Synergizing Deconfounding and Temporal Generalization For Time-series Counterfactual Outcome Estimation

Yiling Liu *yiling.liu@duke.edu*
Program in Computational Biology and Bioinformatics, Duke University

Juncheng Dong *juncheng.dong@duke.edu*
Department of Electrical and Computer Engineering, Duke University

Chen Fu *chenfu@meta.com*
AI at Meta, Meta Platforms, Inc.

Wei Shi *weishi0079@meta.com*
AI at Meta, Meta Platforms, Inc.

Ziyang Jiang *jzy95310@meta.com*
AI at Meta, Meta Platforms, Inc.

Qi Xu *xuqi0511@gmail.com*
AI at Meta, Meta Platforms, Inc.

Zhigang Hua *zhua@meta.com*
AI at Meta, Meta Platforms, Inc.

David Carlson[†] *david.carlson@duke.edu*
Department of Civil and Environmental Engineering
Department of Biostatistics and Bioinformatics
Department of Computer Science
Duke University

Reviewed on OpenReview: *https://openreview.net/forum?id=xuJH3BJiNu*

Abstract

Estimating *counterfactual* outcomes from time-series observations is crucial for effective decision-making, e.g. when to administer a life-saving treatment, yet remains significantly challenging because (*i*) the counterfactual trajectory is never observed and (*ii*) confounders evolve with time and distort estimation at every step. To address these challenges, we propose a novel framework that ***synergistically integrates*** two complementary approaches: ***Sub-treatment Group Alignment*** (SGA) and ***Random Temporal Masking*** (RTM). Instead of the coarse practice of aligning marginal distributions of the treatments in latent space, SGA uses iterative *treatment-agnostic* clustering to identify fine-grained *sub-treatment groups*. Aligning these fine-grained groups achieves improved distributional matching, thus leading to more effective deconfounding. We theoretically demonstrate that SGA optimizes a tighter upper bound on counterfactual risk up to an additive constant term, and empirically verify its deconfounding efficacy. RTM promotes temporal generalization by randomly replaces input covariates with Gaussian noises during training. This encourages the model to rely less on potentially noisy or spuriously correlated covariates at the current step and more on stable historical patterns, thereby improving its ability to generalize across time and better preserve underlying *causal relationships*. Our experiments demonstrate that while applying SGA and RTM individually improves counterfactual outcome estimation, their

[†]Corresponding author.

synergistic combination consistently achieves state-of-the-art performance. This success comes from their distinct yet complementary roles: RTM enhances temporal generalization and robustness across time steps, while SGA improves deconfounding at each specific time point.

1 Introduction

Estimating causal effects from time series data is a pivotal task in many fields such as healthcare, politics, and economics (Bisgaard & Kulahci, 2011; Freeman, 1983; Morid et al., 2023). This fundamentally requires the ability to estimate counterfactual outcomes: what would have happened under different interventions over time. For example, in the treatment of *Ductal Carcinoma In Situ*, accurately estimating counterfactual outcomes is crucial for determining the timing of surgical intervention: if surgery is too late, the cancer may progress to an invasive stage; if conducted too early, the procedure may be unnecessarily invasive (Grimm et al., 2022).

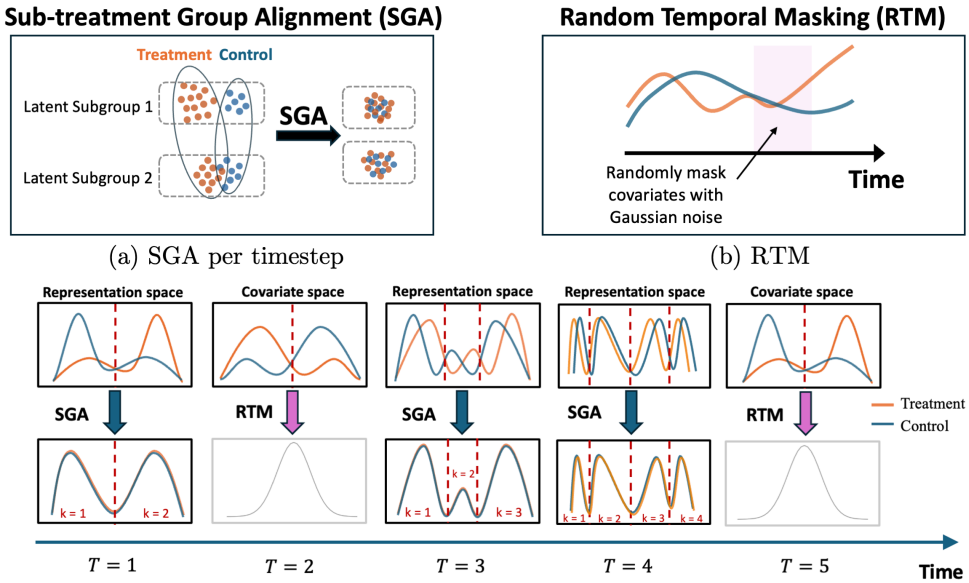
Motivated by this, we explore counterfactual outcome estimation in time series from observational data. The success of causal inference in time series relies on **effective reduction of time-dependent confounding**. However, this task is challenging, primarily because of the unobservability of counterfactual outcomes and the time-varying confounding in time series. A well-established approach for reducing confounding in *static* causal inference is to minimize an upper bound on the counterfactual estimation error (Johansson et al., 2016; 2022; Li & Fu, 2017; Yao et al., 2018), which can be decomposed into two key components: (i) the *factual loss* and (ii) the *statistical discrepancy between treatment and control groups* in the learned representation space. Algorithmically, these methods simultaneously (i) minimize the prediction error of the factual outcomes and (ii) align the treatment groups in the latent space. By ensuring that the representations of multiple treatment groups are brought closer together, they provably reduce the bias introduced by confounders (Johansson et al., 2022). Building on this idea to reduce confounding for time series, existing approaches aim to learn representations that remain invariant to the treatment assignment **at each time step** (Bica et al., 2020; Melnychuk et al., 2022). However, in practice, they typically result in optimizing relatively **loose** upper bounds on the counterfactual error at individual time steps (Arjovsky & Bottou, 2017) with adversarial training. Moreover, models may **over-rely on contemporaneous information** at each time step, potentially hindering their ability to learn stable long-term patterns across time points (Ghouse et al., 2024). This can lead to **compromised generalization** for estimating long-term effects, as errors can propagate and compound over time.

To address these challenges, we introduce our framework with two novel approaches:

- ❖ ***Sub-treatment Group Alignment (SGA)***, which improves deconfounding at each individual time point by identifying and subsequently aligning *sub-treatment groups*.
- ✯ ***Random Temporal Masking (RTM)*** promotes temporal generalization and robust learning from historical patterns by randomly masking covariates at selected time points with Gaussian noise.

Sub-treatment Group Alignment (SGA). SGA first identifies **sub-treatment groups in the representation space** through *treatment-agnostic* clustering at each timestep, and then subsequently aligning the corresponding sub-groups across different treatment groups (see Figure 1). In Section 4, we establish that such sub-group alignment indeed leads to a tighter bound on the counterfactual estimation error up to an additive constant term. This more fine-grained alignment enables **improved matching** of treatment groups, thus allowing us to **reduce the estimation error** more effectively than existing methods.

Random Temporal Masking (RTM). While SGA addresses confounding at **individual** time points, RTM enhances the model's ability to generalize **across** time series. Inspired by masked language modeling, RTM uses random covariate masking, where input covariates at randomly selected time points are replaced with Gaussian noise during training. There are multiple perspectives to understand the benefits of RTM: (i) When input covariates at certain time points are masked, the models must extract useful information from past observations to predict future factual outcomes. In other words, we encourage the model to **focus on the causal relationships that span across time**, leading to better counterfactual predictions. (ii) RTM



(a) SGA per timestep (b) RTM

(c) Synergistically combine SGA and RTM over time

Figure 1: **Conceptual overview of SGA and RTM. (a)** SGA identifies and aligns fine-grained sub-treatment groups at each timestep to improve deconfounding. **(b)** RTM forces the model to leverage historical patterns and enhancing temporal generalization. **(c)** SGA and RTM are synergistically combined to improve counterfactual outcome estimation. Here, k denotes sub-treatment group index.

can prevent model from becoming overly reliant on information from current time points, thus **reducing overfitting to the factual distribution**.

Empirical Validation. Our framework is **broadly applicable** and designed for **straightforward integration** into existing representation learning-based frameworks for time series counterfactual estimation. We empirically validate our framework through comprehensive experiments on synthetic and semi-synthetic datasets, and it consistently demonstrates state-of-the-art (SOTA) performance.

Organization. We first formally define the problem in Section 2 and review related works in Section 3. Then in Section 4, we theoretically establish how sub-treatment group alignment achieves improved deconfounding, thus motivating our SGA approach. In Section 5, we present our framework with SGA and RTM. Finally, experimental results in Section 6 show that applying SGA and RTM individually enhances performance, and that their ***synergistic combination*** achieves SOTA results.

2 Problem Setup

Notations. We use upper-case letters (e.g., A, Y) for scalar random variables and lower-case letters (e.g., a, y) for their corresponding realizations. Multi-dimensional random variables and realizations are denoted using bold fonts (e.g., $\boldsymbol{X}$ and $\boldsymbol{x}$).

Observational Data. We consider a dataset containing N samples following conventional setups (Bica et al., 2020; Li et al., 2020; Melnychuk et al., 2022). Observations are recorded over T time steps, i.e., $t = 1, ..., T$. At each time t, a discrete treatment $A_t \in \mathcal{A} = \{a_0, a_1, ..., a_{|\mathcal{A}|-1}\}$ is assigned. Thus, for each sample $i \in \{1, 2, ..., N\}$, we observe time-varying covariates $\mathbf{X}_t^{(i)} \in \mathbb{R}^d$, factual treatment $A_t^{(i)}$, and outcome $Y_t^{(i)}$ of the factual treatment.

We use the following notation to represent the history up to time step t for each unit i:

$$\bar{\mathbf{H}}_t^{(i)} = \{\bar{\mathbf{X}}_t^{(i)}, \bar{\mathbf{Y}}_t^{(i)}, \bar{\mathbf{A}}_{t-1}^{(i)}, \mathbf{V}^{(i)}\}$$

where $\bar{\mathbf{X}}_t^{(i)} = \{\mathbf{X}_s^{(i)} : s \leq t\}$ denotes the sequence of time-varying covariates up to time t, $\bar{\mathbf{Y}}_t^{(i)} = \{Y_s^{(i)} : s \leq t\}$ represents the sequence of observed outcomes up to time t, $\bar{\mathbf{A}}_{t-1}^{(i)} = \{A_s^{(i)} : s \leq t-1\}$ is the sequence of treatments up to time $t-1$, and $\mathbf{V}^{(i)} \in \mathbb{R}^p$ denotes the static covariates.

Potential Outcomes Framework with Time-Varying Treatments and Outcomes. Building on the potential outcomes framework (Rosenbaum & Rubin, 1983; Rubin, 2005), we extend these assumptions to accommodate time-varying treatments and outcomes, following Robins & Hernan (2008).

Assumption 2.1 (Consistency)**.** If $\bar{\mathbf{A}}_t = \bar{\mathbf{a}}_t$ is a given sequence of treatments for some patient, then $\mathbf{Y}_{t+1}[\bar{\mathbf{a}}_t] = \mathbf{Y}_{t+1}$. This means an individual's potential outcome under the observed treatment history is the observed outcome.

Assumption 2.2 (Sequential Positivity)**.** Positivity states that there is a non-zero probability of receiving any of the counterfactual treatments. It can be expressed as $0 < P(\mathbf{A}_t = \mathbf{a}_t | \bar{\mathbf{H}}_t = \bar{\mathbf{h}}_t) < 1$, if $P(\bar{\mathbf{H}}_t = \bar{\mathbf{h}}_t) > 0$.

Assumption 2.3 (Sequential Ignorability)**.** There is no unobserved confounding of treatment at any time and any future outcome. This can be expressed as $\mathbf{A}_t \perp\!\!\!\perp \mathbf{Y}_{t+1}[\mathbf{a}_t] | \bar{\mathbf{H}}_t, \forall\, \mathbf{a}_t \in \mathcal{A}$.

Using assumptions 2.1–2.3, Robins (1986) establishes the sufficient conditions for identifiability through G-computation, ensuring that causal effects can be appropriately identified.

Causal Directed Acyclic Graphs (DAGs). Figure 2 provides a visual summary of the assumed data-generating mechanisms in both static and time-series settings. These diagrams illustrate the causal relationships among covariates, treatments, and outcomes, and highlight the challenges introduced by time-varying confounding.

Objective. Given the history up to current time t and assuming a specific treatment sequence $\mathbf{a}_{t:t+\tau-1}^{(i)}$ from time t to $t+\tau-1$ applied to sample i, **our goal is to estimate**, for each unit i, **the future outcome at time step** $t+\tau$. To ensure that these counterfactual outcomes are identifiable, we rely on the potential outcomes framework outlined above. Specifically, we estimate:

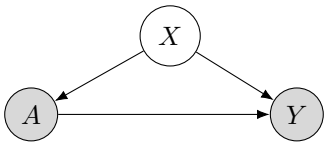
$$\mathbb{E}[Y_{t+\tau}^{(i)}(\mathbf{a}_{t:t+\tau-1}^{(i)}) \,|\, \bar{\mathbf{H}}_t^{(i)}], \tag{1}$$

where $Y_{t+\tau}^{(i)}(\mathbf{a}_{t:t+\tau-1}^{(i)})$ is the potential outcome at $t+\tau$ for unit i with treatment sequence $\mathbf{a}_{t:t+\tau-1}^{(i)}$.

3 Related Work

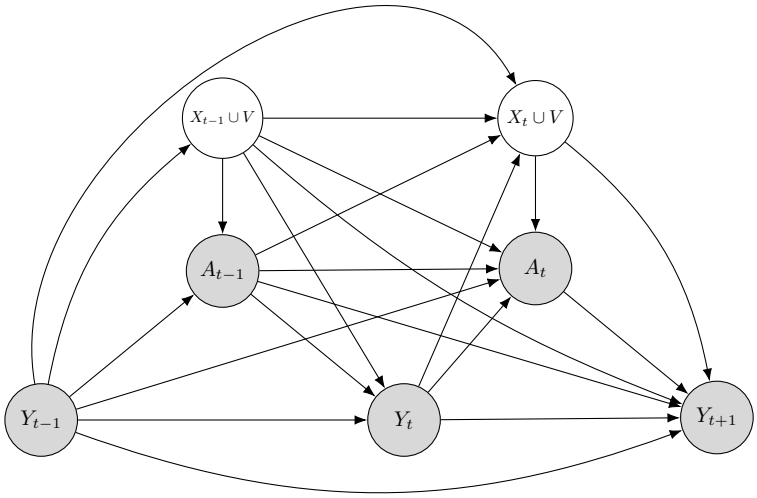
We review the **most** relevant work below and provide a **comprehensive** discussion in Appendix B.

Estimating Counterfactual Outcomes Over Time. Estimating counterfactual outcomes in time-series is challenging due to time-varying confounders. Traditional methods such as G-computation and marginal structural models (Hernán et al., 2001; Robins, 1986; Robins & Hernan, 2008; Robins et al., 2000; Xu et al., 2016) often lack flexibility for complex datasets and rely on strong assumptions. To address these limitations, researchers have developed models that build on the potential outcomes framework initially proposed by Rubin (1978) and extended to time series by Robins & Hernan (2008). Notable among recent methods are Recurrent Marginal Structural Networks (RMSNs) (Lim, 2018), G-Net (Li et al., 2020), Counterfactual Recurrent Networks (CRN) (Bica et al., 2020), and the Causal Transformer (CT) (Melnychuk et al., 2022), which use approaches such as propensity networks and adversarial learning to mitigate the effects of time-varying confounding. However, practical challenges with adversarial training can affect the stability of causal effect estimations. Specifically, training adversarial networks can be challenging due to issues such as mode collapse and oscillations (Liang et al., 2018). Additionally, adversarial training minimizes the Jensen-Shannon divergence (JSD) only when the discriminator is optimal (Arjovsky & Bottou, 2017), which may not always be achievable in practice; even when the discrminator is optimal, using JSD optimizing relatively loose upper bounds on the counterfactual error. To address these challenges, we propose using the Wasserstein-1 distance and provides stronger theoretical guarantees (Mansour et al., 2012; Redko et al., 2017).



(a) Static (Non-Time-Series) Scenario

In the static setting, A denotes treatment assignment, X the observed covariates, and Y the outcome. The covariates act as confounders affecting both treatment and outcome.



(b) Time-Series Scenario

In the time-series setting, A_t denotes treatment at time t, $X_t \cup V$ represents observed time-varying and static covariates, and Y_t denotes the outcome at time t. Past treatments and outcomes influence future covariates, treatments, and outcomes, inducing time-varying confounding.

Figure 2: **Causal Directed Acyclic Graphs (DAGs) Illustrating Causal Relationships.** (a) Static setting. (b) Time-series setting with dynamic treatment effects and time-varying confounding.

Masked Language Modeling. Masked language modeling (MLM) is a common self-supervised pre-training technique for large language models. It operates by randomly masking certain words or tokens in the input, with the model trained to predict the masked tokens. BERT (Devlin, 2018) is the most well-known model that uses this technique. Recent studies have also demonstrated the effectiveness of MLM in enhancing generalization across sequence-based tasks. For example, Chaudhary et al. (2020) shows that when combined with cross-lingual dictionaries, MLM improves predictions for the original masked word and also generalizes to its cross-lingual synonyms. Inspired by the success of masking strategies in language models, we introduce Random Temporal Masking (RTM) for time-series data. Unlike MLM, which focuses on predicting the masked inputs, RTM encourages the model to focus on information that is crucial for both the current time point and future time points, preserve causal information, and reduce the risk of overfitting to factual outcomes.

4 Theoretical Motivation for Sub-treatment Group Alignment

In this section, we provide a theoretical motivation for our proposed Sub-treatment Group Alignment (SGA) method, rigorously illustrating that **aligning sub-treatment groups in the latent space leads to more effective deconfounding in estimating counterfactual outcomes** over time series.

From Static to Time Series. By aligning the corresponding sub-treatment groups, **SGA achieves improved alignment** and thus more effective deconfounding. Since existing time series methods often apply alignment independently at each time step (Bica et al., 2020; Melnychuk et al., 2022), demonstrating SGA's superiority in a static context provides a strong foundation for its benefits in time-series. In other words, achieving better deconfounding at *individual time steps* will consequently lead to more effective deconfounding *over the entire time series*.

Section Organization. Given the rationale that time-series alignment can be decomposed into a collection of static sub-problems, **it is sufficient to consider static settings**. Thus, in Section 4.1 we briefly review (i) representation learning-based models that use alignment (ii) why alignment mitigates confounder bias in the static setting. Subsequently, in Section 4.2, we theoretically establish that SGA indeed improves alignment in static settings. This implies that **integrating SGA into existing time-series frameworks can improve deconfounding at each step**, leading to overall improvements.

4.1 Alignment for Static Setting

Since there is only one time step $t = 1$ in the static setting, we will omit all notations about the time step for clarity. To establish the theoretical bounds, we first define the necessary mathematical components, including the distance metric and the objective functions. Due to space constraints, we defer the discussion of SGA's computational cost to Appendix F.

Wasserstein Distance. The Kantorovich-Rubenstein dual representation of the Wasserstein-1 distance (Villani, 2009) between two distributions p^0_Φ and p^1_Φ is defined as:

$$W_1(p^0_\Phi, p^1_\Phi) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim p^0_\Phi}[f(x)] - \mathbb{E}_{x\sim p^1_\Phi}[f(x)],$$

where the supremum is over the set of 1-Lipschitz functions (all Lipschitz functions f with Lipschitz constant $L \leq 1$). For notational simplicity, we use $D(X_1, X_2)$ to denote a distance between the distributions of any pair of random variables X_1 and X_2.

Lipschitz Functions. For some $K \geq 0$, the set of K-Lipschitz functions consists of all functions f satisfying:

$$\|f(x) - f(x')\| \leq K\|x - x'\|, \quad \forall x, x' \in \mathcal{X}.$$

We assume that the hypothesis class $\mathbb{H}$ is a subset of λ_H-Lipschitz functions for some constant $\lambda_H > 0$, and that the true labeling functions are λ-Lipschitz for some $\lambda > 0$.

Factual and Counterfactual Losses. Following Shalit et al. (2017), let $\Phi : \mathcal{X} \to \mathcal{R}$ be a representation function and $h : \mathcal{R} \times \{0, 1\} \to \mathcal{Y}$ be a hypothesis defined over the representation space. Let $Y_a \in \mathcal{Y}$ denote the potential outcome under treatment assignment $a \in \{0, 1\}$. We define the expected loss for unit-treatment pair (x, a) as:

$$\ell_{h,\Phi}(x, a) = \int_{\mathcal{Y}} L\big(Y_a, h(\Phi(x), a)\big)\, p(Y_a \mid x)\, dY_a,$$

where $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function. The expected factual and counterfactual losses are defined as:

$$\epsilon_F(h, \Phi) = \int_{\mathcal{X}\times\{0,1\}} \ell_{h,\Phi}(x, a)\, p(x, a)\, dxda,$$

$$\epsilon_{CF}(h, \Phi) = \int_{\mathcal{X}\times\{0,1\}} \ell_{h,\Phi}(x, a)\, p(x, 1 - a)\, dxda,$$

where $p(x, a)$ denotes the joint distribution over $\mathcal{X} \times \{0, 1\}$.

Let $u = p(a = 1)$ with $0 < u < 1$. We additionally define the reweighted factual loss:

$$\epsilon^\star_F(h, \Phi) = (1 - u)\, \epsilon^{a=1}_F(h, \Phi) + u\, \epsilon^{a=0}_F(h, \Phi),$$

which symmetrizes the factual error across treatment groups and will be used in the generalization bound.

Representation Learning-based Models. With these definitions established, we have $h(\Phi(x), a)$ as a predictor for an individual x's potential outcome under treatment assignment a. The goal is to find a pair of (h, Φ) that optimizes both the *factual loss* $\epsilon_F^\star(h, \Phi)$ and *counterfactual loss* $\epsilon_{CF}(h, \Phi)$. Note that low factual and counterfactual losses are both necessary and sufficient conditions for accurate potential outcome prediction (Aloui et al., 2023).

Counterfactual Error Estimation. However, the **counterfactual loss $\boldsymbol{\epsilon_{CF}(h, \Phi)}$ cannot be directly optimized** because the counterfactual outcomes are not observable in real-world scenarios. To this end, a group of well-established approaches minimize **upper bounds** of $\epsilon_{CF}(h, \Phi)$. These approaches are mainly based on the following result from Shalit et al. (2017), restated here.

Theorem 4.1 (Simplified Lemma A8 from Shalit et al. (2017), complete version provided in Appendix C.12.)**.** *Let $\Phi : \mathcal{X} \to \mathcal{R}$ be a one-to-one and Jacobian-normalized representation function. Let $h : R \times \{0, 1\} \to Y$ be a hypothesis with Lipschitz constant:*

$$\epsilon_{CF}(h, \Phi) \leq \epsilon_F^\star(h, \Phi) + 2 \cdot B_\Phi \cdot W_1(p_\Phi^0, p_\Phi^1), \tag{2}$$

where B_Φ is a constant and p_Φ^a is the distribution of the random variable $\Phi(X)$ conditioned on $A = a$, that is, representations for individuals receiving treatment $a \in \{0, 1\}$.

Motivation for Alignment. This theorem implies that a model (Φ, h) has low counterfactual error if ***(a)*** it has **low factual error** (which can be easily achieved by minimizing the prediction error on the observational data) and ***(b)*** the covariates of individuals from distinct treatment groups are **statistically similar to each other in the latent (representation) space**. Motivated by these, representation learning-based methods aim to align the treatment and control groups in the latent space while minimizing the factual error. In particular, successful alignment and low factual error guarantee a small value for the upper bound in Equation (2), implying the model has low counterfactual error. However, in practice, ***the error bound may be loose***, leaving the model performance suboptimal (Arjovsky & Bottou, 2017).

4.2 Benefits of Sub-treatment Group Alignment

To this end, we propose to use the **sub-treatment group structures** to achieve **tighter** counterfactual error bound up to an additive constant term, thus supporting **more effective** alignment.

Sub-treatment Groups. We hypothesize that each treatment group is a mixture of K sub-treatment groups in the latent space, and that **the sub-treatment groups across different treatment groups correspond to one another**. For example, in medical studies, patients may naturally form sub-groups **before** the beginning of experiments, based on latent variables such as demographic characteristics or genetic factors. Consider a scenario where patients are sub-grouped according to age (e.g., children, adults, seniors), gender, or genetic markers that influence their response to treatment. Even though these patients receive different treatments, the underlying characteristics defining the sub-groups are consistent across treatment groups. By aligning these corresponding sub-groups in the latent space, we can more effectively account for **hidden confounders** like genetic predispositions or socio-demographic factors, leading to more accurate estimation of treatment effects.

Recall p_Φ^a is the distribution of representations for individuals receiving treatment $a \in \{0, 1\}$, thus

$$p_\Phi^0 = \textstyle\sum_{k=1}^K w_k^0 P_{\Phi,k}^0, \quad p_\Phi^1 = \sum_{k=1}^K w_k^1 P_{\Phi,k}^1,$$

where for $a \in \{0, 1\}$, w_k^a represents the proportion of the k-th sub-group in treatment group a, and $P_{\Phi,k}^a$ is the distribution of representations of individuals in the k-th sub-group under treatment a.

Intuition of Sub-treatment Group Alignment (SGA). In many real-world settings, each treatment group is not homogeneous but instead forms a mixture of latent sub-populations. Marginal alignment minimizes the Wasserstein distance between the overall treatment distributions, but this coarse alignment can still permit substantial mismatch between corresponding latent sub-populations. In contrast, SGA decomposes each treatment distribution into sub-treatment components and aligns them pairwise. Intuitively,

instead of forcing two heterogeneous clouds of representations to overlap globally, SGA aligns them locally at the sub-population level, leading to more effective deconfounding.

Sub-treatment Group Alignment (SGA). SGA has the following alignment objective:

$$\sum_{k=1}^{K} w_k^1 W_1(P_{\Phi,k}^0, P_{\Phi,k}^1). \tag{3}$$

In particular, SGA minimizes the **weighted sum** of the Wasserstein distances between these **corresponding sub-treatment groups**. By aligning on a sub-treatment group level, SGA achieves more refined alignment. Indeed, motivated by the generalization bound in the field of *domain adaptation* (Liu et al., 2025), we next prove in Theorem 4.2 that SGA optimizes a finer-grained conditional objective that is theoretically tighter than marginal alignment (Theorem 4.1) up to an additive constant term, while empirically yielding a strictly tighter constraint on counterfactual risk in practice.

Theorem 4.2 (SGA Improves Generalization Bounds)**.** *Under the following assumptions:*

A1. *For all k, the sub-distributions $P_{\Phi,k}^0$ and $P_{\Phi,k}^1$ are Gaussian distributions with means m_k^0 and m_k^1, and covariances Σ_k^0 and Σ_k^1, respectively. The distance between corresponding sub-distributions is less than or equal to the distance between non-corresponding sub-distributions, i.e., $W_1(P_{\Phi,k}^0, P_{\Phi,k}^1) \leq W_1(P_{\Phi,k}^0, P_{\Phi,k'}^1)$ for $k \neq k'$.*

A2. *The sub-treatment group weights are identical across treatment groups, i.e., $w_k^1 = w_k^0$ for all $k \in [K]$. Moreover, there exists a small constant $\epsilon > 0$, such that $\max_{k\in[K]}(tr(\Sigma_k^0)) \leq \epsilon$ and $\max_{k\in[K]}(tr(\Sigma_k^1)) \leq \epsilon$.*

The following inequalities hold:

$$\epsilon_{CF}(h,\Phi) \leq \epsilon_F^\star(h,\Phi) + 2B_\Phi(\textstyle\sum_{k=1}^{K} w_k^1 W_1(P_{\Phi,k}^0, P_{\Phi,k}^1))$$
$$\textstyle\sum_{k=1}^{K} w_k^1 W_1(P_{\Phi,k}^0, P_{\Phi,k}^1) \leq W_1(p_\Phi^0, p_\Phi^1) + \delta_c,$$

where B_Φ is the same constant in Theorem 4.1 and δ_c is $4\sqrt{\epsilon}$.

See Appendix C for proof of Theorem 4.2 and detailed assumption statements and interpretations.

Remark 4.3. Because the identification of the sub-treatment group is *treatment-agnostic*, each unit's subgroup index is determined *solely* by the representation $\Phi(x)$, independent of the treatment label A. Consequently, whenever treatment assignment is randomized or conditionally ignorable given the representation ($A \perp k \mid \Phi(x)$), the probability of falling into cluster k is identical across arms: $\Pr(k \mid A{=}0) = \Pr(k \mid A{=}1) = w_k$, thus the use of w_k, w_k^0 and w_k^1 is interchangeable.

Remark 4.4. Theorem 4.2 establishes that SGA optimizes an upper bound that is theoretically tighter than the original bound in Theorem 4.1 up to an additive constant term δ_c. Equivalently, the SGA bound can be looser than the marginal one by at most δ_c, so its tightness is controlled by this additive constant. Crucially, however, δ_c is small in practice: as shown in Appendix E.1.3, the SGA alignment loss consistently attains strictly lower discrepancy values than marginal alignment across varying confounding levels. This empirical reduction in the Wasserstein-based discrepancy term directly supports the claim that SGA achieves a tighter practical bound and leads to improved counterfactual prediction accuracy.

Remark 4.5. Theorem 4.2 uses the assumptions regarding the distributional properties of sub-treatment groups in the learned representation space $\mathcal{R}$. Specifically, Assumption **A1** posits that sub-distributions $P_{\Phi,k}^0$ and $P_{\Phi,k}^1$ (for treatment arms $0, 1$) are Gaussian and that $W_1(P_{\Phi,k}^0, P_{\Phi,k}^1) \leq W_1(P_{\Phi,k}^0, P_{\Phi,k'}^1)$ for $k \neq k'$. Assumption **A2** states that the trace of their covariances is bounded by a small ϵ. We justify the validity of these assumptions below in Appendix C.

5 Framework

We propose a framework that ***synergistically integrates*** our two novel approaches: Sub-treatment Group Alignment (SGA) and Random Temporal Masking (RTM). Figure 3 illustrates the general architecture of our framework.

Model Architecture. A key feature of our framework is its versatility; it is not restricted to a specific architecture and can be integrated with various representation-based methods for time series causal inference. Typically, such methods consist of a *time series encoder* ϕ_E (parameterized by θ_E) that learns representations from input time series data, and an *outcome regressor* f_Y (parameterized by θ_Y) that predicts outcomes at the next time point. We note that the encoder ϕ_E can be instantiated with any sequence model architecture, such as RNNs, LSTMs (Hochreiter, 1997), or transformers (Vaswani, 2017). In Section 6, we demonstrate this flexibility by integrating our approaches with two well-established methods for time series causal inference: *Causal Transformer* (**CT**) (Melnychuk et al., 2022) and *Counterfactual Recurrent Networks* (**CRN**) (Bica et al., 2020).

Random Temporal Masking (RTM) *is applied to the observational data before the training of models.* To implement RTM, we mask covariates at a set of randomly selected time steps by replacing them with Gaussian noise. The model is subsequently trained to predict the outcomes despite these masked covariates, encouraging it to focus on causal information that is robust over time. RTM also reduces the risk of overfitting to factual outcomes at those selected time steps because, after masking, the covariates at current time is independent of the outcome. *This is particularly helpful when the potential outcomes at the current time steps are strongly correlated with current covariates* because under these scenarios the models are inclined to heavily rely on current covariates, thus overfitting to the factual distribution.

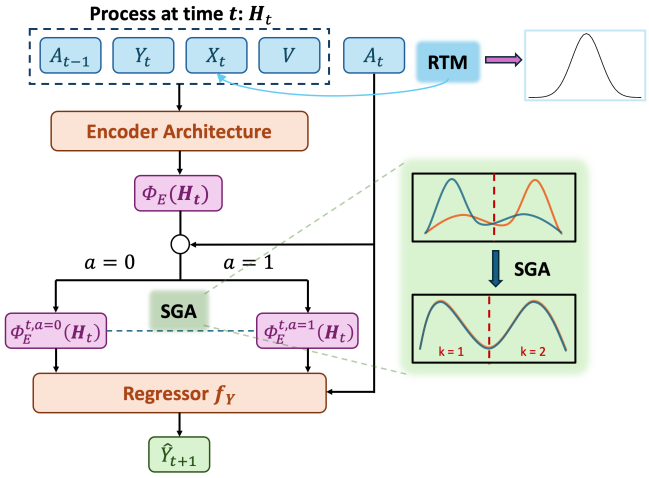


Figure 3: **Overview of SGA & RTM** at each timepoint. For simplicity, we show a binary treatment scenario.

Importantly, RTM does not explicitly model temporal dynamics; rather, it functions as a training-time regularization mechanism. By randomly masking covariates at selected time steps with Gaussian noise, RTM reduces the availability of contemporaneous information during training. Consequently, the model must rely more heavily on historical observations when predicting future outcomes. This encourages the learning of representations that capture dependencies stable across time rather than spurious correlations localized to a single time step.

In settings where current covariates are strongly correlated with factual outcomes, models without RTM may overfit to these contemporaneous signals, potentially impairing long-horizon counterfactual prediction. RTM mitigates this tendency by regularizing the hypothesis space and promoting temporal robustness. Empirically, this behavior is reflected in improved long-horizon performance and increased attention to past time steps as shown in Sec 6.5.

Objective Function. At each time step t, our framework optimizes the objective

$$\min_{\theta_Y,\theta_E} L_Y^t(\theta_Y,\theta_E) + \lambda L_D^t(\theta_E),$$

where L_Y^t represents the **factual outcome loss** and L_D^t denotes the **SGA loss** calculated with the Wasserstein-1 distance, balanced by λ. Details are provided below.

Factual Outcome Loss. At each time step t, the model learns to predict the observed outcomes, conditioned on $\mathbf{H}_t^{(i)}$ which contains the information from previous steps and the current covariates, by optimizing

$$L_Y^t(\theta_Y,\theta_E) = \tfrac{1}{N}\sum_{i=1}^{N}(\ell(y_i^{t+1},\hat{y}_i^{t+1})),$$

where $\hat{y}_i^{t+1} = f_Y\left(\phi_E\left(\mathbf{H}_t^{(i)}, A_t^{(i)}\right)\right)$ and $\ell(\cdot,\cdot)$ denotes the loss function (e.g., mean squared error).

SGA Loss. Motivated by Section 4, our framework aligns the sub-treatment groups across distinct treatment groups. To this end, at each time step t and for each treatment group a, we use Gaussian Mixture Models (GMMs) to cluster the individuals' features in the representation space into K sub-treatment groups. Let the random variable $\phi_E^{t,a,k}(\mathbf{H}_t)$ denote the representations of samples in the k-th sub-group of treatment group a at time step t.

To accommodate the applications with **multiple treatment groups** (more than two), for each time step t and each corresponding sub-treatment group, we align the sub-treatment groups with **the uniform mixtures of them**. That is, for all the k-th sub-treatment groups in all $|\mathcal{A}|$ treatment groups where $|\mathcal{A}|$ is the total number of treatments, we first create a mixture of them with uniform weights and align all of them with the uniform mixture. Note that by triangle inequality this is a **sufficient condition** to align multiple groups well. Specifically, the SGA loss is defined as:

$$L_D^t(\theta_E) = \sum_{k=1}^{K} \sum_{a \in \mathcal{A}} w_k^{t,a} W_1(\phi_E^{t,a,k}(\mathbf{H}_t), \phi_E^{t,k}(\mathbf{H}_t)),$$

where $w_k^{t,a}$ represents the proportion of samples in sub-group k of treatment group a, and $\phi_E^{t,k}(\mathbf{H}_t)$ is the uniform mixture of $\{\phi_E^{t,a,k}(\mathbf{H}_t)\}_{a \in \mathcal{A}}$. Note that all the quantities in $L_D^t(\theta_E)$ can be estimated from the observational data. We provide implementation details and our algorithm in Appendix D.

6 Experiments

Main Results. We evaluate our framework on a fully-synthetic Pharmacokinetic-Pharmacodynamic (PK-PD) benchmark (Bica et al., 2020; Geng et al., 2017) and a semi-synthetic dataset derived from MIMIC-III (Johnson et al., 2016). We integrate our framework into the architectures of the LSTM-based *Counterfactual Recurrent Networks (CRN)* (Bica et al., 2020) and the Transformer-based *Causal Transformer (CT)* (Melnychuk et al., 2022). Across both datasets, our experiments consistently demonstrate that the ***synergistic combination*** of SGA and RTM achieves state-of-the-art (SOTA) performance in counterfactual outcome estimation (Section 6.1 and 6.2). Full details are deferred to Appendix E.

Analysis. We run comprehensive ablation studies, showing **(a)** individual contributions of SGA and RTM (Section 6.3), **(b)** detailed analyses of RTM's masking strategy (Section 6.4), **(c)** RTM's impact on attention mechanisms (Section 6.5), and **(d)** SGA's sensitivity to clustering choices (Section 6.6).

Baseline Methods. We compare against SOTA baselines: Marginal Structural Models (**MSMs**) (Hernán et al., 2001; Robins et al., 2000), Recurrent Marginal Structural Networks (**RMSNs**) (Lim, 2018), G-Net (Li et al., 2020), Counterfactual Recurrent Networks (**CRN**) (Bica et al., 2020), and Causal Transformer (**CT**) (Melnychuk et al., 2022).

6.1 Experiments on Fully-Synthetic Data from a PK-PD Model of Tumor Growth

We first evaluate on the fully-synthetic data frequently used in the literature (Bica et al., 2020; Melnychuk et al., 2022), which allows simulation of treatment-response dynamics and varying levels of time-dependent confounding.

Tasks and Evaluation Metrics. Following Melnychuk et al. (2022), we report normalized Root Mean Squared Error (RMSE) on both *one-step-ahead* and *τ-step-ahead* counterfactual predictions under *varying levels* of time-varying confounding, indexed by γ. Details on dataset generation, hyperparameters, visualization of the representation space are in Appendix E.1.

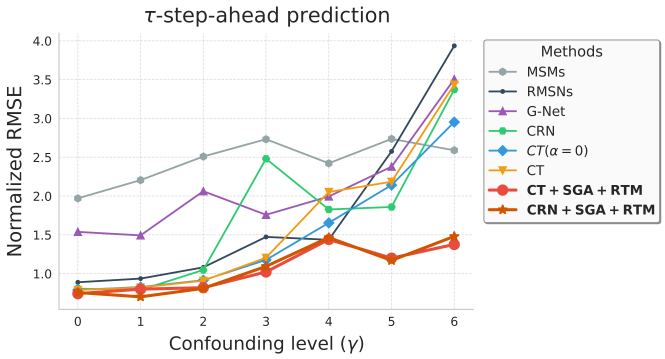


Figure 4: Performances on τ-step-ahead (τ=6) prediction. Note that CT (α=0) refers to CT w/o alignment.

Statistical Significance and Robustness. To further assess statistical significance and robustness, we additionally report results with error bars in Figure 5.

This figure extends the analysis of Figure 4 by incorporating mean ± standard deviation across multiple runs for the challenging long-term ($\tau = 6$) counterfactual prediction task under varying confounding levels γ.

Results. As shown in Figure 4 and Table 1, integrating both SGA and RTM into CRN and CT **significantly improves** their performance compared to the vanilla models. Notably, our framework performs exceptionally well in scenarios with **high levels of confounding**, indicating our effectiveness in deconfounding. Baselines results for confounding levels $\gamma \in [0, 4]$ are cited from Melnychuk et al. (2022), and we extend experiments to higher confounding levels to more thoroughly test the robustness of our methods in complex causal scenarios. As shown in Figure 5, both CRN+SGA+RTM and CT+SGA+RTM maintain competitive or superior performance with relatively tight error bars as γ increases. This demonstrates that our synergistic integration of Sub-treatment Group Alignment (SGA) and Random Temporal Masking (RTM) not only improves predictive accuracy but also enhances stability under increasing time-varying confounding.

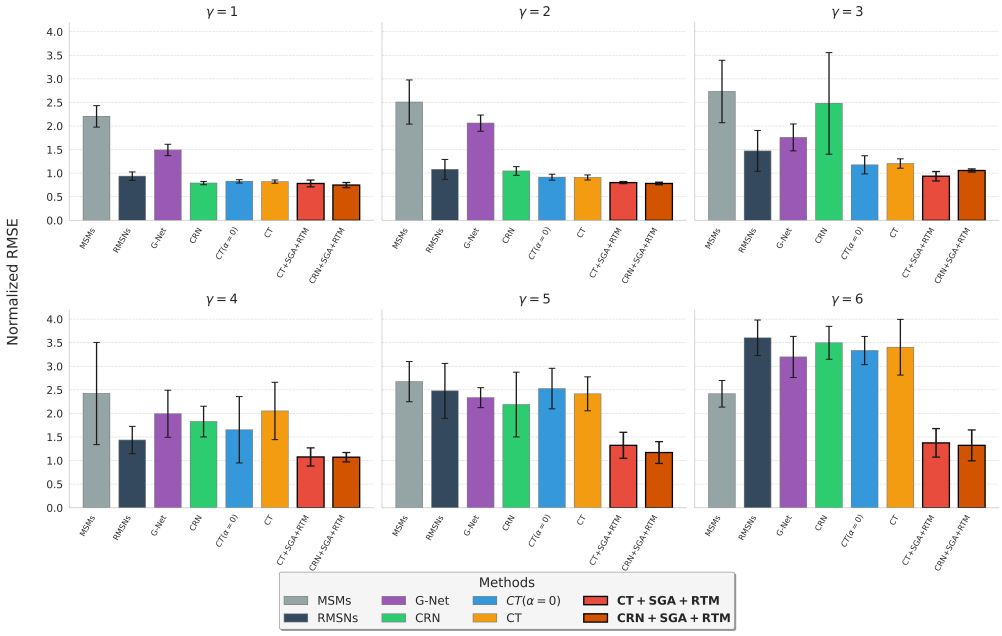


Figure 5: **Impact of varying confounding level γ on long-term ($\tau = 6$) prediction performance with error bars**. This figure extends Figure 4 by reporting mean ± standard deviation over multiple runs.

6.2 Experiments on Semi-Synthetic Data

We further validate our framework on a semi-synthetic dataset based on real-world medical data from intensive care units. This dataset is generated following the approach of Melnychuk et al. (2022), which builds upon the MIMIC-III dataset (Johnson et al., 2016) to simulate patient trajectories with outcomes that reflect both endogenous and exogenous dependencies while incorporating treatment effects.

Results and Analysis. As shown in Table 2, applying our framework on top of CRN and CT **consistently improves performance** compared to the vanilla models. Furthermore, our methods yield performance comparable to the SOTA models. This is consistent with the findings reported in Melnychuk et al. (2022) that confounding may not be the primary challenge in this specific dataset, as there is also minimal difference between the performance of CT and CT($\alpha = 0$). This observation implies that, in this semi-synthetic dataset, the level of confounding may be relatively low. To this end, given that the strength of our approach lies in reducing confounding,it is expected that the performance gain is marginal compared to existing state-of-the-art methods. Details are in Appendix E.2.

Table 1: Normalized RMSE for one-step-ahead and τ-step-ahead predictions on fully synthetic data, comparing vanilla CRN/CT with CRN/CT enhanced with SGA + RTM. The values in **black** indicate lower RMSE for CRN-based models, and values in **violet** indicate lower RMSE for CT-based models.

τ	Method	$\gamma = 0$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$	$\gamma = 6$
$\tau = 1$	**CRN**	0.755	0.788	0.881	1.062	1.358	1.403	2.031
	CRN + SGA + RTM	**0.720**	**0.741**	**0.822**	**0.802**	**0.949**	**1.107**	**1.808**
	CT	0.770	0.783	0.864	1.098	1.413	1.497	2.204
	CT + SGA + RTM	**0.750**	**0.736**	**0.765**	**0.883**	**0.983**	**1.340**	**1.975**
$\tau = 2$	**CRN**	0.671	0.666	0.741	1.668	1.151	1.456	3.117
	CRN + SGA + RTM	**0.629**	**0.648**	**0.674**	**0.822**	**0.934**	**0.948**	**1.308**
	CT	0.681	0.677	0.713	0.908	1.274	1.718	3.053
	CT + SGA + RTM	**0.610**	**0.673**	**0.688**	**0.751**	**0.786**	**0.911**	**0.983**
$\tau = 3$	**CRN**	0.700	0.692	0.818	1.959	1.360	1.684	3.451
	CRN + SGA + RTM	**0.643**	**0.666**	**0.706**	**0.862**	**0.981**	**0.997**	**1.404**
	CT	0.703	0.712	0.770	1.010	1.536	1.948	3.367
	CT + SGA + RTM	**0.630**	**0.691**	**0.718**	**0.805**	**0.858**	**1.023**	**1.127**
$\tau = 4$	**CRN**	0.734	0.722	0.898	2.201	1.573	1.802	3.594
	CRN + SGA + RTM	**0.662**	**0.681**	**0.740**	**0.922**	**1.016**	**1.060**	**1.478**
	CT	0.726	0.748	0.822	1.089	1.762	2.095	3.516
	CT + SGA + RTM	**0.650**	**0.727**	**0.748**	**0.848**	**0.919**	**1.107**	**1.231**
$\tau = 5$	**CRN**	0.769	0.755	0.976	2.361	1.730	1.870	3.560
	CRN + SGA + RTM	**0.682**	**0.700**	**0.771**	**0.983**	**1.046**	**1.118**	**1.497**
	CT	0.756	0.786	0.870	1.154	1.922	2.168	3.570
	CT + SGA + RTM	**0.677**	**0.761**	**0.779**	**0.889**	**0.995**	**1.162**	**1.315**
$\tau = 6$	**CRN**	0.807	0.790	1.047	2.480	1.827	1.859	3.372
	CRN + SGA + RTM	**0.701**	**0.715**	**0.792**	**1.031**	**1.064**	**1.170**	**1.480**
	CT	0.789	0.821	0.909	1.205	2.052	2.184	3.436
	CT + SGA + RTM	**0.705**	**0.800**	**0.801**	**0.935**	**1.058**	**1.200**	**1.376**

Table 2: RMSE for one-step-ahead and τ-step-ahead predictions on semi-synthetic data based on real-world medical data (MIMIC-III).

	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
MSMs	0.37	0.57	0.74	0.88	1.14	1.95	3.44	> 10.0	> 10.0	> 10.0
RMSNs	0.24	0.47	0.60	0.70	0.78	0.84	0.89	0.94	0.97	1.00
G-Net	0.34	0.67	0.83	0.94	1.03	1.10	1.16	1.21	1.25	1.29
CRN	0.30	0.48	0.59	0.65	0.68	0.71	0.72	0.74	0.76	0.78
CRN + SGA + RTM	**0.27**	**0.43**	**0.52**	**0.58**	**0.62**	**0.65**	**0.67**	**0.69**	**0.72**	**0.73**
CT ($\alpha = 0$)	**0.20**	**0.38**	0.46	**0.50**	**0.52**	0.54	0.56	**0.57**	0.59	0.60
CT	0.21	**0.38**	0.46	**0.50**	0.53	0.54	**0.55**	**0.57**	**0.58**	0.59
CT + SGA + RTM	0.21	**0.38**	**0.44**	**0.50**	**0.52**	**0.52**	0.56	**0.57**	**0.58**	**0.58**

6.3 Ablation Studies

We run ablation experiments to evaluate the individual contributions of **SGA** and **RTM**.

Effectiveness of SGA. The left panel of Table 3 demonstrates that incorporating SGA **alone** consistently improves performance over baseline models. The improvements are particularly evident in settings with **higher** confounding levels (larger γ). This supports our claim that SGA's fine-grained sub-treatment group alignment leads to more effective deconfounding.

Effectiveness of RTM. The right panel of Table 3 shows that adding RTM consistently improves model performance. RTM's impact is more pronounced on **longer-term** predictions (larger τ) and under **stronger** confounding (larger γ), confirming that RTM promotes better temporal generalization.

6.4 Analysis of Masking Strategies

To evaluate our choice of Gaussian noise for RTM, we compare it against alternative masking strategies: (*i*) *masking with zeros* and (*ii*) *interpolation-masking* that replaces input covariates from selected time-point t with the **average** covariates of $t - 1$ and $t + 1$. We observe the followings:

O1. Gaussian noise significantly outperforms other choices. We hypothesize that this is because Gaussian noise acts as a more effective regularizer. It discourages the model from overfitting to specific input

Table 3: Normalized RMSE for τ-step-ahead predictions on fully-synthetic data, CRN/CT vs. CRN/CT+SGA (left) and CRN/CT vs. CRN/CT+RTM (right).

γ	Method	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
2	**CRN**	0.818	0.898	0.976	1.047
	CRN + SGA	**0.698**	**0.743**	**0.782**	**0.810**
	CT	0.770	0.822	0.870	0.909
	CT + SGA	**0.762**	**0.813**	**0.854**	**0.876**
6	**CRN**	3.451	3.594	3.560	3.372
	CRN + SGA	**3.157**	**3.284**	**3.231**	**3.034**
	CT	3.367	3.516	3.570	3.436
	CT + SGA	**3.231**	**3.345**	**3.238**	**3.014**

γ	Method	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
2	**CRN**	0.818	0.898	0.976	1.047
	CRN + RTM	**0.791**	**0.862**	**0.907**	**0.934**
	CT	0.770	0.822	0.870	0.909
	CT + RTM	**0.720**	**0.752**	**0.787**	**0.819**
6	**CRN**	3.451	3.594	3.560	3.372
	CRN + RTM	**2.316**	**2.469**	**2.541**	**2.523**
	CT	3.367	3.516	3.570	3.436
	CT + RTM	**1.883**	**2.009**	**2.039**	**2.005**

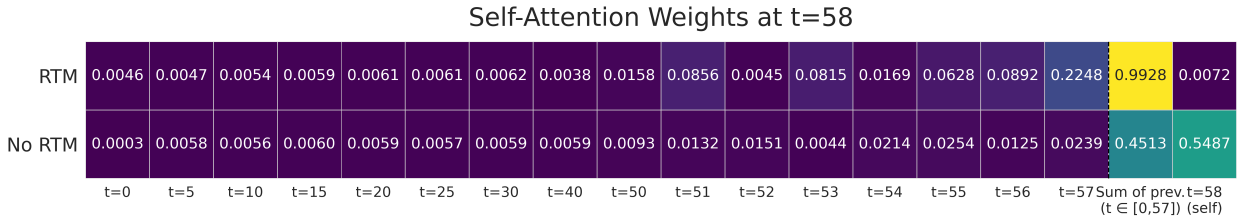


Figure 6: **Heatmap of self-attention weights. Left:** attention to past time points. **Rightmost columns:** sum of total past attention (`Sum of prev.`), and attention to the current time point (`t=58`).

values at the masked time steps and forces it to learn more generalizable representations from the historical context.

Table 4: Comparison of masking strategies on fully-synthetic data (γ=6) with normalized RMSE. Gaussian noise refers to CT+RTM. Other strategies are also applied to the CT model architecture.

Masking Strategy	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$
Zero masking	2.821	3.072	3.122	2.986
Interpolation	3.435	3.556	3.602	3.480
No Masking	3.367	3.516	3.570	3.436
Gaussian noise	**1.883**	**2.009**	**2.039**	**2.005**

O2. Zero masking performs sub-optimally. It introduces a strong, and often unnatural signal at zero. Models risk using these artificial zeros or misinterpreting them, which negatively affects the learning of true underlying patterns.

O3. Interpolation yields the highest error. It creates artificially smooth data. This obscures natural data variability and can potentially impose a deterministic bias, preventing the model from capturing true temporal dynamics.

6.5 RTM Attention Weight Analysis

To investigate how RTM promotes learning of long-range dependencies, we show the softmaxed self-attention weights from the final attention layer of Causal Transformer model. Figure 6 compares attention patterns for CT+RTM versus CT (no RTM) when predicting the outcome at the last time-point ($t = 58$) under high confounding ($\gamma = 6$). This analysis confirms RTM's success in shifting the model's reliance from current time step towards historical context, thereby promoting temporal generalization and the learning of long-term causal patterns, as Figure 6 shows:

- **With RTM, the model allocates over 99% of its attention to previous time-points**. This distributed attention across historical data demonstrates effective leveraging of past information.

- **Without RTM, attention to past steps is significantly lower,** with the model focusing over 54% of its attention on the current time-point. This highlights the tendency of models without RTM to **over-rely on information at the current time step** rather than learn long-term causal patterns.

6.6 Clustering Sensitivity

In our experiments, SGA uses Gaussian Mixture Models (GMMs) for *treatment-agnostic* clustering to identify sub-treatment groups. We analyze SGA's sensitivity to the choice of clustering algorithm and the number of clusters (K). Specifically, we compare GMM with k-means using varying K.

Evaluations are performed on the fully-synthetic dataset under the challenging setting of **long-term predictions** (τ=6) and **higher confounding** (γ=6). We observe from Table 5:

O1. SGA's performance **remains relatively stable** across different numbers of clusters, suggesting low sensitivity to this hyperparameter.

O2. Both k-means and GMM yield comparable performance, indicating robustness to the specific clustering algorithm used.

Table 5: Sensitivity analysis of SGA to the choice of clustering algorithm and the number of clusters (K). The table reports the normalized RMSE with $\gamma = 6$ at $\tau = 6$.

Model	Clustering Algorithm	No SGA (K=1)	SGA with varying K			
			K=2	K=3	K=5	K=10
CRN+SGA	**k-means**	3.372	3.568	3.301	**3.259**	3.504
	GMM	3.372	**3.034**	3.258	3.276	3.595
CT+SGA	**k-means**	3.436	3.154	**3.008**	3.206	3.059
	GMM	3.436	**3.014**	3.423	3.158	3.082

These findings confirm SGA's **robustness** to the clustering algorithm and the number of clusters, making it practical for diverse real-world settings where the optimal number of clusters may be unknown or require tuning.

7 Conclusion

We address the critical challenge of counterfactual outcome estimation in time series by introducing two novel, synergistic approaches: Sub-treatment Group Alignment (SGA) and Random Temporal Masking (RTM). SGA tackles time-varying confounding at each time point by aligning fine-grained sub-treatment group distributions, leading to tighter counterfactual error bound up to an additive constant term and more effective deconfounding. Complementarily, RTM promotes temporal generalization and robust learning from historical patterns by randomly masking covariates, encouraging the model to preserve underlying causal relationships across time steps and rely less on potentially noisy contemporaneous time steps.

Our comprehensive experiments demonstrate that SGA and RTM are broadly applicable and significantly enhance existing state-of-the-art methods. While each approach individually improves performance, their synergistic combination consistently achieves SOTA performance on both synthetic and semi-synthetic benchmark datasets. Together, SGA and RTM offer a flexible and effective framework for improving causal inference from observational time series data.

References

Ahmed Aloui, Juncheng Dong, Cat P Le, and Vahid Tarokh. Transfer learning for individual treatment effect estimation. In *Uncertainty in Artificial Intelligence*, pp. 56–66. PMLR, 2023.

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Advances in Neural Information Processing Systems*, 30, 2017.

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

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

Søren Bisgaard and Murat Kulahci. *Time series analysis and forecasting by example*. John Wiley & Sons, 2011.

Aditi Chaudhary, Karthik Raman, Krishna Srinivasan, and Jiecao Chen. Dict-mlm: Improved multilingual pre-training using bilingual dictionaries. *arXiv preprint arXiv:2010.12566*, 2020.

Tamara Czinczoll, Christoph Hönes, Maximilian Schall, and Gerard de Melo. NextLevelBERT: Investigating masked language modeling with higher-Level representations for long documents. *arXiv preprint arXiv:2402.17682*, 2024.

Julie Delon and Agnes Desolneux. A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *Data Mining and Knowledge Discovery*, 35(4):1713–1738, 2021.

John R Freeman. Granger causality and the times series analysis of political relationships. *American Journal of Political Science*, pp. 327–358, 1983.

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

Ghulam Ghouse, Atiq Ur Rehman, and Muhammad Ishaq Bhatti. Understanding of causes of spurious associations: Problems and prospects. *Journal of Statistical Theory and Applications*, 23(1):44–66, 2024.

Lars J Grimm, Habib Rahbar, Monica Abdelmalak, Allison H Hall, and Marc D Ryser. Ductal carcinoma in situ: state-of-the-art review. *Radiology*, 302(2):246–255, 2022.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems*, 30, 2017.

Miguel A Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96(454): 440–448, 2001.

Miguel Ángel Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men, 2000.

S Hochreiter. Long Short-term Memory. *Neural Computation MIT-Press*, 1997.

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

Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23 (166):1–50, 2022.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pp. 5067–5077. PMLR, 2020.

Rui Li, Zach Shahn, Jun Li, Mingyu Lu, Prithwish Chakraborty, Daby Sow, Mohamed Ghalwash, and Li-wei H Lehman. G-net: a deep learning approach to g-computation for counterfactual outcome prediction under dynamic treatment regimes. *arXiv preprint arXiv:2003.10551*, 2020.

Sheng Li and Yun Fu. Matching on balanced nonlinear representations for treatment effects estimation. *Advances in Neural Information Processing Systems*, 30, 2017.

Kevin J Liang, Chunyuan Li, Guoyin Wang, and Lawrence Carin. Generative adversarial network training is a continual learning problem. *arXiv preprint arXiv:1811.11083*, 2018.

Bryan Lim. Forecasting treatment responses over time using recurrent marginal structural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Yiling Liu, Jiayi Dong, Ziyang Jiang, Ahmed Aloui, Kun Li, Michael H. Klein, and David Carlson. Understanding and robustifying sub-domain alignment for domain adaptation. *Transactions on Machine Learning Research*, March 2025. 2025-March.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rényi divergence. *arXiv preprint arXiv:1205.2628*, 2012.

Mohammad Ali Mansournia, Goodarz Danaei, Mohammad Hossein Forouzanfar, Mahmood Mahmoodi, Mohsen Jamali, Nasrin Mansournia, and Kazem Mohammad. Effect of physical activity on functional performance and knee pain in patients with osteoarthritis: analysis with marginal structural models. *Epidemiology*, 23(4):631–640, 2012.

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

Mohammad Amin Morid, Olivia R Liu Sheng, and Joseph Dunbar. Time series prediction using deep learning methods in healthcare. *ACM Transactions on Management Information Systems*, 14(1):1–29, 2023.

Khai Nguyen, Tongzheng Ren, Huy Nguyen, Litu Rout, Tan Nguyen, and Nhat Ho. Hierarchical sliced Wasserstein Distance. *arXiv preprint arXiv:2209.13570*, 2022.

Ofir Pele and Michael Werman. Fast and robust Earth Mover's Distances. In *2009 IEEE 12th international Conference on Computer Vision*, pp. 460–467. IEEE, 2009.

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

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

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

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

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

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

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

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

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

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

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 222–235, 2020.

Yanbo Xu, Yanxun Xu, and Suchi Saria. A bayesian nonparametric approach for estimating individualized treatment-response curves. In *Machine Learning for Healthcare Conference*, pp. 282–300. PMLR, 2016.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.