Adversary-Free Counterfactual Prediction via Information-Regularized Representations

Shiqin Tang[*,a] **Rong Feng**[*,a,b] **Shuxin Zhuang**[a,b] **Youzhi Zhang**[a] **Hongzong Li**[†,c]
[a] CAIR-CAS [b] CityUHK [c] HKUST

Abstract

We study counterfactual prediction under assignment bias and propose a mathematically grounded, information-theoretic approach that removes treatment–covariate dependence without adversarial training. Starting from a bound that links the counterfactual–factual risk gap to mutual information, we learn a stochastic representation Z that is predictive of outcomes while minimizing $I(Z;T)$. We derive a tractable variational objective that upper-bounds the information term and couples it with a supervised decoder, yielding a stable, provably motivated training criterion. The framework extends naturally to dynamic settings by applying the information penalty to sequential representations at each decision time. We evaluate the method on controlled numerical simulations and a real-world clinical dataset, comparing against recent state-of-the-art balancing, reweighting, and adversarial baselines. Across metrics of likelihood, counterfactual error, and policy evaluation, our approach performs favorably while avoiding the training instabilities and tuning burden of adversarial schemes.

1 INTRODUCTION

Causal effect estimation addresses the "what if" (counterfactual) question: *What outcome would have occurred for a patient, possibly contrary to fact, had the patient received a different treatment?* In real-world (observational) data, however, who receives a treatment is rarely random. For example, clinicians may preferentially administer a medication to patients with more severe symptoms, while patients with milder illness are less likely to receive it. If we then compare outcomes between the treated (on average sicker) and untreated (on average healthier) groups, the difference reflects both the medication's effect and the baseline severity gap; as a result, the effect we would predict for mildly ill patients can be biased. This systematic distortion—arising because treatment choice depends on prognostic factors—is often called *assignment bias* (a form of confounding) (Rosenbaum and Rubin, 1983; Robins, 1986; Robins et al., 2000).

Existing solutions. A large body of work mitigates such bias via balanced representations (Johansson et al., 2016; Shalit et al., 2017), adversarial domain-invariance (Ganin et al., 2016; Kazemi and Ester, 2024), and reweighting or orthogonalization (Hainmueller, 2012; Huang et al., 2006; Chernozhukov et al., 2018). Yet, adversarial approaches often introduce instability and substantial tuning overhead, and balancing via integral probability metrics can become brittle in complex, high-dimensional settings.

An information-theoretic lens. We offer a unifying perspective that explains why these strategies help. The key idea is to view bias reduction through the amount of information covariates carry about treatment choice. When the features of an individual strongly reveal which treatment they received, treated and untreated groups can look systematically different, and naive comparisons are unreliable. By contrast, when a learned representation removes or attenuates this treatment-revealing signal—while still keeping what matters for predicting outcomes—treated and untreated cases become more comparable, and effect estimates improve. This information-theoretic viewpoint clarifies the goals underlying balancing and domain invariance and provides principled guidance for how to build representations that support counterfactual prediction.

* Equal contribution. † Correspondence to: Hongzong Li <lihongzong@ust.hk>.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

Our approach in a nutshell. We introduce an information-theoretic representation learning method that explicitly discourages treatment-revealing features without relying on adversarial training. Concretely, we learn a stochastic representation that is predictive of outcomes yet carries little mutual information with treatment assignment. This direct control of the dependence between representation and treatment provides a simple, stable alternative to adversarial objectives, and it can be seamlessly extended to the dynamic case with time-varying covariates. We refer to the static model as the **S**tatic **I**nformation-theoretic **C**ounterfactual **E**stimator (*SICE*) and its sequential extension as the **D**ynamic **I**nformation-theoretic **C**ounterfactual **E**stimator (*DICE*).

Paper structure. Section 2 lays the foundation by formalizing the setting and assumptions, and develops risk–gap bounds that motivate our approach. Section 3 and 4 respectively introduce the proposed *SICE* and *DICE* models. Section 5 presents experiments on controlled simulations and a real clinical dataset, followed by discussion and limitations. Supplementary proofs and additional experimental results are deferred to the appendix.

Contributions. ① We present an information-theoretic formulation that unifies balancing, reweighting, and invariance by linking the risk gap to treatment–covariate dependence. ② We propose two models – *SICE* (static) and *DICE* (sequential) – that avoid adversarial training by directly approximating the mutual-information term. ③ In both controlled simulations and real data, we vary treatment dimensionality from small to ultra–high and show the proposed models remain competitive as dimensionality grows.

2 BACKGROUND

We first fix notation and explicitly state the assumptions and conditions we rely on, so that the learning problem is pinpointed precisely and the regimes in which our method applies are rigorously defined.

Notations. We denote covariates (or confounders) by X (e.g., a patient's profile), treatment by T, and outcome by Y, with $X \in \mathcal{X}$, $T \in \mathcal{T}$, and $Y \in \mathcal{Y}$.

Associational and causal estimands. In causal inference, we are mostly interested in establishing the following models (Neuberg, 2003; Hern'an and Robins, 2020):

$$m_t(x) := \mathbb{E}[Y \mid T = t, X = x] \tag{1}$$

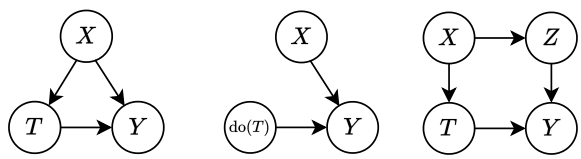
$$g_t(x) := \mathbb{E}[Y \mid \mathrm{do}(T = t), X = x] \tag{2}$$

where $m_t(x)$ refers to the expected values of Y among all subjects with $(X, T) = (x, t)$, and $g_t(x)$ represents a hypothetical average value of Y among all subjects with $X = x$ intervened on treatment t (no matter what treatments they actually receive); we refer to m_t and g_t as associational and causal outcome models, respectively. The associational expectation (1) cannot be used for decision making because, for instance, if a certain pattern x is always associated with treatment t, then the model never observes outcomes under any alternative t' for those x. Any prediction at $X = x$ and $T = t'$ is pure extrapolation driven only by parametric/inductive assumptions, not data. Hence plug-in decisions based on those counterfactual predictions can be arbitrarily wrong. The causal expectation model (2), on the other hand, can be used to facilitate decision-making, but it is difficult to estimate, especially without extra information like instrumental variables. Therefore, our optimal strategy here is to consider the scenarios where the two models coincide.

Potential outcome framework. We adopt the Neyman–Rubin framework: for $t \in \mathcal{T}$, let $Y(t)$ denote the potential outcome had treatment t been assigned. We assume the standard identification conditions: (i) *Consistency/SUTVA* (the observed outcome under the received treatment equals the corresponding potential outcome, with well-defined interventions); (ii) *Positivity/overlap* ($0 < p(T = t \mid X = x)$ on the support of (X, t)); and (iii) *Conditional ignorability* ($Y(t) \perp\!\!\!\perp T \mid X$). Under the consistency and ignorability assumptions, the following identity holds (Robins, 1986; Neuberg, 2003; Hern'an and Robins, 2020):

$$\begin{aligned}\mathbb{E}[Y(t)|X = x] &:= \mathbb{E}[Y|\mathrm{do}(T = t), X = x] \\ &= \mathbb{E}[Y|T = t, X = x].\end{aligned} \tag{3}$$

The *structural causal model (SCM)* in Figure 1 (a) aligns with the settings in most clinical research as the covariates X causally affect both the treatment T and the outcome Y (e.g. patients who receive transfusions are often clinically sicker). If we further assume that X and T are independent, i.e., $p(x|t) = p(x)$, we obtain the setting of a *randomized controlled trial (RCT)*, as illustrated in Figure 1 (b). RCTs are regarded as the gold standard for causal inference, but they can be very costly to conduct and, in some cases, ethically problematic. In the remainder of this section, we discuss how to identify and adjust for the bias introduced by the causal dependence between X and T.



(a) (b) (c)

Figure 1: Structural causal models: (a) an SCM with covariates satisfying the ignorability assumption, (b) randomized controlled trial (RCT), (c) Z as a learned representation of X.

Identifying the bias. Assume consistency, positivity, and ignorability hold, and let $T \sim \pi$ denote the treatment with law π. Write $p_t(x) := p(x|T = t)$ for the covariate law conditional on $T = t$, and $p_x(x) := \int p_t(x)\pi(dt)$ for the marginal of X. Fix a predictor $g_t : \mathcal{X} \to \mathcal{Y}$ for arm t and define the (deterministic) loss profile as

$$\begin{aligned} \phi_t(x) &:= \mathbb{E}\big[L\big(Y(t), g_t(x)\big) \,\big|\, X = x\big] \\ &= \int L(y, g_t(x))p(y|x, \mathrm{do}(T = t))dy. \end{aligned} \tag{4}$$

We define the factual and cross-arm risks for arm t as

$$R_t^F := \mathbb{E}_{p(x|t)}[\phi_t(x)], \quad R_{t\to t'}^{CF} := \mathbb{E}_{p(x|t')}[\phi_t(x)]. \tag{5}$$

We also define the counterfactual (marginal) risk as

$$R_t^{CF} := \mathbb{E}_{\pi(t')}[R_{t\to t'}^{CF}] = \mathbb{E}_{p_x(x)}[\phi_t(x)], \tag{6}$$

and the population aggregates $R^F := \mathbb{E}_{t\sim\pi}[R_t^F]$ and $R^{CF} := \mathbb{E}_{t\sim\pi}[R_t^{CF}]$. Let $\mathcal{F}$ be a function class and suppose there exists $\lambda > 0$ such that $\phi_t/\lambda \in \mathrm{conv}(\mathcal{F})$ for all $t \in \mathcal{T}$. Then, for any $t, t' \in \mathcal{T}$,

$$\begin{aligned} \big|R_{t\to t'}^{CF} - R_t^F\big| &= \big|\mathbb{E}_{p_{t'}}[\phi_t(x)] - \mathbb{E}_{p_t}[\phi_t(x)]\big| \\ &\le \lambda\, \mathrm{IPM}_{\mathcal{F}}(p_{t'}, p_t), \end{aligned} \tag{7}$$

where $\mathrm{IPM}_{\mathcal{F}}(p, q) := \sup_{f\in\mathcal{F}} \big|\mathbb{E}_p[f(x)] - \mathbb{E}_q[f(x)]\big|$. Averaging over $t' \sim \pi$ yields

$$\big|R_t^{CF} - R_t^F\big| \le \lambda \int \mathrm{IPM}_{\mathcal{F}}(p_{t'}, p_t)\, \pi(dt'). \tag{8}$$

A further average over $t \sim \pi$ gives the population bound (Kazemi and Ester, 2024)

$$R^{CF} \le R^F + \lambda \iint \mathrm{IPM}_{\mathcal{F}}(p_t, p_{t'})\pi(dt)\pi(dt'). \tag{9}$$

In particular, if the conditional covariate laws $\{p_t\}_{t\in\mathcal{T}}$ are close in the $\mathrm{IPM}_{\mathcal{F}}$ sense, the counterfactual risk cannot exceed the factual risk by much.

Relations with information theory. Given the function class $\mathcal{F} = \{f : \|f\|_\infty \le 1\}$, we have $\mathrm{IPM}_{\mathcal{F}}(p_t, p_{t'}) = 2\mathrm{TV}(p_t, p_{t'})$. Based on (9), we have

$$\begin{aligned} R^{CF} &\le R^F + 2\lambda \iint \mathrm{TV}(p_t, p_{t'})\pi(dt)\,\pi(dt') \\ &\overset{(a)}{\le} R^F + 4\lambda \int \mathrm{TV}(p_t, p_x)\pi(dt) \\ &\overset{(b)}{\le} R^F + 2\sqrt{2}\lambda \int \sqrt{D_{\mathrm{KL}}(p_t\|p_x)}\pi(dt) \\ &\overset{(c)}{\le} R^F + 2\sqrt{2}\lambda\sqrt{\mathbb{E}_{\pi(t)}[D_{\mathrm{KL}}(p(x|t)\|p(x))]} \\ &= R^F + 2\sqrt{2}\lambda\sqrt{I(X;T)}, \end{aligned} \tag{10}$$

where inequalities (a), (b), and (c) are due to triangular inequality, Pinsker's inequality, and Jensen's inequality. Thus, reducing the mutual information $I(X;T)$ tightens the counterfactual–factual risk gap.

Related works. The bound established in (10) motivates the learning of representations $Z = \phi(X)$ that support factual prediction and bring the treated/control representation distributions closer. In the static setting, balanced-representation methods such as CFR/TARNet explicitly penalize a discrepancy (e.g., $\mathrm{disc}_{\mathcal{H}}$ or MMD) between $p(Z \mid T{=}1)$ and $p(Z \mid T{=}0)$ while fitting outcomes, thereby shrinking the bound (10) (Johansson et al., 2016; Shalit et al., 2017). Beyond these, we highlight two complementary strategies that fit naturally within our bound:

(a) **Information-regularized representations.** Learn features $Z = \phi(X)$ that are predictive for Y yet uninformative about T by solving

$$\begin{aligned} &\min_{\phi}\ \mathbb{E}\big[L\big(y, g(\phi(x), t)\big)\big] + \lambda\, I\big(T; \phi(X)\big) \\ &\propto \min_{\phi}\ \max_{\theta}\ \mathbb{E}\big[L\big(y, g(\phi(x), t)\big)\big] + \lambda\mathbb{E}\big[\log p_\theta(t|\phi(x))\big], \end{aligned} \tag{11}$$

where a variational classifier $p_\theta(t \mid \phi(x))$ provides a tractable surrogate for $I(T; Z)$.

(b) **Treatment-invariant joint encodings.** Learn a joint feature $Z = \phi(X, T)$ and enforce invariance across (possibly clustered) treatments by adversarially discouraging a discriminator from predicting the treatment cluster c (Berrevoets et al., 2020):

$$\begin{aligned} &\min_{\phi}\ \mathbb{E}\big[L\big(y, (g\circ\phi)(x, t)\big)\big] + \lambda\, I\big(C; \phi(X, T)\big) \\ &\propto \min_{\phi}\ \max_{\theta}\ \mathbb{E}\big[L\big(y, (g\circ\phi)(x, t)\big)\big] \\ &\qquad + \lambda\, \mathbb{E}\big[\log p_\theta\big(c \mid \phi(x, t)\big)\big]. \end{aligned} \tag{12}$$

Pushing $p(\phi(X, T) \mid C{=}c)$ together across clusters contracts the discrepancy between treated/-control representations and likewise tightens the bound.

Orthogonal lines include reweighting and orthogonalization—IPTW/stabilized weights, entropy balancing/kernel mean matching, and double/debiased ML—which address assignment bias via importance weighting or Neyman-orthogonal scores rather than representation constraints (Rosenbaum and Rubin, 1983; Robins et al., 2000; Hainmueller, 2012; Huang et al., 2006; Chernozhukov et al., 2018). In the *dynamic setting* with time-dependent confounding, representation learning is extended along trajectories (e.g., CRN, Causal Transformer) to maintain balance sequentially (Bica et al., 2020; Melnychuk et al., 2022b).

3 SICE FORMULATION

Let Z be a representation of X via a stochastic map $z \sim q_\phi(z|x)$; thus $T \leftarrow X \rightarrow Z$ is a Markov structure. For any treatment t, the induced conditional law of Z is $p(z|t) = \int p(x|t) q_\phi(z|x) dx$, and if $T \sim \pi$ we write the T-marginalized representation distribution as $p_z(z) = \int p(z|t)\pi(dt)$. We redefine the loss profile in (4) as

$$\phi_t(z) := \mathbb{E}[L(Y(t), g_t(z))|Z = z]. \tag{13}$$

and the risks as

$$R^F = \mathbb{E}_{\pi(t)p(z|t)}[\phi_t(z)], \quad R^{CF} = \mathbb{E}_{\pi(t)p_z(z)}[\phi_t(z)]. \tag{14}$$

By adapting the argument in Sec. 2 to the representation space and comparing $p(z|t)$ with the marginal $p_z(z)$, we obtain

$$R^{CF} \leq R^F + \sqrt{2}\lambda\sqrt{I(Z;T)} \approx \hat{R}^F + \sqrt{2}\lambda\sqrt{I(Z;T)}, \tag{15}$$

where due to consistency

$$\hat{R}^F = \mathbb{E}_{p_\mathcal{D}(x,y,t)q_\phi(z|x)}[L(y, g_t(z))]. \tag{16}$$

By minimizing the bound (15), we compress away T-predictive idiosyncrasies while retaining outcome-relevant signal via the decoder term, encouraging Z to retain outcome-relevant information while reducing treatment-revealing information. The proposed model is illustrated in Figure 1 (c) with a Markov chain $X \rightarrow Z \rightarrow Y$.

Assumptions for SICE. We retain Consistency/-SUTVA. Overlap is required on the representation: there exists $\delta \in (0,1)$ such that for all (z,t) with $p_Z(z) > 0$, $\delta \leq p(T = t|Z = z) \leq 1 - \delta$. The ignorability assumption is updated as $Y(t) \perp\!\!\!\perp T|Z$, for all $t \in \mathcal{T}$. We assume $Z = \phi(X)$ is outcome-sufficient and confounding-sufficient so that $Y(t) \perp\!\!\!\perp T|Z$. We enforce approximate treatment-agnosticity via the MI regularizer to reduce assignment bias, while preserving outcome signal via the decoder loss.

Adjusting for the bias. Motivated by Eq. (10), we seek a representation that is maximally predictive for Y while minimally informative about T. Let $z \sim q_\phi(z \mid x)$ be a stochastic encoder and write expectations over $(x,y,t) \sim p_\mathcal{D}$ and $z \sim q_\phi(\cdot|x)$. We optimize

$$\min \quad \mathbb{E}[L(y, g(z,t))] + \lambda I(t;z) \tag{17}$$

Following the approach of (Moyer et al., 2018), we decompose the mutual information as

$$I(z;t) = I(z;t|x) + I(z;x) - I(z;x|t), \tag{18}$$

where

$$I(z;t|x) = H(z|x) - H(z|x,t) = 0,$$

where $H(z|x,t) = H(z|x)$ because $p(z|x,t) = p(z|x)$; this can also be inferred by noting the conditional independence between z and t given x. We now focus on lower-bounding $I(z;x|t)$,

$$\begin{aligned} I(z;x|t) &= H(x|t) - H(x|z,t) \\ &= H(x|t) + \mathbb{E}[\log p(x|z,t)] \\ &\geq \mathbb{E}[\log p_\psi(x|z,t)] + \text{Const.} \end{aligned} \tag{19}$$

Similarly to variational information bottleneck (VIB) (Alemi et al., 2016), we can upper-bound $I(x;z)$ as

$$I(x;z) \leq \mathbb{E}[\log q_\phi(z|x) - \log r(z)]. \tag{20}$$

Therefore, we have

$$I(z;t) \leq \mathbb{E}[\log q_\phi(z|x) - \log r(z) - \log p_\psi(x|z,t)] + c. \tag{21}$$

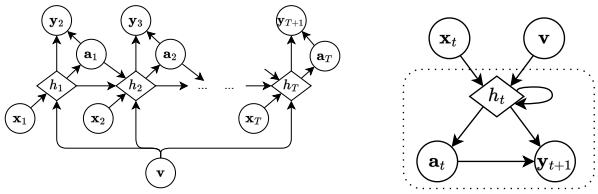
Putting it all together, we have the formulation,

$$\begin{aligned} \min_{g,\phi,\psi} \quad & \mathbb{E}_{p_\mathcal{D}(x,y,t)q_\phi(z|x)}\big[L(y, g(z,t)) - \lambda \log p_\psi(x|z,t)\big] \\ & + \lambda \mathbb{E}_{p_\mathcal{D}(x)}[D_{\mathrm{KL}}(q_\phi(z|x) \| r(z))] \end{aligned} \tag{22}$$

The learning problem is cast as a single objective with a scalar information penalty, rather than a min–max game against a domain classifier. Moreover, thanks to the tractable variational bound in (19), SICE naturally accommodates complex treatment spaces—including high-dimensional or continuous t (as is verified in Section 5)—where enumerating or clustering "domains" becomes impractical and standard DAT methods struggle to scale.

Intervention. To estimate the individualized treatment effect at a given covariate x, we compute the posterior over z given x, then plug the treatment intervention into the outcome head while holding z fixed. Mathematically, we have

$$\begin{aligned} \mathrm{ITE}(x;t,t') &= \mathbb{E}[Y(t) - Y(t')|X = x] \\ &= \mathbb{E}_{q_\phi(z|x)}[g(z,t) - g(z,t')]. \end{aligned} \tag{23}$$



(a) (b)

Figure 2: DICE: (a) Graphical representation of DICE, where circular nodes denote observed variables and diamond nodes represent RNN hidden states; (b) simplified illustration of its single recurrent segment.

This is a valid "do"-intervention under the representation-sufficiency assumption $Y(t) \perp\!\!\!\perp T|Z$; under this assumption, intervening on T changes the outcome head while leaving the posterior over z given x unchanged.

4 DICE FORMULATION

Notations. Suppose there are T decision points. We use $\mathbf{v} \in \mathcal{V}$ to denote the time-invariant covariates, which could contain patient demographics and pre-operative assessments. Let $\mathbf{x}_t \in \mathcal{X}$ denote the time-varying covariates at decision point t, such as patients' vital signs. We use $\mathbf{a}_t \in \mathcal{A}$ to denote the action at time t. We use $\mathbf{y}_{t+1}$ to denote the response at each time point, e.g. the tumor size. We define the history up to time t as $h_t = \{\mathbf{v}, \mathbf{x}_{1:t}, \mathbf{a}_{1:t-1}\}$. In practice, we can use RNN hidden states to encode the historical information. A graphical representation of DICE is shown in Figure 2. In this Section, we use the term "RNN" to encompass a spectrum of recurrent architectures such as vanilla RNNs, gated recurrent units (GRUs) (Cho et al., 2014), and long short-term memory networks (LSTMs)(Hochreiter and Schmidhuber, 1997), and the RNN backbone can be easily replaced by a transformer and other structures of the same purpose (Melnychuk et al., 2022a).

Assumptions. For each $t = 1, \ldots, T$, we assume (i) *consistency*: if the realized treatment history equals $\bar{\mathbf{a}}_{1:t}$, then $Y_{t+1} = Y_{t+1}(\bar{\mathbf{a}}_{1:t})$; (ii) *sequential ignorability*: conditional on feature z_t, the current treatment is independent from the potential outcome, $Y_{t+1}(a_t) \perp\!\!\!\perp a_t|z_t$; and (iii) *sequential positivity*: for any h with $p(h_t = h) > 0$, we have $p(\mathbf{a}_t = a|h_t = h) > 0$ for every feasible action a under evaluation.

Adjusting for the bias. As shown in Figure 2 (b), the recurrent segment of DICE is comparable to the SCM of its static counterpart shown in Figure 1 (a), with h_t taking the role of x in the static case. The treatment-assignment bias in the dynamic model mostly comes from the association between the history h_t containing time-varying confounders $\mathbf{x}_t$ and the current treatment $\mathbf{a}_t$. To remove such bias, we can learn a stochastic representation $z_t \sim q_\phi(z_t|h_t)$ that is treatment-agnostic and predictive on y_{t+1}. We adopt the following parameterization:

$$h_t = d_h(h_{t-1}, \mathbf{v}, \mathbf{a}_{t-1}, \mathbf{x}_t),\ \mathbf{z}_t \sim q_\phi(\mathbf{z}_t|h_t),\ \hat{\mathbf{y}}_{t+1} = g(\mathbf{z}_t, \mathbf{a}_t), \tag{24}$$

for $t \in [T]$, where d_h and g are neural network predictors, and q_ϕ is stochastic embedding function. We establish the objective:

$$\min \quad \sum_{t=1}^{T} \mathbb{E}[L(\mathbf{y}_{t+1}, g(\mathbf{z}_t, \mathbf{a}_t))] + \lambda \sum_{t=1}^{T} I(\mathbf{z}_t; \mathbf{a}_t). \tag{25}$$

The training objective is given by

$$\begin{aligned} \min_{d_h, \phi, g, \psi} \quad \sum_{t=1}^{T} \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x}_t, \mathbf{a}_t, \mathbf{y}_{t+1}) q_\phi(\mathbf{z}_t|h_t)} [L(\mathbf{y}_{t+1}, g(\mathbf{z}_t, \mathbf{a}_t)) \\ - \lambda \log p_\psi(h_t|\mathbf{z}_t, \mathbf{a}_t)] \\ + \lambda \mathbb{E}_{p_{\mathcal{D}}(h_t)} [D_{\mathrm{KL}}(q_\phi(\mathbf{z}_t|h_t) \| r(\mathbf{z}_t))]. \end{aligned} \tag{26}$$

5 EXPERIMENTAL VALIDATION

The code for replicating the experiments in this section can be found at https://github.com/chrisfngr/AFCP.

5.1 SICE: Static Model

We evaluate the static counterfactual prediction model SICE against strong baselines across a broad spectrum of treatment dimensionalities, including *ultra-high-dimensional* regimes. To stress-test scalability and realism, we vary d_t from small/medium scales to hundreds of dimensions—settings commonly encountered in personalized medicine (multi-drug or multi-exposure regimens) and recommendation/targeted advertising (large action catalogs and multi-action bundles). On synthetic benchmarks where counterfactuals are available, we report **ATE Error**, **RMSE**$_y$, and **PEHE** (all lower is better). Baselines include TARNet/CFR with IPM regularization (incl. Wasserstein) Shalit et al. (2017), a CFR variant with HSIC independence regularization Gretton et al. (2007), DragonNet augmented with a generalized propensity-score head Shi et al. (2019); Hirano and Imbens (2004), and an adversarial learner with a gradient reversal layer (GRL) Ganin et al. (2016).All baselines follow the same preprocessing and training budget as SICE.

5.1.1 Synthetic Data

Unless otherwise noted, all models are trained for **50 epochs** using the **Adam optimizer** (learning rate 5×10^{-4}) under a unified preprocessing pipeline and

Table 1: Comparison across treatment dimensions d_t (lower is better). Best is **bold** $+\star$; second-best is underlined $+\square$.

(a) ATE Error

Method	2	5	10	12	14	16	18
SICE	**0.0006***	0.0019$^\square$	**0.0001***	0.0141	**0.0005***	**0.0012***	**0.0069***
TARNet	0.0054$^\square$	0.0022	0.0089	0.0006$^\square$	0.0133	0.0210	0.0406
CFR-HSIC	0.0054$^\square$	0.0081	0.0068	0.0133	0.0326	0.0131	0.0211
DragonNet-GPS	0.0099	0.0076	0.0196	**0.0001***	0.0128	0.0219	0.0179
GRL	0.0099	0.0043	0.0013$^\square$	0.0169	0.0116	0.0149	0.0131$^\square$
CFR-Wass	0.0086	**0.0008***	0.0213	0.0080	0.0029$^\square$	0.0088$^\square$	0.0187

(b) RMSE_y

Method	2	5	10	12	14	16	18
SICE	**0.5375***	0.6241$^\square$	0.8261$^\square$	**0.6534***	0.6855	**0.8022***	**0.7361***
TARNet	0.5901	0.6297	0.8818	0.6994	**0.6635***	0.8196	0.7682
CFR-HSIC	0.6047	0.6999	0.9412	0.8373	0.7452	0.9896	0.8906
DragonNet-GPS	0.5565$^\square$	**0.6215***	0.8658	0.6886$^\square$	0.6957	0.8037$^\square$	0.7617
GRL	0.5874	0.6288	**0.7895***	0.6978	0.6698$^\square$	0.9102	0.7396$^\square$
CFR-Wass	0.6064	0.7119	0.9097	0.7630	0.6916	1.0837	0.9585

(c) PEHE

Method	2	5	10	12	14	16	18
SICE	0.1635$^\square$	**0.2385***	0.5480	**0.3769***	0.5042$^\square$	**0.5387***	**0.4675***
TARNet	**0.1610***	0.2498$^\square$	**0.4811***	0.4112$^\square$	**0.4983***	0.5942	0.4896
CFR-HSIC	0.1738	0.3553	0.6717	0.6039	0.5888	0.8894	0.7399
DragonNet-GPS	0.1774	0.2682	0.4983	0.4144	0.5178	0.5467$^\square$	0.4737$^\square$
GRL	0.1872	0.2577	0.4843$^\square$	0.4456	0.5080	0.6132	0.4830
CFR-Wass	0.1886	0.3333	0.5285	0.4946	0.5382	0.8298	0.8059

training budget. Unless specified, experiments run on a single **NVIDIA A100 (80GB)** GPU. Baseline models employ 128-dimensional representation layers. Across experiments, we report three metrics: **ATE Error**, **RMSE**$_y$, and **PEHE** (all lower is better).

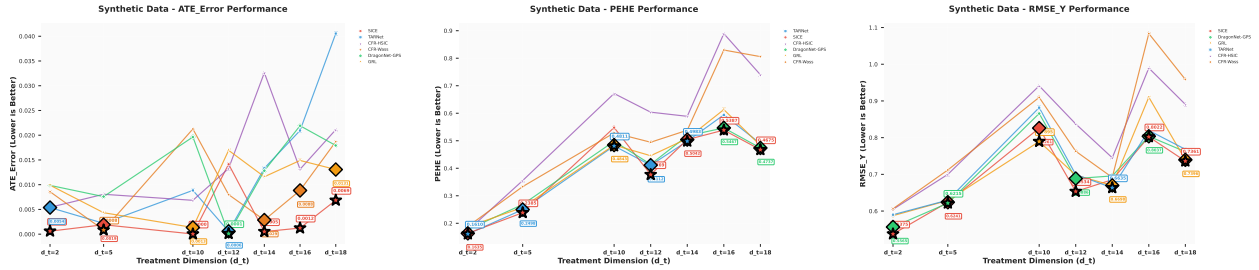
Across three metrics and seven treatment dimensions, SICE attains **13/21** best and **5/21** second-best results (three metrics $\times$ $d_t \in \{2, 5, 10, 12, 14, 16, 18\}$). Focusing on the higher-dimensional regime $d_t{>}2$ ($d_t \in \{5, 10, 12, 14, 16, 18\}$), SICE is best in **11/18** and second-best in **4/18** metric–dimension pairs (only **3/18** cases outside the top two). Breakdown for $d_t{>}2$: on **ATE Error** SICE is best in **4/6** (second at $d_t{=}5$); on **RMSE**$_y$ it is best in **3/6** (second at $d_t{=}5, 10$); on **PEHE** it is best in **4/6** (second at $d_t{=}14$). The only clear non–top-two outliers are ATE at $d_t{=}12$ and PEHE at $d_t{=}10$; elsewhere SICE is either the best or runner-up. Overall, performance degrades much more slowly than the baselines as d_t increases and remains competitive through $d_t{=}18$, highlighting robustness in higher-dimensional treatments.

Table 2: Comparison of PEHE on the Synthetic Dataset across different treatment dimensions d_t (lower is better).

Synthetic Dataset

Method	$d_t = 200$	$d_t = 500$	$d_t = 1000$
SICE	**1.18***	**3.44***	**8.34***
CFR–HSIC	1.25$^\square$	3.53$^\square$	8.52$^\square$
TARNet	1.30	3.65	8.59
GRL	1.36	3.67	8.65
CFR–Wass	1.33	3.71	8.63
DragonNet–GPS	1.31	3.69	8.70

Discussion on High-Dimensional Treatment Experiment We now evaluate SICE in the setting



(a) ATE Error (b) PEHE (c) RMSE_y

Figure 3: Comparison across treatment dimensions d_t. Panels (a)–(c) compare SICE with other methods on ATE Error, PEHE, and RMSE_y (all lower is better). Star markers denote the best method; square markers denote the second-best.

Table 3: NHANES 2017–2018 with multi-dimensional treatments ($T \in \{0,1\}^{82}$) and covariates ($X \in \mathbb{R}^{14}$). We report factual prediction (RMSE/MAE; lower is better) and off-policy targeting (AUUC; higher is better). Best is **bold**+$\star$, second-best is underlined+$\square$.

(a) Factual prediction: RMSE / MAE

Method	RMSE ↓	MAE ↓
SICE (ours)	**0.0285***	**0.0227***
TARNet	0.0938	0.0823
CFR–HSIC	0.0899$^\square$	0.0817$^\square$
DragonNet–GPS	0.1332	0.1225
GRL	0.1001	0.0902
CFR–Wass	0.1008	0.0912

(b) Off-policy targeting: AUUC

Method	AUUC ↑
SICE (ours)	0.2761$^\square$
TARNet	0.2383
CFR–HSIC	0.2393
DragonNet–GPS	**0.3021***
GRL	0.2034
CFR–Wass	0.2150

Notes. NHANES lacks ground-truth counterfactuals, so PEHE is not applicable. Arrows indicate the preferred direction for each metric. Best/second-best are evaluated across listed methods within each panel.

of *ultra–high-dimensional treatments*, which frequently arise in real-world applications. Examples include personalized medicine (e.g., multi-drug or multi-exposure regimens) and recommendation/targeted advertising (e.g., large action catalogs or multi-action bundles), where the treatment can be a vector with hundreds of components. To stress-test scalability, we report supplementary results on a high-dimensional synthetic benchmark with $d_t \in \{200, 500, 1000\}$ (Table 2). SICE attains the lowest PEHE across all three ultra-high-dimensional settings, indicating strong robustness as treatment dimensionality grows.

As shown in the table above, the results are clear: **SICE consistently outperforms all other baseline methods** across all tested treatment dimensions ($d_t = 200, 500, 1000$). For each dimension, SICE achieves the lowest PEHE (Precision in Estimating Heterogeneous Effects), a crucial metric for evaluating individual treatment effect estimation. This finding strongly supports our hypothesis that SICE is a superior and more robust choice for causal effect estimation in the presence of high-dimensional and non-linear data.

The ability of SICE to maintain its performance advantage as the treatment dimensionality scales up confirms its robustness and adaptability. This makes it a highly effective and reliable solution for causal inference tasks in today's complex, data-rich domains, where traditional methods often struggle.

5.1.2 Real Data (NHANES 2017–2018)

We use the **NHANES 2017–2018** U.S. health survey (CDC/NCHS) Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) (2020). In our setup, the *treatment* is a multi-dimensional vector of medication/exposure indicators with $t_{\text{dim}} = 82$; *covariates* include core demographics/exam features with $x_{\text{dim}} = 14$; the *outcome* Y is continuous (e.g., a laboratory biomarker), so squared and absolute errors are meaningful for factual fit. Counterfactuals are unobserved; thus we pair factual prediction (RMSE/MAE) with an off-policy targeting metric (AUUC) Rzepakowski and Jaroszewicz (2012) that evaluates ranking quality for beneficial interventions.

Table 4: DICE: Comparison across treatment dimensions d_t (lower is better). Best is **bold**+$\star$; second-best is underlined+$\square$.

(a) ATE Error

Method	$d_t = 2$	$d_t = 4$	$d_t = 6$	$d_t = 8$	$d_t = 10$
DICE	**0.0187***	**0.0029***	**0.0702***	**0.0951***	**0.0236***
TARNet	0.0550$^\square$	0.2469$^\square$	0.2496$^\square$	0.1722$^\square$	1.6088$^\square$
CFR–HSIC	0.2363	0.1811	1.1234	2.1101	0.7033
DragonNet–GPS	0.2381	0.2237	1.1421	2.2098	0.7269
GRL	0.2385	0.4012	1.1048	2.0916	0.7451

(b) RMSE_y

Method	$d_t = 2$	$d_t = 4$	$d_t = 6$	$d_t = 8$	$d_t = 10$
DICE	**0.2288***	**0.2968***	**0.3538***	**0.3306***	**0.2215***
TARNet	0.4801$^\square$	0.6326$^\square$	0.8753$^\square$	0.5056$^\square$	1.0363$^\square$
CFR–HSIC	1.8095	10.5588	7.5341	6.4323	2.4929
DragonNet–GPS	1.8041	10.5683	7.5353	6.4346	2.4953
GRL	1.8116	10.4921	7.5832	6.3553	2.4920

(c) PEHE

Method	$d_t = 2$	$d_t = 4$	$d_t = 6$	$d_t = 8$	$d_t = 10$
DICE	**0.0293***	**0.1003***	**0.1180***	**0.1548***	**0.0823***
TARNet	0.1161$^\square$	0.2624	0.3561$^\square$	0.3030$^\square$	1.6686
CFR–HSIC	0.2364	0.2081$^\square$	1.1264	2.1105	0.7035$^\square$
DragonNet–GPS	0.2383	0.2384	1.1476	2.2105	0.7271
GRL	0.2386	0.4320	1.1162	2.0928	0.7453

Interpretation. As summarized in Table 3, **SICE** achieves *state-of-the-art factual accuracy* under high-dimensional multi-treatment inputs (t_{dim}=82, x_{dim}=14), attaining the best RMSE and MAE. On the causal targeting axis, SICE ranks *second* by AUUC, close to the uplift-specialized DragonNet–GPS. This profile—top factual calibration with competitive off-policy ranking—is particularly attractive for clinical decision support, where accurate outcome modeling is primary and reliable targeting quality is an added benefit in high-dimensional treatment settings.

5.2 DICE: Dynamic Model

We next assess the **Dynamic Information-theoretic Counterfactual Estimator (DICE)** in dynamic treatment settings. Synthetic data ($N = 1000$ samples over $T = 10$ time steps) is generated via a configurable **dynamic simulation** with $\mathbf{X} \in \mathbb{R}^8$, $\mathbf{Y} \in \mathbb{R}^1$, and treatment dimension d_t varied up to 10. All baselines follow the same setup as above. DICE employs a 128-dimensional hidden layer and learns a 32-dimensional latent representation $\mathbf{Z}$, controlled by $\lambda = 10^{-5}$.

Table 4 reports the outcomes. Across all d_t, DICE consistently achieves the **best scores** on ATE Error, $\mathbf{RMSE}_y$, and PEHE, clearly outperforming TARNet, CFR-HSIC, DragonNet-GPS, and GRL. Unlike baselines, which deteriorate sharply in higher treatment dimensions, DICE remains stable and accurate, demonstrating superior robustness for estimating ITEs in dynamic high-dimensional settings.

6 CONCLUSION

We presented an adversary-free framework for counterfactual prediction grounded in an information-theoretic risk bound that links the counterfactual–factual gap to mutual information. Building on this insight, we introduced SICE (static) and DICE (sequential), which learn stochastic representations that remain predictive for outcomes while directly penalizing treatment–representation dependence via a tractable variational surrogate, avoiding the instability of min–max training. Empirically, the approach is competitive across synthetic settings and a real clinical dataset, and it remains robust as treatment dimensionality grows, supporting complex, multi-attribute inter-

ventions. Future work includes sensitivity-aware training and proximal/IV-style extensions, tighter MI estimators, removal of assignment bias in the case of unmeasured confoundings, and policy-learning objectives that leverage our representation for decision-making under sequential dynamics.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Jeroen Berrevoets, James Jordon, Ioana Bica, Mihaela van der Schaar, et al. Organite: Optimal transplant donor organ offering using an individual treatment effect. *Advances in neural information processing systems*, 33:20037–20050, 2020.

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

Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS). National health and nutrition examination survey (nhanes) data, 2017–2018, 2020. URL https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017. Accessed 2025-10-03.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

Yaroslav Ganin, Victor Lempitsky, et al. Domain-adversarial training of neural networks. *JMLR*, 17 (59):1–35, 2016.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46, 2012.

Miguel A. Hern'an and James M. Robins. *Causal Inference: What If*. Chapman Hall/CRC, 2020. Available at https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

Keisuke Hirano and Guido W. Imbens. The propensity score with continuous treatments. *Biometrika*, 91 (2):331–346, 2004. doi: 10.1093/biomet/91.2.331.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In *Neural Computation*, volume 9, pages 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *NeurIPS*, 2016.

Amirreza Kazemi and Martin Ester. Adversarially balanced representation for continuous treatment effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13085–13093, 2024.

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

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *ICML*, 2022b.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. *Advances in neural information processing systems*, 31, 2018.

Leland Gerson Neuberg. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685, 2003.

James M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application of the g-formula. *Mathematical Modelling*, 7(9–12):1393–1512, 1986. doi: 10.1016/0270-0255(86)90088-6.

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

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

Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple

treatments. *Knowledge and Information Systems*, 32(2):303–327, August 2012. ISSN 0219-1377. doi: 10.1007/s10115-011-0434-0. URL https://doi.org/10.1007/s10115-011-0434-0.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML*, 2017.

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *arXiv preprint arXiv:1906.02120*, 2019. URL https://arxiv.org/abs/1906.02120.

Shiqin Tang[*,a], **Rong Feng**[*,a,b], **Shuxin Zhuang**[a,b], **Youzhi Zhang**[a], **Hongzong Li**[†,c]

Adversary-Free Counterfactual Prediction via Information-Regularized Representations: Supplementary Materials

Standing notation. X covariates; $T \in \mathcal{T}$ with law π; representation $Z \sim q_\phi(\cdot|X)$; outcome Y. Per-arm Z-profile $\varphi_t(z) := \mathbb{E}[L(Y(t), g_t(z))|Z = z]$. Factual and counterfactual risks $R_F := \mathbb{E}_{t\sim\pi}\mathbb{E}_{p(z|t)}[\varphi_t(z)]$ and $R_{CF} := \mathbb{E}_{t\sim\pi}\mathbb{E}_{p_Z}[\varphi_t(z)]$. Total variation $\mathrm{TV}(\cdot,\cdot)$; Kullback–Leibler $D_{\mathrm{KL}}(\cdot\|\cdot)$; mutual information $I(Z;T) = \int D_{\mathrm{KL}}(p(z|t)\|p_Z)\,\pi(dt)$.

7 Main Z-space bound

Theorem 1 (Z-space risk gap via MI)**.** Assume there exists $\lambda > 0$ and $\mathcal{F} \subset \{f : \|f\|_\infty \le 1\}$ such that $\phi_t/\lambda \in \mathrm{conv}(\mathcal{F})$ for all $t \in \mathcal{T}$. Then

$$|R^{CF} - R^F| \le \sqrt{2}\,\lambda\,\sqrt{I(Z;T)}.$$

Proof. Write $p_t(\cdot) := p(z \mid T = t)$. Then

$$\begin{aligned}
|R^{CF} - R^F| &= \left|\int \big(\mathbb{E}_{p_Z}[\phi_t] - \mathbb{E}_{p_t}[\phi_t]\big)\,\pi(dt)\right| \\
&\le \int |\mathbb{E}_{p_Z}[\phi_t] - \mathbb{E}_{p_t}[\phi_t]|\;\pi(dt) \\
&\le \lambda \int \mathrm{IPM}_{\mathcal{F}}(p_t, p_Z)\,\pi(dt) \\
&\le 2\lambda \int \mathrm{TV}(p_t, p_Z)\,\pi(dt) \\
&\le \sqrt{2}\,\lambda \int \sqrt{D_{\mathrm{KL}}(p_t\|p_Z)}\,\pi(dt) \\
&\le \sqrt{2}\,\lambda \sqrt{\int D_{\mathrm{KL}}(p_t\|p_Z)\,\pi(dt)} \\
&= \sqrt{2}\,\lambda\,\sqrt{I(Z;T)}.
\end{aligned}$$

□

Equality/tightness. Exact equality when $I(Z;T) = 0$ (perfect balance). Otherwise, the chain of inequalities is tight if simultaneously: (i) all arm-wise TV maximizers align, (ii) the divergences $D_{\mathrm{KL}}(p(z|t)\|p_Z)$ are π-a.s. constant , (iii) D_{KL} is small so that Pinsker is near-tight.

8 Low MI limits treatment predictability

Proposition 1 (Binary case)**.** Assume $T \in \{0,1\}$ and $\pi(0) = \pi(1) = 1/2$. Let $e^\star$ denote the Bayes error of predicting T from Z. Then

$$e^\star \ge \frac{1}{2} - \sqrt{\frac{I(Z;T)}{2}}.$$

Proof. For equal priors,

$$e^\star = \frac{1}{2}\big(1 - \mathrm{TV}(p_0, p_1)\big), \qquad p_i := p(z \mid T = i).$$

Also $p_Z = \frac{1}{2}(p_0 + p_1)$, so by triangle inequality,

$$\mathrm{TV}(p_0, p_1) \le \mathrm{TV}(p_0, p_Z) + \mathrm{TV}(p_1, p_Z).$$

Applying Pinsker's inequality and then Cauchy–Schwarz gives

$$\mathrm{TV}(p_0, p_1) \le \sqrt{2\, I(Z;T)}.$$

Substituting into the expression for $e^\star$ yields the claim. □

Remark 1 (Multiclass uniform prior)**.** If T is uniform on $K \ge 2$ classes, then Fano's inequality yields

$$e^\star \ge 1 - \frac{I(Z;T) + \log 2}{\log K}.$$

9 Additional Experiments

We study how the mutual–information regularization $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10\}$ affects four metrics on a fixed synthetic dataset: test RMSE_y, PEHE, ATE error, and $\mathrm{HSIC}(z,t)$. Let $z = f_\phi(x)$ denote the learned representation of x. Following standard kernel–based independence testing Gretton et al. (2007), we use $\mathrm{HSIC}(z,t)$ as a dependence proxy between z and the treatment t (smaller indicates weaker z–t coupling).

Findings. Across representation dimensions t_{dim}, accuracy metrics favor small–to–moderate regularization: RMSE_y and PEHE typically attain minima near $\lambda \approx 10^{-4}$–10^{-3}, and ATE error is smallest for $\lambda \in [10^{-4}, 10^{-2}]$. Larger regularization $(0.1, 1, 10)$ consistently degrades accuracy, whereas $\mathrm{HSIC}(z,t)$ decreases (monotonically or near-monotonically) with λ, indicating weaker z–t coupling. For example, at t_{dim}=2, RMSE_y improves from 0.5639 (10^{-5}) to 0.5388 (10^{-4}) but rises to 1.4231 at λ=10; PEHE is best at 10^{-3} (0.1529) yet increases to 0.6808 at λ=10; HSIC drops from 2.98×10^{-3} (10^{-5}) to 6.81×10^{-4} (10). At t_{dim}=10, RMSE_y is 0.8246 at 10^{-4} but 2.9369 at 10, while HSIC declines from 9.79×10^{-3} (10^{-5}) to 1.09×10^{-3} (10). These results validate the intended role of the regularizer: λ acts as a knob on the shared information between the treatment t and the learned representation $z = f_\phi(x)$. As λ increases, the empirical dependence between z and t—estimated on the test set via HSIC as a kernel-based proxy for $I(z,t)$ Gretton et al. (2007)—consistently decreases across treatment dimensions. Operationally, we train the model on the training split, embed test inputs to obtain z, and then compute $\mathrm{HSIC}(z,t)$; the near-monotone decline of HSIC with larger λ demonstrates that stronger regularization yields representations z that carry less information about t.

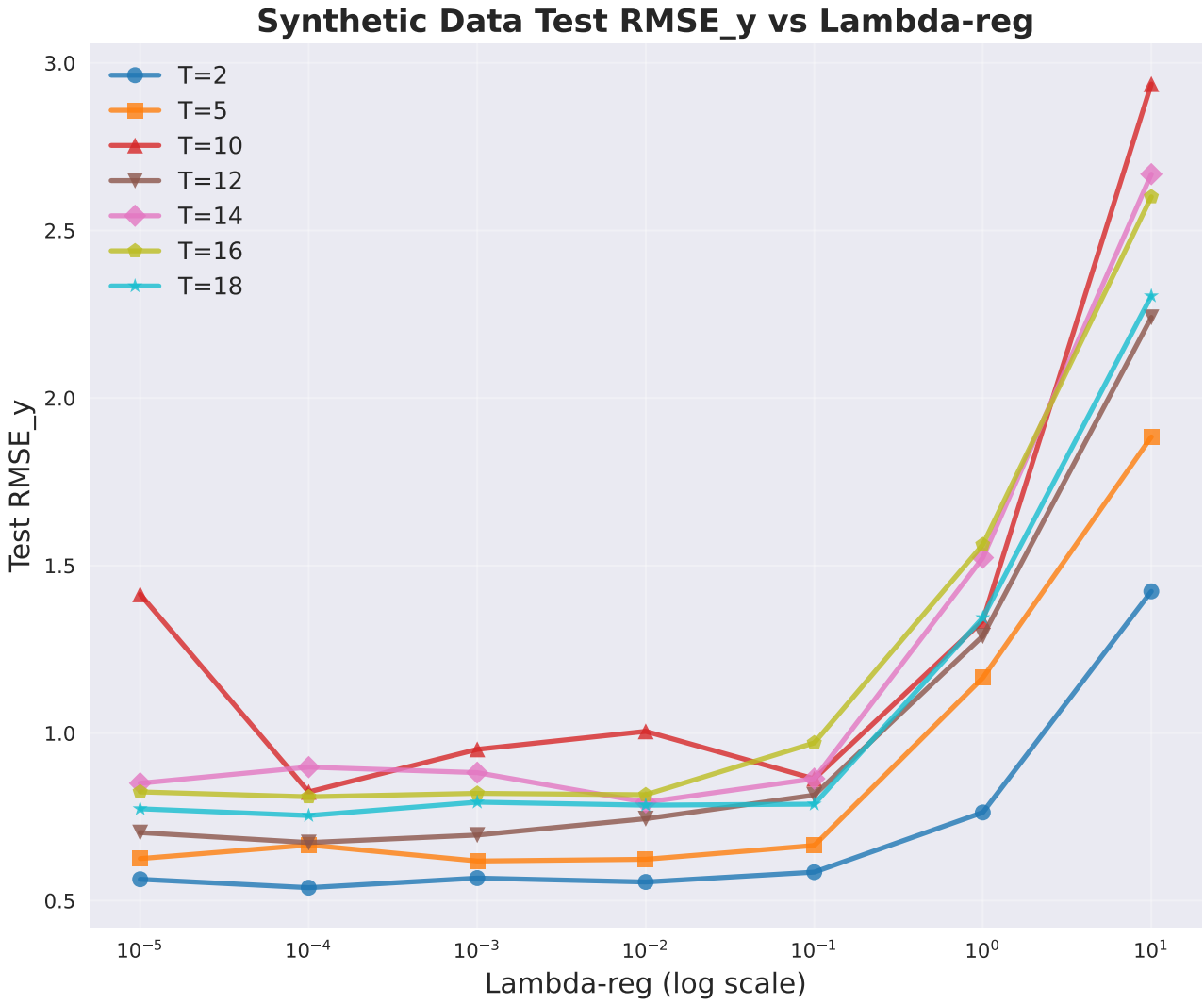


Figure 4: RMSE_y versus $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10\}$ on the synthetic dataset (lower is better).

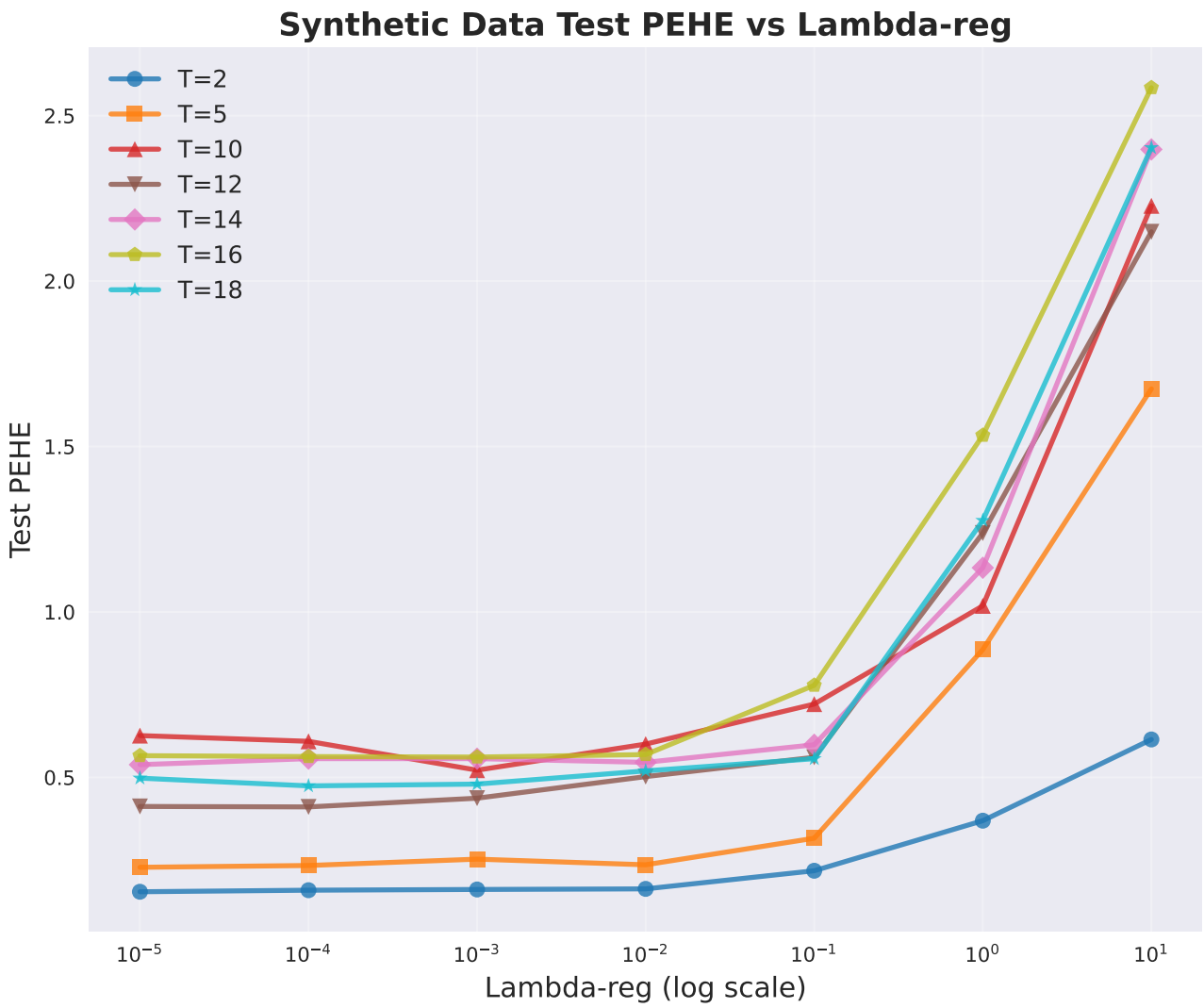


Figure 5: PEHE versus λ on the synthetic dataset (lower is better).

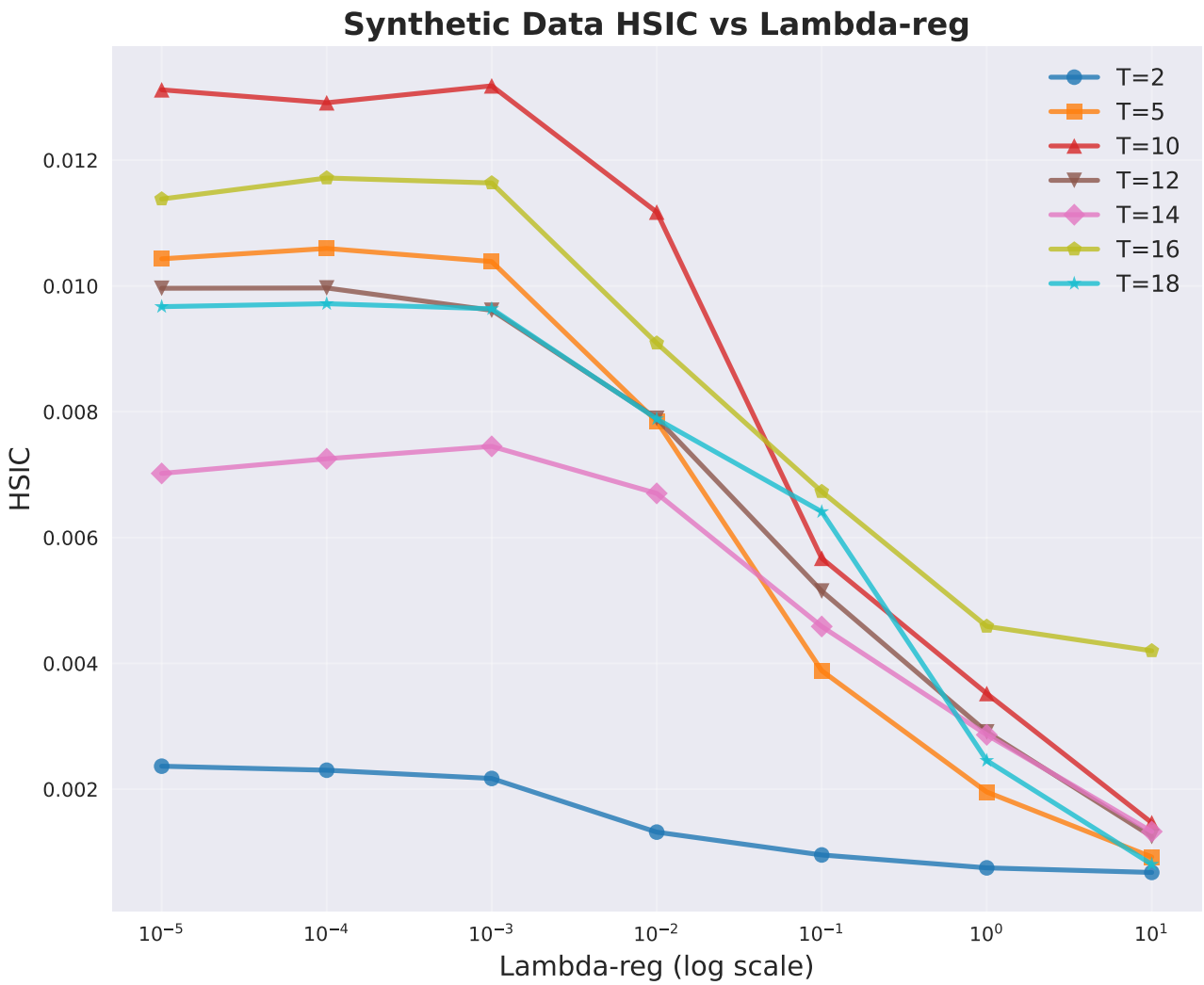


Figure 6: HSIC(z, t) versus λ as a kernel-based dependence proxy between $z = f_\phi(x)$ and the treatment t (smaller indicates weaker dependence).

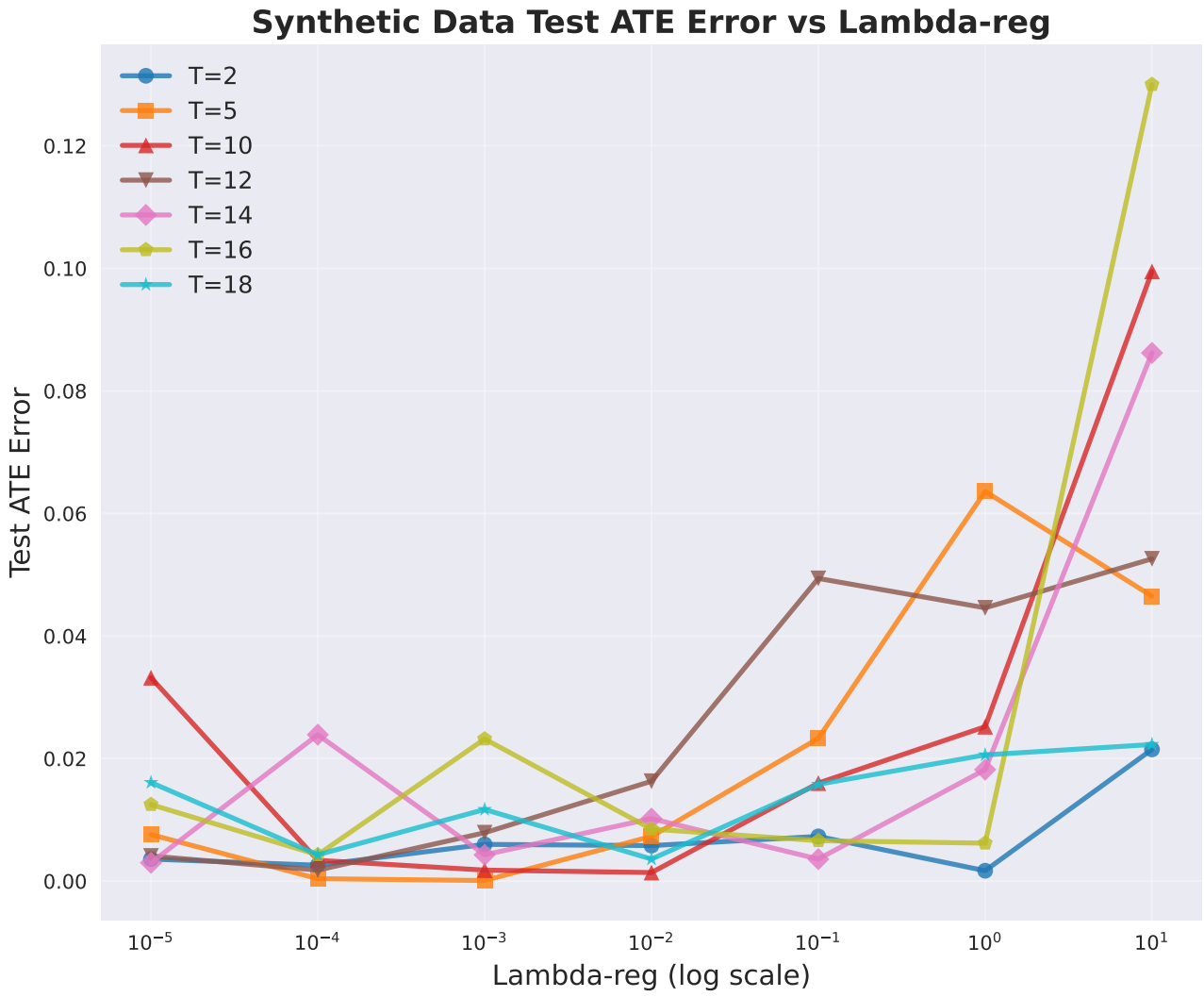


Figure 7: ATE error versus λ on the synthetic dataset (lower is better).

Table 5: Results across t_{dim} and λ (averaged over runs; standard deviations omitted for space). Rows are t_{dim}; columns are $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10\}$.

(a) RMSE_y

t_{dim}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	0.1	1	10
2	0.563900	0.538800	0.567400	0.555800	0.585200	0.763300	1.423100
5	0.625600	0.666200	0.618400	0.623500	0.664700	1.165100	1.884000
10	1.414200	0.824600	0.951600	1.005300	0.863600	1.335200	2.936900
12	0.703200	0.673400	0.695800	0.744600	0.815000	1.289800	2.240700
14	0.850500	0.898600	0.882100	0.793700	0.864000	1.523500	2.668000
16	0.824600	0.810000	0.820300	0.816100	0.970900	1.562300	2.600400
18	0.774200	0.754100	0.794000	0.785000	0.788000	1.344500	2.304600

(b) PEHE

t_{dim}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	0.1	1	10
2	0.154300	0.159000	0.152900	0.155300	0.167500	0.250400	0.680800
5	0.276600	0.286900	0.277000	0.284100	0.349200	0.833200	1.369800
10	0.707000	0.555600	0.548500	0.561900	0.612500	0.939200	2.183300
12	0.468800	0.462500	0.468700	0.486200	0.539900	0.914100	1.862400
14	0.598900	0.639700	0.562600	0.581600	0.601300	1.096400	2.221100
16	0.703800	0.675100	0.695600	0.708400	0.730900	1.188400	2.359900
18	0.736200	0.724800	0.736700	0.737300	0.740400	1.162100	2.226800

(c) ATE error

t_{dim}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	0.1	1	10
2	0.003700	0.000100	0.000500	0.000200	0.000800	0.005100	0.048200
5	0.002200	0.002600	0.000100	0.000100	0.000900	0.013400	0.090000
10	0.053600	0.013700	0.019000	0.006000	0.009800	0.032800	0.203200
12	0.009200	0.009400	0.007200	0.003700	0.009200	0.047700	0.206200
14	0.021700	0.028800	0.015700	0.013200	0.014000	0.091000	0.302500
16	0.019700	0.019800	0.018800	0.013500	0.015600	0.079600	0.316400
18	0.019000	0.019300	0.016700	0.015000	0.014900	0.079500	0.303300

(d) HSIC(z, t)

t_{dim}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	0.1	1	10
2	0.002981	0.002975	0.002821	0.002678	0.001845	0.001208	0.000681
5	0.004485	0.004276	0.003993	0.003615	0.001637	0.000781	0.000603
10	0.009785	0.009740	0.008795	0.007895	0.005640	0.001608	0.001089
12	0.004377	0.004383	0.004090	0.003659	0.003045	0.001790	0.000969
14	0.008351	0.008332	0.007280	0.006456	0.004819	0.001721	0.001290
16	0.005608	0.005535	0.004826	0.004629	0.003576	0.002113	0.003355
18	0.006674	0.006705	0.006611	0.005721	0.005841	0.002936	0.005055

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]