_k^{(s,a)}.$$

We then select $a_m^{\mathrm{cf}} \in \arg\max_{a \in \mathcal{A}} \widehat{Q}_{\mathrm{cf}}(a \mid \mathcal{H}^i, m)$.

J Predictive Evaluation and Robustness

J.1 Behavioral action prediction

Figure 8 compares held-out action prediction. The global baseline uses one cost profile; the IID mixture uses the same regime count as the switching model but history-independent mixing weights. Both use the same anchor and cost features as the switching model.

For predicted candidate probabilities $p_m(a)$, NLL is the mean $-\log p_m(a_m)$, and Brier is the mean $\sum_a [p_m(a) - \mathbf{1}\{a = a_m\}]^2$. Top-1 and Top-5 are the fractions of decisions whose observed action ranks first or among the five highest-probability candidates. ECE compares top-action confidence with empirical accuracy in ten equal-width confidence bins, weighted by their decision counts. Changed-action NLL uses decisions whose action differs from the preceding decision.

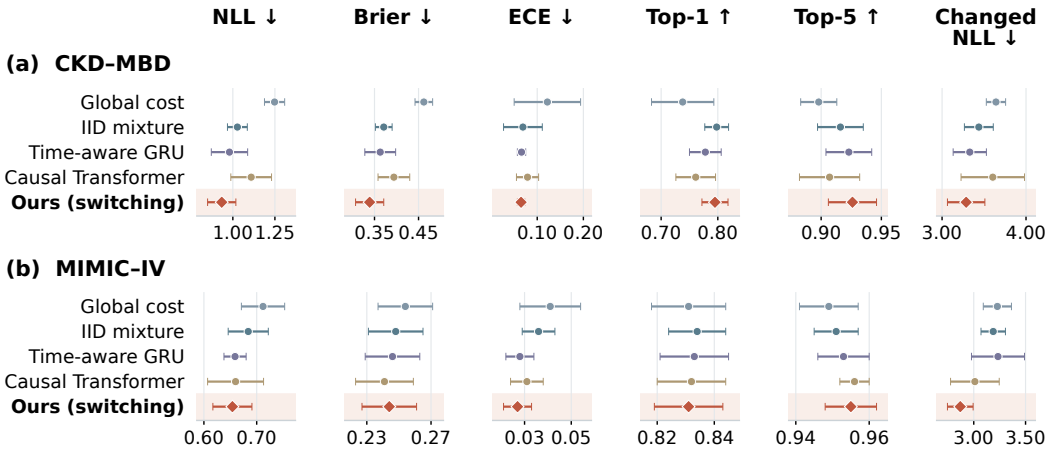


Figure 8: Held-out behavioral action prediction on (a) CKD–MBD and (b) MIMIC–IV. Points and whiskers show means and ± 1 sample standard deviation over five patient/subject splits. Changed NLL uses decisions with an action change. Axis ranges vary by dataset and metric.

J.2 Stage I physiological forecasting

We evaluate Stage I by predicting future observed low/normal/high physiological categories, comparing it with calibrated persistence, first-order Markov, a three-state HMM, multinomial logistic regression, a time-aware GRU, and ODE-RNN. All methods share forecast origins and targets and use only information available at the origin, holding the origin treatment fixed during action-conditioned rollouts. Targets occur exactly at the specified horizon on the monthly (CKD–MBD) or hourly (MIMIC–IV) grid, with at least one physiological observation at or before the origin. Within each bin, CKD–MBD retains the last input record per variable and MIMIC–IV the latest measurement. Missing target variables are excluded without imputation during evaluation.

Figure 9 shows the lowest reported mean NLL and Brier scores and the highest mean accuracy for our model across all four dataset–horizon settings. Its mean classwise calibration error ranges from 0.0314 to 0.0390 and is lowest among the compared methods in all four settings. These results support physiological forecasting with the dynamics used for action valuation and longitudinal clinical audit.

J.3 Ablations of the shared anchor and regime switching

Synthetic recovery. Table 10 compares regime recovery, switch localization, and cost recovery. With the correct anchor, switching improves all three over the IID mixture. Removing the anchor degrades cost recovery; misspecifying it distorts costs despite similar regime recovery. Boundary F1 indicates only partial recovery of transition timing. Appendix D.5 further compares during-fit and post-hoc anchoring.

Clinical fit and interpretation. Table 11 compares clinical behavioral fit. The no-anchor model sets $Q_{\text{anc}} = 0$; the fixed-class model sets $\Pi(\Delta) = I$ but updates its class posterior after each action. Anchor + switching achieves the lowest mean NLL, improving over fixed classes by 0.081 in CKD–MBD and 0.048 in MIMIC–IV.

Sensitivity to carried-forward treatment records. Using the same five patient-level splits, we refit and evaluate Stage II using only recorded, encodable CKD–MBD prescriptions, retaining recorded unchanged treatments and actual time gaps. Stage I, the anchor, candidate actions, and feature scaling

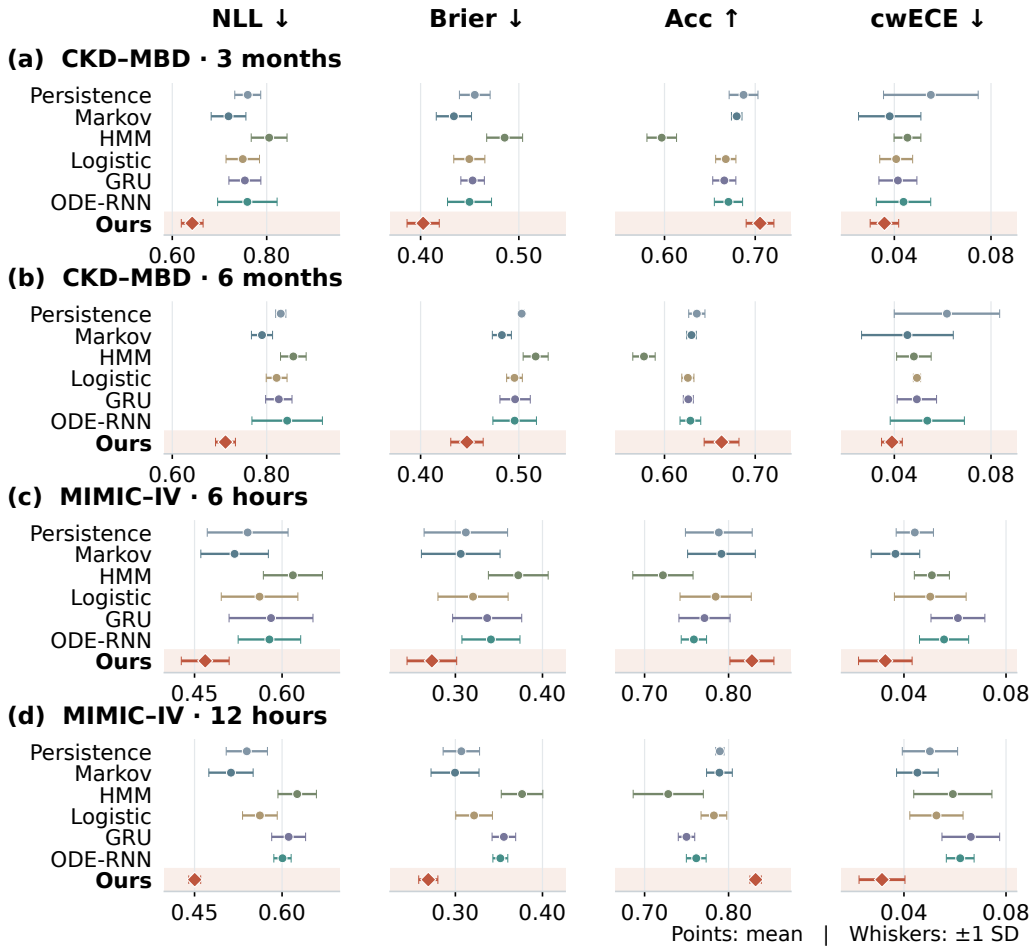


Figure 9: Held-out Stage I physiological forecasting. Panels show CKD–MBD at (a) three and (b) six months, and MIMIC–IV at (c) six and (d) twelve hours. NLL and Brier evaluate predicted probabilities over the three observed physiological categories; Acc is category accuracy, and cwECE averages calibration error across classes. Lower values are better for NLL, Brier, and cwECE; higher accuracy is better. Points show means and whiskers indicate one standard deviation over ten seeds for every method. Warm-red diamonds highlight our model. Axes are shared across horizons within each dataset for each metric.

Table 10: Synthetic ablations under the nonnegative-cost protocol (Tables 1 and 3). Mean ± sample SD over ten seeds. ARI measures regime recovery; boundary F1 allows a one-decision tolerance; RMSE evaluates action-cost contrasts.

Model	ARI ↑	Boundary F1 ↑	Cost RMSE ↓
Correct anchor + switching	0.555 ± 0.083	0.390 ± 0.114	0.068 ± 0.018
Correct anchor + IID mixture	0.078 ± 0.026	0.188 ± 0.025	0.471 ± 0.185
No anchor + switching	0.498 ± 0.109	0.344 ± 0.109	0.470 ± 0.060
Wrong anchor + switching	0.558 ± 0.077	0.395 ± 0.111	0.808 ± 0.031

remain fixed. The three anchored models use four regimes, the same regularization grid and per-fit optimization budget, with selection by validation NLL. Switching retains lower NLL than both controls on all five splits, both overall and on treatment-change decisions (Table 11).

Table 11: Clinical ablations: held-out pre-action NLL (↓), mean ± sample SD over five patient/subject splits. Right: Stage II refits using only recorded CKD–MBD prescriptions; Changed NLL uses prescriptions differing from the preceding recorded regimen. Bold marks the lowest mean in each column; dashes denote comparisons not included here.

	Full decision sequences		Recorded CKD–MBD prescriptions	
Model	CKD–MBD	MIMIC–IV	NLL	Changed NLL
Anchor + switching	$\mathbf{0.934 \pm 0.084}$	$\mathbf{0.654 \pm 0.037}$	$\mathbf{1.191 \pm 0.107}$	$\mathbf{3.020 \pm 0.150}$
No anchor + switching	0.945 ± 0.086	0.667 ± 0.042	—	—
Anchor + fixed trajectory class	1.015 ± 0.111	0.702 ± 0.037	1.261 ± 0.115	3.115 ± 0.145
Anchor + IID mixture	—	—	1.299 ± 0.092	3.230 ± 0.139

J.4 Behavioral-regime count

Table 12 compares regime counts on validation data. The selected $K_u = 4$ for CKD–MBD and $K_u = 3$ for MIMIC–IV are within 0.005 and 0.003 NLL of $K_u = 5$, respectively, with higher minimum occupancy and greater separation.

Stability compares the complete action-effect profiles $R_{u,j}$ (Figure 4). Regimes are matched by maximum mean cosine similarity; the split most similar to the others is the reference, and stability averages the four non-reference scores. Maximum pairwise cosine measures within-split regime similarity, averaged across splits. Minimum occupancy is the smallest validation mean of $\rho^{-}_{m,u}$ across regimes and splits.

Table 12: Validation sensitivity to regime count. Bold K_u marks the selected model. Lower maximum profile cosine indicates greater separation; occupancy and stability are defined in the text.

(a) CKD–MBD

K_u	NLL ↓	Top-1 ↑	Brier ↓	Min. occ. ↑	Stability ↑	Max. cosine ↓
2	1.153 ± 0.083	0.747	0.416	9.18%	0.914	−0.224
3	1.112 ± 0.071	0.745	0.416	9.17%	0.832	0.297
4	1.086 ± 0.070	0.745	0.412	7.10%	0.820	0.460
5	1.081 ± 0.070	0.746	0.413	6.52%	0.854	0.577

(b) MIMIC–IV

K_u	NLL ↓	Top-1 ↑	Brier ↓	Min. occ. ↑	Stability ↑	Max. cosine ↓
2	0.658 ± 0.092	0.844	0.239	12.72%	0.979	0.921
3	0.640 ± 0.088	0.842	0.237	6.52%	0.973	0.933
4	0.640 ± 0.087	0.842	0.238	5.35%	0.954	0.948
5	0.637 ± 0.087	0.841	0.238	5.06%	0.952	0.950

J.5 Sensitivity to the Anchor and Stage I Values

Table 13 tests four anchor variants with the same splits and latent counts. Relative to equation 12, the equal-channel reward sets target-alignment and low/high penalty weights to one and removes joint-risk terms; the amplified-risk reward scales low/high penalties by 1.5 and joint-risk penalties by 2, leaving target-alignment weights unchanged. Both retain treatment burden 0.05. One-step continuation values only the next transition; value noise adds action-centered Gaussian perturbations with SD 20% of the training median action-value range. Reward variants reselect Stage II hyperparameters on validation data; continuation and noise variants refit with the primary split-specific hyperparameters and initializations.

Across variants, mean held-out NLL changes by at most 0.006, fitted-action agreement is at least 99.1%, and profile cosine is 0.972–1.000. Under continuation and value-noise changes, aligned regime assignments agree on 98.1–99.4% of decisions.

Table 13: Anchor sensitivity over five patient/subject splits. NLL: mean ± sample SD. Anchor and fitted action agreement: mean top-action agreement with the primary analysis. Profile cosine: mean similarity of complete w_u vectors after maximum-cosine matching, averaged over regimes and splits.

Dataset	Anchor specification	NLL ↓	Anchor action agr.	Fitted action agr.	Profile cos.
CKD–MBD	Primary	0.934 ± 0.084	100.0%	100.0%	1.000
	Equal-channel reward	0.934 ± 0.083	92.8%	99.6%	0.975
	Amplified-risk reward	0.936 ± 0.081	95.2%	99.8%	0.972
	One-step continuation	0.934 ± 0.085	57.0%	99.3%	0.996
	Stage I value noise (20%)	0.940 ± 0.083	40.3%	99.1%	0.992
MIMIC–IV	Primary	0.654 ± 0.037	100.0%	100.0%	1.000
	Equal-channel reward	0.655 ± 0.037	99.6%	99.7%	0.991
	Amplified-risk reward	0.655 ± 0.037	83.6%	99.8%	1.000
	One-step continuation	0.653 ± 0.037	97.0%	99.8%	0.999
	Stage I value noise (20%)	0.658 ± 0.034	27.4%	99.1%	0.999

K Model Selection and Reproducibility

Unless otherwise stated, the behavioral prediction and robustness analyses use the same five fixed patient/subject-level splits. Table 14 summarizes the model-selection and optimization settings. For each outer split, Stage I is fit on the training partition; the Stage II training values are obtained by five-fold cross-fitting within that partition, while validation and test values use the outer-training Stage I fit. Held-out actions are scored from pre-action histories. Tables report the mean and sample standard deviation across splits. Clinical data and model inputs are described in Appendices E–G; predictive metrics and comparisons are defined in Appendix J.

Table 14: Core model-selection and optimization settings. Candidate actions and normalizers are estimated from training data; hyperparameters are selected on validation data.

Setting	**CKD–MBD**	**MIMIC–IV**
Split unit (70/15/15%)	Patient	Subject
Decision interval / anchor return	Monthly / 6 months ($\delta = 0.99$)	Hourly / 6 hours ($\delta = 0.99$)
Candidate-action vocabulary	20 most frequent training tuples + OTHER	Up to 50 most frequent training tuples + OTHER
Final latent counts	$K_z = 3$, $K_u = 4$	$K_z = 3$, $K_u = 3$
Behavioral-regime candidates	$K_u \in \{2, 3, 4, 5\}$	$K_u \in \{2, 3, 4, 5\}$
Persistence candidates	$\tau \in \{9, 18, 36\}$ months	$\tau \in \{96, 192\}$ hours
Stage II penalties (ℓ_2, ℓ_1)	$(.001, 0), (.002, 0)$ $(.005, 0), (.01, 0)$	$(.0005, 0), (.001, 0)$ $(.002, 0), (.002, .0002)$
Stage II optimization	Global: 1,200 steps; switching: 35 outer iterations, 25 Adam steps per iteration, 2 restarts	Global: 1,200 steps; switching: 20 outer iterations, 15 Adam steps per iteration, 2 restarts
Stage I particle EM	48 particles, 12 smoothing samples; up to 6 epochs, batch size 16	16 particles, 4 smoothing samples; up to 4 epochs, batch size 32
Stage I update steps	8 transition / 12 emission steps per batch; 6 emission warm-up steps	5 transition / 8 emission steps per batch; 4 emission warm-up steps

Regime persistence. Stage II selects the persistence scale, regularization, and restart by validation NLL. For a gap Δ between decisions, the clinical regime transition is

$$\Pi(\Delta) = e^{-\Delta/\tau} I + \left(1 - e^{-\Delta/\tau}\right) \mathbf{1}\pi^\top, \tag{16}$$

where π is the fitted stationary regime distribution. The scale τ is shared by regimes within each fitted model and controls how quickly the previous regime information decays toward π. It is a transition-memory parameter, distinct from the physiological mean dwell time obtained from the Stage I CTMC in Figure 3.

Regime-layer updates and regularization. For each candidate τ, forward–backward uses $\Pi(\Delta)$ at the observed gaps. Writing $n = \sum_i M_i$, the initial and stationary distributions are updated as

$$\mu_u = \frac{\sum_i \gamma_1^i(u) + 1/2}{N + K_u/2}, \qquad \pi_u = \frac{\sum_{i,m} \gamma_m^i(u) + 1/2}{n + K_u/2}.$$

Cost updates use the mean posterior-weighted negative log-likelihood plus $\sum_u [\ell_2 \|w_u\|_2^2/2 + \ell_1 \|w_u\|_1]$, optimized with Adam; the ℓ_1 term uses the sign subgradient. Validation selects τ, penalty coefficients, and restart. All five selected CKD–MBD fits use $\ell_1 = 0$; two of the five MIMIC–IV fits use $\ell_1 = 0.0002$ and the others use zero. Every selected fit uses a positive ℓ_2 coefficient. The occupancy updates are moment estimates for this transition kernel and do not imply EM likelihood monotonicity.

Both neural baselines receive the pre-action covariates, a one-hot encoding of the previous action, and $\log(1 + \text{gap})$ and $\log(1 + \text{decision time})$, standardized using training data. Predictions use the history through the current pre-action context, with a causal attention mask for the Transformer. The GRU uses hidden width 32 or 64, and the causal Transformer uses one layer, four heads, model width 32 or 64, feed-forward width 64 or 128, and dropout 0.1. Both train for at most 120 epochs with patience 12; learning rate, weight decay, architecture, and stopping epoch are selected by validation NLL.

L Limitations and Broader Impacts

Our model is intended as a retrospective modeling and auditing framework rather than an autonomous decision-making system. The recovered regimes are interpreted relative to the specified anchor reward, fitted latent dynamics, and domain feature map. Synthetic recovery tests evaluate this interpretation under known ground truth; clinical sensitivity analyses examine its dependence on the chosen reward and fitted values. Although the two clinical cohorts span different temporal scales, further validation on additional institutions, cohorts, and operational environments would strengthen evidence for broader generalization. In clinical applications, the results are meant to support structured review and hypothesis generation, not direct prospective decision support without expert oversight.

The broader goal of this work is to improve transparency in sequential treatment decisions by linking physiological evolution to persistent behavioral patterns under a shared reference. This helps reviewers examine when treatment choices depart from the reference and which inferred action-cost contributions account for these differences within the patient's course. Potential risks include over-interpreting anchor-optimal counterfactuals as direct recommendations, and practical deployment should include domain expert review, privacy safeguards, and prospective validation.