An Empirical Examination of Balancing Strategy for Counterfactual Estimation on Time Series

Qiang Huang [1 2] **Chuizheng Meng** [3] **Defu Cao** [3] **Biwei Huang** [4] **Yi Chang** [1 2 5] **Yan Liu** [3]

Abstract

Counterfactual estimation from observations represents a critical endeavor in numerous application fields, such as healthcare and finance, with the primary challenge being the mitigation of treatment bias. The balancing strategy aimed at reducing covariate disparities between different treatment groups serves as a universal solution. However, when it comes to the time series data, the effectiveness of balancing strategies remains an open question, with a thorough analysis of the robustness and applicability of balancing strategies still lacking. This paper revisits counterfactual estimation in the temporal setting and provides a brief overview of recent advancements in balancing strategies. More importantly, we conduct a critical empirical examination for the effectiveness of the balancing strategies within the realm of temporal counterfactual estimation in various settings on multiple datasets. Our findings could be of significant interest to researchers and practitioners and call for a reexamination of the balancing strategy in time series settings.

1. Introduction

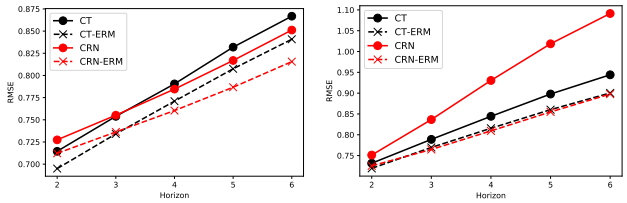
Temporal counterfactual outcome estimation (Morgan & Winship, 2015; Hernán & Robins, 2010; Pearl, 2009; Brodersen et al., 2015; Liu et al., 2022) aims to predict what the outcome would have been under different treatment scenarios. It is a crucial task in various real-world applications, such as health care (Prosperi et al., 2020; Richens et al., 2019), finance (Lundberg & Frost, 1992; Dhar, 1998; Castro-Iragorri, 2019), and e-commerce (Hua et al., 2021; Goswami et al., 2022). For example, in personalized medicine (Jain, 2002; Pazzagli et al., 2018; Porcher et al., 2019), counterfactual outcome provides a more comprehensive understanding of how the patient may respond to each treatment over time, allowing for a data-driven decision regarding the most suitable treatment strategy; In e-commerce, counterfactual inference can give companies guidance on when and to whom to issue coupons to different groups of users for increasing sales.

Estimating counterfactual outcomes is inherently challenging, principally for two reasons. Firstly, the intrinsic nature of observed data precludes the direct observation of counterfactuals, rendering their estimation a complex endeavor as acknowledged in the literature (Mandel & Lehman, 1996; Pearl, 2009; Boninger et al., 1994; Hassanpour & Greiner, 2019). These outcomes, delineating the hypothetical scenario under alternative treatments, are intrinsically unobservable. Furthermore, the confounding variables further convolute this task (McNamee, 2003; VanderWeele & Shpitser, 2013; Jager et al., 2008), since they affect both the treatment and the outcome, leading to treatment bias and obscuring the true causal effect. This results in a disparity between the distributions of observed (factual) and unobserved (counterfactual) outcome and makes accurate estimation of counterfactuals extremely difficult.

To address above challenges, the research community harnessed balancing technologies and developed a series of work to address the confounding issue. Strategies like Inverse Probability of Treatment Weighting (IPTW) (Chesnaye et al., 2022; Allan et al., 2020), stratification (Imbens & Rubin, 2015; Miratrix et al., 2013) and matching (Abadie et al., 2004; Abadie & Imbens, 2006), create pseudo-populations to mitigate distributional discrepancies between treatment groups. Alternative approaches, such as G-computation (Robins, 1986; 1987), iteratively simulate potential outcomes based on Monte Carlo. With the advance of deep learning, a new wave of methodologies has emerged to uncover complex, nonlinear relationships in data. These methods, as explored in studies by (Shalit et al., 2017; Hartford et al., 2017; Atan et al., 2018; Yao et al., 2018; Hassanpour & Greiner, 2019), are geared towards encoding

[1]School of Artificial Intelligence, Jilin University, Changchun, Jilin, China [2]International Center of Future Science, Jilin University, Changchun, Jilin, China [3]Department of Computer Science, University of Southern California, California, Los Angeles, United States [4]Halicioğlu Data Science Institute, University of California San Diego, San Diego, California, United States [5]Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MOE, Changchun, Jilin, China. Correspondence to: Yi Chang <yichang@jlu.edu.cn>, Yan Liu <yanliu.cs@usc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



(a) Slight treatment bias. (b) Significant treatment bias.

Figure 1. Performance comparison between balanced models (Causal Transformer and CRN) and their ERM variants in multi-step counterfactual estimation on Tumor Growth dataset, the X-axis represents Horizon, Y-axis represents RMSE, lower is better.

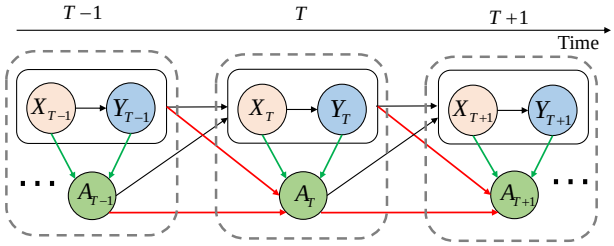


Figure 2. Causal structure in temporal setting, take time step $T-1$ to $T+1$ as an example, where green and red arrows denote the current and time-varying treatment bias, respectively.

covariates into a latent space. The goal is to derive representations that are devoid of treatment-related information and minimize the correlations between the representations and the treatment.

Despite the outlined effectiveness of balancing strategies in estimating counterfactual outcomes, our initial empirical investigation into the time series scenarios unveiled a counterintuitive phenomenon: in a temporal setting, models that forego balancing strategies—specifically, empirical risk minimization (ERM)—demonstrate superior performance in counterfactual estimation tasks compared to their counterparts that impose balancing strategies, even in the presence of treatment bias. Figure 1 illustrates two examples, i.e., Causal Transformer (Melnychuk et al., 2022) and CRN (Bica et al., 2020a). We observed that their ERM variants consistently outperforms Causal Transformer and CRN in multi-step prediction tasks (in either slight and significant treatment bias conditions). This surprising observation deviates from the prevailing perception on the effectiveness of balancing strategies and calls for an in-depth examination.

In this study, we undertake a comprehensive examination of contemporary methodologies devised to mitigate treatment bias via balancing techniques in time series analysis. We not only demonstrate how these balancing strategies perform but also discuss the underlying reasons through more in-depth empirical analyses. This paper elucidates the development in balancing methods for temporal counterfactual estimation, sheds light into when and how balance can be achieved in time series scenario, and discusses when it may be an elusive goal through empirical studies. We hope that the analysis results provide better understanding on balancing strategy in temporal counterfactual estimation and useful insights to researchers and practitioners.

The organization of this paper is as follows: Section 2 provides a review of the problem setting and basic assumptions of temporal counterfactual outcomes. In Section 3, we briefly revisit the advancements in balancing strategies from the perspective of balancing types and discusses popular methods in practical applications. Section 4 conducts empirical studies of these temporal causal models on benchmark datasets and discusses the effectiveness of balancing strategies in a temporal setting. Finally, Section 5 summarizes our empirical analysis and sheds light into promising directions to fundamentally address the problem of counterfactual estimation for time series.

2. Foundations

Notations. Let i denote the i-th individual (a single unit of analysis within a study, e.g., a patient) with historical trajectories over multiple time steps from $t = 1, ...T^{(i)}$. For each time step t, each individual i has the following observations: the time-varying covariates $\boldsymbol{X}_t^{(i)} \in \mathcal{R}^{d_x}$, where d_x is the dimension of the time-varying covariates; the static covariates $\boldsymbol{V}^{(i)}$ that do not change over time; the treatment assignments $\boldsymbol{A}_t^{(i)} \in \{a_1, ..., a_{d_a}\}$ where d_a is the number of treatment variables; the outcome $\boldsymbol{Y}_t^{(i) \in \mathcal{R}^{d_Y}}$. Then we let $\mathcal{D} = \{\{\boldsymbol{x}_t^{(i)}, \boldsymbol{a}_t^{(i)}, \boldsymbol{y}_t^{(i)}\}_{t=1}^{T^{(i)}}\}_{i=1}^{N}$ denote the observation dataset. For simplicity, we omit the superscript (i) for each individual unless needed and use $\boldsymbol{X}_t$ to represent both static and time-varying covariates at time step t.

Causal Structure. Figure 2 illustrates the causal structure and relationships between key variables in a temporal setting. X_t represents the time-varying covariates at time step t, Y_t denotes the outcome at time t, and A_t is the treatment assigned at time step t. The diagram shows that at each time step, the treatment assignment A_t is influenced by the time-varying covariates X_t as well as the history, indicating the presence of time-varying confounders. The outcome Y_t is affected by the historical treatment A_{t-1}, previous outcome Y_{t-1}, and current covariates X_t.

Target: Estimating $\mathbb{E}(\boldsymbol{y}_{t+\tau}[\bar{\boldsymbol{a}}_{t:t+\tau-1}]|\bar{\boldsymbol{H}}_t)$, where $\bar{\boldsymbol{H}}_t = (\bar{\boldsymbol{X}}_t, \bar{\boldsymbol{A}}_{t-1}, \bar{\boldsymbol{Y}}_t, \boldsymbol{V})$ is the observed history. $\bar{\boldsymbol{X}}_t = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t)$, $\bar{\boldsymbol{A}}_{t-1} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{t-1})$, $\bar{\boldsymbol{Y}}_t = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_t)$, and $\boldsymbol{V} = \boldsymbol{v}$. $\tau \geq 1$ denotes the projection horizon for a τ-step-ahead prediction. $\bar{\boldsymbol{a}}_{t:t+\tau-1} = (\boldsymbol{a}_t, \boldsymbol{a}_{t+1}, \ldots, \boldsymbol{a}_{t+\tau-1})$ is the sequence of the applied treatments in the future τ discrete time steps.

Table 1. Summary of partial literature for counterfactual estimation.

Taxonomy	Methods	Blancing Type	Balancing Tech	Backbone Type
Non-temporal	Statistical Method (Re-weighting,Stratification,Matching)	Pseudo-Population	Re-sampling	N/A
	CB-IV	IV-Regression	IPM Balancing	DNN
	Deep-Treat	Representation-Based	Propensity Inverse	DNN
	CITE	Representation-Based	Contrastive Balancing	DNN
	CFR	Representation-Based	IPM Balancing	DNN
	MIM-DRCFR	Representation-Based	MI Minimization	DNN
Temporal	MSM	Reweighting	Propensity Inverse	Linear Regression
	GNET	G-compupation	Monte-Carlo	LSTM
	RMSN	Reweighting	Propensity Inverse	LSTM
	CRN	Representation-Based	Gradient Reversal	LSTM
	Causal Transformer (CT)	Representation-Based	Domain Confusion	Transformer

Statistical methods (Rosenbaum & Rubin, 1983; Miratrix et al., 2013; Abadie et al., 2004), CB-IV (Wu et al., 2022), Deep-Treat (Atan et al., 2018), CITE (Li & Yao, 2022), CFR (Shalit et al., 2017), MIM-DRCFR (Cheng et al., 2022), MSM(Robins et al., 2000),GNET(Li et al., 2021),RMSN (Lim, 2018),CRN (Bica et al., 2020b),CT(Melnychuk et al., 2022), IV denotes the Instrumental Variable, IPM represents the Integral Probability Metric and MI denotes the Mutual Information.

3. Balancing Techniques

Table 1 provides a concise overview of counterfactual estimation methods from the perspective of balancing strategy. These methods vary in their balancing techniques, ranging from re-sampling and propensity inverse to more complex strategies like gradient reversal and domain confusion. It can be seen that the development of the solutions follow the paths towards leveraging advanced machine learning for causal inference. The table only lists a selection of representative methods. Many other models based on the balancing strategies not mentioned in the table, such as (Hartford et al., 2017; Yao et al., 2018; Hassanpour & Greiner, 2019; Gilbertson et al., 2016; Wodtke, 2020; Liu et al., 2020), because the balancing strategies adopted by these unlisted methods are already included in Table 1.

Representation-based balancing strategies have received significant attention in recent years, especially those based on deep learning, given its high performance in estimation accuracy. Therefore, in this work, we focus our discussion on representation-based balancing strategies for counterfactual estimation. Specifically, we choose the following three popular balancing methods for our empirical study:

Adversarial Gradient Reversal (AGR) (Bica et al., 2020b). It aims to build a treatment classifier G_A taking the representation $\boldsymbol{h}_t$ as input, and maximize the following classification loss to obtain the representation $\boldsymbol{h}_t$ that is invariant to the treatment assignment:

$$\mathcal{L}_{G_A}(\theta_A, \theta_R) = -\sum_{j=1}^{d_a} \mathbb{I}_{(\boldsymbol{A}_t = a_j)} log G_A(\boldsymbol{h}_t; \theta_A), \quad (1)$$

where θ_R denotes the parameters for generating the representation $\boldsymbol{h}_t$, θ_A denotes the parameter of the treatment classifier G_A, $\mathbb{I}$ is the indicator function. The adversarial loss aims to make the inferred representation $\boldsymbol{h}_t$ is not predictive to the treatment assignment.

Counterfactual Domain Confusion (CDC) (Melnychuk et al., 2022). The CDC balancing method is designed to ensure these representations are non-predictive of the current treatment assignment. First, the CDC method is also developed to fit the treatment classifier network G_A using the representation $\boldsymbol{h}_t$ by minimizing classification loss:

$$\mathcal{L}_{G_A}(\theta_A, \theta_R) = -\sum_{j=1}^{d_a} \mathbb{I}_{(\boldsymbol{A}_t = a_j)} log G_A(\boldsymbol{h}_t; \theta_A), \quad (2)$$

then CDC method proposes to minimize the cross-entropy between a uniform distribution over treatment categorical space and predictions of G_A via the following objective:

$$\mathcal{L}_{conf}(\theta_A, \theta_R) = -\sum_{j=1}^{d_a} \frac{1}{d_a} log G_A(\boldsymbol{h}_t; \theta_A), \quad (3)$$

PS-based Contrastive Balancing (PCB) (Li & Yao, 2022). The balancing method utilizes a contrastive learning paradigm to infer the balanced representation based on the propensity score, which is a two-stage procedure. First, we pre-train a propensity score estimator $e(\cdot)$ supervised by the treatment and get the propensity score for every instance.

Then we can build the positive and negative sample sets $\boldsymbol{X}^+$, $\boldsymbol{X}^-$ based on the above calculated propensity scores, where $\boldsymbol{X}^+$ are identified by propensity scores near 0.5, suggesting a higher likelihood of random treatment assignment, and $\boldsymbol{X}^-$ are characterized by propensity scores closer to 0 or 1, indicating a more pronounced inclination towards a specific treatment. Then, for every anchor sample i, the contrastive

loss for balancing the representation is as follows:

$$\mathcal{L}_c = \sum_{i=1}^{N} -log \frac{exp(\boldsymbol{h}_i \cdot \boldsymbol{h}_k^+ / \tau)}{\sum_{j=1}^{K} exp(\boldsymbol{h}_i \cdot \boldsymbol{h}_j^- / \tau)}, \tag{4}$$

where $\boldsymbol{h}_k^+$ is the representation of the randomly selected one positive sample from $\boldsymbol{X}^+$, $\boldsymbol{h}_j^-$ is the representation for every negative sample from $\boldsymbol{X}^-$, τ determines how much the contrastive loss inclines to the hard negative samples. Minimizing the loss can force the representation of units similar to that of positive samples while different from that of negative ones.

4. Empirical Study

4.1. Experiment Settings

4.1.1. Baselines and Metric

Although there are many proposed sequential models for temporal counterfactual estimation, we focus on several state-of-the-art models for examination, including (1) Causal Transformer (CT) (Melnychuk et al., 2022): A transformer-based counterfactual outcome prediction model with domain confusion module (CDC) for balanced representations; (2) Counterfactual Recurrent Network(CRN) (Bica et al., 2020b): a sequence-to-sequence model with adversarial gradient reversal (AGR) for balancing; (3) Recurrent Marginal Structural Networks(RMSN) (Lim, 2018): an LSTM-based model with propensity reweighting to adjust for time-dependent confounders; (4) G-Net (Li et al., 2021): a sequential deep learning model based on G-computation; and (5) Marginal Structural Model (MSM) (Robins et al., 2000): a linear marginal structural model based on the inverse-probability-of-treatment weighted estimator.

To ensure a thorough evaluation, we also explore the impact of the PCB balancing strategy, which is not inherently temporal, on both CT and CRN models. We refer to the balanced representation module as BRM and designate models without BRM as ERM (Empirical Risk Minimization) for clarity in our discussion. We evaluate these models by the Root Mean Square Error (RMSE) (i.e., lower is better).

4.1.2. Datasets

Pure synthetic dataset. Synthetic datasets are generated through autoregressive iterations based on a predefined causal structure between variables, such as the data generation process in (Bica et al., 2020a). Such data closely aligns with the causal assumptions formulated by practitioners, making it highly suitable for use in evaluating temporal counterfactual estimation.

Tumor Growth Simulator (Lim, 2018; Bica et al., 2020b; Melnychuk et al., 2022): The simulator employs a currently

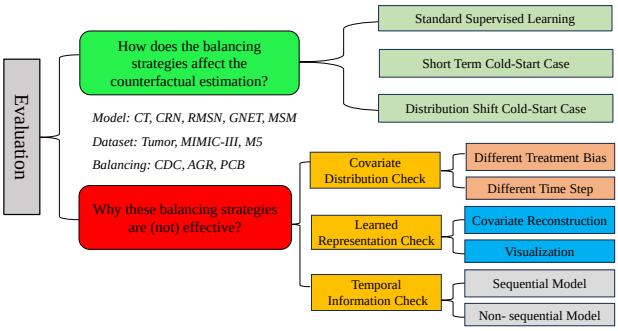


Figure 3. Road map of the experimental evaluation in this study.

popular and widely accepted biomedical model to simulate the temporal evolution of tumor volume, adhering to recognized practices in the field of biomedical research. The model consists of two types of binary treatment: (1) Radiotherapy when assigned to a patient has an immediate effect on the outcome of the next step; (2) Chemotherapy affects several future outcomes with certain exponentially decaying effects, and the tumor volume as outcome.

Semi-synthetic MIMIC III (Johnson et al., 2016): This dataset underwent standardized preprocessing procedures specifically designed for MIMIC-III data. It comprises ICU data aggregated at hourly intervals. The covariates within this dataset encompass 25 vital signs (dynamic features over time) and 3 static attributes, namely gender, ethnicity, and age. For additional information regarding to the data generation process, please refer to the Appendix of (Melnychuk et al., 2022).

M5 dataset (Makridakis et al., 2022). This dataset, provided by Walmart, captures the unit sales data of a diverse range of products sold across the United States, structured as grouped time series. It encompasses sales figures for 3,049 distinct products, which are segmented into three main categories: Hobbies, Foods, and Household. The original dataset does not have any label information on counterfactual outcomes. Therefore, we resort to factual evaluation and discuss the results in the Appendix due to space limit.

4.1.3. Evaluation

In temporal counterfactual estimation, we evaluate the models in a variety of scenarios to verify the effectiveness of different balancing strategies, including:

Standard supervised learning. The scenario involves using the full historical data, including previous covariates and treatment sequences, to predict future counterfactual outcomes over several time steps by using the standard train-test training procedure.

Short-term history cold-start. The scenario involves developing a short-term history test set for assessing the causal models. In this context, the model has access only to the covariates and treatment sequence information from K steps

before the predicted time step. This limitation in the extent of historical information available to the model is referred to as the truncation size.

Distribution shift cold-start. The scenario involves the case where there is distributional difference between the training data and the test data, particularly relevant in counterfactual estimation. It specifically pertains to the discrepancy in treatment bias between the data used during the training phase and that used in the testing phase.

This empirical study is structured around three critical questions: (1) *How do these temporal models perform for the counterfactual estimation?* (2) *Does the balanced representation module contribute to the temporal counterfactual estimation?* (3) *Why is (not) the balanced representation module effective for counterfactual estimation?* To enhance readability, the road map of the experimental evaluation for this examination study has been shown in Figure 3.

4.2. Counterfactual Estimation in Standard Supervised Learning

Results on synthetic Tumor dataset. Here, we report the performance of the counterfactual prediction on the Tumor Growth dataset. Specifically, let γ control the magnitude of the time-dependent confounding bias, i.e., the extent of correlation between covariates and treatment. Let τ denote the projection horizon for multi-step ahead prediction. In this evaluation setting, we set the $\gamma = 1, 3, 10$ to test the model performance in the varying magnitude of confounding bias, projection horizon to $\tau = 5$, and we report the average RMSE and standard deviation (STD) for 5 runs, the values of these parameters in this setting (also the subsequent setting) are chosen following previous work (Melnychuk et al., 2022). The results are shown in Table 2. From this table, we can have the following observations:

(1) When γ is relatively small (e.g., 1 and 3), the performance of balanced (e.g., CT and CT-PCB) and ERM models is similar. However, when γ becomes larger (e.g., 10), the performance of the non-balanced models is better than the balanced ones in most cases, and the PCB balancing module performs the worst.

(2) With the increase of γ, the variance of the performance for the models with BRM increases significantly. That means, the existence of BRM will introduce the high variance, making the models unstable in the case where the confounding bias is large.

(3) The performance of GNET, RMSN and MSM is undesirable when the γ is small. However, they perform better and are more stable methods when γ gets large (e.g., 10).

Overall, we did not observe a positive contribution from the balanced representation module to counterfactual estimation, instead tending to make the model's predictions unstable.

Experiments on semi-synthetic MIMIC III dataset. We adopt similar settings as in Tumor dataset to evaluate the performance of these sequential models on the semi-synthetic MIMIC III dataset, as shown in Table 3. We omit the experiment results of the MSM model because the model cannot converge on this dataset. Similar to the Tumor dataset, the balanced representation module fails to improve the performance on the MIMIC III dataset.

4.3. Counterfactual Estimation in Cold-start Case

Experimental results in standard supervised learning contradict previous claims about the effectiveness and robustness of balanced representation. Cold-start cases pose significant challenges to counterfactual estimation, while the balanced representation module could help alleviate the issue. Thus we further examine the following evaluation scenarios:

Short term history cold-start. We created a short-term history test set to evaluate causal models' performance in cold start situations. We used a history of truncation size, assessing models' ability to predict counterfactual outcomes with limited historical data. We report the results for $\gamma = 1, 3, 10$ in Table 4. It can seen that even under the cold start case, the models with balancing modules fail to help with counterfactual estimation.

Distribution shift cold-start. In this case, we trained models using data with significant confounding bias ($\gamma = 10$) and tested their performance on data with varying bias levels ($\gamma = 1, 3, 8$). This approach aims to determine if balancing modules enhance counterfactual prediction when there's a distribution shift between the source and target domains, as those balancing modules are designed to balance representations across different groups. Results in Table 5 show that similar to the short-term history and standard supervised learning cases, there is little difference between balanced models and non-balanced models with low γ (e.g., 1). Meanwhile, the balancing module introduces high variance in cases of significant treatment bias (e.g., $\gamma = 8$).

4.4. Why Balancing Module Does Not Work for the Counterfactual Estimation?

The above study shows that using a balanced representation module in models does not help (and often worsens) in various scenarios. In this section, we analyze the covariate distribution among treatment groups and the models' learned representations, in order to offer insights into why the balancing module is ineffective.

Checking on covariate distribution. We investigate how covariate distributions vary among groups under different treatment bias levels. The balancing module aims to equalize these distributions for counterfactual estimation, but

Table 2. RMSE (mean ± std) with $\gamma = 1, 3, 10$ on Tumor dataset.

	Methods	τ=1	2	3	4	5
$\gamma = 1$	CT	0.808±0.064	0.715±0.071	0.754±0.067	0.791±0.059	0.832±0.059
	CT-PCB	0.901±0.056	0.829±0.060	0.898±0.078	0.965±0.097	1.033±0.117
	CT w/o BRM	0.794±0.068	0.695±0.057	0.734±0.056	0.771±0.054	0.808±0.051
	CRN	0.818±0.044	0.727±0.046	0.755±0.044	0.785±0.041	0.817±0.041
	CRN-PCB	1.071±0.103	0.962±0.154	0.961±0.144	0.974±0.133	0.992±0.120
	CRN w/o BRM	0.845±0.062	0.712±0.035	0.736±0.031	0.760±0.028	0.787±0.030
	RMSN	1.116±0.116	1.047±0.154	1.072±0.096	1.090±0.065	1.113±0.050
	GNET	0.867±0.070	0.982±0.079	1.150±0.119	1.219±0.146	1.255±0.176
	MSM	1.204±0.042	1.693±0.123	2.028±0.151	2.227±0.164	2.314±0.167
$\gamma = 3$	CT	1.033±0.122	0.827±0.097	0.913±0.127	0.984±0.135	1.043±0.145
	CT-PCB	1.615±0.292	1.750±0.553	2.003±0.638	2.178±0.686	2.290±0.701
	CT w/o BRM	1.003±0.108	0.785±0.084	0.855±0.109	0.916±0.128	0.965±0.142
	CRN	1.070±0.108	1.139±0.358	1.380±0.518	1.560±0.604	1.690±0.641
	CRN-PCB	1.582±0.120	1.426±0.259	1.472±0.286	1.549±0.322	1.599±0.350
	CRN w/o BRM	1.046±0.111	0.795±0.056	0.853±0.079	0.907±0.105	0.957±0.129
	RMSN	1.266±0.078	1.154±0.132	1.240±0.143	1.290±0.166	1.327±0.189
	GNET	1.334±0.299	1.114±0.092	1.243±0.095	1.286±0.106	1.311±0.116
	MSM	1.749±0.139	2.404±0.445	2.742±0.516	2.896±0.545	2.924±0.547
$\gamma = 10$	CT	5.746±1.709	6.722±2.875	7.326±2.989	7.497±2.941	7.493±2.829
	CT-PCB	5.742±1.603	6.387±2.988	6.896±3.292	7.015±3.433	7.132±3.479
	CT w/o BRM	4.395±0.958	4.338±1.939	4.840±1.997	4.991±1.818	5.080±1.736
	CRN	4.963±0.345	6.526±2.435	7.085±3.204	7.289±3.384	7.326±3.210
	CRN-PCB	5.280±0.619	8.982±2.904	10.403±3.541	10.617±3.603	10.304±3.394
	CRN w/o BRM	4.707±0.394	6.502±2.026	7.342±2.174	7.690±2.174	7.727±2.093
	RMSN	5.109±0.400	5.339±1.815	5.479±1.895	5.331±1.801	5.125±1.621
	GNET	3.893±0.367	4.010±1.307	4.958±1.595	5.375±1.725	5.465±1.730
	MSM	5.837±0.616	2.040±0.672	3.039±0.999	3.870±1.274	4.617±1.525

altering distributions might lower the model's accuracy for individual samples. If group differences in covariate distributions are not significant, the benefits of balancing may not outweigh its drawbacks. We examine the two covariates on this dataset averaging on history, using kernel density estimation (KDE) and a Gaussian kernel function to plot their distribution shapes for $\gamma = 0, 1, 3, 10$. Figure 4 shows the Gaussian fit distributions for two covariates. For the first covariate, significant distribution differences among groups only appear at higher treatment bias levels (γ equals 0, 1, 3 show minimal differences). In contrast, the distribution of the second covariate among different groups is almost identical, regardless of treatment bias intensity.

To check if our observation is general, we generate time-series data using the universal method from (Bica et al., 2020a) and analyze the covariate distribution across treatment groups. This involves generating initial covariates and treatments using Gaussian and Bernoulli distributions, and then creating time-series under the proposed causal structure. The generated dataset features 10-dimensional covariates, a maximum sequence length of 30, and two-dimensional treatment variables, following a p-order autoregressive process, the detailed data generation can be found in Appendix.

We analyzed covariate distribution differences across treatment groups under various gamma values (indicating treatment bias strength) in our generated time-series data. Due to space constraints, we only show the distribution for one covariate, but similar patterns exist in others. Using KDE, we plotted covariate distributions for time steps 1, 4, 8 and 16 under $\gamma = 0.2$ and 0.4. Figure 5 reveals that early in the time series (e.g., time-step=1,4), covariate distribution differences across groups grow with increasing gamma value. However, at later time steps (e.g., time-step=8, 16), these differences diminish, suggesting treatment bias lessens over time and making balance-focused methods less effective.

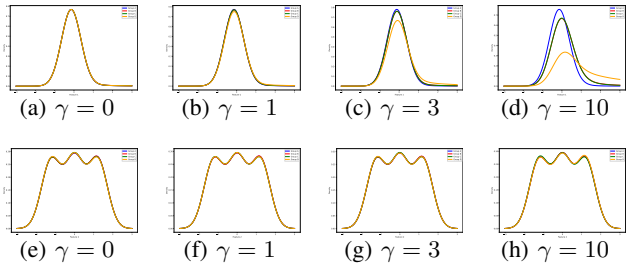
Learned representations for reconstruction of covariates. We analyze if models, specifically the Causal Transformer and CRN, can accurately reconstruct original covariates with and without the balancing module. The goal is to determine if the balancing module leads to information loss. After training these models using standard supervised learning, we compare the representations of test data before and after applying the balance module. We then use an LSTM-based decoder to reconstruct the covariates from these representations. The reconstruction's accuracy is measured using the MSE between the reconstructed and original covariates.

Table 3. RMSE (mean ± std) on MIMIC-III dataset.

Methods	τ=1	2	3	4	5	6
CT	0.25±0.06	0.42±0.09	0.55±0.15	0.65±0.21	0.73±0.25	0.79±0.29
CT-PCB	0.70±0.24	1.15±0.67	1.30±0.70	1.41±0.74	1.50±0.78	1.56±0.81
CT w/o BRM	0.23±0.06	0.41±0.08	0.53±0.13	0.63±0.17	0.70±0.21	0.75±0.23
CRN	0.24±0.03	0.50±0.08	0.68±0.15	0.82±0.24	0.97±0.38	1.13±0.54
CRN-PCB	0.30±0.06	0.97±0.74	1.13±0.81	1.30±0.86	1.45±0.92	1.59±0.99
CRN w/o BRM	0.22±0.02	0.47±0.13	0.63±0.20	0.77±0.29	0.89±0.39	1.01±0.49
RMSN	0.29±0.08	0.59±0.19	0.78±0.27	0.91±0.30	0.99±0.29	1.06±0.28
GNET	0.40±0.12	0.72±0.14	0.98±0.23	1.17±0.30	1.34±0.35	1.47±0.40

Table 4. RMSE (mean ± std) with $\gamma = 1, 3, 10$ for the cold start case of short-term history on Tumor dataset, lower is better.

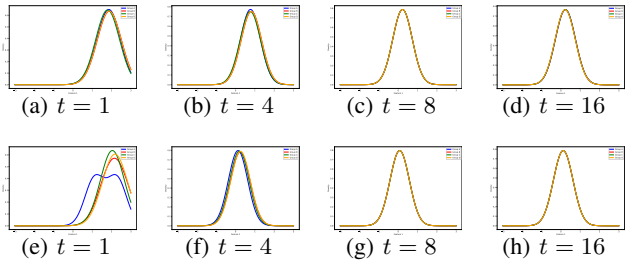
	Methods	τ=1	2	3	4	5
$\gamma = 1$	CT	0.827±0.052	0.715±0.074	0.770±0.077	0.819±0.075	0.869±0.075
	CT w/o BRM	0.822±0.055	0.716±0.064	0.770±0.062	0.820±0.062	0.873±0.066
	CRN	0.842±0.037	0.752±0.049	0.794±0.048	0.836±0.048	0.879±0.049
	CRN w/o BRM	0.868±0.077	0.733±0.055	0.769±0.055	0.805±0.055	0.841±0.057
	RMSN	1.134±0.117	1.031±0.143	1.079±0.099	1.121±0.084	1.165±0.083
	GNET	0.899±0.053	6.062±1.338	7.162±1.425	7.385±1.442	7.180±1.423
	MSM	1.242±0.050	1.906±0.102	2.286±0.122	2.508±0.129	2.600±0.125
$\gamma = 3$	CT	1.080±0.103	0.902±0.068	1.045±0.093	1.157±0.117	1.257±0.137
	CT w/o BRM	1.082±0.105	0.888±0.065	1.034±0.069	1.146±0.071	1.241±0.071
	CRN	1.147±0.103	1.330±0.274	1.660±0.418	1.900±0.477	2.073±0.493
	CRN w/o BRM	1.127±0.103	0.923±0.116	1.036±0.137	1.138±0.156	1.226±0.171
	RMSN	1.336±0.067	1.230±0.177	1.343±0.178	1.414±0.190	1.476±0.208
	GNET	1.394±0.287	4.869±0.286	6.032±0.519	6.546±0.749	6.746±0.947
	MSM	1.866±0.101	3.158±0.233	3.613±0.263	3.821±0.273	3.859±0.270
$\gamma = 10$	CT	5.385±1.876	5.732±2.804	7.312±3.635	8.222±3.640	8.573±3.542
	CT w/o BRM	4.234±0.828	2.706±0.483	3.959±0.510	4.112±0.456	4.873±0.772
	CRN	5.057±0.474	8.377±1.832	9.189±3.031	9.527±3.497	9.619±3.488
	CRN w/o BRM	4.798±0.508	8.620±1.676	9.737±2.008	10.177±2.083	10.199±1.986
	RMSN	5.313±0.562	7.433±1.207	7.429±1.171	7.116±1.060	6.737±0.913
	GNET	3.854±0.370	6.614±0.892	7.774±1.029	8.222±1.069	8.264±1.054
	MSM	5.980±0.742	2.663±0.465	3.962±0.684	5.042±0.864	6.016±1.029



(a) $\gamma = 0$ (b) $\gamma = 1$ (c) $\gamma = 3$ (d) $\gamma = 10$

(e) $\gamma = 0$ (f) $\gamma = 1$ (g) $\gamma = 3$ (h) $\gamma = 10$

Figure 4. Distribution shape for the two covariates under different strengths of treatment bias on Tumor. (a)-(f) for the first covariate, (g)-(l) for the second covariate.



(a) $t = 1$ (b) $t = 4$ (c) $t = 8$ (d) $t = 16$

(e) $t = 1$ (f) $t = 4$ (g) $t = 8$ (h) $t = 16$

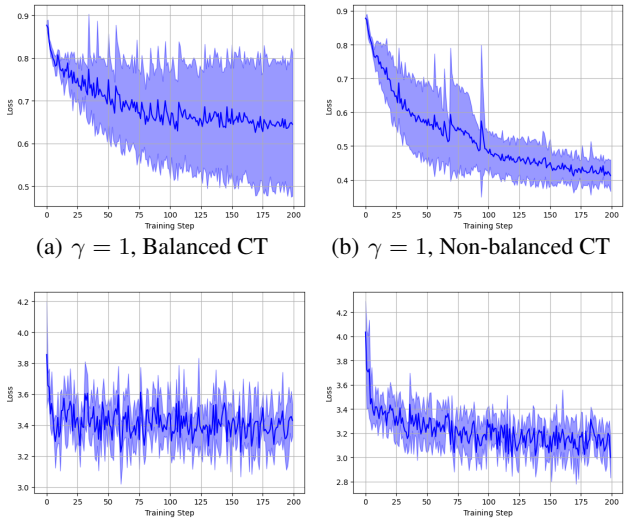
Figure 5. Distribution shape for the covariate under different strengths of treatment bias and time-steps on pure synthetic dataset, (a)-(d) for $\gamma = 0.2$, (e)-(h) for $\gamma = 0.4$.

Figure 6 shows the reconstruction loss for the Causal Transformer with balanced and non-balanced representations. At a high treatment bias ($\gamma = 10$), the balanced representation's error doesn't converge, suggesting a significant loss of original covariate information. Conversely, the non-balanced representation's error decreases, indicating better retention of covariate information. When $\gamma = 1$, both representations perform adequately, but the non-balanced one has a lower final loss, because, at low treatment biases, the balanced representation's negative impact is less noticeable. The results for the CRN model, not shown due to space constraints, follow a similar pattern.

Learned representation visualization with or without balanced module. We visualized the representations learned

Table 5. RMSE (mean ± std) with $\gamma = 1, 3, 8$ for the cold start case of distribution shift on Tumor dataset, lower is better.

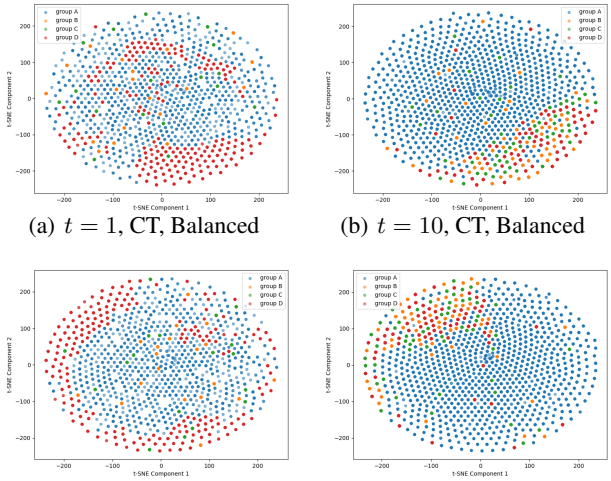
	Methods	τ=1	2	3	4	5
$\gamma = 1$	CT	1.654±1.190	2.491±2.630	3.030±3.469	3.452±4.001	3.595±4.049
	CT w/o BRM	0.941±0.077	0.991±0.084	1.076±0.105	1.177±0.129	1.263±0.138
	CRN	1.566±0.069	2.065±0.146	2.067±0.322	2.095±0.376	2.127±0.390
	CRN w/o BRM	1.339±0.079	1.308±0.068	1.466±0.098	1.575±0.103	1.631±0.094
	RMSN	1.374±0.097	1.188±0.072	1.220±0.061	1.273±0.074	1.338±0.115
	GNET	1.074±0.050	4.238±0.043	4.884±0.117	5.023±0.233	5.050±0.369
	MSM	1.160±0.033	0.507±0.043	0.763±0.063	0.976±0.081	1.167±0.096
$\gamma = 3$	CT	1.889±0.987	2.600±2.463	3.148±3.207	3.554±3.676	3.718±3.693
	CT w/o BRM	1.211±0.141	1.138±0.148	1.285±0.170	1.405±0.200	1.521±0.216
	CRN	1.810±0.098	2.231±0.396	2.375±0.668	2.456±0.752	2.506±0.740
	CRN w/o BRM	1.619±0.103	1.734±0.289	1.971±0.309	2.106±0.300	2.162±0.287
	RMSN	1.704±0.117	1.504±0.216	1.548±0.224	1.570±0.199	1.595±0.169
	GNET	1.297±0.096	3.909±0.075	4.313±0.202	4.372±0.335	4.387±0.458
	MSM	1.605±0.185	0.617±0.104	0.930±0.161	1.193±0.207	1.430±0.250
$\gamma = 8$	CT	4.942±1.938	5.723±3.801	6.532±4.389	6.937±4.525	7.091±4.416
	CT w/o BRM	3.495±0.622	3.287±1.403	3.776±1.577	3.925±1.445	4.080±1.442
	CRN	4.124±0.235	5.346±1.846	5.782±2.433	5.957±2.570	6.019±2.427
	CRN w/o BRM	3.871±0.281	5.218±1.717	5.932±1.849	6.249±1.847	6.311±1.775
	RMSN	4.163±0.255	4.201±1.535	4.326±1.581	4.240±1.486	4.123±1.324
	GNET	3.167±0.260	4.643±0.749	5.269±1.086	5.539±1.289	5.625±1.401
	MSM	4.788±0.438	1.608±0.544	2.403±0.813	3.065±1.040	3.659±1.249



(a) $\gamma = 1$, Balanced CT (b) $\gamma = 1$, Non-balanced CT

(c) $\gamma = 10$, Balanced CT (d) $\gamma = 10$, Non-balanced CT

Figure 6. Reconstruction loss curves for Causal Transformer.



(a) $t = 1$, CT, Balanced (b) $t = 10$, CT, Balanced

(c) $t = 1$, CT, Non-balanced (d) $t = 10$, CT, Non-balanced

Figure 7. Visualization for balanced and non-balanced representation for different time steps from Causal Transformer.

with and without balancing modules at various time steps to investigate their balancing differences. Given that previous results showed minor covariate distribution differences among treatment groups at low treatment bias intensities, we focus on visualizations at $\gamma = 10$ in the synthetic tumor dataset for a clearer demonstration of the balancing module's distribution adjustment capabilities. The representation visualization for the Causal Transformer is illustrated in Figure 7. We can see that the representations with or without the balancing module do not differ much in terms of balancing degree across different groups (similar observations are also made for CRN). However, it is evident that the distribution of representations after balancing significantly differs from that before balancing. In other words, the balancing module not only fails to eliminate treatment bias, but also leads to a decline in predictive performance due to the altered distribution of covariate representations.

The experiments above suggest that models with balanced module struggle to reconstruct covariates under high treatment bias since they lose essential information during training. Conversely, with lower treatment bias and minor differences in group covariate distributions, the balanced module can result in over-adjustment of variables and decreased predictive performance for counterfactual estimation.

Table 6. RMSE (mean ± std) comparison between sequential and non-sequential models in terms of different strengths of treatment bias.

Tumor		γ=0	1	2	3	6	8	10
Sequential	CT	0.307±0.072	0.259±0.049	0.444±0.142	0.935±0.178	3.657±1.670	4.336±1.335	3.827±1.952
	CT w/o BRM	0.299±0.071	0.325±0.105	0.456±0.109	0.852±0.234	2.406±0.646	3.108±0.307	3.891±3.241
	CRN	0.239±0.017	0.295±0.088	0.323±0.049	0.488±0.124	0.812±0.164	1.146±0.151	1.483±0.173
	CRN w/o BRM	0.243±0.018	0.271±0.042	0.282±0.042	0.496±0.220	0.902±0.127	1.197±0.173	1.324±0.133
Non-Sequential	CFR	0.340±0.024	0.390±0.036	0.511±0.103	0.749±0.139	1.727±0.235	2.107±0.278	2.419±0.297
	TARnet	0.330±0.025	0.393±0.040	0.512±0.097	0.767±0.124	1.827±0.236	2.382±0.286	2.842±0.330
	OLS1	0.557±0.017	0.617±0.033	0.738±0.056	0.951±0.075	1.667±0.131	1.872±0.136	1.866±0.128
	OLS2	0.594±0.036	0.590±0.009	0.595±0.031	0.629±0.038	0.800±0.021	0.892±0.020	0.959±0.046
	Random	0.824±0.025	0.868±0.034	0.940±0.055	1.076±0.062	1.538±0.120	1.724±0.122	1.835±0.131
MIMIC III		γ=0.25	1.0	2.5	5	10	15	50
Sequential	CT	0.084±0.015	0.108±0.016	0.128±0.012	0.152±0.023	0.188±0.049	0.226±0.066	0.490±0.128
	CT w/o BRM	0.085±0.014	0.106±0.017	0.125±0.010	0.145±0.014	0.177±0.045	0.202±0.031	0.480±0.130
	CRN	0.104±0.013	0.159±0.009	0.225±0.044	0.265±0.048	0.310±0.054	0.352±0.057	0.445±0.094
	CRN w/o BRM	0.099±0.012	0.122±0.013	0.133±0.016	0.147±0.028	0.188±0.081	0.238±0.108	0.503±0.137
Non-Sequential	CFR	0.265±0.058	0.307±0.055	0.328±0.051	0.336±0.049	0.341±0.049	0.341±0.044	0.366±0.052
	TARnet	0.271±0.057	0.318±0.059	0.349±0.061	0.358±0.060	0.364±0.054	0.375±0.048	0.428±0.046
	OLS1	0.465±0.091	0.525±0.071	0.535±0.072	0.545±0.076	0.559±0.080	0.566±0.083	0.586±0.097
	OLS2	0.300±0.042	0.353±0.038	0.385±0.049	0.403±0.065	0.405±0.073	0.403±0.065	0.437±0.078
	Random	0.353±0.004	0.427±0.038	0.440±0.070	0.443±0.081	0.452±0.087	0.458±0.088	0.470±0.090

4.5. When Are the Temporal Dependencies Informative for Counterfactual Estimation?

We examined the effectiveness of temporal dependency modeling in counterfactual outcome estimation by comparing the performance of sequential models (Causal Transformer, CRN, and their ERM variants) against classical non-sequential models. The non-sequential baselines included: (1) Counterfactual Regression (CFR) (Shalit et al., 2017); (2) TarNet (Shalit et al., 2017), a CFR variant without the balancing module; (3) OLS1, a linear regression model combining treatment assignment and covariates; (4) OLS2, with separate linear regression models for treatment and control groups; and (5) the Random method, predicting outcomes by randomly selecting from historical outcomes.

The results are shown in Table 6. When the treatment bias is relatively low, sequential models outperform non-sequential models and the Random method. However, as the treatment bias increases to a certain extent, the sequential models, as mentioned in the previous analysis, tend to lose a considerable amount of original covariate information due to excessive treatment bias. Consequently, the performance of counterfactual estimation falls behind that of non-sequential models, and in some cases, even the Random method. This suggests that temporal dependency information can be overshadowed when the treatment bias is too large, rendering it ineffective for model generalization.

5. Summary and Future Work

In this paper, we conduct a thorough empirical study on existing popular methods for counterfactual estimation in temporal settings. It has been observed that the effectiveness and applicability of representation-based balancing strategies could differ from previously reported results. We identify the limitations of existing representation balance techniques in various scenarios. Our analysis results suggest three promising directions to pursue when developing new solutions to temporal counterfactual estimation:

- **Trade-off between balancing and prediction accuracy**. Although balancing strategies can mitigate the impact of treatment bias, they inevitably disrupt the data distribution, which in turn affects the accuracy of outcome predictions. Therefore, one promising future direction is to investigate appropriate strategies to balance these two factors.

- **More stable balancing technologies**. Our analysis reveals that balancing strategies can lead to instability in temporal model training, resulting in high variance in prediction. Therefore, new balancing strategies that can offer strong stability should be explored.

- **Treatment bias checking in advance**. Excessive treatment bias can results in various challenges, such as masking important temporal information in the data and leading to overall model collapse. Furthermore, treatment bias may change over time (e.g., gradually diminishing). Therefore, exploring strategies to assess how significant treatment bias is and decide if the balancing strategies are needed could effectively help to mitigate the model collapse issue.

Acknowledgements

The authors would like to thank the anonymous referees for their valuable comments. The work is done while the first author is a visiting student at USC. Qiang Huang and Yi Chang are supported by the National Key R&D Program of China under Grant No.2023YFF0905400 and the National Natural Science Foundation of China (No.U2341229). Chuizheng Meng, Defu Cao, and Yan Liu are supported in part by NSF Research Grant # 2226087 and #1837131.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning and Causal Inference. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abadie, A. and Imbens, G. W. Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267, 2006.

Abadie, A., Drukker, D., Herr, J. L., and Imbens, G. W. Implementing matching estimators for average treatment effects in stata. *The stata journal*, 4(3):290–311, 2004.

Allan, V., Ramagopalan, S. V., Mardekian, J., Jenkins, A., Li, X., Pan, X., and Luo, X. Propensity score matching and inverse probability of treatment weighting to address confounding by indication in comparative effectiveness research of oral anticoagulants. *Journal of comparative effectiveness research*, 9(9):603–614, 2020.

Atan, O., Jordon, J., and Van der Schaar, M. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Bica, I., Alaa, A., and Van Der Schaar, M. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International Conference on Machine Learning*, pp. 884–895. PMLR, 2020a.

Bica, I., Alaa, A. M., Jordon, J., and van der Schaar, M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=BJg866NFvB.

Boninger, D. S., Gleicher, F., and Strathman, A. Counterfactual thinking: From what might have been to what may be. *Journal of personality and social psychology*, 67(2): 297, 1994.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. Inferring causal impact using bayesian structural time-series models. 2015.

Castro-Iragorri, C. Does the market model provide a good counterfactual for event studies in finance? *Financial Markets and Portfolio Management*, 33:71–91, 2019.

Cerulli, G. treatrew: A user-written command for estimating average treatment effects by reweighting on the propensity score. *The Stata Journal*, 14(3):541–561, 2014.

Cheng, M., Liao, X., Liu, Q., Ma, B., Xu, J., and Zheng, B. Learning disentangled representations for counterfactual regression via mutual information minimization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1802–1806, 2022.

Chesnaye, N. C., Stel, V. S., Tripepi, G., Dekker, F. W., Fu, E. L., Zoccali, C., and Jager, K. J. An introduction to inverse probability of treatment weighting in observational research. *Clinical Kidney Journal*, 15(1):14–20, 2022.

Dhar, V. Data mining in finance: using counterfactuals to generate knowledge from organizational information systems. *Information Systems*, 23(7):423–437, 1998.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Gilbertson, D. T., Bradbury, B. D., Wetmore, J. B., Weinhandl, E. D., Monda, K. L., Liu, J., Brookhart, M. A., Gustafson, S. K., Roberts, T., Collins, A. J., et al. Controlling confounding of treatment effects in administrative data in the presence of time-varying baseline confounders. *Pharmacoepidemiology and drug safety*, 25(3):269–277, 2016.

Goswami, P., Hoyos-Idrobo, A., Deschamps, T., and Feuillet, N. From theory to practice: Testing harvesting-based propensity estimators on counterfactual learning-to-rank from implicit feedback for e-commerce search. 2022.

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pp. 1414–1423. PMLR, 2017.

Hassanpour, N. and Greiner, R. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019.

Hernán, M. A. and Robins, J. M. Causal inference, 2010.

Hua, J., Yan, L., Xu, H., and Yang, C. Markdowns in e-commerce fresh retail: a counterfactual prediction and multi-period optimization approach. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3022–3031, 2021.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Jager, K., Zoccali, C., Macleod, A., and Dekker, F. Confounding: what it is and how to deal with it. *Kidney international*, 73(3):256–260, 2008.

Jain, K. K. Personalized medicine. *Current opinion in molecular therapeutics*, 4(6):548–558, 2002.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

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

Li, R., Hu, S., Lu, M., Utsumi, Y., Chakraborty, P., Sow, D. M., Madan, P., Li, J., Ghalwash, M., Shahn, Z., 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.

Li, X. and Yao, L. Contrastive individual treatment effects estimation. In *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 1053–1058. IEEE, 2022.

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

Liu, L., Wang, Y., and Xu, Y. A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 2022.

Liu, R., Yin, C., and Zhang, P. Estimating individual treatment effects with time-varying confounders. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 382–391. IEEE, 2020.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Lundberg, C. G. and Frost, D. E. Counterfactuals in financial decision making. *Acta Psychologica*, 79(3):227–244, 1992.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022.

Mandel, D. R. and Lehman, D. R. Counterfactual thinking and ascriptions of cause and preventability. *Journal of personality and social psychology*, 71(3):450, 1996.

McNamee, R. Confounding and confounders. *Occupational and environmental medicine*, 60(3):227–234, 2003.

Melnychuk, V., Frauen, D., and Feuerriegel, S. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, pp. 15293–15329. PMLR, 2022.

Miratrix, L. W., Sekhon, J. S., and Yu, B. Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(2):369–396, 2013.

Morgan, S. L. and Winship, C. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

Pazzagli, L., Linder, M., Zhang, M., Vago, E., Stang, P., Myers, D., Andersen, M., and Bahmanyar, S. Methods for time-varying exposure related problems in pharmacoepidemiology: an overview. *Pharmacoepidemiology and drug safety*, 27(2):148–160, 2018.

Pearl, J. Causal inference in statistics: An overview. 2009.

Porcher, R., Jacot, J., Wunder, J. S., and Biau, D. J. Identifying treatment responders using counterfactual modeling and potential outcomes. *Statistical methods in medical research*, 28(10-11):3346–3362, 2019.

Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., and Bian, J. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375, 2020.

Richens, J. G., Lee, C. M., and Johri, S. Counterfactual diagnosis. *arXiv preprint arXiv:1910.06772*, 2019.

Robins, J. 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.

Robins, J. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of chronic diseases*, 40:139S–161S, 1987.

Robins, J. and Hernan, M. Estimation of the causal effects of time-varying exposures. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*, pp. 553–599, 2008.

Robins, J. M. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods*, 23(8):2379–2412, 1994.

Robins, J. M., Hernan, M. A., and Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pp. 550–560, 2000.

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

Schulam, P. and Saria, S. Reliable decision support using counterfactual models. *Advances in neural information processing systems*, 30, 2017.

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

Soleimani, H., Subbaswamy, A., and Saria, S. Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. *arXiv preprint arXiv:1704.02038*, 2017.

Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4068–4076, 2015.

VanderWeele, T. J. and Shpitser, I. On the definition of a confounder. *Annals of statistics*, 41(1):196, 2013.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wodtke, G. T. Regression-based adjustment for time-varying confounders. *Sociological Methods & Research*, 49(4):906–946, 2020.

Wu, A., Kuang, K., Li, B., and Wu, F. Instrumental variable regression with confounder balancing. In *International Conference on Machine Learning*, pp. 24056–24075. PMLR, 2022.

Xu, Y., Xu, Y., and Saria, S. A bayesian nonparametric approach for estimating individualized treatment-response curves. In *Machine learning for healthcare conference*, pp. 282–300. PMLR, 2016.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. *Advances in neural information processing systems*, 31, 2018.

Appendix

A. Assumptions

The task of counterfactual outcome estimation is based on the potential outcome framework, where three fundamental assumptions ensure its causal identifiability. Here, we introduce the three basic assumptions of causal identifiability in a temporal setting.

Assumption A.1. Consistency. If $\bar{\boldsymbol{A}}_t$=$\bar{\boldsymbol{a}}_t$ represents a sequence of treatments assigned to a specific individual, then under this treatment sequence, the potential outcome $\bar{\boldsymbol{Y}}_{t+1}(\bar{\boldsymbol{A}}_t = \bar{\boldsymbol{a}}_t) = \boldsymbol{Y}_t$, where $\boldsymbol{Y}_t$ is the observed outcome conditioned on $\bar{\boldsymbol{A}}_t = \bar{\boldsymbol{a}}_t$. This implies that the potential outcome for an individual, given their observed exposure history, corresponds to the actual (factual) outcome that will be observed for that individual.

Assumption A.2. Overlap. In the entire historical timeline, there is a nonzero likelihood of either receiving or not receiving any treatment, i.e., given the history of an individual, the probability of treatment assignment holds $0 < P(\boldsymbol{A}_t = \boldsymbol{a}_t | \bar{\boldsymbol{H}}_t = \bar{\boldsymbol{h}}_t) < 1$ if $P(\bar{\boldsymbol{H}}_t = \bar{\boldsymbol{h}}_t) > 0$.

Assumption A.3. Unconfoundedness. Given the observed history, the current treatment assignment is independent of the potential outcome, i.e., $\boldsymbol{A}_t \perp \boldsymbol{Y}_{t+1}(\boldsymbol{a}_t) | \bar{\boldsymbol{H}}_t, \forall \boldsymbol{a}_t$. This implies that all the possible confounders that affect both treatment and outcome are observed.

B. Data Generation of Pure Synthetic Dataset

We generate random time-series data using the universal method from literature (Bica et al., 2020a), analyzing covariate distribution across treatment groups. This involves generating initial covariates and treatments using Gaussian and Bernoulli distributions, and then creating time-series data under the proposed causal structure. The generated dataset features 10-dimensional covariates, a maximum sequence length of 30, and the two-dimensional treatment variables, following a p-order autoregressive process:

$$\begin{aligned} X_t =& \frac{1}{p} \sum_{i=1}^{p} (\beta_i X_{t-i} + \sum_{j=1}^{k} \lambda_{i,j} A_{t-i,j} + \epsilon_t), \\ \pi_{tj} =& \gamma \sum_{i=1}^{p} X_{t-i}, \quad A_{tj} | \pi_{tj} \sim B(\sigma(\pi_{tj})), \\ Y_t =& \lambda_Y \cdot f(W \cdot Z_t) + \epsilon_Y, \end{aligned} \tag{5}$$

where k denotes the dimension of treatment variables, $\beta_i \sim \mathcal{N}(1 - (\frac{i}{p}), \frac{1}{p})$, $\epsilon \sim \mathcal{N}(0, 0.01)$; σ denotes the sigmoid active function, f denotes the nonlinear mapping function (e.g., tanh function), γ represents the strength of treatment bias and γ_Y controls the amount of confounding applied to the outcome.

C. Related works

Counterfactual outcome estimation under I.I.D setting. Due to the prohibitive costs and potential ethical concerns associated with randomized experiments (Guo et al., 2020), there has been a growing interest in recent years in causal inference from observational data, with a majority of existing studies primarily focusing on independent and identically distributed (i.i.d.) data. Traditional statistical methods aim to construct one or more pseudo-populations that differ from the original data to achieve a balance in the distribution of sample covariates among different groups, like sample re-weighting (Rosenbaum & Rubin, 1983; Cerulli, 2014) based on IPTW (Chesnaye et al., 2022), Stratification (Imbens & Rubin, 2015; Miratrix et al., 2013) and matching (Abadie et al., 2004; Abadie & Imbens, 2006). Some instrument variables-based methods have also been introduced. DeepIV (Hartford et al., 2017), for instance, introduced a two-stage approach comprising a first-stage network for treatment prediction and a second-stage network incorporating a loss function that encompasses integration over the conditional treatment distribution for counterfactual predictions. Confounder Balanced IV Regression (CB-IV) (Wu et al., 2022) jointly remove the bias from the unmeasured confounders and balance the observed confounders by twp stage IV regression. Recently, more methods have been proposed based on representation learning to enable more effective and more accurate counterfactual estimation. CFR (Shalit et al., 2017) frames counterfactual inference as a form of domain adaptation and employs neural networks to learn ITEs by creating balanced representations through the minimization of distribution differences between control and treated groups. CEVAE (Louizos et al., 2017) proposes a novel approach, mapping the original observed features to a latent space to capture hidden confounders using a variational

autoencoder (Kingma & Welling, 2014). Atan et al. (Atan et al., 2018) introduce Deep-Treat, which mitigates bias by learning representations and crafting effective treatment policies using deep neural networks on transformed data for counterfactual estimation. Yao et al. (Yao et al., 2018) introduce a method for counterfactual estimation, known as SITE, grounded in deep representation learning, capable of capturing hidden confounders and preserving the local similarity of data. Hassanpour et al. (Hassanpour & Greiner, 2019) proposed to learn three underlying sources of the observation data as the disentangled representations to account for the treatment bias and then conduct unbiased counterfactual outcome estimation. Then MIM-DRCFR (Cheng et al., 2022) was proposed to identify disentangled representations by mutual information minimization for the counterfactual outcome estimation.

Counterfactual outcome estimation over time. Early methods for temporal counterfactual outcome estimation primarily included G-computation, marginal structural models (MSMs), and structural nested models (Robins, 1986; 1994; Robins et al., 2000; Robins & Hernan, 2008). However, these methods heavily rely on linear assumptions to predict counterfactual outcomes, and they become inadequate when dealing with data that contains intricate temporal dependencies. Subsequent research endeavors have aimed to overcome the limitations of model expressiveness. This has been accomplished through the adoption of Bayesian non-parametric methods (Xu et al., 2016; Soleimani et al., 2017; Schulam & Saria, 2017) or more sophisticated deep neural networks, such as recurrent neural networks (RNNs). Notably, in the realm of Bayesian non-parametric methods, recurrent marginal structural networks (RMSNs) (Lim, 2018) have emerged, replacing the linear model of MSM with an RNN-based architecture to forecast treatment outcomes. Similarly, G-Net (Li et al., 2021) integrates RNNs into the g-computation framework in place of classical regression models. Inspired by the successes of representation learning for domain adaptation and generalization (Ganin et al., 2016; Tzeng et al., 2015), recent research ventures have delved into the development of learned representations that serve both predictive purposes for outcome estimation and alleviating treatment bias within training data. For instance, the Counterfactual Recurrent Network (CRN) (Bica et al., 2020b) employs an RNN-based model trained with dual objectives: the factual outcome regression loss and the gradient reversal (Ganin et al., 2016) with respect to treatment prediction. The former fosters the development of informative outcome-predictive representations, while the latter encourages the creation of uniform representations across different treatments. This joint training target yields representations that are both informative and balanced. Motivated by similar objectives, a recent study by (Melnychuk et al., 2022) replaces the RNN-based architecture with a Transformer-based (Vaswani et al., 2017) one and employs a domain confusion loss (Tzeng et al., 2015) to facilitate the learning of treatment-agnostic representations.

D. Additional Experiment Results

D.1. Detailed Description of Baselines

Transformer-based:

- Causal Transformer (CT) (Melnychuk et al., 2022): A transformer-based counterfactual outcome prediction model, which is specifically designed to capture complex, long-range dependencies among time-varying confounders. The model proposed a domain confusion module (CDC) to obtain the adversarial balanced representations for addressing confounding bias.

LSTM-based:

- Counterfactual Recurrent Network(CRN) (Bica et al., 2020b): A sequence-to-sequence model that leverages the increasingly available patient observational data to estimate treatment effects over time and maximize the loss of treatment classifier for eliminating the confounding bias by adversarial gradient reversal.
- Recurrent Marginal Structural Networks(RMSN) (Lim, 2018): A LSTM-based model for time series data, which uses propensity weighting to adjust for time-dependent confounders.
- G-Net (Li et al., 2021): A sequential deep learning framework for counterfactual prediction under dynamic time-varying treatment strategies in complex longitudinal settings based on G-computation.

Linear-based:

- Marginal Structural Model (MSM) (Robins et al., 2000). A linear marginal structural model based on the inverse-probability-of-treatment weighted estimator.

Table 7. RMSE (mean $\pm$ std) with $\gamma = 2, 6, 8$ on Tumor dataset in standard supervised learning.

	Methods	τ=1	2	3	4	5	6
$\gamma = 2$	CT	0.8607±0.0636	0.7316±0.0381	0.7889±0.0362	0.8443±0.0497	0.8979±0.0605	0.9441±0.0765
	CT w/o BRM	0.8589±0.0605	0.7193±0.0399	0.7695±0.0543	0.8156±0.0608	0.8603±0.0709	0.9004±0.079
	CRN	0.8874±0.0643	0.7512±0.042	0.8365±0.0647	0.9305±0.087	1.0186±0.1041	1.0914±0.1144
	CRN w/o BRM	0.9081±0.0833	0.7253±0.0648	0.7647±0.0711	0.8093±0.0791	0.8549±0.0865	0.8978±0.09
	RMSN	1.1192±0.1139	0.9905±0.1068	1.0193±0.1189	1.0485±0.1388	1.0817±0.1596	1.1184±0.1782
	GNET	1.0014±0.0542	1.0301±0.07	1.1816±0.0999	1.2527±0.135	1.2944±0.1635	1.3194±0.1811
	MSM	1.4126±0.1091	2.0351±0.2576	2.387±0.3079	2.5739±0.3327	2.6381±0.3357	2.6038±0.3251
$\gamma = 6$	CT	2.9329±1.0845	3.6032±1.7021	4.1373±1.9675	4.4685±2.1355	4.6556±2.2183	4.6891±2.2337
	CT w/o BRM	2.8848±0.2065	2.8963±1.0772	3.1507±1.1061	3.2695±1.1335	3.2695±1.1335	3.3876±1.146
	CRN	2.5019±0.2458	3.0962±1.2731	3.5662±1.6316	3.763±1.7179	3.8416±1.7144	3.8249±1.66
	CRN w/o BRM	2.452±0.2341	3.1151±1.1059	3.5853±1.3143	3.7225±1.3755	3.6914±1.3697	3.5311±1.3101
	RMSN	2.4291±0.2032	1.9708±0.6216	2.3183±0.8332	2.5352±0.8919	2.6429±0.866	2.6645±0.7951
	GNET	2.1158±0.15	1.6777±0.3713	2.099±0.4588	2.3431±0.5212	2.4959±0.5712	2.5896±0.6363
	MSM	2.9828±0.2093	2.1097±0.67	1.8604±0.605	1.8862±0.5998	1.8448±0.5613	1.656±0.4577
$\gamma = 8$	CT	4.045±0.7598	4.9419±1.9872	5.5432±2.1325	5.9236±2.2364	6.1506±2.346	6.189±2.445
	CT w/o BRM	4.1487±1.0624	4.4726±1.9057	4.8615±1.8993	5.05±1.8558	5.1302±1.7835	5.0791±1.6694
	CRN	3.8468±0.2727	5.7004±2.4747	6.294±3.0082	6.4696±3.0877	6.4373±2.9626	6.2295±2.7222
	CRN w/o BRM	3.5992±0.2349	4.6192±1.3005	5.3959±1.5165	5.709±1.5527	5.7371±1.504	5.5302±1.3913
	RMSN	3.4263±0.4333	2.8244±1.2188	3.5567±1.4583	3.7758±1.4774	3.7013±1.5067	3.5593±1.3595
	GNET	3.1364±0.3073	2.856±0.8067	3.5623±1.0225	3.9387±1.143	4.1182±1.1899	4.1359±1.1703
	MSM	4.1831±0.371	1.5761±0.5207	2.1267±0.7158	2.5203±0.853	2.7747±0.9489	2.9511±1.0315

D.2. Factual Outcome Estimation on Real-world M5 dataset

The dataset for M5 Forecasting, as referenced in (Makridakis et al., 2022), encompasses daily transaction figures from Walmart outlets in three American states, supplemented by detailed information on products and stores, in addition to factors like pricing and notable occurrences. This dataset is restructured for the purpose of estimating treatment outcomes, identifying the pricing of products as the treatment factor and product sales figures as the result variable. All other attributes are considered as covariate variables.

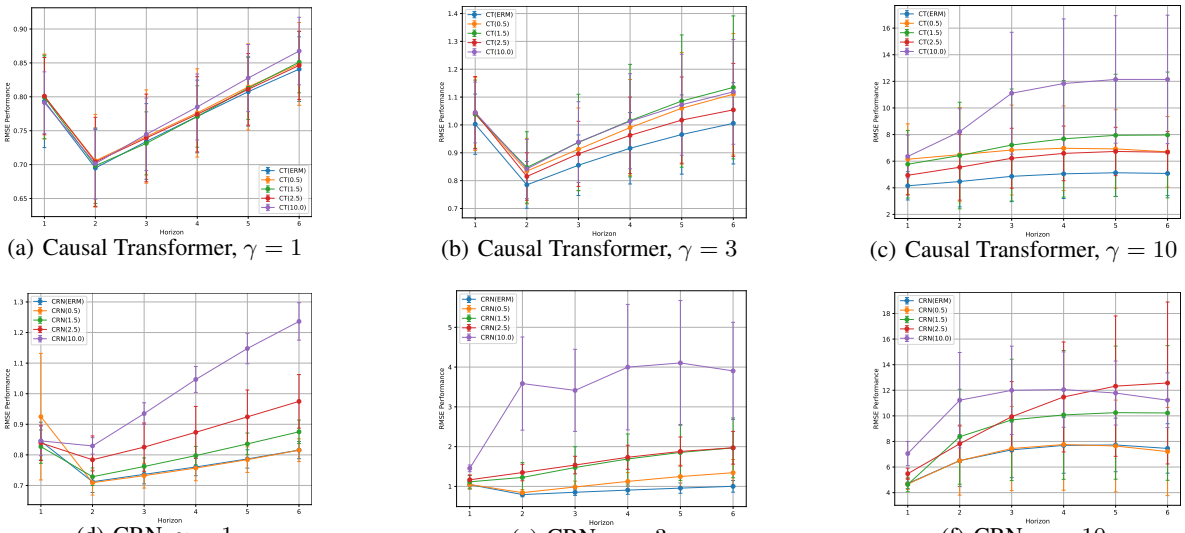
Given that this dataset is derived from real-world data and lacks information on counterfactual outcomes, we demonstrate the performance of various models and their Empirical Risk Minimization (ERM) variants in estimating factual outcomes. We omit GENT and MSM here because they can not converge on this dataset. Similarly, we set the horizon τ from 1 to 6, and the results, as shown in Table 8, indicate that ERM variants without a balancing module generally outperform those with a balancing module. This outcome is intuitive since we are dealing with factual outcomes, and the presence of a balancing module alters the original distribution of covariates, causing a loss of information beneficial to outcome prediction.

D.3. Multi-step Counterfactual Outcome in Standard Supervised Learning

reports on the performance of various models and their ERM variants in counterfactual estimation on the Tumor dataset, with the treatment bias strength parameter γ, set as 2, 6, and 8. The results, illustrated in Table 7, corroborate the phenomena discussed in the main text. It was found that ERM variants without a balancing module generally exhibit superior predictive performance compared to those with a balancing module. Moreover, as the intensity of treatment bias increases, the negative impact on model performance brought about by the balancing strategy becomes more pronounced (as evidenced by the increase in the Root Mean Square Error (RMSE) average) and more unstable (as can be discerned from the standard deviation of the RMSE). This highlights the challenges in managing treatment bias in models, especially when balancing mechanisms are employed, suggesting a careful consideration of the trade-off between bias correction and predictive accuracy.

Table 8. RMNSE (mean ± std) for M5 real-world dataset in factual outcome estiamtion.

Methods	τ=1	2	3	4	5	6
CT	4.5934±0.0823	8.9863±0.1888	9.5868±0.1934	9.9135±0.2266	10.1454±0.2535	10.3476±0.2838
CT w/o BRM	4.5908±0.0855	8.9802±0.1814	9.5806±0.1789	9.9029±0.2047	10.1342±0.2273	10.3354±0.2551
CRN	4.9843±0.316	9.1488±0.1674	9.7878±0.1619	10.1161±0.1683	10.3767±0.1921	10.5927±0.2228
CRN w/o BRM	4.6808±0.0554	9.0483±0.1636	9.6803±0.1535	10.004±0.1625	10.2566±0.1779	10.4691±0.1971
RMSN	5.1912±0.1169	9.7308±0.2557	10.4807±0.3374	11.069±0.4258	11.632±0.532	12.1559±0.6548



(a) Causal Transformer, $\gamma = 1$ (b) Causal Transformer, $\gamma = 3$ (c) Causal Transformer, $\gamma = 10$

(d) CRN, $\gamma = 1$ (e) CRN, $\gamma = 3$ (f) CRN, $\gamma = 10$

Figure 8. Model performance comparison in different weights of balancing module, where CT with Counterfactual Domain Confusion, CRN with Adversarial Gradient Reversal.

D.4. Additional Short-term History Cold-start Case for Counterfactual Estimation

In this report, we detail additional findings on the performance of counterfactual outcome estimation in the context of short-term history cold-start cases on the Tumor dataset, under varying intensities of treatment bias. Specifically, we compared the performance of different counterfactual estimation models and their Empirical Risk Minimization (ERM) variants in multi-step prediction scenarios, with γ values set at 2, 6, and 8, as depicted in Table 9. Our observations reveal that while models with a balancing strategy may exhibit lower average Root Mean Square Error (RMSE) values in certain cases compared to those without a balancing strategy, the significant difference in the standard deviation between the two groups undermines the potential advantages of balanced models in these instances. Furthermore, it is noteworthy that in the majority of cases, models lacking a balancing strategy tend to perform better. This suggests that the purported benefits of implementing a balancing mechanism for counterfactual outcome prediction in cold-start scenarios with short-term history may not be consistently demonstrable across different levels of treatment bias, especially when considering the variability in model performance.

D.5. Additional Distribution Shift Cold-start Case for Counterfactual Estimation

In this report, we present additional experimental results on counterfactual outcome estimation under the scenario of distribution shift cold-start Case, involving various models and their ERM variants. These models were trained with a treatment bias strength of $\gamma = 10$, and then tested on datasets constructed with the treatment bias strength of $\gamma = 2, 6$ to evaluate the effectiveness of balancing strategies when there are distributional differences between the source and target domains. The findings, as displayed in Table 10, consistently show that ERM variants without a balancing module generally outperform those with a balancing module. We did not observe any benefit from the balancing strategy in the context of distribution shift cases for counterfactual estimation tasks; instead, it appears to have a detrimental effect, introducing

Table 9. RMSE (mean ± std) performance on Tumor for short-term history cold-start case.

	Methods	τ=1	2	3	4	5	6
$\gamma = 2$	CT	0.9233±0.0605	0.7834±0.0572	0.8694±0.0643	0.9514±0.0669	1.0158±0.0741	1.0733±0.0801
	CT w/o BRM	0.9215±0.0527	0.7955±0.0575	0.8797±0.0698	0.9564±0.077	1.0177±0.0863	1.0709±0.0884
	CRN	0.9411±0.045	0.848±0.0838	0.9751±0.0981	1.1063±0.1163	1.2235±0.1292	1.3169±0.1323
	CRN w/o BRM	0.9632±0.0622	0.8061±0.0956	0.8835±0.1173	0.9631±0.1374	1.0368±0.1517	1.1018±0.1573
	RMSN	1.1651±0.0931	1.0225±0.1023	1.0943±0.1043	1.1698±0.1118	1.2383±0.1237	1.3018±0.1319
	GNET	1.0559±0.058	5.0544±0.1537	6.1204±0.2822	6.4523±0.4056	6.4235±0.5026	6.174±0.5921
	MSM	1.4994±0.0684	2.4804±0.1105	2.9169±0.1328	3.1447±0.1411	3.2205±0.142	3.1734±0.1437
$\gamma = 6$	CT	3.1077±1.0199	4.3539±3.144	5.0768±3.6509	5.4948±3.9159	5.7599±3.951	5.7262±3.8615
	CT w/o BRM	2.8129±0.3076	3.0805±1.2024	3.4053±1.2094	3.6102±0.9977	3.8604±0.8948	3.991±0.8221
	CRN	2.7107±0.295	4.5699±1.1237	5.1567±1.2849	5.356±1.289	5.3911±1.2332	5.2859±1.1368
	CRN w/o BRM	2.6608±0.2909	4.7513±0.871	5.4321±0.8854	5.5918±0.8353	5.4986±0.7715	5.2123±0.6938
	RMSN	2.6146±0.174	2.7629±0.6637	3.4115±0.7996	3.7641±0.8695	3.9148±0.8473	3.9114±0.7832
	GNET	2.2245±0.2132	4.6513±0.4998	5.4084±0.6527	5.6466±0.7452	5.6574±0.7861	5.5463±0.7823
	MSM	3.2872±0.2421	3.0459±0.4111	2.7305±0.3868	2.7901±0.4116	2.7366±0.4222	2.4671±0.4189
$\gamma = 8$	CT	4.1338±0.5542	5.7549±2.494	6.7353±2.8143	7.2575±2.9254	7.4752±2.8458	7.6408±2.7925
	CT w/o BRM	4.0003±0.6324	4.0751±1.4148	4.568±1.5578	4.8±1.5702	4.8626±1.5929	5.1146±1.4664
	CRN	4.0312±0.2933	8.3633±3.9551	9.1368±4.2186	9.3202±4.063	9.23±3.8222	8.8752±3.4921
	CRN w/o BRM	3.7808±0.2239	6.6869±1.2202	7.7847±1.3929	8.2069±1.4314	8.217±1.4154	7.8825±1.3351
	RMSN	3.6366±0.4517	3.8651±1.5634	5.0449±1.6389	5.3907±1.5291	5.248±1.5433	5.0034±1.3334
	GNET	3.1837±0.2193	5.6514±0.9173	6.5778±1.0536	6.9325±1.0519	6.987±0.9693	6.8128±0.8408
	MSM	4.3752±0.4191	2.2088±0.3828	2.9991±0.5359	3.5619±0.6461	3.9347±0.7295	4.2067±0.8001

instability into the model's performance. This suggests that in scenarios where the training and testing data distributions differ significantly, employing a balancing strategy may not enhance, and could potentially impair, the stability and accuracy of counterfactual outcome predictions.

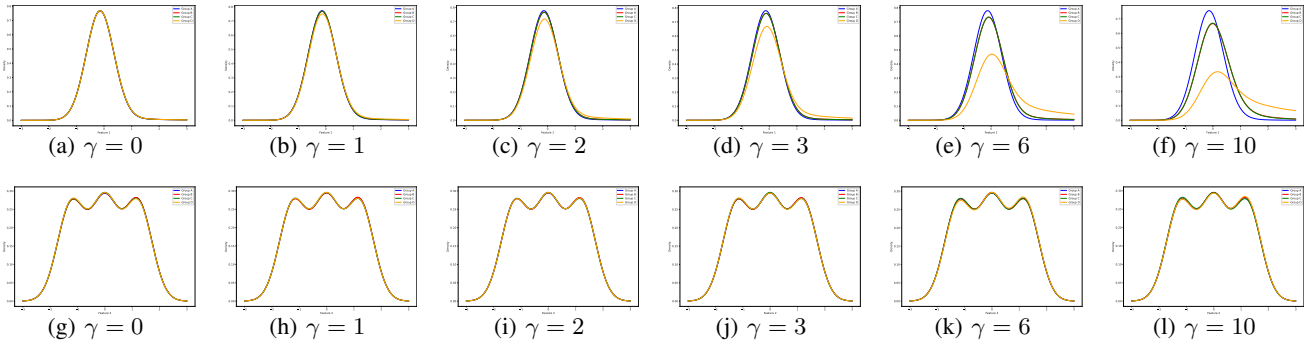
D.6. Exploring the Impact of Contribution Weight for Balancing Module

In this section, we conduct the experiments to explore the impact of the contribution weights of the two balanced representation methods: Adversarial Gradient Reversal, Counterfactual Domain Confusion on model performance in temporal counterfactual outcome estimation tasks for the synthetic Tumor Growth dataset. Due to the poor performance of PS-based Contrastive Balancing, we omit the exploration for it. These methods were initially mentioned in Section 3. In our analysis, we denote the contribution weights of these balancing modules as α, with values set within the range of [0.5, 1.5, 2.5, 10.0]. We then compare their results with the ERM version (where $\alpha = 0$) of each model, while keeping other hyperparameters consistent with those used in standard supervised learning settings. Experimental results, as illustrated in Figure 8, demonstrate the performance of different balanced representation module weights under varying values of $\gamma = [1, 3, 10]$.

Our observations indicate that when γ values are relatively low (e.g., 1 or 3), signifying less significant treatment bias, the weight values of the balanced representation modules have a minimal impact on the performance of counterfactual outcome estimation, the model performance in terms of different balancing weights is almost the same. However, as the γ value increases, a general trend emerges: with the rise in the balancing module's weight, there is a corresponding decline in performance. Notably, compared to the ERM models, the performance of models with higher balancing module weights tends to be less effective, which is consistent with the previous observation in the standard supervised learning setting. This finding underscores the negative role that balancing module weights play in temporal counterfactual outcome estimation, particularly in scenarios characterized by more pronounced treatment biases.

Table 10. RMSE (mean ± std) performance on Tumor for distribution shift cold-start case.

	Methods	τ=1	2	3	4	5	6
$\gamma = 2$	CT	1.7095±1.0786	2.4904±2.5066	3.0249±3.2811	3.4384±3.7743	3.5949±3.8012	3.702±3.8142
	CT w/o BRM	1.0343±0.1086	1.0491±0.102	1.1595±0.1364	1.2747±0.1549	1.3739±0.1682	1.4078±0.1844
	CRN	1.626±0.0804	2.0827±0.2604	2.1672±0.4988	2.2263±0.5693	2.2678±0.563	2.284±0.5354
	CRN w/o BRM	1.4259±0.0824	1.4815±0.1531	1.6756±0.1583	1.7945±0.1491	1.8476±0.1399	1.834±0.1368
	RMSN	1.4809±0.1192	1.3126±0.1	1.3464±0.1102	1.3807±0.1043	1.4222±0.1102	1.4675±0.1519
	GNET	1.1463±0.0576	4.022±0.0451	4.5349±0.1598	4.6286±0.2901	4.6283±0.4205	4.6054±0.5494
	MSM	1.3067±0.1205	0.5554±0.0677	0.8374±0.1064	1.0729±0.1382	1.2844±0.1669	1.4894±0.1901
$\gamma = 6$	CT	3.3542±1.0787	4.0092±2.9922	4.7749±3.7048	5.2517±4.0412	5.4713±4.013	5.5775±3.9294
	CT w/o BRM	2.2628±0.1719	2.0385±0.6964	2.4144±0.8376	2.5979±0.8042	2.7848±0.8568	2.8334±0.8094
	CRN	2.9282±0.0577	3.831±1.3974	4.1685±1.8389	4.3323±1.9419	4.4152±1.8572	4.4311±1.6918
	CRN w/o BRM	2.7046±0.0917	3.5286±1.2803	4.0347±1.3757	4.2852±1.3728	4.3668±1.3267	4.2864±1.2416
	RMSN	2.8807±0.1817	2.8372±1.121	2.9339±1.1156	2.9181±1.0219	2.893±0.8884	2.8541±0.7413
	GNET	2.1707±0.1345	4.068±0.4381	4.5058±0.6647	4.6623±0.8378	4.723±0.9577	4.7171±1.0359
	MSM	3.1887±0.2687	1.0831±0.3845	1.6306±0.5812	2.0902±0.7477	2.5028±0.9024	2.9196±1.0679



(a) $\gamma = 0$ (b) $\gamma = 1$ (c) $\gamma = 2$ (d) $\gamma = 3$ (e) $\gamma = 6$ (f) $\gamma = 10$

(g) $\gamma = 0$ (h) $\gamma = 1$ (i) $\gamma = 2$ (j) $\gamma = 3$ (k) $\gamma = 6$ (l) $\gamma = 10$

Figure 9. Distribution shape for the two covariates under different strengths of treatment bias. (a)-(f) for the first covariate, (g)-(l) for the second covariate.

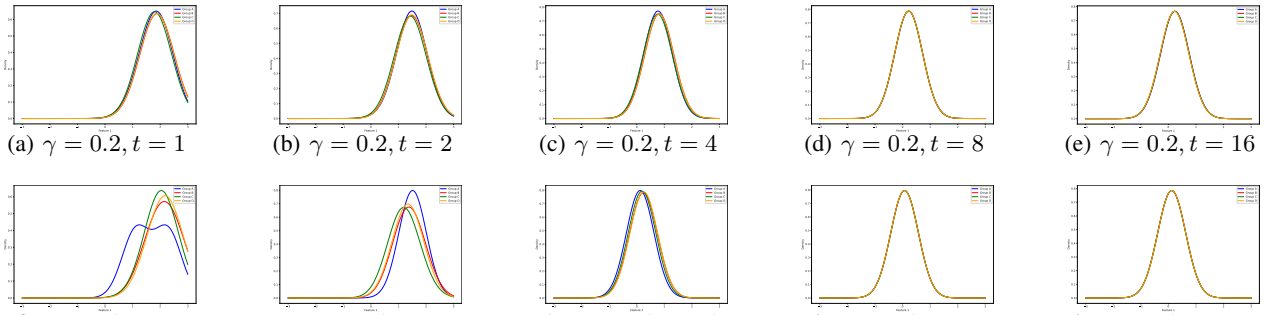
D.7. Detailed Analysis for the Covariate Distribution

Tumor Growth. Here we present the covariate distribution averaged on the all-time steps for treatment bias strength $gamma = \{0, 1, 2, 3, 6, 10\}$. The Gaussian fit distributions of the first covariate and the second covariate are shown in Figures 9. We can observe that for the first covariate, when the treatment bias intensity is relatively low (e.g., γ equals 0, 1, 2, 3), the differences in the distribution of this variable among different groups are not very pronounced. It's only when the treatment bias is relatively high that noticeable differences in distribution between groups emerge. As for the second covariate, regardless of the intensity of the treatment bias, the distribution of this covariate among different groups remains nearly identical.

Therefore, based on the results presented above, this may be the reason why the balanced representation module of the causal time-series model doesn't work: in this dataset, the distribution of one covariate is the same across different groups, and the distribution of the other covariate only exhibits noticeable differences among groups when the treatment bias is relatively high. As mentioned at the outset, the Balanced Representation Module aims to reduce the distance between covariate distributions among different groups. However, if the disparities in covariate distributions among different groups in the data are not sufficiently large, this module is likely to become a burden to the model, leading to a decrease in predictive performance.

Pure Synthetic Data. Furthermore, to avoid losing generality, we present the visualization of covariate distribution for more

time-step selection. We set the sampled time step to {1,2,4,8,16}. We plot the visualization of one certain covariate over the sampled time steps as shown in Figure 10. The results reveal that at the onset of the timeline (e.g., at time-steps 1 and 2), an increase in the gamma value correlates with a more pronounced disparity in covariate distributions among different treatment groups. Yet, as the timeline progresses to later stages (e.g., at time step 8 and 16), the variance in covariate distributions between treatment groups diminishes. This pattern suggests that, in line with typical time series assumptions, treatment bias tends to decrease over time, thereby reducing the efficacy of strategies aimed at balancing representation distributions.



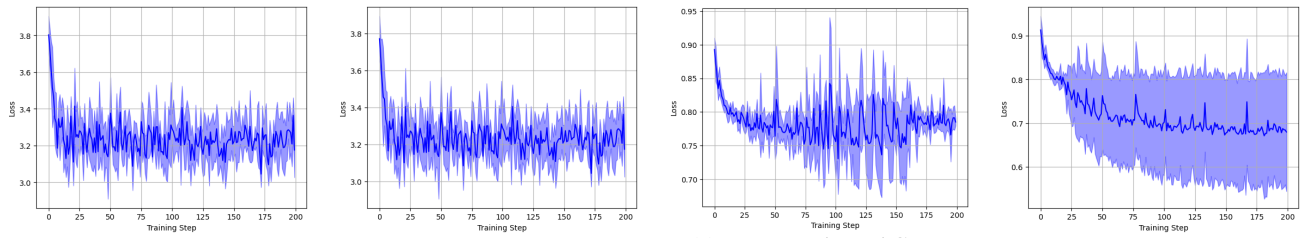
(a) $\gamma = 0.2, t = 1$ (b) $\gamma = 0.2, t = 2$ (c) $\gamma = 0.2, t = 4$ (d) $\gamma = 0.2, t = 8$ (e) $\gamma = 0.2, t = 16$

(f) $\gamma = 0.4, t = 1$ (g) $\gamma = 0.4, t = 2$ (h) $\gamma = 0.4, t = 4$ (i) $\gamma = 0.4, t = 8$ (j) $\gamma = 0.4, t = 16$

Figure 10. Distribution shape for the covariate under different strengths of treatment bias and time-steps on pure synthetic dataset.

D.8. Learned Representations for Reconstruction of Covariates

In the main text, we present the reconstruction results of the original covariates by the Causal Transformer under scenarios of both mild and significant treatment bias, indicated by $\gamma = 1$ and 10, respectively. This is done to validate the impact of the balanced representation strategy on the loss of covariate information when employing balancing strategies versus not. Here, we also display the capability of the CRN model to reconstruct the original covariates under the same settings, using two types of representations as illustrated in Figure 11. We observe that under conditions of mild treatment bias ($\gamma = 1$), the training for the task of reconstructing original covariates with non-balanced representation is more stable than with balanced representation. The loss for the former reasonably and steadily declines, whereas the latter's loss fluctuates continuously. However, under conditions of significant treatment bias ($\gamma = 10$), both balanced and non-balanced representations show poor convergence in training. This suggests that excessive treatment bias can distort the training process of capturing useful information from the original covariates, leading to unstable prediction outcomes.

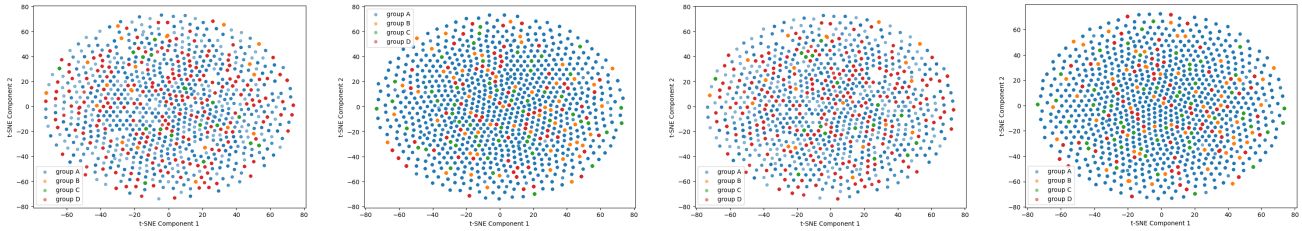


(a) $\gamma = 10$, Balanced CRN (b) $\gamma = 10$, Non-balanced CRN (c) $\gamma = 1$, Balanced CRN (d) $\gamma = 1$, Non-balanced CRN

Figure 11. Reconstruction loss curves for CRN model.

D.9. Representation Visualization for CRN model

Here, we report on the visualization of representations derived from the CRN model, both with and without the implementation of a balancing strategy. Similarly, we set the intensity of treatment bias, γ, to 10 and obtained representations for time steps $t = 1$ and $t = 10$. The visualization results, as shown in Figure 12, indicate that the distribution differences between the balanced and non-balanced representations, after dimensionality reduction, are not significant. Under conditions of pronounced treatment bias, the balancing module does not appear to be effective. This observation corroborates our previous discussion and analysis that excessive treatment bias can disrupt the model's learning process. The visualizations provide a



(a) $t = 1$, CRN, Balanced (b) $t = 10, CRN, Balanced$ (c) $t = 1$, CRN, Non-balanced (d) $t = 10$, CRN, Non-balanced

Figure 12. Visualization for balanced and non-balanced representation for different time steps for CRN models.

tangible insight into how, in the presence of significant treatment bias, both balanced and non-balanced strategies result in similar distributions of representations. This suggests a limitation in the balancing module's ability to mitigate the effects of treatment bias when it is substantial, affirming the notion that an overly pronounced treatment bias poses a challenge to the model's ability to learn and differentiate effectively between treated and untreated groups based on the original covariates.