Shared-Benchmark Regime Decomposition for Nonstationary Clinical Decisions

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

Abstract

Structured clinical decision sequences, such as tabular EHRs and irregularly sampled physiological signals, are partially observed and often nonstationary: similar measured states can lead to different actions at different times. We introduce *Shared-Benchmark Regime Decomposition* (SBRD), a latent-variable framework that separates a shared benchmark value function from switching constraint-regime deviations. A continuous-time latent dynamics layer constructs belief states and benchmark action values from structured health records, while a sparse regime layer recovers persistent, interpretable wedges in observed decisions. On ICU sepsis treatment (MIMIC-IV) and chronic CKD-MBD management, SBRD improves held-out action prediction over stationary and interpretable behavioral baselines and yields clinically meaningful regime profiles. The recovered wedges support benchmark-relative constraint-relaxation diagnostics that reveal where treatment complexity or protocol-related pressures account for recoverable benchmark value.

1. Introduction

Structured clinical data, including tabular EHRs and irregularly sampled physiological time series, underlie a growing share of high-stakes sequential decisions, from chronic disease management to ICU stabilization (Johnson et al., 2023). The downstream choices taken on top of these signals (drug selection, dose adjustment, escalation/de-escalation) are partially observed, irregularly timed, and notably *nonstationary*: under similar measured conditions, clinicians frequently choose different actions (Jung & Shah, 2015). Standard responses absorb this variation into drifting objectives, unstructured noise, or misspecified rewards (Cheung et al., 2020; Lecarpentier & Rachelson, 2019), which obscures *when* latent pressures (acuity, protocol adherence, regimen complexity) drive treatment choices and limits clinical interpretability.

We develop *Shared-Benchmark Regime Decomposition* (SBRD), a structured latent-variable framework that decomposes nonstationary clinical behavior into a shared, outcome-grounded benchmark and a switching regime layer. A continuous-time latent state $z(t)$ supports a benchmark action-value Q_{base} that captures stable decision logic under partial observability and irregular timing (Rubanova et al., 2019; Habiba & Pearlmutter, 2020; Ansari et al., 2023; Ghahramani & Hinton, 2000). Observed actions deviate from this benchmark through a regime-specific shadow-cost profile $C(\cdot\,; u)$, with the latent regime u evolving as a sticky Markov chain on the decision clock (Hamilton, 1989; Kim & Nelson, 2017). Each regime is a sparse linear combination of interpretable action and context features, so regime contrasts read as switching wedges in the spirit of Chari et al. (2007) and Sims (2003). Holding Q_{base} fixed while the wedge layer switches keeps benchmark values comparable across irregular timing and provides an identifying restriction that grounds counterfactual analysis. We use the term "constraint" regime broadly, as a label for regime-persistent frictions around the shared benchmark, rather than only as a literal hard constraint.

We make three contributions. (i) We introduce a decomposition of nonstationary clinical decisions into a shared benchmark and switching regime-dependent wedges under partial observability and irregular timing, contrasting with stationary IRL/behavior-cloning baselines (Ng et al., 2000; Ziebart et al., 2008; Bloem & Bambos, 2015; Yu et al., 2019) and discrete-time regime-switching models (Hamilton, 1989; Kim & Nelson, 2017) (extended discussion in Appendix A). (ii) We provide regime identification, sparse wedge recovery, and plug-in stability guarantees, with counterfactual evaluation grounded in the SCM framework (Pearl, 2009b;a) (Appendix B). (iii) On MIMIC-IV sepsis (Johnson et al., 2024; 2023) and a private CKD-MBD chronic care cohort, SBRD improves held-out action prediction over stationary and interpretable behavioral baselines (Ziebart et al., 2008; Frosst & Hinton, 2017), with constraint-relaxation diagnostics that quantify which regime-dependent wedges drive the largest recoverable benchmark value.

[1]The Chinese University of Hong Kong, Shenzhen, China [2]City University of Hong Kong, Hong Kong SAR, China. Correspondence to: Shuang Li <lishuang@cuhk.edu.cn>.

Accepted to the Workshop on Structured Data for Health at ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

2. Shared-Benchmark Regime Decomposition

We model nonstationary clinical behavior through a shared benchmark and a switching regime layer. The benchmark is supported by an unobserved system state $z(t)$ in continuous time, while the regime layer is supported by a latent regime u_m that evolves on the decision clock. Holding the benchmark fixed across regimes makes regime-dependent deviations comparable across irregularly sampled trajectories, in contrast to formulations that absorb nonstationarity into objective drift or unconstrained policy mixtures. Figure 1 summarizes the overall structure of the model.

2.1. Setup and shared benchmark

We observe N trajectories $\mathcal{D}^i = \{(t^i_m, o^i_m, a^i_m)\}_{m=1}^{M_i}$ at irregular decision epochs t^i_m, with categorical observations $o^i_m \in \mathcal{O}$ and finite actions $a^i_m \in \mathcal{A}$. A latent state $z(t) \in \{1, \dots, K_z\}$ evolves between decisions and drives outcomes; because $z(t)$ is unobserved, the agent acts on the belief $b^i_m(k) = \Pr(z(t^i_m) = k \mid o^i_{1:m}, a^i_{1:m-1})$ together with continuous outcome proxies $\lambda^i_m \in (0,1)^{|\mathcal{L}|}$. We write the augmented decision state as $x^i_m = (b^i_m, \lambda^i_m)$.

Let $r_{\text{base}}(z, \lambda, a) = r(\lambda)$ denote an outcome-grounded benchmark reward. The shared benchmark action-value is

$$Q_{\text{base}}(b, \lambda, a) = \mathbb{E}\Big[\int_0^\infty e^{-\rho t}\, r(\lambda(t))\, dt \,\Big|\, b(0) = b, \lambda(0) = \lambda, a(0) = a\Big], \tag{1}$$

with discount $\rho > 0$ and piecewise-constant actions on $[t_m, t_{m+1})$. By construction Q_{base} is common across trajectories and time and *independent* of the regime layer; this is the structural anchor that allows regime-dependent deviations to be compared across irregular timing.

2.2. Continuous-time latent dynamics

Between decision epochs, $z(t)$ evolves as an action-modulated CTMC with generator entries $q_{jk}(a) = \exp(\theta^\top_{jk}\phi(a))$, so the propagator over $\Delta_m = t_{m+1} - t_m$ is $P(\Delta_m; a_m) = \exp(Q(a_m)\Delta_m)$ and irregular timing is handled directly. Outcome proxies follow state-dependent ODEs $\dot{\lambda}_\ell(t) = \alpha^{(z(t))}_\ell(1 - \lambda_\ell(t)) - \beta^{(z(t))}_\ell \lambda_\ell(t)$, and categorical emissions $p_\theta(o \mid z, \lambda)$ distinguish an "on-target" (Normal) category from off-target ones via split parameters $\phi_{k\ell,c} \in (0,1)$. Beliefs are updated by a standard predict–update step on the CTMC; precise emission and filtering equations are deferred to Appendix J.

2.3. Switching regime layer

A latent regime $u^i_m \in \{1, \dots, K_u\}$ evolves on the *decision clock* as a sticky Markov chain with transition matrix Π ($\Pr(u^i_m = v \mid u^i_{m-1} = u) = \Pi_{uv}$). Given a feature map $g(b, \lambda, a) \in \mathbb{R}^d$ on action and context features, regime-dependent wedges are linear shadow costs

$$C(b, \lambda, a; u) = w^\top_u g(b, \lambda, a), \tag{2}$$

and observed actions follow the random-utility logit

$$p(a \mid b, \lambda, u) = \frac{\exp\{\beta\big(Q_{\text{base}}(b, \lambda, a) - w^\top_u g(b, \lambda, a)\big)\}}{\sum_{a'} \exp\{\beta\big(Q_{\text{base}}(b, \lambda, a') - w^\top_u g(b, \lambda, a')\big)\}}, \tag{3}$$

with inverse-temperature $\beta > 0$. Holding Q_{base} fixed across regimes while $C(\cdot; u)$ switches is the core identifying restriction; an ℓ_1 penalty on w_u yields sparse, interpretable regime profiles that double as targets for counterfactual relaxation.

2.4. Three-stage learning and counterfactuals

We fit SBRD in three stages. **Stage I** learns the dynamics parameters $\Theta_{\text{dyn}} = \{\theta^{(z)}, \alpha, \beta, \phi, \pi_0\}$ by EM on the CTMC–ODE–emission model (smoothing E-step, parameter M-step) and produces belief–proxy pairs $\{x^i_m\}$ and benchmark values $\{Q_{\text{base}}(x^i_m, a)\}$ via discounted CTMC–ODE rollouts under intervened actions. **Stage II** fixes these benchmark objects and maximizes the regime-layer log-likelihood

$$\mathcal{L}(\Theta_{\text{reg}}) = \sum_{i=1}^{N} \log \sum_{u_{1:M_i}} p_\Pi(u_{1:M_i}) \prod_{m=1}^{M_i} p(a^i_m \mid x^i_m, u^i_m), \tag{4}$$

with $\Theta_{\text{reg}} = \{\Pi, w_1, \dots, w_{K_u}, \psi\}$ (ψ parameterizes a learned feature component of g_ψ, kept distinct from the emission split parameters $\phi_{k\ell,c}$), under the ℓ_1 penalty $\lambda \sum_u \|w_u\|_1$. We use generalized EM: the E-step is forward–backward over regimes, the M-step updates Π by expected transition counts and $(\{w_u\}, \psi)$ by maximizing the posterior-weighted, ℓ_1-penalized multinomial logit. **Stage III** treats the fitted layers as a structural causal model: it computes counterfactual benchmark-optimal actions and evaluates "constraint-relaxation" interventions that scale or remove wedge groups while holding the dynamics and Q_{base} fixed (Appendix H).

Theoretical guarantees. Conditional on Q_{base} and a fixed feature map, the regime sequence and pairwise wedge differences are identifiable up to label permutation and per-regime additive constants (Theorem B.3); the sparse regime layer is stable to plug-in error in Stage I quantities, with the regime log-likelihood perturbed by $O\big(\beta(\varepsilon_Q + (L_Q + WL_g)(\varepsilon_b + \varepsilon_\lambda))\big)$ (Theorem B.5); and Stage III interventions admit an SCM-consistent counterfactual interpretation. Precise statements, additional consistency, oracle, and segmentation results, and proofs are in Appendix B.

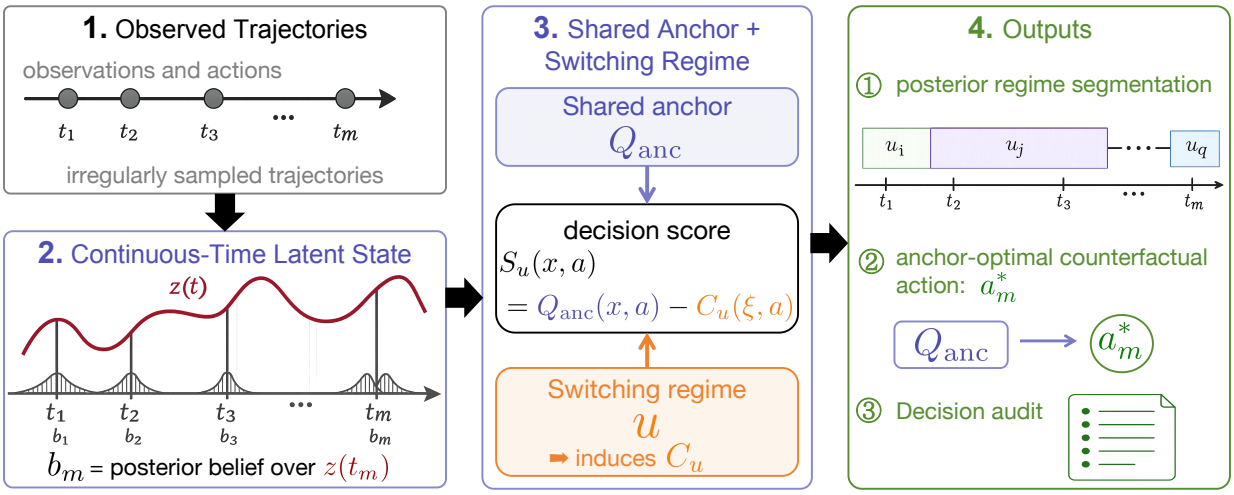


Figure 1. **Shared-Benchmark Regime Decomposition (SBRD).** Panel 1: irregularly sampled trajectories of structured observations and actions. Panel 2: a continuous-time latent state $z(t)$ supporting benchmark construction under partial observability and irregular timing. Panel 3: the core decomposition, in which a shared benchmark Q_{base} captures stable decision logic, while a switching regime u_m induces persistent deviations through $C(\cdot; u)$. Panel 4: outputs include posterior regime segmentation, counterfactual benchmark-optimal action a^*, and an LLM-based audit of latent decision pressures.

3. Experiments

We evaluate SBRD on two clinical decision-making cohorts with markedly different timing and operational character: high-frequency ICU sepsis treatment from MIMIC-IV (Johnson et al., 2024; 2023) and long-horizon chronic care from a private CKD-MBD cohort. Cohort details, action/observation spaces, and preprocessing are deferred to Appendix C. We ask whether the regime layer (i) recovers persistent, interpretable deviations from the shared benchmark, (ii) improves held-out action prediction over stationary and interpretable behavioral baselines, and (iii) supports counterfactual benchmark-optimal actions for diagnosing where regime-dependent wedges account for recoverable benchmark value.

3.1. Interpretable regime decomposition

The fitted dynamics layer recovers K_z outcome-grounded phenotypes (Appendix E) and produces Q_{base}. Conditional on these benchmark objects, we estimate the regime layer with $K_u = 3$ in each cohort. The wedges w_u act on a 16-dimensional MIMIC-IV feature map grounded in sepsis-management protocols (compliance violations, treatment-group counts, regimen-change features) and an 8-dimensional CKD-MBD map covering long-term safety and resource trade-offs (mineral-imbalance synergies, polypharmacy, medication cost); see Appendix D. We use MIMIC-IV as the main-text example.

The recovered MIMIC-IV regimes (Figure 2(1)) split into three persistent profiles. *Regime 1*, the most prevalent, loads heavily on `added_count` and `n_groups_active`, consistent with acute escalation through multi-modal intervention. *Regime 0* concentrates on `forbid_violate` indicators, consistent with protocol-adherent safety boundaries. *Regime 2* loads on `drug_count` and `removed_count`, consistent with complexity management or de-escalation. The transition matrix is sticky ($\Pi_{uu} > 0.85$ for all u; Figure 2(3)), so the regimes describe persistent decision modes rather than per-step action noise. Trajectory-level diagnostics that show coupled and decoupled dynamics between $z(t)$ and u_m appear in Appendix K; analogous CKD-MBD regime profiles appear in Appendix M.

3.2. Held-out predictive performance

We compare SBRD against (a) a stationary one-regime SBRD (K_u=1); (b) Maximum Entropy IRL (Ziebart et al., 2008); and (c) Soft Decision Trees (Frosst & Hinton, 2017). SBRD with K_u=3 uniformly improves held-out action log-likelihood and top-5 accuracy on both cohorts (Table 1); MaxEnt IRL is restricted to a stationary reward and cannot represent goal shifts, while SDT lacks generative structure for intent inference. A sweep over $K_u \in \{1, \ldots, 6\}$ plateaus at small K_u with persistently high self-transitions, indicating parsimonious, stable regime structure (Appendix L); the advantage of K_u>1 persists across three reward families (Appendix F). CTMC+ODE dynamics reliability against ODE-RNN (Habiba & Pearlmutter, 2020) and NCDSSM (Ansari et al., 2023) is reported in Appendix G.

3.3. Constraint-relaxation counterfactuals

We use the fitted regime layer as a benchmark-relative diagnostic: relaxing wedge groups identifies which regime-

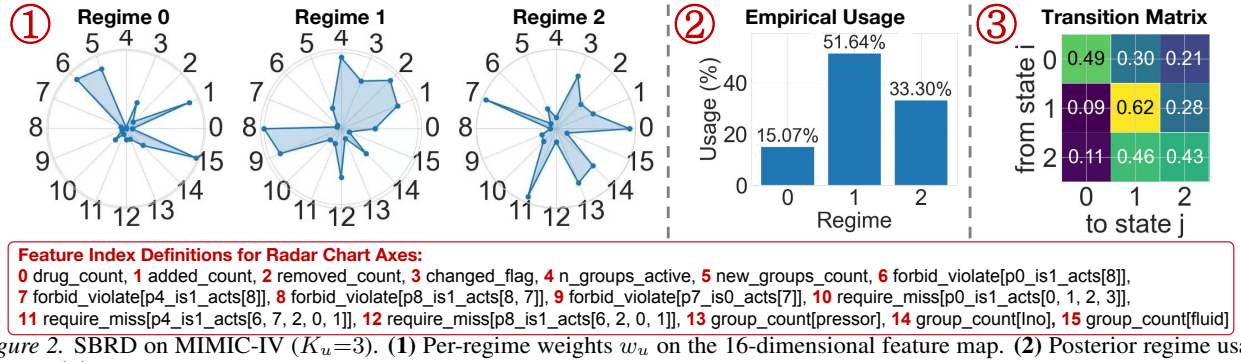


Figure 2. SBRD on MIMIC-IV (K_u=3). **(1)** Per-regime weights w_u on the 16-dimensional feature map. **(2)** Posterior regime usage $\sum_m \gamma_m(u)$. **(3)** Inter-regime transition matrix Π.

Table 1. Held-out action prediction across clinical domains. SBRD improves over interpretable behavioral baselines on both cohorts (mean $\pm$ SD).

Model	Log-L ↑	Acc@5 ↑
CKD-MBD		
SBRD (Ours)*	$\mathbf{-4.52 \pm 0.85}$	$\mathbf{50.5\% \pm 5.3\%}$
MaxEnt IRL†	-5.64 ± 0.00	$29.3\% \pm 0.7\%$
SDT‡	-5.23 ± 0.25	$15.4\% \pm 4.2\%$
MIMIC-IV		
SBRD (Ours)*	$\mathbf{-0.74 \pm 0.02}$	$\mathbf{96.8\% \pm 0.2\%}$
MaxEnt IRL†	-5.29 ± 0.07	$80.1\% \pm 0.7\%$
SDT‡	-1.84 ± 0.00	$93.4\% \pm 0.0\%$

* Optimal action (a^*) + regimes (u) + phenotypes (z).
† State–action reward $r(s, a)$ on empirical FH-MDP.
‡ Hierarchical, discriminative decision paths.

dependent deviations account for the largest recoverable benchmark value. For MIMIC-IV we group the 16-dimensional features into *Compliance* (protocol violations, required-treatment omissions), *Complexity* (regimen size, step-to-step changes), and *Resource* (high-intensity therapeutic breadth); for each group G we scale the corresponding wedge weights by $\alpha \in [0, 1]$, recompute the action distribution under (3), and report the relative gain in Q_{base}.

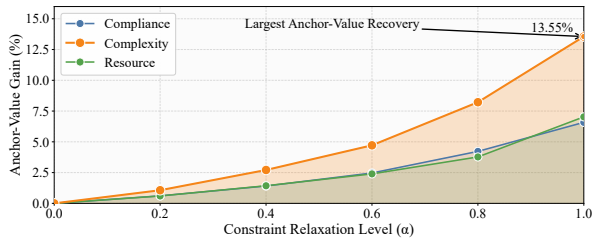


Figure 3. Benchmark recovery under group-wise constraint relaxation (MIMIC-IV). Curves report the relative increase in Q_{base} as the relaxation level α increases.

Figure 3 shows that complexity-related wedges yield the largest recovery, 13.55% at full relaxation, compared with 6.58% for compliance and 7.03% for resource-related wedges. Conditional on the fitted shared benchmark, treatment-complexity and adjustment frictions therefore account for the largest recoverable deviation in the MIMIC-IV cohort. Clinically, this is consistent with sepsis management, where the operational burden of multi-modal escalation, rather than outright protocol violation, is the dominant source of benchmark-relative slack. An analogous CKD-MBD analysis (Appendix M) shows risk-mitigation and safety/simplification wedges dominating the recoverable gap, indicating that the diagnostic generalizes from acute to chronic care without re-tuning the wedge map.

3.4. Case-level decision auditing

SBRD also couples the structural outputs $\{a^*_m, u_m, z_m\}$ with an LLM-based auditor producing, per trajectory, a rationality check on the synthesized benchmark a^*, a wedge-level interpretation of the a vs. a^* gap via the active regime u_m, and a fidelity check linking $z(t)$ and u_m shifts to clinical events. The synthesized a^* uses an SCM-consistent abduction–action–prediction rollout with fixed exogenous noise (Appendix H). On a representative CKD-MBD trajectory, the auditor flags a sustained Regime R_2 (Safety/Simplification) episode driven by hypercalcemia risk, attributing the a vs. a^* gap to toxicity-avoidance pressure rather than to violation of the underlying KDIGO benchmark, an explanation that purely predictive baselines cannot produce. Prompts and the full case study are in Appendix I.

4. Conclusion

Shared-Benchmark Regime Decomposition (SBRD) disentangles an outcome-grounded shared benchmark from switching regime-dependent wedges in nonstationary clinical trajectories. Holding the benchmark fixed while a sparse regime layer absorbs persistent deviations, SBRD yields interpretable regime profiles, improves held-out action prediction over stationary and interpretable baselines, and enables benchmark-relative constraint-relaxation diagnostics that localize recoverable value across decision frictions.

References

Aguirregabiria, V. and Mira, P. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38–67, 2010. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2009.09.007. URL https://www.sciencedirect.com/science/article/pii/S0304407609001985. Structural Models of Optimization Behavior in Labor, Aging, and Health.

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

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

Chari, V. V., Kehoe, P. J., and McGrattan, E. R. Business cycle accounting. *Econometrica*, 75(3):781–836, 2007.

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

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

Choi, S., Kang, H., Kim, N., and Kim, J. How does artificial intelligence improve human decision-making? evidence from the ai-powered go program. *Strategic Management Journal*, 46(6):1523–1554, 2025. doi: https://doi.org/10.1002/smj.3694. URL https://sms.onlinelibrary.wiley.com/doi/abs/10.1002/smj.3694.

Congdon, W., Kling, J., Ludwig, J., and Mullainathan, S. Chapter 4 - social policy: Mechanism experiments and policy evaluationsa. In Banerjee, A. V. and Duflo, E. (eds.), *Handbook of Economic Field Experiments*, volume 2 of *Handbook of Economic Field Experiments*, pp. 389–426. North-Holland, 2017. doi: https://doi.org/10.1016/bs.hefe.2016.10.004. URL https://www.sciencedirect.com/science/article/pii/S2214658X16300186.

Deshpande, S., Walambe, R., Kotecha, K., Selvachandran, G., and Abraham, A. Advances and applications in inverse reinforcement learning: a comprehensive review. *Neural Computing and Applications*, pp. 1–53, 2025.

Evans, L., Rhodes, A., Alhazzani, W., Antonelli, M., Coopersmith, C. M., French, C., Machado, F. R., Mcintyre, L., Ostermann, M., Prescott, H. C., et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Critical care medicine*, 49 (11):e1063–e1143, 2021.

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

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

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

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

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

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

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

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Johnson, D. G. Algorithmic accountability in the making. *Social Philosophy and Policy*, 38(2):111–127, 2021. doi: 10.1017/S0265052522000073.

Jung, K. and Shah, N. H. 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.

Keane, M. P., Todd, P. E., and Wolpin, K. I. Chapter 4 - the structural estimation of behavioral models: Discrete choice dynamic programming methods and applications. volume 4 of *Handbook of Labor Economics*, pp. 331–461. Elsevier, 2011. doi: https://doi.org/10.1016/S0169-7218(11)00410-2.

URL https://www.sciencedirect.com/science/article/pii/S0169721811004102.

Kim, C.-J. and Nelson, C. R. *State-space models with regime switching: classical and Gibbs-sampling approaches with applications*. MIT press, 2017.

Koene, A., Clifton, C., Hatada, Y., Webb, H., and Richardson, R. A governance framework for algorithmic accountability and transparency. 2019.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

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

Maskin, E. S. Mechanism design: How to implement social goals. *American Economic Review*, 98(3):567–76, June 2008. doi: 10.1257/aer.98.3.567. URL https://www.aeaweb.org/articles?id=10.1257/aer.98.3.567.

Moe, S. M., Drüeke, T., Lameire, N., and Eknoyan, G. Chronic kidney disease–mineral-bone disorder: a new paradigm. *Advances in chronic kidney disease*, 14(1): 3–12, 2007.

Ng, A. Y., Russell, S., et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

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

Pearl, J. *Causality*. Cambridge University Press, 2 edition, 2009b.

Rubanova, Y., Chen, R. T., and Duvenaud, D. K. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Rudin, C. 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.

Rust, J. Chapter 51 structural estimation of markov decision processes. volume 4 of *Handbook of Econometrics*, pp. 3081–3143. Elsevier, 1994. doi: https://doi.org/10.1016/S1573-4412(05)80020-0. URL https://www.sciencedirect.com/science/article/pii/S1573441205800200.

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

Sims, C. A. Implications of rational inattention. *Journal of Monetary Economics*, 50 (3):665–690, 2003. ISSN 0304-3932. doi: https://doi.org/10.1016/S0304-3932(03)00029-1. URL https://www.sciencedirect.com/science/article/pii/S0304393203000291. Swiss National Bank/Study Center Gerzensee Conference on Monetary Policy under Incomplete Information.

Vassiliades, A., Polatidis, N., Samaras, S., Diplaris, S., Martin, I. C., Manolopoulos, Y., Vrochidis, S., and Kompatsiaris, I. Utilizing large language models for machine learning explainability, 2025. URL https://arxiv.org/abs/2510.06912.

Yu, L., Yu, T., Finn, C., and Ermon, S. Meta-inverse reinforcement learning with probabilistic context variables. *Advances in neural information processing systems*, 32, 2019.

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

A. Related Work

Our work interfaces with research programs in behavioral modeling, structural econometrics, and algorithmic governance. On the normative side, our approach is motivated by the foundational literature on value alignment, as formalized in Cooperative Inverse Reinforcement Learning (CIRL) (Hadfield-Menell et al., 2016), while building on empirical evidence that high-performing AI recommendations can measurably improve expert decisions (Choi et al., 2025). Unlike standard imitation learning or behavior cloning that prioritizes predictive accuracy, SBRD follows the structural econometrics tradition of explaining observed choices through latent primitives and frictions. This tradition is rooted in dynamic discrete choice models (Keane et al., 2011; Aguirregabiria & Mira, 2010) and the structural estimation of Markov decision processes (Rust, 1994). By interpreting systematic deviations from a frictionless benchmark as time-varying distortions, or "wedges" (Chari et al., 2007), which admit a natural interpretation as shadow prices of binding constraints, we rationalize seemingly sub-optimal actions as optimal responses to latent institutional or operational constraints (Sims, 2003).

We specifically contrast SBRD with Inverse Reinforcement Learning (IRL) (Ng et al., 2000), which typically assumes a stationary reward function or handles heterogeneity by learning distinct, multiple objectives (Bloem & Bambos, 2015; Yu et al., 2019). While prior work has utilized mixture-of-experts or state-clustering to identify diverse behavioral modes (Deshpande et al., 2025), SBRD posits a shared normative objective across agents. This allows us to decouple transient physiological fluctuations from persistent strategic shifts, moving beyond black-box behavior cloning toward a model where key frictions are explicitly parameterized and auditable (Rudin, 2019). This framing aligns with contemporary models of boundedly rational agents operating under organizational constraints, shifting the focus from identifying what an agent intends to do to identifying what prevents them from reaching the normative optimum.

The non-stationarity of decision logic, where both clinical practices and institutional incentives evolve over time (Jung & Shah, 2015), poses a challenge for longitudinal modeling. Drawing on advances in continuous-time modeling for irregularly-sampled data, including Neural Ordinary Differential Equations (Rubanova et al., 2019) and continuous-discrete state space models (Habiba & Pearlmutter, 2020; Ansari et al., 2023), SBRD adopts a latent Continuous-Time Markov Chain (CTMC) to capture underlying system dynamics. Unlike discrete-time regime-switching models that conflate behavioral persistence with observation granularity, the continuous-time formulation enables SBRD to disentangle genuine decision "stickiness" from artifacts of asynchronous logging and delayed feedback (Ghahramani & Hinton, 2000). By evaluating the persistence of inferred regimes in real time rather than per decision step, we distinguish stochastic decision noise from stable, latent organizational anchors, yielding a robust framework for longitudinal clinical decision analysis under irregular and asynchronously observed records.

Finally, SBRD provides a methodology for welfare gap analysis and algorithmic accountability (Koene et al., 2019; Johnson, 2021). We ground our counterfactual engine in the Structural Causal Model (SCM) framework (Pearl, 2009b), adopting the standard *abduction–action–prediction* workflow for counterfactual queries (Pearl, 2009a). This connects our work to Mechanism Design, which studies how institutional rules implement social objectives (Maskin, 2008), and Counterfactual Policy Evaluation (CPE), which quantifies the welfare consequences of alternative interventions (Congdon et al., 2017). We further extend this paradigm by utilizing Large Language Models (LLMs) as automated institutional auditors (Vassiliades et al., 2025), bridging the gap between latent structural parameters and actionable governance insights. This transformation enables clinical decision-makers to retrospectively quantify the value lost to institutional frictions, shifting the dialogue from simple treatment recommendation toward systemic clinical optimization.

B. Theoretical Properties, Identifiability, and Proofs

This appendix collects the theoretical content referenced in the "Theoretical guarantees" callout at the end of Section 2.4. It restates the three main results (Theorem B.3, Proposition B.4, Theorem B.5) together with their setup and informal interpretation, and then provides precise assumptions and proofs, plus additional statistical guarantees: (quasi-)MLE consistency for the regime layer, oracle-type bounds for sparse wedge recovery, localization bounds for regime segmentation under stickiness and emission separation, strict dominance over stationary wedge models unless stationarity holds, and SCM-based counterfactual correctness and consistency of counterfactual policy evaluation.

Throughout, we treat (x_m, Q_{base}) as given when analyzing the regime layer, and then return to plug-in robustness when (x_m, Q_{base}) are estimated in Stage I.

B.1. Main-text results: setup and statements

Setup. We observe trajectories $\mathcal{D} = \{(x_m^i, a_m^i)\}_{i=1..N,\ m=1..M_i}$ where $x_m = (b_m, \lambda_m)$ is the augmented decision state produced by the dynamics layer (Section 2) and $a_m \in \mathcal{A}$ is a discrete action. The regime layer posits a latent regime $u_m \in \{1, \dots, K\}$ (with $K = K_u$) evolving as a Markov chain with transition matrix Π. Conditional on (x_m, Q_{base}), actions follow the same logit rule as in Section 2.3, written here for convenience:

$$p_\Theta(a \mid x, u) = \frac{\exp\{\beta\,[Q_{\text{base}}(x,a) - C_\Theta(x,a;u)]\}}{\sum_{a' \in \mathcal{A}} \exp\{\beta\,[Q_{\text{base}}(x,a') - C_\Theta(x,a';u)]\}}, \tag{5}$$

where $C_\Theta(x,a;u) = w_u^\top g_\phi(x,a)$ and $\Theta = \big(\Pi, \{w_u\}_{u=1}^K, \phi\big)$. The induced HMM likelihood is

$$p_\Theta(a_{1:M} \mid x_{1:M}) = \sum_{u_{1:M}} p_\Pi(u_1) \prod_{m=2}^{M} p_\Pi(u_m \mid u_{m-1}) \prod_{m=1}^{M} p_\Theta(a_m \mid x_m, u_m). \tag{6}$$

Identifiability of regimes and wedges (conditional on the benchmark).

Definition B.1 (Equivalence)**.** Two parameter sets Θ and Θ' are equivalent, written $\Theta \sim \Theta'$, if they induce the same conditional action distributions for all (x, a) and the same regime-transition law up to a permutation of regime labels.

Assumption B.2 (Regularity for identification (informal))**.** (i) $Q_{\text{base}}(x, \cdot)$ is not constant across actions on a set of positive probability; (ii) different regimes induce different wedge differences on a set of positive probability; and (iii) the regime Markov chain is ergodic and emissions satisfy a full-rank condition sufficient for finite-state HMM identifiability. A precise statement appears later in this appendix (Section B.4).

Theorem B.3 (Identifiability up to label swapping and regime-wise offsets)**.** *Under Assumption B.2, and conditional on Q_{base} and the feature map g_ϕ, the regime-layer model (5)–(6) is identifiable up to: (i) permutation of regime labels, and (ii) regime-wise additive constants $C(x,a;u) \mapsto C(x,a;u) + \kappa_u$ that do not depend on a. Equivalently, for each regime u, the collection of pairwise wedge differences $\Delta C_u(x;a,a') := C(x,a;u) - C(x,a';u)$ is identifiable (up to label permutation).*

Economic interpretation: wedges as shadow prices.

Proposition B.4 (Entropy-regularized benchmark-minus-wedge)**.** *Fix (x, u) and define $V_u(x,a) := Q_{\text{base}}(x,a) - w_u^\top g_\phi(x,a)$. Then the choice probabilities (5) are the unique maximizer of*

$$\max_{\pi \in \Delta(\mathcal{A})} \Big\{ \mathbb{E}_{a\sim\pi}[V_u(x,a)] + \frac{1}{\beta} H(\pi) \Big\}, \qquad H(\pi) = -\sum_{a \in \mathcal{A}} \pi(a) \log \pi(a).$$

End-to-end robustness to plug-in benchmark error. Let $\widehat{x}_m = (\widehat{b}_m, \widehat{\lambda}_m)$ and $\widehat{Q}_{\text{base}}$ be Stage I outputs and suppose

$$\sup_{m,a} \big|\widehat{Q}_{\text{base}}(\widehat{x}_m, a) - Q_{\text{base}}(x_m, a)\big| \le \varepsilon_Q, \quad \sup_m \|\widehat{b}_m - b_m\|_1 \le \varepsilon_b, \quad \sup_m \|\widehat{\lambda}_m - \lambda_m\|_1 \le \varepsilon_\lambda.$$

Theorem B.5 (Stability of regime likelihood under baseline errors)**.** *Assume $g_\phi(\cdot, a)$ is L_g-Lipschitz in $x = (b, \lambda)$ and $Q_{\text{base}}(\cdot, a)$ is L_Q-Lipschitz. Let $W = \max_u \|w_u\|_1$ on the parameter set. Then for any fixed regime-layer parameters,*

$$\sup_{m,u} \Big| \log p_\Theta(a_m \mid \widehat{x}_m, u; \widehat{Q}_{\text{base}}) - \log p_\Theta(a_m \mid x_m, u; Q_{\text{base}}) \Big| \;\le\; 2\beta \Big(\varepsilon_Q + (L_Q + W L_g)(\varepsilon_b + \varepsilon_\lambda) \Big).$$

B.2. Preliminaries and notation

Fix the action set $\mathcal{A}$ with $|\mathcal{A}| \ge 2$. Let K denote the number of constraint regimes ($K = K_u$ in the main text). For a fixed feature map g_ϕ, define for each regime $u \in \{1, \dots, K\}$ the wedge $C(x,a;u) = w_u^\top g_\phi(x,a)$, and the regime-specific (scaled) utility $U_u(x,a) := \beta\big(Q_{\text{base}}(x,a) - C(x,a;u)\big)$. The corresponding multinomial-logit emission kernel is $e_u(a \mid x) := p_\Theta(a \mid x, u) = \exp\{U_u(x,a)\} / \sum_{a' \in \mathcal{A}} \exp\{U_u(x,a')\}$.

B.3. Useful lemmas

Lemma B.6 (Logit identifies utility differences)**.** *Fix x and regime u. If $e_u(\cdot \mid x)$ is strictly positive on $\mathcal{A}$, then for any $a, a' \in \mathcal{A}$, $\log\big(e_u(a \mid x)/e_u(a' \mid x)\big) = U_u(x,a) - U_u(x,a')$. Consequently, $\{U_u(x,a)\}_{a\in\mathcal{A}}$ is identified up to an additive constant (independent of a), and all pairwise differences are uniquely identified.*

Lemma B.7 (Log-sum-exp is 1-Lipschitz in $\|\cdot\|_\infty$)**.** *For any $r, s \in \mathbb{R}^{|\mathcal{A}|}$, $\big|\log\sum_a \exp(r_a) - \log\sum_a \exp(s_a)\big| \le \|r-s\|_\infty$.*

Lemma B.8 (A sufficient identifiability condition for finite-state HMMs (informal))**.** *Consider a K-state HMM with ergodic transition matrix Π and emission vectors $\{e_u(\cdot \mid x^{(t)})\}_{u=1}^K$ that are linearly independent for enough contexts $\{x^{(t)}\}$. Then $(\Pi, \{e_u\}_{u=1}^K)$ is identifiable from the joint law of $(a_{1:M} \mid x_{1:M})$ up to a permutation of hidden-state labels.*

B.4. Precise identification assumptions

Assumption B.9 (Non-degeneracy of Q_{base})**.** There exists a measurable set $\mathcal{X}_0$ with $\mathbb{P}(x \in \mathcal{X}_0) > 0$ such that for every $x \in \mathcal{X}_0$, the vector $\{Q_{\text{base}}(x,a)\}_{a\in\mathcal{A}}$ is not constant in a.

Assumption B.10 (Positivity / interiority)**.** There exists $\underline{p} > 0$ such that for all regimes u, all contexts x in the support, and all actions a, $e_u(a \mid x) \ge \underline{p}$ and $\Pi_{uv} \ge \underline{p}$ for all u, v.

Assumption B.11 (Regime separation through wedge differences)**.** For any two regimes $u \ne v$, there exists (x, a, a') in the support such that $(w_u - w_v)^\top\big(g_\phi(x,a) - g_\phi(x,a')\big) \ne 0$.

Assumption B.12 (Full-rank emissions on separating contexts)**.** There exist contexts $\{x^{(t)}\}_{t=1}^T$ in the support of x, with $T \ge K$ and $|\mathcal{A}| \ge K+1$, such that for each t the $|\mathcal{A}| \times K$ matrix $E^{(t)}$ whose uth column is $\big(e_u(a \mid x^{(t)})\big)_{a\in\mathcal{A}}$ has rank K.

Assumption B.13 (Ergodicity and stickiness of regime transitions)**.** Π is irreducible and aperiodic. In addition, $\min_u \Pi_{uu} \ge 1-\epsilon$ for some $\epsilon \in (0,1)$.

B.5. Proof of Theorem B.3

Proof. Fix a regime u and a context x. By Lemma B.6, $U_u(x,a) - U_u(x,a') = \log\big(e_u(a \mid x)/e_u(a' \mid x)\big)$, so $\{U_u(x,a)\}_{a\in\mathcal{A}}$ is identified up to an additive constant. Since $U_u(x,a) = \beta(Q_{\text{base}}(x,a) - C(x,a;u))$ and Q_{base} is treated as known, Assumption B.9 implies the action-dependent term cannot be absorbed into an action-independent constant. Hence pairwise wedge differences $\Delta C_u(x;a,a')$ are identified, and adding a regime-wise constant κ_u leaves $e_u(\cdot \mid x)$ unchanged, so identification is up to such offsets. Assumptions B.11–B.13 and Lemma B.8 yield identifiability of $(\Pi, \{e_u\})$ up to permutation of regime labels. □

B.6. Proof of Proposition B.4

Proof. The Lagrangian first-order condition for $\max_{\pi\in\Delta(\mathcal{A})} \sum_a \pi(a)V_u(x,a) + \beta^{-1}H(\pi)$ gives $\log\pi(a) = \beta(V_u(x,a) + \nu) - 1$, i.e., $\pi(a) \propto \exp\{\beta V_u(x,a)\}$. Normalizing yields the logit in (5). Uniqueness follows from strict concavity. □

B.7. Proof of Theorem B.5

Proof. Let $r_a := \beta(\widehat{Q}_{\text{base}}(\widehat{x}_m, a) - w_u^\top g_\phi(\widehat{x}_m, a))$ and $s_a := \beta(Q_{\text{base}}(x_m, a) - w_u^\top g_\phi(x_m, a))$. Then the difference of log choice probabilities is $(r_{a_m} - s_{a_m}) - (\log\sum_a e^{r_a} - \log\sum_a e^{s_a})$, whose absolute value is at most $2\|r-s\|_\infty$ by Lemma B.7. Lipschitz bounds on Q_{base} and g_ϕ, together with $\|w_u\|_1 \le W$, give $\|r-s\|_\infty \le \beta(\varepsilon_Q + (L_Q + WL_g)(\varepsilon_b + \varepsilon_\lambda))$, which yields the claim. □

B.8. Additional statistical guarantees for the regime layer

(Quasi-)MLE consistency. Let $\ell_i(\Theta) := \log p_\Theta(a^i_{1:M_i} \mid x^i_{1:M_i})$. Under i.i.d. sampling with bounded (or exponentially-tailed) trajectory lengths, compactness, continuity, positivity (B.10), and ergodicity (B.13), any MLE $\hat{\Theta}_N \in \arg\max_\Theta \sum_i \ell_i(\Theta)$ converges (modulo $\sim$) to $\Theta^\star$ in the well-specified case (Theorem B.14), and to the KL projection in the misspecified case (Theorem B.15); proofs are standard M-estimation.

Theorem B.14 (Strong consistency of the regime-layer MLE (well-specified case))**.** *Under the assumptions above, if the true process is in the model class and the model is identifiable modulo $\sim$, then $\hat{\Theta}_N \to \Theta^\star$ a.s. as $N \to \infty$, in the quotient space.*

Theorem B.15 (Quasi-MLE: convergence to the KL projection)**.** *If the true process is not in the model class, $\hat{\Theta}_N$ converges (modulo $\sim$) to the maximizers of $\mathbb{E}[\ell_1(\Theta)]$, equivalently the KL projection of the true conditional law onto the model class.*

Sparse wedge recovery (oracle-type bounds). The Stage II M-step for $\{w_u\}$ is a posterior-weighted multinomial logit with an ℓ_1 penalty.

Assumption B.16 (RSC and score concentration for weighted logit)**.** Fix a regime u and assume $\|g_\phi(x,a)\|_\infty \le G$ a.s. Assume restricted strong convexity over s-sparse directions and concentration of the weighted score with rate $G\sqrt{\log d/(N\bar{M})}$, where $\bar{M} = N^{-1}\sum_i M_i$.

Theorem B.17 (Oracle inequality for sparse wedge vectors)**.** *Assume g_ϕ is fixed and each $w_u^\star$ is s-sparse. Choose $\lambda \asymp G\sqrt{\log d/(N\bar{M})}$. Then with high probability $\sum_u \|\hat{w}_u - w_u^\star\|_2 \lesssim K\sqrt{s\log d/(N\bar{M})}$ and $\sum_u \|\hat{w}_u - w_u^\star\|_1 \lesssim Ks\sqrt{\log d/(N\bar{M})}$.*

B.9. Segmentation localization under stickiness

Let $u_{1:M}^\star$ be the true regime path and S the number of switches. Let $\hat{u}_{1:M}$ be the Viterbi path.

Assumption B.18 (Emission separation)**.** There exists $\delta > 0$ such that for any $u \ne v$, $\mathbb{E}[\log(e_u(a \mid x)/e_v(a \mid x)) \mid u^\star = u] \ge \delta$.

Theorem B.19 (Segmentation errors localize near switch points)**.** *Under positivity, ergodicity, and emission separation, with probability at least $1-\eta$ the misclassified set $\{m : \hat{u}_m \ne u_m^\star\}$ has total length at most $C\,S\log(M/\eta)$. In particular, the misclassification fraction vanishes if $S = o(M/\log M)$.*

B.10. Dominance over stationary wedge models

Theorem B.20 (Strict likelihood dominance unless stationarity holds)**.** *If the true regime-layer process belongs to $\mathcal{M}_K$ ($K \ge 2$) and is not observationally equivalent to any model in $\mathcal{M}_1$ (the stationary wedge submodel), then the population log-likelihood gap $\sup_{\Theta\in\mathcal{M}_K}\mathbb{E}[\ell_1(\Theta)] - \sup_{\theta\in\mathcal{M}_1}\mathbb{E}[\ell_1(\theta)]$ is strictly positive and equal to the minimum KL divergence to $\mathcal{M}_1$. Empirical likelihood-ratio differences grow at the effective sample size rate.*

B.11. Counterfactual validity under SCM semantics

Assumption B.21 (SCM structure and invariance)**.** There exists a structural causal model in which $(z(t), \lambda(t))$ and observations are deterministic functions of actions $a(t)$ and exogenous noise ε, and ε is invariant to interventions on $a(t)$.

Theorem B.22 (Unit-level counterfactual correctness)**.** *Under Assumption B.21, fixing an abducted noise realization ε consistent with the factual history and applying $\mathrm{do}(a \leftarrow \tilde{a})$ yields counterfactual samples satisfying consistency and composition.*

Theorem B.23 (Consistency of counterfactual policy evaluation)**.** *For any counterfactual policy π, if Stage III dynamics coincide with the true SCM and abduction–action–prediction is used, then the Monte Carlo estimator of discounted benchmark return under π converges to $\mathbb{E}[R(\pi)]$ as the number of rollouts $S \to \infty$.*

C. Experimental Contexts and Data Sources

We evaluate SBRD across two clinical decision-making contexts that span markedly different timing and operational character. CKD-MBD captures the intricate, slow-evolving "balancing act" of chronic outpatient mineral management, while MIMIC-IV provides a high-frequency view of acute physiological stabilization in the ICU. Together they test the framework against both persistent cognitive load and rapid environmental shifts under partial observability and irregular timing.

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. 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 (Evans et al., 2021), clinician decisions often deviate from stationary protocols. The rich, irregularly sampled clinical trajectories in MIMIC-IV provide an ideal testbed for SBRD, 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 drive these clinical decisions over time.

- **Patient cohort.** We extract 2,475 sequences of patients diagnosed with sepsis (Saria, 2018), using a 72-hour observation

Table 2. Treatments and output indicators in the 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 (creatinine_normal)* Normal arterial lactate (arterial_lactate_normal)*
Output	Low-urine

window with complete data records.

- **Treatments (actions).** The primary interventions are vasopressor therapy and fluid resuscitation, essential for stabilizing blood pressure and ensuring organ perfusion (Komorowski et al., 2018); these are the clinician's actions a_t. Vasopressors increase cardiac output via vasoconstriction; fluids restore intravascular volume. We categorize treatments into 3 fluid types, 4 vasopressors, and 2 inotropes, as detailed in Table 2.
- **Outcomes and observations.** We use real-time urine output and survival status as primary indicators (low urine output is a critical signal of impending septic shock). We further include eight informative lab measurements relevant to sepsis: four representing poor physiological states and four representing favorable outcomes. These variables let the latent dynamics model evaluate the counterfactual consequences of clinician actions.

Chronic Management (CKD-MBD). Chronic Kidney Disease–Mineral Bone Disorder is a systemic syndrome characterized by complex interactions between serum calcium (Ca), phosphorus (P), and parathyroid hormone (PTH) (Moe et al., 2007). 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 (Gaweda et al., 2022). We apply SBRD 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 of longitudinal clinical records for 150 patients collected between January 2020 and January 2024. Each trajectory includes laboratory measurements and the corresponding pharmacological interventions prescribed by attending physicians.
- **Treatments (actions).** The action space consists of 13 discrete markers representing dosage levels across calcium-based binders, phosphate binders, vitamin D analogs, and calcimimetics; the specific weighting of one drug over another typically reflects the clinician's underlying decision regime.
- **Outcomes and observations.** We monitor corrected calcium (cCa), phosphate (P), and parathyroid hormone (PTH); these are discretized into low, normal, and high categories, forming 10 outcome markers used as state observations. By

evaluating counterfactual outcomes for these three variables, SBRD can infer when a clinician shifts from prioritizing one metabolic target (e.g., phosphorus control) to another (e.g., PTH suppression) due to changing latent constraints.

- **Rationale.** CKD-MBD management 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 (Chen & Bushinsky, 2017). The dataset captures these subtle shifts in treatment strategy that are not explicitly recorded in the EHR but are reflected in diverging decision trajectories.

D. Structural Definitions of Regime-Defining Features

A central component of the SBRD framework is the feature vector $g(b, \lambda, a)$, which captures the structural "frictions" or institutional constraints that rationalize observed agent behavior. These features allow the model to distinguish between purely normative actions and those driven by latent decision regimes. The clinical features below 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.
  - **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`**: Binary indicator of whether any medication change occurred.
  - **4. `n_groups_active`**: Count of active coarse treatment groups (*pressor, fluid, or inotrope*).
  - **5. `new_groups_count`**: Number of newly introduced treatment groups.
- **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.
  - **7. `forbid_violate`[**$p_4_is1_acts[8]$]: Risk penalty for using *Water* (8) when lactate levels are high.
  - **8. `forbid_violate`[**$p_8_is1_acts[8, 7]$]: Risk penalty for using *Water* (8) or *Colloid* (7) during low urine output.
  - **9. `forbid_violate`[**$p_7_is0_acts[7]$]: Risk penalty for using *Colloid* (7) when creatinine is abnormal.
- **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 vasopressors during hypotension.
  - **11. `require_miss`[**$p_4_is1_acts[6, 7, 2, 0, 1]$]: Omission penalty for failing to provide either fluid or specific vasopressors during high lactate.
  - **12. `require_miss`[**$p_8_is1_acts[6, 2, 0, 1]$]: Omission penalty for failing to provide either crystalloid or specific vasopressors during low urine output.
- **D) Group Intensity Counts (Dim 13–15):**
  - **13. `group_count`[Ino]**: Active drugs within the *Inotrope* group ($[4, 5]$).
  - **14. `group_count`[fluid]**: Active drugs within the *Fluid* group ($[6, 7, 8]$).
  - **15. `group_count`[pressor]**: 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. 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 4, these phenotypes are characterized through their stationary distributions, temporal stability, and population-level evolution.

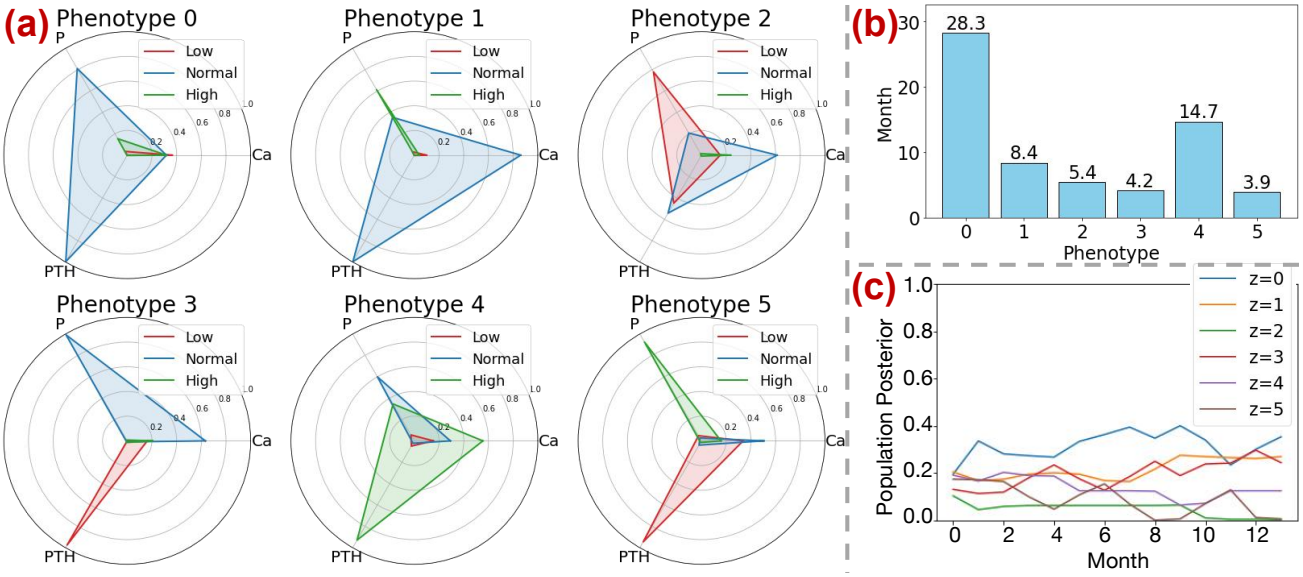


Figure 4. Comprehensive analysis of latent patient phenotypes in the CKD-MBD cohort. **(a) Top radar charts**: Emission probabilities of Calcium (Ca), Phosphorus (P), and PTH across three ranges (Low, Normal, High) for each phenotype. **(b) Middle bar chart**: γ-weighted average dwell time in months for each phenotype. **(c) Bottom line chart**: Temporal evolution of population posterior probabilities for each latent state z_i across the treatment horizon.

(a) Phenotype radar profiles. The radar charts reveal distinct biochemical signatures: phenotypes z_0 and z_1 exhibit a dominant "Normal" region (well-controlled mineral metabolism), while phenotypes z_4 and z_5 expand along the "High" axes for Phosphorus 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 state. Phenotype z_0 maintains an average dwell time of 28.3 months, designating it as a quasi-stable maintenance state. High-risk states such as z_5 have a dwell time of only 3.9 months; this volatility suggests they are transient and trigger more aggressive clinical interventions intended to push patients toward more stable phenotypes.

(c) Population posterior evolution. As the observation period increases, posterior probabilities of stable states (z_0 and z_1) trend upward, while high-risk states gradually decline. This validates the ability of the CTMC layer to capture the macro-level cohort transition from decompensated phases toward more compensated physiological states.

F. Detailed Sensitivity Analysis for Reward Specifications

F.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_{clin} as a *normative proxy* used to define a reference action-value function Q_{clin}. Taking our clinical applications as the primary evaluative benchmark, we explicitly evaluate the robustness of SBRD 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:

1. **Outcome-based terminal baseline.** We implement an episodic terminal signal (with discounting over decision epochs):

$$r_{\text{clin}}^{\text{term}}(z, a) = 0,$$
$$R_T = \mathbb{I}(\text{favorable outcome}) - \alpha\, \mathbb{I}(\text{adverse outcome}),$$

   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:

$$r_{\text{clin}}^{\text{stab}}(z, a) = -\big(\lambda_1 \,\text{shock}(z) + \lambda_2 \,\text{resp_fail}(z) + \lambda_3 \,\text{renal}(z)\big),$$

   where the right-hand-side terms 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 $Q_{\text{clin}}^{\text{CVaR}}$ or moment-penalized variants $r_{\text{clin}}(z, a) - \rho\, \text{Var}[\Delta\text{Score} \mid z, a]$.

Importantly, all choices yield the same modeling structure: the reference Q_{clin} changes, while the inferred switching constraint regimes explain residual systematic deviations.

Practical computation of Q_{clin} under continuous-time dynamics. Given action-held intervals $[t_m, t_{m+1})$, the CTMC defines a transition kernel $P^{(a_m)}(\Delta t_m) = \exp\big(Q^{(a_m)} \Delta t_m\big)$, which enables value iteration on the latent state space (or Monte Carlo rollouts) to compute $Q_{\text{clin}}(b, a)$ for each belief b. When the emission model is complex, we compute Q_{clin} on a discretized belief representation (e.g., via clustering beliefs) or via rollouts initialized from the posterior belief state b.

F.2. Performance under different reward specifications

This section provides a comprehensive performance breakdown of SBRD under the three reward families. Across all specifications, models with switching regimes ($K_u \in \{3, 5\}$) consistently outperform the stationary baseline ($K_u = 1$), substantiating the robustness of the framework (Table 3).

G. Detailed Comparative Benchmarking of System Dynamics Models

This section benchmarks the CTMC+ODE dynamics model against ODE-RNN (Habiba & Pearlmutter, 2020) and NCDSSM (Ansari et al., 2023) on a hold-out prediction task. In Table 4, we evaluate model fidelity across diverse temporal scales: 3-month and 6-month horizons for the chronic CKD-MBD cohort, and 6-hour and 12-hour horizons for the acute MIMIC-IV sequences. This multi-horizon assessment shows that the CTMC+ODE dynamics layer is competitive with or improves on neural temporal baselines while additionally recovering interpretable latent phenotypes used by the regime layer.

H. Counterfactual Policy Optimization and SCM-based Sampling

To ground sequential decision-making in realistic counterfactuals, we formalize the process as a Structural Causal Model (SCM). This allows us to evaluate alternative policy meta-rules while preserving the system's underlying exogenous characteristics, ensuring that estimated welfare gains are not artifacts of stochastic noise.

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	$\mathbf{0.98 \pm 0.00}$

H.1. Counterfactual reward function

We define the objective for counterfactual optimization using a cumulative λ-based reward. At each discrete observation epoch t_m^i, a meta-rule prescribes an action $a_{t_m}^i = f_\theta(o_m^i, z(t_m^i))$, where f_θ is a differentiable function. The instantaneous reward r_τ at time τ is defined as a function of the deterministic target-alignment probabilities $\lambda_\ell(\tau)$:

$$r_\tau \;=\; \sum_{\ell \in L} w_\ell \, \lambda_\ell(\tau) \;-\; w_{\text{risk}} \prod_{j \in J} p_j^{\text{hi}}(\tau, z(\tau)), \tag{7}$$

where $w_\ell > 0$ weights desirability of remaining target-aligned, $w_{\text{risk}} > 0$ penalizes synergistic risks, and $p_j^{\text{hi}}(\tau, z) = (1 - \phi_{z,j})(1 - \lambda_j(\tau))$. The cumulative discounted reward is $R(\theta) = \sum_{\tau=1}^{T} \gamma^\tau r(\lambda_\tau)$, with $\gamma = \exp(-\rho\Delta)$ when discretized at step Δ to match continuous-time discounting.

H.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 realizations of Gumbel noise that drove CTMC jumps, and ε^{em} denotes the Gumbel noise associated with categorical emissions o_m^i.

Step 2: Action (meta-rule intervention). We replace the observed sequence with actions prescribed by the meta-rule: $\tilde{a}_{t_m}^i = f_\theta(o_m^i, z(t_m^i))$. To ensure unit-level consistency, we maintain the latent Poisson arrival times s_k and the Gumbel realizations ε^{tr} as fixed exogenous traits of the individual unit.

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

1. **Latent state $z(t)$.** Transitions at uniformization jump times s_k use the identical Gumbel realizations $g_{k,j} \in \varepsilon^{\text{tr}}$ from

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

abduction:

$$z_{k+1} = \arg\max_j \{\log \tilde{P}_{z_k j}(\tilde{a}_{s_k}) + g_{k,j}\}, \quad \tilde{P}(\tilde{a}_{s_k}) = I + \frac{1}{\Omega} Q(\tilde{a}_{s_k}). \tag{8}$$

2. **Outcome dynamics λ_τ.** Between epochs, λ_ℓ evolves deterministically:

$$\lambda_\ell(\tau+1) = \pi_{1,\ell}^{(z_\tau)} + (\lambda_\ell(\tau) - \pi_{1,\ell}^{(z_\tau)}) e^{-\rho_\ell^{(z_\tau)}}. \tag{9}$$

3. **Emissions at observations.** Counterfactual signals $\tilde{o}_{m,\ell}^i$ are reconstructed using the fixed emission Gumbels ε^{em} and the new latent trajectory.

Optimization. We average over S posterior samples to compute $\hat{R}(\theta) = S^{-1} \sum_{s=1}^{S} R^{(s)}(\theta)$, and update θ by gradient ascent. Gradients backpropagate through the deterministic λ recursion and the Gumbel-softmax relaxation of z. This procedure ensures that counterfactual audits start from the unit's actual inferred condition, quantifying the welfare recovery achievable through constraint relaxation.

I. Detailed Counterfactual Auditing Prompts and Case Study

We use GPT-4o as the automated auditor, configured with a temperature of 0.1 and a fixed seed of 42 to ensure consistency and reproducibility across longitudinal clinical trajectories. Below, Section I.1 and Section I.2 list the prompt templates used for the MIMIC-IV and CKD-MBD cohorts, followed by a representative CKD-MBD case study (Section I.3).

I.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 **Counterfactual Optimal Strategy** (a^*) derived from the SBRD framework to validate its effectiveness.

- **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).

SBRD 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).
- **Structured GAP**: Systematic deviations where the observed action a diverges from the optimal a^* due to the penalty costs imposed by the active regime u.

Input Format

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

Your Analytical Task

1. **Rationality Analysis**: Determine if a^* would effectively stabilize vital signs (e.g., MAP) and clear lactate more than a. Explain if the model's suggested intervention is clinically sound for sepsis management.
2. **GAP Interpretation**: Identify unobserved constraints (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. Verify if shifts in u align with the timing of behavioral "GAPs" observed in the actual care.

Output (Strict JSON Report)

```
{
  "clinical_rationality": "Assessment of why a* is superior or sound.",
  "gap_interpretation": "Explanation of hidden clinician constraints via u.",
  "fidelity_validation": "Evidence linking z, u, and clinician behavior
  shifts."
}
```

I.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 **Counterfactual Optimal Strategy** (a^*). Validate if the SBRD framework effectively captures clinical logic and hidden constraints.
- **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).

SBRD 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).
- **Structured GAP**: Systematic deviations where the observed action a diverges from the optimal a^* due to penalty costs imposed by the active regime u.

Input Format

- `patient_biomarkers` = {MONTHLY_CA_P_PTH_DATA}
- `observed_actions_a` = {ACTUAL_MARKER_TRAJECTORY}
- `optimal_actions_a_star` = {MODEL_COUNTERFACTUAL_STRATEGY_A_STAR}

- `inferred_regimes_u` = {TIME_SERIES_OF_U}
- `latent_trajectories_z` = {TIME_SERIES_OF_Z}

Your Analytical Task

1. **Rationality Analysis**: Determine if a^* would effectively stabilize biochemical markers more than a. Explain why the model's suggested dosage is clinically sound.
2. **GAP Interpretation**: Identify unobserved constraints 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. Verify if shifts in u align with the timing of behavioral "GAPs" in the actual care.

Output (Strict JSON Report)

```
{
  "clinical_rationality": "Assessment of why a* is superior or sound.",
  "gap_interpretation": "Explanation of hidden clinician constraints via u.",
  "fidelity_validation": "Evidence linking z, u, and clinician behavior
  shifts."
}
```

I.3. Representative CKD-MBD Case Study

To illustrate SBRD as an auditing tool, Figure 5 shows a representative CKD-MBD trajectory: observed clinician actions a_m vs. the SBRD-synthesized counterfactual benchmark a_m^* over time, the inferred latent state z_m and constraint regime u_m, and the structured LLM audit report assembled from these components.

J. Detailed Continuous-Time Dynamics: Emissions and Filtering

This appendix gives the detailed observation model and belief filtering equations summarized in Section 2.2 of the main text. At each decision epoch t_m^i, the analyst observes a categorical covariate vector o_m^i. We assume a conditional emission model $p_\theta(o_m^i \mid z, \lambda)$ that factors across dimensions. For dimension ℓ with C categories, we distinguish an "on-target" category (Normal) and allocate residual mass across off-target categories using split parameters $\phi_{k\ell,c} \in (0,1)$ satisfying $\sum_{c\neq\text{Normal}} \phi_{k\ell,c} = 1$. Conditional on $z(t_m^i) = k$ and $\lambda_\ell(t_m^i)$,

$$\Pr(o_{m,\ell}^i = c \mid z = k, \lambda) = \begin{cases} \lambda_\ell, & c = \text{Normal}, \\ \phi_{k\ell,c}\big(1-\lambda_\ell\big), & c \neq \text{Normal}. \end{cases} \tag{10}$$

Beliefs are propagated forward under the action-modulated CTMC and updated by Bayes' rule. Given b_{m-1}^i, the predictive propagation is

$$\tilde{b}_m^i = b_{m-1}^i \, P(\Delta_{m-1}^i; a_{m-1}^i), \tag{11}$$

and after observing o_m^i,

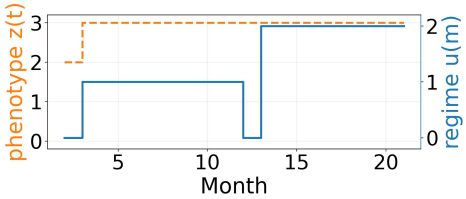
$$b_m^i(k) = \frac{\tilde{b}_m^i(k)\, p_\theta(o_m^i \mid z(t_m^i) = k, \lambda(t_m^i))}{\sum_{j=1}^{K_z} \tilde{b}_m^i(j)\, p_\theta(o_m^i \mid z(t_m^i) = j, \lambda(t_m^i))}. \tag{12}$$

The pair (b_m^i, λ_m^i) summarizes the payoff-relevant information used by the regime layer in the main text. Stage I parameters $\Theta_{\text{dyn}} = \{\theta^{(z)}, \alpha, \beta, \phi, \pi^{(0)}\}$ are learned by EM (smoothing E-step using forward–backward on the CTMC–ODE–emission model and parameter updates in the M-step), and benchmark values $Q_{\text{base}}(x_m, a)$ are computed by intervening on the current action and propagating $(z(t), \lambda(t))$ forward to compute the discounted reward $\int_0^\infty e^{-\rho t} r(\lambda(t))\, dt$.

(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 5. Automated decision auditing and structural counterfactual analysis for a CKD-MBD trajectory. **(a) Observed vs. counterfactual actions.** Observed clinician actions (a_m) and SBRD-synthesized benchmark actions (a_m^*), alongside longitudinal biomarkers (Ca, P, PTH). Treatment markers: M0–M1 are calcium carbonate, M4–M6 are sevelamer, M7 is 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** generated from the structured outputs in (a)–(b).

K. Trajectory-Level Regime Diagnostics

Figure 6 shows representative test-set trajectories of latent state $z(t)$ and inferred regime path u_m from the MIMIC-IV cohort. The examples illustrate three qualitative patterns: regime switches that co-occur with inferred system-state changes (a), regime switches that occur under stable latent state (b), and stable regimes despite frequent short-run state volatility (c). These coupled and decoupled patterns visualize why SBRD separates the decision-clock regime layer from the continuous-time system-dynamics layer rather than collapsing them into a single discrete process.

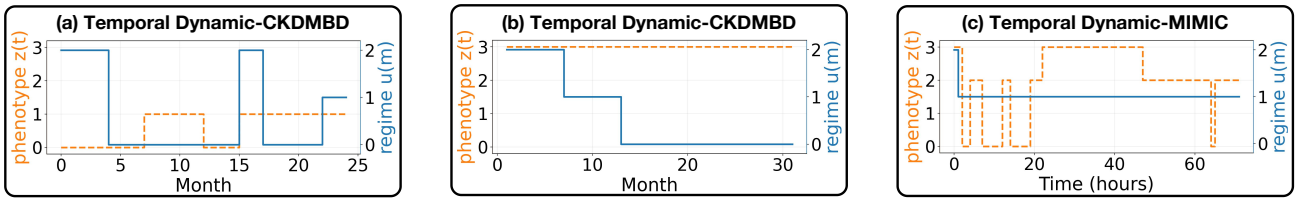


Figure 6. Representative test-set trajectories of latent state $z(t)$ and regime path u_m (K_z=4, K_u=3). **(a)** A regime transition co-occurs with a $z(t)$ shift. **(b)** A regime transition occurs under stable $z(t)$. **(c)** The regime remains stable despite frequent $z(t)$ shifts.

L. Sensitivity to the Number of Regimes

Figure 7 reports held-out predictive performance as a function of $K_u \in \{1, \ldots, 6\}$. SBRD with K_u>1 uniformly improves over the stationary baseline (K_u=1) on both the CKD-MBD and MIMIC-IV cohorts in log-likelihood and top-k accuracy, and the gain plateaus at small K_u, consistent with parsimonious regime structure rather than overfitting. Self-transition probabilities remain above 0.85 across the sweep, indicating that the recovered regimes describe stable strategic intent rather than per-step action noise.

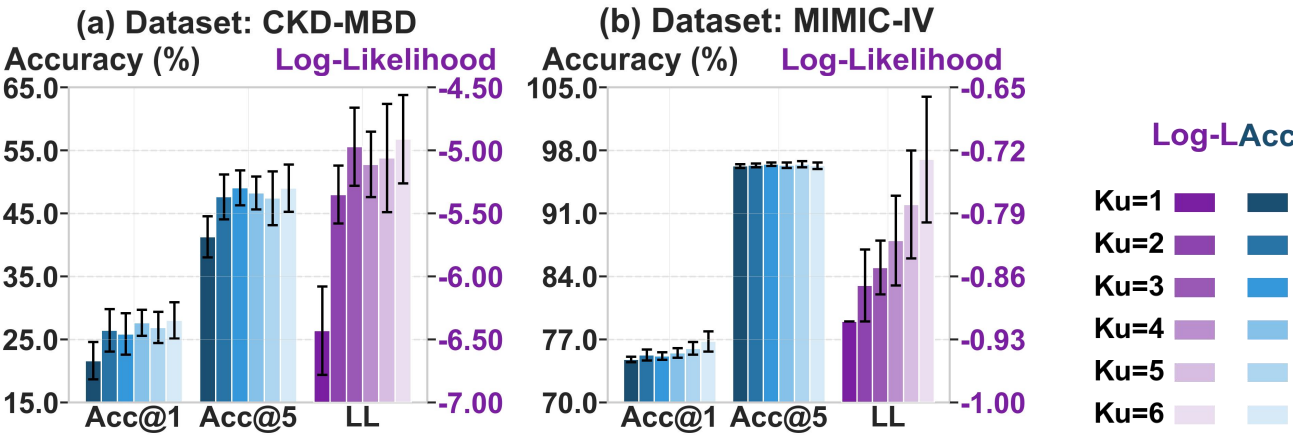


Figure 7. Predictive performance across $K_u \in \{1, \ldots, 6\}$ on the clinical cohorts. Accuracy (blue bars) and log-likelihood (purple bars) consistently exceed the stationary baseline (K_u=1), supporting the necessity of latent switching logic for modeling nonstationary clinical decision trajectories. Bar shading goes from dark (K_u=1) to light (K_u=6); the legend is reproduced on the right.

M. CKD-MBD Regime Profiles

For completeness, we report the recovered constraint regimes on the chronic CKD-MBD cohort, which complement the MIMIC-IV regimes shown in the main text. With K_u=3 and the 8-dimensional CKD-MBD wedge map of Section D, the inferred regimes correspond to interpretable trade-offs between long-term risk prevention and immediate safety:

- R_0 **(Aggressive/Responsive):** large weights on f_6 (Treatment escalation) and f_4 (Uncontrolled labs proxy), consistent with a proactive stance on stabilization;
- R_1 **(Risk Mitigation):** concentrated on f_2 ($Ca \times P$ risk proxy), targeting long-term cardiovascular prevention;

- R_2 **(Safety/Simplification):** peaks at f_1 (Hypercalcemia $\times$ Ca-binder risk) and f_3 (Polypharmacy burden), prioritizing toxicity avoidance and regimen compliance.

The transition matrix Π is sticky ($\Pi_{uu} > 0.85$ for all u), mirroring the persistence pattern observed on MIMIC-IV.