

---

# Causal Contrastive Learning for Counterfactual Regression Over Time

---

Mouad El Bouchattaoui<sup>\*1,2</sup>, Myriam Tami<sup>1</sup>, Benoit Lepetit<sup>2</sup>, and Paul-Henry Cournède<sup>1</sup>

<sup>1</sup>Paris-Saclay University, CentraleSupélec, MICS Lab, Gif-sur-Yvette, France

<sup>2</sup>Saint-Gobain, Paris, France

## Abstract

Estimating treatment effects over time holds significance in various domains, including precision medicine, epidemiology, economy, and marketing. This paper introduces a unique approach to counterfactual regression over time, emphasizing long-term predictions. Distinguishing itself from existing models like Causal Transformer, our approach highlights the efficacy of employing RNNs for long-term forecasting, complemented by Contrastive Predictive Coding (CPC) and Information Maximization (InfoMax). Emphasizing efficiency, we avoid the need for computationally expensive transformers. Leveraging CPC, our method captures long-term dependencies in the presence of time-varying confounders. Notably, recent models have disregarded the importance of invertible representation, compromising identification assumptions. To remedy this, we employ the InfoMax principle, maximizing a lower bound of mutual information between sequence data and its representation. Our method achieves state-of-the-art counterfactual estimation results using both synthetic and real-world data, marking the pioneering incorporation of Contrastive Predictive Encoding in causal inference.

## 1 Introduction

It's vital in real-world applications to estimate potential responses, i.e., responses under hypothetical treatment strategies. Individuals show diverse responses to the same treatment, emphasizing the need to quantify individual response trajectories. This enables personalized interventions, enhancing decision-making efficacy. In medical contexts, precise response estimation enables tailored treatments for patients [2, 70, 49]. This paper focuses on *counterfactual regression over time*, estimating responses under hypothetical treatment plans based on individual records, including past covariates, responses, and treatment sequences up to the current prediction time [64, 62]. In addressing the challenges of this time-varying setting, we tackle: (1) **Time-dependent confounding** [55]: confounders influenced by past treatment, impacting subsequent treatments and responses. (2) **Selection bias**: imbalanced covariate distributions across treatment regimes in observational data, requiring time-aware handling beyond methods in static settings [64, 67, 41]. (3) **Long-term dependencies**: enduring interdependencies among covariates, treatments, and responses, enabling long-range interactions [15, 54].

Recent advancements in neural networks, such as Recurrent Marginal Structural Networks (RMSNs) [42], Counterfactual Recurrent Networks (CRN) [7], and G-Net [39], have tackled these causal inference challenges. However, their reliance on RNNs limits their ability to capture long-term dependencies. Recent studies [48] propose integrating transformers to better represent temporal dynamics. Rather than viewing this as a limitation of RNNs, we see it as an opportunity to emphasize their strengths. We design specific architectures for counterfactual regression over large horizons, avoiding complex, hard-to-interpret models like transformers. Our approach leverages

---

<sup>\*</sup>[mouad.el-bouchattaoui@centralesupelec.fr](mailto:mouad.el-bouchattaoui@centralesupelec.fr)

the computational efficiency of RNNs, incorporating Contrastive Predictive Coding (CPC) [52, 29] for learning data history representations. This enhances model performance while maintaining efficiency, offering a compelling alternative to transformer-based approaches. Furthermore, we usually formulate identification assumptions of counterfactual responses over the original process history space (Appendix B.1). However, these assumptions may not hold in the representation space for arbitrary functions. Since identification often involves conditional independence, it applies when using an invertible representation function. Current models for time-varying settings [42, 7, 48] do not enforce representation invertibility. To address this, instead of adding complexity with a decoder, we *implicitly* push the history process to be "reconstructable" from the encoded representation by maximizing Mutual Information (MI) between representation and input, following the InfoMax principle [43], akin to Deep InfoMax [32].

## 2 Contributions

Our approach is inspired by self-supervised learning using MI objectives [32]. We aim to maximize MI between different views of the same input, introducing counterfactual regression over time through CPC to capture long-term dependencies. Additionally, we propose a tractable lower bound to the original InfoMax objectives for more efficient representations. This is challenging due to the sequential nature and high dimensionality, marking a novelty. We demonstrate the importance of regularization terms via an ablation study. Previous work leveraging contrastive learning for causal inference applies only to the static setting with no theoretical grounding [16]. To our knowledge, we frame for the first time the representation balancing problem from an information-theoretic perspective and show that the suggested adversarial game (Theorem 5.4) yields theoretically balanced representations using the Contrastive Log-ratio Upper Bound (CLUB) of MI, computed efficiently. Key innovations of our Causal CPC model include: (1) We showcase the capability of leveraging CPC to capture long-term dependencies in the process history using InfoNCE [25, 26, 52], an unexplored area in counterfactual regression over time where its integration into process history modeling is not straightforward in causality. (2) We enforce input reconstruction from representation by contrasting the representation with its input. Such quality is generally overlooked in baselines, yet it ensures that confounding information is retained, preventing biased counterfactual estimation. (3) Applying InfoMax to process history while respecting its dynamic nature is challenging. We provide a tractable lower bound to the original InfoMax problem, also bringing theoretical insights on the bound's tightness. (4) We suggest minimizing an upper bound on MI between representation and treatment to make the representation non-predictive of the treatment, using the CLUB of MI [13]. This novel information-theoretic perspective results in a theoretically balanced representation across all treatment regimes. (5) By using a simple Gated Recurrent Unit (GRU) layer [14] as the model backbone, we demonstrate that well-designed regularizations can outperform more complex models like transformers. Finally, our experiments on synthetic data (cancer simulation [24]) and semi-synthetic data based on real-world datasets (MIMIC-III [35]) show the superiority of Causal CPC at accurately estimating counterfactual responses.

## 3 Related Work

**Models for Counterfactual Regression Through Time** Traditionally, causal inference addresses time-varying confounders using Marginal Structural Models (MSMs) [64], which rely on inverse probability of treatment weighting [62]. However, MSMs can yield high variance estimates, especially with extreme values, and are limited to pooled logistic regression, impractical for high-dimensional, dynamic data. RMSNs [41] enhance MSMs by integrating RNNs for propensity and outcome modeling. CRN [7] employs adversarial domain training with a gradient reversal layer [23] to establish a treatment-invariant representation space, reducing bias induced by time-varying confounders. Similarly, G-Net [39] combines g-computation and RNNs to predict counterfactuals in dynamic treatment regimes. Causal Transformer (CT) [48] uses transformers to estimate counterfactuals over time and handles selection bias by learning a treatment-invariant representation via Counterfactual Domain Confusion loss (CDC) [78]. These models, like ours, assume *sequential ignorability* [62], in contrast to a body of work which does not fully verify our assumptions [45, 73, 68, 60, 80, 79, 8, 28, 12, 38, 58, 69, 19, 10, 31, 34, 6, 11, 82, 22], which we discuss in detail in Appendix C.1.2. In contrast, we introduce a contrastive learning approach to capture long-term dependencies while maintaining a simple model and ensuring high computational efficiency in both training and prediction. This demonstrates that simple models with well-designed regularization terms can still achieve high

prediction quality. Additionally, previous works [62, 64, 41, 39, 48] did not consider the role of invertible representation in improving counterfactual regression. Here, we introduce an InfoMax regularization term to make our encoder easier to invert. Appendix C.1 provides a detailed overview of counterfactual regression models.

**InfoMax Principle** The InfoMax principle aims to learn a representation that maximizes MI with its input [43, 5]. Estimating MI for high-dimensional data is challenging, often addressed by maximizing a simple and tractable lower bound on MI [32, 56, 40]. Another approach involves maximizing MI between two lower-dimensional representations of different views of the same input [3, 29, 75, 77], offering a more practical solution. We adopt this strategy by dividing our process history into two views, past and future, and maximizing a tractable lower bound on MI between them. This encourages a "reconstructable" representation of the process history. To our knowledge, the only work applying an InfoMax approach to counterfactual regression, albeit in static settings, is [16]. They propose maximizing MI between an individual's representation and a global representation, aggregating information from all individuals into a single vector. However, the global representation lacks clarity and interpretability, raising uncertainties about its theoretical underpinnings in capturing confounders. Furthermore, there's a lack of theoretical analysis on how minimizing MI between individual and treatment-predictive representations could yield a treatment-invariant representation. As a novelty, we extend the InfoMax principle to longitudinal data, providing a theoretical guarantee of learning balanced representations. Appendix C.2 discusses self-supervision and MI, with all proofs in Appendix G.

## 4 Problem Formulation

**Setup** In the framework of Potential Outcomes (PO) [65], and following [63], we track a cohort of individuals (units)  $i \in \{1, 2, \dots, N\}$  over  $t_{max}$  time steps. At each time  $t \in \{1, 2, \dots, t_{max}\}$ , we observe the following: (1) **Discrete treatment**  $W_{it} \in \mathcal{W} = \{0, 1, \dots, K-1\}$ , e.g., in medical contexts,  $W_{it}$  may represent treatments like radiotherapy or chemotherapy. (2) **Outcome of interest**  $Y_{it} \in \mathcal{Y} \subset \mathbb{R}$ , such as tumor volume. (3) **Time-varying context**  $\mathbf{X}_{it} \in \mathcal{X} \subset \mathbb{R}^{d_x}$ , containing information about the individual that may influence treatment decisions and outcomes.  $\mathbf{X}_{it}$  is a  $d_x$ -dimensional vector of confounders, such as health records or clinical measurements. (4) **Static confounders**  $\mathbf{V} \in \mathcal{V} \subset \mathbb{R}^{d_v}$ , such as gender, which remain constant over time. (5) **Partially observed potential outcomes**  $Y_{it}(\omega_{i,\leq t})$ , representing the outcomes that *would have been* observed for individual  $i$  at time  $t$  under treatment sequence  $\omega_{i,\leq t} = (\omega_{i,1}, \dots, \omega_{i,t}) \in \mathcal{W}^t$ . We define the history process up to time  $t+1$  as  $\mathbf{H}_{t+1} = [\mathbf{V}, \mathbf{X}_{\leq t+1}, W_{\leq t}, Y_{\leq t}]$ , capturing all information prior to the assignment of treatment  $W_{t+1}$ . This history is illustrated in the causal graph shown in Figure 1.

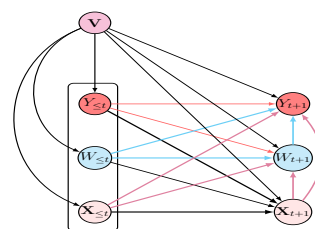


Figure 1: Causal graph over  $\mathbf{H}_{t+1}$

**Goal** Given a training dataset  $\{\mathbf{H}_{i,t+1}, i = 1, \dots, N\}$  sampled from the empirical distribution  $\mathbb{P}_{\mathbf{H}_{t+1}}$ , we address the following causal inference problem: *Given a history process  $\mathbf{H}_{t+1}$ , how can we efficiently estimate counterfactual responses up to time  $t + \tau$  (where  $\tau \geq 1$  is the prediction horizon) for a potential treatment sequence  $\omega_{t+1:t+\tau} = (\omega_{t+1}, \dots, \omega_{t+\tau})$ ?* The goal is to estimate the causal quantity:  $\mathbb{E}(Y_{t+\tau}(\omega_{t+1:t+\tau}) \mid \mathbf{H}_{t+1})$ , i.e., the expected outcome at time  $t + \tau$ , given the history  $\mathbf{H}_{t+1}$  and a sequence of treatments  $\omega_{t+1:t+\tau}$ . We identify this causal quantity from observational data using the assumption of sequential ignorability [62, 42, 39, 48], which is implicitly assumed in Figure 1 and explicitly discussed in Appendix B.1. This allows us to express the counterfactual as:

$$\mathbb{E}(Y_{t+\tau}(\omega_{t+1:t+\tau}) \mid \mathbf{H}_{t+1}) = \mathbb{E}(Y_{t+\tau} \mid \mathbf{H}_{t+1}, W_{t+1:t+\tau} = \omega_{t+1:t+\tau}).$$

## 5 Causal CPC

### 5.1 Representation Learning

**Contrastive Predictive Coding** We employ contrastive learning to efficiently represent the process history  $\mathbf{H}_t$ . Causal forecasting over multiple time horizons requires representations that capture variability in  $\mathbf{H}_t$ . For short-term predictions, local features and smooth signal variations are critical,

while long-term predictions rely on capturing global structures and long-term dependencies, as shared information between history and future points diminishes.

To achieve this, we learn a representation of  $\mathbf{H}_t$  that predicts future components over multiple time steps. For each horizon  $j = 1, \dots, \tau$ , future components are defined as  $\mathbf{U}_{t+j} = [\mathbf{V}, \mathbf{X}_{t+j}, W_{t+j-1}, Y_{t+j-1}] \in \mathcal{U} \subset \mathbb{R}^{(d_v+d_x+K+1)}$ . First, local features are extracted by encoding  $[\mathbf{V}, \mathbf{X}_t, W_{t-1}, Y_{t-1}]$  into  $\mathbf{Z}_t = \Phi_{\theta_1}([\mathbf{V}, \mathbf{X}_t, W_{t-1}, Y_{t-1}])$ . Then, the full process history  $\mathbf{H}_t$  is summarized into a *context representation*  $\mathbf{C}_t$ , given by an autoregressive model:  $\Phi_{\theta_2}^{ar}(\mathbf{Z}_{\leq t}) = \mathbf{C}_t$ , where  $\Phi_{\theta_2}^{ar}$  is implemented with a GRU [14]. This results in the representation function:  $\Phi_{\theta_1, \theta_2}(\mathbf{H}_t) = \mathbf{C}_t$ . To train the model, we use a contrastive loss that encourages the context  $\mathbf{C}_t$  to predict future local features  $\mathbf{Z}_{t+1}, \dots, \mathbf{Z}_{t+\tau}$  while distinguishing them from the features of other individuals. This is done by minimizing the InfoNCE loss  $\mathcal{L}_j^{(InfoNCE)}$  for each horizon  $j$ :

$$\mathcal{L}_j^{(InfoNCE)}(\theta_1, \theta_2, \Gamma_j) := -\mathbb{E}_{\mathcal{B}} \left[ \log \frac{\exp(T_j(\mathbf{U}_{t+j}, \mathbf{C}_t))}{\sum_{l=1}^{|\mathcal{B}|} \exp(T_j(\mathbf{U}_{l,t+j}, \mathbf{C}_t))} \right], \quad (1)$$

where  $\mathcal{B}$  is a batch containing individual histories, and  $\Gamma_j$  is a weight matrix. The discriminator  $T_j(\cdot, \cdot)$  classifies the correct future feature among negative samples from other individuals:

$$T_j(\mathbf{U}_{t+j}, \Phi_{\theta_1, \theta_2}(\mathbf{H}_t)) = \Phi_{\theta_1}(\mathbf{U}_{t+j})^T \Gamma_j \mathbf{C}_t = \mathbf{Z}_{t+j}^T \Gamma_j \mathbf{C}_t. \quad (2)$$

In practice,  $\mathbf{C}_t$  predicts  $\hat{\mathbf{Z}}_{t+j}$ , and prediction quality is measured by the dot product  $\mathbf{Z}_{t+j}^T \hat{\mathbf{Z}}_{t+j}$ . Minimizing the InfoNCE loss  $\mathcal{L}_j^{(InfoNCE)}$  provides a lower bound on the MI between the context and future features  $I(\mathbf{U}_{t+j}, \mathbf{C}_t)$  [52]:

$$I(\mathbf{U}_{t+j}, \mathbf{C}_t) \geq \log(|\mathcal{B}|) - \mathcal{L}_j^{(InfoNCE)}. \quad (3)$$

For multiple forecasting horizons  $j = 1, 2, \dots, \tau$ , we learn long-term dependencies by minimizing the InfoNCE loss across all horizons:

$$\mathcal{L}^{CPC}(\theta_1, \theta_2, \{\Gamma_j\}_{j=1}^{\tau}) := \frac{1}{\tau} \sum_{j=1}^{\tau} \mathcal{L}_j^{(InfoNCE)}(\theta_1, \theta_2, \Gamma_j). \quad (4)$$

Thus, minimizing  $\mathcal{L}^{CPC}$  maximizes the *shared information* between the context and future components as shown in Eq. (5), pushing the model to capture the global structure of the process over large horizons—crucial for counterfactual regression across multiple time steps:

$$\frac{1}{\tau} \sum_{j=1}^{\tau} I(\mathbf{U}_{t+j}, \mathbf{C}_t) \geq \log(|\mathcal{B}|) - \mathcal{L}^{CPC}. \quad (5)$$

**InfoMax Principle** We introduce a regularization term to make the context representation of the process history  $\mathbf{H}_t$  "reconstructable." We leverage the InfoMax principle to maximize the MI between  $\mathbf{H}_t$  and the context  $\mathbf{C}_t$ . However, we avoid computing the contrastive loss between  $\mathbf{C}_t$  and  $\mathbf{H}_t$  for two main reasons. First,  $\mathbf{H}_t$  is a high-dimensional sequence, making the loss computation very demanding. Secondly, we are still interested in incorporating inductive bias toward capturing global dependencies, this time by pushing any subsequence to be predictive of any future subsequence within  $\mathbf{H}_t$ . Hence, we divide the process history into two non-overlapping views,  $\mathbf{H}_t^f := \mathbf{U}_{1:t_0}$ ,  $\mathbf{H}_t^h := \mathbf{U}_{t_0+1:t}$  representing a *historical subsequence* and a *future subsequence* within the process history  $\mathbf{H}_t$ , with  $t_0$  randomly chosen per batch. We then maximize the MI between the representations of these views,  $\mathbf{C}_t^h$  and  $\mathbf{C}_t^f$ , resulting in a lower bound to the InfoMax objective as formulated below:

**Proposition 5.1.**  $I(\mathbf{C}_t^h, \mathbf{C}_t^f) \leq I(\mathbf{H}_t, (\mathbf{C}_t^h, \mathbf{C}_t^f))$ .

We provide an intuitive discussion of the inequality by providing an exact writing of the gap in 5.1:

**Theorem 5.2.**

$$I(\mathbf{H}_t; (\mathbf{C}_t^f, \mathbf{C}_t^h)) - I(\mathbf{C}_t^f, \mathbf{C}_t^h) = I(\mathbf{H}_t; \mathbf{C}_t^f | \mathbf{C}_t^h) + \mathbb{E}_{\mathbf{h}_t \sim \mathbb{P}_{\mathbf{H}_t}} \mathbb{E}_{\mathbf{c}_t^f \sim \mathbb{P}_{\mathbf{C}_t^f | \mathbf{h}_t}} \left[ D_{KL}[\mathbb{P}_{\mathbf{C}_t^h | \mathbf{h}_t} || \mathbb{P}_{\mathbf{C}_t^h | \mathbf{c}_t^f}] \right]. \quad (6)$$

Both terms on the RHS of Eq. (6) are positive, providing an alternative proof to Proposition 5.1. When equality holds, it implies  $I(\mathbf{H}_t; \mathbf{C}_t^f | \mathbf{C}_t^h) = 0$ , indicating that  $\mathbf{H}_t$  is independent of  $\mathbf{C}_t^f$  given  $\mathbf{C}_t^h$ . This suggests that  $\mathbf{C}_t^h$  retains sufficient information from  $\mathbf{H}_t$  that is predictive of  $\mathbf{C}_t^f$ . The symmetry of MI also leads to the occurrence of the second term on the RHS when conditioning on  $\mathbf{C}_t^f$ . The equality in Proposition 5.1 implies  $\mathbb{P}_{\mathbf{C}_t^h | \mathbf{h}_t} = \mathbb{P}_{\mathbf{C}_t^h | \mathbf{c}_t^f}$ , suggesting that  $\mathbf{C}_t^f$  efficiently encodes its subsequence while sharing maximum information with  $\mathbf{C}_t^h$ .

By considering the proposed variant of the InfoMax principle, we can compute a contrastive bound to  $I(\mathbf{C}_t^h, \mathbf{C}_t^f)$  more efficiently, as the random vectors reside in a low-dimensional space thanks to the encoding. We define a contrastive loss using InfoNCE similar to Eq. (1):

$$\mathcal{L}^{(InfoMax)}(\theta_1, \theta_2, \eta) := -\mathbb{E}_{\mathcal{B}} \left[ \log \frac{\exp(T_{\eta}(\mathbf{C}_t^f, \mathbf{C}_t^h))}{\sum_{i=1}^{|\mathcal{B}|} \exp(T_{\eta}(\mathbf{C}_{i,t}^f, \mathbf{C}_t^h))} \right]. \quad (7)$$

We use a non-linear discriminator  $T_{\eta}$  parametrized by  $\eta$  (detailed in Appendix I). The representation of the past subsequence  $\mathbf{C}_t^h$  is mapped to a prediction of the future subsequence  $\hat{\mathbf{C}}_t^f := F_{\eta}(\mathbf{C}_t^h)$  and  $T_{\eta} = \mathbf{C}_t^{f\top} \hat{\mathbf{C}}_t^f$ .

Theorem 5.2 and Proposition 5.1 justify using the loss in Eq. (7) by showing that our InfoMax simplification provides a valid lower bound. Thus, the contrastive loss in Eq. (7) is valid, as:

$$\log(|\mathcal{B}|) - \mathcal{L}^{(InfoMax)} \leq I(\mathbf{C}_t^h, \mathbf{C}_t^f) \leq I(\mathbf{H}_t, (\mathbf{C}_t^h, \mathbf{C}_t^f)). \quad (8)$$

The "mental model" behind our regularization term comes from the MI,  $I(\mathbf{H}_t, (\mathbf{C}_t^h, \mathbf{C}_t^f)) = H(\mathbf{H}_t) - H(\mathbf{H}_t | (\mathbf{C}_t^h, \mathbf{C}_t^f))$ , which can be written using entropy. Since the entropy term is constant and parameter-free, minimizing the conditional entropy  $H(\mathbf{H}_t | (\mathbf{C}_t^h, \mathbf{C}_t^f)) \geq 0$  ensures that  $\mathbf{H}_t$  is almost surely a function of  $(\mathbf{C}_t^h, \mathbf{C}_t^f)$  (Appendix G.4, Proposition G.2). When MI is maximized, the theoretical existence of such a function suggests that the learned context  $\mathbf{C}_t$  can decode and reconstruct  $\mathbf{H}_t$ .

Beyond the idea of reconstruction, it was shown that the InfoNCE objective implicitly learns to invert the data's generative model under mild assumptions [86]. Recent works [18, 44] extend this insight to multi-modal settings, which can reframe our InfoMax problem:  $\mathbf{H}_t^h$  and  $\mathbf{H}_t^f$  can be seen as two coupled modalities, allowing us to identify latent generative factors up to some mild indeterminacies (e.g. rotations, affine mappings). We plan to extend multi-modal causal representation learning to the longitudinal setting, where we anticipate minimizing our InfoMax objective, in the limit of infinite data, will effectively invert the data generation process up to a class of indeterminacies that we conjecture to be broader and under weaker assumptions than those in current causal representation learning literature, given our focus is on causal inference rather than the identification of causal latent variables. We initiate a formal basis for this claim in Appendix G.5.

## 5.2 Balanced representation learning

**Motivation** Our goal is counterfactual regression, specifically estimating  $\mathbb{E}(Y_{t+\tau}(\omega_{t+1:t+\tau}) | \mathbf{H}_{t+1})$ . For simplicity, with  $\tau = 1$ , we estimate the potential outcome for a given treatment  $W_{t+1} = \omega_{t+1}$ , where  $W_{t+1} \in \{0, 1, \dots, K-1\}$ , expressed as  $\mathbb{E}(Y_{t+1}(\omega_{t+1}) | \mathbf{H}_{t+1})$ , which under standard assumptions is identified as:

$$\mathbb{E}(Y_{t+1}(\omega_{t+1}) | \mathbf{H}_{t+1}) = \mathbb{E}(Y_{t+1} | \mathbf{H}_{t+1}, W_{t+1} = \omega_{t+1}).$$

The RHS can be estimated from data as  $\mathbb{E}(Y_{t+1} | \mathbf{H}_{t+1}, W_{t+1}) = f(\mathbf{H}_{t+1}, W_{t+1})$ . Since only one treatment is observed per individual at each time step,  $W_{i,t+1} = \omega_{i,t+1}$ , our model  $\hat{f}$  generates counterfactual responses by switching treatments  $\hat{f}(\mathbf{h}_{i,t+1}, \omega'_{t+1})$ , where  $\omega'_{t+1} \neq \omega_{t+1}$  (e.g., chemotherapy vs. radiotherapy). The challenge is that  $\mathbf{H}_{t+1}$  and  $W_{t+1}$  are not independent, introducing potential bias in counterfactual estimation [61], leading to covariate shift or selection bias. To address this, we learn a representation  $\Phi(\mathbf{H}_{t+1})$  that enforces distributional balance during decoding.

**Setup** To mitigate selection bias, we leverage the context representation  $\mathbf{C}_t$  of  $\mathbf{H}_t$  and introduce two sub-networks: one for response prediction and one for treatment prediction, both using a mapping of the context representation:

$$\Phi_t = \text{SELU}(\text{Linear}(\mathbf{C}_t)) = \Phi_{\theta_R}(\mathbf{H}_t),$$

where SELU represents the Scaled Exponential Linear Unit [37], and  $\theta_R$  denotes all parameters of the representation learner, i.e.,  $\theta_R = [\theta_1, \theta_2]$ . Following [7, 48], our objective is to learn a representation that accurately predicts outcomes while remaining *distributionally balanced* across all possible treatment choices  $W_t = 0, 1, \dots, K-1$ . To achieve this, we frame the problem as an adversarial game: one network learns to predict the next treatment from the representation, while a regularization term ensures that the representation is non-predictive of the treatment.

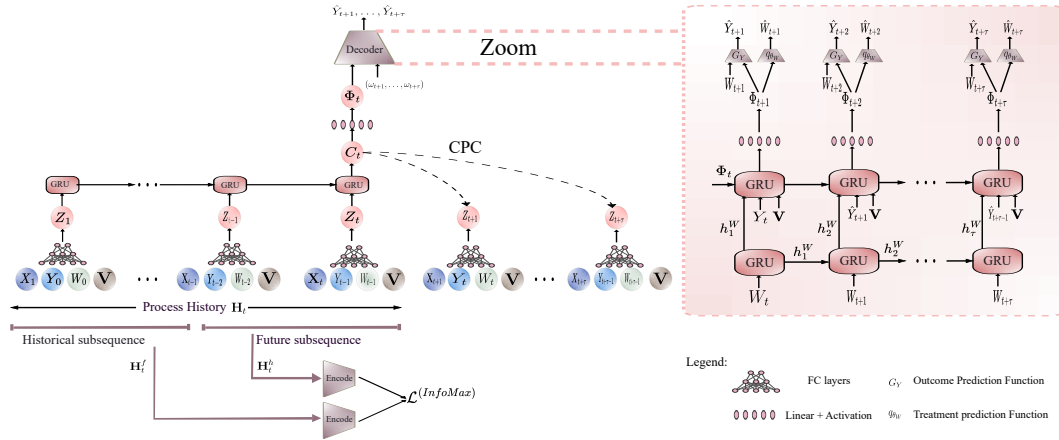


Figure 2: Causal CPC architecture: The left shows the encoder, which learns context  $C_t$  from process history  $H_t$ , with CPC and InfoMax objectives used for pretraining. The right shows the decoder, which autoregressively predicts the future outcome sequence from  $C_t$ .

**Factual response prediction** Since we intend to predict counterfactual responses for  $\tau$  steps ahead in time, we train a decoder to predict the factual responses  $Y_{t+1}, \dots, Y_{t+\tau}$  given the sequence of treatments  $(W_{t+1}, \dots, W_{t+\tau})$ . We minimize the negative conditional likelihood

$$\begin{aligned} \mathcal{L}_Y(\theta_R, \theta_Y) &= -\log p_{\theta_Y}(y_{t+1:t+\tau} | \Phi_t, \omega_{t+1:t+\tau}) \\ &= -\sum_{j=1}^{\tau} \log p_{\theta_Y}(y_{t+j} | y_{t+1:t+j-1}, \Phi_t, \omega_{t+1:t+j}). \end{aligned}$$

We denote  $\mathcal{I}_t^j := [Y_{t+1:t+j-1}, \Phi_t, W_{t+1:t+j}]$  and assume a Gaussian distribution for the conditional responses  $Y_{t+j} | \mathcal{I}_t^j \sim \mathcal{N}(G_Y(\mathcal{I}_t^j), \sigma^2)$ , where  $G_Y$  models the mean of the conditional response (see right side of Figure 2). We set  $\sigma = 0.05$  throughout our experiments. The response sequence is estimated autoregressively using a GRU-based decoder without teacher forcing [81] to ensure model training's consistency with testing in real-world scenarios (Figure 2 and Algorithm 2 in Appendix H).

**Treatment prediction** We learn a treatment prediction sub-network parameterized by  $\theta_W$  that takes as input the representation  $\Phi_{t+1}$  and predicts a distribution  $q_{\theta_W}(\omega_{t+1} | \Phi_{t+1})$  over the treatment  $W_{t+1}$  by minimizing the negative log-likelihood,  $\mathcal{L}_W = -\log q_{\theta_W}(\omega_{t+1} | \Phi_{t+1})$ . To assess the quality of the representation in predicting the treatment, the gradient from  $\mathcal{L}_W$  only updates the treatment network parameters  $\theta_W$  and is not backpropagated through the response of the parameters for the representation  $\Phi_{t+1}$  (Algorithm 2, Appendix H).

**Adversarial learning** To create an adversarial game, we update the representation learning parameters, and in the next step, the treatment network  $q_{\theta_W}(\cdot | \Phi_{t+1})$  with adverse losses such that the representation  $\Phi_{t+1}$  becomes invariant with respect to the assignment of  $W_{t+1}$ . Different from SOTA models (as highlighted in related work) and in line with our information guidelines principles, learning a balanced representation  $\Phi_{t+1}$  amounts to ensuring  $\Phi_{t+1} \perp\!\!\!\perp W_{t+1}$ , which is equivalent to  $I(\Phi_{t+1}, W_{t+1}) = 0$ . Hence, we minimize the MI as a way to confuse the treatment classifier. Specifically, we minimize an upper bound on  $I(\Phi_{t+1}, W_{t+1})$ , namely the CLUB of MI [13].

$$\begin{aligned} I_{\text{CLUB}}(\Phi(H_t), W_{t+1}; q_{\theta_W}) &:= \mathbb{E}_{\mathbb{P}_{(\Phi(H_t), W_{t+1})}} [\log q_{\theta_W}(W_{t+1} | \Phi(H_{t+1}))] \\ &\quad - \mathbb{E}_{\mathbb{P}_{\Phi(H_t)}} \mathbb{E}_{\mathbb{P}_{W_{t+1}}} [\log q_{\theta_W}(W_{t+1} | \Phi(H_{t+1}))]. \end{aligned} \quad (9)$$

We use the objective in Eq. (9) to update the representation learner  $\Phi(\cdot)$  [9, 32]. This update aims to minimize the discrepancy between the conditional likelihood of treatments for units sampled from  $\mathbb{P}_{(H_t, W_{t+1})}$  and the conditional likelihood of treatments under the assumption of independent sampling from the product of marginals  $\mathbb{P}_{H_{t+1}} \otimes \mathbb{P}_{W_{t+1}}$ . In practice, we generate samples from the product of marginals by shuffling the treatment  $W_{t+1}$  across the batch dimension similar to [9, 32]. When minimizing  $\mathcal{L}_W$ ,  $q_{\theta_W}(\omega_{t+1} | \Phi_{t+1})$  gets closer to the true conditional distribution  $p(\omega_{t+1} | \Phi_{t+1})$ , and, in this case, the objective in Eq. (9) provides an upper bound of the MI between representation and treatment. We formalize the intuition by adapting the result of [13]:

**Theorem 5.3.** [13] Let  $q_{\theta_W}(\Phi_{t+1}, \omega_{t+1}) := q_{\theta_W}(\omega_{t+1} | \Phi_{t+1})p(\Phi_{t+1})$  be the joint distribution induced by  $q_{\theta_W}(\omega_{t+1} | \Phi_{t+1})$  over the representation space of  $\Phi_{t+1}$ . If:

$$D_{KL}(p(\Phi_{t+1}, \omega_{t+1}) || q_{\theta_W}(\Phi_{t+1}, \omega_{t+1})) \leq D_{KL}(p(\Phi_{t+1})p(\omega_{t+1}) || q_{\theta_W}(\Phi_{t+1}, \omega_{t+1})),$$

then  $I(\Phi_{t+1}, W_{t+1}) \leq I_{CLUB}(\Phi_{t+1}, W_{t+1}; q)$ .

Based on Theorem 5.3, our adversarial training is interpretable and can be explained as follows: the treatment classifier seeks to minimize  $\mathbb{E}_{\mathbb{P}_{(\mathbf{H}_t, W_{t+1})}} [\mathcal{L}_W]$ , which is equivalent to minimizing Kullback-Leibler divergence  $D_{KL}(p(\Phi_{t+1}, \omega_{t+1}) || q_{\theta_W}(\Phi_{t+1}, \omega_{t+1}))$ . Therefore,  $q_{\theta_W}(\Phi_{t+1}, \omega_{t+1})$  could get closer to  $p(\Phi_{t+1}, \omega_{t+1})$  than, ultimately, to  $p(\Phi_{t+1})p(\omega_{t+1})$ , as we train the network to predict  $W_{t+1}$  from  $\Phi_{t+1}$ . In such a case and by Theorem 5.3,  $I_{CLUB}$  provides an upper bound on the MI. Hence, in a subsequent step, we minimize  $I_{CLUB}$  w.r.t the representation parameters, minimizing the MI  $I(\Phi_{t+1}, W_{t+1})$  and achieving balance. We theoretically formulate such behavior by proving in the following theorem that, at the Nash equilibrium of this adversarial game, the representation is exactly balanced across the different treatment regimes provided by  $W_{t+1}$ .

**Theorem 5.4.** Let  $t \in \{1, 2, \dots, t_{max}\}$ ,  $\Phi = \Phi_{\theta_R}$  and  $q = q_{\theta_W}$  are, respectively, any representation and treatment network. Let  $\mathbb{P}_{\Phi(\mathbf{H}_t)}$  be the probability distribution over the representation space and  $\mathbb{P}_{\Phi(\mathbf{H}_t) | W_{t+1}}$  its conditional counterpart. Then, there exist  $\Phi^*$  and  $q^*$  such that:

$$\begin{aligned} \Phi^* &= \arg \min_{\Phi} I_{CLUB}(\Phi(\mathbf{H}_t), W_{t+1}; q^*) \\ q^* &= \arg \max_q \mathbb{E}_{\mathbb{P}_{\Phi^*}(\mathbf{H}_t)} [\log q(W_{t+1} | \Phi^*(\mathbf{H}_t))]. \end{aligned} \quad (10)$$

Such an equilibrium holds if and only if  $\mathbb{P}_{\Phi(\mathbf{H}_t) | W_{t+1}=0} = \mathbb{P}_{\Phi(\mathbf{H}_t) | W_{t+1}=1} = \dots = \mathbb{P}_{\Phi(\mathbf{H}_t) | W_{t+1}=k-1}$ , almost surely.

Since we target multi-timestep forecasting, covariate balancing in the representation space extends beyond  $t+1$ . For simplicity, we presented it for  $t+1$ , but in practice, the adversarial game applies the balancing across all forecasting horizons (Algorithm 2). The theorem also holds for other horizons by replacing  $\Phi(\mathbf{H}_t)$  with  $\Phi(\mathbf{H}_{t+j-1})$  and  $W_{t+1}$  with  $W_{t+j}$ , for  $2 \leq j \leq \tau$ .

**Causal CPC Training** The Causal CPC model is trained in two stages: (1) **Encoder pretraining:** We first learn an efficient representation of the process history by minimizing loss:

$$\mathcal{L}_{enc} = \mathcal{L}^{CPC}(\theta_1, \theta_2, \{\Gamma_j\}_{j=1}^{\tau}) + \mathcal{L}^{(InfoMax)}(\theta_1, \theta_2, \gamma).$$

(2) **Decoder training:** After pretraining, we fine-tune the encoder by optimizing the factual outcome and treatment networks in the adversarial game from Theorem 5.4. Formally:

$$\begin{aligned} \min_{\theta_R, \theta_Y} \mathcal{L}_{dec}(\theta_R, \theta_Y, \theta_W) &= \mathcal{L}_Y(\theta_R, \theta_Y) + I_{CLUB}(\Phi_{\theta_R}(\mathbf{H}_t), W_{t+1}; q_{\theta_W}), \\ \min_{\theta_W} \mathcal{L}_W(\theta_W, \theta_R) &= -\mathbb{E}_{\Phi_{\theta_R}(\mathbf{H}_t)} [\log q_{\theta_W}(W_{t+1} | \Phi_{\theta_R}(\mathbf{H}_t))]. \end{aligned}$$

## 6 Experiments

We compare Causal CPC with SOTA baselines: MSMs [64], RMSN [41], CRN [7], G-Net [39], and CT [48]. All models are fine-tuned via a grid search over hyperparameters, including architecture and optimizers. Model selection is based on mean squared error (MSE) on factual outcomes from a validation set, and the same criterion is used for early stopping. Further details on hyperparameters and training procedures are provided in Appendices J and D.

### 6.1 Experiments with Synthetic Data

**Tumor Growth** We use the Pharmacokinetic-Pharmacodynamic (PK-PD) model [24] to simulate responses of non-small cell lung cancer patients, following previous works [41, 7, 48]. We evaluate our approach on simulated counterfactual trajectories, varying the confounding level via the parameter  $\gamma$  (Appendix E.1). Unlike [48], who used larger datasets (10,000 for training, 1,000 for testing), we use a smaller, more challenging dataset (1,000 for training, 500 for testing) to reflect real-world data limitations. For long-horizon forecasting, we set the prediction horizon to 10 and evaluate two training sequence lengths: 60 and 40, with covariates of dimension 4.

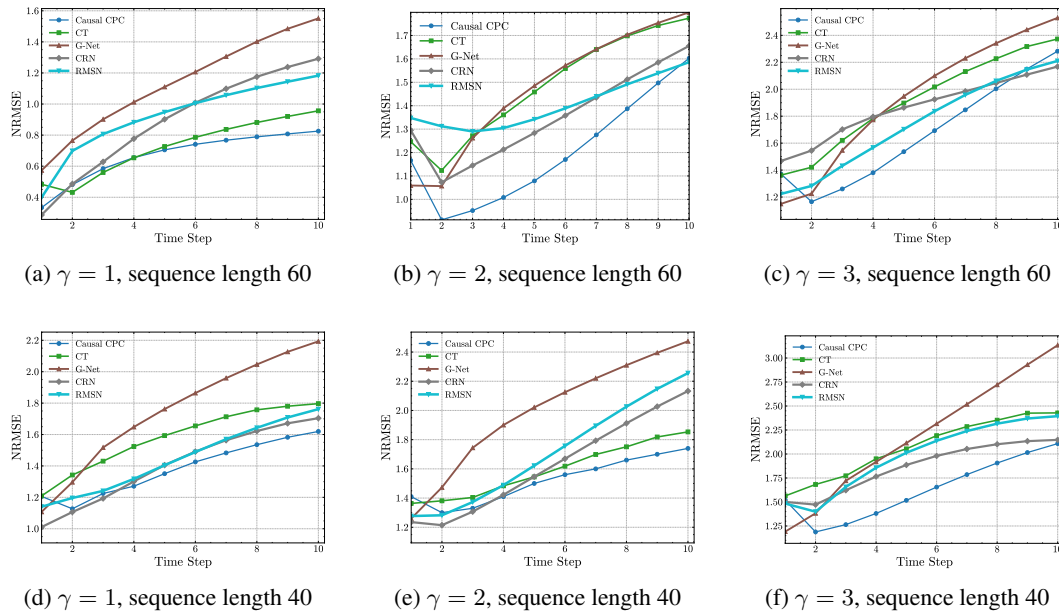


Figure 3: Evolution of error (NRMSE) in estimating counterfactual responses for cancer simulation data. Top: training sequence length 60. Bottom: training sequence length 40. In both cases,  $\tau = 10$ . MSM is excluded due to high prediction errors.

**Results** We tested all models on the cancer simulation data across three confounding levels,  $\gamma = 1, 2, 3$ . Figures 3a, 3b and 3c show the evolution of Normalized Root Mean Squared Error (NRMSE) over counterfactual tumor volume as the prediction horizon increases. Causal CPC consistently outperforms all baselines at larger horizons, demonstrating its effectiveness in long-term predictions. This confirms the quality of  $C_t$  in predicting future components across multiple time steps, capturing the global structure of the process as discussed in Eq. (5). Extended results are provided in Appendix E.2.1. In the more challenging case where the maximum sequence length is 40 (Figures 3d, 3e and 3f), the error evolution remains similar to Figure 3a, 3b and 3c. Our model maintains its advantage, outperforming most baselines in long-term forecasting. However, Causal CPC does not outperform other models in short-term forecasting, a consistent limitation across experiments. Still, the model's superior long-term performance highlights its potential in applications requiring long-term accuracy. At higher levels of confounding, the model does not always outperform SOTA models at certain time steps. This may be due to the low dimensionality of the time-varying components and static covariates,  $\mathbf{U}_t = [\mathbf{V}, \mathbf{X}_t, \mathbf{W}_{t-1}, \mathbf{Y}_{t-1}]$ , which have only four dimensions. Our model leverages contrastive learning-based regularization to excel on datasets with higher confounding dimensions, as demonstrated on MIMIC-III where  $\mathbf{U}_t$  has 72 dimensions. In this setting, our model consistently outperforms baselines at longer prediction horizons. The occasional underperformance of Causal CPC at the final horizon is due to  $\tau = 10$  being the last contrasted horizon, not an issue specific to  $\tau = 10$ . To support this, we reran all models with a sequence length of 60,  $\tau = 15$ , and  $\gamma = 2$ . As shown in Figure 4, Causal CPC still outperforms SOTA for horizons beyond  $\tau = 10$  due to the encoder's retraining, where the InfoNCE loss is computed over all 15 time steps. The last prediction error remains close to SOTA, suggesting that training over larger horizons than initially intended may be beneficial.

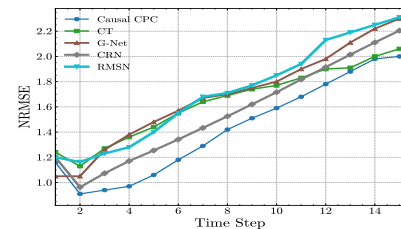


Figure 4: Models' performance for cancer simulation,  $\gamma = 2$ ,  $\tau = 15$ .

## 6.2 Experiments with semi-synthetic and real data

**Semi-synthetic MIMIC-III** We used a semi-synthetic dataset constructed by [48] based on the MIMIC-III dataset [35], incorporating both endogenous temporal dependencies and exogenous dependencies from observational patient trajectories, as detailed in Appendix F.1. The

Table 1: Evolution of RMSEs for the semi-synthetic MIMIC III, sequence length 100.

| Model                    | $\tau = 1$       | $\tau = 2$       | $\tau = 3$       | $\tau = 4$        | $\tau = 5$       | $\tau = 6$       | $\tau = 7$       | $\tau = 8$       | $\tau = 9$       | $\tau = 10$      |
|--------------------------|------------------|------------------|------------------|-------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| <b>Causal CPC (ours)</b> | 0.32±0.04        | 0.45±0.08        | 0.54±0.06        | 0.61±0.10         | <b>0.66±0.10</b> | <b>0.69±0.11</b> | <b>0.71±0.11</b> | <b>0.73±0.06</b> | <b>0.75±0.05</b> | <b>0.77±0.10</b> |
| <b>CT</b>                | 0.42±0.38        | <b>0.40±0.06</b> | <b>0.52±0.08</b> | <b>0.60±0.005</b> | 0.67±0.10        | 0.72±0.12        | 0.77±0.13        | 0.81±0.14        | 0.85±0.16        | 0.88±0.17        |
| <b>G-Net</b>             | 0.54±0.13        | 0.72±0.14        | 0.85±0.16        | 0.96±0.17         | 1.05±0.18        | 1.14±0.18        | 1.24±0.17        | 1.33±0.16        | 1.41±0.16        | 1.49±0.16        |
| <b>CRN</b>               | <b>0.27±0.03</b> | 0.45±0.08        | 0.58±0.09        | 0.72±0.11         | 0.82±0.15        | 0.92±0.20        | 1.00±0.25        | 1.06±0.28        | 1.12±0.32        | 1.17±0.35        |
| <b>RMSN</b>              | 0.40±0.16        | 0.70±0.21        | 0.80±0.19        | 0.88±0.17         | 0.94±0.16        | 1.00±0.15        | 1.05±0.14        | 1.10±0.14        | 1.14±0.13        | 1.18±0.13        |

patient trajectories are high-dimensional and exhibit long-range dependencies. Similar to the cancer simulation, the training data consisted of relatively few sequences (500 for training, 100 for validation, and 400 for testing). Table 1 presents the mean and standard deviation of counterfactual predictions across multiple horizons ( $\tau = 10$ ). We test two maximum sequence lengths, 100 and 60, to assess the models' robustness for long-horizon forecasting.

**Results** Causal CPC consistently outperformed the baselines, especially at larger horizons, both with a sequence length of 100 and a reduced length of 60 (Figures 1, 5). Its superior performance at longer horizons is likely due to the high number of covariates, making it well-suited to contrastive-based training. We also tested the models with 800/200/200 individuals for training/validation/testing, as in [48] (Appendix F.2.2), where Causal CPC achieved state-of-the-art (SOTA) results comparable to CT but with much shorter training and prediction times.

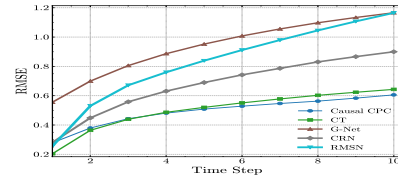


Figure 5: Performance for MIMIC III semi-synthetic, sequence length 60.

Table 2: Models complexity and the running time averaged over five seeds. Results are reported for tumor growth simulation ( $\gamma = 1$ ). Hardware: GPU-1xNVIDIA Tesla M60.

| Model                                 | Trainable parameters (k) | Training time (min) | Prediction time (min) |
|---------------------------------------|--------------------------|---------------------|-----------------------|
| <b>Causal CPC (encoder + decoder)</b> | 8.2                      | 16±3                | 4±1                   |
| <b>CT</b>                             | 11                       | 12±2                | 30±3                  |
| <b>G-Net</b>                          | 1.2                      | 2±0.5               | 35±3                  |
| <b>CRN</b>                            | 5.2                      | 13±2                | 4±1                   |
| <b>RMSN</b>                           | 1.6                      | 22±2                | 4±1                   |
| <b>MSM</b>                            | <0.1                     | 1±0.5               | 1±0.5                 |

Table 3: Ablation study with NRMSE averaged across ( $1 \leq \tau \leq 10$ ) for cancer simulation ( $\gamma = 1$ ) and MIMIC III.

| Model                                       | Cancer_Sim  | MIMIC III   |
|---------------------------------------------|-------------|-------------|
| Causal CPC (Full)                           | <b>1.05</b> | <b>0.62</b> |
| Causal CPC (w/o $\mathcal{L}^{(InfoNCE)}$ ) | 1.07        | 0.68        |
| Causal CPC (w/o $\mathcal{L}^{(InfoMax)}$ ) | 1.13        | 0.74        |
| Causal CPC (w CDC loss)                     | 1.07        | 0.73        |
| Causal CPC (w/o balancing)                  | 1.08        | 0.69        |

**Computational Efficiency and Model Complexity** Efficient execution is crucial for practical deployment, especially with periodic retraining. Beyond training, challenges arise in evaluating multiple counterfactual trajectories per individual, which grow exponentially with the forecasting horizon as  $K^\tau$ , where  $K$  is the number of possible treatments. This is particularly relevant when generating multiple treatment plans, such as minimizing tumor volume. Table 2 shows the models' complexity (number of parameters) and running time, split between model fitting and prediction. Causal CPC is highly efficient during prediction due to its simple 1-layer GRU, similar to CRN (1-layer LSTM), while providing better ITEs estimation. In contrast, CT is less efficient due to its transformer architecture and teacher forcing, which requires recursive data loading during inference. G-Net also has longer prediction times due to Monte Carlo sampling. Overall, Causal CPC strikes a strong balance between accuracy and efficiency, making it well-suited under constrained resources.

**Real MIMIC-III Data** We evaluated our model on real MIMIC-III data, where counterfactual trajectories cannot be assessed due to the absence of observed counterfactual responses. However, performance can still be measured by forecasting factual (observed) responses over time. Our model estimates responses for each individual based on their observed treatment trajectory. As shown in Figure 6, Causal CPC consistently outperforms all baselines, especially at larger horizons, demonstrating its robustness and effectiveness in real-world settings.

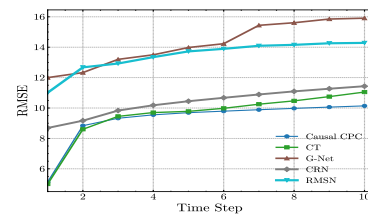


Figure 6: Evolution of RMSEs, Real MIMIC III, sequence length 100.

## 7 Discussion

**Why does Causal CPC outperform SOTA at large horizons?** Our context  $C_t$  is designed to capture shared information across future representations, particularly covariates, by minimizing the

InfoNCE loss over multiple time steps (Eq. 1). As shown in Eq. (5), minimizing  $\mathcal{L}^{CPC}$  maximizes shared information between the context and future components, helping capture the *global structure* of the process. This is especially beneficial for counterfactual regression over long horizons, explaining the model’s superior performance. However, it may not always outperform SOTA in shorter-term predictions due to its focus on long-term dependencies.

**Short-term Counterfactual Regression** While our model is designed for long-term predictions, it may not consistently outperform SOTA for short-horizon tasks. However, the use of contrastive loss, particularly InfoNCE (Eq. 4), suggests potential adaptability to balance both short- and long-term predictions without retraining. A trade-off could be achieved by adjusting the contrastive term weights across time steps in Eq. (4), which we leave for future work.

**Ablation study** We examined the model’s performance in various configurations—full model, without CPC, and without InfoMax. Table 3 and Figure 7 show that removing either term reduces counterfactual accuracy across all horizons, underscoring their significance. Additionally, replacing our ICLUB objective with CDC loss [48] or removing balancing increases errors. We also tested different MI lower bounds like NWJ [50] and MINE [4] for both CPC and InfoMax (Appendix C.2), finding that InfoNCE yielded the best results (Table 10). Full ablation results are in Appendices E.2.2 and F.2.1.

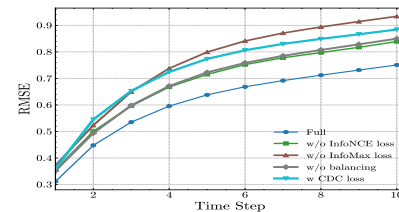


Figure 7: Ablation study of Causal CPC on MIMIC III.

**Falsifiability Test** This study assumes sequential ignorability, common in causal inference [41, 7, 39, 48]. To assess robustness, we performed a falsifiability test by omitting certain confounders during training, while they remained in MIMIC-III data construction. As seen in Table 4, violating sequential ignorability increased prediction errors for Causal CPC, CT, and CRN, though RMSN was less affected but underperformed at  $\tau \geq 2$ . Despite this, Causal CPC maintained its lead at larger horizons, demonstrating strong encoding of long-term dependencies.

Table 4: Results on the MIMIC III when sequential ignorability is violated reported by RMSEs

| Model              | $\tau = 1$       | $\tau = 2$       | $\tau = 3$       | $\tau = 4$       | $\tau = 5$       | $\tau = 6$       | $\tau = 7$       | $\tau = 8$       | $\tau = 9$       | $\tau = 10$      |
|--------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| Causal CPC         | 0.44±0.04        | 0.56±0.07        | 0.66±0.07        | 0.73±0.08        | 0.78±0.08        | 0.83±0.06        | <b>0.86±0.10</b> | <b>0.88±0.08</b> | <b>0.91±0.08</b> | <b>0.95±0.07</b> |
| Causal Transformer | <b>0.34±0.07</b> | <b>0.48±0.07</b> | <b>0.60±0.07</b> | <b>0.68±0.06</b> | <b>0.75±0.06</b> | <b>0.80±0.07</b> | <b>0.86±0.09</b> | 0.91±0.11        | 0.95±0.13        | 1.00±0.15        |
| CRN                | 0.40±0.07        | 0.54±0.09        | 0.70±0.09        | 0.84±0.09        | 0.97±0.09        | 1.08±0.13        | 1.18±0.16        | 1.26±0.19        | 1.33±0.21        | 1.39±0.23        |
| RMSN               | 0.38±0.08        | 0.67±0.21        | 0.78±0.16        | 0.84±0.14        | 0.91±0.14        | 0.98±0.15        | 1.04±0.16        | 1.09±0.18        | 1.15±0.19        | 1.20±0.23        |

**Tightness of MI Upper Bounds** Estimating MI bounds for high-dimensional variables is challenging and expensive [59, 57], often limited to low-dimensional inputs or Gaussian assumptions. In MI-constrained models, batch size is crucial. As shown in Eq. (3), increasing the batch size  $\mathcal{B}$  tightens the lower bound via  $\log(|\mathcal{B}|)$ , and to balance memory and performance, we chose batch sizes of 256 for the encoder and 128 for the decoder. While these bounds may not be perfectly tight, mutual information and self-supervision biases significantly enhance performance (Appendix C.2), as confirmed by ablation studies. Other MI estimators like NWJ and MINE [50, 4] did not improve performance; our initial setup consistently performed better (Appendix F.2.1).

**Extending Causal CPC to Continuous Treatment** Our approach could be extended to continuous treatments by replacing the treatment classifier with a regressor. Since the method maximizes likelihood, the equilibrium in Theorem 5.4 remains valid. However, in practice, continuous treatments will be represented by a single dimension, unlike discrete treatments with  $K$ -dimensional one-hot encoding. This risks losing important treatment information in counterfactual predictions. A simpler adaptation to our model could involve discretizing continuous treatments.

**Conclusion** We introduced a novel approach to long-term counterfactual regression, combining RNNs with CPC to achieve SOTA results without relying on complex transformer models. Prioritizing computational efficiency, we incorporated contrastive loss-based regularization guided by mutual information (MI). Our method consistently outperforms existing models on both synthetic and real-world datasets, marking the first application of CPC in causal inference. Future work could focus on improving interpretability by integrating Shapley values into the causal framework [30]. Additionally, developing uncertainty-aware models tailored for longitudinal data is crucial for enhancing the reliability of predictions in our causal framework [19, 33, 84].

## References

- [1] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.
- [2] O. Atan, J. Jordon, and M. Van der Schaar. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [3] P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.
- [4] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.
- [5] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.
- [6] J. Berrevoets, A. Curth, I. Bica, E. McKinney, and M. van der Schaar. Disentangled counterfactual recurrent networks for treatment effect inference over time. *arXiv preprint arXiv:2112.03811*, 2021.
- [7] I. Bica, A. M. Alaa, J. Jordon, and M. van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083*, 2020.
- [8] I. Bica, A. M. Alaa, and M. van der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International Conference on Machine Learning*, 2020.
- [9] P. Brakel and Y. Bengio. Learning independent features with adversarial nets for non-linear ica. *arXiv preprint arXiv:1710.05050*, 2017.
- [10] D. Cao, J. Enouen, and Y. Liu. Estimating treatment effects in continuous time with hidden confounders. *arXiv e-prints*, pages arXiv–2302, 2023.
- [11] Y. Chen, A. Prati, J. Montgomery, and R. Garnett. A multi-task gaussian process model for inferring time-varying treatment effects in panel data. In *International Conference on Artificial Intelligence and Statistics*, pages 4068–4088. PMLR, 2023.
- [12] L. Cheng, R. Guo, and H. Liu. Causal mediation analysis with hidden confounders. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, null:null, 2021.
- [13] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020.
- [14] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [15] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.
- [16] Z. Chu, S. L. Rathbun, and S. Li. Learning infomax and domain-independent representations for causal effect inference with real-world data. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 433–441. SIAM, 2022.
- [17] T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.
- [18] I. Daunhawer, A. Bizeul, E. Palumbo, A. Marx, and J. E. Vogt. Identifiability results for multimodal contrastive learning. *arXiv preprint arXiv:2303.09166*, 2023.

- [19] E. De Brouwer, J. Gonzalez, and S. Hyland. Predicting the impact of treatments over time with uncertainty aware neural differential equations. In *International Conference on Artificial Intelligence and Statistics*, pages 4705–4722. PMLR, 2022.
- [20] M. D. Donsker and S. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212, 1983.
- [21] W. Falcon and The PyTorch Lightning team. PyTorch Lightning, Mar. 2019.
- [22] D. Frauen, T. Hatt, V. Melnychuk, and S. Feuerriegel. Estimating average causal effects from patient trajectories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7586–7594, 2023.
- [23] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.
- [24] C. Geng, H. Paganetti, and C. 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):1–12, 2017.
- [25] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.
- [26] M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of machine learning research*, 13(2), 2012.
- [27] J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.
- [28] T. Hatt and S. Feuerriegel. Sequential deconfounding for causal inference with unobserved confounders. *arXiv preprint arXiv:2104.09323*, 2021.
- [29] O. Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020.
- [30] T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789, 2020.
- [31] Ç. Hızlı, S. John, A. T. Juuti, T. T. Saarinen, K. H. Pietiläinen, and P. Marttinen. Causal modeling of policy interventions from treatment-outcome sequences. In *International Conference on Machine Learning*, pages 13050–13084. PMLR, 2023.
- [32] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.
- [33] A. Jesson, S. Mindermann, U. Shalit, and Y. Gal. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33:11637–11649, 2020.
- [34] S. Jiang, Z. Huang, X. Luo, and Y. Sun. Cf-gode: Continuous-time causal inference for multi-agent dynamical systems. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD ’23, page 997–1009, New York, NY, USA, 2023. Association for Computing Machinery.
- [35] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

- [36] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [37] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.
- [38] M. Kuzmanovic, T. Hatt, and S. Feuerriegel. Deconfounding temporal autoencoder: estimating treatment effects over time using noisy proxies. In *Machine Learning for Health*, pages 143–155. PMLR, 2021.
- [39] R. Li, S. Hu, M. Lu, Y. Utsumi, P. Chakraborty, D. M. Sow, P. Madan, J. Li, M. F. Ghalwash, Z. Shahn, and L. wei H. Lehman. G-net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In *ML4H@NeurIPS*, 2021.
- [40] P. P. Liang, Z. Deng, M. Ma, J. Zou, L.-P. Morency, and R. Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, 2023.
- [41] B. Lim. Forecasting treatment responses over time using recurrent marginal structural networks. *advances in neural information processing systems*, 31, 2018.
- [42] B. Lim. Forecasting treatment responses over time using recurrent marginal structural networks. *advances in neural information processing systems*, 31, 2018.
- [43] R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.
- [44] Y. Liu, Z. Zhang, D. Gong, B. Huang, M. Gong, A. v. d. Hengel, K. Zhang, and J. Q. Shi. Revealing multimodal contrastive representation learning through latent partial causal models. *arXiv preprint arXiv:2402.06223*, 2024.
- [45] M. J. Lopez and R. Gutman. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, pages 432–454, 2017.
- [46] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [47] S. Matthes, Z. Han, and H. Shen. Towards a unified framework of contrastive learning for disentangled representations. *Advances in Neural Information Processing Systems*, 36:67459–67470, 2023.
- [48] V. Melnychuk, D. Frauen, and S. Feuerriegel. Causal transformer for estimating counterfactual outcomes. *ArXiv*, abs/2204.07258, 2022.
- [49] S. Mueller and J. Pearl. Personalized decision making—a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050, 2023.
- [50] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.
- [51] S. Nowozin, B. Cseke, and R. Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.
- [52] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [53] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems*, 2019.
- [54] T. Pham, T. Tran, D. Phung, and S. Venkatesh. Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of biomedical informatics*, 69:218–229, 2017.

- [55] R. W. Platt, E. F. Schisterman, and S. R. Cole. Time-modified confounding. *American journal of epidemiology*, 170(6):687–694, 2009.
- [56] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.
- [57] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.
- [58] Z. Qian, Y. Zhang, I. Bica, A. Wood, and M. van der Schaar. Synctwin: Treatment effect estimation with longitudinal outcomes. *Advances in Neural Information Processing Systems*, 34:3178–3190, 2021.
- [59] T. Rainforth, A. Kosiorek, T. A. Le, C. Maddison, M. Igl, F. Wood, and Y. W. Teh. Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, pages 4277–4285. PMLR, 2018.
- [60] R. Ranganath and A. Perotte. Multiple causal inference with latent confounding. *arXiv preprint arXiv:1805.08273*, 2018.
- [61] J. M. Robins. Association, causation, and marginal structural models. *Synthese*, 121(1/2):151–179, 1999.
- [62] J. M. Robins and M. A. Hernán. Estimation of the causal effects of time-varying exposures. *Longitudinal data analysis*, 553:599, 2009.
- [63] J. M. Robins and M. A. Hernán. Estimation of the causal effects of time-varying exposures. *Longitudinal data analysis*, 553:599, 2009.
- [64] J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology, 2000.
- [65] D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.
- [66] N. Saunshi, J. Ash, S. Goel, D. Misra, C. Zhang, S. Arora, S. Kakade, and A. Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR, 2022.
- [67] E. F. Schisterman, S. R. Cole, and R. W. Platt. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass.)*, 20(4):488, 2009.
- [68] P. Schulam and S. Saria. Reliable decision support using counterfactual models. *Advances in neural information processing systems*, 30, 2017.
- [69] N. Seedat, F. Imrie, A. Bellot, Z. Qian, and M. van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. *arXiv preprint arXiv:2206.08311*, 2022.
- [70] U. Shalit. Can we learn individual-level treatment policies from clinical data? *Biostatistics*, 21(2):359–362, 2020.
- [71] R. Shwartz Ziv and Y. LeCun. To compress or not to compress—self-supervised learning and information theory: A review. *Entropy*, 26(3):252, 2024.
- [72] H. Soleimani, J. Hensman, and S. Saria. Scalable joint models for reliable uncertainty-aware event prediction. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1948–1963, 2017.
- [73] H. Soleimani, A. Subbaswamy, and S. Saria. Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. *ArXiv*, abs/1704.02038, 2017.

- [74] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2012.
- [75] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.
- [76] C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.
- [77] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.
- [78] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*, pages 4068–4076, 2015.
- [79] V. Veitch, D. Sridhar, and D. Blei. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR, 2020.
- [80] Y. Wang and D. M. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.
- [81] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.
- [82] S. Wu, W. Zhou, M. Chen, and S. Zhu. Counterfactual generative models for time-varying treatments. *arXiv preprint arXiv:2305.15742*, 2023.
- [83] Y. Yazıcı, C.-S. Foo, S. Winkler, K.-H. Yap, G. Piliouras, and V. Chandrasekhar. The unusual effectiveness of averaging in GAN training. In *International Conference on Learning Representations*, 2019.
- [84] M. Yin, C. Shi, Y. Wang, and D. M. Blei. Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*, 119(545):122–135, 2024.
- [85] N. Zhao, Z. Wu, R. W. Lau, and S. Lin. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*, 2020.
- [86] R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.

## A Impact Statements

Our paper seeks to advance the field of Trustworthy Machine Learning by focusing on the accurate estimation of counterfactual trajectories. This capability holds significant potential to enhance decision-making processes across various domains, particularly in healthcare, where clinicians can leverage models designed to mitigate bias and promote fairness. Additionally, by focusing on efficiency, our contributions extend beyond traditional machine learning considerations to address environmental concerns associated with energy consumption. By advocating for the prudent use of computational resources, especially in training complex models deployed in real-world scenarios, we aim to promote sustainability in developing and applying machine learning solutions.

## B Causal assumptions

### B.1 Identifiability Assumptions in Causal CPC

In this section, we detail the assumptions used for the identifiability of the counterfactual responses  $\mathbb{E}(Y_{t+\tau}(\omega_{t+1:t+\tau}) \mid \mathbf{H}_{t+1})$ . As briefly stated in Section 3, we follow similar assumptions to [62, 64, 7, 48], namely

**Assumption B.1** (Consistency). For every time step  $t$  and given any manner by which a unit  $i$  receives the sequence of treatment  $\omega_{i,\leq t}$ , we always observe the potential outcome  $Y_{it}(\omega_{i,\leq t})$ . Formally:

$$W_{i,\leq t} = w_{i,\leq t} \implies Y_{it} = Y_{it}(w_{i,\leq t}).$$

**Assumption B.2** (Sequential Ignorability). Given any time step  $t$ , we have the conditional independence:

$$Y_{it}(\omega_{it}) \perp\!\!\!\perp W_{it} \mid \mathbf{H}_{it} = \mathbf{h}_{it} \quad \forall (\omega_{it}, \mathbf{h}_{it})$$

**Assumption B.3** (Overlap/positivity). Given any time step  $t$ , and for any possible historical context  $\mathbf{h}_t$ , the probability of observing any of the possible treatment regimes is strictly positive but not deterministic:

$$p(\mathbf{h}_t) \neq 0 \implies 0 < p(W_t = \omega_t \mid \mathbf{h}_t) < 1$$

The three assumptions are sufficient for the identification of the counterfactual responses from observational data, which we formulate in the following proposition.

**Proposition B.4.** Assuming consistency, overlap, and ignorability (assumptions B.1, B.2, B.3), the causal quantity  $\mathbb{E}(Y_{t+\tau}(\omega_{t+1:t+\tau}) \mid \mathbf{H}_{t+1})$  is identifiable from observational data following

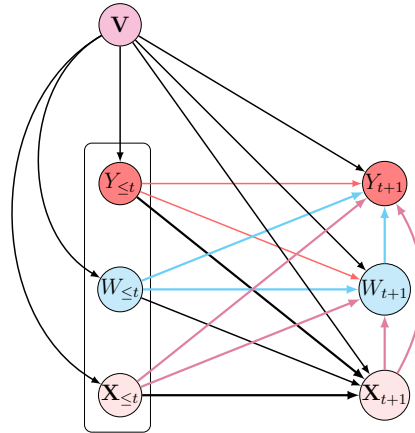
$$\mathbb{E}(Y_{t+\tau}(\omega_{t+1:t+\tau}) \mid \mathbf{H}_{t+1}) = \mathbb{E}(Y_{t+\tau} \mid \mathbf{H}_{t+1}, W_{t+1:t+\tau} = \omega_{t+1:t+\tau})$$

*Proof.* See [62] □

### B.2 On the Causal Graph

We repeat the causal graph introduced in Figure 1 to explain the data generation process. here, all of the past observed data encompassed in  $\mathbf{H}_{t+1}$  confounds future treatments and responses,  $W_{t+1}, W_{t+2}, \dots, W_{t_{max}}$  and  $Y_{t+1}, Y_{t+2}, \dots, Y_{t_{max}}$ , which create long-term dependencies. The fact that post-covariates are affected by past treatments creates time-dependent confounding. The static covariates are assumed to be affecting all of the time-varying variables. Since we suppose sequential ignorability, there are no possible exogenous noises affecting both treatments and responses. However, such noise may possibly affect responses, time-varying covariates, and response variables.

In the figure, for simplicity, we represent past treatments as  $W_{\leq t}$  such that each element in that sub-sequence confounds the next treatment and response  $W_{t+1}$  and  $Y_{t+1}$ . Idem for  $Y_{\leq t}$  and  $\mathbf{X}_{\leq t}$ . The static covariates  $\mathbf{V}$  are assumed to be affecting all the time-varying variables. We omit the representation of exogenous noise for simplicity. Interactions between  $W_{\leq t}$ ,  $\mathbf{X}_{\leq t}$ , and  $Y_{\leq t}$  were also omitted for simplicity.



## C Extended related work

### C.1 Counterfactual regression over time: Methods overview

#### C.1.1 Methods included in experiments

In this section, we give a brief overview of models included in our experiments: MSMs [62], RMSN [41], CRN [7], G-Net [39], and CT [48]. To delineate the differences between these models and Causal CPC, we detail in Table 5 the main design differences between all these models.

Table 5: A summary of the methods included in our experiments

| Model                    | Model Backbone        | Tailored to long-term forecast? | Learning of long-term dependencies | Use of contrastive learning | Prediction of counterfactuals | handling selection bias | Invertibility of representation         |
|--------------------------|-----------------------|---------------------------------|------------------------------------|-----------------------------|-------------------------------|-------------------------|-----------------------------------------|
| <b>Causal CPC (ours)</b> | GRU                   | yes                             | CPC                                | learn long-term relations   | Autoregressive                | Balanced representation | yes, contrast representation with input |
| <b>CT</b>                | 3 Transformers        | yes                             | Transformer architecture           | N/A                         | Autoregressive                | Balanced representation | N/A                                     |
| <b>G-Net</b>             | LSTM                  | No                              | N/A                                | N/A                         | Autoregressive                | G-Computation           | Current covariates $X_t$                |
| <b>CRN</b>               | LSTM                  | No                              | N/A                                | N/A                         | Autoregressive                | Balanced representation | N/A                                     |
| <b>RMSN</b>              | LSTM                  | No                              | N/A                                | N/A                         | Autoregressive                | Weighting               | N/A                                     |
| <b>MSM</b>               | Logistic+linear model | No                              | N/A                                | N/A                         | Autoregressive                | Weighting               | N/A                                     |

#### C.1.2 Methods Violating Our Assumptions

Our work relies on the assumption of sequential ignorability, yet several alternative models operate under different assumptions, often addressing the presence of unobserved confounders. Some of these models are rooted in deconfounding theory [45, 60, 80], which has been extended to time-varying settings. Deconfounding involves imposing a factor model on treatment assignment, where each treatment becomes conditionally independent given latent variables that act as proxies for unobserved confounders. Examples of this approach include [8, 28, 10]. Other models assume the presence of proxy variables, inferring a representation of unobserved confounders through probabilistic models based on these proxies [79, 12, 38].

In contrast to our setting, which is governed by the three causal assumptions in Appendix B.1, many models assume a data-generating process similar to [73, 72, 58]. These methods, often non- or semi-parametric, tend to either ignore static covariates or handle them linearly, leading to computational inefficiencies and scalability issues. Nevertheless, some non- or semi-parametric approaches—such as [68, 69, 19, 31]—align with our causal assumptions but extend them to continuous time, treating sequential ignorability in a continuous setting.

Additionally, models like [34] incorporate continuous-time and assume interactions between units, where an individual’s outcome depends on both their treatment and the treatments of others. [6], focusing on binary treatment sequences, requires a stronger version of sequential ignorability—conditional on current covariates—whereas our model assumes a weaker version, conditioning on the entire history of covariates to account for long-lasting confounding effects.

Furthermore, [11] focuses solely on binary treatments and targets the estimation of the average treatment effect on the treated (ATT). The authors assume a specific treatment regime, where individuals enter a post-treatment state after a defined point in time. This assumption is restrictive compared to our framework, which allows for complex, individualized treatment assignment mechanisms

and non-binary treatments, where treatment values can fluctuate over time. As a result, [11] is incompatible with our causal assumptions.

Other methods, like [82], address high-dimensional counterfactual generation based on time-varying treatment plans under the same sequential ignorability assumption. However, they are not designed for causal forecasting over multiple time steps, as required in our setting. Similarly, [22] focuses on estimating the average causal effect and is not suited for predicting individual treatment effects or conditional counterfactual responses, as it targets marginal counterfactual expectations via g-computation.

## C.2 Mutual Information and Self-Supervision

**Self-Supervised Learning and Mutual Information** In self-supervised learning, Deep InfoMax [32] uses MI computation between input images and their representations, focusing on maximizing MI to improve reconstruction quality. Local MI between representations and image patches captures detailed patterns across regions, enhancing encoding. By maximizing average MI with local regions, Deep InfoMax significantly boosts downstream task performance, while global MI plays a key role in reconstructing the entire input from the representation.

CPC aligns with the MI-based approach seen in Deep InfoMax, emphasizing the maximization of MI between global and local representation pairs. Distinct from Deep InfoMax, CPC processes local features sequentially, constructing partial "summary features" to predict specific local features in the future. While classical self-supervised paradigms often focus on tasks like classification or reconstruction-based objectives—favoring either local or global MI maximization—integrating both approaches becomes essential for downstream tasks such as counterfactual regression over time, which justifies why Causal CPC is designed to support both local and global MI maximization to improve temporal predictions.

Several other methods share similarities with CPC, such as Contrastive Multiview Coding [75]. This method emphasizes maximizing mutual information between representations of different views of the same observation. Augmented Multiscale Deep InfoMax [3], akin to CPC, makes predictions across space but differs by predicting representations across layers in the model. While Instance Discrimination [85] encourages representations capable of discriminating between individual examples in the dataset, our preference for CPC arises from its adaptability in processing sequential features in an ordered and autoregressive manner, which aligns with the requirements of our specific context, especially when dealing with counterfactual regression over time.

**Mutual Information and Inductive Bias.** Mutual information (MI) estimation success relies not only on MI's properties but also on the inductive biases from feature representation choices and MI estimator parameterization [77]. Experimental evidence shows that, although MI remains invariant under homeomorphisms, maximization with an invertible encoder during random initialization enhances downstream performance. While higher-capacity critics yield tighter MI bounds, findings consistent with [59] suggest that simpler critics provide better representations, even with looser MI bounds. Accordingly, we selected a simple bilinear critic function for contrastive losses. In vision tasks, augmentations and contrastive loss properties are crucial for representation efficiency [1, 76, 27], and [66] highlights that inductive bias, via function class representation and optimizers, significantly affects downstream performance, offering theoretical, non-vacuous guarantees on representation quality.

**Variational Approaches and MI Estimation Challenges** The estimation of MI faces inherent challenges, particularly within variational lower bounds. These bounds often degrade as MI increases, creating a delicate trade-off between high bias and high variance. To address this, methods that utilize upper bounds on MI have been developed, attempting to mitigate challenges associated with variational bounds. One strategy for MI maximization involves computing gradients of a lower MI bound concerning the parameters of a stochastic encoder. This computational approach potentially eliminates the need for direct MI estimation, providing a more tractable solution. However, estimating MI from samples remains challenging, and traditional approaches encounter scalability issues in modern machine-learning problems.

It's crucial to note that higher estimated MI between observations and learned representations does not consistently translate to improved predictive performance in downstream supervised learning tasks. CPC is an example, exhibiting less variance but more bias, with estimates capped at  $\log |\mathcal{B}|$ . Strategies to reduce bias, such as increasing the batch size, introduce higher computational complexity, requiring additional evaluations for estimating each batch with the encoding function.

In our empirical approach, we adopt a specific sampling strategy for sequences, considering a one-time step per batch. This facilitates computing the InfoNCE between local summary features at time  $t$  and the future prediction of local features, leading to a reduction in algorithmic complexity for contrastive loss computation. Empirical observations demonstrate non-decreased representation quality and improved prediction of factual and counterfactual outcomes.

**Other MI lower bounds.** The Mutual Information Neural Estimator (MINE) [4] leverages the relationship between MI and the Kullback-Leibler (KL) divergence. MI can be expressed as the KL divergence between the joint distribution and the product of marginals:

$$I(X; Z) := D_{KL}(\mathbb{P}_{(X,Z)} \parallel \mathbb{P}_X \otimes \mathbb{P}_Z)$$

MINE employs the Donsker-Varadhan representation [20] of the KL divergence:

$$D_{KL}(\mathbb{P} \parallel \mathbb{Q}) = \sup_{T: \Omega \rightarrow \mathbb{R}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T]) \quad (11)$$

Here, the supremum is taken over all functions  $T$  where the expectations exist. For a specific class of functions  $\mathcal{F}$ , potentially represented by a class of neural networks, we obtain the lower bound:

$$D_{KL}(\mathbb{P} \parallel \mathbb{Q}) \geq \sup_{T \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T]) \quad (12)$$

In practice, we maximize

$$\hat{I}_{\gamma}^{\text{MINE}}(\mathbb{P} \parallel \mathbb{Q}) = \mathbb{E}_{\mathbb{P}}[T_{\gamma}] - \log(\mathbb{E}_{\mathbb{Q}}[e^{T_{\gamma}}]),$$

where  $T_{\gamma}$  is a discriminator parameterized by  $\gamma$ , representing neural network parameters. The MINE estimator is a strongly consistent estimator of the true MI (Theorem 2, [4]).

Alternatively, the **f-divergence** representation of  $D_{KL}$  [51] allows us to derive another MI lower bound, known as the Nguyen, Wainwright, and Jordan (NWJ) estimator [50]:

$$D_{KL}(\mathbb{P} \parallel \mathbb{Q}) \geq \sup_{T \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^{T-1}]) \quad (13)$$

This results in the estimator:

$$\hat{I}_{\gamma}^{\text{NWJ}}(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\mathbb{P}}[T_{\gamma}] - \log(\mathbb{E}_{\mathbb{Q}}[e^{T_{\gamma}-1}]).$$

Unlike the InfoNCE estimator, which exhibits high bias and low variance, the NWJ estimator has a low bias but high variance [57].

## D Experimental protocol

All models were implemented using PyTorch [53] and PyTorch Lightning [21]. In contrast to the approach in [48], we employed early stopping for all models. The stopping criterion was defined as the Mean Squared Error over factual outcomes for a dedicated validation dataset. Specifically, for the Causal CPC encoder, the stopping criterion was determined by the validation loss of the encoder.

While all models in the benchmark were trained using the Adam optimizer [36], we opted for training Causal CPC (encoder plus decoder without the treatment subnetwork) with AdamW [46] due to its observed stability during training. Similar to the common practice in training GAN discriminators, the treatment subnetwork was optimized using SGD with momentum [74].

The CT employed the Exponential Moving Average (EMA) [83] of parameters to enhance training stability. However, this technique was not applied to Causal CPC, as experimental evidence suggested only marginal improvements. Weight decay was set to zero for all models.

For each experiment, the models were trained over five different seeds, and the reported performance metrics include the mean and standard deviation of the results.

The counterfactual trajectories are generated following two strategies:

- **Single sliding treatment** [7, 48]: Trajectories are generated with a single treatment per trajectory while the treatment slides over the forecasting range to generate multiple trajectories. Similar to [7], we apply such a generation scheme to cancer simulation data.
- **Random trajectories**: Trajectories are generated such that at each time step, treatment is generated randomly. We apply random trajectories to semi-synthetic MIMIC data.

For the falsifiability test on MIMIC III dataset, we mask two confounders from the inputs of the benchmark models, namely sodium and glucose measurements.

## E Experiments on synthetic data: Details

### E.1 Description of the Simulation Model

We present a tumor growth simulation model, focusing on the Pharmacokinetic-Pharmacodynamic (PK-PD) model as discussed in [24], a recent approach to predicting treatment responses in non-small cell lung cancer patients. This simulation models the evolution of tumor volume, denoted by  $V(t)$ , in discrete time, where  $t$  represents the number of days since diagnosis:

$$V(t) = \left( 1 + \underbrace{\Lambda \log \left( \frac{K}{V(t-1)} \right)}_{\text{Tumor Growth}} - \underbrace{\kappa_c C(t)}_{\text{Chemotherapy}} - \underbrace{(\kappa_{rd} R_d(t) + v R_d(t)^2)}_{\text{Radiation}} + \underbrace{e_t}_{\text{Noise}} \right) V(t-1) \quad (14)$$

Here, the model parameters  $\Lambda, K, \kappa_c, \kappa_{rd}, v$  are sampled for each patient based on prior distributions from [24]. Additionally,  $R_d(t)$  represents the radiation dose applied at time  $t$ , and  $C(t)$  denotes the drug concentration.

We introduce confounding into the assignment of radiotherapy/chemotherapy treatment by making it dependent on past tumor volume evolution. Treatment is simulated using a Bernoulli distribution with probability  $\sigma(\pi_t)$ , where:

$$\pi_t = \frac{\gamma}{D_{\max}} (\bar{D}(t) - \delta) \quad (15)$$

In this context:  $\bar{D}(t)$  represents the average tumor diameter over the last 15 days,  $D_{\max} = 13$  cm is the maximum tumor diameter,  $\delta$  is set to  $\delta = D_{\max}/2$ .

The parameter  $\gamma$  controls the level of time-dependent confounding, with a higher  $\gamma$  value assigning more weight to the history of tumor diameter in treatment assignment.

### E.2 Additional results

#### E.2.1 Comparison to benchmark models

We report in this section detailed counterfactual errors for Causal CPC and baselines over the cancer simulation dataset, which are responsible for Figure 3.

Table 6: Results on the synthetic data set with sequence length 60: mean±standard deviation of NRMSEs. The best value for each metric is given in bold: smaller is better.

| Model             | $\gamma = 1$     |                  |                  |                  |                  |                  |                  |                  |                  |                  |
|-------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
|                   | $\tau = 1$       | $\tau = 2$       | $\tau = 3$       | $\tau = 4$       | $\tau = 5$       | $\tau = 6$       | $\tau = 7$       | $\tau = 8$       | $\tau = 9$       | $\tau = 10$      |
| Causal CPC (ours) | <b>0.83±0.06</b> | <b>0.86±0.09</b> | <b>0.94±0.09</b> | <b>0.97±0.08</b> | <b>1.03±0.10</b> | <b>1.07±0.10</b> | <b>1.12±0.10</b> | <b>1.17±0.09</b> | <b>1.22±0.08</b> | <b>1.26±0.08</b> |
| CT                | 0.99±0.13        | 0.92±0.14        | 0.98±0.14        | 1.05±0.15        | 1.11±0.18        | 1.11±0.11        | 1.21±0.17        | 1.26±0.16        | 1.31±0.005       | 1.35±0.16        |
| G-Net             | 0.91±0.15        | 1.1±0.16         | 1.24±0.16        | 1.33±0.17        | 1.40±0.18        | 1.47±0.19        | 1.52±0.18        | 1.57±0.22        | 1.63±0.22        | 1.7±0.25         |
| CRN               | 0.84±0.10        | 0.83±0.09        | 0.92±0.10        | 1.00±0.11        | 1.09±0.12        | 1.17±0.14        | 1.25±0.16        | 1.32±0.18        | 1.37±0.23        | 1.43±0.26        |
| RMSN              | 0.99±0.13        | 0.91±0.04        | 1.30±0.65        | 1.43±0.76        | 1.56±0.83        | 1.66±0.88        | 1.73±0.91        | 1.77±0.89        | 1.81±0.88        | 1.84±0.86        |
| MSM               | 1.20±0.10        | 1.83±0.26        | 2.07±0.44        | 2.38±0.44        | 2.54±0.45        | 2.90±0.37        | 3.01±0.38        | 3.06±0.36        | 3.08±0.36        | 3.08±0.36        |
| Model             | $\gamma = 2$     |                  |                  |                  |                  |                  |                  |                  |                  |                  |
|                   | $\tau = 1$       | $\tau = 2$       | $\tau = 3$       | $\tau = 4$       | $\tau = 5$       | $\tau = 6$       | $\tau = 7$       | $\tau = 8$       | $\tau = 9$       | $\tau = 10$      |
| Causal CPC (ours) | 1.16±0.22        | <b>0.91±0.10</b> | <b>0.95±0.13</b> | <b>1.00±0.15</b> | <b>1.07±0.19</b> | <b>1.17±0.24</b> | <b>1.27±0.25</b> | <b>1.38±0.28</b> | <b>1.49±0.30</b> | <b>1.60±0.34</b> |
| CT                | 1.24±0.20        | 1.13±0.15        | 1.27±0.21        | 1.36±0.28        | 1.44±0.29        | 1.55±0.27        | 1.64±0.28        | 1.69±0.20        | 1.74±0.28        | 1.77±0.29        |
| G-Net             | <b>1.05±0.21</b> | 1.05±0.08        | 1.26±0.16        | 1.38±0.23        | 1.48±0.27        | 1.57±0.31        | 1.64±0.33        | 1.70±0.36        | 1.75±0.39        | 1.8±0.42         |
| CRN               | 1.25±0.25        | 1.08±0.06        | 1.14±0.12        | 1.21±0.17        | 1.30±0.21        | 1.41±0.25        | 1.54±0.32        | 1.67±0.41        | 1.8±0.51         | 1.92±0.63        |
| RMSN              | 1.47±0.27        | 1.33±0.25        | 1.30±0.23        | 1.33±0.24        | 1.38±0.26        | 1.45±0.28        | 1.52±0.31        | 1.60±0.25        | 1.67±0.38        | 1.75±0.42        |
| MSM               | 1.43±0.27        | 2.22±0.53        | 2.67±0.63        | 2.98±0.70        | 3.19±0.74        | 3.33±0.77        | 3.41±0.79        | 3.44±0.25        | 3.45±0.78        | 3.34±0.77        |
| Model             | $\gamma = 3$     |                  |                  |                  |                  |                  |                  |                  |                  |                  |
|                   | $\tau = 1$       | $\tau = 2$       | $\tau = 3$       | $\tau = 4$       | $\tau = 5$       | $\tau = 6$       | $\tau = 7$       | $\tau = 8$       | $\tau = 9$       | $\tau = 10$      |
| Causal CPC (ours) | 1.37±0.31        | <b>1.16±0.27</b> | <b>1.26±0.30</b> | <b>1.38±0.35</b> | <b>1.53±0.40</b> | <b>1.69±0.47</b> | <b>1.84±0.52</b> | <b>2.00±0.51</b> | 2.14±0.61        | 2.28±0.66        |
| CT                | 1.36±0.32        | 1.42±0.36        | 1.62±0.46        | 1.78±0.53        | 1.89±0.58        | 2.01±0.63        | 2.13±0.66        | 2.22±0.69        | 2.31±0.69        | 2.37±0.73        |
| G-Net             | <b>1.14±0.24</b> | 1.22±0.15        | 1.54±0.26        | 1.77±0.33        | 1.94±0.36        | 2.09±0.40        | 2.23±0.43        | 2.34±0.47        | 2.44±0.52        | 2.52±0.56        |
| CRN               | 1.46±0.29        | 1.54±0.38        | 1.70±0.48        | 1.79±0.53        | 1.86±0.92        | 1.92±0.58        | 1.98±0.59        | 2.04±0.61        | <b>2.10±0.63</b> | <b>2.16±0.64</b> |
| RMSN              | 1.22±0.26        | 1.28±0.29        | 1.43±0.40        | 1.56±0.48        | 1.70±0.53        | 1.83±0.57        | 1.95±0.59        | 2.06±0.61        | 2.14±0.61        | 2.21±0.61        |
| MSM               | 1.70±0.35        | 2.73±0.88        | 3.22±1.03        | 3.25±1.12        | 3.71±1.18        | 3.85±1.22        | 3.91±1.23        | 3.95±1.24        | 3.96±1.24        | 3.94±1.23        |

#### E.2.2 Ablation study

We detail here the results of the ablation study conducted on the cancer simulation dataset (Table 3). The (full) Causal CPC model, as presented in the core paper, gives, in most cases, better results than any ablation configuration.

Table 7: Results on the synthetic data set with sequence length 40: mean±standard deviation of NRMSEs. The best value for each metric is given in bold: smaller is better.

| Model             | $\gamma = 1$     |                  |                  |                  |                  |                  |                  |                  |                  |                  |
|-------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
|                   | $\tau = 1$       | $\tau = 2$       | $\tau = 3$       | $\tau = 4$       | $\tau = 5$       | $\tau = 6$       | $\tau = 7$       | $\tau = 8$       | $\tau = 9$       | $\tau = 10$      |
| Causal CPC (ours) | 1.21±0.07        | 1.13±0.12        | 1.22±0.12        | <b>1.27±0.17</b> | <b>1.35±0.19</b> | <b>1.43±0.21</b> | <b>1.48±0.22</b> | <b>1.54±0.22</b> | <b>1.58±0.23</b> | <b>1.62±0.23</b> |
| CT                | 1.21±0.09        | 1.34±0.10        | 1.43±0.14        | 1.52±0.19        | 1.59±0.22        | 1.65±0.23        | 1.71±0.23        | 1.76±0.21        | 1.78±0.21        | 1.80±0.18        |
| G-Net             | 1.11±0.10        | 1.30±0.17        | 1.52±0.20        | 1.65±0.21        | 1.76±0.24        | 1.86±0.29        | 1.96±0.34        | 2.05±0.40        | 2.13±0.46        | 2.19±0.52        |
| CRN               | <b>1.01±0.12</b> | <b>1.11±0.14</b> | <b>1.19±0.14</b> | 1.30±0.13        | 1.41±0.12        | 1.49±0.09        | 1.56±0.07        | 1.62±0.05        | 1.67±0.04        | 1.70±0.04        |
| RMSN              | 1.14±0.03        | 1.20±0.08        | 1.24±0.07        | 1.32±0.08        | 1.40±0.08        | 1.49±0.08        | 1.57±0.08        | 1.64±0.06        | 1.71±0.05        | 1.76±0.04        |
| Model             | $\gamma = 2$     |                  |                  |                  |                  |                  |                  |                  |                  |                  |
|                   | $\tau = 1$       | $\tau = 2$       | $\tau = 3$       | $\tau = 4$       | $\tau = 5$       | $\tau = 6$       | $\tau = 7$       | $\tau = 8$       | $\tau = 9$       | $\tau = 10$      |
| Causal CPC (ours) | 1.41±0.09        | 1.30±0.17        | 1.33±0.20        | <b>1.41±0.25</b> | <b>1.50±0.28</b> | <b>1.56±0.31</b> | <b>1.60±0.23</b> | <b>1.66±0.27</b> | <b>1.70±0.26</b> | <b>1.74±0.25</b> |
| CT                | 1.36±0.15        | 1.38±0.19        | 1.40±0.25        | 1.48±0.28        | 1.54±0.30        | 1.62±0.31        | 1.70±0.31        | 1.75±0.31        | 1.82±0.32        | 1.85±0.29        |
| G-Net             | 1.26±0.09        | 1.47±0.20        | 1.74±0.28        | 1.90±0.34        | 2.02±0.40        | 2.12±0.44        | 2.22±0.50        | 2.31±0.55        | 2.39±0.61        | 2.47±0.68        |
| CRN               | <b>1.24±0.22</b> | <b>1.21±0.18</b> | <b>1.31±0.22</b> | 1.42±0.25        | 1.55±0.29        | 1.67±0.34        | 1.79±0.41        | 1.91±0.50        | 2.03±0.59        | 2.13±0.70        |
| RMSN              | 1.28±0.15        | 1.28±0.13        | 1.37±0.12        | 1.49±0.15        | 1.62±0.19        | 1.76±0.25        | 1.89±0.30        | 2.03±0.36        | 2.15±0.41        | 2.26±0.46        |
| Model             | $\gamma = 3$     |                  |                  |                  |                  |                  |                  |                  |                  |                  |
|                   | $\tau = 1$       | $\tau = 2$       | $\tau = 3$       | $\tau = 4$       | $\tau = 5$       | $\tau = 6$       | $\tau = 7$       | $\tau = 8$       | $\tau = 9$       | $\tau = 10$      |
| Causal CPC (ours) | 1.52±0.19        | <b>1.19±0.16</b> | <b>1.26±0.23</b> | <b>1.38±0.25</b> | <b>1.52±0.24</b> | <b>1.66±0.25</b> | <b>1.79±0.27</b> | <b>1.91±0.32</b> | <b>2.02±0.38</b> | <b>2.11±0.46</b> |
| CT                | 1.56±0.18        | 1.68±0.33        | 1.77±0.54        | 1.95±0.75        | 2.05±0.83        | 2.19±0.92        | 2.29±1.00        | 2.35±1.04        | 2.43±1.09        | 2.43±1.07        |
| G-Net             | 1.19±0.02        | 1.38±0.04        | 1.72±0.02        | 1.92±0.08        | 2.11±0.18        | 2.32±0.29        | 2.52±0.41        | 2.72±0.54        | 2.93±0.68        | 3.13±0.83        |
| CRN               | 1.50±0.01        | 1.47±0.07        | 1.62±0.24        | 1.76±0.40        | 1.89±0.51        | 1.98±0.57        | 2.05±0.61        | 2.10±0.62        | 2.13±0.61        | 2.15±0.58        |
| RMSN              | <b>1.48±0.19</b> | 1.40±0.13        | 1.66±0.35        | 1.86±0.52        | 2.01±0.61        | 2.14±0.65        | 2.24±0.67        | 2.32±0.67        | 2.37±0.66        | 2.39±0.65        |

Table 8: Results of the ablation study on the synthetic data set: mean±standard deviation of Normalized Rooted Mean Squared Errors (NRMSEs). The best value for each metric is given in bold: smaller is better.

| Model                                       | $\tau = 1$         | $\tau = 2$         | $\tau = 3$       | $\tau = 4$         | $\tau = 5$         | $\tau = 6$         | $\tau = 7$         | $\tau = 8$         | $\tau = 9$         | $\tau = 10$        |
|---------------------------------------------|--------------------|--------------------|------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| CAUSAL CPC (FULL)                           | <b>0.83 ± 0.06</b> | <b>0.86 ± 0.06</b> | 0.94 ± 0.09      | <b>0.97 ± 0.08</b> | <b>1.03 ± 0.10</b> | <b>1.07 ± 0.10</b> | <b>1.12 ± 0.10</b> | <b>1.17 ± 0.06</b> | <b>1.22 ± 0.08</b> | <b>1.26 ± 0.08</b> |
| Causal CPC (w/o $\mathcal{L}^{(InfoNCE)}$ ) | 0.84±0.04          | 0.91±0.07          | 0.95±0.07        | 0.99±0.09          | 1.03±0.10          | 1.10±0.07          | 1.15±0.14          | 1.20±0.14          | 1.23±0.14          | 1.28±0.15          |
| Causal CPC (w/o $\mathcal{L}^{(InfoMax)}$ ) | 0.84±0.04          | 0.86±0.09          | <b>0.91±0.08</b> | 0.99±0.10          | 1.07±0.08          | 1.16±0.08          | 1.24±0.10          | 1.31±0.12          | 1.38±0.08          | 1.46±0.10          |
| Causal CPC (w CDC loss)                     | 0.83±0.02          | 0.89±0.07          | 0.96±0.07        | 1.03±0.07          | 1.07±0.08          | 1.10±0.07          | 1.13±0.10          | 1.18±0.09          | 1.24±0.11          | 1.28±0.11          |
| Causal CPC (w Balancing)                    | 0.84±0.04          | 0.88±0.05          | 0.97±0.05        | 1.04±0.07          | 1.08±0.10          | 1.13±0.08          | 1.15±0.14          | 1.20±0.10          | 1.25±0.08          | 1.29±0.12          |

## F Experiments on semi-synthetic data: Details

### F.1 Description of the simulation model

In this section, we provide a concise overview of the simulation model built upon the MIMIC III dataset, as introduced by [48]. Initially, a cohort of 1,000 patients is extracted from the MIMIC III data, and the simulation proposed by [48] extends the model of [68].

Let  $d_y$  be the dimension of the outcome variable. In the case of multiple outcomes, untreated outcomes, denoted as  $\mathbf{Z}_t^{j,(i)}$  for  $j = 1, \dots, d_y$ , are generated for each patient  $i$  within the cohort. The generation process is defined as follows:

$$\mathbf{Z}_t^{j,(i)} = \underbrace{\alpha_S^j \mathbf{B}\text{-spline}(t)}_{\text{endogenous}} + \underbrace{\alpha_g^j g^{j,(i)}(t) + \alpha_f^j f_Z^j(\mathbf{X}_t^{(i)})}_{\text{exogenous}} + \underbrace{\varepsilon_t}_{\text{noise}} \quad (16)$$

where: the B-spline  $\mathbf{B}\text{-spline}(t)$  is an endogenous component,  $g^{j,(i)}(\cdot)$  is sampled independently for each patient from a Gaussian process with a Matérn kernel and  $f_Z^j(\cdot)$  is sampled from a Random Fourier Features (RFF) approximation of a Gaussian process.

To introduce confounding in the assignment mechanism, current time-varying covariates are incorporated via a random function  $f_Y^l(\mathbf{X}_t)$  and the average of the subset of the previous  $T_l$  treated outcomes,  $\bar{A}_{T_l}(\bar{\mathbf{Y}}_{t-1})$ . For  $d_a$  binary treatments  $\mathbf{A}_t^l$ , where  $l = 1, \dots, d_a$ , the assignment mechanism is modeled as:

$$p_{\mathbf{A}_t^l} = \sigma(\gamma_A^l \bar{A}_{T_l}(\bar{\mathbf{Y}}_{t-1}) + \gamma_X^l f_Y^l(\mathbf{X}_t) + b_l), \\ \mathbf{A}_t^l \sim \text{Bernoulli}(p_{\mathbf{A}_t^l}).$$

Subsequently, treatments are applied to the untreated outcomes using the following expression:

$$E^j(t) = \sum_{i=t-w^l}^t \frac{\min_{l=1,\dots,d_a} \mathbb{1}_{[\mathbf{A}_i^l=1]} p_{\mathbf{A}_i^l} \beta_{lj}}{(w^l - i)^2} \quad (17)$$

The final outcome combines the treatment effect and the untreated simulated outcome:

$$Y_t^j = Z_t^j + E^j(t). \quad (18)$$

## F.2 Additional results

### F.2.1 Ablation study

We detail here the results of the ablation study conducted on the MIMIC III semi-synthetic dataset (Table 3). The (full) Causal CPC model, as presented in the core paper, gives consistently better results than any ablation configuration.

Table 9: Results on MIMIC III semi-synthetic data set: mean±standard deviation of Normalized Rooted Mean Squared Errors (NRMSEs). The best value for each metric is given in bold: smaller is better.

| Model                                       | $\tau = 1$       | $\tau = 2$       | $\tau = 3$       | $\tau = 4$       | $\tau = 5$       | $\tau = 6$       | $\tau = 7$       | $\tau = 8$       | $\tau = 9$       | $\tau = 10$      |
|---------------------------------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| Causal CPC (ful)                            | <b>0.32±0.04</b> | <b>0.45±0.08</b> | <b>0.54±0.06</b> | <b>0.61±0.10</b> | <b>0.66±0.10</b> | <b>0.69±0.11</b> | <b>0.71±0.11</b> | <b>0.73±0.06</b> | <b>0.75±0.05</b> | <b>0.77±0.10</b> |
| Causal CPC (w/o $\mathcal{L}^{(InfoNCE)}$ ) | 0.35±0.04        | 0.50±0.05        | 0.59±0.06        | 0.66±0.06        | 0.71±0.08        | 0.75±0.06        | 0.77±0.07        | 0.79±0.08        | 0.81±0.07        | 0.83±0.07        |
| Causal CPC (w/o $\mathcal{L}^{(InfoMax)}$ ) | 0.36±0.02        | 0.53±0.03        | 0.64±0.04        | 0.71±0.05        | 0.77±0.05        | 0.77±0.05        | 0.83±0.05        | 0.86±0.05        | 0.88±0.08        | 0.90±0.05        |
| Causal CPC (CDC loss)                       | 0.36±0.02        | 0.54±0.03        | 0.65±0.05        | 0.72±0.05        | 0.77±0.05        | 0.70±0.04        | 0.83±0.04        | 0.85±0.03        | 0.86±0.03        | 0.88±0.08        |
| Causal CPC (w/o balancing)                  | 0.35±0.03        | 0.50±0.05        | 0.60±0.06        | 0.67±0.06        | 0.72±0.06        | 0.76±0.06        | 0.78±0.06        | 0.80±0.06        | 0.83±0.06        | 0.85±0.06        |

Furthermore, We replace the InfoNCE objective used to compute the CPC term and InfoMax terms with that of NWJ and MINE (Section C.2). We repeat the same MIMIC III experimentation while varying the objective used for CPC and InfoMax. Table 10 shows the counterfactual errors for each configuration compared to the original formulation of Causal CPC. In all cases, The InfoNCE objective performs better with notable error reduction at large horizons.

Table 10: Results of NWJ and MINE MI lower bounds when used for CPC and InfoMax for MIMIC III semi-synthetic data set: mean±standard deviation of Normalized Rooted Mean Squared Errors (NRMSEs). The best value for each metric is given in bold: smaller is better.

| Model             | $\tau = 1$       | $\tau = 2$       | $\tau = 3$       | $\tau = 4$       | $\tau = 5$       | $\tau = 6$       | $\tau = 7$       | $\tau = 8$       | $\tau = 9$       | $\tau = 10$      |
|-------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| Original Model    | <b>0.34±0.04</b> | <b>0.45±0.08</b> | <b>0.54±0.06</b> | <b>0.61±0.10</b> | <b>0.66±0.10</b> | <b>0.69±0.11</b> | <b>0.71±0.11</b> | <b>0.73±0.06</b> | <b>0.75±0.05</b> | <b>0.77±0.10</b> |
| CPC with NWJ      | 0.34±0.04        | 0.48±0.05        | 0.58±0.06        | 0.66±0.07        | 0.71±0.08        | 0.75±0.07        | 0.78±0.07        | 0.81±0.06        | 0.84±0.06        | 0.87±0.06        |
| CPC with MINE     | 0.35±0.03        | 0.50±0.05        | 0.61±0.04        | 0.69±0.04        | 0.75±0.04        | 0.79±0.03        | 0.82±0.03        | 0.85±0.02        | 0.88±0.02        | 0.91±0.02        |
| InfoMax with NWJ  | 0.42±0.08        | 0.56±0.04        | 0.69±0.07        | 0.77±0.08        | 0.83±0.09        | 0.87±0.09        | 0.90±0.09        | 0.92±0.09        | 0.94±0.08        | 0.96±0.08        |
| InfoMax with MINE | 0.37±0.05        | 0.52±0.03        | 0.65±0.06        | 0.73±0.8         | 0.80±0.10        | 0.84±0.11        | 0.87±0.11        | 0.89±0.10        | 0.91±0.10        | 0.93±0.09        |

### F.2.2 Comparison to benchmark models: standard train/test split

As mentioned in Section 6.2, We also tested Causal CPC on MIMIC III semi-synthetic data using the same experimental protocol as [48], namely by using the split of patients into train/validation/test as 800/200/200. As a result, baseline performances in Table 11 are exactly the same as in [48].

Table 11: Results over the MIMIC III semi-synthetic data set (same experimental protocol as in [48]): mean±standard deviation of Rooted Mean Squared Errors (RMSEs). The best value for each metric is given in bold: smaller is better.

| Model             | $\tau = 1$       | $\tau = 2$       | $\tau = 3$       | $\tau = 4$       | $\tau = 5$       | $\tau = 6$       | $\tau = 7$       | $\tau = 8$       | $\tau = 9$       | $\tau = 10$      |
|-------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| Causal CPC (ours) | 0.25±0.03        | <b>0.37±0.02</b> | <b>0.40±0.01</b> | <b>0.45±0.01</b> | <b>0.49±0.02</b> | <b>0.52±0.02</b> | <b>0.55±0.03</b> | <b>0.56±0.03</b> | <b>0.58±0.04</b> | <b>0.60±0.03</b> |
| CT                | <b>0.20±0.01</b> | 0.38±0.01        | 0.45±0.01        | 0.49±0.01        | 0.52±0.02        | 0.53±0.02        | 0.55±0.02        | <b>0.56±0.02</b> | <b>0.58±0.02</b> | <b>0.59±0.02</b> |
| G-Net             | 0.34±0.01        | 0.67±0.03        | 0.83±0.04        | 0.94±0.04        | 1.03±0.05        | 1.10±0.05        | 1.16±0.05        | 1.21±0.06        | 1.25±0.06        | 1.29±0.06        |
| CRN               | 0.30±0.01        | 0.48±0.02        | 0.59±0.02        | 0.65±0.02        | 0.68±0.02        | 0.71±0.01        | 0.72±0.01        | 0.74±0.01        | 0.76±0.01        | 0.78±0.02        |
| RMSN              | 0.24±0.01        | 0.47±0.01        | 0.60±0.01        | 0.70±0.02        | 0.78±0.04        | 0.84±0.05        | 0.89±0.06        | 0.94±0.08        | 0.97±0.09        | 1.00±0.11        |
| MSM               | 0.37±0.01        | 0.57±0.03        | 0.74±0.06        | 0.88±0.03        | 1.14±0.10        | 1.95±1.48        | 3.44±4.57        | > 10.0           | > 10.0           | > 10.0           |

### F.2.3 Running time and model complexity

In this section, we complement the table about complexity and running time given for cancer simulation in the core paper by providing the exact same table but for MIMIC III semi-synthetic data.

## G Proofs of theoretical results

### G.1 Relation between InfoNCE loss and mutual information

**Proposition G.1.**

$$I(\mathbf{U}_{t+j}, \mathbf{C}_t) \geq \log(|\mathcal{B}|) - \mathcal{L}_j^{(InfoNCE)}$$

Table 12: The number of parameters to train for each model after hyper-parameters fine-tuning and the corresponding running time averaged over five seeds. Results are reported for semi-synthetic MIMIC III data; the processing unit is GPU - 1 x NVIDIA Tesla M60 .

| MODEL             | TRAINABLE PARAMETERS (K) | TRAINING TIME (MIN) | PREDICTION TIME (MIN) |
|-------------------|--------------------------|---------------------|-----------------------|
| CAUSAL CPC (OURS) | <b>9.8</b>               | 12±2                | <b>4±1</b>            |
| CT                | 12                       | 14±1                | 38±2                  |
| G-NET             | 14.7                     | 7±1                 | 40±3                  |
| CRN               | 15.1                     | 21±2                | 5±1                   |
| RMSN              | 20                       | 48±4                | 5±1                   |

*Proof.* In the following, we draw inspiration from the proof of [52]. The InfoNCE loss corresponds to the categorical cross-entropy of classifying the positive sample  $\mathbf{U}_{t+j}$  correctly, given the context  $\mathbf{C}_t$ , with a probability:

$$\frac{\exp(T_j(\mathbf{U}_{t+j}, \mathbf{C}_t))}{\sum_{l=1}^{|\mathcal{B}|} \exp(T_j(\mathbf{U}_{l,t+j}, \mathbf{C}_t))}.$$

The positive sample  $\mathbf{U}_{t+j}$  is one element in the batch  $\mathcal{B}$ , where the remaining elements serve as negative samples. Let  $\text{pos} \in \{1, \dots, |\mathcal{B}|\}$  be the indicator of the positive sample  $\mathbf{U}_{t+j}$ . The optimal probability is given by:

$$p(\text{Index} = \text{pos} \mid \mathcal{B}, \mathbf{C}_t) = \frac{p(\mathbf{u}_{\text{pos},t+j} \mid \mathbf{C}_t) \prod_{l=1, \dots, |\mathcal{B}|; l \neq \text{pos}} p(\mathbf{u}_{l,t+j})}{\sum_{j=1}^{|\mathcal{B}|} \left[ p(\mathbf{u}_{j,t+j} \mid \mathbf{C}_t) \prod_{l=1, \dots, |\mathcal{B}|; l \neq j} p(\mathbf{u}_{l,t+j}) \right]} = \frac{\frac{p(\mathbf{u}_{\text{pos},t+j} \mid \mathbf{C}_t)}{p(\mathbf{u}_{\text{pos},t+j})}}{\sum_{j=1}^{|\mathcal{B}|} \frac{p(\mathbf{u}_{j,t+j} \mid \mathbf{C}_t)}{p(\mathbf{u}_{j,t+j})}}.$$

For the score  $\exp(T_j(\mathbf{U}_{t+j}, \mathbf{C}_t))$  to be optimal, it should be proportional to  $\frac{p(\mathbf{u}_{\text{pos},t+j} \mid \mathbf{C}_t)}{p(\mathbf{u}_{\text{pos},t+j})}$ . The mutual information (MI) lower bound arises from the fact that  $\exp(T_j(\mathbf{U}_{t+j}, \mathbf{C}_t))$  estimates the density ratio  $\frac{p(\mathbf{u}_{\text{pos},t+j} \mid \mathbf{C}_t)}{p(\mathbf{u}_{\text{pos},t+j})}$ .

$$\begin{aligned} \mathcal{L}_j^{(\text{InfoNCE})} &= -\mathbb{E}_{\mathcal{B}} \log \left[ \frac{\frac{p(\mathbf{u}_{t+j} \mid \mathbf{C}_t)}{p(\mathbf{u}_{t+j})}}{\frac{p(\mathbf{u}_{t+j} \mid \mathbf{C}_t)}{p(\mathbf{u}_{t+j})} + \sum_{\mathbf{u}_{l,t+j} \in \mathcal{B}_{\text{neg}}} \frac{p(\mathbf{u}_{l,t+j} \mid \mathbf{C}_t)}{p(\mathbf{u}_{l,t+j})}} \right] \\ &= \mathbb{E}_{\mathcal{B}} \log \left[ 1 + \frac{p(\mathbf{u}_{t+j})}{p(\mathbf{u}_{t+j} \mid \mathbf{C}_t)} \sum_{\mathbf{u}_{l,t+j} \in \mathcal{B}_{\text{neg}}} \frac{p(\mathbf{u}_{l,t+j} \mid \mathbf{C}_t)}{p(\mathbf{u}_{l,t+j})} \right] \\ &\approx \mathbb{E}_{\mathcal{B}} \log \left[ 1 + \frac{p(\mathbf{u}_{t+j})}{p(\mathbf{u}_{l,t+j} \mid \mathbf{C}_t)} (|\mathcal{B}| - 1) \mathbb{E}_{\mathbf{U}_{t+j}} \frac{p(\mathbf{u}_{l,t+j} \mid \mathbf{C}_t)}{p(\mathbf{u}_{l,t+j})} \right] \quad (19) \\ &= \mathbb{E}_{\mathcal{B}} \log \left[ 1 + \frac{p(\mathbf{u}_{t+j})}{p(\mathbf{u}_{t+j} \mid \mathbf{C}_t)} (|\mathcal{B}| - 1) \right] \\ &\geq \mathbb{E}_{\mathcal{B}} \log \left[ \frac{p(\mathbf{u}_{t+j})}{p(\mathbf{u}_{t+j} \mid \mathbf{C}_t)} |\mathcal{B}| \right] \\ &= -I(\mathbf{u}_{t+j}, \mathbf{C}_t) + \log(|\mathcal{B}|), \end{aligned}$$

The approximation in the third equation, Eq. (19), becomes more precise as the batch size increases.  $\square$

## G.2 Relation between InfoMax and Input Reconstruction

We now prove Proposition 5.1, which states that:

$$I(\mathbf{C}_t^h, \mathbf{C}_t^f) \leq I(\mathbf{H}_t, (\mathbf{C}_t^h, \mathbf{C}_t^f)).$$

*Proof.* This follows from two applications of the data processing inequality [17], which states that for random variables  $A$ ,  $B$ , and  $C$  satisfying the Markov relation  $A \rightarrow B \rightarrow C$ , the inequality  $I(A; C) \leq I(A; B)$  holds.

First, since  $\mathbf{C}_t^h = \Phi_{\theta_1, \theta_2}(\mathbf{H}_t^h)$  and  $\mathbf{C}_t^f = \Phi_{\theta_1, \theta_2}(\mathbf{H}_t^f)$ , we can write  $\mathbf{H}_t^h = \text{trunc}_f(\mathbf{H}_t)$  and  $\mathbf{H}_t^f = \text{trunc}_h(\mathbf{H}_t)$ , where  $\text{trunc}_f$  and  $\text{trunc}_h$  truncate the future and history processes, respectively, given a splitting time  $t_0$ .

Thus, we have the Markov relation:

$$\mathbf{C}_t^h \xleftarrow{\Phi_{\theta_1, \theta_2} \circ \text{trunc}_f} \mathbf{H}_t \xrightarrow{\Phi_{\theta_1, \theta_2} \circ \text{trunc}_h} \mathbf{C}_t^f,$$

which is Markov equivalent to:

$$\mathbf{C}_t^h \xrightarrow{\Phi_{\theta_1, \theta_2} \circ \text{trunc}_f} \mathbf{H}_t \xrightarrow{\Phi_{\theta_1, \theta_2} \circ \text{trunc}_h} \mathbf{C}_t^f.$$

By the data processing inequality, this results in  $I(\mathbf{C}_t^h, \mathbf{C}_t^f) \leq I(\mathbf{H}_t, \mathbf{C}_t^h)$ . On the other hand, we have the trivial Markov relation  $\mathbf{H}_t \rightarrow (\mathbf{C}_t^h, \mathbf{C}_t^f) \rightarrow \mathbf{C}_t^h$ , which implies  $I(\mathbf{H}_t, \mathbf{C}_t^h) \leq I(\mathbf{H}_t, (\mathbf{C}_t^h, \mathbf{C}_t^f))$ . Combining these two inequalities proves the proposition.  $\square$

### G.3 Proof of Theorem 5.2

To begin, we split the process history into two non-overlapping views (Figure 2):  $\mathbf{H}_t^h := \mathbf{U}_{1:t_0}$  and  $\mathbf{H}_t^f := \mathbf{U}_{t_0+1:t}$ , representing a historical subsequence and a future subsequence within the process history  $\mathbf{H}_t$ , respectively. We then computed representations of these two views denoted  $\mathbf{C}_t^h$  and  $\mathbf{C}_t^f$ , respectively. This naturally gives rise to the Markov chain, as in showed in the proof of proposition 5.1:

$$\mathbf{C}_t^h \leftarrow \mathbf{H}_t \rightarrow \mathbf{C}_t^f$$

which is Markov equivalent to:

$$\mathbf{C}_t^h \rightarrow \mathbf{H}_t \rightarrow \mathbf{C}_t^f$$

Following this Markov chain, we can show that [71]:

$$I(\mathbf{C}_t^f, \mathbf{C}_t^h) = I(\mathbf{H}_t, \mathbf{C}_t^h) - \mathbb{E}_{\mathbf{h}_t \sim \mathbb{P}_{\mathbf{H}_t}} \mathbb{E}_{\mathbf{c}_t^f \sim \mathbb{P}_{\mathbf{C}_t^f | \mathbf{h}_t}} \left[ D_{KL}[\mathbb{P}_{\mathbf{C}_t^h | \mathbf{h}_t} \| \mathbb{P}_{\mathbf{C}_t^h | \mathbf{c}_t^f}] \right]$$

On the other hand, by applying the chain rule of the mutual information to  $I(\mathbf{H}_t; (\mathbf{C}_t^h, \mathbf{C}_t^f))$  we get:

$$I(\mathbf{H}_t; (\mathbf{C}_t^f, \mathbf{C}_t^h)) = I(\mathbf{H}_t, \mathbf{C}_t^h) + I(\mathbf{H}_t; \mathbf{C}_t^f | \mathbf{C}_t^h)$$

Combining these equations, the tightness of our bounds can be written as:

$$I(\mathbf{H}_t; (\mathbf{C}_t^f, \mathbf{C}_t^h)) - I(\mathbf{C}_t^f, \mathbf{C}_t^h) = I(\mathbf{H}_t; \mathbf{C}_t^f | \mathbf{C}_t^h) + \mathbb{E}_{\mathbf{h}_t \sim \mathbb{P}_{\mathbf{H}_t}} \mathbb{E}_{\mathbf{c}_t^f \sim \mathbb{P}_{\mathbf{C}_t^f | \mathbf{h}_t}} \left[ D_{KL}[\mathbb{P}_{\mathbf{C}_t^h | \mathbf{h}_t} \| \mathbb{P}_{\mathbf{C}_t^h | \mathbf{c}_t^f}] \right]$$

### G.4 On the Relation Between Conditional Entropy and Reconstruction

We now prove the statement in the core paper, which asserts that the conditional entropy  $H(\mathbf{H}_t | (\mathbf{C}_t^h, \mathbf{C}_t^f)) \geq 0$  is minimized if  $\mathbf{H}_t$  is almost surely a function of  $(\mathbf{C}_t^h, \mathbf{C}_t^f)$ . The proof is adapted from [17].

**Proposition G.2.** *If  $H(\mathbf{A} | \mathbf{B}) = 0$ , then  $\mathbf{A} = f(\mathbf{B})$  almost surely.*

*Proof.* For simplicity, suppose  $\mathbf{A}$  and  $\mathbf{B}$  are discrete random variables. Assume, by contradiction, that there exists  $\mathbf{b}_0$  and two distinct values  $\mathbf{a}_1$  and  $\mathbf{a}_2$  such that  $p(\mathbf{a}_1 | \mathbf{b}_0) > 0$  and  $p(\mathbf{a}_2 | \mathbf{b}_0) > 0$ . Then, the conditional entropy is given by:

$$H(\mathbf{A} | \mathbf{B}) = - \sum_{\mathbf{b}} p(\mathbf{b}) \sum_{\mathbf{a}} p(\mathbf{a} | \mathbf{b}) \log p(\mathbf{a} | \mathbf{b}).$$

In particular, we have:

$$H(\mathbf{A} | \mathbf{B}) \geq p(\mathbf{b}_0) (-p(\mathbf{a}_1 | \mathbf{b}_0) \log p(\mathbf{a}_1 | \mathbf{b}_0) - p(\mathbf{a}_2 | \mathbf{b}_0) \log p(\mathbf{a}_2 | \mathbf{b}_0)) > 0.$$

Since  $-\log t \geq 0$  for  $0 \leq t \leq 1$  and is strictly positive for  $t$  not equal to 0 or 1, the conditional entropy  $H(\mathbf{A} | \mathbf{B}) = 0$  if and only if  $\mathbf{A}$  is almost surely a function of  $\mathbf{B}$ .  $\square$

### G.5 On the benefit of the InfoMax loss on inverting the data generation process

To ensure identifiability in the latent space, we leverage recent advances in causal and disentangled representation learning. Suppose the true data-generating process is given by  $\mathbf{H}_t = g(\mathbf{z}_t)$ , where  $\mathbf{z}_t$  represents the true latent factors. In the sequential context, we assume that the same function  $g$  generates two historical subsequences:

$$\mathbf{H}_t^f = g(\mathbf{z}_t^f), \quad \mathbf{H}_t^h = g(\mathbf{z}_t^h).$$

We assume a general dependency of the form:

$$p(\mathbf{z}_t^f | \mathbf{z}_t^h) = \frac{Q(\mathbf{z}_t^f)}{Z(\mathbf{z}_t^h)} \exp(-d(\mathbf{z}_t^f, \mathbf{z}_t^h)).$$

Here,  $\Phi$  is an encoder, and we use the InfoMax regularization term as follows:

$$\mathcal{L}^{(InfoMax)}(\Phi, d, \mathcal{B}) := -\mathbb{E}_{\mathcal{B}} \left[ \log \frac{\exp(-d(\Phi(\mathbf{H}_t^f), \Phi(\mathbf{H}_t^h)))}{\sum_{l=1}^{|\mathcal{B}|} \exp(-d(\Phi(\mathbf{H}_{l,t}^f), \Phi(\mathbf{H}_t^h)))} \right]$$

According to [47], under certain conditions, if the encoder  $f$  minimizes  $\mathcal{L}^{(InfoMax)}$ , then  $h = g \circ f$  is a scaled permutation matrix. This result suggests that when the encoder achieves a minimizer for  $\mathcal{L}^{(InfoMax)}$ , the encoder function  $f$  closely approximates an invertible transformation of  $g$ .

From a causal inference perspective, if  $Y_{it}(\omega_{it}) \perp\!\!\!\perp W_{it} | \mathbf{H}_{it}$  and  $\mathbf{H}_{it} = g(\mathbf{Z}_{it})$ , then an invertible function  $g \circ f$  ensures that:

$$Y_{it}(\omega_{it}) \perp\!\!\!\perp W_{it} | g \circ f(\mathbf{H}_{it}).$$

Thus,  $Y_{it}(\omega_{it}) \perp\!\!\!\perp W_{it} | g(\mathbf{C}_{it})$  and since  $g$  is invertible, we have:

$$Y_{it}(\omega_{it}) \perp\!\!\!\perp W_{it} | \mathbf{C}_{it}.$$

This demonstrates that the representation  $\mathbf{C}_{it}$  retains the essential independence structure, facilitating accurate counterfactual inference.

### G.6 Proof of theorem 5.4

To prove the Theorem 5.4, we first prove the following lemma and proposition.

**Lemma G.3.** *Let  $\Phi$  be a fixed representation function. Given that  $q(W_{t+1} | \Phi(\mathbf{H}_t))$  is the conditional likelihood of observing the treatment  $W_{t+1}$ , denote the probability of observing each treatment value as  $q^j = q(\Phi(\mathbf{H}_t)) := q(W_{t+1} = j | \Phi(\mathbf{H}_t))$  for  $j \in \{0, 1, \dots, K-1\}$ . Then, the optimal treatment prediction function is such that*

$$q^{j,*}(\Phi(\mathbf{H}_t)) = \frac{p(\Phi(\mathbf{H}_t) | W_{t+1} = j)}{\sum_{l=0}^{K-1} p(\Phi(\mathbf{H}_t) | W_{t+1} = l)p(W_{t+1} = l)} \quad (20)$$

*Proof.* For a fixed representation  $\Phi$ , finding the optimal treatment probabilities amounts to solving the following optimization problem:

$$\max_q \mathbb{E}_{\mathbb{P}(\Phi(\mathbf{H}_t), W_{t+1})} [\log q(W_{t+1} | \Phi(\mathbf{H}_t))] \quad \text{subject to} \quad \sum_{l=0}^{K-1} q^l(\Phi(\mathbf{H}_t)) = 1 \quad (21)$$

First, we write the likelihood  $q(W_{t+1} | \Phi(\mathbf{H}_t))$  using the conditional probabilities  $q^j(\Phi(\mathbf{H}_t))$ .

$$q(W_{t+1} | \Phi(\mathbf{H}_t)) = \prod_{j=0}^{K-1} q^j(\Phi(\mathbf{H}_t))^{\mathbb{1}_{\{W_{t+1}=j\}}}$$

Then, the treatment likelihood can be written as

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}(\Phi(\mathbf{H}_t), W_{t+1})} [\log q(W_{t+1} \mid \Phi(\mathbf{H}_t))] &= \mathbb{E}_{\mathbb{P}(\Phi(\mathbf{H}_t), W_{t+1})} \left[ \sum_{l=0}^{K-1} \log(q^l(\Phi(\mathbf{H}_t))) \mathbb{1}_{\{W_{t+1}=j\}} \right] \\
&= \sum_{l=0}^{K-1} \int \log(q^l(\Phi(\mathbf{H}_t))) \mathbb{1}_{\{W_{t+1}=j\}} p(W_{t+1} \mid \Phi(\mathbf{H}_t)) p(\Phi(\mathbf{H}_t)) dW_{t+1} d\Phi(\mathbf{H}_t) \\
&= \sum_{l=0}^{K-1} \int \log(q^l(\Phi(\mathbf{H}_t))) p(W_{t+1} = l \mid \Phi(\mathbf{H}_t)) p(\Phi(\mathbf{H}_t)) d\Phi(\mathbf{H}_t) \\
&= \sum_{l=0}^{K-1} \int \log(q^l(\Phi(\mathbf{H}_t))) p(\Phi(\mathbf{H}_t) \mid W_{t+1} = l) p(W_{t+1} = l) d\Phi(\mathbf{H}_t)
\end{aligned}$$

Let's denote  $\alpha_l = p(W_{t+1} = l)$ , the marginal probability of observing the  $l$ -th treatment regime, and  $p_l^\Phi(\mathbf{H}_t) = p(\Phi(\mathbf{H}_t) \mid W_{t+1} = l)$  with a corresponding probability distribution  $\mathbb{P}_l^\Phi$ . We intend to maximize point-wise the objective in Eq. (21). Plugging the latter formulation of the conditional likelihood in Eq. (21) and writing the Lagrangian function, we get

$$\max_q \sum_{l=0}^{K-1} \log(q^l(\Phi(\mathbf{H}_t))) p_j^\Phi(\mathbf{H}_t) \alpha_l + \lambda \left( \sum_{l=0}^{K-1} q^l(\Phi(\mathbf{H}_t)) - 1 \right) \quad (22)$$

Computing the gradient w.r.t  $q^l(\Phi(\mathbf{H}_t))$  for  $l \in \{0, 1, \dots, K-1\}$  and setting to zero, we have

$$q^{l,*}(\Phi(\mathbf{H}_t)) = -\frac{\alpha_l p_j^\Phi(\mathbf{H}_t)}{\lambda} \quad (23)$$

Then, by the equality constraint, we find that  $\lambda = -\sum_{l=0}^{K-1} \alpha_l p_j^\Phi(\mathbf{H}_t)$ .  $\square$

**Proposition G.4.** Let  $\Phi$  be a fixed representation function. The  $I_{CLUB}$  objective when the treatment prediction function is optimal (i.e.  $q = q^*$ ) has the following form:

$$I_{CLUB} = \sum_{j=0}^{K-1} \alpha_j D_{KL}(\mathbb{P}_j^\Phi \parallel \sum_{l=0}^{K-1} \alpha_l \mathbb{P}_l^\Phi) + \mathbb{E}_{\mathbb{P}_{\Phi(\mathbf{H}_t)}} [D_{KL}(\mathbb{P}_{W_{t+1}} \parallel \mathbb{P}_{W_{t+1} \mid \Phi(\mathbf{H}_t)})] \quad (24)$$

*Proof.* First, recall that

$$\begin{aligned}
I_{CLUB}(\Phi(\mathbf{H}_t), W_{t+1}; q^*) &= \mathbb{E}_{\mathbb{P}(\Phi(\mathbf{H}_{t+1}), W_{t+1})} [\log q^*(W_{t+1} \mid \Phi(\mathbf{H}_{t+1}))] - \mathbb{E}_{\mathbb{P}_{\Phi(\mathbf{H}_{t+1})}} \mathbb{E}_{\mathbb{P}_{W_{t+1}}} (\log q^*(W_{t+1} \mid \Phi(\mathbf{H}_{t+1}))) \\
I_{CLUB}(\Phi(\mathbf{H}_t), W_{t+1}; q^*) &= A - B
\end{aligned}$$

Let's detail  $A$  and  $B$  separately,

$$\begin{aligned}
A &= \sum_{j=0}^{K-1} \int \alpha_j \log(q^{j,*}(\Phi(\mathbf{H}_t))) p_j^\Phi(\mathbf{H}_t) d\Phi(\mathbf{H}_t) \\
&= \sum_{j=0}^{K-1} \int \alpha_j \log\left(\frac{\alpha_j p_j^\Phi(\mathbf{H}_t)}{\sum_{l=0}^{K-1} p_l^\Phi(\mathbf{H}_t) \alpha_l}\right) p_j^\Phi(\mathbf{H}_t) d\Phi(\mathbf{H}_t) \\
&= \sum_{j=0}^{K-1} \int \alpha_j \log\left(\frac{p_j^\Phi(\mathbf{H}_t)}{\sum_{l=0}^{K-1} p_l^\Phi(\mathbf{H}_t) \alpha_l}\right) p_j^\Phi(\mathbf{H}_t) d\Phi(\mathbf{H}_t) + \log(\alpha_j) \alpha_j \\
&= \sum_{j=0}^{K-1} \alpha_j D_{KL}(\mathbb{P}_j^\Phi \parallel \sum_{l=0}^{K-1} \alpha_l \mathbb{P}_l^\Phi) + \sum_{j=0}^{K-1} \log(\alpha_j) \alpha_j
\end{aligned}$$

Finally, we can write

$$A = \sum_{j=0}^{K-1} \alpha_j D_{KL}(\mathbb{P}_j^\Phi \parallel \sum_{l=0}^{K-1} \alpha_l \mathbb{P}_l^\Phi) - H(W_{t+1}) \quad (25)$$

For the remaining term  $B$ , we have

$$\begin{aligned}
B &= \mathbb{E}_{\mathbb{P}_{\Phi(\mathbf{H}_t)}} \mathbb{E}_{\mathbb{P}_{W_{t+1}}} (\log q^*(W_{t+1} \mid \Phi(\mathbf{H}_{t+1}))) \\
&= \sum_{j=0}^{K-1} \mathbb{E}_{\mathbb{P}_{\Phi(\mathbf{H}_t)}} \mathbb{E}_{\mathbb{P}_{W_{t+1}}} [\log(q^j(\Phi(\mathbf{H}_t))) \mathbb{1}_{\{W_{t+1}=j\}}] \\
&= \sum_{j=0}^{K-1} \mathbb{E}_{\mathbb{P}_{\Phi(\mathbf{H}_t)}} [\alpha_j \log(q^j(\Phi(\mathbf{H}_t)))] \\
&= \sum_{j=0}^{K-1} \alpha_j \int \log \left[ \frac{\alpha_j p_j^\Phi(\mathbf{H}_t)}{\sum_{l=0}^{K-1} p_l^\Phi(\mathbf{H}_t) \alpha_l} \right] p(\Phi(\mathbf{H}_t)) d\Phi(\mathbf{H}_t) \\
&= \sum_{j=0}^{K-1} \alpha_j \int \log \left[ \frac{p(\Phi(\mathbf{H}_t))}{\sum_{l=0}^{K-1} p_l^\Phi(\mathbf{H}_t) \alpha_l} \frac{p(W_{t+1}=j \mid \Phi(\mathbf{H}_t))}{p(W_{t+1}=j)} \right] p(\Phi(\mathbf{H}_t)) d\Phi(\mathbf{H}_t) \\
&\quad - H(W_{t+1}) \\
&= \sum_{j=0}^{K-1} \alpha_j \underbrace{\int \log \left[ \frac{p(\Phi(\mathbf{H}_t))}{\sum_{l=0}^{K-1} p_l^\Phi(\mathbf{H}_t) \alpha_l} \right] p(\Phi(\mathbf{H}_t)) d\Phi(\mathbf{H}_t)}_{=0} \\
&\quad + \sum_{j=0}^{K-1} \alpha_j \int \log \left[ \frac{p(W_{t+1}=j \mid \Phi(\mathbf{H}_t))}{p(W_{t+1}=j)} \right] p(\Phi(\mathbf{H}_t)) d\Phi(\mathbf{H}_t) - H(W_{t+1})
\end{aligned}$$

The final form of  $B$  is therefore

$$B = - \int D_{KL}(\mathbb{P}_{W_{t+1}} \parallel \mathbb{P}_{W_{t+1} \mid \Phi(\mathbf{H}_t)}) p(\Phi(\mathbf{H}_t)) d\Phi(\mathbf{H}_t) - H(W_{t+1}) \quad (26)$$

The proposition follows immediately from Equations (25) and (26).  $\square$

*Proof.* (Theorem 5.4) Since by lemma G.3, the  $I_{CLUB}$  formulation in proposition G.4 holds, then to prove that the representation is balanced, it is enough to see that by the positivity of  $D_{KL}$

$$I_{CLUB} \geq \mathbb{E}_{\mathbb{P}_{\Phi(\mathbf{H}_t)}} [D_{KL}(\mathbb{P}_{W_{t+1}} \parallel \mathbb{P}_{W_{t+1} \mid \Phi(\mathbf{H}_t)})] \geq 0 \quad (27)$$

$I_{CLUB}$  is minimal when  $I_{CLUB} = 0$ , which happens if and only if for  $j \in \{0, 1, \dots, K-1\}$   $p(W_{t+1}=j) = p(W_{t+1}=j \mid \Phi(\mathbf{H}_t))$  almost surely which, by Bayes rule is equivalent to say  $p(\Phi(\mathbf{H}_t)) = p(\Phi(\mathbf{H}_t) \mid W_{t+1}=j)$ .  $\square$

## H Causal CPC Pseudo algorithm

In this section, we present a detailed overview of the training procedure for Causal CPC. Initially, we train the Encoder using only the contrastive terms, as outlined in Algorithm 1. Our primary objective is to ensure that, for each time step  $t$ , the process history  $\mathbf{H}_t$  is predictive of future local features  $\mathbf{Z}_t$ . However, calculating the InfoNCE loss for a batch across all possible time steps  $t = 0, \dots, t_{\max}$  can be computationally demanding.

To address this, we adopt a more efficient approach by uniformly sampling a single time step  $t$  per batch. Subsequently, the corresponding process history  $\mathbf{H}_t$  is contrasted. The sampled  $\mathbf{H}_t$  is then employed as input for the InfoMax objective and randomly partitioned into future  $\mathbf{H}_t^f$  and past  $\mathbf{H}_t^h$  sub-processes.

The decoder is trained while taking the encoder as input (Algorithm 2), utilizing a lower learning rate compared to the untrained part of the decoder. It is trained autoregressively and without teacher forcing. This implies that for each time step  $t$ , our GRU-based decoder should predict the future sequence of treatments  $\hat{Y}_{t+1:t+\tau}$  with its hidden state initialized to the representation  $\Phi_t$  of the historical process up to time  $t$ .

---

**Algorithm 1** Pretraining of the encoder

---

**Require:** Encoder parameters  $\theta_{1,2,3}$ , learning rate  $\mu$ **Input:** data  $\{\mathbf{H}_{i,t_{max}}, i = 1, \dots, N\}$ **for**  $p \in \{1, \dots, \text{epoch}_{\max}\}$  **do**  **for**  $\mathcal{B} = \{\mathbf{H}_{i,t_{max}}, i = 1, \dots, |\mathcal{B}|\}$  **do**     $\mathbf{Z}_t = \Phi_{\theta_1}([\mathbf{X}_t, W_{t-1}, Y_{t-1}])$  for  $t = 0, \dots, t_{max}$ .    Choose  $t \sim \mathcal{U}([1, t_{max} - 1])$ .    Compute  $\mathbf{C}_t = \Phi_{\theta_1, \theta_2}(\mathbf{H}_t)$ .    Compute  $\mathcal{L}^{CPC}(\theta_1, \theta_2, \{\Gamma_j\}_{j=1}^\tau)$ .    Choose  $t_0 \sim \mathcal{U}([1, t])$ .    Compute  $\mathbf{C}_t^h = \Phi_{\theta_1, \theta_2}(\mathbf{H}_t^h)$ ,  $\mathbf{C}_t^f = \Phi_{\theta_1, \theta_2}(\mathbf{H}_t^f)$ ,    Compute  $\mathcal{L}^{(InfoMax)}(\theta_1, \theta_2, \gamma)$ .

Update parameters

$$\theta_{1,2,3} \leftarrow \theta_{1,2,3} - \mu \left( \frac{\partial \mathcal{L}^{CPC}(\theta_1, \theta_2, \{\Gamma_j\}_{j=1}^\tau)}{\partial \theta_{1,2,3}} + \frac{\partial \mathcal{L}^{(InfoMax)}(\theta_1, \theta_2, \gamma)}{\partial \theta_{1,2,3}} \right)$$

**end for****end for****Return:** Trained encoder.

---

To enhance training efficiency, instead of predicting  $\hat{Y}_{i,t+1:t+\tau}$  for all individuals  $i$  in a batch and for all possible time steps  $t$ , we randomly select  $m$  time indices  $t_{i,1}, \dots, t_{i,m}$  for each individual  $i$ . From these indices, we compute future treatment response sequences  $\hat{Y}_{i,t_{i,1}+1:t_{i,1}+\tau}, \dots, \hat{Y}_{i,t_{i,m}+1:t_{i,m}+\tau}$ . We found that is enough to train while selecting randomly 10% of the time steps.

---

**Algorithm 2** Training of the decoder

---

**Require:** Encoder parameters  $\theta_{1,2,3}$ , Decoder parameters  $\theta_4, \theta_Y, \theta_W$ .**Require:** Encoder learning rate  $\mu_{enc}$ , Treatment learning rate  $\mu_W$ , Outcome learning rate  $\mu_Y$ .**Require:** Number of random time indices  $m$ .**Input:** data  $\{\mathbf{H}_{i,t_{max}}, i = 1, \dots, N\}$ **for**  $p \in \{1, \dots, \text{epoch}_{\max}\}$  **do**  **for**  $\mathcal{B} = \{\mathbf{H}_{i,t_{max}}, i = 1, \dots, |\mathcal{B}|\}$  **do**    Compute  $\mathbf{C}_{i,t} = \text{encoder}(\mathbf{H}_{i,t})$  for  $t = 0, \dots, t_{max}, i = 1, \dots, |\mathcal{B}|$ .    Compute  $\Phi_t = \Phi_{\theta_R}(\mathbf{H}_t)$ .  **for**  $i = 1, \dots, |\mathcal{B}|$  **do**    Choose  $t_{i,1}, \dots, t_{i,m} \sim \mathcal{U}([1, t_{max} - \tau])$ .    **for**  $t \in \{t_{i,1}, \dots, t_{i,m}\}$  **do**      Compute  $\hat{Y}_{i,t+1:t+\tau}, \hat{W}_{i,t+1:t+\tau}, \Phi_{i,t+1:t+\tau-1} = \text{decoder}(\Phi_t, \mathbf{V}_i, W_{i,t}, Y_{i,t}, W_{i,t+1:t+\tau})$     **end for**  **end for**  Compute  $\mathcal{L}_{dec}(\theta_R, \theta_Y, \theta_W)$  and  $\mathcal{L}_W(\theta_W, \theta_R)$ .

Update parameters in the order.

$$\theta_{1,2,3} \leftarrow \theta_{1,2,3} - \mu_{enc} \left( \frac{\partial \mathcal{L}_{dec}(\theta_R, \theta_Y, \theta_W)}{\partial \theta_{1,2,3}} \right)$$

$$\theta_{4,Y} \leftarrow \theta_{4,Y} - \mu_Y \left( \frac{\partial \mathcal{L}_{dec}(\theta_R, \theta_Y, \theta_W)}{\partial \theta_{4,Y}} \right)$$

$$\theta_W \leftarrow \theta_W - \mu_W \left( \frac{\partial \mathcal{L}_W(\theta_W, \theta_R)}{\partial \theta_W} \right)$$

**end for****end for****Return:** Trained decoder.

---

## I Causal CPC: Architecture details

|                                                                                    |
|------------------------------------------------------------------------------------|
| <b>Inputs:</b> $[\mathbf{X}_t, W_{t-1}, Y_{t-1}]$                                  |
| Linear Layer                                                                       |
| WeightNorm                                                                         |
| SELU                                                                               |
| Linear Layer                                                                       |
| WeightNorm                                                                         |
| <b>Outputs:</b> $\mathbf{Z}_t = \Phi_{\theta_1}([\mathbf{X}_t, W_{t-1}, Y_{t-1}])$ |

Table 13: Architecture for learning local features  $\mathbf{Z}_t$

|                                                                                         |
|-----------------------------------------------------------------------------------------|
| <b>Inputs:</b> $\mathbf{Z}_{\leq t}$                                                    |
| GRU (1 layer)                                                                           |
| <b>Outputs:</b> Hidden state $\mathbf{C}_t = \Phi_{\theta_2}^{ar}(\mathbf{Z}_{\leq t})$ |

Table 14: Architecture for learning context representation  $\mathbf{C}_t$

|                                |
|--------------------------------|
| <b>Inputs:</b> $[\Phi_t, W_t]$ |
| Linear Layer                   |
| WeightNorm                     |
| SELU                           |
| Linear Layer                   |
| WeightNorm                     |
| <b>Outputs:</b> $\hat{Y}_t$    |

Table 15: Architecture for outcome prediction

|                             |
|-----------------------------|
| <b>Inputs:</b> $\Phi_t$     |
| Linear Layer                |
| SpectralNorm                |
| SELU                        |
| Linear Layer                |
| SpectralNorm                |
| <b>Outputs:</b> $\hat{W}_t$ |

Table 16: Architecture for treatment prediction

## J Models hyperparameters

In this section, we report the range of all hyperparameters to be fine-tuned, as well as fixed hyperparameters for all models and across the different datasets used in experiments. Best hyperparameter values are reported in the configuration files in the code repository.

Table 17: Hyper-parameters search range for RMSN

| Model                      | Sub-model                    | Hyperparameter             | Cancer simulation          | MIMIC III (SS)             |
|----------------------------|------------------------------|----------------------------|----------------------------|----------------------------|
| RMSNs                      | Propensity Treatment Network | LSTM layers                | 1                          | 1                          |
|                            |                              | Learning rate              | 0.01, 0.005, 0.001, 0.0001 | 0.01, 0.005, 0.001, 0.0001 |
|                            |                              | Batch size                 | 32, 64, 128                | 32, 64, 128                |
|                            |                              | LSTM hidden units          | 4, 6, ..., 12              | 4, 6, ..., 30              |
|                            |                              | LSTM dropout rate          | -                          | -                          |
|                            |                              | Max gradient norm          | 0.5, 1, 2                  | 0.5, 1, 2                  |
|                            |                              | Early Stopping (min delta) | 0.0001                     | 0.0001                     |
|                            |                              | Early Stopping (patience)  | 30                         | 30                         |
| Propensity History Network |                              | LSTM layers                | 1                          | 1                          |
|                            |                              | Learning rate              | 0.01, 0.005, 0.001, 0.0001 | 0.01, 0.005, 0.001, 0.0001 |
|                            |                              | Batch size                 | 32, 64, 128                | 64, 128, 256               |
|                            |                              | LSTM hidden units          | 4, 6, ..., 20              | 4, 6, ..., 30              |
|                            |                              | LSTM dropout rate          | -                          | -                          |
|                            |                              | Early Stopping (min delta) | 0.0001                     | 0.0001                     |
|                            |                              | Early Stopping (patience)  | 30                         | 30                         |
| Encoder                    |                              | LSTM layers                | 1                          | 1                          |
|                            |                              | Learning rate              | 0.01, 0.005, 0.001, 0.0001 | 0.01, 0.005, 0.001, 0.0001 |
|                            |                              | Batch size                 | 32, 64, 128                | 32, 64, 128                |
|                            |                              | LSTM hidden units          | 4, 6, ..., 20              | 4, 6, ..., 30              |
|                            |                              | LSTM dropout rate          | -                          | -                          |
|                            |                              | Early Stopping (min delta) | 0.0001                     | 0.0001                     |
|                            |                              | Early Stopping (patience)  | 30                         | 30                         |
| Decoder                    |                              | LSTM layers                | 1                          | 1                          |
|                            |                              | Learning rate              | 0.01, 0.005, 0.001, 0.0001 | 0.01, 0.005, 0.001, 0.0001 |
|                            |                              | Batch size                 | 32, 64, 128                | 128, 512, 1024             |
|                            |                              | LSTM hidden units          | 4, 6, ..., 20              | 4, 6, ..., 30              |
|                            |                              | LSTM dropout rate          | -                          | -                          |
|                            |                              | Max gradient norm          | 0.5, 1, 2                  | 0.5, 1, 2                  |
|                            |                              | Early Stopping (min delta) | 0.0001                     | 0.0001                     |
|                            |                              | Early Stopping (patience)  | 30                         | 30                         |

Table 18: Hyper-parameters search range for CRN

| Model   | Sub-model | Hyperparameter             | Cancer simulation          | MIMIC III (SS)             |
|---------|-----------|----------------------------|----------------------------|----------------------------|
| CRN     | Encoder   | LSTM layers                | 1                          | 1                          |
|         |           | Learning rate              | 0.01, 0.005, 0.001, 0.0001 | 0.01, 0.005, 0.001, 0.0001 |
|         |           | Batch size                 | 32, 64, 128                | 32, 64, 128                |
|         |           | LSTM hidden units          | 4, 6, ..., 30              | 4, 6, ..., 30              |
|         |           | LSTM dropout rate          | -                          | -                          |
|         |           | BR size                    | 4, 6, ..., 20              | 4, 6, ..., 30              |
|         |           | Early Stopping (min delta) | 0.0001                     | 0.0001                     |
|         |           | Early Stopping (patience)  | 30                         | 30                         |
| Decoder |           | LSTM layers                | 1                          | 1                          |
|         |           | Learning rate              | 0.01, 0.005, 0.001, 0.0001 | 0.01, 0.005, 0.001, 0.0001 |
|         |           | Batch size                 | 128, 256, 512              | 256, 512, 1024             |
|         |           | LSTM hidden units          | 4, 6, ..., 30              | 4, 6, ..., 30              |
|         |           | LSTM dropout rate          | -                          | -                          |
|         |           | BR size                    | 4, 6, ..., 20              | 4, 6, ..., 30              |
|         |           | Early Stopping (min delta) | 0.0001                     | 0.0001                     |
|         |           | Early Stopping (patience)  | 30                         | 30                         |

Table 19: Hyper-parameters search range for G-Net

| Hyperparameter             | Cancer simulation          | MIMIC III (SS)             |
|----------------------------|----------------------------|----------------------------|
| LSTM layers                | 1                          | 1                          |
| Learning rate              | 0.01, 0.005, 0.001, 0.0001 | 0.01, 0.005, 0.001, 0.0001 |
| Batch size                 | 32, 64, 128                | 32, 64, 128                |
| LSTM hidden units          | 4, 6, ..., 30              | 4, 6, ..., 30              |
| FC hidden units            | 4, 6, ..., 30              | 4, 6, ..., 30              |
| LSTM dropout rate          | -                          | -                          |
| R size                     | 4, 6, ..., 20              | 4, 6, ..., 30              |
| MC samples                 | 10                         | 10                         |
| Early Stopping (min delta) | 0.0001                     | 0.0001                     |
| Early Stopping (patience)  | 30                         | 30                         |

Table 20: Hyper-parameters search range for Causal Transformer

| Hyperparameter             | Cancer simulation          | MIMIC III (SS)             |
|----------------------------|----------------------------|----------------------------|
| Transformer blocks         | 1                          | 1                          |
| Learning rate              | 0.01, 0.005, 0.001, 0.0001 | 0.01, 0.005, 0.001, 0.0001 |
| Batch size                 | 32, 64, 128                | 32, 64, 128                |
| Attention heads            | 2                          | 2                          |
| Transformer units          | 4, 6, ..., 20              | 4, 6, ..., 20              |
| LSTM dropout rate          | -                          | -                          |
| BR size                    | 4, 6, ..., 20              | 4, 6, ..., 20              |
| FC hidden units            | 4, 6, ..., 20              | 4, 6, ..., 20              |
| Sequential dropout rate    | 0.1, 0.2, 0.3              | 0.1, 0.2, 0.3              |
| Max positional encoding    | 15                         | 15                         |
| Early Stopping (min delta) | 0.0001                     | 0.0001                     |
| Early Stopping (patience)  | 30                         | 30                         |

Table 21: Hyper-parameters search range for Causal CPC

| Model      | Sub-model | Hyperparameter                                    | Cancer simulation              | MIMIC III (SS)                 |
|------------|-----------|---------------------------------------------------|--------------------------------|--------------------------------|
| Causal CPC | Encoder   | GRU layers                                        | 1                              | 1                              |
|            |           | Learning rate                                     | 0.01, 0.005, 0.001, 0.0001     | 0.01, 0.005, 0.001, 0.0001     |
|            |           | Batch size                                        | 32, 64, 128                    | 64, 128, 256                   |
|            |           | GRU hidden units                                  | 4, 6, ..., 30                  | 4, 6, ..., 30                  |
|            |           | GRU dropout rate                                  | -                              | -                              |
|            |           | Local features (LF) size                          | 4, 6, ..., 20                  | 4, 6, ..., 20                  |
|            |           | Context Representation (CR) size                  | 4, 6, ..., 20                  | 4, 6, ..., 20                  |
|            |           | Early Stopping (min delta)                        | 0.001                          | 0.001                          |
|            |           | Early Stopping (patience)                         | 100                            | 100                            |
| Decoder    |           | GRU layers                                        | 1                              | 1                              |
|            |           | Learning rate (decoder w/o treatment sub-network) | 0.01, 0.005, 0.001, 0.0001     | 0.01, 0.005, 0.001, 0.0001     |
|            |           | Learning rate (encoder fine-tuning)               | 0.001, 0.0005, 0.0001, 0.00005 | 0.001, 0.0005, 0.0001, 0.00005 |
|            |           | Learning rate (treatment sub-network)             | 0.05, 0.01, 0.005, 0.0001      | 0.05, 0.01, 0.005, 0.0001      |
|            |           | Batch size                                        | 32, 64, 128                    | 32, 64, 128                    |
|            |           | GRU hidden units                                  | CR size                        | CR size                        |
|            |           | GRU dropout rate                                  | -                              | -                              |
|            |           | BR size                                           | CR size                        | CR size                        |
|            |           | GRU layers (Treat Encoder)                        | 1                              | 1                              |
|            |           | GRU hidden units (Treat Encoder)                  | 6                              | 6                              |
|            |           | FC hidden units                                   | 4, 6, ..., 20                  | 4, 6, ..., 20                  |
|            |           | Random time indices (m)                           | 10%                            | 10%                            |
|            |           | Early Stopping (min delta)                        | 0.001                          | 0.001                          |
|            |           | Early Stopping (patience)                         | 50                             | 50                             |

Table 22: Hyper-parameters search range for Causal CPC

| Model      | Sub-model | Hyperparameter                                    | Cancer simulation              | MIMIC III (SS)                 |
|------------|-----------|---------------------------------------------------|--------------------------------|--------------------------------|
| Causal CPC | Encoder   | GRU layers                                        | 1                              | 1                              |
|            |           | Learning rate                                     | 0.01, 0.005, 0.001, 0.0001     | 0.01, 0.005, 0.001, 0.0001     |
|            |           | Batch size                                        | 32, 64, 128                    | 64, 128, 256                   |
|            |           | GRU hidden units                                  | 4, 6, ..., 30                  | 4, 6, ..., 30                  |
|            |           | GRU dropout rate                                  | -                              | -                              |
|            |           | Local features (LF) size                          | 4, 6, ..., 20                  | 4, 6, ..., 20                  |
|            |           | Context Representation (CR) size                  | 4, 6, ..., 20                  | 4, 6, ..., 20                  |
|            |           | Early Stopping (min delta)                        | 0.001                          | 0.001                          |
|            |           | Early Stopping (patience)                         | 100                            | 100                            |
| Decoder    |           | GRU layers                                        | 1                              | 1                              |
|            |           | Learning rate (decoder w/o treatment sub-network) | 0.01, 0.005, 0.001, 0.0001     | 0.01, 0.005, 0.001, 0.0001     |
|            |           | Learning rate (encoder fine-tuning)               | 0.001, 0.0005, 0.0001, 0.00005 | 0.001, 0.0005, 0.0001, 0.00005 |
|            |           | Learning rate (treatment sub-network)             | 0.05, 0.01, 0.005, 0.0001      | 0.05, 0.01, 0.005, 0.0001      |
|            |           | Batch size                                        | 32, 64, 128                    | 32, 64, 128                    |
|            |           | GRU hidden units                                  | CR size                        | CR size                        |
|            |           | GRU dropout rate                                  | -                              | -                              |
|            |           | BR size                                           | CR size                        | CR size                        |
|            |           | GRU layers (Treat Encoder)                        | 1                              | 1                              |
|            |           | GRU hidden units (Treat Encoder)                  | 6                              | 6                              |
|            |           | FC hidden units                                   | 4, 6, ..., 20                  | 4, 6, ..., 20                  |
|            |           | Random time indices (m)                           | 10%                            | 10%                            |
|            |           | Early Stopping (min delta)                        | 0.001                          | 0.001                          |
|            |           | Early Stopping (patience)                         | 50                             | 50                             |

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The paper introduces a novel method combining RNNs with CPC for long-term counterfactual regression, leveraging MI objectives for efficient representation learning, and demonstrates state-of-the-art results on both synthetic and real-world data. These claims are substantiated by the detailed theoretical 5 and empirical analyses 6 provided in the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: While Causal CPC excels at large horizon predictions, it does not outperform SOTA models on short-term predictions (Table 1).

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Detailed proofs are provided in Appendix G.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: All the experimental details related to the training and evaluation protocol (D), datasets descriptions (E.1 and F.1) are provided. A detailed description of the Causal CPC architecture is provided in I. Pseudo-algorithms for both the encoder and the decoder are also provided in H. Code is provided in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: Code is provided in the supplementary material at submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: Models' hyperparameter search range is provided in Appendix J, the selection method is provided at the beginning of Section 6, and remaining details about training and testing are provided in the experimental protocol (Appendix D).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[No\]](#)

Justification: It is computationally demanding to compute errors bars for all neural network models in our benchmark. However, we reported the mean and standard deviation of metrics for each experiment, computed from multiple runs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: We provided the computation resources used in the title of Table 2 as well as the time of execution in the same table for cancer simulation data. A similar table is provided in Appendix F.2 for MIMIC III data (Table 12).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Authors acknowledge conducting research in conformity with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: An impact statement is included in Appendix A.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Authors of models and datasets are appropriately cited in the paper in the introduction (1) and experiment (6) sections. Original owners of some model implementations are properly credited in our code.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Code for Causal CPC is provided in the supplementary material.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

#### 14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

#### 15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.