Structured Behavioral Heterogeneity as Latent Regime Constraints

Yuting Yan[1] **Haozhou Gao**[1] **Xinye Chen**[1] **Yinghao Fu**[2,1] **Shuang Li**[1*]

[1]The Chinese University of Hong Kong, Shenzhen, China
[2]City University of Hong Kong, Hong Kong SAR, China
lishuang@cuhk.edu.cn

Abstract

Sequential decision data often exhibit *structured behavioral heterogeneity*: similar observed conditions can lead to different actions across trajectories and time. Standard approaches typically attribute this variation to drifting rewards or unstructured noise, which can produce models that are difficult to interpret and compare across settings. We instead model this heterogeneity as arising from **latent constraint regimes**, which induce persistent, regime-dependent deviations from a shared decision anchor. We introduce a framework that decomposes behavior into (i) a *shared anchor value function* capturing stable, outcome-relevant decision logic and (ii) a *switching regime process* that generates sparse, interpretable deviations from this anchor. This decomposition separates stable decision structure from systematic variation in observed actions. To construct the anchor under partial observability and irregular decision timing, we use a continuous-time latent-state model that infers belief states from irregularly timed decision data and produces comparable action values across trajectories. The regime layer then captures structured deviations without modifying the underlying objective. Across clinical care and retail pricing datasets, the proposed approach improves held-out action prediction relative to stationary models and interpretable behavioral baselines, while identifying persistent regime assignments and sparse, feature-level deviations. These findings show that structured heterogeneity can be captured by a shared anchor with switching, regime-dependent deviations.

1 Introduction

Sequential decisions in clinical care, dynamic pricing, and operations are commonly formalized as Markov decision processes [1], with offline reinforcement learning used to learn from logged trajectories despite distribution shift [2, 3, 4]. A recurring empirical pattern, however, is *apparent behavioral heterogeneity*: under similar observed conditions, agents take different actions across time and contexts [5, 6].

In ICU sepsis management, for example, patients with nearly identical clinical measurements may still receive different treatment decisions. Importantly, these differences often do not reflect disagreement about medical objectives. Instead, the decision environment is continuously reshaped by latent factors: medication shortages, staffing fluctuations across shifts, evolving hospital protocols, or temporary restrictions on intervention options. The observed behavior therefore reflects not only clinical reasoning, but also persistent structure in the surrounding decision context.

Standard modeling approaches typically absorb such variation into drifting rewards, flexible policy classes, or unstructured noise [7, 8]. While expressive, this view effectively attributes all heterogeneity to changes in *what is optimized*, conflating stable decision logic with systematic distortions induced

*Corresponding author.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

by context. As a result, models may fit behavior well but offer limited insight into why similar situations lead to different actions.

Major paradigms in offline and inverse reinforcement learning explain heterogeneity by recovering different rewards [9, 10], latent tasks [11, 12, 13], or cognitive constraints [14, 15]. Meanwhile, switching state-space models drive behavior through latent regime dynamics [16, 17]. Across these views.Across these views, heterogeneity is embedded in the decision problem itself rather than separated from the mechanisms that systematically distort decisions within a shared structure.

We propose a different perspective: much of observed heterogeneity can be understood as arising from *latent constraint regimes* acting on a shared decision anchor. Here, "constraint" denotes persistent deviations from Q_{anc}, not necessarily literal hard constraints. Instead of allowing rewards or policies to vary freely, we decompose behavior into (*i*) *a shared action-value anchor capturing stable, outcome-relevant decision structure* and (*ii*) *a switching regime process that induces sparse, structured deviations from this anchor*. These regimes are not alternative tasks or objectives; they represent persistent modes of constraint that systematically reshape action selection relative to a common decision rule.

A central difficulty in making this decomposition operational is that *decision data are often partially observed and irregularly timed*. Constructing a meaningful shared anchor therefore requires aligning value estimates across uneven decision processes while preserving comparability. To address this, we infer decision-time belief states using a *continuous-time latent-state model*, which produces consistent state representations under irregular observation patterns [18, 19, 20]. Conditioned on these beliefs, the anchor defines a stable value function, while the regime layer captures deviations without modifying the underlying decision structure.

This separation yields a representation in which heterogeneity is expressed explicitly as structured deviations around a shared anchor, rather than implicitly through changing rewards or latent tasks. Across clinical care and retail pricing datasets, this formulation improves held-out action prediction over stationary models and interpretable behavioral baselines, while recovering persistent regime structure with sparse, feature-level interpretations. Beyond predictive performance, the learned regimes provide a direct lens for understanding systematic departures from shared decision logic in terms of recurring constraint patterns.

Finally, this viewpoint contrasts with prior approaches along a simple axis: rather than explaining heterogeneity by changing the decision problem (rewards, tasks, policies, or latent regimes), *we keep the decision structure fixed and explain variation through regime-dependent constraints acting on top of it*. This leads to a more factored interpretation of sequential behavior, where stability lies in the anchor and heterogeneity lies in structured deviations. An extended discussion of related work is provided in Appendix A.

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 (as illustrated in Figure 1). **Stage I** constructs decision-time latent states and a shared anchor value function. **Stage II** learns latent constraint regimes that modulate action selection relative to this anchor. At a high level, the action score under regime u is

$$S_u(x,a) = \underbrace{Q_{\mathrm{anc}}(x,a)}_{\text{shared anchor value}} - \underbrace{C_u(\xi,a)}_{\text{regime-dependent deviation}},$$

where x is the inferred decision-time state and ξ collects the variables available to the regime layer, such as x itself and optional side information. The anchor Q_{anc} is shared across all trajectories and regimes. Heterogeneity enters only through the regime-dependent deviation C_u. This gives a common coordinate system for comparing behavior: regimes differ by how they modulate the same value anchor, not by changing the value function itself.

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, we observe covariates o_m^i and an action $a_m^i \in \mathcal{A}$, where $\mathcal{A}$ is finite:

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

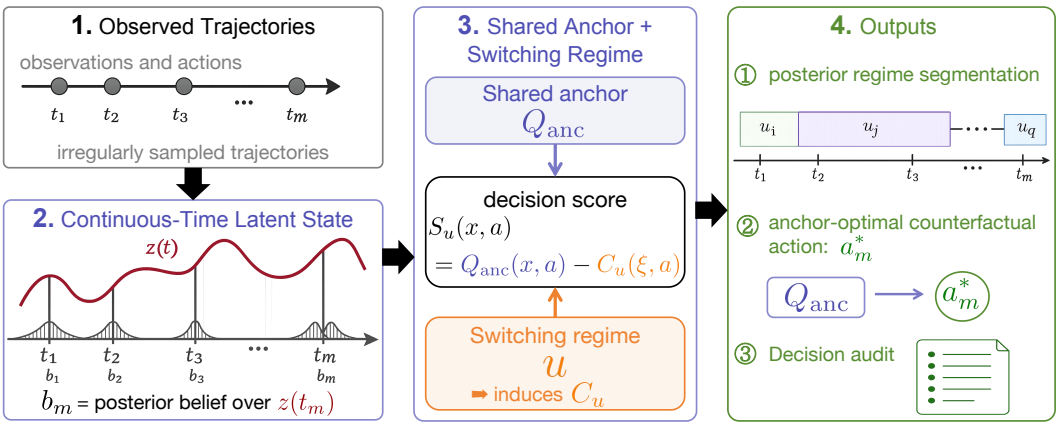


Figure 1: Shared-anchor regime decomposition. Irregular observations are mapped to continuous-time latent beliefs, which define a shared anchor Q_{anc}. A switching regime u_m induces sparse deviations $C_u(\xi, a)$ around this anchor, yielding posterior regime segmentation, anchor-optimal counterfactual actions, and decision audits.

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 about the underlying system when the action is chosen.

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. Observations are generated from a likelihood

$$p_\theta(o_m \mid z(t_m), \lambda(t_m)).$$

Belief updating then takes the standard prediction-correction form

$$\tilde{b}_m = b_{m-1} \exp\{A(a_{m-1})\Delta_{m-1}\},$$
$$b_m(k) \propto \tilde{b}_m(k)\, p_\theta(o_m \mid z(t_m) = k, \lambda(t_m)).$$

The exact emission model is application-specific; the key idea is that Stage I learns latent system dynamics from irregularly observed trajectories, while treating observed actions as inputs that influence state evolution.

Given the fitted dynamics, we define the shared anchor value

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

where r_{anc} is a reference reward and π_0 is a fixed continuation rule. The anchor is not learned separately for each regime. It is fixed before Stage II, so the regime layer cannot explain heterogeneity by changing the underlying value function.

This separation is central. Stage I answers: given the observed history, what is the inferred system state, and how do actions compare under a shared value anchor? Stage II answers: how do observed actions systematically depart from that anchor?

2.3 Stage II: switching latent constraint regimes

Stage II models structured behavioral heterogeneity through a latent switching regime layer. A *constraint regime* is a persistent latent mode under which certain actions become systematically more or less likely relative to the shared anchor value function Q_{anc}. The term "constraint" is used broadly: regimes may reflect unobserved protocol pressure, treatment complexity, operational burden, workload, or other persistent factors that influence action selection without changing the anchor itself.

We introduce a latent regime

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

which evolves along the decision sequence according to

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

The Markov structure encourages regimes to explain coherent behavioral segments rather than isolated action noise. Each regime defines a *deviation function*

$$C_u(\xi, a),$$

which modulates action selection relative to the shared anchor. Here, ξ_m is the introduced *additional context* for explaining behavioral variation, such as protocol indicators, workload proxies, or action-history features, which can contain latent decision state x_m.

Actions are generated according to

$$p(a_m = a \mid x_m, \xi_m, u_m = u) = \frac{\exp\{\beta[Q_{\text{anc}}(x_m, a) - C_u(\xi_m, a)]\}}{\sum_{a' \in \mathcal{A}} \exp\{\beta[Q_{\text{anc}}(x_m, a') - C_u(\xi_m, a')]\}}, \tag{1}$$

where $\beta > 0$ is an inverse-temperature parameter. Intuitively, the constraint term shifts action preferences relative to the shared anchor: some actions become systematically suppressed, while others become more likely under particular regimes. For example, one regime may discourage complex multi-drug changes, while another may suppress actions associated with protocol burden or resource-intensive interventions. The deviation function C_u can take flexible forms. In the main interpretable specification, we use a sparse linear parameterization $C_u(\xi, a) = w_u^\top g_\psi(\xi, a)$, where $g_\psi : \Xi \times \mathcal{A} \to \mathbb{R}^d$ is a feature map and w_u is sparse. Sparsity makes each regime interpretable as a small set of active behavioral deviations.

This parameterization (1) admits a simple decision-theoretic interpretation. For fixed (x, ξ, u), the softmax distribution is the unique solution to

$$\max_{\pi \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} \pi(a) \left[Q_{\text{anc}}(x, a) - C_u(\xi, a)\right] + \frac{1}{\beta} H(\pi),$$

where $H(\pi) = -\sum_a \pi(a) \log \pi(a)$. Thus, regimes act as structured soft constraints that reshape action probabilities around a shared anchor value function.

This distinguishes our approach from mixture-policy or latent-task models. Those methods explain behavioral heterogeneity through separate rewards, policies, or transition structures across latent groups. In contrast, all regimes in our framework share the same latent dynamics, state representation, and anchor value function; only the deviation term C_u changes across regimes.

After fitting, Stage II produces

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

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:

$$\widehat{\Theta}_{\mathrm{dyn}} \in \arg\max_{\Theta_{\mathrm{dyn}}} \sum_{i=1}^{N} \log p_{\Theta_{\mathrm{dyn}}}\big(o^i_{1:M_i} \mid a^i_{1:M_i-1}, t^i_{1:M_i}\big)\,.$$

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

$$\widehat{x}^i_m = (\widehat{b}^i_m, \widehat{\lambda}^i_m)$$

at every decision epoch. We then evaluate

$$\widehat{Q}_{\mathrm{anc}}(\widehat{x}^i_m, a), \qquad a \in \mathcal{A},$$

using dynamic programming, numerical integration, or Monte Carlo rollouts under the fitted dynamics and fixed continuation rule π_0.

The output of Stage I is the table

$$\Big\{\widehat{x}^i_m,\ \widehat{Q}_{\mathrm{anc}}(\widehat{x}^i_m, a) : i = 1, \ldots, N,\ m = 1, \ldots, M_i,\ a \in \mathcal{A}\Big\}\,.$$

These quantities are held fixed in Stage II, ensuring that learned regimes are deviations from a common anchor rather than alternative value functions.

3.2 Stage II: fitting switching constraint regimes

Stage II estimates the latent regime process and the regime-dependent deviations. For the sparse linear specification,

$$C_u(\xi, a) = w_u^{\top} g_\psi(\xi, a), \tag{2}$$

the regime parameters are

$$\Theta_{\mathrm{reg}} = \{\mu, \Pi, w_1, \ldots, w_{K_u}, \psi\},$$

where μ is the initial regime distribution, Π is the transition matrix, w_u are regime-specific deviation weights, and ψ parameterizes the feature map.

Conditional on Stage I outputs, the action likelihood is an HMM with softmax emissions:

$$\mathcal{L}(\Theta_{\mathrm{reg}}) = \sum_{i=1}^{N} \log \sum_{u^i_{1:M_i}} \mu_{u^i_1} \prod_{m=2}^{M_i} \Pi_{u^i_{m-1}, u^i_m} \prod_{m=1}^{M_i} p_{\Theta_{\mathrm{reg}}}(a^i_m \mid \widehat{x}^i_m, \widehat{\xi}^i_m, u^i_m),$$

where $p_{\Theta_{\mathrm{reg}}}(a_m \mid \widehat{x}_m, \widehat{\xi}_m, u)$ is evaluated as (1).

We maximize the sparsity-regularized objective using generalized EM. In the E-step, forward–backward computes posterior regime probabilities

$$\gamma^i_m(u) = \Pr(u^i_m = u \mid \mathcal{D}^i)$$

and expected transition counts. In the M-step, μ and Π are updated from normalized expected counts. The deviation parameters are updated by solving

$$\max_{\{w_u\},\psi} \sum_{i,m} \sum_{u=1}^{K_u} \gamma^i_m(u) \log p_{\Theta_{\mathrm{reg}}}(a^i_m \mid \widehat{x}^i_m, \widehat{\xi}^i_m, u) - \zeta \sum_{u=1}^{K_u} \|w_u\|_1.$$

For hand-crafted feature g_ψ, this is a posterior-weighted sparse multinomial logit problem. If g_ψ contains learned components (e.g., simple neural networks), ψ is updated by gradient steps inside the generalized M-step, with objective improvement checked after each update.

The single-regime $K_u = 1$ is an important ablation. It allows one global deviation around the anchor but does not permit switching behavioral modes. Comparing $K_u = 1$ with $K_u > 1$ tests whether the data contain persistent heterogeneity that cannot be represented by a single deviation profile.

3.3 What the theory guarantees

We summarize the main theoretical consequences because they clarify what can be interpreted from the fitted model. Formal assumptions and proofs are provided in the appendix.

Identifiable action-level deviations. Condition on the anchor Q_{anc} and feature map g_ψ. Under standard finite-state HMM identifiability conditions, the regime transition law and emissions are

identifiable up to permutation of regime labels. Within each regime, the identifiable quantities are action-dependent deviation differences. This follows from the log-odds identity

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

Since Q_{anc} is fixed, the data identify how each regime shifts action scores relative to the anchor, up to action-independent offsets.

Sparse recovery of regime profiles. For the sparse linear specification $C_u(\xi, a) = w_u^\top g_\psi(\xi, a)$, suppose the true regime profiles are s-sparse and the posterior-weighted logit objective satisfies standard restricted strong convexity and score concentration conditions. Then, with an appropriate choice of ζ,

$$\sum_{u=1}^{K_u} \|\widehat{w}_u - w_u^\star\|_2 \lesssim K_u \sqrt{\frac{s \log d}{N \bar{M}}},$$

where d is the feature dimension and $\bar{M} = N^{-1} \sum_i M_i$. This supports reporting compact feature-level regime profiles rather than treating regimes as uninterpreted latent clusters.

Stability to Stage I approximation error. Stage II uses plug-in estimates of the state and anchor values. Let

$$\epsilon_Q = \sup_{m,a} \left| \widehat{Q}_{\mathrm{anc}}(\widehat{x}_m, a) - Q_{\mathrm{anc}}(x_m, a) \right|, \qquad \epsilon_x = \sup_m \|\widehat{x}_m - x_m\|_1.$$

If Q_{anc} and g_ψ are Lipschitz, and $\|w_u\|_1 \le W$, then for fixed regime parameters,

$$\sup_{m,u} \left| \log \widehat{p}(a_m \mid \widehat{x}_m, \widehat{\xi}_m, u) - \log p(a_m \mid x_m, \xi_m, u) \right| \le 2\beta \left[\epsilon_Q + (L_Q + W L_g) \epsilon_x \right].$$

Thus, moderate Stage I error induces a controlled perturbation of the Stage II likelihood.

Why switching regimes can improve fit. The $K_u = 1$ model is nested inside the $K_u > 1$ model. If the true conditional action law contains separated persistent regimes that are not observationally equivalent to a single deviation profile, then the population likelihood of the switching model is strictly larger than that of the single-regime model by a positive KL gap. This justifies the empirical comparison between one shared deviation profile and multiple latent constraint regimes.

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-optimal action

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

which provides a reference point for comparing observed actions with the shared anchor. 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 our model on two clinical settings with markedly different timing scales: high-frequency ICU sepsis treatment from MIMIC-IV [21, 22] and long-horizon chronic care from a private CKD-MBD cohort; we further use retail pricing from Dominick's Finer Foods (DFF) [23, 24] as a cross-domain validation. Detailed data summaries are provided in Appendix D. We ask whether the constraint-regime layer (i) recovers interpretable, persistent deviations from a shared anchor, (ii) improves held-out action prediction over stationary and interpretable behavioral baselines while remaining robust to reward specifications, and (iii) supports counterfactual anchor-optimal actions under constraint relaxation and case-level auditing.

4.1 Interpretable Phenotypes and Constraint Regimes

We examine interpretability in both latent layers: outcome-grounded phenotypes and sparse anchor-relative wedges.

Fitted dynamics layer and recovered phenotypes. The fitted continuous-time dynamics layer recovers K_z outcome-grounded phenotypes, providing an interpretable view of patient heterogeneity through phenotype signatures, dwell times, and population trajectories, with $\lambda(t)$ dynamics and detailed experimental results reported in Appendices C and F. The proxy $\lambda(t)$ tracks Ca/P/PTH for CKD-MBD, hemodynamic/metabolic indicators for MIMIC-IV, and market-state indicators for DFF. For example, in CKD-MBD, the recovered phenotypes show clinically coherent Ca/P/PTH signatures, separate stable from transient high-risk states through distinct dwell times, and trend toward more stable mineral-control profiles over the treatment horizon.

Regime layer and domain-grounded feature maps. We fit the regime layer in each domain using the sparse wedge $C_u(\xi, a) = w_u^\top g_\psi(\xi, a)$, where ξ may include $x = (b, \lambda)$ and optional side information. The feature map g_ψ is shared across regimes, while only w_u is regime-specific; because a-independent features cancel in the choice rule (1), we construct g_ψ from action-side and context-action features.

The instantiated g_ψ is 16-dim for MIMIC-IV (e.g., Sepsis-3 compliance, regimen complexity, and therapeutic-modality intensity) [25], 8-dim for CKD-MBD (e.g., mineral-imbalance safety synergies, polypharmacy burden, and treatment-escalation inertia) [26], and 8-dim for DFF (e.g., menu-cost inertia, competitive-response gaps, and brand-equity guardrails) [27]; full definitions are in Appendix E.

Recovering interpretable regimes. For MIMIC-IV with $K_u = 3$, Figure 2 shows sparse and clinically interpretable regime profiles, with elevated transition diagonals ($\Pi_{uu} \in [0.43, 0.62]$) indicating persistent regimes rather than per-step noise:

- Regime 1, acute escalation: high weights on `added_count` and `n_groups_active`, indicating decisions in which clinicians introduce new treatments and activate multiple therapeutic groups. We interpret this as an acute stabilization mode.
- Regime 0, protocol adherence: concentration on `forbid_violate`, which encodes contraindication or safety-rule violations. This regime captures protocol- and safety-shaped action selection.
- Regime 2, complexity management: weights on `drug_count` and `removed_count`, marking decisions involving regimen size and treatment removal. We interpret this as a complexity-management or de-escalation mode.

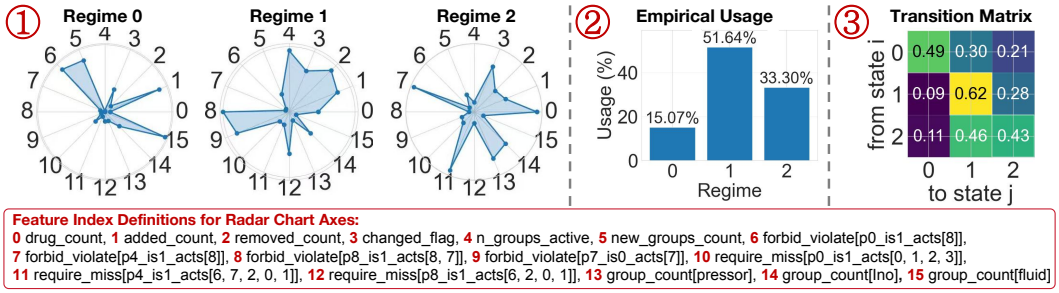


Figure 2: Proposed model on MIMIC-IV ($K_u = 3$). **(1)** Per-regime weights w_u on the 16-dimensional feature map. **(2)** Posterior regime usage $\sum_m \gamma_m(u)$. **(3)** Inter-regime transition matrix Π.

Coupled and decoupled regime–state dynamics. Trajectory-level diagnostics (Figure 3) show three qualitative patterns, motivating the model's separation of decision-clock regimes from continuous-time dynamics.

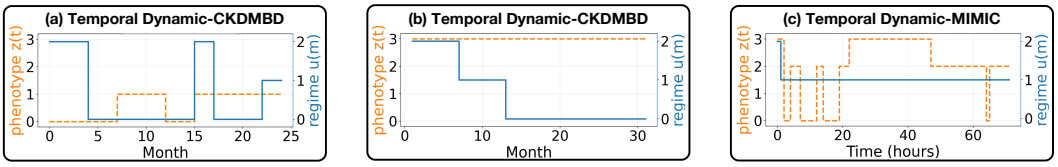


Figure 3: Representative test-set trajectories of latent state $z(t)$ and regime path u_m ($K_z = 4, K_u = 3$). **(a)** A regime transition co-occurs with a $z(t)$ shift. **(b)** A regime transition occurs under stable $z(t)$. **(c)** The regime remains stable despite frequent $z(t)$ shifts.

4.2 Predictive Performance and Robustness

Having established interpretability, we next evaluate held-out predictive performance and robustness along four axes: **interpretable behavioral competitors**, **switching-regime depth**, **reward-specification robustness** , and **continuous-time dynamics validation**.

Interpretable behavioral baselines. We compare our model with Maximum Entropy IRL (MaxEnt IRL) [10], which fits a stationary reward representation, and Soft Decision Trees (SDT) [28], which provide discriminative hierarchical decision rules. As shown in Table 1, Our model achieves the best held-out log-likelihood and Acc@5 in all three domains. The baselines capture different parts of the problem but not the full decomposition: MaxEnt IRL cannot represent persistent regime-dependent wedges around a shared anchor, while SDT does not provide a generative regime process or anchor-optimal counterfactual actions.

Table 1: Held-out action-prediction performance across clinical and economic domains.

Model	CKD-MBD Log-L / Acc@5 (%)	MIMIC-IV Log-L / Acc@5 (%)	DFF Log-L / Acc@5 (%)
Ours*	$\mathbf{-4.52 \pm 0.85 \,/\, 50.5 \pm 5.3}$	$\mathbf{-0.74 \pm 0.02 \,/\, 96.8 \pm 0.2}$	$\mathbf{-0.64 \pm 0.02 \,/\, 98.5 \pm 0.1}$
MaxEnt IRL†	$-5.64 \pm 0.00 \,/\, 29.3 \pm 0.7$	$-5.29 \pm 0.07 \,/\, 80.1 \pm 0.7$	$-3.22 \pm 0.00 \,/\, 70.8 \pm 2.7$
SDT‡	$-5.23 \pm 0.25 \,/\, 15.4 \pm 4.2$	$-1.84 \pm 0.00 \,/\, 93.4 \pm 0.0$	$-2.35 \pm 0.06 \,/\, 83.8 \pm 1.0$

* Anchor-optimal action (a^*), constraint regimes (u), and latent phenotypes.
† Stationary state-action reward $r(s, a)$ on an empirical finite-horizon MDP.
‡ Hierarchical discriminative decision paths.

Predictive gain from switching regimes. Figure 4 reports held-out performance as K_u varies from 1 (stationary wedge baseline) to 6 on CKD-MBD, MIMIC-IV, and DFF. Allowing $K_u > 1$ consistently improves log-likelihood and top-k accuracy, with gains saturating around $K_u = 3$–5, indicating predictive structure not captured by a single stationary deviation profile.

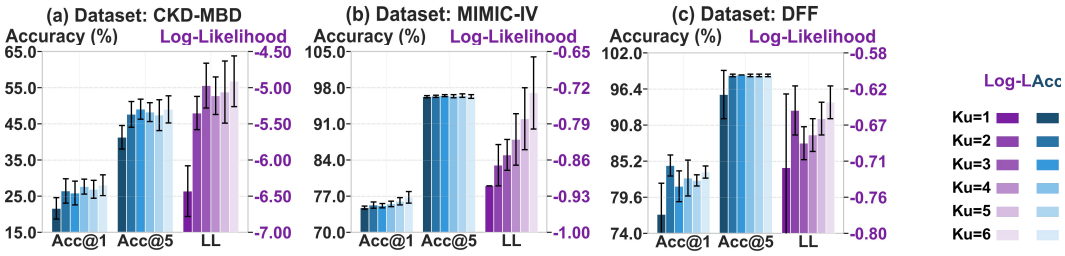


Figure 4: Held-out action prediction across $K_u \in \{1, \ldots, 6\}$ for the three datasets. $K_u = 1$ is the stationary wedge baseline; $K_u > 1$ allows switching constraint regimes.

Reward-specification robustness. We next evaluate whether the predictive advantage depends on a particular reference reward. We refit our model under three reward families: outcome-based, stability-based, and risk-sensitive. Across these specifications, switching-regime models ($K_u > 1$) consistently improve held-out likelihood over the stationary $K_u = 1$ baseline, indicating that the gains are not artifacts of a single reward design. Reward definitions and metrics are in Appendix G.

Dynamics-layer validation. Finally, we evaluate the fitted continuous-time dynamics layer, which supplies the anchor values Q_{anc} used by the regime model and counterfactual diagnostics; errors in this layer would propagate directly to all anchor-relative analyses. Against ODE-RNN [19] and NCDSSM [20], the CTMC–ODE layer remains competitive on hold-out prediction while retaining interpretable latent phenotypes; detailed dynamics benchmarks are provided in Appendix H.

4.3 Counterfactual Diagnostics and Decision Auditing

After the two estimation stages, a diagnostic layer turns the value anchor and constraint-regime layers into counterfactual diagnostics at two levels. At the **population level**, *constraint relaxation* scales down selected wedge groups and reads off the change in Q_{anc}, ranking which regime-dependent

deviations carry the largest recoverable value. At the **unit level**, we synthesize **anchor-optimal counterfactual actions** $\{a_m^*\}$, i.e., the ideal per-trajectory actions under the shared anchor, via an SCM-grounded abduction-action-prediction rollout; decision auditing then interprets each gap $a_m - a_m^*$ through the active regime u_m and latent state z_m. Together these diagnostics support policy refinement, stress-testing, and case-level review, conditional on the fitted shared anchor.

Population-level constraint relaxation. Constraint relaxation produces an anchor-relative ranking of intervention targets, indicating which wedge groups would yield the largest recoverable gain in Q_{anc} if reduced. On MIMIC-IV, we partition the 16-dimensional wedge feature map into three clinical groups: *Compliance* (protocol violations and required-treatment omissions), *Complexity* (regimen size and step-to-step treatment changes), and *Resource* (high-intensity therapeutic breadth). For each group, we scale weights by $(1-\alpha)$, $\alpha \in [0,1]$, and report the relative increase in Q_{anc}. Figure 5 shows that *Complexity* dominates: full relaxation ($\alpha = 1$) recovers 13.5% of Q_{anc}, versus 6.6% for *Compliance* and 7.0% for *Resource*, identifying treatment complexity and step-to-step adjustment as the largest recoverable deviation in MIMIC-IV.

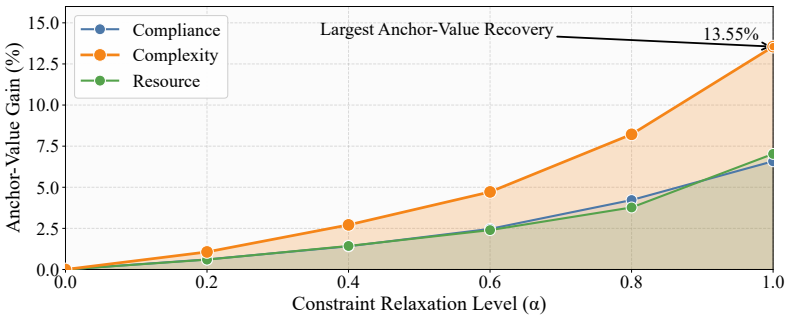


Figure 5: Anchor-value recovery under group-wise constraint relaxation for MIMIC-IV. Curves report the relative increase in Q_{anc} as α increases.

anchor-optimal counterfactual synthesis. A core capability of our model, beyond imitation-based [29] or regression-based [30] recommenders, is the synthesis of *anchor-optimal counterfactual action sequences* $\{a_m^*\}$ via an **SCM-grounded abduction-action-prediction cycle** on each observed trajectory. The cycle abducts the exogenous noises of $(z(t), \lambda(t))$ through the fitted action-modulated CTMC and outcome ODE, forward-rolls candidate actions with these noises held fixed, and selects the Q_{anc}-maximizing action as a_m^*. Full derivation and sampling are in Appendix I. The resulting $\{a_m^*\}$ are unit-level consistent: each counterfactual respects the individual's observed history rather than averaging across the population. The gap $a_m - a_m^*$ supplies the anchor-relative quantities used in the unit-level auditing below.

Unit-level auditing. While population-level relaxation identifies which wedge groups are most influential overall, for individual trajectories we pair the synthesized $\{a_m^*\}$ with an LLM-based reporting layer. The audit input consists of five structured quantities: observations o_m, observed actions a_m, anchor-optimal actions a_m^*, inferred regimes u_m, and latent states z_m. The LLM receives these structured model outputs and produces a three-part audit summarizing (i) whether a_m^* is domain-plausible and operationally feasible, (ii) which constraint-regime wedges explain the deviation between a_m and a_m^*, and (iii) whether regime and state transitions align with recorded trajectory events. The LLM is used only as a reporting layer. Appendix J presents a representative CKD-MBD trajectory audit, and Appendix K provides the full prompt templates.

5 Conclusion

We introduced a shared-anchor regime-decomposition framework for modeling nonstationary sequential decisions through a common action-value anchor and switching constraint-regime deviations. This separation represents persistent behavioral variation as sparse, interpretable wedges rather than arbitrary policy drift, enabling posterior regime segmentation, constraint-relaxation diagnostics, and SCM-grounded anchor-optimal counterfactuals. Across clinical care and retail pricing, the model improves held-out action prediction over stationary and interpretable behavioral baselines while retaining an auditable counterfactual interpretation.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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*, pages 1843–1854. PMLR, 2020.

[8] 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.

[9] Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

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

[11] 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.

[12] 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.

[13] 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.

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

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

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

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

[18] 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.

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

[20] 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*, pages 926–951. PMLR, 2023.

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

[22] 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.

[23] Stephen J Hoch, Byung-Do Kim, Alan L Montgomery, and Peter E Rossi. Determinants of store-level price elasticity. *Journal of marketing Research*, 32(1):17–29, 1995.

[24] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning: with applications in R*, volume 103. Springer, 2013.

[25] Andrew Rhodes, Laura E Evans, Waleed Alhazzani, Mitchell M Levy, Massimo Antonelli, Ricard Ferrer, Anand Kumar, Jonathan E Sevransky, Charles L Sprung, Mark E Nunnally, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. *Intensive care medicine*, 43(3):304–377, 2017.

[26] 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.

[27] Daniel Levy, Mark Bergen, Shantanu Dutta, and Robert Venable. The magnitude of menu costs: direct evidence from large us supermarket chains. *The Quarterly Journal of Economics*, 112(3): 791–824, 1997.

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

[29] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[30] Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

[31] 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*, pages 147–163. PMLR, 2017.

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

[33] 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.

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

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

[36] 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.

[37] 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.

[38] Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

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

[40] 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.

[41] 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.

[42] Zining Zhu, Hanjie Chen, Xi Ye, Qing Lyu, Chenhao Tan, Ana Marasovic, and Sarah Wiegreffe. Explanation in the era of large language models. In Rui Zhang, Nathan Schneider, and Snigdha Chaturvedi, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 19–25, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-tutorials.3. URL `https://aclanthology.org/2024.naacl-tutorials.3/`.

[43] Alexandros Vassiliades, Nikolaos Polatidis, Stamatios Samaras, Sotiris Diplaris, Ignacio Cabrera Martin, Yannis Manolopoulos, Stefanos Vrochidis, and Ioannis Kompatsiaris. Utilizing large language models for machine learning explainability, 2025. URL `https://arxiv.org/abs/2510.06912`.

[44] Ansgar Koene, Chris Clifton, Yohko Hatada, Helena Webb, and Rashida Richardson. A governance framework for algorithmic accountability and transparency. 2019.

[45] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

[46] 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.

[47] Suchi Saria. Individualized sepsis treatment using reinforcement learning. *Nature medicine*, 24 (11):1641–1642, 2018.

[48] 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.

[49] Adam E Gaweda, Eleanor D Lederer, and Michael E Brier. Artificial intelligence–guided precision treatment of chronic kidney disease–mineral bone disorder. *CPT: Pharmacometrics & Systems Pharmacology*, 11(10):1305–1315, 2022.

[50] Wei Chen and David A Bushinsky. Kdigo ckd–mbd guideline update: evolution in the face of uncertainty. *Nature Reviews Nephrology*, 13(10):600–602, 2017.

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 [2, 3]. In healthcare, MDP and deep RL models have been used to derive treatment policies from MIMIC sepsis trajectories [4, 31, 32], but subsequent work emphasizes that clinician actions are heterogeneous, confounded, and not necessarily an optimal gold standard [33]. Off-policy evaluation provides tools for estimating policy value from observational logs [34, 35], and hidden-parameter MDPs model patient-level variation during transfer [36]. 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 [9, 10], while later work handles diverse demonstrations through latent contexts, mixtures, or heterogeneous environments [11, 12, 13]. Cooperative IRL and bounded-rational-agent models further recognize that observed behavior can reflect uncertainty, limited inference, or interaction between human and algorithmic objectives [14, 15]. These methods usually explain heterogeneity by changing rewards, policies, or latent task identities. Our model instead keeps a shared anchor fixed and assigns systematic behavioral variation to a sparse switching deviation layer. This makes the learned regimes closer to auditable decision pressures than to opaque policy clusters, aligning with calls for interpretable models in high-stakes settings [37, 28].

Finally, 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 [5, 6, 8, 7]; switching state-space models provide a classical latent-regime mechanism for piecewise-stationary structure [17, 16]. For irregularly sampled clinical records, latent ODEs and neural continuous-discrete state-space models learn dynamics without forcing observations onto an artificial grid [18, 19, 20]. Counterfactual time-series methods and structural causal models provide semantics for asking how outcomes or values would change under alternative interventions [38, 39, 40, 41]. The proposed framework couples these threads by placing the anchor on a continuous-time latent state while letting a discrete regime process govern deviations on the decision clock; the resulting anchor-relative counterfactuals can then be summarized by an LLM-based auditor for case-level explanation [42, 43, 44].

B Theoretical Properties and Proofs

This appendix formalizes the theoretical claims used in Section 3. 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 $Q_{\mathrm{anc}}(x_m, a)$. The regime-dependent deviation is denoted by $C_u(\xi_m, a)$, with sparse linear specification $C_u(\xi, a) = w_u^\top g_\psi(\xi, a)$.

We treat Q_{anc}, β, and g_ψ as fixed when proving identifiability and sparse recovery. If β is estimated jointly, a scale normalization is required; otherwise the product $\beta[Q_{\mathrm{anc}} - C_u]$ is identifiable but its scale is not separately determined.

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 latent constraint regimes. A regime $u_m \in \{1, \ldots, K\}$ evolves as a Markov chain with initial distribution μ and transition matrix Π. Conditional on x_m, ξ_m, and $u_m = u$, the action emission model is

$$e_u(a \mid x, \xi) := p_\Theta(a \mid x, \xi, u) = \frac{\exp\{\beta[Q_{\mathrm{anc}}(x, a) - C_u(\xi, a)]\}}{\sum_{a' \in \mathcal{A}} \exp\{\beta[Q_{\mathrm{anc}}(x, a') - C_u(\xi, a')]\}}. \tag{3}$$

For a trajectory of length M, the conditional HMM likelihood is

$$p_\Theta(a_{1:M} \mid x_{1:M}, \xi_{1:M}) = \sum_{u_{1:M}} \mu_{u_1} \prod_{m=2}^{M} \Pi_{u_{m-1}, u_m} \prod_{m=1}^{M} e_{u_m}(a_m \mid x_m, \xi_m). \tag{4}$$

In the sparse linear model, $C_u(\xi, a) = w_u^\top g_\psi(\xi, a)$, where $g_\psi : \Xi \times \mathcal{A} \to \mathbb{R}^d$ is a fixed feature map and $w_u \in \mathbb{R}^d$.

Two parameter values are considered equivalent if they induce the same conditional distribution of $a_{1:M}$ given $x_{1:M}, \xi_{1:M}$, 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-deviation)**.** *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 (3) *is the unique maximizer of*

$$\max_{\pi \in \Delta(\mathcal{A})} \left\{ \sum_{a \in \mathcal{A}} \pi(a) S_u(x, \xi, a) + \frac{1}{\beta} H(\pi) \right\}, \qquad H(\pi) = -\sum_{a \in \mathcal{A}} \pi(a) \log \pi(a). \tag{5}$$

Proof. The objective in (5) is strictly concave on the simplex because the entropy term is strictly concave and $\beta > 0$. Form the Lagrangian

$$\mathcal{L}(\pi, \nu) = \sum_a \pi(a) S_u(x, \xi, a) - \frac{1}{\beta} \sum_a \pi(a) \log \pi(a) + \nu \left(\sum_a \pi(a) - 1 \right).$$

At the interior optimum,

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

Thus $\log \pi(a) = \beta[S_u(x, \xi, a) + \nu] - 1$, so $\pi(a) \propto \exp\{\beta S_u(x, \xi, a)\}$. Normalizing over $\mathcal{A}$ gives (3). Strict concavity gives uniqueness. □

This proposition justifies viewing $C_u(\xi, a)$ as a regime-specific soft constraint or deviation term: the regime does not replace Q_{anc}; it modifies action probabilities around that shared anchor.

B.3 Identifiability of regime-dependent deviation contrasts

We next state conditions under which the learned regimes and action-level deviation contrasts are identifiable.

Assumption 1 (Conditional HMM identifiability)**.** Conditional on the context sequence (x_m, ξ_m), the finite-state HMM in (4) is identifiable up to permutation of hidden-state labels. A sufficient condition is that Π is full rank and ergodic, regimes induce distinct emission distributions, and the emission vectors $\{e_u(\cdot \mid x, \xi)\}_{u=1}^K$ are linearly independent on a set of contexts with positive probability.

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 (Feature-difference rank for the sparse linear model)**.** For the sparse linear specification, define the feature-difference set

$$\mathcal{G} = \{g_\psi(\xi, a) - g_\psi(\xi, a') : (\xi, a, a') \text{ in the support}\}.$$

The identifiable part of w_u is its projection onto $\mathrm{span}(\mathcal{G})$. Full identification of w_u requires that the corresponding feature-difference design has full rank on the active coordinates.

Theorem 2 (Identifiability of deviation contrasts)**.** *Under Assumptions 1–2, and conditional on Q_{anc}, β, and g_ψ, the regime transition law and emissions are identifiable up to permutation of regime labels. For each identified regime u, the action-level deviation contrasts*

$$C_u(\xi, a) - C_u(\xi, a')$$

are identifiable for all (ξ, a, a') in the support, up to the same label permutation. In the sparse linear model, w_u is identifiable modulo the null space of the feature-difference design in Assumption 3; if the active feature-difference design has full rank, then w_u is identifiable on its active coordinates.

Proof. By Assumption 1, the transition matrix Π and the emission kernels $\{e_u(\cdot \mid x, \xi)\}_{u=1}^K$ are identified up to a common permutation of regime labels.

Fix an identified regime u and context (x, ξ). For any two actions $a, a' \in \mathcal{A}$, the log-odds under (3) satisfy

$$\log \frac{e_u(a \mid x, \xi)}{e_u(a' \mid x, \xi)} = \beta \left[Q_{\mathrm{anc}}(x, a) - Q_{\mathrm{anc}}(x, a') - \{C_u(\xi, a) - C_u(\xi, a')\} \right]. \tag{6}$$

The left-hand side is identified from the emission distribution, and Q_{anc} and β are fixed. Therefore $C_u(\xi, a) - C_u(\xi, a')$ is identified.

No stronger statement is possible without additional restrictions, because adding any action-independent offset $\kappa_u(\xi)$ to $C_u(\xi, a)$ leaves all action probabilities unchanged. For the sparse

linear model, if two weight vectors w_u and w'_u induce the same emissions, then for every feature difference in the support,

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

Thus $w_u - w'_u$ lies in the null space of the feature-difference design. Full rank on the active coordinates removes this ambiguity and identifies w_u on those coordinates. □

B.4 Stability to Stage-I plug-in error

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 3 (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 $q_m(a) = Q_{\text{anc}}(x_m, a)$ and $\widehat{q}_m(a) = \widehat{Q}_{\text{anc}}(\widehat{x}_m, a)$. 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 4 (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 sparse linear model, if $\|w_u\|_1 \le W$ and $g_\psi(\cdot, a)$ is L_g-Lipschitz in ξ, then one may take $L_C \le W L_g$.

Proof. Fix m and u. Define scaled utility vectors over actions:

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

Then

$$\log \widehat{e}_u(a_m \mid \widehat{x}_m, \widehat{\xi}_m) - \log e_u(a_m \mid x_m, \xi_m) = (r_{a_m} - s_{a_m}) - \left(\log \sum_a e^{r_a} - \log \sum_a e^{s_a} \right).$$

By Lemma 3, the absolute value is at most $2\|r - s\|_\infty$. For any action a,

$$|r_a - s_a| \le \beta \left(|\widehat{q}_m(a) - q_m(a)| + |C_u(\widehat{\xi}_m, a) - C_u(\xi_m, a)| \right) \le \beta(\epsilon_Q + L_C \epsilon_\xi).$$

Taking the supremum over a, m, u gives the result. In the sparse linear case,

$$|C_u(\widehat{\xi}, a) - C_u(\xi, a)| = |w_u^\top \{g_\psi(\widehat{\xi}, a) - g_\psi(\xi, a)\}| \le \|w_u\|_1 L_g \|\widehat{\xi} - \xi\|_1 \le W L_g \|\widehat{\xi} - \xi\|_1.$$

□

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

B.5 Sparse recovery for regime profiles

We state a standard oracle inequality for the sparse linear Stage-II M-step. This result is conditional on fixed responsibilities $\gamma_m^i(u)$ and a fixed feature map g_ψ. It therefore describes the weighted sparse-logit subproblem solved inside generalized EM; global finite-sample guarantees for the nonconvex HMM likelihood would require additional initialization and landscape assumptions.

Let $n = \sum_{i=1}^N M_i = N\bar{M}$. For regime u, define the normalized negative weighted log-likelihood

$$\mathcal{R}_u(w) = -\frac{1}{n} \sum_{i,m} \gamma_m^i(u) \log e_{u,w}(a_m^i \mid \widehat{x}_m^i, \widehat{\xi}_m^i),$$

where $e_{u,w}$ denotes (3) with $C_u(\xi, a) = w^\top g_\psi(\xi, a)$. The penalized estimator is

$$\widehat{w}_u \in \arg\min_w \{\mathcal{R}_u(w) + \zeta \|w\|_1\}.$$

Assumption 4 (Sparse truth and bounded features)**.** The oracle target $w_u^\star$ is s-sparse with support S_u, and $\|g_\psi(\xi, a)\|_\infty \le G$ almost surely.

Assumption 5 (Score concentration)**.** The regularization level satisfies

$$\|\nabla \mathcal{R}_u(w_u^\star)\|_\infty \le \zeta/2$$

with high probability.

Assumption 6 (Restricted strong convexity)**.** For all Δ in the cone $\|\Delta_{S_u^c}\|_1 \le 3\|\Delta_{S_u}\|_1$,

$$\mathcal{R}_u(w_u^\star + \Delta) - \mathcal{R}_u(w_u^\star) - \langle \nabla \mathcal{R}_u(w_u^\star), \Delta \rangle \ge \frac{\kappa}{2}\|\Delta\|_2^2.$$

Theorem 5 (Oracle inequality for sparse regime deviations)**.** *Under Assumptions 4–6,*

$$\|\widehat{w}_u - w_u^\star\|_2 \le \frac{3\zeta\sqrt{s}}{\kappa}, \qquad \|\widehat{w}_u - w_u^\star\|_1 \le \frac{12\zeta s}{\kappa}.$$

If $\zeta \asymp G\sqrt{\log d/n}$, then summing over K regimes gives

$$\sum_{u=1}^{K} \|\widehat{w}_u - w_u^\star\|_2 \lesssim K\sqrt{\frac{s \log d}{N\bar{M}}}.$$

Proof. Let $\Delta = \widehat{w}_u - w_u^\star$. By optimality of $\widehat{w}_u$,

$$\mathcal{R}_u(w_u^\star + \Delta) + \zeta\|w_u^\star + \Delta\|_1 \le \mathcal{R}_u(w_u^\star) + \zeta\|w_u^\star\|_1.$$

Thus

$$\mathcal{R}_u(w_u^\star + \Delta) - \mathcal{R}_u(w_u^\star) \le \zeta(\|w_u^\star\|_1 - \|w_u^\star + \Delta\|_1).$$

Using the support S_u, the right-hand side is at most

$$\zeta(\|\Delta_{S_u}\|_1 - \|\Delta_{S_u^c}\|_1).$$

By convexity and the score bound,

$$\mathcal{R}_u(w_u^\star + \Delta) - \mathcal{R}_u(w_u^\star) \ge \langle \nabla \mathcal{R}_u(w_u^\star), \Delta \rangle \ge -\frac{\zeta}{2}\|\Delta\|_1.$$

Combining these inequalities yields

$$-\frac{\zeta}{2}(\|\Delta_{S_u}\|_1 + \|\Delta_{S_u^c}\|_1) \le \zeta(\|\Delta_{S_u}\|_1 - \|\Delta_{S_u^c}\|_1),$$

which implies the cone condition $\|\Delta_{S_u^c}\|_1 \le 3\|\Delta_{S_u}\|_1$.

Applying restricted strong convexity,

$$\frac{\kappa}{2}\|\Delta\|_2^2 \le \mathcal{R}_u(w_u^\star + \Delta) - \mathcal{R}_u(w_u^\star) - \langle \nabla \mathcal{R}_u(w_u^\star), \Delta \rangle.$$

Using the basic inequality again and the score bound gives

$$\frac{\kappa}{2}\|\Delta\|_2^2 \le \frac{3\zeta}{2}\|\Delta_{S_u}\|_1 \le \frac{3\zeta}{2}\sqrt{s}\|\Delta\|_2.$$

Therefore $\|\Delta\|_2 \le 3\zeta\sqrt{s}/\kappa$. The cone condition then gives

$$\|\Delta\|_1 = \|\Delta_{S_u}\|_1 + \|\Delta_{S_u^c}\|_1 \le 4\|\Delta_{S_u}\|_1 \le 4\sqrt{s}\|\Delta\|_2 \le \frac{12\zeta s}{\kappa}.$$

The stated rate follows from the standard choice $\zeta \asymp G\sqrt{\log d/n}$ and $n = N\bar{M}$. □

B.6 Consistency of the regime-layer MLE

Let

$$\ell_i(\Theta) = \log p_\Theta(a_{1:M_i}^i \mid x_{1:M_i}^i, \xi_{1:M_i}^i)$$

denote the conditional regime-layer log-likelihood contribution for trajectory i.

Assumption 7 (Sampling and trajectory lengths)**.** Trajectories $\{(x_{1:M_i}^i, \xi_{1:M_i}^i, a_{1:M_i}^i)\}_{i=1}^N$ are independent across i. Either $\sup_i M_i \le M_{\max} < \infty$, or M_i has an exponential tail and the likelihood is analyzed on events $M_i \le C \log N$, whose probability tends to one.

Assumption 8 (Compactness and continuity)**.** The parameter space for $\Theta = (\mu, \Pi, \{C_u\}_{u=1}^K)$, or for the sparse linear parameterization $(\mu, \Pi, \{w_u\}_{u=1}^K, \psi)$, is compact. The map $\Theta \mapsto \ell_i(\Theta)$ is almost surely continuous and dominated by an integrable envelope.

Theorem 6 (Consistency of the regime-layer MLE)**.** *Assume the Stage-II model is well specified, with true parameter $\Theta^\star$, and Assumptions 1, 7, and 8 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 label permutation and action-independent deviation offsets.

Proof. By Assumptions 7 and 8, 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 7 (Quasi-MLE under misspecification)**.** *Under Assumptions 7 and 8, 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 KL projection of the true conditional law onto the model class.*

Proof. The same uniform convergence argument applies. The population objective can be written as a constant minus the KL divergence from the true conditional law to the model law. Hence its maximizers are exactly the KL projections. Argmax consistency gives convergence to this set. □

B.7 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 8 (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 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 $\theta \in \mathcal{M}_1$,

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

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 7 and 8, the empirical likelihood gap converges to the positive population gap. This justifies the empirical comparison between $K_u = 1$ and $K_u > 1$.

B.8 Optional localization result for regime segmentation

The next result is an auxiliary statement about the Viterbi path under known parameters. It is not needed for estimating the model, but it explains why persistent regimes can yield stable segmentations when emissions are well separated.

Let $u^\star_{1:M}$ be the true regime path for one trajectory, let $\widehat{u}_{1:M}$ be the Viterbi path under the true parameters, and let S be the number of true switches.

Assumption 9 (Emission separation)**.** There exists $\delta > 0$ such that for any $u \neq v$,

$$\mathbb{E}\left[\log \frac{e_u(a_m \mid x_m, \xi_m)}{e_v(a_m \mid x_m, \xi_m)} \,\middle|\, u^\star_m = u \right] \geq \delta.$$

The log-likelihood ratio increments are bounded or sub-exponential uniformly over contexts.

Assumption 10 (Regime persistence)**.** The transition matrix has large diagonal entries, $\min_u \Pi_{uu} \geq 1 - \epsilon$, and all transition probabilities are bounded away from zero on the support.

Proposition 9 (Segmentation errors localize near switches)**.** *Under Assumptions 9 and 10, there exist constants $c, C > 0$ such that, for any $\eta \in (0,1)$, with probability at least $1 - \eta$, Viterbi errors are confined to neighborhoods of the true switch points with total length at most $CS\log(M/\eta)$. Consequently,*

$$\frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\{\widehat{u}_m \neq u^\star_m\} \leq C \frac{S \log(M/\eta)}{M}.$$

Thus, if $S = o(M/\log M)$, the misclassification fraction vanishes.

Proof sketch. Consider an interior block of length L lying inside a true segment with regime u. For any competing regime $v \neq u$, the cumulative log-likelihood ratio favoring u over v is a sum of increments with mean at least δ. By concentration for bounded or sub-exponential increments, the probability that v dominates u on this block is at most $\exp(-cL)$. A union bound over blocks and competing regimes shows that, with probability at least $1-\eta$, no interior block longer than $C\log(M/\eta)$ is misassigned. Therefore errors can occur only within logarithmic neighborhoods of true switch points. The transition penalty from Assumption 10 prevents rapid alternating paths unless supported by strong emission evidence, yielding the stated localization. □

B.9 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 $a_m^* \in \arg\max_a Q_{\mathrm{anc}}(x_m, a)$, 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.

C Stage I: Latent Dynamics Specification

This appendix details the continuous-time dynamics layer summarized in Section 2. 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 $z(t) \in \{1, \ldots, K_z\}$ 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

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

where $\phi(a)$ is an action-feature embedding; diagonal entries satisfy $[A(a)]_{jj} = -\sum_{k \neq j} [A(a)]_{jk}$. 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 $\lambda_\ell(t) \in (0,1)$ ($\ell = 1, \ldots, L$) evolves between observations according to a state-dependent first-order ODE

$$\frac{d}{dt}\lambda_\ell(t) = \alpha_\ell^{(z(t))}\big(1 - \lambda_\ell(t)\big) - \kappa_\ell^{(z(t))}\lambda_\ell(t), \tag{7}$$

with state-conditioned rates $\alpha_\ell^{(k)}, \kappa_\ell^{(k)} \geq 0$ for each $k \in \{1, \ldots, K_z\}$. Equation (7) 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. Conditional on $z(t_m) = k$ and $\lambda(t_m)$, the observation o_m is drawn from an application-specific likelihood $p_\theta(o_m \mid z(t_m) = k, \lambda(t_m))$. Concrete instantiations for each dataset are described in Appendix D.

Belief filtering. Beliefs over z are propagated through the CTMC and updated by Bayes' rule:

$$\tilde{b}_m = b_{m-1}\exp\{A(a_{m-1})\Delta_{m-1}\}, \quad b_m(k) \propto \tilde{b}_m(k)\, p_\theta(o_m \mid z(t_m) = k, \lambda(t_m)).$$

The decision-time state used by the choice model is $x_m = (b_m, \lambda_m)$.

D Experimental Contexts and Data Sources

To evaluate our model, we use three sequential decision settings with different timing scales and operational contexts: high-frequency ICU sepsis treatment, long-horizon CKD-MBD chronic care, and retail pricing. These domains provide heterogeneous trajectories for testing whether a shared anchor plus switching constraint-regime layer improves prediction and yields interpretable anchor-relative deviations.

Table 2: Treatments and Output Indicators in MIMIC-IV Experiment

Category	Items
Vasopressor	Epinephrine Phenylephrine Norepinephrine Dopamine
Inotropic	Dobutamine Milrinone
Fluid	Crystalloid Colloid Water
Lab Measurement	Low system blood pressure (sysbp_low) Low saturation levels (spo2_sao2_low) Low white blood cell count (wbc_count_low) High arterial lactate (arterial_lactate_high) Normal central venous pressure (cvp_normal)* Normal systemic vascular resistance (svr_normal)* Normal serum creatinine concentration (creatinine_normal)* Normal arterial lactate (arterial_lactate_normal)*
Output	Low-urine

D.1 Clinical Decision-Making: MIMIC-IV and CKD-MBD

These two clinical datasets offer complementary perspectives on medical decision-making: while CKD-MBD captures the intricate, slow-evolving 'balancing act' of chronic outpatient care, MIMIC-IV provides a high-frequency view of acute physiological stabilization in the ICU. Together, they test the framework's robustness against both persistent cognitive load and rapid environmental shifts.

Acute Care (MIMIC-IV):

Sepsis is a life-threatening condition resulting from a dysregulated host response to infection, leading to widespread inflammation, organ failure, and high mortality rates [45]. To study clinician behavior in this high-stakes setting, we utilize the MIMIC-IV database, a comprehensive resource containing de-identified, high-resolution health data from ICU patients. Due to the significant uncertainty and dynamic nature of sepsis management [46], clinician decisions often deviate from stationary protocols. The rich, irregularly sampled clinical trajectories in MIMIC-IV provide an ideal testbed for our framework, allowing us to learn a continuous-time latent patient dynamics model and discover the underlying constraint regimes, such as resource pressures or changing risk postures, that are associated with persistent deviations in clinical decisions over time.

- **Patient Cohort:** We extracted 2,475 sequences of patients diagnosed with sepsis [47], using a 72-hour observation window with complete data records.
- **Treatments (Actions):** The primary interventions in our study are vasopressor therapy and fluid resuscitation, which are essential for stabilizing blood pressure and ensuring organ perfusion [4]. In our model, these represent the clinician's actions a_t. Vasopressors serve to increase cardiac output via vasoconstriction, while fluid administration restores intravascular volume. We categorize these treatments into 3 types of fluids, 4 types of vasopressors, and 2 types of inotropes, as detailed in Table 2.
- **Outcomes and Observations:** To ground the clinical objective and learn the patient dynamics model, we utilize real-time urine output and survival status as primary indicators. Low urine output is a critical signal of kidney dysfunction and impending septic shock. Additionally, we include eight informative lab measurements relevant to sepsis: four representing poor physiological states and four representing favorable outcomes. These variables allow our latent dynamics model to evaluate the counterfactual consequences of clinician actions.

Chronic Management (CKD-MBD):

Chronic Kidney Disease-Mineral Bone Disorder (CKD-MBD) is a systemic syndrome characterized by a complex interplay of biochemical abnormalities, including serum calcium (Ca), phosphorus (P), and parathyroid hormone (PTH) levels [26, 48]. Because these variables are deeply interconnected and influence long-term risks of cardiovascular events and bone fractures, clinical management requires frequent adjustments of multiple medications [49]. We apply our framework to this setting to understand the latent factors, such as shifting risk postures or institutional guidelines, that lead clinicians to change their treatment priorities over the course of long-term care.

- **Data Source and Patients:** We utilize a private dataset collected from a hospital, consisting of longitudinal clinical records for 150 patients captured between January 2020 and January 2024. This cohort provides a unique, high-resolution view of real-world clinical practice. Each patient trajectory includes a series of laboratory measurements and the corresponding pharmacological interventions prescribed by attending physicians.
- **Treatments (Actions):** The treatment space encompasses 13 discrete action markers representing dosage levels across five essential drug classes: calcium-based binders, phosphate binders, vitamin D analogs, and calcimimetics. In our model, these prescriptions represent the clinician's actions a_t. These agents are used concurrently to manage the mineral balance, but the specific weighting of one drug over another often reflects the clinician's underlying decision regime.
- **Outcomes and Observations:** We monitor three primary lab-based indicators: corrected calcium (cCa), phosphate (P), and parathyroid hormone (PTH). Following clinical standards, these are discretized into low, normal, and high categories, forming 10 outcome markers. These biomarkers serve as the state observations x_t used to train our continuous-time patient dynamics model. By evaluating counterfactual outcomes for these three variables, our framework can infer when a clinician has shifted from prioritizing one metabolic target (e.g., phosphorus control) to another (e.g., PTH suppression) through persistent anchor-relative deviations captured by the regime layer.
- **Rationale:** The management of CKD-MBD is often described as a "balancing act" where guidelines (such as the KDIGO updates) provide targets, but individual clinician behavior fluctuates based on perceived patient risk or cognitive load [50]. This dataset is ideal for our regime-switching behavioral model because it captures these subtle shifts in treatment strategy that are not explicitly recorded in the electronic health record but are reflected in the diverging decision trajectories.

D.2 Economic Decision-Making: Dominick's Finer Foods (DFF)

To evaluate our model beyond healthcare, we apply it to retail dynamic pricing using the Dominick's Finer Foods (DFF) dataset from the Chicago Booth Kilts Center for Marketing [23, 24]. In this setting, retail managers balance anchor revenue considerations against operational frictions such as menu costs, promotional inertia, and psychological pricing barriers. The dataset tests whether our model can recover persistent pricing regimes as anchor-relative deviations rather than treating all variation as demand noise or stationary reward misspecification.

- **Data Source and Market Units:** We focus on the Orange Juice category, a standard benchmark in the pricing literature. From the broader DFF database, we select a representative cohort of stores with the most complete and consistent longitudinal records for our experimental analysis. Our final processed dataset consists of trajectories from key retail units monitored over a 52-week period, providing a high-resolution view of real-world pricing dynamics.
- **Pricing (Actions):** The action space a_t encompasses 8 discrete markers representing pricing interventions and promotional activities. These include three price tiers for the focal brand (Citrus Hill), indicators for in-store special displays and significant discount bursts, and dynamic pricing markers for price inertia, price hikes, and price drops. Competitor pricing and promotion variables for Minute Maid enter the market-state/context indicators rather than the action space.
- **Outcomes and Indicators:** Consistent with our structural modeling approach, we define 9 market state indicators that serve as observations x_t. The core indicator is Customer Brand Loyalty (LoyalCH), a persistent latent state reflecting the brand's resilience to competitive pricing. Following clinical analogies, we discretize loyalty into four levels: "crisis," "at-risk," "stable," and "robust". We also track relative market position through price-difference markers and store-specific fixed effects, while incorporating a real-time purchase signal as a

feedback indicator. By evaluating anchor-relative pricing deviations, our model estimates constraint-labeled wedges associated with observed pricing inertia and market-state changes.

- **Rationale:** Retail pricing is often modeled as an MDP in which the agent seeks long-term revenue. In practice, observed prices may deviate from an anchor policy because of menu costs, promotional calendars, and other operational frictions. This dataset allows us to test whether our model recovers switching pricing regimes that improve prediction and provide interpretable anchor-relative deviations.

E Structural Definitions of Regime-Defining Features

A central component of our model is the domain feature map $g(b, \lambda, a)$, which parameterizes anchor-relative wedges through interpretable action and context features. These features do not by themselves identify literal constraints; rather, they provide constraint-labeled coordinates on which the regime layer places sparse shadow-cost weights.

E.1 Clinical Management Features (MIMIC-IV and CKD-MBD)

The clinical features are grounded in medical safety protocols and the cognitive burden of managing multi-drug regimens.

MIMIC-IV ICU Protocol Features The 16-dimensional feature vector for MIMIC-IV is grounded in clinical protocols for sepsis and hemodynamic management.

- **A) History & Complexity (Dim 0–5):**
  - **0.** `drug_count`: Total number of base drugs in the current action. It reflects the complexity and burden of the medication combination.
  - **1.** `added_count`: Number of drugs added relative to the previous action.
  - **2.** `removed_count`: Number of drugs removed relative to the previous action.
  - **3.** `changed_flag`: A binary indicator (0/1) of whether any medication change occurred in this step.
  - **4.** `n_groups_active`: Count of active coarse treatment groups (*pressor, fluid, or inotrope*), reflecting the breadth of therapeutic modalities.
  - **5.** `new_groups_count`: Number of newly introduced treatment groups (transitioning from unused to active).
- **B)** `forbid_if` **Violations (Dim 6–9):** Defined as $1[\text{trigger}] \cdot 1[\text{used forbidden}]$.
  - **6.** `forbid_violate`**[**$p_0_is1_acts[8]$**]**: Risk penalty for using *Water* (8) when blood pressure is critically low (p_0 is True).
  - **7.** `forbid_violate`**[**$p_4_is1_acts[8]$**]**: Risk penalty for using *Water* (8) when lactate levels are high (p_4 is True).
  - **8.** `forbid_violate`**[**$p_8_is1_acts[8, 7]$**]**: Risk penalty for using *Water* (8) or *Colloid* (7) during low urine output (p_8 is True).
  - **9.** `forbid_violate`**[**$p_7_is0_acts[7]$**]**: Risk penalty for using *Colloid* (7) when creatinine is abnormal (p_7 is False).
- **C)** `require_if` **Omissions (Dim 10–12):** Defined as $1[\text{trigger}] \cdot 1[\text{miss all required}]$.
  - **10.** `require_miss`**[**$p_0_is1_acts[0, 1, 2, 3]$**]**: Omission penalty for failing to provide any vasopressors (0/1/2/3) during hypotension.
  - **11.** `require_miss`**[**$p_4_is1_acts[6, 7, 2, 0, 1]$**]**: Omission penalty for failing to provide either fluid (6/7) or specific vasopressors (0/1/2) during high lactate.
  - **12.** `require_miss`**[**$p_8_is1_acts[6, 2, 0, 1]$**]**: Omission penalty for failing to provide either crystalloid (6) or specific vasopressors (0/1/2) during low urine output.
- **D) Group Intensity Counts (Dim 13–15):**
  - **13.** `group_count`**[Ino]**: Count of active drugs within the *Inotrope* group ([4, 5]).
  - **14.** `group_count`**[fluid]**: Count of active drugs within the *Fluid* group ([6, 7, 8]).
  - **15.** `group_count`**[pressor]**: Count of active drugs within the *Vasopressor* group ([0, 1, 2, 3]).

CKD-MBD Chronic Management Features The 8-dimensional feature set for CKD-MBD focuses on long-term safety and resource trade-offs:

- f_1**: Hypercalcemia** $\times$ **Ca-binder risk**: Safety penalty for calcium-based binders during high-calcium states.

- f_2**: $Ca \times P$ risk proxy**: Captures synergistic risk of vascular calcification.
- f_3**: Polypharmacy burden**: Structural penalty on regimen complexity.
- f_4**: Uncontrolled labs proxy**: Penalizes under-treatment relative to biomarker elevation.
- f_5**: Medication cost**: Reflects economic/resource constraints.
- f_6**: Treatment escalation**: The "inertia" cost of increasing strategy intensity.
- f_7**: Specialty drug flag**: Marks the use of high-tier, specialized therapies.
- f_8**: Overall intensity**: Aggregated continuous "dosage" level.

E.2 Economic Pricing Features (DFF)

For the retail pricing domain, we define an 8-dimensional feature set for constraint-labeled pricing wedges. These features capture operational and market frictions that can induce persistent deviations from the shared pricing anchor.

- f_1**: Menu Costs (Inertia):** A penalty for changing prices between weeks ($Price_t \neq Price_{t-1}$). This reflects the operational cost of updating tags and systems.
- f_2**: Psychological Pricing Barrier:** Captures the "stickiness" of ending prices (e.g., .99). Managers are less likely to cross these thresholds.
- f_3**: Competitive Response Gap:** Measures the delay in responding to a significant price drop by a primary competitor (Minute Maid).
- f_4**: Promotion Fatigue:** A penalty for maintaining "Special" status for too many consecutive weeks, reflecting diminishing marginal returns.
- f_5**: Cannibalization Risk:** Reflects the latent cost of pricing the focal brand (CH) too close to the generic house brand, which might erode overall store margins.
- f_6**: Brand Equity Guardrail:** A structural constraint that prevents the retail price from falling below a "floor" to preserve premium brand image.
- f_7**: Inventory-driven Pressure:** Proxy for the urgency to lower prices when historical volume decreases (M9/M10 indicators).
- f_8**: Site-specific pricing rigidity**: An interaction between the Store 7 indicator and price-adjustment actions, capturing site-specific policy frictions in changing prices.

F Detailed Visualization of Phenotypes

To provide a granular understanding of the physiological dynamics captured by the Continuous-Time Markov Chain (CTMC), this section presents a detailed visualization of the six latent patient phenotypes of CKD-MBD. As illustrated in Figure 6, these phenotypes are characterized through their stationary distributions, temporal stability, and population-level evolution.

(a) Phenotype Radar Profiles The radar charts reveal distinct biochemical signatures for each phenotype. Phenotypes z_0 and z_1 exhibit a dominant "Normal" (blue) region, representing stages where mineral metabolism is relatively well-controlled. In contrast, Phenotypes z_4 and z_5 show significant expansions along the "High" (green/red) axes for Phosphorus (P) and PTH, identifying high-risk physiological profiles characterized by severe mineral dysregulation.

(b) γ-weighted Dwell Time Variations in dwell time reflect the inherent stability of each physiological state. Phenotype z_0 maintains an average dwell time of 28.3 months, significantly longer than other states, designating it as a quasi-stable maintenance state within the cohort. Conversely, high-risk states such as Phenotype z_5 have a dwell time of only 3.9 months. This volatility suggests that such states are transient and often trigger more aggressive clinical interventions intended to shift the patient toward a more stable phenotype.

(c) Population Posterior Evolution The population-level trends provide empirical evidence of treatment progression over time. As the observation period increases, the posterior probabilities for stable states (z_0 and z_1) show a clear upward trajectory, while the density of high-risk states gradually declines. This evolution provides qualitative evidence that the CTMC layer captures population-level movement from decompensated phases toward more compensated physiological states.

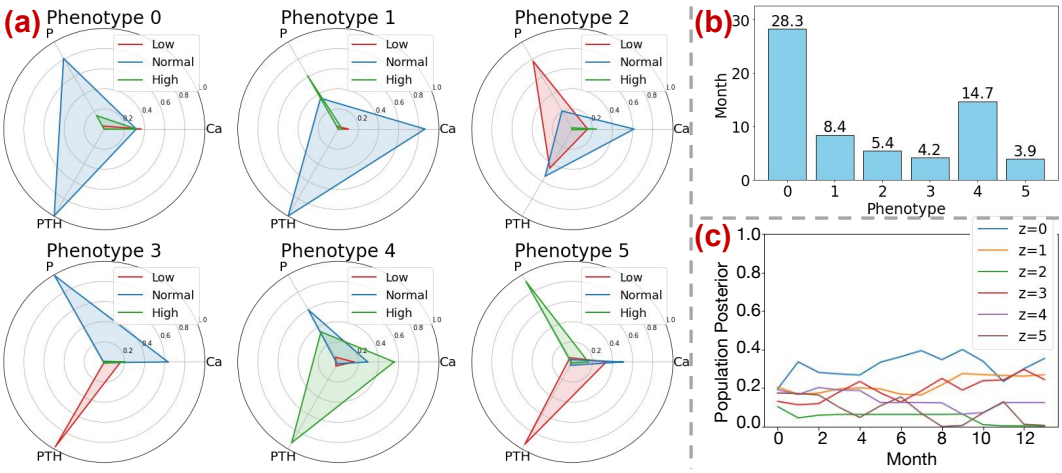


Figure 6: Comprehensive analysis of latent patient phenotypes in the CKD-MBD cohort. **(a) Top Radar Charts**: Emission probabilities of Calcium (Ca), Phosphorus (P), and PTH across three ranges (Low, Normal, High) for each phenotype. **(b) Middle Bar Chart**: γ-weighted average dwell time in months for each phenotype, indicating physiological stability. **(c) Bottom Line Chart**: Temporal evolution of the population posterior probabilities for each latent state z_i across the treatment horizon.

G Detailed Sensitivity Analysis for Reward Specifications

G.1 Reward Specification

A common critique in inverse decision modeling is that the "true reward" is not observable and may vary across stakeholders. We therefore treat r_{anc} as a *anchor proxy* used to define a reference action-value function Q_{anc}. Taking our clinical applications as the primary evaluative anchor, we explicitly evaluate the robustness of our model to plausible baseline families within high-acuity and chronic care settings.

Baseline families. We consider baseline specifications that depend only on available outcomes and clinical stability signals. Examples include:

1. **Outcome-based terminal baseline.** We implement an episodic terminal signal (with discounting over decision epochs), e.g.,
$$r_{\text{anc}}^{\text{term}}(z,a) = 0,$$
$$R_T = \mathbb{I}(\text{favorable outcome}) - \alpha\,\mathbb{I}(\text{adverse outcome}),$$
where "favorable/adverse" outcomes are instantiated by the application (e.g., survival vs. mortality, remission vs. progression).
2. **Stability/shaping baselines.** Using physiologic stability proxies derived from observations (or emissions), e.g.,
$$r_{\text{anc}}^{\text{stab}}(z,a) = -\big(\omega_1\,\text{shock}(z) + \omega_2\,\text{resp_fail}(z) + \omega_3\,\text{renal}(z)\big),$$
where $\text{shock}(\cdot)$, $\text{resp_fail}(\cdot)$, etc. are latent-state risk scores learned from the emission model or defined via clinically motivated mappings.
3. **Risk-sensitive variants.** To reflect aversion to catastrophic deterioration, we consider conservative baselines such as a tail-risk action-value $Q_{\text{anc}}^{\text{CVaR}}(\cdot)$ or moment-penalized variants of the form
$$r_{\text{anc}}(z,a) - \kappa_{\text{var}}\,\text{Var}[\Delta\text{Score} \mid z,a],$$
where ΔScore denotes a clinically relevant change (e.g., in an organ-failure proxy) and the variance is estimated under model-based rollouts.

Importantly, all choices yield the same modeling structure: the reference Q_{anc} changes, while the inferred switching constraint regimes capture residual anchor-relative deviations.

Practical computation of Q_{anc} under continuous-time dynamics. Given action-held intervals $[t_m, t_{m+1})$, the CTMC defines a transition kernel
$$P^{(a_m)}(\Delta t_m) = \exp\big(A^{(a_m)}\Delta t_m\big),$$

Table 3: Predictive performance across reward specifications and regime counts (K_u).

Reward Type	K_u	Log-Likelihood ↑	Top-1 Acc. ↑	Top-5 Acc. ↑	Top-10 Acc. ↑
Dataset: CKD-MBD					
Outcome-based	1	-5.57 ± 0.39	0.17 ± 0.01	0.46 ± 0.04	0.56 ± 0.06
	3	$\mathbf{-4.52 \pm 0.85}$	$\mathbf{0.22 \pm 0.05}$	$\mathbf{0.51 \pm 0.05}$	$\mathbf{0.63 \pm 0.06}$
	5	-4.96 ± 0.49	0.20 ± 0.03	0.46 ± 0.06	0.58 ± 0.05
Stability-based	1	-5.89 ± 0.24	0.22 ± 0.03	0.41 ± 0.03	0.51 ± 0.04
	3	-4.77 ± 0.23	$\mathbf{0.27 \pm 0.03}$	0.45 ± 0.02	0.53 ± 0.04
	5	$\mathbf{-4.52 \pm 0.27}$	0.26 ± 0.03	$\mathbf{0.47 \pm 0.05}$	$\mathbf{0.54 \pm 0.05}$
Risk-sensitive	1	-5.57 ± 0.39	0.17 ± 0.02	0.44 ± 0.04	0.59 ± 0.04
	3	-5.40 ± 0.26	0.17 ± 0.01	0.51 ± 0.05	0.63 ± 0.06
	5	$\mathbf{-5.26 \pm 0.33}$	$\mathbf{0.19 \pm 0.03}$	$\mathbf{0.51 \pm 0.03}$	$\mathbf{0.64 \pm 0.02}$
Dataset: MIMIC-IV					
Outcome-based	1	-0.97 ± 0.00	0.76 ± 0.01	0.95 ± 0.00	0.98 ± 0.00
	3	-0.77 ± 0.08	0.81 ± 0.04	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.98 \pm 0.00}$
	5	$\mathbf{-0.68 \pm 0.06}$	$\mathbf{0.83 \pm 0.05}$	0.96 ± 0.01	0.98 ± 0.00
Stability-based	1	-0.91 ± 0.00	0.75 ± 0.00	0.96 ± 0.00	0.98 ± 0.00
	3	-0.85 ± 0.03	0.75 ± 0.00	$\mathbf{0.96 \pm 0.00}$	0.99 ± 0.00
	5	$\mathbf{-0.78 \pm 0.06}$	$\mathbf{0.76 \pm 0.01}$	0.96 ± 0.00	$\mathbf{0.99 \pm 0.00}$
Risk-sensitive	1	-0.91 ± 0.00	0.76 ± 0.01	0.96 ± 0.00	0.98 ± 0.00
	3	-0.78 ± 0.02	$\mathbf{0.77 \pm 0.00}$	$\mathbf{0.97 \pm 0.00}$	0.98 ± 0.00
	5	$\mathbf{-0.74 \pm 0.02}$	0.77 ± 0.01	0.97 ± 0.00	$\mathbf{0.98 \pm 0.00}$

which enables value iteration on the latent state space (or Monte Carlo rollouts) to compute $Q_{\text{anc}}(b, \lambda, a)$ for each belief b. When the emission model is complex, we compute Q_{anc} on a discretized belief representation (e.g., via clustering beliefs) or via rollouts initialized from the posterior belief state b.

G.2 Performance under Different Reward Specifications

This section provides a comprehensive performance breakdown of the proposed framework under different clinical reward specifications. To ensure that the inferred switching regimes are not artifacts of a specific objective, we evaluate the framework using three distinct reward families: (i) Outcome-based, (ii) Stability-based, and (iii) Risk-sensitive.

Table 3 summarizes the predictive metrics for both datasets. Across all reward types, models with switching regimes ($K_u > 1$) consistently outperform the stationary baseline ($K_u = 1$), substantiating the robustness of our framework.

H Detailed Comparative Benchmarking of System Dynamics Models

This section presents a comparative benchmarking of the CTMC+ODE dynamics model against ODE-RNN and NCDSSM baselines through a hold-out prediction task. In Table 4, we evaluate model fidelity across diverse temporal scales to ensure robustness across domains: specifically 3-month and 6-month horizons for the chronic CKD-MBD cohort, 6-hour and 12-hour horizons for the acute MIMIC-IV sequences, and a 3-week out-of-sample window for the DFF retail pricing dataset. This multi-horizon assessment validates the model's capacity to capture both rapid state fluctuations and long-term systemic trends.

I SCM-based Counterfactual Sampling and Anchor-Optimal Action Synthesis

To ground sequential decision-making in realistic counterfactuals, we formalize the process as a Structural Causal Model (SCM). This allows us to evaluate anchor-optimal counterfactual actions while preserving the system's underlying exogenous characteristics, so anchor-recovery estimates are evaluated under unit-level counterfactual consistency rather than fresh stochastic draws.

Table 4: Comparison of System Dynamics Models (Hold-out Prediction). Results represent mean ± SD across all biomarkers; ECE_{avg} denotes mean Expected Calibration Error.

Model	NLL ↓	Brier ↓	Acc ↑	ECE_{avg} ↓
Dataset: CKD-MBD (3-Month Horizon)				
CTMC+ODE	$\mathbf{0.668 \pm 0.012}$	0.261 ± 0.001	$\mathbf{0.726 \pm 0.006}$	$\mathbf{0.098 \pm 0.006}$
ODE-RNN	0.672 ± 0.001	$\mathbf{0.239 \pm 0.001}$	0.672 ± 0.006	0.199 ± 0.009
NCDSSM	0.690 ± 0.009	0.248 ± 0.004	0.528 ± 0.039	0.160 ± 0.021
Dataset: CKD-MBD (6-Month Horizon)				
CTMC+ODE	$\mathbf{0.667 \pm 0.009}$	0.282 ± 0.002	$\mathbf{0.717 \pm 0.011}$	$\mathbf{0.091 \pm 0.008}$
ODE-RNN	0.673 ± 0.001	$\mathbf{0.240 \pm 0.000}$	0.664 ± 0.008	0.185 ± 0.009
NCDSSM	0.700 ± 0.008	0.253 ± 0.004	0.523 ± 0.018	0.124 ± 0.021
Dataset: MIMIC-IV (6-Hour Horizon)				
CTMC+ODE	$\mathbf{0.195 \pm 0.004}$	$\mathbf{0.112 \pm 0.003}$	$\mathbf{0.936 \pm 0.002}$	$\mathbf{0.031 \pm 0.003}$
ODE-RNN	0.645 ± 0.000	0.226 ± 0.000	0.876 ± 0.002	0.448 ± 0.000
NCDSSM	0.861 ± 0.019	0.145 ± 0.003	0.847 ± 0.004	0.283 ± 0.002
Dataset: MIMIC-IV (12-Hour Horizon)				
CTMC+ODE	$\mathbf{0.204 \pm 0.004}$	$\mathbf{0.121 \pm 0.003}$	$\mathbf{0.925 \pm 0.003}$	$\mathbf{0.034 \pm 0.004}$
ODE-RNN	0.645 ± 0.000	0.226 ± 0.000	0.877 ± 0.001	0.449 ± 0.000
NCDSSM	1.447 ± 0.038	0.154 ± 0.002	0.847 ± 0.003	0.220 ± 0.002
Dataset: DFF (3-Week Horizon)				
CTMC+ODE	$\mathbf{0.536 \pm 0.000}$	0.336 ± 0.000	$\mathbf{0.791 \pm 0.001}$	$\mathbf{0.124 \pm 0.001}$
ODE-RNN	0.701 ± 0.008	0.249 ± 0.003	0.690 ± 0.021	0.439 ± 0.003
NCDSSM	0.551 ± 0.011	$\mathbf{0.189 \pm 0.004}$	0.733 ± 0.000	0.325 ± 0.012

I.1 Anchor-Optimal Action and Counterfactual Reward

At each decision epoch t_m^i, we synthesize an anchor-optimal counterfactual action by evaluating candidate actions under the fitted structural dynamics. For each candidate $a \in \mathcal{A}$, we intervene on the current action, propagate the latent trajectory forward under the SCM-consistent rollout, and compute the corresponding anchor return. The anchor-optimal action is then

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

where $Q_{\text{anc}}(x_m, a)$ is evaluated by Monte Carlo rollouts initialized from the fitted belief state and the abducted unit-level noise. The instantaneous reward r_τ at time τ is defined as a function of the deterministic target-alignment probabilities $\lambda_\ell(\tau)$:

$$r_\tau \;=\; \sum_{\ell \in L} w_\ell \, \lambda_\ell(\tau) \;-\; w_{\text{risk}} \prod_{j \in J} p_j^{\text{hi}}(\tau, z(\tau)), \tag{8}$$

where $w_\ell > 0$ weights the desirability of remaining target-aligned across dimensions $\ell \in L$, and $w_{\text{risk}} > 0$ penalizes synergistic risks. Consistent with our emission model, the probability of a high-risk violation for dimension j is defined as $p_j^{\text{hi}}(\tau, z) = \eta_{z,j,\text{hi}}(1 - \lambda_j(\tau))$, where $\eta_{z,j,\text{hi}}$ denotes the emission split mass assigned to the high-risk off-target category. The cumulative discounted reward is:

$$R_m(a) = \sum_{\tau=m}^{T} \delta^{\tau - m} \, r_\tau, \quad \delta \in (0, 1], \tag{9}$$

For regularly spaced epochs, the discrete discount factor satisfies $\delta = e^{-\rho\Delta}$; with irregular timing we use $\delta^{\tau-m}$ as the epoch-indexed approximation to the continuous-time discount in the anchor-value definition in Section 2. Here, λ_τ acts as a continuous-time surrogate for target alignment along each rollout.

I.2 Abduction-Action-Prediction Framework

Following the standard causal inference workflow, we generate counterfactuals by fixing the exogenous noise inferred from the observed history $\mathcal{H}^i = (\{o_m^i, a_m^i\}_{m=1}^{M_i})$.

Step 1: Abduction. Using current model parameters, we infer the posterior over latent states and exogenous noises $p(z_{1:T}, \varepsilon^{\text{tr}}, \varepsilon^{\text{em}} \mid \mathcal{H}^i)$ via a backward-smoothing particle-EM procedure. Here, ε^{tr} denotes the specific realizations of Gumbel noise that drove CTMC jumps, and ε^{em} denotes the Gumbel noise associated with categorical emissions o^i_m.

Step 2: Action Intervention. For each candidate action $a \in \mathcal{A}$ at decision epoch t^i_m, we replace the factual action with the intervention $\tilde{a}^i_{t_m} = a$. Subsequent actions can either follow the factual trajectory for single-step action evaluation or be recursively replaced by anchor-optimal actions for sequence synthesis. To ensure unit-level consistency, we maintain the latent Poisson arrival times s_k and the specific Gumbel realizations ε^{tr} as fixed exogenous traits of the individual unit.

Step 3: Prediction (Forward Rollout). The system is rolled forward under the intervention $\text{do}(a_t := \tilde{a}_t)$ while keeping exogenous noises fixed:

1. **Latent State** $z(t)$: Transitions at each uniformization jump time s_k are computed using the identical Gumbel realizations $g_{k,j} \in \varepsilon^{\text{tr}}$ from the abduction step:
$$z_{k+1} = \arg\max_j \{\log \tilde{P}_{z_k j}(\tilde{a}_{s_k}) + g_{k,j}\}, \quad \tilde{P}(\tilde{a}_{s_k}) = I + \frac{1}{\Omega} A(\tilde{a}_{s_k}). \tag{10}$$
where $\Omega \geq \max_{j,a} |A_{jj}(a)|$ is the uniformization rate.
2. **Outcome Dynamics** λ_τ: Between epochs, λ_ℓ evolves deterministically via the state-dependent ODE used in the fitted dynamics layer. Over an interval of length Δ_τ with latent state z_τ, the closed-form update is
$$\lambda_\ell(\tau+\Delta_\tau) = \bar{\lambda}^{(z_\tau)}_\ell + \big(\lambda_\ell(\tau) - \bar{\lambda}^{(z_\tau)}_\ell\big) \exp\Big[-\big(\alpha^{(z_\tau)}_\ell + \kappa^{(z_\tau)}_\ell\big)\Delta_\tau\Big], \qquad \bar{\lambda}^{(z)}_\ell = \frac{\alpha^{(z)}_\ell}{\alpha^{(z)}_\ell + \kappa^{(z)}_\ell}. \tag{11}$$
3. **Emissions at Observations**: Counterfactual signals $\tilde{o}^i_{m,\ell}$ are reconstructed using the fixed emission Gumbels ε^{em} and the newly generated latent trajectory.

anchor-action synthesis. For each decision epoch, we average over S posterior samples to estimate the anchor value of each candidate action:
$$\widehat{Q}_{\text{anc}}(x_m, a) = \frac{1}{S} \sum_{s=1}^{S} \sum_{\tau=m}^{T} \delta^{\tau-m} r^{(s,a)}_\tau.$$
We then select $a^*_m \in \arg\max_{a \in \mathcal{A}} \widehat{Q}_{\text{anc}}(x_m, a)$. This procedure produces anchor-optimal counterfactual actions from the unit's inferred condition and supports anchor-recovery diagnostics under constraint relaxation.

J Case-Level Auditing Example

We formalize the policy evaluation process by utilizing a Large Language Model (LLM) as an automated reporting layer that translates structured model outputs into natural-language audit reports. The auditor receives five structured inputs: system observations o_m, observed actions a_m, anchor-optimal counterfactual actions a^*_m, inferred regimes u_m, and latent states z_m. It then generates a three-tiered report designed for systemic auditing:

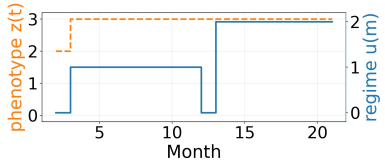
1. **Rationality Analysis**: Assesses the plausibility of the synthesized anchor-optimal action a^*, checking whether the anchor action is operationally reachable under the fitted trajectory context.
2. **Wedge Interpretation**: Explains the anchor-relative deviation between a_m and a^*_m by mapping the active regime u_m to specific institutional, clinical, or market frictions.
3. **Fidelity Summary**: Relates latent transitions in regimes and states to recorded environmental shocks or clinical events as a trajectory-level consistency check.

This auditing layer extends our model from a predictive model into a structured diagnostic tool, enabling stakeholders to retrospectively identify when, where, and how constraint-labeled deviations from the shared anchor arise. Complete prompt templates are provided in Appendix K. To illustrate the practical auditing capability of our model, Figure 7 presents a representative case study from the CKD-MBD cohort, integrating anchor-optimal counterfactual synthesis, latent regime inference, and an LLM-based reporting layer.

(a) Observed vs. Counterfactual Trajectory

Month	Ca	P	PTH	action _doc	action _cf
0	NA	NA	NA	M1	M1
1	NA	NA	NA	M1	M1
2	High	High	NA	M4	M7
3	High	Normal	Low	M4	M7
4	Low	High	Low	M1+M4	M1
5	Low	High	Low	M1+M4	M1
6	High	Normal	Low	M1+M4	M0
7	High	Normal	Low	M1+M4	M0
8	Normal	Normal	Low	M1+M4	M0
9	Normal	Normal	Low	M1+M4	M0
10	Normal	Normal	Low	M1+M4	M0
11	Normal	High	Low	M1+M4	M4
12	Normal	High	Low	M1+M5	M4
13	Low	High	Low	M1+M5	M4
14	Normal	High	Low	M1+M6	M4
15	Normal	High	Low	M1+M6	M4
16	Normal	Normal	Low	M1+M4	M0
17	Normal	Normal	Low	M1+M4	M0
18	High	Normal	Low	M1+M4	M0
19	High	Normal	Low	M1+M5	M0
20	High	Normal	Low	M1+M5	M1
21	High	Normal	Low	M1+M5	M1

(b) Temporal dynamics (K_z=4, K_u=3)



(c) LLM-based Report

1. clinical_rationality:
a* ensures stability by:
(1) preempting iatrogenic PTH suppression via non-calcium M7 substitution (Month 2-3);
(2) resolving 'Polypharmacy Traps' (Month 14-21) by simplifying redundant clinician stacks into parsimonious, targeted titration;
(3) preemptively protecting PTH levels to prevent low-turnover bone disease.

2. gap_interpretation:
(1) Regime u=1 (Month 3-12) indexes 'Clinical Inertia' and 'Cost Sensitivity' (f5, f6), explaining the reliance on cheap calcium despite failure;
(2) Regime u=2 (Month 13+) captures 'Target Fixation' (f4), quantifying irrational drug stacking as an over-reaction to phosphate fluctuations.

3. fidelity_validation:
Empirical grounding is demonstrated by:
(1) the z-transition at Month 03 correctly indexing chronic PTH suppression;
(2) the u=2 jump at Month 13 being perfectly synchronized with the observed onset of medication hyper-complexity;
(3) the temporal alignment of u shifts with behavioral 'GAPs' validates the causal interpretability of the latent trajectories.

Figure 7: Automated decision auditing and structural counterfactual analysis for a CKD-MBD trajectory. The figure integrates three structural perspectives: **(a) Observed vs. Counterfactual Actions.** Comparison between observed clinician actions (a_m) and model-synthesized anchor-optimal actions (a_m^*), alongside longitudinal biomarkers (Ca, P, PTH). Treatment markers denote specific medications and dosages: M0–M1 represent calcium carbonate, M4–M6 represent sevelamer, and M7 represents lanthanum carbonate. Combined markers indicate concurrent therapies. **(b) Latent Structural Dynamics.** Temporal evolution of the inferred latent physiological state z_m and constraint regime u_m. **(c) LLM-based Institutional Audit Report.**

K Detailed Counterfactual Auditing Prompts for LLMs

We utilize GPT-4o as the automated reporting layer, configured with a temperature of 0.1 and a fixed seed of 42 to improve consistency and reproducibility across longitudinal trajectories. The LLM receives structured model outputs only; anchor actions, regimes, and counterfactual quantities are computed by our model.

K.1 Detailed Counterfactual Auditing Prompt for MIMIC-IV

Complete Prompt Template for Counterfactual Auditing and Clinical Validation (MIMIC-IV)

Task Description

- **Role**: You are an expert intensivist specializing in Sepsis and Septic Shock.
- **Objective**: Audit a patient's treatment trajectory by comparing the **Observed Clinical Action** (a) against the **anchor-optimal Counterfactual Strategy** (a^*) derived from the framework, and assess whether the anchor-relative deviations are clinically plausible under the inferred regime.
- **Context**: Hourly longitudinal records including hemodynamic vital signs, laboratory measurements, and complex intervention markers (Markers 0–102).

Clinical Knowledge Base: Marker Definitions (Action and Observation Space)

- **(A) Treatment Markers (Actions 0–8)**:
  - **Vasopressors/Inotropes**: Marker 0 (Epinephrine), 1 (Phenylephrine), 2 (Norepinephrine), 3 (Dopamine), 4 (Dobutamine), 5 (Milrinone).
  - **Fluid Resuscitation**: Marker 6 (Crystalloid), 7 (Colloid), 8 (Water).
- **(B) Body Predicate Markers (Observations 9–101)**:
  - This section covers 31 laboratory or vital sign items. Each item has **3 markers** representing `low`, `normal`, and `high` ranges.

- Totaling 93 markers (Markers 9–101). For example: Markers 9–11 (Heart Rate), 15–17 (Mean Arterial Pressure), and 99–101 (Lactate).

- **(C) Head Predicate Marker (Clinical Event 102)**:
  - Marker 102: Low Urine Output (a critical indicator of sepsis-associated acute kidney injury or impaired organ perfusion).

Model Component Definitions

- **Latent Patient State (z)**: A continuous-time Markov Chain (CTMC) physiological phenotype (e.g., septic shock progression, fluid responsiveness) that determines the baseline clinical reward.
- **Latent Constraint Regime (u)**: A hidden decision mode reflecting time-varying clinician pressures (e.g., risk-aversion, institutional policy, or resource constraints).
- **anchor-relative Deviation**: Systematic deviations where the observed action a differs from the anchor-optimal action a^* due to the shadow costs associated with the active regime u.

Input Format

- `patient_biomarkers` = {HOURLY_HEMODYNAMIC_AND_LAB_DATA}
- `observed_actions_a` = {ACTUAL_INTERVENTION_TRAJECTORY}
- `optimal_actions_a_star` = {A_STAR_TRAJECTORY}
- `inferred_regimes_u` = {TIME_SERIES_OF_U}
- `latent_trajectories_z` = {TIME_SERIES_OF_Z}

Your Analytical Task

1. **Rationality Analysis**: Determine whether a^* would plausibly stabilize vital signs (e.g., MAP) and clear lactate relative to a. Explain whether the model's suggested intervention (e.g., vasopressor titration vs. fluid bolus) is clinically sound for sepsis management.
2. **Deviation Interpretation**: Identify plausible unobserved frictions or decision pressures (e.g., fear of pulmonary edema, institutional protocols) using the inferred regime u to explain the deviation between a and a^*.
3. **Trajectory-Regime Correlation**: Correlate transitions in z and u with clinical events. Assess whether shifts in u align with the timing of anchor-relative deviations observed in the actual care.

Output (Strict JSON Report)

```
{
  "rationality": "Plausibility of a*.",
  "deviation": "Deviation explained by u.",
  "fidelity": "Links among z, u, and actions."
}
```

K.2 Detailed Counterfactual Auditing Prompt for CKD-MBD

Complete Prompt Template for Counterfactual Auditing and Clinical Validation (CKD-MBD)

Task Description

- **Role**: You are an expert nephrologist and clinical auditor specializing in Chronic Kidney Disease-Mineral and Bone Disorder (CKD-MBD).
- **Objective**: Audit a patient's treatment trajectory by comparing the **Observed Clinical Action (a)** against the **anchor-optimal Counterfactual Strategy (a^*)**. Assess whether our model captures clinically plausible anchor-relative deviations through the inferred constraint regime.
- **Context**: Monthly longitudinal records including Calcium (Ca), Phosphorus (P), and PTH. Medication is represented by Markers 0–12.

Clinical Knowledge Base: Medication Marker Definitions

- **(A) Calcium Supplements (Calcium Carbonate/Caltrate)**:
  - Marker 0: 1.0–3.0 g/day (Low dose).
  - Marker 1: 4.0–8.0 g/day (Medium dose).
  - Marker 2: 9.0–12.0 g/day (High dose).
  - Marker 3: >12.0 g/day (Ultra-high dose, rare).
- **(B) Phosphorus Binders (Sevelamer)**:
  - Marker 4: 0.25–3.0 g/day (Low dose).
  - Marker 5: 4.0–8.0 g/day (Medium dose).
  - Marker 6: 9.0–12.0 g/day (High dose).
- **(C) Lanthanum Carbonate**: Marker 7: 1.5 g/day (Specialized binder).
- **(D) Active Vitamin D (Calcitriol and analogs)**:

- Marker 8: <1.0 μg/week (Low dose).
- Marker 9: 1.0–2.0 μg/week (Medium dose).
- Marker 10: >2.0 μg/week (High dose).

- **(E) Calcimimetics (Cinacalcet)**: Marker 11 (<0.2 mg/day), Marker 12 (>0.2 mg/day).

Model Definitions

- **Latent Patient State** (z): A continuous-time physiological phenotype (e.g., hyperparathyroid or mineral imbalance state) that determines the baseline clinical reward.
- **Latent Constraint Regime** (u): A hidden decision mode reflecting time-varying clinician pressures (e.g., risk-aversion, institutional policy, or resource constraints).
- **anchor-relative Deviation**: Systematic deviations where the observed action a differs from the anchor-optimal action a^* due to the shadow costs associated with the active regime u.

Input Format

- `patient_biomarkers` = {MONTHLY_CA_P_PTH_DATA}
- `observed_actions_a` = {ACTUAL_MARKER_TRAJECTORY}
- `optimal_actions_a_star` = {A_STAR_TRAJECTORY}
- `inferred_regimes_u` = {TIME_SERIES_OF_U}
- `latent_trajectories_z` = {TIME_SERIES_OF_Z}

Your Analytical Task

1. **Rationality Analysis**: Determine whether a^* would plausibly stabilize biochemical markers relative to a. Explain why the model's suggested dosage is clinically sound.
2. **Deviation Interpretation**: Identify plausible unobserved frictions or decision pressures (e.g., compliance, institutional protocols) using the inferred regime u to explain the deviation between a and a^*.
3. **Trajectory-Regime Correlation**: Correlate transitions in z and u with clinical events. Assess whether shifts in u align with the timing of anchor-relative deviations in the actual care.

Output (Strict JSON Report)

```
{
  "rationality": "Plausibility of a*.",
  "deviation": "Deviation explained by u.",
  "fidelity": "Links among z, u, and actions."
}
```

K.3 Detailed Counterfactual Auditing Prompt for DFF

Complete Prompt Template for Counterfactual Auditing and Economic Validation (DFF)

Task Description

- **Role**: You are an expert retail economist and pricing strategist specializing in the Orange Juice category (Citrus Hill vs. Minute Maid).
- **Objective**: Audit a retail unit's pricing trajectory by comparing the **Observed Pricing Action** (a) against the **anchor-optimal Counterfactual Strategy** (a^*) derived from the framework, and interpret anchor-relative deviations through inferred pricing frictions.
- **Context**: Weekly longitudinal records of pricing, promotional activities, and consumer loyalty (Markers 0–17).

Economic Knowledge Base: Marker Definitions (Action and Market Space)

- **(A) Pricing Interventions (Actions 0–7)**:
  - **Price Tiers**: Marker 0 (Ultra-low price, aggressive promotion), Marker 1 (Low price), Marker 2 (Normal/High price, margin protection).
  - **Promotional Activity**: Marker 3 (In-store special display), Marker 4 (Significant discount depth > 15%).
  - **Dynamic Interventions**: Marker 5 (Price Inertia: maintaining previous week's price), Marker 6 (Price Hike), Marker 7 (Price Drop).
- **(B) Market State Indicators (Observations 8–16)**:
  - **Customer Brand Loyalty (LoyalCH)**: Marker 8 (Crisis: <0.25), Marker 9 (At-risk: 0.25–0.5), Marker 10 (Stable: 0.5–0.8), Marker 11 (Robust: >0.8).
  - **Competitive Pressure (Minute Maid)**: Marker 12 (Competitor low-price challenge), Marker 13 (Competitor normal/high price), Marker 14 (Competitor in-store promotion).
  - **Market Positioning**: Marker 15 (Price advantage: CH is significantly cheaper), Marker 16 (Price disadvantage: CH is significantly more expensive).
- **(C) Feedback Signal (Outcome 17)**:

- Marker 17: Purchase success for focal brand (Citrus Hill).

Model Component Definitions

- **Latent Market State** (z): A hidden environment phenotype (e.g., price elasticity, competitive intensity) determining the baseline revenue potential.
- **Latent Constraint Regime** (u): A hidden decision mode reflecting institutional priorities or frictions (e.g., menu costs, rigid promotional calendars, or cognitive load from pricing updates).
- **anchor-relative Deviation**: Systematic deviations where the observed pricing a differs from the anchor-optimal pricing action a^* due to shadow costs associated with the active regime u.

Input Format

- `market_indicators` = {WEEKLY_LOYALTY_AND_COMPETITIVE_DATA}
- `observed_actions_a` = {ACTUAL_PRICING_TRAJECTORY}
- `optimal_actions_a_star` = {A_STAR_TRAJECTORY}
- `inferred_regimes_u` = {TIME_SERIES_OF_U}
- `latent_trajectories_z` = {TIME_SERIES_OF_Z}

Your Analytical Task

1. **Rationality Analysis**: Determine whether a^* would plausibly improve purchase probability (Marker 17) and maintain brand loyalty (Markers 10–11) relative to a. Evaluate whether the suggested move is economically sound.
2. **Deviation Interpretation**: Identify latent frictions (e.g., menu costs, psychological pricing barriers, or institutional "stickiness") using the inferred regime u to explain why a differs from the anchor-optimal action a^*.
3. **Trajectory-Regime Correlation**: Correlate shifts in z and u with market events. Assess whether changes in u align with the timing of anchor-relative deviations.

Output (Strict JSON Report)

```
{
  "rationality": "Plausibility of a*.",
  "deviation": "Deviation explained by u.",
  "fidelity": "Links among z, u, and actions."
}
```

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; this conditional interpretation is part of the model design and is supported empirically by dynamics-layer validation and reward-sensitivity analyses. Although we evaluate the framework across clinical and retail settings with 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 nonstationary sequential decisions by separating shared anchor behavior from persistent anchor-relative deviations. This can help analysts audit when operational, institutional, or resource-related pressures are associated with deviations from a reference policy. Potential risks include over-interpreting anchor-optimal counterfactuals as direct recommendations or treating LLM-generated audit summaries as definitive conclusions. To mitigate these risks, the LLM is used only as a reporting layer over structured model outputs; all regimes, anchor actions, and counterfactual quantities are computed by the model. Private patient data are not released, and practical deployment should include domain expert review, privacy safeguards, and prospective validation.

M Computing Infrastructure

All experiments are performed on an Ubuntu 20.04.3 LTS server with an Intel(R) Xeon(R) Gold 6248R CPU @ 3.00 GHz, 227 GB of system memory, and NVIDIA GeForce RTX 3090 GPUs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are supported by Sections 2–4 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Appendix L.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions, theorem statements, and proofs are provided in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The model (Section 2), the three-stage estimation pipeline (Section 3), the datasets (Appendix D), the feature maps (Appendix E), the reward families (Appendix G), and the counterfactual rollout procedure (Appendix I) jointly enable reproduction of the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

Justification: The code to reproduce our framework and experiments will be made available upon acceptance. MIMIC-IV and DFF are available through their original credentialed or licensed access mechanisms.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper describes the core estimation and evaluation protocol in Sections 3–4, with dataset construction, feature definitions, reward variants, and counterfactual rollout details provided in Appendices D, E, G, and I.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All quantitative tables report mean $\pm$ one standard deviation across multiple random seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are reported in Appendix M.

   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses only de-identified retrospective records and is positioned as an anchor-based auditing tool rather than a treatment or pricing recommender.

   Guidelines:
   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive and negative societal impacts are discussed in Appendix L.

    Guidelines:
    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper poses no such risks. The paper releases no high-risk pretrained models or scraped datasets; the LLM auditor only consumes model outputs through a standard API.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: MIMIC-IV (PhysioNet credentialed license), DFF (Chicago Booth Kilts Center academic terms), and all baselines are properly cited in Section 4 and Appendix D and used under their respective licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release a new dataset or pretrained model asset.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or newly recruited human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The clinical data used are fully de-identified retrospective records; no prospective human-subjects research was conducted.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: GPT-4o is used only as a structured reporting layer over computed outputs (not as part of the core estimator); details and prompts are in Section 4.3 and Appendix K.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.