Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Enhancing Counterfactual Estimation: A Focus on Temporal Treatments

Xin Wang[1] , **Shengfei Lyu***[2] , **Kangyang Luo**[1] , **Lishan Yang**[1] , **Huanhuan Chen***[1] and **Chunyan Miao**[2]

[1]University of Science and Technology of China
[2]Nanyang Technological University
{wz520, ky_luo, ylishan}@mail.ustc.edu.cn, hchen@ustc.edu.cn, {shengfei.lyu, ascymiao}@ntu.edu.sg,

Abstract

In the medical field, treatment sequences significantly influence future outcomes through complex temporal interactions. Therefore, highlighting the role of temporal treatments within the model is crucial for accurate counterfactual estimation, which is often overlooked in current methods. To address this, we employ the Koopman operator, known for its capability to model complex dynamic systems, and introduce a novel model named the Counterfactual Temporal Dynamics Network via Neural Koopman Operators (CTD-NKO). This model utilizes Koopman operators to encapsulate sequential treatment data, aiming to capture the causal dynamics within the system induced by temporal interactions between treatments. Moreover, CTD-NKO implements a weighting strategy that aligns joint and marginal distributions of the system state and the current treatment to mitigate time-varying confounding bias. This deviates from the balanced representation strategy employed by existing methods, as we demonstrate that such a strategy may suffer from the potential information loss of historical treatments. These designs allow CTD-NKO to exploit treatment information more thoroughly and effectively, resulting in superior performance on both synthetic and real-world datasets.

1 Introduction

Accurate estimation of counterfactuals over time is crucial for evaluating the temporal effects of different treatment strategies, which can optimize medical decision-making and significantly impact healthcare [Yazdani and Boerwinkle, 2015]. While randomized controlled trials are the gold standard for causal inference [Hariton and Locascio, 2018], estimating counterfactuals from observational data is gaining attention due to the high costs and ethical constraints associated with conducting these trials in real-world settings.

Recent neural network techniques have advanced this field, particularly by focusing on integrating treatments within network architectures, supported by evidence in static causal inference settings. For example, [Shalit *et al.*, 2017] develop a two-head network architecture for binary treatment observational data, influencing many subsequent studies [Shi *et al.*, 2019; Hassanpour and Greiner, 2020; Johansson *et al.*, 2022]. For continuous treatments, [Nie *et al.*, 2020] introduce the Varying Coefficient Network (VCNet), enabling the prediction network to function continuously with treatment, thus strengthening the impact of treatment on predictions.

In longitudinal settings, current and historical treatments, viewed as time series, often jointly influence future patient outcomes through complex temporal interactions. For example, understanding the intricate temporal interactions between antibiotics is crucial for optimizing antibiotic usage and minimizing the development of antibiotic resistance [Roemhild *et al.*, 2022]. However, existing methods often overlook the role of treatments in their model design. For example, Counterfactual Recurrent Network (CRN) [Bica *et al.*, 2020] and Causal Transformer (CT) [Melnychuk *et al.*, 2022] learn representations from historical information and then concatenate them with current treatments as input to a feedforward neural network for counterfactual estimation. This simple concatenation of current treatments with historical information embedded in representations may overly simplify interactions between temporal treatments, thus undermining the model's ability to capture the complex dynamics that evolve over time.

Another challenge in counterfactual estimation when observational data as time series is the complex confounding bias introduced by time-varying confounders. CRN and CT attempt to mitigate this issue by learning a balanced representation that excludes current treatment assignment information. However, in real-world applications, current treatments are often closely related to historical treatments. Our research demonstrates that adopting such a balanced representation can lead to the loss of historical treatment information, which may adversely affect counterfactual estimation.

The challenges identified make it difficult for existing methods to effectively and comprehensively utilize temporal treatment information. Our study models the evolution of patient states as a dynamical system. Subsequently, using the Koopman operator [Koopman, 1931], which linearizes nonlinear systems to effectively model complex dynamics, we propose a novel model named the Counterfactual Temporal Dynamics Network via Neural Koopman Operators (CTD-NKO). Figure 1 shows CTD-NKO's integration of two Recurrent Neural Network (RNN) modules. One RNN is tasked

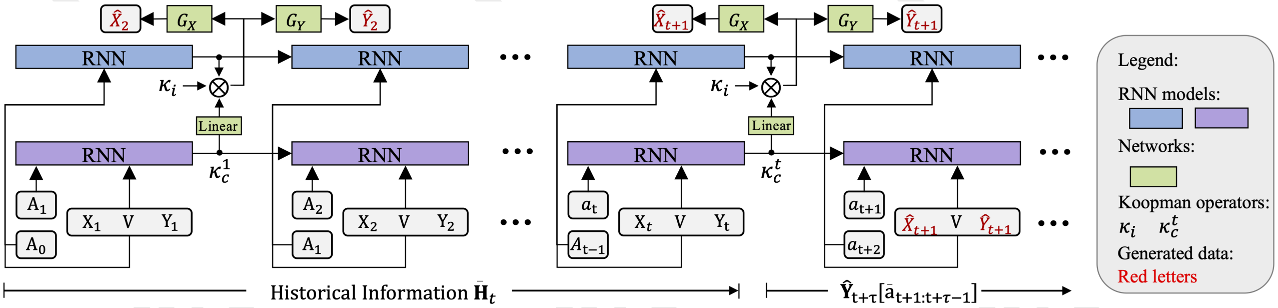


Figure 1: The architectural diagram of CTD-NKO is illustrated as follows: An RNN utilizes historical information to learn the system state (A_0 is zero-padded); another RNN combines current treatment and historical information to learn the causal Koopman operator $\mathcal{K}_c^t$, while $\mathcal{K}_i$ is a trainable matrix. The symbol $\bigotimes$ represents operations within the Koopman space. Subsequently, the system employs feedforward networks G_X and G_Y to estimate counterfactuals based on the predicted next system state. CTD-NKO is trained utilizing observational data from previous time points $(1, \cdots, t)$, and it employs an autoregressive strategy for inference at subsequent time points $(t+1, \cdots, t+\tau)$.

with learning the system state, while the other combines current and historical treatment information to learn a causal Koopman operator. This architecture capitalizes on the ability of Koopman operators to model complex nonlinear systems, thus enhancing the capacity of capturing causal dynamics induced by temporal treatments. Additionally, CTD-NKO learns an intrinsic Koopman operator to capture inherent system dynamics, such as the circadian rhythm of heart rate [Massin *et al.*, 2000].

Furthermore, CTD-NKO introduces a novel weighting method to mitigate confounding bias by aligning the joint and marginal distributions of the system state and the current treatment. This approach avoids the loss of historical treatment information caused by learning a balanced representation that excludes current treatment assignment information, further improving counterfactual estimation performance. CTD-NKO offers several key advantages over existing methods, including unified treatment representation, separation of causal and non-causal dynamics, and enhanced expressiveness in modeling complex temporal dependencies (see Appendix A for detailed discussion). Overall, our main contributions are threefold:

- CTD-NKO improves the capability of counterfactual estimation over time by effectively learning complex interactions among temporal treatment using Koopman operators.
- We theoretically point out that using treatment-invariant balanced representations to mitigate confounding bias may lead to the loss of historical treatment information. To avoid this issue, we alternatively propose a novel weighting method.
- Experimental results on both synthetic and real-world datasets demonstrate that CTD-NKO achieves state-of-the-art performance and efficiency.

2 Related Work

Counterfactual estimation over time. Current leading-edge methodologies for analyzing time-varying outcomes utilize advancements in deep neural networks. Notable examples include RMSN [Lim *et al.*, 2018], CRN [Bica *et al.*, 2020], G-Net [Li *et al.*, 2021], and CT [Melnychuk *et al.*, 2022]. RMSN incorporates two propensity networks and uses a training approach based on Inverse Probability of Treatment Weighting (IPTW) for its prediction models. G-Net enhances the conventional G-computation technique via a deep learning framework. Both CT and CRN, on the other hand, focus on creating balanced representations that effectively predict outcomes while not being predictive of current treatment allocations. These methods often learn historical and current treatment information separately, which is not conducive to exploring complex dynamics, especially the causal dynamics driven by temporal interactions between treatments. A recent study [Kacprzyk *et al.*, 2024] introduces a method based on ordinary differential equations (ODE) rather than neural networks for estimating counterfactuals over time. This method offers improved interpretability and the ability to handle irregularly sampled data, opening new avenues for research. However, its reliance on ODE discovery may impose limitations due to strong assumptions, e.g., the functional form of the ODE involved. A more comprehensive survey of relevant literature on causal inference can be found in Appendix B.

Koopman operator. Koopman operator [Koopman, 1931] explores how nonlinear dynamics can be linearized through an infinite-dimensional operator on Koopman space, often approximated by Dynamic Mode Decomposition (DMD) [Brunton *et al.*, 2016]. Recent advancements in integrating Koopman operator with machine learning, particularly through the use of Deep Neural Networks (DNNs) [Takeishi *et al.*, 2017; Morton *et al.*, 2019; Yeung *et al.*, 2019; Han *et al.*, 2020; Fan *et al.*, 2022], have greatly enhanced the ability to directly derive measurement functions from data. These innovations encompass the utilization of DNNs to construct Koopman invariant subspaces and the dynamic adaptation of Koopman operators to accommodate evolving system dynamics, marking a significant departure from conventional static approaches. For example, [Takeishi *et al.*, 2017] propose a data-driven approach to learn Koopman invariant subspaces (LKIS), enhancing the ability of DNNs for Koopman spectral analysis. [Brunton *et al.*, 2022] explore various methods, including data-driven techniques such as DNNs, that utilize predefined functions to improve the Koopman operator's ability to model and analyze nonlinear dynamical systems by approximating their underlying linear structure.

However, our work differs from these studies in its focus. Rather than developing novel theories related to Koopman operator, we are more inclined towards its application. This aligns with recent works that employ Koopman operators to address temporal domain adaptation. For instance, KNF [Wang *et al.*, 2023] employs predefined measurement functions for learning a Koopman operator and an attention map, targeting temporal distribution changes in forecasting. Concurrently, Koopa [Liu *et al.*, 2023] introduces an inventive architecture, merging deep residual structures with Koopman operators to enhance efficiency and performance in managing non-stationary time series. We observe that current temporal counterfactual estimation methods often overlook the interaction between treatments over time in their model design. Our paper applies Koopman operators to encapsulate sequential treatment data over time, distinguishing it from existing methods by effectively addressing this oversight.

3 Background

3.1 Problem Formulation

Consider an i.i.d. observational dataset $\mathcal{D}$, which contains detailed information of N patients. Mathematically, it can be represented as $\mathcal{D} = \left\{ \mathbf{x}_t^{(i)}, \mathbf{a}_t^{(i)}, \mathbf{y}_t^{(i)}{}_{t=1}^{T^{(i)}} \cup \mathbf{v}^{(i)} \right\}_{i=1}^{N}$. For each patient indexed by i, the observations include time-varying covariates $\mathbf{X}_t^{(i)} \in \mathcal{X}$, treatments received $\mathbf{A}_t^{(i)} \in \mathcal{A}$, and outcomes $\mathbf{Y}_t^{(i)} \in \mathcal{Y}$ at discrete time steps $T^{(i)}$. Additionally, patients' static covariates (e.g., gender and age) are denoted as $\mathbf{V}^{(i)} \in \mathcal{V}$. To simplify the notation, the patient-specific superscript (i) will be omitted when it does not affect the context understanding.

Based on the potential outcome framework proposed by Robin [Rubin, 1978] and its extension to accommodate time-varying treatments by Robins and Hernan [Robins and Hernan, 2008], we aim to estimate time-varying counterfactual outcomes, following previous studies [Bica *et al.*, 2020; Lim *et al.*, 2018; Li *et al.*, 2021]. Let the patient's historical information be $\bar{\mathbf{H}}_t = (\bar{\mathbf{X}}_t, \bar{\mathbf{A}}_{t-1}, \bar{\mathbf{Y}}_t, \mathbf{V})$, where $\bar{\mathbf{X}}_t = (\mathbf{X}_1, \cdots, \mathbf{X}_t)$, $\bar{\mathbf{Y}}_t = (\mathbf{Y}_1, \cdots, \mathbf{Y}_t)$, and $\bar{\mathbf{A}}_{t-1} = (\mathbf{A}_1, \cdots, \mathbf{A}_{t-1})$. Our goal is to estimate the potential outcome $\mathbf{Y}_{t+\tau}[\bar{\mathbf{a}}_{t:t+\tau-1}]$ after a sequence of treatments $\bar{\mathbf{a}}_{t:t+\tau-1} = (\mathbf{a}_t, \cdots, \mathbf{a}_{t+\tau-1})$, given the patient's historical information $\bar{\mathbf{H}}_t$, i.e.,

$$\mathbb{E}[\mathbf{Y}_{t+\tau}[\bar{\mathbf{a}}_{t:t+\tau-1}] | \bar{\mathbf{H}}_t]. \tag{1}$$

The identifiability of treatment effects from observational data relies on assumptions outlined in prior research [Melnychuk *et al.*, 2022; Lim *et al.*, 2018]: consistency, sequential ignorability, and sequential overlap. These are detailed in Appendix C.

3.2 Koopman Operator

A discrete-time dynamical system is represented by $u_{t+1} = \mathbf{F}(u_t)$, where u_t denotes the system state and $\mathbf{F}$ describes the dynamics. Identifying direct transitions between states can be challenging due to nonlinearity or noise [Eivazi *et al.*, 2021; Morton *et al.*, 2018]. Koopman theory [Koopman, 1931], however, allows for the projection of the state u_t into a measurement function space (a.k.a. Koopman space) g, managed by an infinite-dimensional linear operator $\mathcal{K}$, such that:

$$\mathcal{K}g(u_t) = g(\mathbf{F}(u_t)) = g(u_{t+1}). \tag{2}$$

Koopman operators can effectively model complex nonlinear dynamics [Koopman, 1931; Yeung *et al.*, 2019; Li and Jiang, 2021], leveraging linear operators to advance understanding of dynamical systems.

4 Methods

The CTD-NKO, illustrated in Figure 1, comprises two primary RNN modules. One RNN is tasked with learning the system state, while the other is dedicated to learning the Koopman operator that governs state evolution. Subsequently, the model employs the evolved state for counterfactual prediction. We will next detail the problem modeling and the specifics of our method. Code available at https://github.com/wangxin0126/CTD-NKO_IJCAI.

4.1 Modeling Patient State Dynamics via Koopman Theory

In this study, we model patient state evolution as a dynamical system, with transitions intricately linked to historical data, especially prior treatments. These transitions usually display complex dynamics due to temporal treatment interactions. To capture this complexity, we utilize Koopman theory in our modeling approach.

Inspired by the synergy between Koopman theory and machine learning [Takeishi *et al.*, 2017; Morton *et al.*, 2019; Fan *et al.*, 2022; Wang *et al.*, 2023; Liu *et al.*, 2023], we employ DNNs alongside a set of measurement functions $\mathcal{G} := [g_1, \cdots, g_n]$, each mapping $\mathbb{R} \to \mathbb{R}$, to learn the dynamics of complex systems. These functions include nonlinear mappings such as exponential functions, enhancing the DNNs' ability to capture the inherent nonlinearity of dynamical systems [Wang *et al.*, 2023; Kutz *et al.*, 2016; Brunton *et al.*, 2022]. Specifically, an encoder Φ_r is used to extract the latent state vector $\mathbf{r}_t = \Phi_r(\bar{\mathbf{H}}_t) \in \mathbb{R}^n$ from historical data. We then apply these measurement functions to project the learned state into the Koopman space:

$$\mathcal{G}(\mathbf{r}_t) = [g_1(\mathbf{r}_t^1), g_2(\mathbf{r}_t^2), \cdots, g_n(\mathbf{r}_t^n)], \tag{3}$$

where upper indices indicate vector components. For clarity, we denote the system state in the Koopman space by $\mathbf{s}_t = \Psi(\bar{\mathbf{H}}_t) = \mathcal{G}(\Phi_r(\bar{\mathbf{H}}_t)) \in \mathcal{S}$, where Ψ is the composition of $\mathcal{G}$ and Φ_r.

In counterfactual estimation for longitudinal data, the evolution of a system can be naturally divided into two types: causal and non-causal. Causal evolution primarily arises from the interaction between the treatment sequence and other historical information, while non-causal evolution reflects the inherent patterns of the system itself. For instance, in the absence of medication, a patient's heart rate or blood pressure typically exhibits periodic variations. As such, we define causal $\mathcal{K}_c^t$ and intrinsic $\mathcal{K}_i$ Koopman operators to capture

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

these evolutionary processes, allowing us to describe the system's evolution in the Koopman space as follows:

$$\mathbf{s}_{t+1} = (\mathcal{K}_c^t + \mathcal{K}_i)\mathbf{s}_t. \tag{4}$$

4.2 Counterfactual Estimation

Consistent with prior studies [Wang *et al.*, 2023; Liu *et al.*, 2023], we utilize DNNs to derive a matrix representation of the Koopman operator. For causal evolution, we employ an encoder Φ_c to generate the low-dimensional representation $\Phi_c(\bar{\mathbf{H}}_t, \mathbf{A}_t)$, which is subsequently transformed as an $n \times n$ matrix representing the causal Koopman operator $\hat{\mathcal{K}}_c^t$. In our implementation, both Φ_r and Φ_c are implemented using Long Short-Term Memory units (LSTM), with the parameters denoted as θ_Φ. For non-causal dynamics, a separate matrix forms the intrinsic Koopman operator $\hat{\mathcal{K}}_i$. Given Equation 4, we compute the estimated system state $\hat{\mathbf{s}}_{t+1}$ in the Koopman space for the next time step.

Subsequently, this allows for counterfactual predictions based on $\hat{\mathbf{s}}_{t+1}$. CTD-NKO adopts an autoregressive recursive strategy [Chevillon, 2007; Taieb and Atiya, 2015] for multi-step-ahead prediction, which is also employed in G-Net [Li *et al.*, 2021]. Therefore, during the training process, we need to predict the output and time-varying covariates for the next step. To perform output prediction, we define a feedforward neural network G_Y, parameterized by θ_Y, to decode the expected output from $\hat{\mathbf{s}}_{t+1}$. We use the Mean Squared Error (MSE) to define the following loss function:

$$\mathcal{L}_Y(\theta_Y, \theta_\Phi, t) = \|\mathbf{Y}_{t+1} - G_Y(\hat{\mathbf{s}}_{t+1}|\theta_Y)\|^2. \tag{5}$$

To predict the covariates, we design a module F_X that consists of two feedforward neural networks, G_X and J_X, with parameters represented by θ_X. Similar to the output prediction, we use G_X to decode the expected covariates from $\hat{\mathbf{s}}_{t+1}$. Some covariates may change at a slower pace, such as cholesterol levels. Inspired by the gating mechanism in GRUs, we utilize J_X to design a smoothing mechanism that adapts to this change trend:

$$F_X(\theta_X, \theta_\Phi, t) = \rho G_X(\hat{\mathbf{s}}_{t+1}) + (1-\rho)\mathbf{X}_t, \tag{6}$$

where $\rho = \text{Sigmoid}(J_X(\hat{\mathbf{s}}_{t+1}))$ is used to regulate the smoothing degree, achieving a balance between historical observations and predicted values. Similarly, we define the following loss function:

$$\mathcal{L}_X(\theta_X, \theta_\Phi, t) = \|\mathbf{X}_{t+1} - F_X(\theta_X, \theta_\Phi, t)\|^2. \tag{7}$$

Furthermore, to encourage the model to learn Koopman operators correctly, we define the following loss function based on the 'true' $\mathbf{s}_{t+1}$ learned by the encoder:

$$\mathcal{L}_\mathcal{K}(\theta_\Phi, t) = \|\mathbf{s}_{t+1} - \hat{\mathbf{s}}_{t+1}\|^2. \tag{8}$$

4.3 Balancing via Weighted Factual Loss

The aforementioned modeling approach facilitates accurate next-step outcome predictions. Let $\mathcal{K}^t = \mathcal{K}_c^t + \mathcal{K}_i \in \Omega_\mathcal{K}$, and let $f : \mathcal{S} \times \Omega_\mathcal{K} \to \mathcal{Y}$ and L represent the prediction function and the loss function, respectively. The marginal error that we aim to minimize is then defined as:

$$\epsilon_\text{M} := \mathbb{E}_{\bar{\mathbf{H}}_t \sim P(\bar{\mathbf{H}}_t)}[L(f(\mathbf{S}_t, \mathcal{K}^t), \mathbf{Y}_{t+1}(\mathbf{A}_t))], \tag{9}$$

which is consistent with the evaluation method proposed by [Melnychuk *et al.*, 2022] and reflects the model's performance in estimating counterfactual distributions. Given the unobservability of counterfactual outcomes, we are constrained to estimate only the factual error ϵ_F from the observed data:

$$\epsilon_\text{F} := \mathbb{E}_{\bar{\mathbf{H}}_t \sim P(\bar{\mathbf{H}}_t|\mathbf{A}_t)}[L(f(\mathbf{S}_t, \mathcal{K}^t), \mathbf{Y}_{t+1}(\mathbf{A}_t))]. \tag{10}$$

However, due to confounding bias, the distribution during counterfactual evaluation often diverges from the factual distribution, suggesting that using factual loss as a surrogate for marginal loss could introduce biases. To address this issue, [Bica *et al.*, 2020] and [Melnychuk *et al.*, 2022] propose to learn a treatment-invariant balanced representation $\Phi(\bar{\mathbf{H}}_t)$ through an encoder Φ. This aims to equalize $P(\Phi(\bar{\mathbf{H}}_t))$ and $P(\Phi(\bar{\mathbf{H}}_t)|\mathbf{A}_t)$, decoupling the representations of historical data from current treatments, thereby reducing confounding. However, this approach may compromise the precision of counterfactual estimates, leading us to propose Lemma 4.1 for further clarification.

Lemma 4.1. *Let $\mathbf{Z}_t = \Phi(\bar{\mathbf{H}}_t)$, and let $I(\cdot;\cdot)$ denote the mutual information. Suppose $I(\bar{\mathbf{A}}_{t-1}; \mathbf{A}_t) > 0$. When completely eliminating the correlation between the representation of the historical information and the current treatment, i.e., $I(\mathbf{Z}_t; \mathbf{A}_t) = 0$, the information about $\bar{\mathbf{A}}_{t-1}$ contained in $\mathbf{Z}_t$ must be lossy relative to the historical information $\bar{\mathbf{H}}_t$:*

$$I(\bar{\mathbf{A}}_{t-1}; \mathbf{Z}_t) < I(\bar{\mathbf{A}}_{t-1}; \bar{\mathbf{H}}_t). \tag{11}$$

Lemma 4.1 suggests that using treatment-invariant balanced representations leads to the loss of historical treatment information, which may have a negative impact on accurately predicting future treatment outcomes. To address this, we propose an alternative approach to mitigate confounding bias, allowing the weighted-adjusted ϵ_F to be used as an estimate of ϵ_M. The key idea of our method is to minimize the distance between the weighted joint distribution and the marginal distribution of the system state $\mathbf{S}_t$ and the current treatment $\mathbf{A}_t$. To achieve this, we introduce the Integral Probability Metric (IPM), a measure of the distance between two probability distributions Q and P:

$$\text{IPM}_\mathcal{M}(Q, P) = \sup_{m \in \mathcal{M}} \left| \int m(\zeta)(Q(\zeta) - P(\zeta))d\zeta \right|, \tag{12}$$

where ζ is the concatenation of $\mathbf{S}_t$ and $\mathbf{A}_t$, and $\mathcal{M}$ is a family of functions $m : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$.

Theorem 4.2. *Assume Ψ be a one-to-one representation function. When the weight w satisfies IPM($P_w(\mathbf{S}_t, \mathbf{A}_t), P(\mathbf{S}_t)P(\mathbf{A}_t)$)=0, then w can equate factual and marginal errors, i.e.,*

$$\begin{aligned}&\mathbb{E}_{\bar{\mathbf{H}}_t \sim P(\bar{\mathbf{H}}_t|\mathbf{A}_t)}[wL(f(\mathbf{S}_t, \mathcal{K}^t), \mathbf{Y}_{t+1}(\mathbf{A}_t))] = \\ &\mathbb{E}_{\bar{\mathbf{H}}_t \sim P(\bar{\mathbf{H}}_t)}[L(f(\mathbf{S}_t, \mathcal{K}^t), \mathbf{Y}_{t+1}(\mathbf{A}_t))].\end{aligned} \tag{13}$$

Theorem 4.2 indicates that by minimizing the distance between the weighted distribution $P_w(\mathbf{S}_t, \mathbf{A}_t)$ and $P(\mathbf{S}_t)P(\mathbf{A}_t)$, we can obtain a set of weights (over time) that make the adjusted factual error close to the marginal error. In this study, we choose the family of 1-Lipschitz functions, where the employed IPM is known as the wasserstein distance [Villani and others, 2009]. For detailed proofs, please see Appendix D.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Preprint

Algorithm 1 Pseudocode of Training CTD-NKO

```
Require: $\mathcal{D} = \left\{\{\mathbf{x}_t^{(i)}, \mathbf{a}_t^{(i)}, \mathbf{y}_t^{(i)}\}_{t=1}^{T^{(i)}} \cup \{\mathbf{v}^{(i)}\}\right\}_{i=1}^{N}$, $\theta := \{\theta_\Phi, \theta_X, \theta_Y\}$, $l$, $p_{\text{max}}$, $c$, $j = 0$, $\lambda_\mathcal{K}$, $\lambda_X$, $\beta$
 1: Initialize $\theta^0_{\text{EMA}}$, set weights $\left\{\{w_t^{(i)}\}_{t=1}^{T^{(i)}}\right\}_{i=1}^{N}$ uniformly
 2: for $p = 1$ to $p_{\text{max}}$ do
 3:     if $p \bmod c = 0$ then
 4:         Update weights $\left\{\{w_t^{(i)}\}_{t=1}^{T^{(i)}}\right\}_{i=1}^{N}$
 5:     end if
 6:     $\lambda_{\mathcal{K}_p} = \lambda_\mathcal{K}\left(\frac{2}{1+\exp(-10\cdot(p/p_{\text{max}}))} - 1\right)$
 7:     for each batch $\mathcal{B}$ do
 8:         $\mathcal{L}_Y^\mathcal{B} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{t=1}^{T^{(i)}} w_t^{(i)}\mathcal{L}_Y^{(i)}(\theta_Y, \theta_\Phi, t)$
 9:         $\mathcal{L}_X^\mathcal{B} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{t=1}^{T^{(i)}} \mathcal{L}_X^{(i)}(\theta_X, \theta_\Phi, t)$
10:         $\mathcal{L}_\mathcal{K}^\mathcal{B} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{t=1}^{T^{(i)}} \mathcal{L}_\mathcal{K}^{(i)}(\theta_\Phi, t)$
11:         $\theta_\Phi \leftarrow \theta_\Phi - l(\frac{\partial\mathcal{L}_Y^\mathcal{B}}{\partial\theta_\Phi} + \lambda_X\frac{\partial\mathcal{L}_X^\mathcal{B}}{\partial\theta_\Phi} + \lambda_{\mathcal{K}_p}\frac{\partial\mathcal{L}_\mathcal{K}^\mathcal{B}}{\partial\theta_\Phi})$
12:         $\theta_Y \leftarrow \theta_Y - l\frac{\partial\mathcal{L}_Y^\mathcal{B}}{\partial\theta_Y}$
13:         $\theta_X \leftarrow \theta_X - l\lambda_X\frac{\partial\mathcal{L}_X^\mathcal{B}}{\partial\theta_X}$
14:         $j \leftarrow j + 1$
15:         $\theta^j_{\text{EMA}} = \beta\theta^{j-1}_{\text{EMA}} + (1-\beta)\theta$
16:     end for
17: end for
Ensure: Optimized parameters $\theta$, EMA parameters $\theta_{\text{EMA}}$
```

4.4 Training and Inference

Algorithm 1 presents the pseudocode for the training process of CTD-NKO, with inputs including the observed data $\mathcal{D}$ and other necessary parameters, and outputs being the optimized model parameters. For detailed parameter settings, please refer to Appendix H. During training, we adopt the Exponential Moving Average (EMA) strategy [Tarvainen and Valpola, 2017] to obtain more reliable results, following the CT study [Melnychuk *et al.*, 2022]. The first line of Algorithm 1 initializes the EMA parameters θ^0_{EMA}, and in Line 15, the parameters are iteratively updated as follows:

$$\theta^j_{\text{EMA}} = \beta\theta^{j-1}_{\text{EMA}} + (1-\beta)\theta^j, \tag{14}$$

where j denotes the iteration number and β denotes the exponential smoothing factor.

In Lines 3-5 of Algorithm 1, we update the weights every c epochs. The weight calculation is based on Theorem 4.2, which involves minimizing the IPM distance between the weighted joint distribution and the marginal distribution of the system state representation and the current treatment. We obtain samples from the joint distribution and the marginal distribution by applying the actual treatment and the shuffled treatment to the observed data, respectively. Then, we learn weights by minimizing the wasserstein distance between the weighted joint distribution and the marginal distribution. For specific details, please refer to Appendix G. The model parameters are updated in Lines 8-13 of Algorithm 1, with the aim of minimizing the following objective function:

$$\mathcal{L} = \frac{1}{N}\sum_{i\in\mathcal{D}}\sum_{t=1}^{T^{(i)}}(w^{(i)}\mathcal{L}_Y^{(i)}(\theta_Y, \theta_\Phi, t) + \lambda_X\mathcal{L}_X^{(i)}(\theta_X, \theta_\Phi, t) + \lambda_{\mathcal{K}_p}\mathcal{L}_\mathcal{K}(\theta_\Phi, t). \tag{15}$$

Here, λ_X denotes a predefined weight coefficient, and $\lambda_{\mathcal{K}_p}$ denotes a weight coefficient that gradually increases with the training epochs. This design takes into account that the learning of $\mathbf{s}_t$ may not be accurate enough in the early stages of training; therefore, the importance of $\mathcal{L}_\mathcal{K}(\theta_\Phi, t)$ is reduced during the early training phase. We implement CTD-NKO using the Pytorch Lightning framework and employ the Adam algorithm [Kingma and Ba, 2014] for gradient optimization. After training, CTD-NKO performs one-step-ahead predictions and uses an autoregressive approach for multi-step-ahead predictions.

	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
RMSN	5.18±0.12	10.01±0.27	10.96±0.71	11.64±1.35	12.34±2.01	13.02±2.63
CRN	4.82±0.11	9.13±0.17	9.75±0.16	10.08±0.18	10.33±0.21	10.54±0.23
G-Net	5.05±0.06	11.92±0.19	12.96±0.23	13.65±0.27	14.15±0.28	14.59±0.32
CT	4.62±0.08	9.02±0.21	9.60±0.21	9.92±0.23	10.15±0.25	10.35±0.29
CTD-NKO	**4.57±0.10***	**8.97±0.17***	**9.55±0.18****	**9.86±0.19****	**10.08±0.21****	**10.27±0.24***

Table 1: Performance comparison of CTD-NKO with baseline models on the RW-MIMIC dataset: RMSE reported as mean ± standard deviation across five runs. Statistical significance was assessed using the Wilcoxon signed-rank test, with * and ** indicating p-values < 0.1 and < 0.05, respectively.

5 Experiments

In this section, we conduct a series of experiments to validate the effectiveness of the proposed CTD-NKO. Following the standard workflow of the Counterfactual Estimation Benchmark [Melnychuk *et al.*, 2022], we compare CTD-NKO with existing models on both synthetic and real-world datasets. We then analyze the efficiency of the baseline methods and CTD-NKO on various data settings. Finally, we investigate the role of different components of CTD-NKO through ablation studies.

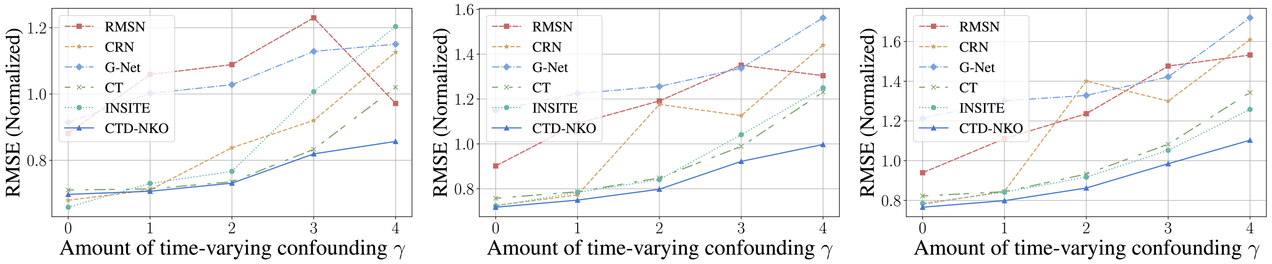
In this study, we select several state-of-the-art models from the recent literature on estimating time-varying counterfactuals to serve as comparative baselines. These include neural network-based models like **RMSN** [Lim *et al.*, 2018], **CRN** [Bica *et al.*, 2020], **G-Net** [Li *et al.*, 2021], and **CT** [Melnychuk *et al.*, 2022], as well as the non-neural network-based model **INSITE** [Kacprzyk *et al.*, 2024]. To guarantee a fair comparison, we perform hyperparameter tuning for these baseline methods (refer to Appendix H for details).

5.1 Counterfactual Estimation Performance Comparison

Experiments with FS-Tumor Dataset

Data. The FS-Tumor dataset has been widely adopted in previous studies evaluating counterfactual estimation over time, such as [Bica *et al.*, 2020; Melnychuk *et al.*, 2022; Lim *et al.*, 2018; Kacprzyk *et al.*, 2024]. In the dataset, a

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.



(a) 2-step-ahead prediction

(b) 4-step-ahead prediction

(c) 6-step-ahead prediction

Figure 2: The performance comparison of CTD-NKO and other baselines on the Tumor dataset for 2-step, 4-step, and 6-step ahead predictions under varying levels of time-varying confounders (γ) is conducted. The results are reported as the average RMSE over five runs.

parameter γ in the biomathematical model of tumor controls the time-varying confounding factors. An increase in the parameter γ signifies a greater influence of historical data on treatment allocation, implying a more pronounced presence of confounding biases. For more details, please refer to Appendix F

Results. In this study, we select the range of γ from 0 to 4. It is worth noting that $\gamma = 0$ represents a scenario where the treatment is assigned completely at random, indicating the absence of confounding bias. Figure 2 illustrates the comparison of τ-step-ahead prediction results, where $\tau \in \{2, 4, 6\}$. For more comprehensive experimental results, please refer to Appendix F. The results in Figure 2 demonstrate that CTD-NKO exhibits the best prediction accuracy in the majority of cases, with the lowest Root Mean Square Error (RMSE). This advantage becomes more pronounced as the confounding bias increases, i.e., when γ is larger.

Experiments with MIMIC-III Datasets

Data. MIMIC-III is a comprehensive database that encompasses electronic health records of patients in the intensive care unit and has been widely utilized to evaluate the performance of various models in complex real-world medical settings. In accordance with the related studies [Hatt and Feuerriegel, 2021; **?**; **?**], we construct a Real-World MIMIC (RW-MIMIC) dataset by selecting 25 dynamic patient covariates and 3 static features from the MIMIC-III database. Our primary focus was to investigate the impact of two common treatments, i.e., vasopressors and mechanical ventilation, on patients' blood pressure.

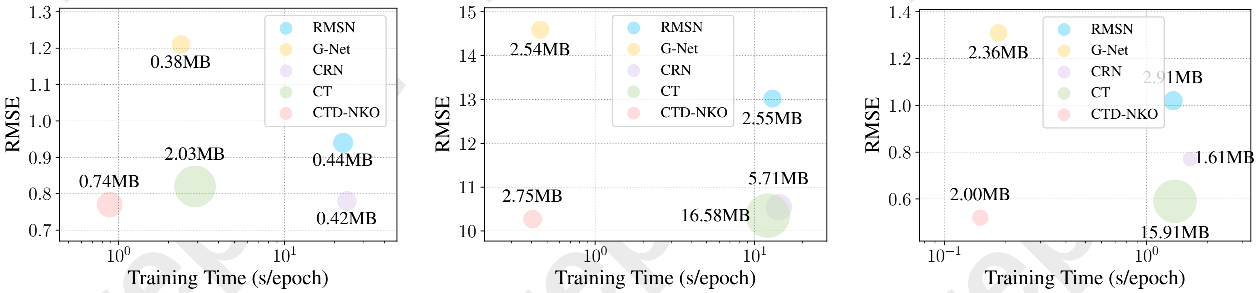
As a real-world data source, MIMIC-III does not contain information on counterfactual outcomes. To control for confounding bias and obtain counterfactuals for evaluation while analyzing high-dimensional patient trajectories, we construct a semi-synthetic dataset based on MIMIC-III, termed SS-MIMIC, following the methodology described in [Melnychuk *et al.*, 2022]. This dataset is generated considering treatment effects, endogenous dependencies, and exogenous dependencies, building upon the research methods of [Schulam and Saria, 2017]. For more details on these two datasets, please refer to Appendix F

Results. Table 1 and Table 2 present the τ-step-ahead prediction results on the RW-MIMIC and SS-MIMIC datasets, respectively. The results demonstrate that CTD-NKO consistently exhibits superior performance on both datasets, achieving the lowest RMSE and a relatively small standard deviation. These experimental findings suggest that CTD-NKO can effectively manage complex long-term dependencies and may be well-suited for applications that closely mirror real-world complexities. Notably, we do not compare INSITE on the MIMIC-III dataset due to the challenges posed by its high-dimensional, time-varying covariates. For more detailed reasons, please refer to Appendix G.

5.2 Model Efficiency Analysis

In practical applications, both operational efficiency and predictive performance are crucial. Hence, we evaluate the CTD-NKO against neural network-based models across several datasets in three key aspects: the $\tau_{\max}$-step-ahead prediction RMSE, training speed, and peak GPU memory usage. The term 'peak GPU memory usage' refers to the maximum amount of memory utilized by a GPU during training. This metric is instrumental in assessing a model's efficiency and scalability [Ikuzawa *et al.*, 2016; Nie *et al.*, 2022; Bergner *et al.*, 2023], thereby enabling optimal hardware utilization and cost-effectiveness in applications.

Figure 3 presents the comparative results: On the FS-Tumor, RW-MIMIC, and SS-MIMIC datasets, compared to the state-of-the-art model CT, CTD-NKO reduces training time per epoch by 69.3%, 96.6%, and 89.1% respectively, while the peak GPU memory usage is only 36.4%, 16.6%, and 12.6% of that consumed by the CT model. The performance advantage of the CT model stems from its adoption of the more powerful Transformer architecture; however, this also leads to larger memory consumption, particularly when handling complex data. In contrast, models based on LSTM architecture, while less demanding in terms of memory, tend to exhibit slightly inferior predictive performance. The CTD-NKO model, applying the Koopman theory and utilizing a straightforward LSTM architecture, achieves a good balance between performance and computational cost. Furthermore, CTD-NKO, similar to G-Net, utilizes an autoregressive recursive strategy for multi-step-ahead predictions. This strategy, focusing solely on encoder training, boosts efficiency compared to the CRN's encoder-decoder method, which requires training multiple components.



(a) FS-Tumor dataset ($\gamma = 0$) (b) RW-MIMIC dataset (c) SS-MIMIC dataset

Figure 3: Comparison of neural network-based model efficiency: Performance metrics are presented for τ_{max}-step-ahead predictions on the FS-Tumor, RW-MIMIC, and SS-MIMIC datasets, with respective τ_{max} values of 6, 6, and 10. To facilitate an equitable comparison across different model configurations, peak GPU memory usage is normalized based on batch size.

	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
RMSN	0.23 ± 0.01	0.46 ± 0.01	0.59 ± 0.02	0.69 ± 0.03	0.77 ± 0.04	0.84 ± 0.06	0.90 ± 0.07	0.96 ± 0.07	0.99 ± 0.07	1.02 ± 0.07
CRN	0.29 ± 0.02	0.46 ± 0.01	0.57 ± 0.01	0.63 ± 0.01	0.66 ± 0.01	0.69 ± 0.01	0.70 ± 0.01	0.73 ± 0.01	0.75 ± 0.01	0.77 ± 0.01
G-Net	0.35 ± 0.01	0.66 ± 0.02	0.82 ± 0.02	0.94 ± 0.03	1.03 ± 0.04	1.10 ± 0.04	1.16 ± 0.05	1.22 ± 0.05	1.27 ± 0.06	1.31 ± 0.06
CT	0.20 ± 0.01	0.37 ± 0.00	0.44 ± 0.00	0.49 ± 0.01	0.52 ± 0.01	0.54 ± 0.01	0.55 ± 0.01	0.57 ± 0.01	0.58 ± 0.01	0.59 ± 0.01
CTD-NKO	**0.17 ± 0.01****	**0.34 ± 0.00****	**0.40 ± 0.01****	**0.44 ± 0.01****	**0.46 ± 0.01****	**0.48 ± 0.01****	**0.49 ± 0.01****	**0.50 ± 0.01****	**0.51 ± 0.01****	**0.52 ± 0.02****

Table 2: Performance comparison of CTD-NKO with baseline models on the SS-MIMIC dataset: RMSE reported as mean ± standard deviation across five runs. Statistical significance was assessed using the Wilcoxon signed-rank test, with ** indicating p-values < 0.05.

	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
CTD-NKO + CT	1.33±0.25	0.87±0.12	0.96±0.15	1.03±0.17	1.09±0.18	1.14±0.19
CTD-NKO + ours	**1.32±0.21**	**0.86±0.09**	**0.93±0.12**	**1.00±0.16**	**1.05±0.19**	**1.10±0.21**

Table 3: CTD-NKO with different balancing strategies: Results for the FS-Tumor dataset with $\gamma = 4$.

	$\gamma = 2$		$\gamma = 4$	
	$\tau = 1$	$\tau = 6$	$\tau = 1$	$\tau = 6$
CTD-NKO	0.88 ± 0.07	**0.86 ± 0.06**	**1.32 ± 0.21**	**1.10 ± 0.21**
w/o weighting	0.89 ± 0.09	0.88 ± 0.08	1.34 ± 0.20	1.16 ± 0.26
w/o global operator	**0.86 ± 0.06**	0.88 ± 0.12	1.33 ± 0.17	1.24 ± 0.25

Table 4: Ablation study results on various tumor datasets, with the results presented as RMSE in the form of mean ± standard deviation over five runs.

5.3 Ablation Study

To assess the significance of various components within the CTD-NKO model, we conduct ablation studies to compare the predictive performance of the full model against variants lacking specific components across different γ settings on the FS-Tumor dataset, as presented in Table 4. Specifically, 'w/o weighting' denotes the model's use of uniform weights by omitting the weighting balancing module; 'w/o global operator' indicates that only the causal Koopman operator is utilized for modeling the temporal state transitions.

When the time-varying confounding bias is larger, the disparity between the counterfactual and factual distributions increases. The experimental results demonstrate that under these conditions, the learned balancing weights of CTD-NKO can mitigate this issue, thus leading to a more pronounced performance improvement. Additionally, the global Koopman operator enhances the model's performance, with its impact growing more substantial at farther prediction steps τ and larger confounding bias levels. Finally, we compare the proposed weighting method with the domain confusion loss strategy introduced by CT on CTD-NKO and test it on FS-Tumor ($\gamma = 4$). The results in Table 3 demonstrate that our weighting strategy outperforms learning balanced representations, which is consistent with our theoretical analysis.

6 Conclusion

Our work emphasizes the importance of effectively modeling temporal treatment interactions and preserving historical treatment information in counterfactual estimation over time. We propose the Counterfactual Temporal Dynamics Network via Neural Koopman Operators (CTD-NKO), which models temporal counterfactual reasoning as an evolutionary process of a nonlinear dynamical system, capturing complex temporal treatment interactions and inherent rhythmic dynamics. Additionally, we introduce a novel weighting method to mitigate confounding bias while reducing the loss of historical treatment information. These designs enable CTD-NKO to outperform state-of-the-art methods in terms of performance and efficiency, providing a powerful tool for improving medical decision-making and optimizing treatment strategies.

Acknowledgements

This research is supported by National Key RD Program of China (No. 2021ZD0111700), National Nature Science Foundation of China (No. 62137002, 62176245), Special

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Foundation for Science and Technology Innovation and Entrepreneurship of CCTEG NO. 2020-2-TD-CXY006, Key Science and Technology Special Project of Anhui Province (No. 202103a07020002). We thank the anonymous reviewers for their constructive comments that help improve the manuscript.

References

[Bergner *et al.*, 2023] Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Iterative patch selection for high-resolution image recognition. In *The Eleventh International Conference on Learning Representations*, 2023.

[Bica *et al.*, 2020] Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020.

[Brunton *et al.*, 2016] Bingni W Brunton, Lise A Johnson, Jeffrey G Ojemann, and J Nathan Kutz. Extracting spatial–temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition. *Journal of neuroscience methods*, 258:1–15, 2016.

[Brunton *et al.*, 2022] Steven L Brunton, Marko Budišić, Eurika Kaiser, and J Nathan Kutz. Modern koopman theory for dynamical systems. *SIAM Review*, 64(2):229–340, 2022.

[Chevillon, 2007] Guillaume Chevillon. Direct multi-step estimation and forecasting. *Journal of Economic Surveys*, 21(4):746–785, 2007.

[Eivazi *et al.*, 2021] Hamidreza Eivazi, Luca Guastoni, Philipp Schlatter, Hossein Azizpour, and Ricardo Vinuesa. Recurrent neural networks and koopman-based frameworks for temporal predictions in a low-order model of turbulence. *International Journal of Heat and Fluid Flow*, 90:108816, 2021.

[Fan *et al.*, 2022] Fletcher Fan, Bowen Yi, David Rye, Guodong Shi, and Ian R Manchester. Learning stable koopman embeddings. In *2022 American Control Conference (ACC)*, pages 2742–2747. IEEE, 2022.

[Han *et al.*, 2020] Yiqiang Han, Wenjian Hao, and Umesh Vaidya. Deep learning of koopman representation for control. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1890–1895. IEEE, 2020.

[Hariton and Locascio, 2018] Eduardo Hariton and Joseph J Locascio. Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13):1716, 2018.

[Hassanpour and Greiner, 2020] Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020.

[Hatt and Feuerriegel, 2021] Tobias Hatt and Stefan Feuerriegel. Sequential deconfounding for causal inference with unobserved confounders. *arXiv preprint arXiv:2104.09323*, 2021.

[Ikuzawa *et al.*, 2016] Takuya Ikuzawa, Fumihiko Ino, and Kenichi Hagihara. Reducing memory usage by the lifting-based discrete wavelet transform with a unified buffer on a gpu. *Journal of Parallel and Distributed Computing*, 93:44–55, 2016.

[Johansson *et al.*, 2022] Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *The Journal of Machine Learning Research*, 23(1):7489–7538, 2022.

[Kacprzyk *et al.*, 2024] Krzysztof Kacprzyk, Samuel Holt, Jeroen Berrevoets, Zhaozhi Qian, and Mihaela van der Schaar. ODE discovery for longitudinal heterogeneous treatment effects inference. In *The Twelfth International Conference on Learning Representations*, 2024.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Koopman, 1931] Bernard O Koopman. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

[Kutz *et al.*, 2016] J Nathan Kutz, Joshua L Proctor, and Steven L Brunton. Koopman theory for partial differential equations. *arXiv preprint arXiv:1607.07076*, 2016.

[Li and Jiang, 2021] Mengnan Li and Lijian Jiang. Deep learning nonlinear multiscale dynamic problems using koopman operator. *Journal of Computational Physics*, 446:110660, 2021.

[Li *et al.*, 2021] Rui Li, Stephanie Hu, Mingyu Lu, Yuria Utsumi, Prithwish Chakraborty, Daby M. Sow, Piyush Madan, Jun Li, Mohamed Ghalwash, Zach Shahn, and Li-wei Lehman. G-net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 282–299. PMLR, 04 Dec 2021.

[Lim *et al.*, 2018] Bryan Lim, Ahmed Alaa, and Mihaela van der Schaar. Forecasting treatment responses over time using recurrent marginal structural networks. *Advances in neural information processing systems*, 31, 2018.

[Liu *et al.*, 2023] Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning non-stationary time series dynamics with koopman predictors. *Advances in Neural Information Processing Systems*, 36, 2023.

[Massin *et al.*, 2000] Martial M Massin, Krystel Maeyns, Nadia Withofs, Françoise Ravet, and Paul Gérard. Circadian rhythm of heart rate and heart rate variability. *Archives of disease in childhood*, 83(2):179–182, 2000.

[Melnychuk *et al.*, 2022] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, pages 15293–15329. PMLR, 2022.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

[Morton *et al.*, 2018] Jeremy Morton, Antony Jameson, Mykel J Kochenderfer, and Freddie Witherden. Deep dynamical modeling and control of unsteady fluid flows. *Advances in Neural Information Processing Systems*, 31, 2018.

[Morton *et al.*, 2019] Jeremy Morton, Freddie D Witherden, and Mykel J Kochenderfer. Deep variational koopman models: inferring koopman observations for uncertainty-aware dynamics modeling and control. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3173–3179, 2019.

[Nie *et al.*, 2020] Lizhen Nie, Mao Ye, Dan Nicolae, et al. VCNet and functional targeted regularization for learning causal effects of continuous treatments. In *Proceedings of the 9th International Conference on Learning Representations*, 2020.

[Nie *et al.*, 2022] Xiaonan Nie, Xupeng Miao, Zhi Yang, and Bin Cui. Tsplit: Fine-grained gpu memory management for efficient dnn training via tensor splitting. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 2615–2628, 2022.

[Robins and Hernan, 2008] James Robins and Miguel Hernan. Estimation of the causal effects of time-varying exposures. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*, pages 553–599, 2008.

[Roemhild *et al.*, 2022] Roderich Roemhild, Tobias Bollenbach, and Dan I Andersson. The physiology and genetics of bacterial responses to antibiotic combinations. *Nature Reviews Microbiology*, 20(8):478–490, 2022.

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

[Schulam and Saria, 2017] Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. *Advances in neural information processing systems*, 30, 2017.

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

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

[Taieb and Atiya, 2015] Souhaib Ben Taieb and Amir F Atiya. A bias and variance analysis for multistep-ahead time series forecasting. *IEEE transactions on neural networks and learning systems*, 27(1):62–76, 2015.

[Takeishi *et al.*, 2017] Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning koopman invariant subspaces for dynamic mode decomposition. *Advances in neural information processing systems*, 30, 2017.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

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

[Wang *et al.*, 2023] Rui Wang, Yihe Dong, Sercan O Arik, and Rose Yu. Koopman neural operator forecaster for time-series with temporal distributional shifts. In *The 11th International Conference on Learning Representations*, 2023.

[Yazdani and Boerwinkle, 2015] Afsaneh Yazdani and Eric Boerwinkle. Causal inference in the age of decision medicine. *Journal of data mining in genomics & proteomics*, 6(1), 2015.

[Yeung *et al.*, 2019] Enoch Yeung, Soumya Kundu, and Nathan Hodas. Learning deep neural network representations for koopman operators of nonlinear dynamical systems. In *2019 American Control Conference (ACC)*, pages 4832–4839. IEEE, 2019.