SURVIVAL POLICY LEARNING AS PROBABILISTIC INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning optimal treatment policies that maximize survival probabilities from longitudinal observational data is a central problem in causal inference for medicine. Existing applications of reinforcement learning (RL) to longitudinal medical data typically adapt standard RL algorithms under a Markov decision process formulation and rely on heuristic rewards specified by clinicians or auxiliary models, thereby weakening the scientific validity and interpretability of the learned policies. In this study, we formulate causal survival policy learning within the framework of control as inference. This formulation yields a principled dense reward which connects reward design directly to survival quantities and induces a behavior-anchored, KL-regularized policy learning objective for offline observational data. We further use an autoregressive model to learn state representations that approximate the Markov property of the sequential decision process. We evaluate the proposed method using synthetic data and semi-synthetic data based on real-world chronic disease epidemiological study, tracking individual health trajectories over a 30-year horizon. Across experiments, the proposed survival-derived reward provides an effective alternative to sparse survival rewards for survival policy learning.

1 INTRODUCTION

Many high-stakes clinical questions are naturally formulated in terms of time-to-event outcomes, making survival analysis a core component of clinical evidence generation. Current clinical evidence is based on causal inference from randomized controlled trials or observational data. A popular method is survival analysis where the survival probability after intervention is estimated using statistical methods such as the Cox proportional hazards model (Cox, 1972). Recent advances in causal inference methods and data availability led to the use of causal survival analysis under sustained interventions (van der Laan & Rose, 2018; Hernan & Robins, 2024). Although several estimators have been proposed within the framework of causal inference (Bang & Robins, 2005; Laan & Gruber, 2012; Frauen et al., 2023; Shirakawa et al., 2024), these are estimators of the causal effects under fixed treatment regimes. To optimize the patient's outcome, we need estimators of the optimal treatment regimes.

Learning optimal treatment policies from longitudinal observational data is a central task in precision medicine. In chronic disease management, long-term sequential decision-making to control health status is essential for preventing future fatal events, such as cardiovascular disease, cancer, dementia, and all-cause mortality. Causal reinforcement learning (RL) provides a framework for applying RL methods to decision-making tasks under causal assumptions (Bareinboim et al., 2025). However, causal RL methods tailored specifically to survival analysis remain underdeveloped. Survival outcomes pose specific methodological challenges that distinguish clinical RL from standard benchmarks: events of interest are typically rare, delayed by years or even decades, and subject to right-censoring; confounders evolve over time, leading to time-varying treatment-confounder feedback; and causal positivity is often violated for parts of the action space, yielding biased value estimates if not handled explicitly. While off-policy evaluation methods such as importance sampling and doubly-robust estimators address some of these issues for fixed treatment regimes, principled extensions to optimal-policy learning under longitudinal survival dynamics remain scarce. Bridging this gap requires reward formulations and policy-improvement objectives explicitly grounded in survival-analytic identities rather than imported from generic RL settings, so that the learned policy is compatible with existing clinical evidence and, by extension, current clinical practice.

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

Reward shaping is a critical challenge in applying RL to medical problems. Because medical evidence is grounded in survival outcomes, the reward should align with them: optimizing cumulative reward should be compatible with optimizing mean counterfactual survival. Yet rewards in clinical RL studies are typically hand-crafted from clinical expertise or specified by large language models, and therefore lack a solid scientific grounding. The per-step survival indicator is consistent with the endpoint but is sparse and delayed, making credit assignment hard. We therefore seek dense rewards with scientifically grounded causal signals whose optimization is compatible with survival optimization.

Given this context, we use the framework of control as inference to formulate survival policy learning as inference in a probabilistic model, with the survival indicator serving as the optimality variable. Under this framework, we treat the behavior policy as the prior. This yields a dense reward given by the log survival increment, whose expected cumulative sum provides a lower bound on the log survival probability. Our contributions include:

- We unify causal survival analysis with behavior-policy-anchored offline RL within a control-as-inference framework. Based on a structural variational inference formulation, the variational policy optimizes mean log-counterfactual survival while remaining anchored to the behavior policy through the Kullback–Leibler (KL) divergence, which helps protect against positivity violations.
- Our method provides a principled approach to dense reward shaping from sparse survival signals. The reward is defined as the log survival increment at each time point; optimizing its expected cumulative sum can be interpreted as optimizing mean log-counterfactual survival, which lower-bounds the log expected survival.
- As a dense alternative for comparison, we derive a simple linear dense reward that is the negative at-risk intensity. We prove that the optimization of the mean cumulative sum of this reward is equivalent to maximizing the marginal survival probability at the end of the time horizon. Although the derivation is straightforward, to the best of our knowledge, this is the first study that clearly states this equivalence in survival RL.
- We demonstrate that the proposed survival-derived reward provides an effective alternative to sparse survival-indicator rewards, improving learning in several settings while maintaining offline credibility through behavior-policy anchoring.

2 Related Work

Clinical RL A central limitation of clinical RL is that many clinically important objectives are often defined by long-horizon survival or other terminal events, so reward is both sparse and delayed, which makes temporal credit assignment difficult and can substantially weaken offline RL because value must be estimated from rare end states rather than frequently observed intermediate consequences. Prior clinical RL studies have addressed treatment optimization, dosing, monitoring, and adaptive intervention design, with reward formulations ranging from sparse terminal outcomes to proxy physiologic signals, multi-objective utility functions, inverse reward learning, and, more recently, preference-based or language-model-derived supervision (Komorowski et al., 2018; Peine et al., 2021; Kondrup et al., 2023; Kalimouttou et al., 2025; Yu et al., 2019; Tan et al., 2026). However, Roggeveen et al. (Roggeveen et al., 2024) showed that reward weighting materially changes learned policies, underscoring that reward specification is itself a major source of instability. In contrast, we start from the survival objective itself and use its causal factorization to decompose long-horizon survival into natural dense components aligned with clinically meaningful intermediate subgoals. This yields a principled dense reward that improves temporal credit assignment without replacing the clinical endpoint with ad hoc proxy outcomes.

Control as Inference and Causal RL We formulate survival policy learning as inference, drawing on the variational inverse control of events (VICE) (Fu et al., 2018) and the broader control-as-inference (CaI) (Levine, 2018). VICE avoids manually specified dense rewards by using sparse binary events as optimality variables. Our setting differs because the future contains at most one absorbing terminal event, such as death or disease onset, governed by longitudinal causal survival dynamics. While causal RL (Bareinboim et al., 2025) has been formulated using structural causal models (SCM), many of its illustrative examples rely on deterministic or highly structured dynamics,

such as navigation or maze-like environments. In contrast, clinical survival policy learning involves rare, delayed, highly stochastic, and confounded terminal outcomes. This makes control as inference particularly natural: conditioning on future survival induces a principled dense objective through cumulative log-survival probabilities.

Dynamic Treatment Regimes and Positivity-Aware Interventions Optimal dynamic treatment regimes (DTRs) have a long history in the statistics literature. Many existing approaches are value-based, using sequential regression to estimate treatment-specific values and selecting treatments that maximize the estimated value at each decision point. However, the resulting policy is not necessarily constrained by the behavior policy, which can exacerbate positivity concerns in longitudinal settings. Modified treatment policies (Díaz et al., 2023) and shift interventions (Kennedy, 2019) address such concerns by defining counterfactual policies as modifications of the behavior policy. Applications of this idea to policy learning, however, have primarily focused on single-decision settings (Zhao et al., 2024).

Recent Survival-based RL Recent work has begun to incorporate ideas from survival analysis into reinforcement learning. For example, Tiofack et al. formulate goal-conditioned RL as a survival learning problem (SVL) for estimating time-to-goal distributions (Tiofack et al., 2026); Wang et al. use cumulative hazard functions as a penalty to stabilize policy optimization in modulated hazard-aware policy optimization (MHPO) (Wang et al., 2026); and related frameworks, such as Evidence-Anchored Reward Attribution (EARA) and Conditional Reward Modeling (CRM), use survival probabilities for process-level credit assignment in language-model training (Zhang et al., 2025). These studies highlight the utility of survival-based quantities for reward design and credit assignment. In contrast, our work directly connects causal survival analysis with control as inference, yielding a KL-regularized offline policy learning algorithm whose reward is derived from survival quantities rather than specified heuristically.

3 METHOD

In this section, we formulate the offline causal survival learning problem following the causal inference and control as inference frameworks, and then propose an offline RL algorithm. In Section 3.1, we define the random variable and data generating process, then introduce the optimization objective which is the sum of the logarithmic survival quantity. In Section 3.2, we introduce the framework of control-as-inference and derive the optimization objective and the log-survival rate type reward function. In Section 3.3, we connect our reward function with the counterfactual survival function, and prove that our objective optimizes a lower bound on survival. In Section 3.5, we introduce an alternative linear dense reward and show the expected cumulative sum of this reward identifies the counterfactual survival up to a constant. In Section 3.4, we introduce an additional step to learn the state space representation to approximate the Markovian property. In Section 3.6, we describe the proposed algorithm and the agent, a behavior-anchored soft actor-critic (BA-SAC), and prove that the algorithm optimizes our objective.

3.1 PROBLEM SETTING

Random variables Let L_t be the time-dependent covariate, A_t be the time-dependent treatment, Y_t be the failure event indicator for $t = 1, \ldots, \tau$. The event process is absorbing, so $Y_t = 1$ implies $Y_{t+1} = 1$. Since the observed health trajectory is in general not the Markov decision process (MDP), i.e. future health status would depend on the entire health trajectory, we assume the partially observed Markov decision process (POMDP) with states S_t as shown in Figure 1. The state-action trajectory is denoted by $X = (S_{1:\tau}, A_{1:\tau})$ and the observed data is denoted by $O = (L_{1:\tau}, A_{1:\tau}, Y_{1:\tau})$.

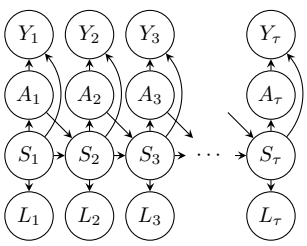


Figure 1: Graphical Model of data generating process.

Data Generating Distribution We consider the decomposition of the joint distributions of states S_t, actions A_t, and outcomes Y_t into

conditional distributions μ, π_β, and λ, according to the graphical model (Figure 1):

$$p(S_{1:\tau}, A_{1:\tau}, Y_{1:\tau}) = \mu(S_1) \prod_{t=1}^{\tau} \mu(S_{t+1} \mid S_t, A_t) \pi_\beta(A_t \mid S_t) \lambda(Y_t \mid S_t, A_t). \tag{1}$$

Here $\mu(S_{t+1} \mid S_t, A_t)$, $\pi_\beta(A_t \mid S_t)$, and $\lambda(Y_t \mid S_t, A_t)$ are the abbreviations of $\mu(S_{t+1} \mid S_t, A_t, Y_t)$, $\pi_\beta(A_t \mid S_t, Y_t)$, and $\lambda(Y_t \mid S_t, A_t, Y_{t-1})$, respectively. In practice, we consider these distributions conditional on $Y_t = 0$, hence omitting Y_t in the condition.

Counterfactual Distributions For any policy π, we define the counterfactual trajectory distribution

$$p^\pi(S_{1:\tau}, A_{1:\tau}, Y_{1:\tau}) = \mu(S_1) \prod_{t=1}^{\tau} \mu(S_{t+1} \mid S_t, A_t) \pi(A_t \mid S_t) \lambda(Y_t \mid S_t, A_t). \tag{2}$$

We also define the at-risk state-action trajectory distribution by enforcing $Y_t = 0$:

$$q^\pi(S_{1:\tau}, A_{1:\tau}) = \mu(S_1) \prod_{t=1}^{\tau} \mu(S_{t+1} \mid S_t, A_t, Y_t = 0) \pi(A_t \mid S_t). \tag{3}$$

The two distributions are linked as

$$p^\pi(X, Y_{1:\tau=0}) = q^\pi(X) \prod_{t=1}^{\tau} \big\{1 - \lambda(Y_t = 1 \mid S_t, A_t)\big\}. \tag{4}$$

Policy Learning We say that a policy π is *feasible* if $\pi_\beta(a_t \mid s_t) = 0$ implies $\pi(a_t \mid s_t) = 0$. Let Π° be the set of all feasible policies. We define the counterfactual survival under the policy π as

$$\mathcal{S}^\pi(\tau) := \mathbb{E}_{p^\pi}[1 - Y_\tau]. \tag{5}$$

We aim to learn the optimal treatment policy π that maximizes the counterfactual survival within the class of feasible policies:

$$\pi^\star := \arg\max_{\pi \in \Pi^\circ} \mathcal{S}^\pi(\tau). \tag{6}$$

By integrating out the state-action trajectory X in Equation 4, we have the following formula:

Lemma 1. *The counterfactual survival probability is the expectation of the cumulative survival probability under the at-risk trajectory distribution:*

$$\mathcal{S}^\pi(\tau) = \mathbb{E}_{S_t, A_t \sim q^\pi}\Big[\prod_{t=1}^{\tau} \big(1 - \lambda(t \mid S_t, A_t)\big)\Big]. \tag{7}$$

3.2 Survival Policy Learning as Probabilistic Inference

We formulate this survival policy learning problem using the framework of control-as-inference (Levine, 2018) with the optimality variable $\mathcal{O}_t = 1 - Y_t$. Then the reward function $\log p(\mathcal{O}_t = 1 \mid S_t, A_t)$ is given by

$$r^{\log}(S_t, A_t) := \log(1 - \lambda(t \mid S_t, A_t)). \tag{8}$$

Proposition 1 (Survival Control as Inference)**.** *Suppose the behavior policy π_β is used as the policy prior. Then the variational policy that minimizes the KL divergence to the survival-conditioned trajectory posterior maximizes the following objective:*

$$\mathcal{J}(\pi) = \mathbb{E}_{q^\pi}\Big[\sum_{t=1}^{\tau} r^{\log}(S_t, A_t)\Big] - \mathbb{E}_{q^\pi}\Big[\sum_{t=1}^{\tau} D_{\mathrm{KL}}\big(\pi(\cdot \mid S_t) \| \pi_\beta(\cdot \mid S_t)\big)\Big]. \tag{9}$$

Furthermore, the corresponding optimal policy is given by

$$\pi^{\star}(A_t \mid S_t) \propto \pi_\beta(A_t \mid S_t) \exp Q^{\pi^{\star}}(S_t, A_t), \tag{10}$$

where the Q function is defined as

$$Q^\pi(S_t, A_t) = \mathbb{E}_{S_{t+1:\tau}, A_{t+1:\tau} \sim q^\pi(\cdot \mid S_t, A_t)}\Big[\sum_{t'=t}^{\tau} r^{\log}(S_{t'}, A_{t'}) - \sum_{t'=t}^{\tau} \log \frac{\pi(A_{t'} \mid S_{t'})}{\pi_\beta(A_{t'} \mid S_{t'})}\Big]. \tag{11}$$

As mentioned in the previous section, the first term in Equation 9 maximizes a lower bound on the log marginal survival probability. The second term regularizes the policy toward the behavior policy, encouraging adherence to the support of the observed treatment distribution. The full derivation of the objective 9 can be found in Appendix A.

Tempered Objective In practice, we optimize the tempered objective with a coefficient $\alpha \geq 0$:

$$\mathcal{J}^{\alpha}(\pi) = \mathbb{E}_{q^{\pi}}\Big[\sum_{t=1}^{\tau} r^{\log}(S_t, A_t)\Big] - \alpha \cdot \mathbb{E}_{q^{\pi}}\Big[\sum_{t=1}^{\tau} D_{\mathrm{KL}}\big(\pi(\cdot \mid S_t) \| \pi_{\beta}(\cdot \mid S_t)\big)\Big]. \tag{12}$$

This gives a tempered KL-regularized control-as-inference objective. The corresponding optimal policy is given by

$$\pi_{\alpha}^{\star}(A_t \mid S_t) \propto \pi_{\beta}(A_t \mid S_t) \exp\big[Q^{\pi^{\star}}(S_t, A_t)/\alpha\big]. \tag{13}$$

One can see that $\alpha \to 0$ yields the standard optimal policy that maximizes the average mean cumulative reward.

Relation to Shift Intervention Our optimal policy equation 13 is related to interventions designed to mitigate positivity violations in causal effect estimation. Approaches to mitigate positivity violations in statistical causal inference include shift interventions (Kennedy, 2019) or modified interventions (Díaz et al., 2023) of the behavior policy. These approaches define counterfactual policies that modify the behavior policy to preserve support. Our CaI-derived optimal policy equation 13 can be regarded as a modified treatment policy since it is an exponential tilt of the behavior policy. More specifically, Kennedy (2019) defines a shift intervention

$$\operatorname{logit} \pi^{*}(A_t = 1 \mid S_t) = \operatorname{logit} \pi_{\beta}(A_t = 1 \mid S_t) + \delta \tag{14}$$

for some scalar δ in the binary treatment setting. In our case, let $\delta(S_t) = (Q(1, S_t) - Q(0, S_t))/\alpha$, then we have

$$\operatorname{logit} \pi_{\alpha}^{\star}(A_t = 1 \mid S_t) = \operatorname{logit} \pi_{\beta}(A_t = 1 \mid S_t) + \delta(S_t) \tag{15}$$

Therefore, one can regard our policy as an individualized incremental propensity score intervention.

3.3 Causal Identifiability

We describe the causal structure of the problem and the identification of our causal target object, the mean log-counterfactual survival, following the roadmap of causal inference Petersen & van der Laan (2014). Suppose the structural causal model underlying the probabilistic model depicted in Figure 1 is defined by introducing external noises $U_t^{\bullet}$ and deterministic production functions $f_t^{\bullet}$:

$$\begin{aligned} S_t &= f_t^{s}(S_{t-1}, A_{t-1}, U_t^{s}), \\ A_t &= f_t^{a}(S_t, U_t^{a}), \text{ and} \\ Y_t &= f_t^{y}(S_t, A_t, U_t^{y}). \end{aligned} \tag{16}$$

For counterfactual policy g which is a function of S_t and U_t^a and produces actions, sequentially define counterfactual variables S_t^g, A_t^g, and Y_t^g. Let π be the conditional distribution of actions given states induced by g. Let us introduce the following assumptions to identify the counterfactual survival from observed data.

Assumption 1 (Sequential Randomization)**.** $Y_{t:\tau}^{g} \perp A_t \mid S_t$ *for all* g *and* t.

Assumption 2 (Positivity)**.** $\pi(\cdot \mid S_t) \ll \pi_{\beta}(\cdot \mid S_t)$ *a.s.*

The survival probability based on the counterfactual outcome $\mathcal{S}^{g}(\tau) = \mathbb{E}[1 - Y_{\tau}^{g}]$ is identified by the observed variables through the following g-formula:

Proposition 2 (Robins 1986)**.** *Under Assumptions 1 and 2, the counterfactual survival probability defined in 5 is the mean of the counterfactual survival outcome:*

$$\mathcal{S}^{g}(\tau) = \mathcal{S}^{\pi}(\tau). \tag{17}$$

Combining Lemma 1, Proposition 2, and Jensen's inequality, we have the following bound:

Corollary 1. *Under Assumptions 1 and 2, the log of counterfactual survival is bounded from below by the mean of the cumulative reward under the counterfactual policy:*

$$\mathbb{E}_{S_t, A_t \sim q^{\pi}}\Big[\sum_{t=1}^{\tau} r^{\log}(S_t, A_t)\Big] \leq \log \mathcal{S}^{\pi}(\tau). \tag{18}$$

Remark. By Corollary 1, the expected cumulative logarithmic reward maximizes a lower bound on the log counterfactual survival. Since $\log(\cdot)$ is concave, this objective has a risk-sensitive interpretation: trajectories with a low survival probability receive disproportionately large penalties. Thus, compared with directly maximizing marginal survival, expected log-survival favors policies that avoid rare but severe low-survival trajectories. Detailed exposition can be found in Appendix C.

3.4 State Space Representation

Since health trajectories generally do not satisfy the Markov property with respect to current covariates alone, we introduce a learned state-space representation. A sufficient but impractical Markov state is the full observed history $H_t = (S_{1:t}, A_{1:t-1})$. Under sequential ignorability with respect to H_t, our goal is to learn a compact causal state $S_t = f_\theta(H_t)$ that preserves the information in the observed history relevant to treatment assignment, censoring, and survival. We parameterize f_θ using a Transformer encoder and assume that S_t is sufficient for treatment assignment and potential outcomes. Under this sufficiency assumption, policy evaluation and optimization can be performed in the learned state space. We also provide a risk-sensitive interpretation of the proposed objective with a supporting inequality.

3.5 Alternative Linear Dense Reward

We can derive another dense reward which is a negative intensity. The optimization of the expected cumulative sum of this reward is equivalent to the optimization of the counterfactual survival. Let $Y_t^* = \mathbb{1}\{Y_{t-1} = 0\}Y_t$ and $\lambda^*(t \mid S_t, A_t) = \mathbb{1}\{Y_{t-1} = 0\}\lambda(t \mid S_t, A_t)$ be the at-risk event indicator and the at-risk intensity. Formally, the average sum of these rewards identifies the survival:

Proposition 3. *For any feasible policy π, we have*

$$\mathbb{E}_{S_t, A_t, Y_t \sim p^\pi}\Big[\sum_{t=1}^{\tau} -\lambda^*(t \mid S_t, A_t)\Big] = \mathcal{S}^\pi(\tau) - 1. \tag{19}$$

Remark. Note that the sparse reward $-\mathbb{1}\{Y_{t-1} = 0\}Y_t$ also satisfies

$$\mathbb{E}_{S_t, A_t, Y_t \sim p^\pi}\Big[\sum_{t=1}^{\tau} -Y_t^*\Big] = \mathcal{S}^\pi(\tau) - 1. \tag{20}$$

We can interpret the linear dense reward $r(S_t, A_t) = -\lambda^*(t \mid S_t, A_t)$ as the "risk score" of the short-term failure given the state and action. In this sense, Eq 20 provides a principled connection between the survival objective with heuristics used in clinical RL applications that use clinical risk scores, such as sequential organ failure assessment (SOFA) score in critical care, as rewards.

Remark When we consider the continuous outcome process $(Y'_t)_t$ with a continuous intensity $\lambda(t \mid S_t, A_t)$, the linear reward can be regarded as the control-as-inference reward when we approximate the probability of no jump during the interval $[t, t+1)$ with the Poisson process as $P(Y_t = 0 \mid S_t, A_t) = \exp\{-\lambda(t \mid S_t, A_t)\}$. The control-as-inference reward is $r^{\text{lin}}(S_t, A_t) = \log P(Y_t = 0 \mid S_t, A_t) = -\lambda(t \mid S_t, A_t)$. The log reward 8 corresponds to the approximation of this probability with the Bernoulli process as $P(Y_t = 0 \mid S_t, A_t) = 1 - \lambda(t \mid S_t, A_t)$.

3.6 Algorithm

Our algorithm consists of two stages: (1) learning a state representation and the associated reward function from observational data, and (2) learning the optimal policy via offline reinforcement learning on the learned state space.

3.6.1 State Space Learning

Since the observed health trajectory does not satisfy the Markov property, we learn a latent state representation S_t that summarizes the relevant history for predicting future observations and outcomes using a Transformer encoder and the flow-matching prediction head for continuous time-dependent variables L_t. The learned reward is denoted by $r_\eta(S_t, A_t)$. Details can be found in Appendix E.

Algorithm 1 Behavior-Anchored Soft Actor-Critic

Require: Offline data $\mathcal{D}$, horizon τ, discount factor γ, temperature α, target update rate τ

1: Learn state representations $S_t = f_\phi(H_t)$ from observed histories $H_t = (W, L_{1:t}, A_{1:t-1})$, hazard model $\hat{\lambda}(t \mid S_t, A_t)$, and behavior policy model $\hat{\pi}_b(A_t|S_t)$
2: Define the survival-derived reward $r_t(S, A) \leftarrow \log\{1 - \hat{\lambda}_t(S, A)\}$
3: Initialize actor $\pi_\theta(A|S)$, critics Q_{ψ_1}, Q_{ψ_2}, and target critics $Q_{\bar{\psi}_1}, Q_{\bar{\psi}_2}$
4: **for** each gradient step **do**
5: Sample a minibatch $(S_t, A_t, D_t, S_{t+1}) \sim \mathcal{D}$
6: Compute the soft survival Bellman target $\tilde{Y}_t$ by equation 22
7: Update critics by minimizing equation 21
8: Update actor by minimizing equation 23
9: Update target critics: $\bar{\psi}_j \leftarrow \tau\psi_j + (1 - \tau)\bar{\psi}_j$ for $j = 1, 2$
10: **end for**
11: **return** learned policy π_θ

3.6.2 Behavior-Anchored Soft Actor-Critic

We now describe a practical offline policy optimization algorithm induced by the control-as-inference formulation. Let $\mathcal{D}$ denote an offline dataset collected with the behavior policy. Let Q_{ψ_1}, Q_{ψ_2} be two critic networks, $Q_{\bar{\psi}_1}, Q_{\bar{\psi}_2}$ be their target networks, and $Q_\psi = (\sum_j Q_{\psi_j})/2$ be their average. The critic is trained by minimizing the Bellman error

$$\mathcal{L}_Q(\psi_i) = \mathbb{E}_{(S_t, A_t, S_{t+1}) \sim \mathcal{D}}\Big[\big(Q_{\psi_i}(S_t, A_t) - \tilde{Y}_t\big)^2\Big], \tag{21}$$

where

$$\tilde{Y}_t = r_\eta(S_t, A_t) + \gamma(1 - D_t)\mathbb{E}_{A_{t+1} \sim \pi_\theta(\cdot|S_{t+1})}\Big[Q_{\bar{\psi}}(S_{t+1}, A_{t+1}) - \alpha \log \frac{\pi_\theta(A_{t+1} \mid S_{t+1})}{\pi_\beta(A_{t+1} \mid S_{t+1})}\Big]. \tag{22}$$

Here $D_t = \mathbb{1}\{t = \tau\}$ OR $\mathbb{1}\{Y_t = 1\}$ is the terminal indicator. The actor is trained to minimize

$$\mathcal{L}_\pi(\theta) = \mathbb{E}_{S_t \sim \mathcal{D},\, A_t^\theta \sim \pi_\theta(\cdot|S_t)}\big[-Q_\psi(S_t, A_t^\theta)\big] + \alpha\mathbb{E}_{S_t \sim \mathcal{D}}\Big[D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid S_t) \| \pi_\beta(\cdot \mid S_t)\big)\Big], \tag{23}$$

where π_β is the learned behavior policy in Stage 1. The proposed algorithm is shown in Algorithm 1.

Remark. The proposed algorithm differs from standard SAC in two ways. First, the reward is not manually specified, but is induced by the causal survival likelihood through $r_\eta(S, A) = \log\{1-\lambda_\eta(t \mid S, A)\}$. Second, the entropy term is replaced by a KL penalty to the behavior policy, reflecting the fact that offline causal policy learning must remain within the support of the observed data. This yields a principled survival-aware SAC+BC algorithm for offline longitudinal treatment policy learning.

Proposition 4. *Assume that the estimated hazard reward equation 8 is fixed, $\gamma = 1$, and that the behavior policy has support over the actions considered by π. Then the KL-regularized actor-critic updates in Algorithm 1 are stochastic approximations to coordinate ascent on the tempered survival objective equation 12. In particular, the critic update estimates the corresponding soft Bellman operator, and the actor update performs policy improvement for this KL-regularized objective.*

Remark. Thus, the algorithm does not optimize an ad hoc reinforcement learning objective. Rather, the critic estimates the soft value function associated with the survival-derived reward $\log(1 - \lambda_t)$, and the actor performs KL-regularized policy improvement relative to the behavior policy prior π_β. The resulting updates are therefore stochastic approximations to coordinate ascent on the tempered control-as-inference survival objective.

3.7 Dataset with Uninformative Right-Censoring

Under conditional independent right censoring, censoring does not alter the full-data causal estimand or the survival-derived reward. It only determines which transitions are observed. We therefore train the hazard model and Bellman updates using uncensored at-risk transitions; censored transitions are masked rather than treated as failure events. Extensions to conditionally informative censoring are discussed in Appendix D.

4 EXPERIMENTS

We evaluate the proposed algorithm along two axes: policy performance and positivity adherence. The central empirical question is whether survival-derived dense rewards improve learning from sparse terminal survival outcomes while keeping the learned policy within the support of the observational behavior policy. We therefore compare the proposed logarithmic reward $\log(1 - \lambda_t)$ with the sparse terminal reward $-Y_t^*$ and the linear intensity reward $-\lambda_t^*$ across DQN, CQL, and behavior-anchored SAC (BA-SAC).

We use two environments. The synthetic environment provides a controlled non-Markovian longitudinal survival setting in which reward design can be isolated. The semi-synthetic CIRCS (Yamagishi et al., 2019) environment uses real baseline health trajectories and a learned autoregressive simulator to evaluate candidate policies over a 30-year horizon. We report mean counterfactual survival and the KL divergence from the behavior policy.

4.1 BASELINES AND EVALUATION METRICS

We compared the proposed method with the deep Q-network (DQN) (Mnih et al., 2013) and the conservative Q-learning (CQL) (Kumar et al., 2020) as baseline. We also compared the proposed reward equation 8 with the original sparse survival reward $r_t(S_t, A_t) = -\mathbb{1}\{Y_{t-1} = 0\}Y_t$, and the alternative linear reward $r_t(S_t, A_t) = -\lambda(t \mid S_t, A_t)$. The performance of the agents and the reward are evaluated by the mean survival, and the mean KL-divergence between the learned policy with the behavior policy.

4.2 DATA

We developed two environments with synthetic data generating process (DGP) and semi-synthetic DGP based on a real-world health dataset.

4.2.1 SYNTHETIC DATA

We built a simulator with the following non-Markovian data generating process described in Appendix F.1 and use 10K samples to fit the proposed method.

4.2.2 SEMISYNTHETIC DATA

We built a simulator based on real-world population-based cardiovascular epidemiological data. We used data from the Circulatory Risk in Communities Study (CIRCS) (Yamagishi et al., 2019). We first fit an autoregressive Transformer with a flow-matching model for time-dependent covariates L_t. We used 13,465 samples from the simulator with initial states from the real-world data to train our models. The action is the reduction of the systolic blood pressure (SBP) categories defined by 10 mmHg strata. Details are described in Section F.2.

4.3 RESULTS

In the synthetic experiment (Table 1a), the survival-derived dense rewards consistently improved policy learning relative to the sparse terminal reward. Across DQN, CQL, and BA-SAC, the linear and logarithmic rewards achieved mean survival of approximately 0.76–0.79, compared with approximately 0.71–0.74 for the sparse reward. The difference between the linear and logarithmic rewards was small, suggesting that the primary benefit comes from providing a dense survival-based learning signal. Behavior anchoring also substantially reduced deviation from the observational policy: for example, with the logarithmic reward, increasing α from 0.001 to 0.05 reduced KL divergence from 0.609 to 0.169, while survival decreased from 0.789 to 0.762, illustrating the trade-off between policy improvement and adherence to the observed treatment distribution.

In the semi-synthetic CIRCS experiment (Table 1b), unconstrained policies often achieved high apparent survival together with very large KL divergence from the behavior policy. DQN policies, for example, had KL values above 8, indicating substantial distribution shift. In contrast, BA-SAC with moderate regularization remained much closer to the behavior policy. At $\alpha = 0.01$, survival

Table 1: Mean counterfactual survival on synthetic and semi-synthetic (CIRCS) data. Mean ± SD across 10 seeds. Reward naming: sparse $= -Y_t^*$, linear $= -\lambda^*(t \mid S_t, A_t)$, log $= \log(1-\lambda)$. For (b) CIRCS semi-synthetic data, cells with KL > 2.0 are grayed out.

		(a) Synthetic		(b) Semi-synthetic (CIRCS)	
Agent	Reward	Mean Survival	KL	Mean Survival	KL
Behavior (π_β)	–	0.702 ± 0.004	0.000	0.480 ± 0.004	0.000
Never-treat	–	0.790 ± 0.004	0.554 ± 0.002	0.447 ± 0.003	8.420 ± 0.026
Always-treat	–	0.588 ± 0.006	0.582 ± 0.002	0.316 ± 0.002	8.365 ± 0.028
Random (p=0.5)	–	0.708 ± 0.006	0.073 ± 0.000	0.688 ± 0.004	7.852 ± 0.009
DQN	sparse	0.730 ± 0.025	0.565 ± 0.080	0.546 ± 0.092	8.718 ± 0.263
	linear	0.787 ± 0.007	0.512 ± 0.035	0.539 ± 0.087	8.549 ± 0.339
	log	0.786 ± 0.006	0.509 ± 0.036	0.524 ± 0.098	8.321 ± 0.371
CQL (α=0.01)	sparse	0.712 ± 0.006	0.374 ± 0.003	0.551 ± 0.024	2.351 ± 0.322
	linear	0.787 ± 0.005	0.350 ± 0.002	0.533 ± 0.039	2.439 ± 0.170
	log	0.785 ± 0.010	0.351 ± 0.002	0.523 ± 0.062	3.936 ± 0.202
BA-SAC (α=0.001)	sparse	0.735 ± 0.031	0.746 ± 0.042	0.596 ± 0.013	3.699 ± 0.181
	linear	0.788 ± 0.005	0.615 ± 0.016	0.560 ± 0.028	3.865 ± 0.250
	log	0.789 ± 0.006	0.609 ± 0.018	0.441 ± 0.040	5.867 ± 0.520
BA-SAC (α=0.01)	sparse	0.718 ± 0.016	0.729 ± 0.063	0.565 ± 0.008	1.039 ± 0.053
	linear	0.784 ± 0.005	0.293 ± 0.017	0.570 ± 0.009	1.149 ± 0.059
	log	0.784 ± 0.006	0.295 ± 0.017	0.567 ± 0.013	1.349 ± 0.094
BA-SAC (α=0.05)	sparse	0.718 ± 0.011	0.544 ± 0.034	0.502 ± 0.004	0.173 ± 0.005
	linear	0.762 ± 0.006	0.169 ± 0.004	0.503 ± 0.003	0.186 ± 0.006
	log	0.762 ± 0.005	0.169 ± 0.004	0.515 ± 0.007	0.381 ± 0.045

was 0.565, 0.570, and 0.567 for the sparse, linear, and logarithmic rewards, respectively, with KL divergence near 1. The logarithmic reward did not uniformly outperform the linear reward, but achieved comparable performance while retaining its direct interpretation under the control-as-inference formulation.

5 Discussion

We formulate the positivity-aware survival policy learning within the framework of control as inference with a principled decomposition of the sparse survival signal into dense rewards which have a natural interpretation in terms of survival analysis widely used in clinical evidence synthesis. We demonstrate that the proposed method can improve survival policy learning compared with naive sparse survival reward in several settings, while keeping the learned policy close to the treatment distribution observed in the data. The similar performance of the logarithmic and linear rewards is also consistent with their theoretical relationship, since $\log(1 - \lambda)$ is well approximated by $-\lambda$ in low-hazard settings.

Nevertheless, our study has several limitations. First, we did not evaluate our method in an online environment. However, online evaluation is infeasible for outcomes that unfold over decades, as in the longitudinal study used in our experiments. Our method offers a way to infer long-term effects through a short-term decomposition, since the proposed reward is defined as a function of the short-term event risk. Second, we did not compare the proposed method with domain-specific hand-crafted rewards. Our framework suggests that information from domain-specific external sources could improve estimation of the immediate risk, λ, and thereby improve the quality of the reward equation 8. Exploring such domain-specific reward enhancements is an important direction for future research.

REFERENCES

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, December 2005. ISSN 0006-341X. doi: 10.1111/j.1541-0420.2005.00377.x.

Elias Bareinboim, Junzhe Zhang, and Sanghack Lee. An Introduction to Causal Reinforcement Learning. Technical Report R-65, CausalAI Lab, Columbia University, March 2025. URL https://causalai.net/r65.pdf.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1972.tb00899.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1972.tb00899.x. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1972.tb00899.x.

Iván Díaz, Nicholas Williams, Katherine L. Hoffman, and Edward J. Schenck. Nonparametric causal effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association*, 118(542):846–857, 2023. doi: 10.1080/01621459.2021.1955691. URL https://doi.org/10.1080/01621459.2021.1955691. tex.eprint: https://doi.org/10.1080/01621459.2021.1955691.

Dennis Frauen, Tobias Hatt, Valentyn Melnychuk, and Stefan Feuerriegel. Estimating average causal effects from patient trajectories. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37 (6):7586–7594, June 2023. doi: 10.1609/aaai.v37i6.25921. URL https://ojs.aaai.org/index.php/AAAI/article/view/25921.

Justin Fu, Avi Singh, Dibya Ghosh, Larry Yang, and Sergey Levine. Variational Inverse Control with Events: A General Framework for Data-Driven Reward Definition, November 2018. URL http://arxiv.org/abs/1805.11686. arXiv:1805.11686 [cs].

M.A. Hernan and J.M. Robins. *Causal inference: What if*. Chapman & hall/CRC monographs on statistics & applied probab. CRC Press, 2024. ISBN 978-1-4200-7616-5. URL https://books.google.co.jp/books?id=_KnHIAAACAAJ. tex.lccn: 2022050839.

Alexandre Kalimouttou, Jason N. Kennedy, Jean Feng, Harvineet Singh, Suchi Saria, Derek C. Angus, Christopher W. Seymour, and Romain Pirracchio. Optimal Vasopressin Initiation in Septic Shock: The OVISS Reinforcement Learning Study. *JAMA*, 333(19):1688–1698, May 2025. ISSN 1538-3598. doi: 10.1001/jama.2025.3046.

Edward H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019. doi: 10.1080/01621459.2017.1422737. URL https://doi.org/10.1080/01621459.2017.1422737. tex.eprint: https://doi.org/10.1080/01621459.2017.1422737.

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, November 2018. ISSN 1546-170X. doi: 10.1038/s41591-018-0213-5. URL https://www.nature.com/articles/s41591-018-0213-5.

Flemming Kondrup, Thomas Jiralerspong, Elaine Lau, Nathan de Lara, Jacob Shkrob, My Duc Tran, Doina Precup, and Sumana Basu. Towards Safe Mechanical Ventilation Treatment Using Deep Offline Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):15696–15702, 2023. ISSN 2374-3468. doi: 10.1609/aaai.v37i13.26862. URL https://ojs.aaai.org/index.php/AAAI/article/view/26862.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 978-1-7138-2954-6.

Mark J. van der Laan and Susan Gruber. Targeted Minimum Loss Based Estimation of Causal Effects of Multiple Time Point Interventions. *The International Journal of Biostatistics*, 8(1), May 2012. ISSN 1557-4679. doi: 10.1515/1557-4679.1370. URL https://www.degruyter.com/document/doi/10.1515/1557-4679.1370/html.

Sergey Levine. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review, May 2018. URL http://arxiv.org/abs/1805.00909. arXiv:1805.00909 [cs].

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The eleventh international conference on learning representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with Deep Reinforcement Learning, December 2013. URL http://arxiv.org/abs/1312.5602. arXiv:1312.5602 [cs].

Arne Peine, Ahmed Hallawa, Johannes Bickenbach, Guido Dartmann, Lejla Begic Fazlic, Anke Schmeink, Gerd Ascheid, Christoph Thiemermann, Andreas Schuppert, Ryan Kindle, Leo Celi, Gernot Marx, and Lukas Martin. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *npj Digital Medicine*, 4(1):32, February 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00388-6. URL https://www.nature.com/articles/s41746-021-00388-6.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: visual reasoning with a general conditioning layer. In *Proceedings of the thirty-second AAAI conference on artificial intelligence and thirtieth innovative applications of artificial intelligence conference and eighth AAAI symposium on educational advances in artificial intelligence*, AAAI'18/IAAI'18/EAAI'18, New Orleans, Louisiana, USA, 2018. AAAI Press. ISBN 978-1-57735-800-8.

Maya L. Petersen and Mark J. van der Laan. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology (Cambridge, Mass.)*, 25(3):418–426, May 2014. ISSN 1531-5487. doi: 10.1097/EDE.0000000000000078.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

Luca F. Roggeveen, Ali El Hassouni, Harm-Jan de Grooth, Armand R. J. Girbes, Mark Hoogendoorn, Paul W. G. Elbers, and Dutch ICU Data Sharing Against COVID-19 Collaborators. Reinforcement learning for intensive care medicine: actionable clinical insights from novel approaches to reward shaping and off-policy model evaluation. *Intensive Care Medicine Experimental*, 12(1):32, March 2024. ISSN 2197-425X. doi: 10.1186/s40635-024-00614-x.

Toru Shirakawa, Yi Li, Yulun Wu, Sky Qiu, Yuxuan Li, Mingduo Zhao, Hiroyasu Iso, and Mark J. Van Der Laan. Longitudinal targeted minimum loss-based estimation with temporal-difference heterogeneous Transformer. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45097–45113. PMLR, July 2024. URL https://proceedings.mlr.press/v235/shirakawa24a.html.

Daniel J. Tan, Kay Choong See, and Mengling Feng. Learning Preference-Based Objectives from Clinical Narratives for Sequential Treatment Decision-Making, April 2026. URL http://arxiv.org/abs/2604.10783. arXiv:2604.10783 [cs].

Franki Nguimatsia Tiofack, Fabian Schramm, Théotime Le Hellard, and Justin Carpentier. SVL: Goal-Conditioned Reinforcement Learning as Survival Learning, April 2026. URL http://arxiv.org/abs/2604.17551. arXiv:2604.17551 [cs].

Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW.

M.J. van der Laan and S. Rose. *Targeted learning in data science: Causal inference for complex longitudinal studies*. Springer series in statistics. Springer International Publishing, 2018. ISBN 978-3-319-65304-4.

Hongjun Wang, Wei Liu, Weibo Gu, Xing Sun, and Kai Han. MHPO: Modulated Hazard-aware Policy Optimization for Stable Reinforcement Learning, March 2026. URL http://arxiv.org/abs/2603.16929. arXiv:2603.16929 [cs].

Kazumasa Yamagishi, Isao Muraki, Yasuhiko Kubota, Mina Hayama-Terada, Hironori Imano, Renzhe Cui, Mitsumasa Umesawa, Yuji Shimizu, Tomoko Sankai, Takeo Okada, Shinichi Sato, Akihiko Kitamura, Masahiko Kiyama, and Hiroyasu Iso. The Circulatory Risk in Communities Study (CIRCS): A Long-Term Epidemiological Study for Lifestyle-Related Disease Among Japanese Men and Women Living in Communities. *Journal of Epidemiology*, 29(3):83–91, March 2019. ISSN 1349-9092. doi: 10.2188/jea.JE20180196.

Chao Yu, Jiming Liu, and Hongyi Zhao. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. *BMC Medical Informatics and Decision Making*, 19(2):57, April 2019. ISSN 1472-6947. doi: 10.1186/s12911-019-0763-6. URL https://doi.org/10.1186/s12911-019-0763-6.

Zheng Zhang, Ziwei Shan, Kaitao Song, Yexin Li, and Kan Ren. Linking Process to Outcome: Conditional Reward Modeling for LLM Reasoning. In *International Conference on Learning Representations (ICLR)*, October 2025. URL https://openreview.net/forum?id=4DJoBOQNd0.

Pan Zhao, Antoine Chambaz, Julie Josse, and Shu Yang. Positivity-free policy learning with observational data. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of the 27th international conference on artificial intelligence and statistics*, volume 238 of *Proceedings of machine learning research*, pp. 1918–1926. PMLR, May 2024. URL https://proceedings.mlr.press/v238/zhao24a.html.

A Structural Variational Inference for Optimal Control

In this section, we derive the control-as-inference objective used in Section 3.2. The key modeling choice is to use the behavior policy π_β as the policy prior. This is natural and safe in the offline observational setting because we cannot identify the causal effect of actions outside the support of π_β.

Proof of Proposition 1. Let $X = (S_{1:\tau}, A_{1:\tau})$ denote a state-action trajectory, and let $\mathcal{O}_{1:\tau}$ denote the sequence of survival optimality variables, where $\mathcal{O}_t = 1 - Y_t$. Under the observational data-generating process, the prior trajectory distribution factorizes as

$$p_\beta(X) = p(S_1) \prod_{t=1}^{\tau} \mu(S_{t+1} \mid S_t, A_t)\, \pi_\beta(A_t \mid S_t).$$

The likelihood of the survival optimality variables is

$$p(\mathcal{O}_{1:\tau} = 1 \mid X) = \prod_{t=1}^{\tau} p(\mathcal{O}_t = 1 \mid S_t, A_t) = \prod_{t=1}^{\tau} \{1 - \lambda_t(S_t, A_t)\}.$$

Equivalently, defining

$$r(S_t, A_t) = \log p(\mathcal{O}_t = 1 \mid S_t, A_t) = \log\{1 - \lambda_t(S_t, A_t)\},$$

we have

$$\log p(\mathcal{O}_{1:\tau} = 1 \mid X) = \sum_{t=1}^{\tau} r(S_t, A_t).$$

The posterior over trajectories conditioned on survival is

$$p(X \mid \mathcal{O}_{1:\tau} = 1) = \frac{p_\beta(X) p(\mathcal{O}_{1:\tau} = 1 \mid X)}{p(\mathcal{O}_{1:\tau} = 1)}.$$

Since this posterior is generally intractable, we introduce a variational trajectory distribution induced by a candidate policy π with the environment dynamics fixed:

$$q_\pi(X) = p(S_1) \prod_{t=1}^{\tau} \mu(S_{t+1} \mid S_t, A_t)\, \pi(A_t \mid S_t).$$

The evidence for survival can be lower-bounded by Jensen's inequality:

$$\begin{aligned}
\log p(\mathcal{O}_{1:\tau} = 1) &= \log \int p_\beta(X) p(\mathcal{O}_{1:\tau} = 1 \mid X)\, dX \\
&= \log \int q_\pi(X) \frac{p_\beta(X) p(\mathcal{O}_{1:\tau} = 1 \mid X)}{q_\pi(X)}\, dX \\
&\geq \mathbb{E}_{q_\pi} \left[\log p(O_{1:\tau} = 1 \mid X) + \log p_\beta(X) - \log q_\pi(X) \right].
\end{aligned}$$

Substituting the factorization of p_β and q_π, the transition terms cancel because the candidate policy changes only the action mechanism, not the environment dynamics. Therefore,

$$\begin{aligned}
\log p_\beta(X) - \log q_\pi(X) &= \sum_{t=1}^{\tau} \log \pi_\beta(A_t \mid S_t) - \sum_{t=1}^{\tau} \log \pi(A_t \mid S_t) \\
&= -\sum_{t=1}^{\tau} \log \frac{\pi(A_t \mid S_t)}{\pi_\beta(A_t \mid S_t)}.
\end{aligned}$$

Thus the variational lower bound becomes

$$\begin{aligned}
\mathcal{L}(\pi) &= \mathbb{E}_{q_\pi} \Big[\sum_{t=1}^{\tau} r(S_t, A_t) - \sum_{t=1}^{\tau} \log \frac{\pi(A_t \mid S_t)}{\pi_\beta(A_t \mid S_t)} \Big] \\
&= \mathbb{E}_{q_\pi} \Big[\sum_{t=1}^{\tau} r(S_t, A_t) \Big] - \mathbb{E}_{S_{1:\tau} \sim q_\pi} \Big[\sum_{t=1}^{\tau} D_{\mathrm{KL}} \big(\pi(\cdot \mid S_t) \,\|\, \pi_\beta(\cdot \mid S_t) \big) \Big].
\end{aligned}$$

Therefore, maximizing the evidence lower bound over feasible policies is equivalent to solving

$$\pi^\star \in \arg\max_{\pi \in \Pi^\circ} \left\{ \mathbb{E}_{q_\pi}\Big[\sum_{t=1}^{\tau} r(S_t, A_t)\Big] - \mathbb{E}_{S_{1:\tau}\sim q_\pi}\Big[\sum_{t=1}^{\tau} D_{\mathrm{KL}}\big(\pi(\cdot \mid S_t) \,\|\, \pi_\beta(\cdot \mid S_t)\big)\Big] \right\}.$$

This is precisely the behavior-anchored objective in Proposition 1. □

Equivalently, the posterior policy induced by conditioning on future survival has the Bayes reweighting of the behavior policy

$$p(A_t \mid S_t, \mathcal{O}_{t:\tau} = 1) \propto \pi_\beta(A_t \mid S_t)\beta_t(S_t, A_t),$$

where $\beta_t(S_t, A_t) = p(\mathcal{O}_{t:\tau} \mid S_t, A_t)$ is the backward message. Hence the behavior policy π_β acts as the reference measure, or prior, and the survival likelihood contributes the reward

$$r(S_t, A_t) = \log\{1 - \lambda_t(S_t, A_t)\}.$$

This derivation shows that the KL penalty is not an ad hoc regularizer: it is the variational cost of changing the action distribution from the observational behavior policy to the learned intervention policy.

B PROOFS

Proof of Proposition 2. Fix a feasible policy π induced by the counterfactual regime g. Recall that the at-risk intensity is defined as

$$\lambda(t \mid S_t, A_t) = \mathbf{1}\{Y_{t-1} = 0\} \Pr(Y_t = 1 \mid S_t, A_t, Y_{t-1} = 0).$$

Therefore,

$$1 - \lambda(t \mid S_t, A_t) = \Pr(Y_t = 0 \mid S_t, A_t, Y_{t-1} = 0)$$

on the at-risk event $\{Y_{t-1} = 0\}$, while $\lambda(t \mid S_t, A_t) = 0$ off the at-risk event.

Under the intervention g, the observational action mechanism is replaced by the policy π, while the state transition and outcome mechanisms are left unchanged. By sequential randomization and positivity, the distribution of the counterfactual trajectory is identified by the longitudinal g-formula. In particular, the at-risk state-action trajectory distribution under g is identified by q^π.

Now consider the counterfactual survival probability

$$S^g(\tau) = \Pr(Y_\tau^g = 0).$$

Since the event process is absorbing, $Y_\tau^g = 0$ is equivalent to

$$Y_1^g = 0, \ldots, Y_\tau^g = 0.$$

Hence, by the chain rule,

$$\begin{aligned} S^g(\tau) &= \Pr(Y_1^g = 0, \ldots, Y_\tau^g = 0) \\ &= \mathbb{E}_{p^\pi}\Big[\prod_{t=1}^{\tau} \Pr(Y_t^g = 0 \mid S_t^g, A_t^g, Y_{t-1}^g = 0)\Big]. \end{aligned}$$

By consistency and the definition of the at-risk intensity,

$$\Pr(Y_t^g = 0 \mid S_t^g, A_t^g, Y_{t-1}^g = 0) = 1 - \lambda(t \mid S_t^g, A_t^g),$$

because on the event $\{Y_{t-1}^g = 0\}$,

$$\lambda(t \mid S_t^g, A_t^g) = \Pr(Y_t^g = 1 \mid S_t^g, A_t^g, Y_{t-1}^g = 0).$$

Therefore,

$$S^g(\tau) = \mathbb{E}_{p^\pi}\Big[\prod_{t=1}^{\tau}\{1 - \lambda(t \mid S_t, A_t)\}\Big].$$

This proves the claim. □

Proof of Proposition 3. This alternative reward is based on the expression

$$Y_\tau = \sum_{t=1}^{\tau} \mathbb{1}\{Y_{t-1} = 0\} Y_t, \tag{24}$$

where $Y_0 = 0$. By the law of iterated expectation, we have the decomposition of the counterfactual survival function into the expectation of the cumulative negative intensity:

$$\begin{aligned}
S^g(\tau) &= 1 + \mathbb{E}_{p^\pi}[-Y_\tau] \\
&= 1 + \mathbb{E}_{p^\pi}\Big[\sum_{t=1}^{\tau} -\mathbb{1}\{Y_{t-1} = 0\} Y_t\Big] \\
&= 1 + \sum_{t=1}^{\tau} \mathbb{E}_{p^\pi}\big[-\mathbb{1}\{Y_{t-1} = 0\} Y_t\big] \\
&= 1 + \sum_{t=1}^{\tau} \mathbb{E}_{p^\pi}\Big[-\mathbb{E}_{p^\pi}\big[\mathbb{1}\{Y_{t-1} = 0\} Y_t \mid S_t, A_t, Y_{t-1} = 0\big]\Big] \\
&= 1 + \sum_{t=1}^{\tau} \mathbb{E}_{p^\pi}\Big[-\mathbb{E}\big[\mathbb{1}\{Y_{t-1} = 0\} Y_t \mid S_t, A_t, Y_{t-1} = 0\big]\Big] \\
&= 1 + \sum_{t=1}^{\tau} \mathbb{E}_{p^\pi}\Big[-\mathbb{1}\{Y_{t-1} = 0\} \mathbb{E}\big[Y_t \mid S_t, A_t, Y_{t-1} = 0\big]\Big] \\
&= 1 + \sum_{t=1}^{\tau} \mathbb{E}_{p^\pi}\big[-\mathbb{1}\{Y_{t-1} = 0\} \lambda(t \mid S_t, A_t)\big] \\
&= 1 + \mathbb{E}_{p^\pi}\Big[\sum_{t=1}^{\tau} -\mathbb{1}\{Y_{t-1} = 0\} \lambda(t \mid S_t, A_t)\Big].
\end{aligned}$$

Hence we have

$$\mathbb{E}_{p^\pi}\Big[\sum_{t=1}^{\tau} -\mathbb{1}\{Y_{t-1} = 0\} \lambda(t \mid S_t, A_t)\Big] = S^g(\tau) - 1.$$

□

Proof of Proposition 4. Fix a time t and state s. Given the current critic Q_t^π, the policy improvement problem is pointwise in s:

$$\max_{\rho \in \Delta(\mathcal{A})} \sum_a \rho(a \mid s) Q_t^\pi(s, a) - \alpha \sum_a \rho(a \mid s) \log \frac{\rho(a \mid s)}{\pi_\beta(a \mid s)}.$$

This is a strictly concave optimization problem in ρ whenever $\alpha > 0$ and $\pi_\beta(a \mid s) > 0$ on the support of admissible actions. Introducing a Lagrange multiplier $\eta(s)$ for the constraint $\sum_a \rho(a \mid s) = 1$, the Lagrangian is

$$\mathcal{L}(\rho, \eta) = \sum_a \rho(a \mid s) Q_t^\pi(s, a) - \alpha \sum_a \rho(a \mid s) \log \frac{\rho(a \mid s)}{\pi_\beta(a \mid s)} + \eta(s) \left(\sum_a \rho(a \mid s) - 1 \right).$$

The first-order condition with respect to $\rho(a \mid s)$ gives

$$Q_t^\pi(s, a) - \alpha \left\{ \log \frac{\rho(a \mid s)}{\pi_\beta(a \mid s)} + 1 \right\} + \eta(s) = 0.$$

Rearranging,

$$\log \frac{\rho(a \mid s)}{\pi_\beta(a \mid s)} = \frac{1}{\alpha} Q_t^\pi(s, a) + c(s),$$

where $c(s)$ absorbs constants that do not depend on a. Therefore

$$\rho(a \mid s) \propto \pi_\beta(a \mid s) \exp\{Q_t^\pi(s, a)/\alpha\}.$$

Normalizing over actions yields

$$\pi_{\mathrm{new}}(a \mid s) = \frac{\pi_\beta(a \mid s)\exp\{Q_t^\pi(s,a)/\alpha\}}{\sum_{a'}\pi_\beta(a' \mid s)\exp\{Q_t^\pi(s,a')/\alpha\}}.$$

It remains to show that this update improves the survival variational objective. By definition of π_{new}, for every state s,

$$\begin{aligned}&\mathbb{E}_{A\sim\pi_{\mathrm{new}}(\cdot|s)}\Big[Q_t^\pi(s,A) - \alpha\log\frac{\pi_{\mathrm{new}}(A \mid s)}{\pi_\beta(A \mid s)}\Big]\\ &\quad\geq \mathbb{E}_{A\sim\pi(\cdot|s)}\Big[Q_t^\pi(s,A) - \alpha\log\frac{\pi(A \mid s)}{\pi_\beta(A \mid s)}\Big] = V_t^\pi(s).\end{aligned}$$

Thus the soft value under the improved policy is no smaller than the soft value under the previous policy at each state.

Using the soft Bellman equation,

$$Q_t^\pi(s,a) = r_t(s,a) + \mathbb{E}\left[V_{t+1}^\pi(S_{t+1}) \mid S_t = s, A_t = a\right],$$

the one-step improvement inequality implies

$$V_t^\pi(s) \leq \mathbb{E}_{A_t\sim\pi_{\mathrm{new}}(\cdot|s)}\Big[r_t(s,A_t) - \alpha\log\frac{\pi_{\mathrm{new}}(A_t \mid s)}{\pi_\beta(A_t \mid s)} + \mathbb{E}\left[V_{t+1}^\pi(S_{t+1}) \mid S_t = s, A_t\right]\Big].$$

Iterating this inequality forward in time gives

$$V_0^\pi(s) \leq \mathbb{E}_{q_{\pi_{\mathrm{new}}}}\Big[\sum_{t=0}^{T-1}\Big\{r_t(S_t,A_t) - \alpha\log\frac{\pi_{\mathrm{new}}(A_t \mid S_t)}{\pi_\beta(A_t \mid S_t)}\Big\}\,\Big|\,S_0 = s\Big].$$

Averaging over the baseline state distribution yields

$$\mathcal{J}(\pi_{\mathrm{new}}) \geq \mathcal{J}(\pi).$$

Therefore the policy improvement step monotonically improves the KL-regularized survival objective.

Finally, the critic update estimates the soft Bellman equation for Q^π, while the actor update performs the pointwise KL-regularized maximization derived above. Hence the alternating critic and actor updates constitute stochastic coordinate ascent on the survival ELBO. Since

$$r_t(s,a) = \log\{1 - \lambda_t(s,a)\},$$

this ELBO is precisely the control-as-inference lower bound associated with longitudinal survival optimality. □

C RISK-SENSITIVE INTERPRETATION.

The logarithmic survival reward also admits a risk-sensitive interpretation. To make this precise, define the cumulative log-survival loss

$$\mathcal{L}_{\log}(\pi) := \mathbb{E}_{p^\pi}\Big[\sum_{t=1}^{\tau} -\log\{1 - \lambda(t \mid S_t, A_t)\}\Big].$$

Proposition 5. *For any feasible policy π and any $\varepsilon \in (0,1)$,*

$$\Pr_{p^\pi}\Big(\max_{1\leq t\leq\tau}\lambda(t \mid S_t, A_t) > \varepsilon\Big) \leq \frac{\mathcal{L}_{\log}(\pi)}{-\log(1-\varepsilon)}.$$

Proof. Let

$$Z := \sum_{t=1}^{\tau} -\log\{1 - \lambda(t \mid S_t, A_t)\}.$$

Since $\lambda(t \mid S_t, A_t) \in [0, 1)$, each term in the sum is nonnegative. Consider the event

$$E_\varepsilon := \Big\{ \max_{1 \le t \le \tau} \lambda(t \mid S_t, A_t) > \varepsilon \Big\}.$$

On E_ε, there exists some time point $t^\star \le \tau$ such that

$$\lambda(t^\star \mid S_{t^\star}, A_{t^\star}) > \varepsilon.$$

Because the function $x \mapsto -\log(1-x)$ is increasing on $[0, 1)$, we have

$$-\log\{1 - \lambda(t^\star \mid S_{t^\star}, A_{t^\star})\} > -\log(1-\varepsilon).$$

Therefore,

$$Z = \sum_{t=1}^{\tau} -\log\{1 - \lambda(t \mid S_t, A_t)\} \ge -\log\{1 - \lambda(t^\star \mid S_{t^\star}, A_{t^\star})\} > -\log(1-\varepsilon).$$

Hence,

$$E_\varepsilon \subseteq \{Z > -\log(1-\varepsilon)\}.$$

By Markov's inequality,

$$\Pr_{p^\pi}(Z > -\log(1-\varepsilon)) \le \frac{\mathbb{E}_{p^\pi}[Z]}{-\log(1-\varepsilon)}.$$

Combining the event inclusion with Markov's inequality gives

$$\Pr_{p^\pi}\Big(\max_{1 \le t \le \tau} \lambda(t \mid S_t, A_t) > \varepsilon \Big) \le \frac{\mathbb{E}_{p^\pi}\left[\sum_{t=1}^{\tau} -\log\{1 - \lambda(t \mid S_t, A_t)\}\right]}{-\log(1-\varepsilon)}.$$

□

Comparison with the linear dense reward. The linear dense reward $-\lambda(t \mid S_t, A_t)$ and the logarithmic reward $\log\{1 - \lambda(t \mid S_t, A_t)\}$ both provide dense survival-derived signals, but they induce different risk profiles. Maximizing the linear reward is equivalent to minimizing

$$\mathcal{L}_{\text{lin}}(\pi) := \mathbb{E}_{p^\pi}\Big[\sum_{t=1}^{\tau} \lambda(t \mid S_t, A_t) \Big],$$

which directly penalizes the expected cumulative event intensity. By contrast, maximizing the logarithmic reward is equivalent to minimizing

$$\mathcal{L}_{\log}(\pi) := \mathbb{E}_{p^\pi}\Big[\sum_{t=1}^{\tau} -\log\{1 - \lambda(t \mid S_t, A_t)\} \Big].$$

The difference follows from the expansion

$$-\log(1-\lambda) = \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{3} + \cdots, \qquad \lambda \in [0, 1).$$

Thus,

$$-\log(1-\lambda) \ge \lambda,$$

with the gap increasing as λ becomes large. Moreover,

$$-\log(1-\lambda) \to \infty \qquad \text{as} \qquad \lambda \to 1.$$

Therefore, the logarithmic objective penalizes high event intensities superlinearly, whereas the linear objective penalizes them only linearly.

This shows that the logarithmic reward is more risk-sensitive than the linear reward. The linear reward mainly controls the average cumulative intensity, while the logarithmic reward places disproportionately greater penalties on high-hazard state–action pairs. Proposition 5 formalizes this intuition: a small value of $\mathcal{L}_{\log}(\pi)$ directly controls the probability that any state-action pair along the trajectory has event intensity exceeding a threshold ε. Hence, compared with optimizing $-\lambda(t \mid S_t, A_t)$, optimizing $\log\{1 - \lambda(t \mid S_t, A_t)\}$ favors policies that avoid high-hazard regions of the trajectory distribution.

D Extension to Dataset with Right Censoring

We extend the formulation to longitudinal data subject to right censoring. Let $C_t \in \{0, 1\}$ indicate whether a trajectory is censored at time t, and define the observed at-risk indicator

$$R_t = \mathbb{1}\{Y_{t-1} = 0,\ C_{t-1} = 0\}.$$

Censoring determines whether subsequent states and outcomes are observed, but does not alter the full-data causal estimand

$$S^\pi(\tau) = \Pr(Y_\tau^\pi = 0)$$

or the survival-derived reward

$$r_t^{\log}(S_t, A_t) = \log\{1 - \lambda(t \mid S_t, A_t)\}.$$

Suppose censoring is conditionally independent of the future counterfactual trajectory given the observed history, or equivalently the learned state and current action:

$$C_t \perp (Y_{t:\tau}^\pi, S_{t+1:\tau}^\pi) \mid S_t, A_t, Y_{t-1} = 0, C_{t-1} = 0,$$

together with censoring positivity,

$$\Pr(C_t = 0 \mid S_t, A_t, R_t = 1) > 0$$

on the support of the target policy. Then the at-risk event intensity remains identified from uncensored transitions as

$$\lambda(t \mid S_t, A_t) = \Pr(Y_t = 1 \mid S_t, A_t, R_t = 1, C_t = 0).$$

Consequently, the control-as-inference derivation and the policy objective in Equation 9 are unchanged. In practice, censored transitions are masked from the outcome loss and from Bellman targets requiring an unobserved subsequent state; censoring is never encoded as a failure event.

D.2 Censoring depending on observed history

More generally, censoring may depend strongly on the observed longitudinal history. Let

$$g_t(S_t, A_t) = \Pr(C_t = 0 \mid S_t, A_t, R_t = 1)$$

denote the conditional probability of remaining observed at time t, and define the probability of remaining uncensored through time t by

$$G_t = \prod_{k=1}^{t} g_k(S_k, A_k).$$

Under conditional independent censoring given the observed history and censoring positivity, inverse-probability-of-censoring weighting (IPCW) recovers expectations under the corresponding full-data distribution. For an integrable quantity h_t observed while uncensored,

$$\mathbb{E}_{p^\pi}[h_t] = \mathbb{E}_{p^\pi}\left[\frac{\mathbb{1}(C_{1:t} = 0)}{G_t} h_t\right].$$

Accordingly, Stage 1 can estimate the event intensity using the IPCW outcome loss

$$\mathcal{L}_Y^{\mathrm{IPCW}} = -\sum_{t=1}^{\tau} \frac{R_t(1 - C_t)}{\widehat{G}_t}\left[Y_t \log \widehat{\lambda}_t + (1 - Y_t)\log(1 - \widehat{\lambda}_t)\right],$$

with stabilized or truncated weights used in practice when necessary. Analogously, Stage-2 critic updates may be weighted by the estimated inverse censoring probability when censoring induces selective observation of transitions:

$$\mathcal{L}_Q^{\mathrm{IPCW}}(\psi_i) = \mathbb{E}_{\mathcal{D}}\left[\frac{R_t(1 - C_t)}{\widehat{G}_t}\left\{Q_{\psi_i}(S_t, A_t) - \widetilde{Y}_t\right\}^2\right],$$

where $\widetilde{Y}_t$ is the behavior-anchored soft Bellman target in Equation (22).

Importantly, IPCW modifies how the reward model and Bellman equations are estimated from censored observational data; it does not modify the survival reward itself. Once the full-data intensity is identified,

$$r_t^{\log}(S_t, A_t) = \log\{1 - \lambda(t \mid S_t, A_t)\}$$

remains the control-as-inference reward, and the behavior-policy KL term is unchanged. Thus right censoring and treatment-policy regularization address distinct identification problems: the former concerns selective observation of future trajectories, whereas the latter restricts policy improvement to treatment decisions supported by the observational data.

If censoring remains informative after conditioning on the available history (or on a sufficient state representation), the full-data survival distribution is not nonparametrically identified without additional assumptions or external information. In that setting, neither masking nor IPCW alone is sufficient, and sensitivity analysis or an explicit model for the dependence between censoring and the unobserved outcome process is required.

E State Space Learning

This section describes Stage 1 of Algorithm 1, which learns the latent state S_t together with the at-risk intensity $\hat{\lambda}_t$ and the behavior policy $\hat{\pi}_\beta$ used in Section 3.6. All three components share a single autoregressive Transformer trained end-to-end on the offline trajectories $\mathcal{D} = \{(W^{(i)}, L_{1:\tau}^{(i)}, A_{1:\tau}^{(i)}, Y_{1:\tau}^{(i)}, C_{1:\tau}^{(i)})\}_{i=1}^N$, where W collects time-invariant baseline covariates and C_t is the censoring indicator. We write $L_t = (L_t^{\text{cont}}, L_t^{\text{cat}})$ for the continuous and categorical components of the time-varying covariate.

E.1 Tokenization.

We construct an interleaved token sequence

$$\big(W,\ L_1,\ A_1,\ L_2,\ A_2,\ \ldots,\ L_\tau,\ A_\tau\big),$$

mapping each token to $\mathbb{R}^d$ via a type-specific linear embedding. Continuous and categorical features within L_t are embedded separately and summed; categorical levels use learned embeddings with a dedicated index reserved for missing values. To every token we add (i) a sinusoidal positional encoding indexed by t and (ii) a learned type embedding e_W, e_L, e_A that signals the modality of the token, allowing the encoder to share parameters across modalities while distinguishing covariate updates from treatment decisions.

E.2 Encoder.

A stack of N pre-norm Transformer layers with multi-head self-attention processes the token sequence under a strict causal mask, so that the representation at any token position depends only on tokens earlier in clinical time. Concretely,

$$S_t \ = \ f_\theta\big(W,\ L_{1:t},\ A_{1:t-1}\big), \tag{25}$$

read off at the L_t-token position immediately preceding the action A_t. The full set of state representations $S_{1:\tau}$ is produced in a single forward pass. As assumed in Section 3.4, $S_t = f_\theta(H_t)$ is taken to be sufficient for treatment assignment and potential outcomes, so that policy evaluation and optimization can be performed in the learned state space.

E.3 Prediction heads.

Five heads operate on the encoder outputs. We write H_t for the hidden state at the L_t position (heads conditioned on S_t alone) and H_t^A for the hidden state at the A_t position (heads conditioned on (S_t, A_t)).

Intensity head. A linear–sigmoid head g_ϕ predicts the at-risk failure probability

$$\hat{\lambda}_t(S_t, A_t) \ = \ \sigma\big(g_\phi(H_t^A)\big) \ \approx \ \Pr(Y_t = 1 \mid S_t, A_t, Y_{t-1} = 0), \tag{26}$$

consistent with the at-risk intensity. This is the quantity that enters the dense reward $\hat{r}_t = \log(1 - \hat{\lambda}_t)$ in equation 8.

Censoring head. A linear–sigmoid head g_c predicts the at-risk censoring probability $\hat{p}(C_t = 1 \mid S_t, A_t)$. It is used only as auxiliary supervision to keep the encoder aware of dropout dynamics; censoring is disabled at evaluation time and never enters the reward.

Behavior-policy head. A linear–sigmoid head outputs the action propensity $\hat{\pi}_\beta(A_t = 1 \mid S_t)$, providing the KL prior in the actor objective equation 23. Because this head shares the encoder f_θ with g_ϕ, the prior is automatically defined on the same state on which the reward is evaluated.

Continuous-covariate flow-matching head. For the continuous block L_{t+1}^{cont} we learn $p(L_{t+1}^{\text{cont}} \mid S_t, A_t)$ with Optimal-Transport Conditional Flow Matching (Lipman et al., 2023; Tong et al., 2024). A velocity network $v_\psi(x, t'; H_t^A)$, implemented as a multilayer perceptron whose layers are FiLM-modulated (Perez et al., 2018) by H_t^A and a sinusoidal embedding of the flow time $t' \in [0, 1]$, parameterizes the drift of a continuous normalizing flow that transports $\mathcal{N}(0, I)$ to the target conditional law. Training minimizes

$$\mathcal{L}_{\text{flow}} \;=\; \sum_{t=1}^{\tau-1} \mathbb{E}_{t' \sim \mathcal{U}(0,1),\, x_0 \sim \mathcal{N}(0,I)} \Big[\big\| v_\psi(x_{t'}, t'; H_t^A) - u_{t'} \big\|_2^2 \Big], \tag{27}$$

along the linear OT path $x_{t'} = (1 - t')x_0 + t' L_{t+1}^{\text{cont}}$ with target velocity $u_{t'} = L_{t+1}^{\text{cont}} - x_0$. Sampling at inference uses an Euler discretization with K steps. The flow head ensures that the encoder retains the information about the continuous covariate dynamics needed for counterfactual rollouts and propensity estimation.

Categorical-covariate head. Per-variable softmax heads predict L_{t+1}^{cat} from H_t^A and are trained with cross-entropy.

E.4 At-risk weighting.

Supervision is restricted to time points at which the patient is alive and observed. With at-risk indicator

$$R_t \;=\; \mathbb{1}\{Y_{t-1} = 0,\; C_{t-1} = 0\}, \qquad R_1 \equiv 1, \tag{28}$$

the per-step losses are

$$\mathcal{L}_Y = -\sum_{t=1}^{\tau} R_t (1 - C_t) \big[Y_t \log \hat{\lambda}_t + (1 - Y_t) \log(1 - \hat{\lambda}_t) \big], \tag{29}$$

$$\mathcal{L}_C = -\sum_{t=1}^{\tau} R_t \big[C_t \log \hat{p}_C + (1 - C_t) \log(1 - \hat{p}_C) \big], \tag{30}$$

$$\mathcal{L}_A = -\sum_{t=1}^{\tau} R_t \big[A_t \log \hat{\pi}_\beta + (1 - A_t) \log(1 - \hat{\pi}_\beta) \big], \tag{31}$$

$$\mathcal{L}_{\text{flow}}^{\text{w}} = \sum_{t=1}^{\tau-1} R_{t+1}\, \ell_{\text{flow}}(t), \qquad \mathcal{L}_{\text{cat}}^{\text{w}} = \sum_{t=1}^{\tau-1} R_{t+1}\, \ell_{\text{cat}}(t), \tag{32}$$

where $\ell_{\text{flow}}(t)$ and $\ell_{\text{cat}}(t)$ are the per-step flow-matching and categorical losses defined above. Weighting the outcome term by the additional factor $1 - C_t$ matches the at-risk intensity and ensures that $\hat{\lambda}_t$ targets the causal estimand of Section 3.3 rather than the marginal hazard.

E.5 Total objective.

The encoder, the intensity, the censoring, the behavior policy, and the observation models are learned jointly by minimizing

$$\mathcal{L}_{\text{state}}(\theta, \phi, \psi) \;=\; w_Y \mathcal{L}_Y + w_C \mathcal{L}_C + w_A \mathcal{L}_A + w_{\text{flow}} \mathcal{L}_{\text{flow}}^{\text{w}} + w_{\text{cat}} \mathcal{L}_{\text{cat}}^{\text{w}}. \tag{33}$$

We set w_Y as the dominant term so that the dense reward $\hat{r}_t = \log(1 - \hat{\lambda}_t)$ retains a faithful interpretation as the log-survival increment of equation 8; the remaining weights are tuned on a held-out split. Joint training, rather than a two-stage scheme that first fits an encoder and then a hazard, is important empirically: it forces the encoder to retain features predictive of Y_t, which is the only signal that matters for the downstream RL objective equation 9.

E.6 Implementation.

We use $N \in \{2, \ldots, 6\}$ Transformer layers, hidden dimension $d \in \{8, \ldots, 256\}$, and $\{2, 4, 8\}$ attention heads chosen as a function of d. Optimization uses AdamW with a cosine-annealed learning rate (peak $\approx 4 \times 10^{-4}$, 5% warmup), batch size 64, and up to 100 epochs with early stopping on validation loss. The flow-matching head uses $K = 20$ Euler steps at inference. Per-dataset hyperparameter configurations are reported with the released code.

E.7 Use in Stage 2.

After Stage 1 training we freeze f_θ, g_ϕ, and $\hat{\pi}_\beta$. Stage 2 (Algorithm 1) consumes (S_t, A_t, S_{t+1}) tuples produced by the frozen encoder, uses $\hat{r}_t = \log\{1 - \hat{\lambda}_t(S_t, A_t)\}$ as the per-step reward, and uses $\hat{\pi}_\beta(\cdot \mid S_t)$ as the KL anchor in equation 23; only the actor and critic networks are updated during offline policy optimization. The flow-matching and categorical heads are not used by Stage 2 directly, but together with the action-propensity head they enable counterfactual rollouts that we use for the synthetic-data ablations in Section F.1.

F Simulator Details

F.1 Synthetic Data

First draw parameters $\alpha_i, \beta_i \sim \text{Normal}((i+1)^{-1}, 0.02)$ and $\gamma_i \sim 2 * \text{Binom}(0.5) - 1$ for $i \in [ht]$, where h is the length of time-dependency with $h = 1$ corresponding to Markovian process. Then, draw error in time-dependent variables $\varepsilon^\ell_{tj} \sim \text{Normal}(0, 0.1)$ for $t \in [\tau]$ and $j \in [p]$, errors in treatment $\varepsilon^A_{t1} \sim \text{Normal}(0, 0.2)$, $\varepsilon^A_{t2} \sim \text{Normal}(0, 0.05)$ for $t \in [\tau]$. For each $t \in [\tau]$, $L_t = \tanh\left(\sum_{k \in [ht]} \alpha_k L_{t-k} + \beta_k \gamma_k (2A_{t-k} - 1)\right) + \varepsilon^\ell_{tj}$, then draw A_t from an indicator function $\mathbb{1}\{(\sigma(s_t+) + \varepsilon^A_{t2}) > 0.5\}$, with $s_t = \tan\left(\prod_{j \in [p]} \bar{L}_j + \bar{A}\right) + \varepsilon^A_{t1}$. The outcome Y_t is drawn from a Bernoulli distribution of a probability $\sigma(p_t)$ with $p_t = \tan\left(\prod_{j \in [p]} \bar{L}_j - 0.7 * (\bar{A} - 0.5)\right) - 4.5$. $Y_t = 1$ if $Y_{t-1} = 1$ for $t > 0$. We set the counterfactual treatment policy as $\mathbb{1}\{\sigma(s_t) > 0.5\}$ for $t \in [\tau]$ and evaluated the counterfactual mean of survival under this policy.

F.2 Semi-Synthetic Data

Data source. The semi-synthetic environment is built on the Circulatory Risk in Communities Study (CIRCS) (Yamagishi et al., 2019), a long-term population-based cardiovascular epidemiology study of Japanese adults conducted across four geographically isolated communities. We use the mortality cohort, comprising $N = 13\,465$ participants followed annually over $\tau = 30$ years; the analysis is based on 403,951 person-time records.

Variable schema. For each participant i we observe a tuple $(W^{(i)}, L^{(i)}_{1:\tau}, A^{(i)}_{1:\tau}, Y^{(i)}_{1:\tau}, C^{(i)}_{1:\tau})$.

- *Baseline covariates* W: age at enrollment, sex, and community of residence.
- *Time-varying covariates* L_t: forty annually measured variables including anthropometrics (BMI), hemodynamic measurements (systolic and diastolic blood pressure), comorbidity indicators and corresponding medication use (hypertension, diabetes, dyslipidemia), serum chemistry (glucose, total protein, albumin, total cholesterol, HDL, triglycerides, uric acid, creatinine, urinary protein), lifestyle variables (alcohol and tobacco history), reproductive status (menopause), and physical-examination findings (ECG ischemia, left-ventricular hypertrophy, atrial fibrillation; Scheie classifications for hypertensive retinopathy). Each variable is paired with a missingness indicator that we treat as an additional categorical channel.
- *Treatment* $A_t \in \{0, 1\}$: a binary indicator of antihypertensive intervention defined by reduction of the systolic blood pressure (SBP) category in 10 mmHg strata, following the operationalization in CIRCS.
- *Outcome* Y_t: all-cause mortality indicator.
- *Censoring* C_t: right-censoring indicator.

Simulator construction. We fit the Stage 1 autoregressive Transformer of Appendix E on the full CIRCS cohort, jointly learning (i) the latent state $S_t = f_\theta(W, L_{1:t}, A_{1:t-1})$, (ii) the at-risk intensity $\hat{\lambda}_t(S_t, A_t)$, (iii) the behavior policy $\hat{\pi}_\beta(A_t \mid S_t)$, and (iv) conditional generators for L_{t+1} (flow-matching for continuous variables, softmax heads for categorical variables and missingness). The fitted model induces a semi-synthetic data-generating process: starting from real baseline tuples $W^{(i)}$ and observed initial covariates $L_1^{(i)}$, we autoregressively roll out (L_{t+1}, A_t, C_t, Y_t) under any candidate policy π by alternating (a) policy sampling $A_t \sim \pi(\cdot \mid S_t)$, (b) hazard sampling $Y_t \sim \mathrm{Bern}(\hat{\lambda}_t(S_t, A_t))$ on the at-risk set, and (c) observation sampling for L_{t+1}. Censoring is disabled at evaluation time. Continuous flow heads use $K = 20$ Euler steps. This construction preserves the marginal distribution of the real baseline population while replacing the unknown true dynamics with a calibrated learned simulator, so that counterfactual survival under a candidate policy can be evaluated to arbitrary Monte-Carlo precision.

Use in experiments. For the in-distribution semi-synthetic results in Table 1b, the cohort contains $13{,}465$ baseline tuples. Stage 1 uses a patient-level training–validation split with early stopping, and the behavior policy is estimated by cross-fitting. Reported metrics are averages of $S^\pi(\tau)$ and $D_{\mathrm{KL}}(\pi_\theta \| \pi_\beta)$ over 10 seeds.