REDUCING DISTRIBUTION DRIFT OVER TIME FOR LONGITUDINAL COUNTERFACTUAL REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating counterfactual outcomes under treatment sequences from longitudinal observational data is essential for time-series personalized decision-making. In this setting, the counterfactual regression model is trained on observations from preceding time steps and used to predict outcomes at future time steps. Therefore, it is crucial for the regression model to generalize across time steps. Existing methods typically address time-varying confounding by balancing representations of different treatment groups, while overlooking the temporal distribution drift of learned representations, which can hinder the generalization ability of the regression model over time. Motivated by this, we propose a longitudinal counterfactual regression method to reduce the temporal distribution drift of learned representations. Specifically, we characterize the temporal distribution drift using the Wasserstein distance between representations at consecutive time steps. Based on this, we establish that the future counterfactual outcome estimation error can be upper bounded by the current factual error and the distribution discrepancies measured across both treatment groups and time steps. Such a theoretical result naturally motivates us to jointly mitigate confounding bias and temporal drift. To achieve this, we propose to learn balanced representations by simultaneously reducing the distribution discrepancies across treatment groups and consecutive time steps. We conduct experiments on benchmark datasets to demonstrate the effectiveness of our method for longitudinal counterfactual regression.

1 INTRODUCTION

Estimating potential outcomes under treatment sequences from longitudinal observational data is fundamental to personalized decision-making (Feuerriegel et al., 2024). In many real-world applications such as healthcare and online advertising, decisions are made at time steps sequentially, with each intervention depending on the evolving state of an individual and affecting future outcomes (Lim, 2018; Bouchattaoui et al., 2024). In healthcare, for example, clinicians often need to assess how a patient's future condition would evolve under different treatment plans given the observed medical history (Murphy, 2003; Bica et al., 2020; Melnychuk et al., 2022).

Compared with static causal effect estimation (Rosenbaum & Rubin, 1983; Shalit et al., 2017; Shi et al., 2019), longitudinal counterfactual regression poses two major challenges: modeling long-range dependencies across time steps and mitigating time-varying confounding bias (Melnychuk et al., 2022; Bouchattaoui et al., 2024). To capture long-range dependencies over time, existing methods typically adopt recurrent network architectures to learn representations of the evolving states of individuals (Lim, 2018; Bica et al., 2020). For example, Melnychuk et al. (2022) design a causal transformer to learn effective representations, and Bouchattaoui et al. (2024) apply a Gated Recurrent Unit network with an information-theoretic regularization to capture long-range dependency. To mitigate time-varying confounding bias, inverse probability reweighting and treatment-balanced representation learning have been leveraged in longitudinal counterfactual regression (Cole & Hernán, 2008; Huang et al., 2024).

Despite their success, these methods usually overlook the temporal distribution drift of the learned representations. Such drift particularly matters in longitudinal counterfactual regression, where the model is trained using observations from preceding time steps and applied to predict counterfactual outcomes at subsequent time steps. Therefore, accurate longitudinal counterfactual regression

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.

requires the generalization ability across time steps, which could be hindered by the temporal distribution drift of the learned representations (Kuznetsov & Mohri, 2015; Du et al., 2021; Qin et al., 2022). Importantly, while individual representations may evolve over time, the overall distribution of representations across the population is expected to remain stable to facilitate generalization across time steps.

Motivated by the above discussion, we seek to address temporal distribution drift in longitudinal counterfactual regression. To this end, we characterize temporal distribution drift through the Wasserstein distance (Courty et al., 2016; Redko et al., 2017) between representations at consecutive time steps, and further establish a theoretical connection between future counterfactual estimation error and the distribution evolution of learned representations. Specifically, we derive an upper bound on future counterfactual estimation error in terms of the current factual estimation error and distribution discrepancies. Here, the discrepancies capture representation imbalance between treatment groups arising from confounding bias as well as distribution discrepancies across consecutive time steps.

Inspired by the theoretical result, we propose a longitudinal counterfactual regression method to address both temporal distribution drift and confounding bias within a unified framework. To achieve this, we propose to learn representations by minimizing the Wasserstein distances between consecutive time steps. At the same time, we reduce the distribution discrepancies across treatment groups by adversarial learning, which trains a treatment group classifier and enforces the representations to maximize the classification error (Ganin et al., 2016). We conduct extensive experiments on different types of benchmark datasets to evaluate the performance of our method for longitudinal counterfactual regression. In addition, we incorporate our method of temporal drift reduction into multiple backbones, demonstrating its general applicability across different model architectures.

We summarize our principal contributions as follows:

- We characterize the temporal distribution drift in longitudinal counterfactual regression, and theoretically establish an upper bound on the counterfactual estimation error in terms of the factual estimation error and the distribution discrepancies across treatment groups and time steps.
- We propose a unified representation learning method named **Re**ducing **C**onfounding bias and **T**emporal drift (ReCT), which simultaneously reduces distribution discrepancies across treatment groups and consecutive time steps.
- We conduct experiments on benchmark datasets to demonstrate the effectiveness of our method and its general applicability across multiple longitudinal counterfactual regression backbones.

2 Related Work

Counterfactual Regression in Static Settings. Counterfactual regression is extensively studied in static settings, where the goal is to estimate potential outcomes of individuals from observational data (Shalit et al., 2017; Shi et al., 2019). Existing approaches mainly focus on learning outcome models and reducing distribution discrepancies between different treatment groups (Shalit et al., 2017; Künzel et al., 2019; Yan et al., 2024). For example, representation-based methods learn treatment-balanced latent representations (Hill, 2011; Johansson et al., 2016; Shalit et al., 2017; Yan et al., 2025), while other approaches develop flexible prediction models for estimating individual treatment responses (Alaa et al., 2017; Curth & Van der Schaar, 2021). Nevertheless, these methods assume that treatments, covariates, and outcomes are observed at a single time step, and therefore cannot directly handle the evolving dependencies and time-varying confounding in longitudinal settings.

Counterfactual Regression Over Time. Counterfactual regression over time aims to estimate future outcomes under hypothetical treatment plans given an individual's observed history, with applications in personalized medicine, epidemiology, and marketing (Bica et al., 2020; Bouchattaoui et al., 2024). Neural methods for this task commonly adopt an autoregressive paradigm, using previously predicted outcomes to generate predictions at subsequent time steps (Bica et al., 2020; Melnychuk et al., 2022; Bouchattaoui et al., 2024). Traditional approaches such as marginal struc-

tural models (MSMs) and G-computation often rely on restrictive parametric models in conventional implementations, limiting the ability to capture complex dependencies in longitudinal data (Robins et al., 2000; Melnychuk et al., 2022). To overcome these limitations, Lim (2018) incorporate recurrent neural networks (RNNs; Elman, 1990) into MSMs for propensity estimation and outcome prediction. Li et al. (2021) combines RNNs with G-computation to predict counterfactual outcomes under dynamic treatment regimes. Another line of work learns treatment-balanced representations that preserve information for outcome prediction while reducing bias induced by treatment assignment (Bica et al., 2020; Melnychuk et al., 2022; Bouchattaoui et al., 2024; Wang et al., 2024). Bica et al. (2020) uses adversarial training to learn treatment-invariant representations with a recurrent encoder–decoder architecture. Melnychuk et al. (2022) build on the Transformer architecture introduced by Vaswani et al. (2017) to capture long-range dependencies through attention mechanisms and learn balanced representations using a counterfactual domain-confusion loss. Bouchattaoui et al. (2024) use a gated recurrent unit (GRU; Cho et al., 2014) and combine contrastive predictive coding and information maximization with mutual information minimization to learn informative, treatment-balanced historical representations. In addition to addressing confounding bias considered in the above studies, the proposed method reduces temporal distribution drift by aligning representation distributions across consecutive time steps to improve generalization across time steps.

3 Problem Formulation

Following the longitudinal potential-outcomes framework (Rubin, 1978; Robins, 1986), we consider N individuals observed over discrete time. For individual i with static covariates $V_i \in \mathcal{V} \subset R^{d_v}$, at time step t, let $X_{i,t} \in \mathcal{X} \subset R^{d_x}$ denote time-varying covariates, $A_{i,t} \in \mathcal{A}$ denote the discrete treatment, and $Y_{i,t} \in \mathcal{Y} \subset R$ denote the outcome. The history available before assigning treatment $A_{i,t+1}$ is

$$H_{i,t+1} = \left[V_i, X_{i,1:t+1}, A_{i,1:t}, Y_{i,1:t}\right]. \tag{1}$$

For a prediction horizon $\tau \geq 1$, let $A_{t+1:t+\tau} = (a_{t+1}, \ldots, a_{t+\tau})$ denote a future treatment sequence, and let $Y_{i,t+\tau}(\mathbf{a}_{t+1:t+\tau})$ denote the potential outcome at time $t+\tau$ under this sequence, with the treatment history up to time t held at its observed values.

Given the observed history $H_{i,t+1}$ and a future treatment sequence $\mathbf{a}_{t+1:t+\tau}$, our goal is to estimate the expected potential outcome at prediction horizon τ:

$$\mu_\tau(H_{i,t+1}, A_{t+1:t+\tau}) = \mathbb{E}[Y_{i,t+\tau}(A_{t+1:t+\tau}) \mid H_{i,t+1}]\,. \tag{2}$$

Evaluating this quantity at successive prediction horizons yields the expected counterfactual response trajectory. Following standard longitudinal causal inference (Robins, 1986; Melnychuk et al., 2022), we assume consistency, positivity, and sequential ignorability. Under these assumptions, the target is identifiable from observational data through the longitudinal g-formula, which accounts for the evolution of time-varying covariates and intermediate outcomes under the specified treatment sequence. Detailed assumptions are provided in Appendix A.

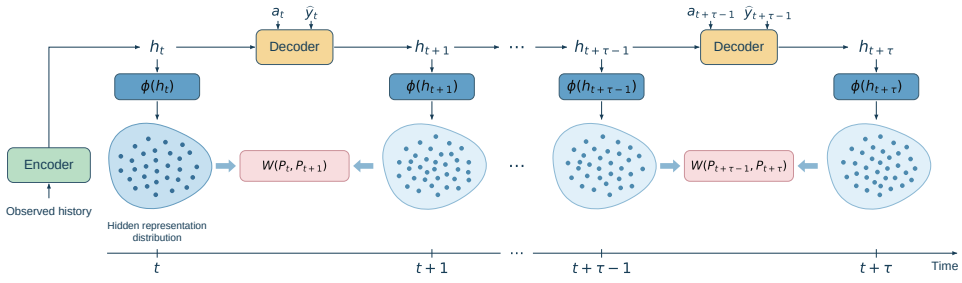


Figure 1: Overview of ReCT. In the figure, $P_t, \ldots, P_{t+\tau}$ denotes temporal representation distribution. To improve the generalization of longitudinal counterfactual regression models to future time steps, we reduce temporal drift by aligning the representation distributions across consecutive time steps.

4 METHOD

In this section, the impact of temporal distribution drift on counterfactual estimation is analyzed by treating representation distributions at different time steps as temporal domains. We derive an upper bound on the longitudinal counterfactual estimation error. Based on the theoretical analysis, a temporal representation learning algorithm is proposed to jointly mitigate time-varying confounding bias, temporal distribution drift, and counterfactual estimation error.

4.1 TEMPORAL DISTRIBUTION DRIFT

Given the observed history $H_{i,t+1}$ and a prescribed treatment sequence $\mathbf{a} = \mathbf{a}_{t+1:t+\tau}$, an autoregressive model estimates the counterfactual outcome at each future time step $k = t+1, \ldots, t+\tau$ using the previous representation, treatment, and predicted outcome (Bica et al., 2020; Bouchattaoui et al., 2024). The encoder E_θ maps the observed history to the initial representation $h_{i,t} = E_\theta(H_{i,t+1})$. The autoregressive update and outcome prediction are given by

$$\begin{aligned} h_{i,k} &= G_\theta\Big(h_{i,k-1}, a_{k-1}, \widehat{Y}_{i,k-1}\Big), \\ z_{i,k} &= \phi_\theta(h_{i,k}), \\ \widehat{Y}_{i,k} &= g_\theta(z_{i,k}, a_k), \end{aligned} \tag{3}$$

where G_θ is the decoder update function, ϕ_θ is the representation function, and g_θ is the outcome predictor.

For the population of observed histories, let P_k denote the overall distribution of the representations at time k, with $Z_k \sim P_k$. The distributions $P_t, \ldots, P_{t+\tau}$ are regarded as temporal domains. In this setting, the counterfactual regression model is trained on observations from preceding time steps and used to predict outcomes at future time steps, requiring the regression model to generalize across time steps. Therefore, the representation distributions $P_t, \ldots, P_{t+\tau}$ are expected to remain close to support accurate counterfactual estimation. Although individual representations evolve over time, temporal distribution alignment ensures that the longitudinal regression model generalizes well across time steps.

To analyze differences between consecutive temporal domains, let $m : \mathcal{Z} \to \mathbb{R}$ be a measurable function with finite expectations under P_k and P_{k-1}. The distribution discrepancy measured by m is defined as

$$\delta_k(m) = \left| \int_{\mathcal{Z}} m(z)\, dP_k(z) - \int_{\mathcal{Z}} m(z)\, dP_{k-1}(z) \right|. \tag{4}$$

The function m measures information contained in the representations, and $\delta_k(m)$ quantifies the change in the expected value of m across consecutive temporal domains.

To characterize the temporal discrepancy, we exploit optimal transport to measure the distance between representation distributions. Optimal transport seeks a transport plan that moves probability mass from one distribution to another with minimum cost and provides a powerful ability to compare distributions (Villani et al., 2009; Gabriel & Marco, 2019). For two probability distributions P and Q on $\mathcal{Z}$ with finite first moments, let $c(z, z') = \|z - z'\|_2$ be the transport cost. The Wasserstein distance is defined by the following Kantorovich problem (KP),

$$W(P, Q) = \inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{Z}\times\mathcal{Z}} c(z, z')\, d\pi(z, z'), \tag{5}$$

where π is a transport plan and $\Pi(P, Q)$ is the set of joint probability distributions with marginals P and Q. The transport plan specifies how probability mass is moved between representations, and $c(z, z')$ measures the cost per unit mass. The minimum total transport cost gives the Wasserstein distance between P and Q.

Based on this, we derive the following upper bound on the temporal distribution discrepancy.

Theorem 1. *Suppose that P_k and P_{k-1} have finite first moments and m is 1-Lipschitz, we have*

$$\delta_k(m) \leq W(P_k, P_{k-1}). \tag{6}$$

The proof is provided in Appendix B.1. Theorem 1 shows that the discrepancy between consecutive temporal domains is upper bounded by the Wasserstein distance between the corresponding representation distributions. The distance $W(P_k, P_{k-1})$ can be estimated from autoregressively generated representations and minimized during training to reduce the upper bound on the temporal discrepancy.

In addition to the temporal distribution drift, we further analyze the counterfactual estimation error over the population, which is expected to be minimized to obtain accurate counterfactual regression. For a treatment value a, let $\ell_k^a(z) = \ell(\widehat{\mu}_k^a(z), \mu_k^a(z))$ denote the pointwise prediction loss at time step k, where $\widehat{\mu}_k^a(z) = g_\theta(z, a)$ is the predicted function and $\mu_k^a(z)$ is the true response function. The counterfactual estimation error at time step k under the overall representation distribution P_k is defined as

$$\epsilon_{P_k}^a = \int_{\mathcal{Z}} \ell_k^a(z)\, dP_k(z). \tag{7}$$

To connect temporal distribution alignment with counterfactual prediction, we establish the following error bound in terms of the temporal distribution discrepancy.

Theorem 2. *For a treatment value a. Suppose that P_k and P_{k-1} have finite first moments and ℓ_k^a is 1-Lipschitz on the common representation space, we have*

$$\epsilon_{P_k}^a \leq \epsilon_{P_{k-1}}^a + W(P_k, P_{k-1}). \tag{8}$$

The proof is provided in Appendix B.2. Theorem 2 shows that the prediction error in the subsequent temporal domain is upper bounded by the prediction error in the preceding temporal domain and the Wasserstein distance between the two representation distributions. The bound therefore motivates aligning representation distributions across temporal domains to constrain prediction errors at future time steps.

4.2 Counterfactual Estimation Error

Since the counterfactual outcomes under non-observed treatments can never be obtained, we cannot directly minimize the counterfactual estimation error in Eq. 7. Instead, we minimize the factual prediction error under the treatment-conditional representation distribution (Yan et al., 2025). Formally, given a treatment value $a \in \mathcal{A}$, let $P_{k,a}$ denote the representation distribution conditional on the observed treatment $A_k = a$, with $(Z_k \mid A_k = a) \sim P_{k,a}$. The corresponding treatment-conditional prediction error is defined as

$$\epsilon_{P_{k,a}}^a = \int_{\mathcal{Z}} \ell_k^a(z)\, dP_{k,a}(z). \tag{9}$$

Due to the confounding bias, the treatment-conditional distribution differs from the population distribution, *i.e.*, $P_k \neq P_{k,a}$. Moreover, prediction errors induced by temporal distribution drift must also be considered in the longitudinal setting. To address this issue in longitudinal counterfactual regression, we establish that the future counterfactual outcome estimation error is upper bounded by the current factual error and the distribution discrepancies across both treatment groups and time steps. Let $\pi_k(a) = \Pr(A_k = a)$ denote the treatment distribution at time k. The overall distribution P_k is the expectation of the conditional distributions $P_{k,a}$ with probabilities $\pi_k(a)$. Now, we analyze the expected future counterfactual estimation error with the following theorem.

Theorem 3. *Let $\mathcal{A}$ be a finite treatment space with $\pi_{k-1}(a) > 0$ for every $a \in \mathcal{A}$. Suppose that the representation distributions have finite first moments and that ℓ_{k-1}^a and ℓ_k^a are 1-Lipschitz for every $a \in \mathcal{A}$, We have*

$$\mathbb{E}_{a\sim\pi_k}\left[\epsilon_{P_k}^a\right] \leq \mathbb{E}_{a\sim\pi_{k-1}}\left[\epsilon_{P_{k-1,a}}^a\right] + \mathbb{E}_{a,b\sim\pi_{k-1}}\left[W(P_{k-1,a}, P_{k-1,b})\right] + W(P_k, P_{k-1}). \tag{10}$$

The proof is provided in Appendix B.3. Theorem 3 bounds the treatment-averaged prediction error at time k by the factual prediction error at time $k-1$, the distances between treatment-group representation distributions at time $k-1$, and the distance between consecutive temporal domains. The treatment-group term controls the discrepancy between conditional and overall prediction errors and the temporal domain alignment term controls the change in overall prediction error across consecutive time steps.

In practice, the empirical distributions of individuals can be implemented using discrete empirical measures (Courty et al., 2016; Gabriel & Marco, 2019). For n independent observed trajectories, we define the empirical distributions as

$$\widehat{P}_{k-1} = \frac{1}{n}\sum_{i=1}^{n} \nu_{z_{i,k-1}}, \qquad \widehat{P}_{k} = \frac{1}{n}\sum_{i=1}^{n} \nu_{z_{i,k}}, \tag{11}$$

where ν_z denotes the Dirac measure at z. The empirical distributions $\widehat{P}_{k-1,a}$ and $\widehat{P}_{k-1,b}$ are constructed analogously using representations from the corresponding treatment groups. And the counterfactual estimation error bound in terms of the empirical distributions is given as follows.

Theorem 4. *Let n be the number of independent observed trajectories. For any $\rho \in (0,1)$, with probability at least $1-\rho$,*

$$\begin{aligned}\mathbb{E}_{a\sim\pi_k}\left[\epsilon^a_{P_k}\right] \le \mathbb{E}_{a\sim\pi_{k-1}}\left[\epsilon^a_{P_{k-1,a}}\right] + \mathbb{E}_{a,b\sim\pi_{k-1}}\left[W\left(\widehat{P}_{k-1,a}, \widehat{P}_{k-1,b}\right)\right] \\ + W\left(\widehat{P}_k, \widehat{P}_{k-1}\right) + O\left(\sqrt{\frac{\log(2/\rho)}{n}}\right).\end{aligned} \tag{12}$$

Theorem 4 bounds the population prediction error using empirical treatment-group alignment and empirical temporal alignment. The finite-sample term accounts for sampling errors in the overall and treatment-conditional representation distributions. For a fixed confidence level and a fixed number of treatment groups, the finite-sample term vanishes as the sample size within every treatment group increases. The bound therefore motivates combining factual outcome supervision with minimization of the empirical treatment-group discrepancy $W(\widehat{P}_{k-1,a}, \widehat{P}_{k-1,b})$ and the empirical temporal discrepancy $W(\widehat{P}_k, \widehat{P}_{k-1})$.

4.3 Algorithm

Motivated by Theorem 4, we jointly minimize the factual prediction loss, the distances between treatment-group representation distributions, and the distances between representation distributions at consecutive time steps. In the following, we discuss the training objectives of our method.

Factual prediction loss. To optimize the factual prediction error $\mathbb{E}_{a\sim\pi_k}[\epsilon^a_{P_{k,a}}]$ in Theorem 4, we train the model over τ future time steps using the factual outcomes $Y_{i,t+1}, \ldots, Y_{i,t+\tau}$. Formally, we minimize the mean squared error as follows

$$\mathcal{L}_Y = \sum_{k=t+1}^{t+\tau} \left\| Y_{i,k} - \widehat{Y}_{i,k} \right\|_2^2, \tag{13}$$

where $\widehat{Y}_{i,k} = g_\theta(z_{i,k}, a_k)$ is the outcome prediction under the observed treatment.

Treatment prediction loss. Now, we minimize the distances across treatment groups shown in the second term $\mathbb{E}_{a,b\sim\pi_k}[W(\widehat{P}_{k,a}, \widehat{P}_{k,b})]$ in Theorem 4. To avoid minimizing the distances between all the group pairs, we adopt adversarial learning to learn representations that are invariant across treatment groups. To this end, we introduce a treatment classifier d_ψ, where $d_\psi(a \mid z_{i,k})$ denotes the predicted probability of treatment a. The treatment classifier is trained by minimizing the following loss

$$\mathcal{L}_{\text{adv}} = -\sum_{k=t+1}^{t+\tau} \sum_{a\in\mathcal{A}} \mathbf{1}\{A_{i,k} = a\} \log d_\psi(a \mid z_{i,k}), \tag{14}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. The treatment classifier minimizes $\mathcal{L}_{\text{adv}}$ to predict treatment assignment, while the representation network maximizes $\mathcal{L}_{\text{adv}}$ through gradient reversal to remove treatment information from the representations. By doing this, the representation distributions of different treatment groups are aligned, minimizing the second term at the right hand in Theorem 4.

Wasserstein distances of temporal representation distribution. To reduce the temporal distribution discrepancy $\sum_{k=t+1}^{t+\tau} W(\widehat{P}_k, \widehat{P}_{k-1})$ in Theorem 4, we learn representations by minimizing the optimal transport cost between representation distributions at consecutive time steps.

Specifically, at the time steps k and $k-1$, let $p_i^{(k)}$ and $p_j^{(k-1)}$ denote the probability masses assigned to the representations of the individual i and j, respectively. The transport cost between them is measured in the representation space as $[C_\theta^{(k)}]_{ij} = \|z_{i,k} - z_{j,k-1}\|_2 = \|\phi_\theta(h_{i,k}) - \phi_\theta(h_{j,k-1})\|_2$. And $\pi_{ij}^{(k)}$ indicates the probability mass transported from $z_{i,k}$ to $z_{j,k-1}$. The set of admissible transport plans is

$$\Pi_k = \left\{ \pi^{(k)} \in \mathbb{R}_+^{n\times n} \;\middle|\; \sum_{j=1}^{n} \pi_{ij}^{(k)} = \frac{1}{n},\ \forall i, \quad \sum_{i=1}^{n} \pi_{ij}^{(k)} = \frac{1}{n},\ \forall j \right\}. \tag{15}$$

The optimal transport plan between the representation distributions at time steps k and $k-1$ is obtained by solving the following problem.

$$\widetilde{\pi}^{(k)} = \underset{\pi^{(k)}\in\Pi_k}{\arg\min} \left\{ \sum_{i=1}^{n}\sum_{j=1}^{n} [C_\theta^{(k)}]_{ij} \pi_{ij}^{(k)} + \varepsilon \Omega(\pi^{(k)}) \right\}, \tag{16}$$

where $\Omega(\pi) = \sum_{i=1}^{n}\sum_{j=1}^{n} \pi_{ij} \log \pi_{ij}$ is the negative entropy, and $\varepsilon > 0$ controls the regularization strength. The regularized optimal transport problem can be efficiently solved by the Sinkhorn algorithm (Sinkhorn & Knopp, 1967; Cuturi, 2013; Gabriel & Marco, 2019).

Based on the resulting optimal transport plan, we compute the optimal transport cost between consecutive representation distributions as

$$\widehat{W}_k = \sum_{i=1}^{n}\sum_{j=1}^{n} [C_\theta^{(k)}]_{ij} \widetilde{\pi}_{ij}^{(k)}, \tag{17}$$

and minimize the temporal distribution alignment loss over the prediction horizon defined as follows

$$\mathcal{L}_{\text{tem}} = \sum_{k=t+1}^{t+\tau} \widehat{W}_k. \tag{18}$$

Minimizing $\mathcal{L}_{\text{tem}}$ with respect to θ reduces the Wasserstein distances between consecutive representation distributions and targets the observable temporal component of the upper bound in Theorem 4.

Joint optimization. Let η denote the parameters of the history encoder, transition model, and representation map, θ the outcome predictor parameters, and ψ the treatment discriminator parameters. The overall training objective is

$$\min_{\theta,\eta} \max_{\psi} \quad \mathcal{L}_Y(\theta,\eta) - \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}(\theta,\psi) + \lambda_{\text{tem}} \mathcal{L}_{\text{tem}}(\theta), \tag{19}$$

where λ_{adv} and λ_{tem} control the trade-offs between the factual prediction loss and the distribution discrepancies across treatment groups and consecutive time steps, respectively. The joint objective learns representations that preserve outcome information while reducing distribution discrepancies across treatment groups and consecutive time steps. The training procedure is summarized in Algorithm 1 in Appendix E.

5 Experiments

We evaluate the proposed method ReCT on synthetic tumor-growth data (Geng et al., 2017), semi-synthetic MIMIC-III data (Melnychuk et al., 2022), and real-world MIMIC-III data (Johnson et al., 2016). We compare ReCT with classic and state-of-the-art methods: MSM (Robins et al., 2000), RMSN (Lim, 2018), G-Net (Li et al., 2021), CRN (Bica et al., 2020), CT (Melnychuk et al., 2022), and Causal CPC (Bouchattaoui et al., 2024). Performance is measured by normalized root mean square error (NRMSE) on synthetic tumor-growth data and root mean square error (RMSE) on the semi-synthetic and real-world data. Further details on data generation and additional experiments are provided in Appendix D, while the computational details of evaluation metrics is detailed in Appendix C.

					$\gamma = 1$					
Model	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
MSM	1.15 ± 0.09	1.93 ± 0.29	2.42 ± 0.37	2.76 ± 0.39	3.00 ± 0.40	3.16 ± 0.41	3.26 ± 0.41	3.32 ± 0.42	3.35 ± 0.42	3.34 ± 0.41
RMSN	0.96 ± 0.17	1.03 ± 0.07	1.11 ± 0.11	1.23 ± 0.13	1.35 ± 0.15	1.46 ± 0.17	1.56 ± 0.19	1.65 ± 0.20	1.73 ± 0.22	1.79 ± 0.23
CRN	$\mathbf{0.90 \pm 0.14}$	1.13 ± 0.33	1.23 ± 0.39	1.36 ± 0.48	1.49 ± 0.57	1.61 ± 0.63	1.72 ± 0.67	1.80 ± 0.68	1.87 ± 0.68	1.93 ± 0.66
G-Net	0.96 ± 0.16	1.24 ± 0.20	1.44 ± 0.21	1.59 ± 0.28	1.71 ± 0.33	1.82 ± 0.36	1.91 ± 0.41	1.99 ± 0.45	2.06 ± 0.50	2.12 ± 0.55
CT	1.06 ± 0.06	1.23 ± 0.13	1.36 ± 0.14	1.48 ± 0.15	1.56 ± 0.16	1.62 ± 0.16	1.67 ± 0.18	1.71 ± 0.20	1.74 ± 0.22	1.77 ± 0.26
Causal CPC	1.20 ± 0.09	1.21 ± 0.10	1.32 ± 0.11	1.42 ± 0.13	1.51 ± 0.14	1.60 ± 0.14	1.69 ± 0.15	1.75 ± 0.16	1.81 ± 0.18	1.85 ± 0.20
ReCT (ours)	0.97 ± 0.11	$\mathbf{1.02 \pm 0.11}$	$\mathbf{1.08 \pm 0.11}$	$\mathbf{1.15 \pm 0.12}$	$\mathbf{1.21 \pm 0.12}$	$\mathbf{1.28 \pm 0.12}$	$\mathbf{1.34 \pm 0.12}$	$\mathbf{1.40 \pm 0.12}$	$\mathbf{1.45 \pm 0.12}$	$\mathbf{1.48 \pm 0.12}$
					$\gamma = 2$					
MSM	1.36 ± 0.13	2.58 ± 0.34	3.17 ± 0.42	3.56 ± 0.45	3.82 ± 0.46	3.98 ± 0.48	4.08 ± 0.48	4.13 ± 0.49	4.14 ± 0.49	4.11 ± 0.49
RMSN	$\mathbf{1.01 \pm 0.08}$	$\mathbf{1.10 \pm 0.07}$	1.35 ± 0.11	1.59 ± 0.15	1.82 ± 0.19	2.03 ± 0.22	2.22 ± 0.26	2.37 ± 0.31	2.51 ± 0.37	2.61 ± 0.44
CRN	1.16 ± 0.12	1.48 ± 0.22	1.76 ± 0.30	1.97 ± 0.36	2.12 ± 0.41	2.25 ± 0.44	2.35 ± 0.46	2.42 ± 0.47	2.48 ± 0.46	2.52 ± 0.44
G-Net	1.05 ± 0.12	1.22 ± 0.22	1.39 ± 0.13	1.51 ± 0.07	1.62 ± 0.07	1.72 ± 0.09	1.81 ± 0.12	1.89 ± 0.15	1.96 ± 0.17	2.11 ± 0.19
CT	1.13 ± 0.07	1.38 ± 0.10	1.59 ± 0.14	1.75 ± 0.17	1.88 ± 0.18	1.97 ± 0.19	2.03 ± 0.22	2.08 ± 0.23	2.12 ± 0.23	2.14 ± 0.25
Causal CPC	1.33 ± 0.22	1.57 ± 0.40	1.71 ± 0.40	1.80 ± 0.39	1.88 ± 0.39	1.97 ± 0.38	2.05 ± 0.36	2.14 ± 0.34	2.21 ± 0.32	2.26 ± 0.31
ReCT (ours)	1.06 ± 0.24	1.14 ± 0.06	$\mathbf{1.27 \pm 0.09}$	$\mathbf{1.39 \pm 0.12}$	$\mathbf{1.51 \pm 0.14}$	$\mathbf{1.63 \pm 0.12}$	$\mathbf{1.75 \pm 0.11}$	$\mathbf{1.85 \pm 0.11}$	$\mathbf{1.95 \pm 0.15}$	$\mathbf{2.03 \pm 0.20}$

Table 1: NRMSE on the synthetic tumor-growth dataset with sequence length 60 (mean $\pm$ standard deviation). Lower is better. Bold values indicate the best result at each horizon.

Model	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
RMSN	0.39 ± 0.07	0.59 ± 0.08	0.70 ± 0.10	0.80 ± 0.13	0.87 ± 0.19	0.93 ± 0.24	0.97 ± 0.28	1.02 ± 0.30	1.06 ± 0.31	1.09 ± 0.32
CRN	0.40 ± 0.09	0.48 ± 0.08	0.61 ± 0.10	0.70 ± 0.13	0.78 ± 0.16	0.85 ± 0.19	0.91 ± 0.23	0.96 ± 0.27	1.01 ± 0.31	1.06 ± 0.35
G-Net	0.41 ± 0.06	0.52 ± 0.05	0.67 ± 0.07	0.78 ± 0.10	0.87 ± 0.13	0.94 ± 0.16	1.00 ± 0.19	1.05 ± 0.21	1.10 ± 0.22	1.14 ± 0.24
CT	$\mathbf{0.26 \pm 0.02}$	$\mathbf{0.36 \pm 0.02}$	$\mathbf{0.46 \pm 0.05}$	0.54 ± 0.08	0.57 ± 0.11	0.62 ± 0.14	0.67 ± 0.15	0.71 ± 0.18	0.75 ± 0.19	0.79 ± 0.22
Causal CPC	0.40 ± 0.05	0.46 ± 0.04	0.51 ± 0.05	0.55 ± 0.06	0.59 ± 0.07	0.62 ± 0.07	0.65 ± 0.08	0.67 ± 0.08	0.69 ± 0.09	0.71 ± 0.10
ReCT (ours)	0.36 ± 0.04	0.42 ± 0.03	0.48 ± 0.03	$\mathbf{0.52 \pm 0.04}$	$\mathbf{0.55 \pm 0.05}$	$\mathbf{0.58 \pm 0.06}$	$\mathbf{0.60 \pm 0.07}$	$\mathbf{0.62 \pm 0.08}$	$\mathbf{0.64 \pm 0.09}$	$\mathbf{0.66 \pm 0.10}$

Table 2: RMSE on the semi-synthetic MIMIC-III dataset with sequence length 100 (mean $\pm$ standard deviation). Lower is better. Bold values indicate the best result at each horizon.

5.1 Synthetic Tumor-Growth Data

Dataset and experimental setup. We evaluate ReCT on a synthetic tumor-growth benchmark based on the model of Geng et al. (2017), which is widely used for longitudinal causal inference (Lim, 2018; Bica et al., 2020; Melnychuk et al., 2022). This benchmark provides ground-truth outcomes under alternative treatment sequences and allows the strength of time-varying confounding to be controlled through γ, with larger values inducing stronger history-dependent treatment assignment. It therefore enables a controlled assessment of counterfactual prediction across both confounding strengths and prediction horizons. We consider $\gamma \in \{1, 2, 3, 4\}$, with 1,000 training trajectories, 500 test trajectories, a maximum sequence length of 60, and prediction horizons $\tau = 1, \ldots, 10$.

Results. Table 1 presents the prediction performance across different horizons and confounding levels. ReCT achieves the best performance at most multi-step horizons and maintains strong long-term prediction performance across the evaluated settings. Compared with the baselines, ReCT maintains a clear advantage at longer prediction horizons, even under stronger confounding. We attribute this improvement to its joint control of treatment-induced imbalance and temporal representation drift. While existing methods mainly balance treatment groups, ReCT also aligns representations across consecutive time steps, limiting distribution shift and error accumulation during multi-step prediction. This allows the outcome predictor to operate on more stable representations and produce more robust long-term counterfactual estimation. We further evaluate performance under stronger confounding ($\gamma \in \{3, 4\}$), with results reported in Appendix D.1.2.

5.2 Semi-synthetic and Real-world MIMIC-III Data

Semi-synthetic data. We use the semi-synthetic MIMIC-III benchmark following (Melnychuk et al., 2022), which combines covariate trajectories from real intensive care unit patients with simulated treatments and outcomes, which allows evaluation on higher-dimensional clinical histories with complex temporal dependencies while retaining ground-truth counterfactual outcomes. We use 1,000/200/200 trajectories for training, validation, and testing, respectively, with a maximum sequence length of 100 and prediction horizons $\tau = 1, \ldots, 10$.

Results. Table 2 shows that ReCT outperforms all baselines at longer prediction horizons on the semi-synthetic MIMIC-III benchmark, with the performance gaps generally widening as the horizon increases. These results demonstrate the effectiveness of ReCT for long-term counterfactual estimation in settings with high-dimensional clinical covariates.

Real-world data. We further evaluate ReCT on real-world MIMIC-III records (Johnson et al., 2016) using observed treatments and outcomes. Since ground-truth counterfactuals are unavailable, this evaluation measures factual prediction accuracy. As shown in Figure 2, ReCT achieves the lowest reported RMSE across all evaluated horizons. Causal CPC remains the closest competitor, while the gaps relative to CT, RMSN, and G-Net widen over longer horizons, supporting ReCT's predictive performance on observed clinical trajectories.

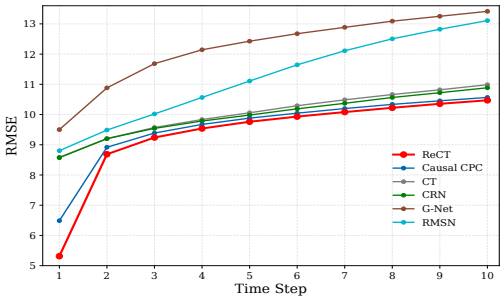


Figure 2: RMSEs on the real-world MIMIC III dataset with sequence length 100.

5.3 Ablation Study

We evaluate four configurations on the synthetic and semi-synthetic benchmarks: (i) the full ReCT model, (ii) ReCT without $\mathcal{L}_{\text{tem}}$, (iii) ReCT without $\mathcal{L}_{\text{adv}}$, and (iv) ReCT without either loss.

Synthetic Data (NRMSE ↓)

Model	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
ReCT w/o $\mathcal{L}_{\text{tem}}$ and $\mathcal{L}_{\text{adv}}$	1.13 ± 0.10	1.32 ± 0.36	1.46 ± 0.43	1.58 ± 0.47	1.69 ± 0.48	1.79 ± 0.47	1.88 ± 0.45	1.96 ± 0.43	2.03 ± 0.41	2.08 ± 0.39
ReCT w/o $\mathcal{L}_{\text{tem}}$	1.20 ± 0.09	1.21 ± 0.10	1.32 ± 0.11	1.42 ± 0.13	1.51 ± 0.14	1.60 ± 0.14	1.69 ± 0.15	1.75 ± 0.16	1.81 ± 0.18	1.85 ± 0.20
ReCT w/o $\mathcal{L}_{\text{adv}}$	1.13 ± 0.08	1.28 ± 0.18	1.39 ± 0.16	1.49 ± 0.15	1.60 ± 0.17	1.70 ± 0.19	1.78 ± 0.21	1.85 ± 0.23	1.90 ± 0.25	1.93 ± 0.27
ReCT	$\mathbf{0.97 \pm 0.11}$	$\mathbf{1.02 \pm 0.11}$	$\mathbf{1.08 \pm 0.11}$	$\mathbf{1.15 \pm 0.12}$	$\mathbf{1.21 \pm 0.12}$	$\mathbf{1.28 \pm 0.12}$	$\mathbf{1.34 \pm 0.12}$	$\mathbf{1.40 \pm 0.12}$	$\mathbf{1.45 \pm 0.12}$	$\mathbf{1.48 \pm 0.12}$

Semi-synthetic MIMIC-III Data (RMSE ↓)

Model	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
ReCT w/o $\mathcal{L}_{\text{tem}}$ and $\mathcal{L}_{\text{adv}}$	0.42 ± 0.06	0.47 ± 0.04	0.54 ± 0.07	0.60 ± 0.10	0.65 ± 0.13	0.69 ± 0.15	0.72 ± 0.17	0.75 ± 0.19	0.77 ± 0.20	0.80 ± 0.21
ReCT w/o $\mathcal{L}_{\text{tem}}$	0.40 ± 0.05	0.46 ± 0.04	0.51 ± 0.05	0.55 ± 0.06	0.59 ± 0.07	0.62 ± 0.07	0.65 ± 0.08	0.67 ± 0.08	0.69 ± 0.09	0.71 ± 0.10
ReCT w/o $\mathcal{L}_{\text{adv}}$	0.39 ± 0.03	0.44 ± 0.01	0.51 ± 0.03	0.56 ± 0.05	0.61 ± 0.08	0.64 ± 0.11	0.68 ± 0.14	0.71 ± 0.17	0.74 ± 0.20	0.77 ± 0.23
ReCT	$\mathbf{0.36 \pm 0.04}$	$\mathbf{0.42 \pm 0.03}$	$\mathbf{0.48 \pm 0.03}$	$\mathbf{0.52 \pm 0.04}$	$\mathbf{0.55 \pm 0.05}$	$\mathbf{0.58 \pm 0.06}$	$\mathbf{0.60 \pm 0.07}$	$\mathbf{0.62 \pm 0.08}$	$\mathbf{0.64 \pm 0.09}$	$\mathbf{0.66 \pm 0.10}$

Table 3: Ablation results (mean $\pm$ standard deviation). $\mathcal{L}_{\text{tem}}$ denotes temporal distribution alignment loss, $\mathcal{L}_{\text{adv}}$ denotes balancing loss. Lower is better and bold indicates the better result within each pair.

Table 3 shows that the complete ReCT model achieves the lowest prediction error across all evaluated horizons on both benchmarks. Removing the temporal distribution alignment loss increases prediction error, demonstrating that mitigating temporal distribution drift improves the model's generalization across time steps. Removing the balancing loss also increases prediction error, underscoring its role in mitigating confounding bias through treatment balancing. Together, these results support the complementary roles of the two loss and show that incorporating temporal alignment into the balanced model is essential for achieving the best performance among the evaluated configurations.

6 Conclusion

In this paper, we study temporal distribution drift in longitudinal counterfactual regression and its effect on generalization across time steps. We theoretically establish that future counterfactual outcome estimation error can be upper-bounded by the current factual error and distribution discrepancies across treatment groups and time steps. Inspired by this analysis, we propose ReCT to learn balanced representations across treatment groups and consecutive time steps. Experiments on benchmark datasets show improved counterfactual prediction at longer horizons, and evaluations with different backbones demonstrate the generality of temporal distribution alignment across network architectures.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing and to identify relevant literature. We have reviewed all AI-assisted work and verified the identified references against their original sources. We take responsibility for the final content of this work, including text and claims produced with the aid of generative AI.

REFERENCES

Ahmed M Alaa, Michael Weisz, and Mihaela Van Der Schaar. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*, 2017.

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

François Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3–4): 541–593, 2007. doi: 10.1007/s00440-006-0004-7.

Mouad E Bouchattaoui, Myriam Tami, Benoit Lepetit, and Paul-Henry Cournède. Causal contrastive learning for counterfactual regression over time. *Advances in Neural Information Processing Systems*, 37:1333–1369, 2024.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1724–1734, 2014.

Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664, 2008.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Alicia Curth and Mihaela Van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International conference on artificial intelligence and statistics*, pp. 1810–1818. PMLR, 2021.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Yuntao Du, Jindong Wang, Wenjie Feng, Sinno Pan, Tao Qin, Renjun Xu, and Chongjun Wang. Adarnn: Adaptive learning and forecasting of time series. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 402–411, 2021.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Changran Geng, Harald Paganetti, and Clemens Grassberger. Prediction of treatment response for combined chemo-and radiation therapy for non-small cell lung cancer patients using a biomathematical model. *Scientific reports*, 7(1):13542, 2017.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Qiang Huang, Chuizheng Meng, Defu Cao, Biwei Huang, Yi Chang, and Yan Liu. An empirical examination of balancing strategy for counterfactual estimation on time series. *arXiv preprint arXiv:2408.08815*, 2024.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):160035, 2016.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Vitaly Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. *Advances in neural information processing systems*, 28, 2015.

Rui Li, Stephanie Hu, Mingyu Lu, Yuria Utsumi, Prithwish Chakraborty, Daby M Sow, Piyush Madan, Jun Li, Mohamed Ghalwash, Zach Shahn, et al. G-net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In *Machine Learning for Health*, pp. 282–299. PMLR, 2021.

Bryan Lim. Forecasting treatment responses over time using recurrent marginal structural networks. *Advances in neural information processing systems*, 31, 2018.

Yiling Liu, Juncheng Dong, Chen Fu, Wei Shi, Ziyang Jiang, Zhigang Hua, and David Carlson. Synergizing deconfounding and temporal generalization for time-series counterfactual outcome estimation. *arXiv preprint arXiv:2511.16006*, 2025.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International conference on machine learning*, pp. 15293–15329. PMLR, 2022.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):331–355, 2003.

Tiexin Qin, Shiqi Wang, and Haoliang Li. Generalizing to evolving domains with latent structure-aware sequential autoencoder. In *International Conference on Machine Learning*, pp. 18062–18082. PMLR, 2022.

Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 737–753. Springer, 2017.

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.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pp. 34–58, 1978.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Xin Wang, Shengfei Lyu, Lishan Yang, Yibing Zhan, and Huanhuan Chen. A dual-module framework for counterfactual estimation over time. In *Forty-first International Conference on Machine Learning*, 2024.

Yuguang Yan, Zeqin Yang, Weilin Chen, Ruichu Cai, Zhifeng Hao, and Michael Kwok-Po Ng. Exploiting geometry for treatment effect estimation via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16290–16298, 2024.

Yuguang Yan, Zongyu Li, Haolin Yang, Zeqin Yang, Hao Zhou, Ruichu Cai, and Zhifeng Hao. Reducing confounding bias without data splitting for causal inference via optimal transport. In *Forty-second International Conference on Machine Learning*, 2025.

A Causal Assumptions and Identification

A.1 Identifiability Assumptions and Target

Following the longitudinal potential outcomes framework (Robins, 1986), we state the standard assumptions used to identify counterfactual outcomes from longitudinal observational data (Bica et al., 2020; Melnychuk et al., 2022; Bouchattaoui et al., 2024). We omit the individual index i for clarity. Consistent with Section 3, the history available before treatment assignment at time t is

$$H_t = \big[V, X_{1:t}, A_{1:t-1}, Y_{1:t-1}\big]. \tag{20}$$

Thus, X_t is observed before treatment A_t is assigned, and Y_t is observed after treatment assignment.

Assumption 1 (Consistency)**.** *For any treatment sequence $\mathbf{a}_{1:t}$, if the observed treatment sequence satisfies $A_{1:t} = \mathbf{a}_{1:t}$, then*

$$Y_t = Y_t(\mathbf{a}_{1:t}). \tag{21}$$

That is, the potential outcome under the observed treatment sequence equals the factual outcome. We also assume that treatments are well defined and that one individual's outcomes are not affected by treatments assigned to other individuals.

Assumption 2 (Sequential ignorability)**.** *At each time t, treatment assignment is independent of future potential outcomes conditional on the observed history. For every treatment sequence $\mathbf{a}_{t:t+\tau}$ under consideration and every $\tau \geq 0$ within the observation period,*

$$A_t \perp\!\!\!\perp Y_{t+\tau}(\mathbf{a}_{t:t+\tau}) \mid H_t. \tag{22}$$

Here, treatments before time t are held at their observed values. This assumption requires that the observed history H_t be sufficient to account for confounding between current treatment assignment and subsequent outcomes.

Assumption 3 (Positivity)**.** *At each time t, every treatment $a \in \mathcal{A}$ has positive probability of being assigned, conditional on almost every history h in the support of H_t. Formally,*

$$\Pr(A_t = a \mid H_t = h) > 0, \qquad a \in \mathcal{A}. \tag{23}$$

This assumption ensures that the treatment assignments required by the interventions under consideration are supported by the observational data.

Identification target. Given the observed history $H_{t+1} = h$ and a prescribed treatment sequence $\mathbf{a}_{t+1:t+\tau}$, our target is

$$\mu_\tau\big(h, \mathbf{a}_{t+1:t+\tau}\big) = \mathbb{E}[Y_{t+\tau}(\mathbf{a}_{t+1:t+\tau}) \mid H_{t+1} = h], \qquad \tau \geq 1. \tag{24}$$

The treatment history up to time t is held at its observed values. Under Assumptions 1–3, this conditional mean is identifiable from observational data through the longitudinal g-formula, which accounts for the evolution of time-varying covariates and intermediate outcomes under the prescribed treatment sequence (Robins, 1986; Li et al., 2021; Melnychuk et al., 2022).

A.2 Causal Directed Acyclic Graphs in Longitudinal Settings

Figure 3 illustrates the causal structure of the longitudinal setting, following the graphical perspective in (Liu et al., 2025). Static covariates V, current covariates X_t, and past treatments and outcomes may affect both treatment assignment A_t and outcome Y_t. Treatments may also affect subsequent covariates, creating treatment–confounder feedback over time.

For example, the path $A_t \to X_{t+1} \to A_{t+1}$ shows how an earlier treatment can influence a later treatment through an intermediate covariate. When X_{t+1} also affects Y_{t+1}, it acts as a confounder for the later treatment while being affected by the earlier treatment (Li et al., 2021).

Under sequential ignorability, conditioning on H_t is sufficient to account for confounding at each treatment decision. The graph respects the temporal ordering of covariate measurement, treatment assignment, and outcome observation. Positivity additionally requires that the treatment assignments under consideration have positive probability at the relevant histories.

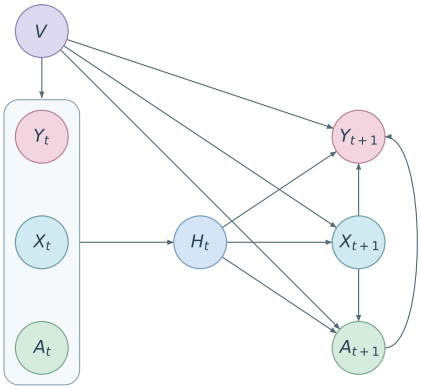


Figure 3: **Causal structure over time.** V denotes static covariates, X_t observed covariates, A_t treatment, and Y_t outcome. Directed edges represent causal dependencies. Earlier treatments can affect later covariates, which may influence subsequent treatment assignments and outcomes.

B Theoretical Analysis and Proofs

Following the notation in Sections 4.1 and 4.2.

B.1 Proof of Theorem 1

Let P_k denote the overall representation distribution at time k, and let $P_{k,a}$ denote the representation distribution conditional on $A_k = a$. For a function $m : \mathcal{Z} \to \mathbb{R}$, the discrepancy between consecutive temporal domains is defined as

$$\delta_k(m) = \left| \int_{\mathcal{Z}} m(z)\, dP_k(z) - \int_{\mathcal{Z}} m(z)\, dP_{k-1}(z) \right|. \tag{25}$$

We use the Wasserstein distance W with transport cost $c(z, z') = \|z - z'\|_2$ to bound this discrepancy.

Theorem 1. Suppose that P_k and P_{k-1} have finite first moments and m is 1-Lipschitz. Then

$$\delta_k(m) \le W(P_k, P_{k-1}). \tag{26}$$

To prove the theorem, we introduce the dual formulation of the Wasserstein distance in the following lemma.

Lemma 1 (Kantorovich–Rubinstein duality (Gabriel & Marco, 2019))**.** *For probability distributions P and Q with finite first moments,*

$$\begin{aligned} \mathrm{DP}(P, Q) &= \sup_{\|f\|_{\mathrm{Lip}} \le 1} \left| \int_{\mathcal{Z}} f(z)\, dP(z) - \int_{\mathcal{Z}} f(z)\, dQ(z) \right| \\ &= \inf_{\gamma \in \Pi(P,Q)} \int_{\mathcal{Z} \times \mathcal{Z}} \|z - z'\|_2 \, d\gamma(z, z') \\ &= W(P, Q), \end{aligned} \tag{27}$$

where $\Pi(P, Q)$ is the set of couplings with marginals P and Q, and the supremum is taken over all 1-Lipschitz functions.

Proof of Theorem 1. By the definition of $\delta_k(m)$ and Lemma 1,

$$\begin{aligned}\delta_k(m) &= \left| \int_{\mathcal{Z}} m(z)\, dP_k(z) - \int_{\mathcal{Z}} m(z)\, dP_{k-1}(z) \right| \\ &\leq \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \left| \int_{\mathcal{Z}} f(z)\, dP_k(z) - \int_{\mathcal{Z}} f(z)\, dP_{k-1}(z) \right| \\ &= \inf_{\gamma \in \Pi(P_k, P_{k-1})} \int_{\mathcal{Z}\times\mathcal{Z}} \|z - z'\|_2 \, d\gamma(z, z') \\ &= W(P_k, P_{k-1}).\end{aligned} \tag{28}$$

The inequality follows from the 1-Lipschitz property of m, and the equality between the supremum and the infimum follows from Lemma 1. □

B.2 Proof of Theorem 2

Theorem 2. Suppose that P_k and P_{k-1} have finite first moments and ℓ_k^a is 1-Lipschitz, we have

$$\epsilon^a_{P_k} \leq \epsilon^a_{P_{k-1}} + W(P_k, P_{k-1}). \tag{29}$$

Proof. Let $\gamma \in \Pi(P_k, P_{k-1})$ be an arbitrary coupling. Using the equality of the pointwise losses and the marginal constraints of γ, we obtain

$$\begin{aligned}\epsilon^a_{P_k} - \epsilon^a_{P_{k-1}} &= \int_{\mathcal{Z}} \ell_k^a(z)\, dP_k(z) - \int_{\mathcal{Z}} \ell_{k-1}^a(z')\, dP_{k-1}(z') \\ &= \int_{\mathcal{Z}} \ell_k^a(z)\, dP_k(z) - \int_{\mathcal{Z}} \ell_k^a(z')\, dP_{k-1}(z') \\ &= \int_{\mathcal{Z}\times\mathcal{Z}} \big(\ell_k^a(z) - \ell_k^a(z')\big)\, d\gamma(z, z') \\ &\leq \int_{\mathcal{Z}\times\mathcal{Z}} |\ell_k^a(z) - \ell_k^a(z')|\, d\gamma(z, z') \\ &\leq \int_{\mathcal{Z}\times\mathcal{Z}} \|z - z'\|_2 \, d\gamma(z, z').\end{aligned} \tag{30}$$

The last inequality follows from the 1-Lipschitz property of ℓ_k^a. Taking the infimum over all admissible couplings yields

$$\begin{aligned}\epsilon^a_{P_k} - \epsilon^a_{P_{k-1}} &\leq \inf_{\gamma \in \Pi(P_k, P_{k-1})} \int_{\mathcal{Z}\times\mathcal{Z}} \|z - z'\|_2 \, d\gamma(z, z') \\ &= W(P_k, P_{k-1}),\end{aligned} \tag{31}$$

which proves the theorem. □

B.3 Proof of Theorem 3

Theorem 3. Let $\mathcal{A}$ be a finite treatment space with $\pi_{k-1}(a) > 0$ for every $a \in \mathcal{A}$. Suppose that the representation distributions have finite first moments and that ℓ_{k-1}^a and ℓ_k^a are 1-Lipschitz for every $a \in \mathcal{A}$, we have

$$\mathbb{E}_{a\sim\pi_k}\left[\epsilon^a_{P_k}\right] \leq \mathbb{E}_{a\sim\pi_{k-1}}\left[\epsilon^a_{P_{k-1,a}}\right] + \mathbb{E}_{a,b\sim\pi_{k-1}}\left[W(P_{k-1,a}, P_{k-1,b})\right] + W(P_k, P_{k-1}). \tag{32}$$

Proof. Subtracting and adding the overall prediction error at time step $k-1$, and applying the Kantorovich–Rubinstein duality, gives

$$\begin{aligned}
\mathbb{E}_{a\sim\pi_k}\left[\epsilon^a_{P_k}\right] &= \left(\mathbb{E}_{a\sim\pi_k}\left[\epsilon^a_{P_k}\right] - \mathbb{E}_{a\sim\pi_{k-1}}\left[\epsilon^a_{P_{k-1}}\right]\right) + \mathbb{E}_{a\sim\pi_{k-1}}\left[\epsilon^a_{P_{k-1}}\right] \\
&= \left(\int_{\mathcal{Z}} \ell_k(z)\, dP_k(z) - \int_{\mathcal{Z}} \ell_{k-1}(z)\, dP_{k-1}(z)\right) + \mathbb{E}_{a\sim\pi_{k-1}}\left[\epsilon^a_{P_{k-1}}\right] \\
&= \left(\int_{\mathcal{Z}} \ell_k(z)\, dP_k(z) - \int_{\mathcal{Z}} \ell_k(z)\, dP_{k-1}(z)\right) + \mathbb{E}_{a\sim\pi_{k-1}}\left[\epsilon^a_{P_{k-1}}\right] \\
&\le W(P_k, P_{k-1}) + \mathbb{E}_{a\sim\pi_{k-1}}\left[\epsilon^a_{P_{k-1}}\right] \\
&= W(P_k, P_{k-1}) + \mathbb{E}_{a\sim\pi_{k-1}}\left[\epsilon^a_{P_{k-1,a}}\right] + \mathbb{E}_{a\sim\pi_{k-1}}\left[\int_{\mathcal{Z}} \ell^a_{k-1}(z)\, dP_{k-1}(z) - \int_{\mathcal{Z}} \ell^a_{k-1}(z)\, dP_{k-1,a}(z)\right] \\
&\le \mathbb{E}_{a\sim\pi_{k-1}}\left[\epsilon^a_{P_{k-1,a}}\right] + \mathbb{E}_{a\sim\pi_{k-1}}\left[W(P_{k-1,a}, P_{k-1})\right] + W(P_k, P_{k-1}).
\end{aligned} \tag{33}$$

The inequalities follow from the 1-Lipschitz properties of $\bar{\ell}_k$ and ℓ^a_{k-1}, respectively.

Since P_{k-1} is the mixture of the treatment-group distributions with weights π_{k-1}, for each treatment value a,

$$\begin{aligned}
W(P_{k-1,a}, P_{k-1}) &= \sup_{\|f\|_{\mathrm{Lip}}\le 1} \left|\mathbb{E}_{b\sim\pi_{k-1}}\left[\int_{\mathcal{Z}} f(z)\, dP_{k-1,a}(z) - \int_{\mathcal{Z}} f(z)\, dP_{k-1,b}(z)\right]\right| \\
&\le \mathbb{E}_{b\sim\pi_{k-1}}\left[\sup_{\|f\|_{\mathrm{Lip}}\le 1} \left|\int_{\mathcal{Z}} f(z)\, dP_{k-1,a}(z) - \int_{\mathcal{Z}} f(z)\, dP_{k-1,b}(z)\right|\right] \\
&= \mathbb{E}_{b\sim\pi_{k-1}}\left[W(P_{k-1,a}, P_{k-1,b})\right].
\end{aligned} \tag{34}$$

Taking the expectation over $a \sim \pi_{k-1}$ and substituting into Eq. (33) proves the theorem. □

B.4 Proof of Theorem 4

Theorem 4. Under the conditions of Theorem 3, consider a fixed model independent of the n sampled trajectories, with pointwise losses bounded in $[0, C_\ell]$. For any $\rho \in (0,1)$ and sufficiently large n, with probability at least $1-\rho$,

$$\begin{aligned}
\mathbb{E}_{a\sim\pi_k}\left[\epsilon^a_{P_k}\right] \le \mathbb{E}_{a\sim\pi_{k-1}}\left[\epsilon^a_{P_{k-1,a}}\right] + \mathbb{E}_{a,b\sim\pi_{k-1}}\left[W\left(\widehat{P}_{k-1,a}, \widehat{P}_{k-1,b}\right)\right] \\
+ W\left(\widehat{P}_k, \widehat{P}_{k-1}\right) + O\left(\sqrt{\frac{\log(2/\rho)}{n}}\right).
\end{aligned} \tag{35}$$

The implicit constant depends on C_ℓ and the fixed treatment distribution, but not on n or ρ.

Lemma 2 (Concentration of empirical loss integrals (Bolley et al., 2007))**.** *Let $Z_1, \dots, Z_m$ be independent samples from Q, and let $\widehat{Q} = m^{-1}\sum_{i=1}^m \nu_{Z_i}$. For a fixed measurable loss function $\ell\colon \mathcal{Z} \to [0, C_\ell]$ and any $r > 0$,*

$$\Pr\left(\left|\int_{\mathcal{Z}} \ell(z)\, dQ(z) - \int_{\mathcal{Z}} \ell(z)\, d\widehat{Q}(z)\right| > r\right) \le 2\exp\left(-\frac{2mr^2}{C_\ell^2}\right). \tag{36}$$

Proof. Let $M = |\mathcal{A}|$, $N = M^2 + 2M$, and

$$r_m = C_\ell\sqrt{\frac{\log(4N/\rho)}{2m}}.$$

Write $n_{k-1,a}$ for the number of sampled trajectories with $A_{k-1} = a$. Apply Lemma 2 to ℓ^a_k under P_k and P_{k-1}, and to ℓ^a_{k-1} under each $P_{k-1,b}$. A union bound controls all N sampling errors simultaneously, with failure probability at most $\rho/2$ when the treatment groups are nonempty.

Using the temporal consistency condition from Theorem 3 and the triangle inequality, we obtain

$$\begin{aligned}\mathbb{E}_{a\sim\pi_k}\left[\epsilon^a_{P_k}\right] - \mathbb{E}_{a\sim\pi_{k-1}}\left[\epsilon^a_{P_{k-1}}\right] &= \mathbb{E}_{a\sim\pi_k}\left[\int_{\mathcal{Z}} \ell^a_k \, dP_k - \int_{\mathcal{Z}} \ell^a_k \, dP_{k-1}\right] \\ &\le \mathbb{E}_{a\sim\pi_k}\left[\left|\int_{\mathcal{Z}} \ell^a_k \, dP_k - \int_{\mathcal{Z}} \ell^a_k \, d\widehat{P}_k\right|\right] \\ &\quad + \mathbb{E}_{a\sim\pi_k}\left[\left|\int_{\mathcal{Z}} \ell^a_k \, d\widehat{P}_k - \int_{\mathcal{Z}} \ell^a_k \, d\widehat{P}_{k-1}\right|\right] \\ &\quad + \mathbb{E}_{a\sim\pi_k}\left[\left|\int_{\mathcal{Z}} \ell^a_k \, d\widehat{P}_{k-1} - \int_{\mathcal{Z}} \ell^a_k \, dP_{k-1}\right|\right] \\ &\le W\Big(\widehat{P}_k, \widehat{P}_{k-1}\Big) + 2r_n.\end{aligned} \tag{37}$$

The final inequality follows from the 1-Lipschitz property of ℓ^a_k and the sampling bounds.

For each pair of treatment values $a, b \in \mathcal{A}$, introducing the conditional empirical distributions gives

$$\begin{aligned}\int_{\mathcal{Z}} \ell^a_{k-1} \, dP_{k-1,b} - \int_{\mathcal{Z}} \ell^a_{k-1} \, dP_{k-1,a} &\le \left|\int_{\mathcal{Z}} \ell^a_{k-1} \, d\widehat{P}_{k-1,b} - \int_{\mathcal{Z}} \ell^a_{k-1} \, d\widehat{P}_{k-1,a}\right| + r_{n_{k-1,b}} + r_{n_{k-1,a}} \\ &\le W\Big(\widehat{P}_{k-1,a}, \widehat{P}_{k-1,b}\Big) + r_{n_{k-1,b}} + r_{n_{k-1,a}}.\end{aligned} \tag{38}$$

Since P_{k-1} is the mixture of the treatment-group distributions with weights π_{k-1},

$$\begin{aligned}\mathbb{E}_{a\sim\pi_{k-1}}\left[\epsilon^a_{P_{k-1}} - \epsilon^a_{P_{k-1,a}}\right] &= \mathbb{E}_{a,b\sim\pi_{k-1}}\left[\int_{\mathcal{Z}} \ell^a_{k-1} \, dP_{k-1,b} - \int_{\mathcal{Z}} \ell^a_{k-1} \, dP_{k-1,a}\right] \\ &\le \mathbb{E}_{a,b\sim\pi_{k-1}}\left[W\Big(\widehat{P}_{k-1,a}, \widehat{P}_{k-1,b}\Big)\right] + 2\mathbb{E}_{a\sim\pi_{k-1}}\left[r_{n_{k-1,a}}\right].\end{aligned} \tag{39}$$

Let $p_{\min} = \min_{a\in\mathcal{A}} \pi_{k-1}(a) > 0$. A Chernoff bound and a union bound give

$$\Pr\left(\min_{a\in\mathcal{A}} n_{k-1,a} < \frac{np_{\min}}{2}\right) \le M \exp\Big(-\frac{np_{\min}}{8}\Big). \tag{40}$$

Thus, for $n \ge 8p_{\min}^{-1}\log(2M/\rho)$, all treatment groups contain at least $np_{\min}/2$ samples with probability at least $1 - \rho/2$. On the intersection with the sampling event,

$$\begin{aligned}2r_n + 2\mathbb{E}_{a\sim\pi_{k-1}}\left[r_{n_{k-1,a}}\right] &\le 2C_\ell\left(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{p_{\min}}}\right)\sqrt{\frac{\log(4N/\rho)}{n}} \\ &= O\left(\sqrt{\frac{\log(2/\rho)}{n}}\right).\end{aligned} \tag{41}$$

This intersection has probability at least $1 - \rho$. Combining Eqs. (37), (39), and (41) proves the theorem. □

C Evaluation Metrics

Let y_i and $\widehat{y}_i$ denote the target outcome and its prediction for the ith test case at prediction horizon τ, with n_τ test cases in total. We calculate

$$\mathrm{RMSE}_\tau = \sqrt{\frac{1}{n_\tau}\sum_{i=1}^{n_\tau}(y_i - \widehat{y}_i)^2}, \qquad \mathrm{NRMSE}_\tau = \frac{100\ \mathrm{RMSE}_\tau}{V_{\max}}. \tag{42}$$

For synthetic tumor growth, NRMSE is reported as a percentage with $V_{\max} = 1150\,\mathrm{cm}^3$. The semi-synthetic and real-world MIMIC-III benchmarks report RMSE. Test targets are simulated counterfactual outcomes for the synthetic and semi-synthetic benchmarks and observed factual outcomes for real-world MIMIC-III.

For model selection and early stopping, we use factual validation MSE

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2. \tag{43}$$

Results are reported as mean $\pm$ standard deviation over five random seeds. Lower values indicate better performance.

D Additional Experimental Details

D.1 Synthetic Tumor Growth Data

D.1.1 Data Generation

We use the tumor-growth simulator based on the PharmacoKinetic-PharmacoDynamic (PK-PD) model of (Geng et al., 2017), following the longitudinal benchmark in (Melnychuk et al., 2022; Bouchattaoui et al., 2024). The simulator describes tumor evolution under chemotherapy and radiotherapy. Specifically, the tumor volume $Y_{i,t}$ of patient i evolves according to

$$Y_{i,t+1} = \left[1 + \rho_i \log\left(\frac{K}{Y_{i,t}}\right) - \beta_{c,i}C_{i,t} - \left(\alpha_{r,i}d_{i,t} + \beta_{r,i}d_{i,t}^2\right) + \eta_{i,t}\right] Y_{i,t}. \tag{44}$$

Here, ρ_i controls tumor growth, K is the carrying capacity, $C_{i,t}$ is the chemotherapy drug concentration, and $d_{i,t}$ is the radiotherapy dose. The parameters governing treatment response vary across patients, while $\eta_{i,t}$ represents stochastic variation. Chemotherapy has a persistent effect through exponentially decaying drug concentration, whereas radiotherapy directly affects the next volume update. The two binary treatments yield four possible treatment combinations: no treatment, chemotherapy only, radiotherapy only, and combined therapy.

Confounding over time is introduced by making treatment assignment depend on the observed history. For each treatment type $q \in \{c, r\}$,

$$A_{i,t}^q \sim \mathrm{Bernoulli}(p_{i,t}^q), \qquad p_{i,t}^q = \sigma\left[\frac{\gamma}{D_{\max}}\left(\overline{D}_{i,t} - \frac{D_{\max}}{2}\right)\right], \tag{45}$$

where σ is the sigmoid function, $\overline{D}_{i,t}$ is the average tumor diameter over the preceding 15 days, and $D_{\max} = 13\,\mathrm{cm}$ is the reference maximum diameter. Larger γ makes treatment assignment more dependent on recent tumor history; $\gamma = 0$ corresponds to randomized assignment. Counterfactual outcomes are obtained by advancing the simulator under specified alternative treatment sequences.

As reported in the main text, we consider $\gamma \in \{1, 2, 3, 4\}$, use 1,000 training trajectories and 500 test trajectories, and limit sequence length to 60. Prediction horizons range from 1 to 10 steps. Hyperparameter selection and early stopping use factual validation MSE, and counterfactual prediction is evaluated using NRMSE.

D.1.2 Results under Stronger Confounding

Table 4 reports the counterfactual estimation results under $\gamma \in \{3, 4\}$. As treatment assignment becomes more dependent on patient history, prediction errors generally increase across methods. ReCT retains its advantage at longer prediction horizons under both confounding levels.

Model	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
					$\gamma = 3$					
MSM	1.68 ± 0.19	3.30 ± 0.57	3.97 ± 0.70	4.37 ± 0.76	4.63 ± 0.80	4.81 ± 0.82	4.92 ± 0.84	4.96 ± 0.85	4.97 ± 0.84	4.94 ± 0.83
RMSN	1.21 ± 0.08	$\mathbf{1.21 \pm 0.18}$	1.53 ± 0.23	1.88 ± 0.34	2.18 ± 0.46	2.45 ± 0.55	2.69 ± 0.61	2.89 ± 0.65	3.05 ± 0.67	3.19 ± 0.66
CRN	$\mathbf{1.16 \pm 0.14}$	1.70 ± 0.68	2.07 ± 0.83	2.37 ± 1.05	2.59 ± 1.22	2.76 ± 1.32	2.89 ± 1.36	2.98 ± 1.37	3.04 ± 1.35	3.08 ± 1.31
G-Net	1.35 ± 0.24	1.32 ± 0.16	1.59 ± 0.15	1.77 ± 0.15	1.93 ± 0.17	2.08 ± 0.20	2.22 ± 0.23	2.34 ± 0.27	2.45 ± 0.31	2.54 ± 0.34
CT	1.29 ± 0.08	1.59 ± 0.29	1.88 ± 0.31	2.08 ± 0.33	2.21 ± 0.35	2.30 ± 0.36	2.36 ± 0.36	2.43 ± 0.33	2.45 ± 0.32	2.48 ± 0.29
Causal CPC	1.45 ± 0.22	1.69 ± 0.49	1.88 ± 0.58	2.02 ± 0.65	2.14 ± 0.68	2.26 ± 0.71	2.37 ± 0.71	2.45 ± 0.72	2.52 ± 0.71	2.56 ± 0.69
ReCT (ours)	1.29 ± 0.17	1.33 ± 0.24	$\mathbf{1.51 \pm 0.24}$	$\mathbf{1.66 \pm 0.25}$	$\mathbf{1.80 \pm 0.24}$	$\mathbf{1.94 \pm 0.23}$	$\mathbf{2.07 \pm 0.21}$	$\mathbf{2.18 \pm 0.19}$	$\mathbf{2.29 \pm 0.17}$	$\mathbf{2.36 \pm 0.14}$
					$\gamma = 4$					
MSM	1.76 ± 0.70	3.71 ± 1.24	4.10 ± 1.41	4.33 ± 1.51	4.48 ± 1.59	4.62 ± 1.63	4.72 ± 1.66	4.81 ± 1.69	4.86 ± 1.70	4.89 ± 1.71
RMSN	$\mathbf{1.36 \pm 0.44}$	$\mathbf{1.58 \pm 0.39}$	2.06 ± 0.62	2.39 ± 0.81	2.64 ± 0.94	2.81 ± 1.05	2.93 ± 1.13	3.02 ± 1.19	3.08 ± 1.25	3.11 ± 1.28
CRN	1.45 ± 0.33	2.14 ± 0.82	2.58 ± 1.03	2.86 ± 1.18	3.07 ± 1.30	3.23 ± 1.40	3.35 ± 1.48	3.44 ± 1.55	3.51 ± 1.61	3.55 ± 1.65
G-Net	1.56 ± 0.34	1.61 ± 0.45	2.09 ± 0.65	2.53 ± 0.82	2.93 ± 0.99	3.27 ± 1.14	3.57 ± 1.29	3.85 ± 1.44	4.10 ± 1.57	4.33 ± 1.71
CT	1.67 ± 0.36	2.00 ± 0.76	2.25 ± 0.88	2.45 ± 0.96	2.62 ± 1.03	2.76 ± 1.09	2.88 ± 1.14	2.97 ± 1.20	3.05 ± 1.25	3.11 ± 1.28
Causal CPC	1.87 ± 0.45	2.34 ± 1.00	2.65 ± 1.20	2.92 ± 1.34	3.19 ± 1.43	3.45 ± 1.49	3.68 ± 1.53	3.88 ± 1.56	4.04 ± 1.59	4.15 ± 1.60
ReCT (ours)	1.59 ± 0.32	1.74 ± 0.52	$\mathbf{1.95 \pm 0.62}$	$\mathbf{2.15 \pm 0.72}$	$\mathbf{2.33 \pm 0.85}$	$\mathbf{2.50 \pm 0.98}$	$\mathbf{2.67 \pm 1.11}$	$\mathbf{2.82 \pm 1.21}$	$\mathbf{2.96 \pm 1.30}$	$\mathbf{3.08 \pm 1.38}$

Table 4: NRMSE on the synthetic tumor-growth dataset with sequence length 60 under stronger confounding (mean $\pm$ standard deviation). Lower is better, and bold values indicate the best result at each horizon.

D.1.3 Generalization Across Backbones

To evaluate the generality of temporal distribution alignment across different network architectures, we incorporate the temporal distribution alignment loss into CRN (Bica et al., 2020) and CT (Melnychuk et al., 2022), both established representation-based models for counterfactual prediction over time. Table 5 reports the results on the synthetic tumor-growth benchmark. Adding the temporal distribution alignment loss improves the prediction performance of both backbones, with more pronounced gains at longer horizons. These results show that temporal distribution alignment improves counterfactual estimation across different temporal causal prediction architectures.

Model	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
CRN	0.90 ± 0.14	1.13 ± 0.33	1.23 ± 0.39	1.36 ± 0.48	1.49 ± 0.57	1.61 ± 0.63	1.72 ± 0.67	1.80 ± 0.68	1.87 ± 0.68	1.93 ± 0.66
CRN + $\mathcal{L}_{\text{tem}}$	$\mathbf{0.90 \pm 0.14}$	$\mathbf{1.02 \pm 0.11}$	$\mathbf{1.13 \pm 0.13}$	$\mathbf{1.24 \pm 0.17}$	$\mathbf{1.35 \pm 0.21}$	$\mathbf{1.44 \pm 0.23}$	$\mathbf{1.52 \pm 0.25}$	$\mathbf{1.57 \pm 0.26}$	$\mathbf{1.61 \pm 0.26}$	$\mathbf{1.64 \pm 0.26}$
CT	1.06 ± 0.06	1.23 ± 0.13	1.36 ± 0.14	1.48 ± 0.15	1.56 ± 0.16	1.62 ± 0.16	1.67 ± 0.18	1.71 ± 0.20	1.74 ± 0.22	1.77 ± 0.26
CT + $\mathcal{L}_{\text{tem}}$	$\mathbf{0.94 \pm 0.11}$	$\mathbf{1.02 \pm 0.15}$	$\mathbf{1.13 \pm 0.18}$	$\mathbf{1.21 \pm 0.20}$	$\mathbf{1.28 \pm 0.20}$	$\mathbf{1.33 \pm 0.21}$	$\mathbf{1.38 \pm 0.19}$	$\mathbf{1.42 \pm 0.18}$	$\mathbf{1.46 \pm 0.17}$	$\mathbf{1.49 \pm 0.15}$

Table 5: Generalization across backbones on the synthetic tumor-growth benchmark, measured by NRMSE. Results are reported as mean $\pm$ standard deviation. $\mathcal{L}_{\text{tem}}$ denotes the temporal distribution alignment loss. Lower values indicate better performance.

D.1.4 Parameter Sensitivity Analysis

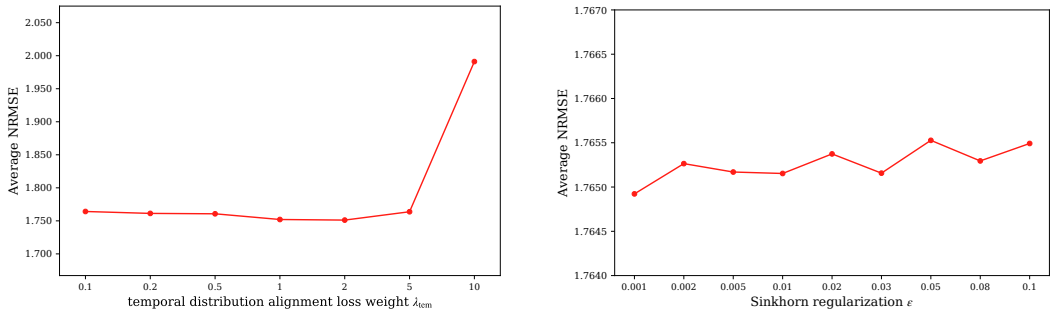
To study the effect of λ_{tem}, which controls the contribution of temporal alignment to the training objective, we vary its value over $\{0.1, 0.2, 0.5, 1, 2, 5, 10\}$. Figure 4(a) shows that ReCT maintains similar performance for values between 0.1 and 5, with the lowest observed NRMSE at $\lambda_{\text{tem}} = 2$. The tested values of 10 yield higher errors, suggesting that an excessive alignment weight can impair prediction accuracy.

We further examine the entropy regularization coefficient ε, which controls the smoothness of the Sinkhorn transport plan, over $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.03, 0.05, 0.08, 0.1\}$. As shown in Figure 4(b), average NRMSE changes only slightly across the tested values. This indicates that predictive performance is relatively insensitive to ε within this range, despite the visible fluctuations on the magnified vertical scale.

D.1.5 Visualization Results

We use UMAP to visualize hidden state representations at time steps 3, 4, and 5. UMAP1, UMAP2, and UMAP3 denote the three coordinates of the learned embedding and have no predefined clinical meaning. Each point represents temporal representation at the indicated step.

Figure 5 shows that the projections without ReCT become concentrated along a narrow curved structure at later steps, whereas those with ReCT retain several locally structured groups. The visualizations provide qualitative evidence of differences in representation structure.



(a) temporal distribution alignment loss weight λ_{tem}.

(b) Sinkhorn regularization ε.

Figure 4: Sensitivity analysis on synthetic data. Panel (a) varies λ_{tem} and panel (b) varies ε. ReCT performs consistently for moderate alignment weights and shows little sensitivity to ε over the tested range. Both panels report NRMSE averaged over horizons 1 to 10, with different vertical scales. Lower is better.

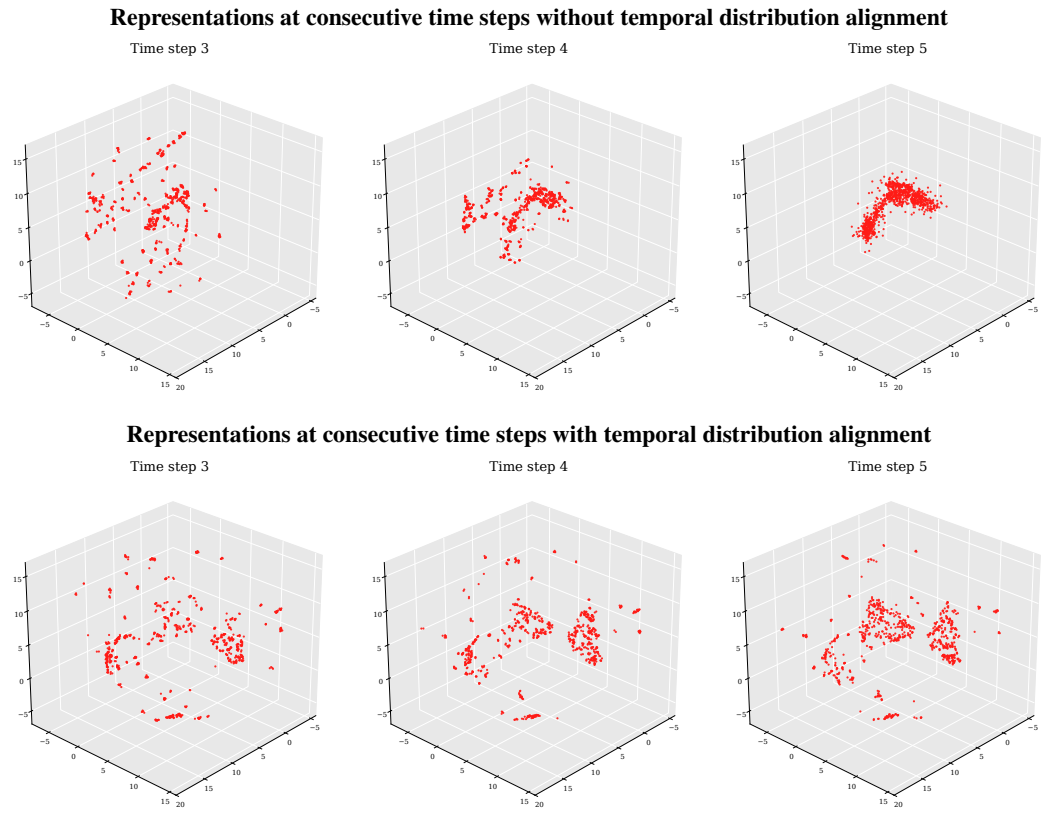


Figure 5: Visualization results on synthetic data. Temporal representations are projected using UMAP at time steps 3, 4, and 5 from left to right. The top row shows representations at consecutive time steps without temporal distribution alignment, and the bottom row shows representations at consecutive time steps with temporal distribution alignment. At later steps, the projections with ReCT retain a more varied local structure than those without alignment.

D.2 SEMI-SYNTHETIC MIMIC-III DATA

D.2.1 DATA GENERATION

The semi-synthetic benchmark combines observed clinical covariates from MIMIC-III (Johnson et al., 2016) with simulated treatment assignments and outcomes, following (Melnychuk et al., 2022; Bouchattaoui et al., 2024). This construction retains realistic clinical covariate trajectories while making potential outcomes available for evaluation.

For patient i and outcome dimension j, the untreated outcome is constructed as

$$U_{i,t}^j = \alpha_S^j s_i(t) + \alpha_g^j g_i^j(t) + \alpha_f^j f^j(X_{i,t}) + \eta_{i,t}^j. \tag{46}$$

Here, $s_i(t)$ is a B-spline component capturing broad temporal trends, $g_i^j(t)$ is a Gaussian process component specific to each patient with a Matérn kernel, and f^j maps a subset of observed covariates to outcomes using a random Fourier feature approximation of a Gaussian process. The coefficients control the contributions of these components, and $\eta_{i,t}^j$ denotes additive noise.

Treatments are then assigned sequentially using current covariates and previously generated treated outcomes. For treatment type l,

$$\begin{aligned} p_{i,t}^l &= \sigma\Big(\gamma_A^l \overline{Y}^l_{i,t-1} + \gamma_X^l f_A^l(X_{i,t}) + b_l\Big), \\ A_{i,t}^l &\sim \text{Bernoulli}(p_{i,t}^l), \end{aligned} \tag{47}$$

where $\overline{Y}^l_{i,t-1}$ averages the relevant past outcomes over a history window specific to each treatment. The coefficients γ_A^l and γ_X^l determine the dependence of treatment assignment on outcome history and current covariates, respectively. This dependence introduces confounding over time.

Following the benchmark simulator, administered treatments generate effects on specific outcomes that persist within a finite time window and decay over time according to an inverse square rule. Let $E_{i,t}^j$ denote their accumulated contribution, including the simulator's aggregation across treatment types. The observed outcome is then

$$Y_{i,t}^j = U_{i,t}^j + E_{i,t}^j. \tag{48}$$

Alternative treatment sequences are applied within the simulator to obtain the corresponding counterfactual outcomes.

Our experiments use 1,000 training, 200 validation, and 200 test trajectories, with a maximum sequence length of 100 and prediction horizons from 1 to 10. These are the splits used in our experiments, as reported in the main text. Model selection and early stopping are based on factual validation MSE, and test performance is reported as RMSE across prediction horizons.

D.2.2 GENERALIZATION ACROSS BACKBONES

We further evaluate the generality of temporal distribution alignment across different network architectures on the semi-synthetic MIMIC-III benchmark by incorporating the temporal distribution alignment loss $\mathcal{L}_{\text{tem}}$ into CRN (Bica et al., 2020) and CT (Melnychuk et al., 2022). As shown in Table 6, adding the temporal distribution alignment loss improves the long-horizon prediction performance of both backbones. The improvements become more pronounced as the prediction horizon increases, supporting the effectiveness of temporal distribution alignment for counterfactual estimation over time. These results further demonstrate the applicability of temporal distribution alignment across different network architectures on the semi-synthetic benchmark.

E TRAINING PROCEDURE

Algorithm 1 summarizes the training procedure of ReCT. For each mini-batch of observed trajectories, the history encoder initializes the autoregressive model at time t. The model then predicts outcomes over the next τ time steps under the observed treatment sequences, using predicted outcomes as inputs to subsequent updates. The observed outcome $Y_{i,t}$ provides the initial outcome

Model	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
CRN	0.40 ± 0.09	0.48 ± 0.08	0.61 ± 0.10	0.70 ± 0.13	0.78 ± 0.16	0.85 ± 0.19	0.91 ± 0.23	0.96 ± 0.27	1.01 ± 0.31	1.06 ± 0.35
CRN + $\mathcal{L}_{\text{tem}}$	$\mathbf{0.40 \pm 0.09}$	$\mathbf{0.42 \pm 0.04}$	$\mathbf{0.51 \pm 0.03}$	$\mathbf{0.58 \pm 0.03}$	$\mathbf{0.63 \pm 0.03}$	$\mathbf{0.68 \pm 0.04}$	$\mathbf{0.72 \pm 0.06}$	$\mathbf{0.76 \pm 0.08}$	$\mathbf{0.79 \pm 0.10}$	$\mathbf{0.82 \pm 0.12}$
CT	$\mathbf{0.26 \pm 0.02}$	$\mathbf{0.36 \pm 0.02}$	0.46 ± 0.05	0.54 ± 0.08	0.57 ± 0.11	0.62 ± 0.14	0.67 ± 0.15	0.71 ± 0.18	0.75 ± 0.19	0.79 ± 0.22
CT + $\mathcal{L}_{\text{tem}}$	0.30 ± 0.07	0.37 ± 0.02	$\mathbf{0.46 \pm 0.02}$	$\mathbf{0.52 \pm 0.04}$	$\mathbf{0.57 \pm 0.05}$	$\mathbf{0.61 \pm 0.07}$	$\mathbf{0.65 \pm 0.08}$	$\mathbf{0.68 \pm 0.10}$	$\mathbf{0.71 \pm 0.11}$	$\mathbf{0.73 \pm 0.12}$

Table 6: Generalization across backbones on the semi-synthetic MIMIC-III benchmark, measured by RMSE. Results are reported as mean $\pm$ standard deviation. $\mathcal{L}_{\text{tem}}$ denotes the temporal distribution alignment loss. Lower values indicate better performance.

Algorithm 1 Training procedure of ReCT

Require: Observed trajectories $\mathcal{D}$, prediction horizon τ, mini-batch size B, loss weights λ_{adv} and λ_{tem}, entropy regularization ε.
Ensure: Trained representation model and outcome predictor.
Initialize model parameters θ, η, and ψ.
repeat
  Sample B factual trajectory segments with observed histories $H_{i,t+1}$ and τ future time steps.
  Set $h_{i,t} = E_\theta(H_{i,t+1})$, $z_{i,t} = \phi_\theta(h_{i,t})$, and $\widehat{Y}_{i,t} = Y_{i,t}$ for all i.
  for $k = t+1$ to $t+\tau$ **do**
    Compute $h_{i,k} = G_\theta(h_{i,k-1}, A_{i,k-1}, \widehat{Y}_{i,k-1})$ for all i.
    Compute $z_{i,k} = \phi_\theta(h_{i,k})$ and $\widehat{Y}_{i,k} = g_\eta(z_{i,k}, A_{i,k})$ for all i.
    Construct $[C_\theta^{(k)}]_{ij} = \|z_{i,k} - z_{j,k-1}\|_2$ for all i, j.
    Compute $\widetilde{\pi}^{(k)}$ by Sinkhorn iterations with row and column marginals $1/B$, following Eq. (16).
    Compute $\widehat{W}_k = \sum_{i=1}^{B} \sum_{j=1}^{B} [C_\theta^{(k)}]_{ij} \widetilde{\pi}_{ij}^{(k)}$.
  end for
  Compute the mini-batch factual prediction loss $\mathcal{L}_Y$ and treatment prediction loss $\mathcal{L}_{\text{adv}}$.
  Compute $\mathcal{L}_{\text{tem}} = \sum_{k=t+1}^{t+\tau} \widehat{W}_k$.
  Update ψ to minimize $\mathcal{L}_{\text{adv}}$ with θ and η fixed.
  Update θ and η to minimize $\mathcal{L}_Y - \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{tem}} \mathcal{L}_{\text{tem}}$ with ψ fixed.
until Convergence

input. At each prediction step, an entropy-regularized transport plan is computed between the empirical representation distributions at consecutive time steps using Sinkhorn iterations. The resulting transport costs are accumulated to obtain the temporal distribution alignment loss.

The treatment classifier is updated to minimize the treatment prediction loss. The representation model and outcome predictor are then updated using the joint objective in Eq. (19), combining factual outcome prediction, adversarial treatment balancing, and temporal distribution alignment. Following the parameter convention in the joint objective, θ denotes the representation model parameters, η denotes the outcome predictor parameters, and ψ denotes the treatment classifier parameters. The outcome predictor is therefore written as g_η in Algorithm 1. The factual prediction and treatment prediction losses are averaged over the mini-batch, with the temporal sums defined in Eqs. (13) and (14).