ESTIMATING INDIVIDUAL TREATMENT EFFECTS UNDER UNOBSERVED CONFOUNDING

Lin Zhao, Ahbab Abeer, Ahmad Choudhary, Hemasai Suhas Kurapati
Duke University

ABSTRACT

The estimation of causal effects from observational data is a cornerstone of computational medicine, allowing for the prediction of individual patient responses to interventions. This project investigates how unobserved confounder variables affect deep causal estimators, specifically focusing on the Treatment-Agnostic Representation Network (TARNet) and its variants. Using synthetic and semi-synthetic data with known Conditional Average Treatment Effects (CATE), we quantify estimation error under varying degrees of confounding. While TARNet variants such as Dragonnet and CFR TARNet (utilizing Wasserstein distance for balancing) have shown significant improvements over TARNet in estimating CATE, they operate under the assumption of "no hidden confounding." We conduct a sensitivity analysis, exploring how confounding effects impact these these advanced models when that assumption is violated.

1 INTRODUCTION

The estimation of causal effects from observational data is a central challenge in modern Machine Learning, especially in clinical settings where data availability is low. In these scenarios, the main goal is to get the Conditional Average Treatment Effect (CATE), which is the the expected difference in outcomes under treatment and control (no treatment) conditions for an individual, based on their observable features, x:

$$\tau(x) = \mathbb{E}[Y_1 - Y_0 | X = x] \tag{1}$$

Y_1 and Y_0 represent the outcomes under treatment and control, and their difference, CATE shown as τ, represents the improvement in the health of an individual. CATE is conditioned on X because an individual's health features can affect the effectiveness of a treatment. For example, chemotherapy might exhibit a positive CATE for a patient in the latter stages of cancer, but a negative CATE (due to severe side effects in the absence of a life-threatening disease) for an individual without cancer. By accurately focusing on CATE, researchers and clinicians are able to model counterfactual scenarios ("what happens when this treatment is applied versus not?"), enabling a more personalized comparison between potential treatment plans for each individual.

However, observational data are very subject to confounding because we only observe Y_0 or Y_1 (either a patient is treated or not). The assignment of treatment T often depends heavily on X (e.g., sick patients receive treatment). Therefore, the distributions of (X) for treated patients differ systematically from the X of untreated patients (e.g', if someone was treated, they may have been very sick). To accurately estimate CATE, any causal inference model must overcome this imbalance. The model must learn to accurately predict both potential outcomes ($\hat{Y}_0$ and $\hat{Y}_1$), regardless of treatment, otherwise the CATE estimates will be biased. Two existing strategies for dealing with this issue are discussed in the related works section.

The main issue that our work examines is the presence of unobserved confounders, variables that affect treatment and outcome, but are not present in X, and are therefore not given to the model. For example, any drug that influences both the doctors decision for treatment and affects the treatment outcome, but is not encoded in the X given to the model. We examine the effect of unobserved confounders, U, on a models ability to predict CATE by evaluating their Precision in Estimating Heterogeneous Effects (PEHE).

$$\text{PEHE} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\tau}(X^{(i)}) - (Y_1^{(i)} - Y_0^{(i)}) \right)^2 \tag{2}$$

1.1 MAIN CHALLENGE - TARNET WITH ONOBSERVED COFOUNDERS

Deep learning architectures like the Treatment-Agnostic Representation Network (TARNet) are given an individuals observed features, X, and trained to predict Y_T, where T is the observed treatment (either 0 or 1). TARNet predicts both Y_1 and Y_0 and is given guidance only on its prediction of the ground-truth Y_T, using the MSE Loss.

TARNet and its variants try to solve the data imbalance problem (caused by confounding in X) by learning a function $z = \Phi(X)$ that outputs a set of hidden features, z, and uses that shared z to predict both Y_1 and Y_0, given X. That way, it must learn an encoding of X that works for both treatment and control, hopefully ensuring good estimates for CATE. However, when an unobserved confounder U exists, the standard TARNet objective (minimizing factual prediction error) becomes biased.

Lets examine a toy example to illustrate this point. Imagine we are trying to predict the effect of tutoring on a student's future GPA. Let X = previous GPA, $T = 1$ if a student enrolled in tutoring, and Y_T = future GPA. Let U = motivation level, a binary variable that affects both whether or no a students chooses to enroll in tutoring, and affects outcomes Y_T.

X **(Prev. GPA)**	**U** **(Motivation)**	$\mathbf{Y_0}$ **No Tutoring**	$\mathbf{Y_1}$ **Tutoring**	τ $(\mathbf{Y_1} - \mathbf{Y_0})$	**Obs. T** **(Treatment)**	**Obs.** $\mathbf{Y_T}$ **(Factual)**
2.5	0 (Low)	2.5	2.7	0.2	0	2.5
3.0	0 (Low)	3.0	3.2	0.2	0	3.0
3.5	0 (Low)	3.5	3.7	0.2	0	3.5
2.5	1 (High)	3.0	3.2	0.2	1	3.2
3.0	1 (High)	3.5	3.7	0.2	1	3.7
3.5	1 (High)	4.0	4.2	0.2	1	4.2

This is an example of a dataset with catastrophic confounding in U. Because motivated students are much more likely to take tutoring, and motivated students also tend to have higher grades, the model learns to output high values for $\hat{Y}_1$ and low values for $\hat{Y}_0$ given X. Despite the true Treatment effect (τ) only being 0.2, the model's predictions for $\hat{Y}_0$ and $\hat{Y}_1$ given any X will be apart by 0.7. The network conflated the causal effect of the treatment with the false correlation induced by U. Thus, the estimated $(\widehat{\text{CATE}})$ is biased:

$$\widehat{\text{CATE}} \approx (\text{true causal effect}) + (\text{effect caused by a difference in } U) \approx 0.2 + 0.5 = 0.7$$

By systematically increasing the confounding effect of U, we will quantify how the estimation bias grows. We will measure the sensitivity of TARNet to unobserved confounding by plotting the PEHE against the dimension of the hidden confounders, U, and the strength of the confounding signal.

1.2 Mitigation

In addition to the base TARNet model, we will be exploring the two following strategies that are used in deep causal learning. We will explore if they are each affected by confounding in a similar manner, and if they can mitigate the effects of confounding.

- **Propensity:** The propensity score, $e(x) = P(T = 1|X)$, represents the probability of receiving treatment given X. In our project, we use the propensity in two complementary ways:
  - 1) First, we concatenate the estimated propensity score to the learned representation or to the prediction heads. This gives the network information about how X influences treatment selection, helping the model correct for selection bias when U is absent.
  - 2) Second, we use the **DragonNet** model Shi et al. (2019), which adds a third "propensity head" trained to predict T. We discuss this model further in Section (2.4).

  These causal deep learning models (TARNet, DragonNet, etc) all assume no unobserved confounding. Propensity mitigation strategies primarily focus on reducing selection bias caused by imbalance in the distribution of X between treated and control groups – thereby reducing PEHE. Our project tests whether these improvements persist when true unobserved confounding is introduced.

- **Wasserstein Distance:** Selection bias results in the distribution of representations for the treated group, $P(\Phi(X)|T = 1)$, differing from the control group's, $P(\Phi(X)|T = 0)$, and standard regression losses do not penalize this discrepancy. Using metrics like MMD, we can force the learned representations of treated and control groups to overlap in feature space. This is supposed to ensure that the model does not rely on features that purely predict treatment.

1.3 Project Scope

Because real-world medical data lacks ground truth counterfactuals (ie: we typically don't know what happens when treatments are NOT applied), we can't empirically measure the bias caused by unobserved confounders in real datasets. To address this we utilize synthetic and semi-synthetic data where X, U, Y_T, confounding strength, and true $CATE$ are known by construction. Then, we will aim to quantify exactly how the estimation bias in TARNet, and its variants grow as confounders are masked (by altering confounder dimensions and changing their strengths), plus whether propensity informed modeling and MMD balancing mitigate it.

2 Related Works

2.1 Representation Learning for Causal Inference

The core of this work is the framework introduced by Shalit, Johansson, and Sontag (2017) Shalit et al. (2017), who formalized the Treatment Agnostic Representation Network (TARNet). They introduced the main architecture of TARNet, and how to train it. They introduced the use a nueral network (NN), to learn a function $z = \phi(x)$ that would output a latent representation of the observable features X. This shared z would be fed into two seperate multi-layer-perceptron networks (heads), that would predict $\hat{Y}_1$ and $\hat{Y}_0$. They chose to have a shared representation layer, z, to reduce the effect of selection bias in predicting $\hat{\tau}_1$

As deep learning became feasible for causal inference, multiple representation learning frameworks came about. Johansson et al (2016) Johansson et al. (2016) investigated domain adaptation theoretic bounds on counterfactuals prediction error. Louizos et al (2017) Louizos et al. (2017) proposed the Deep Latent Variable Model for Causal Inference, which models unobserved confounders using variational inference. This is closely related to the synthetic data design of the project, where latent variables representing unrecorded health factors are used to produce confounding.

2.2 Benchmarks

The literature on synthetic and semi synthetic causal benchmarks provides methodological grounding for the dataset design used in this project. Dorie et al (2019) Dorie et al. (2019) introduced the ACIC challenge datasets, showing the importance of controlled confounding strength. Gentzel et al. (2019) Gentzel et al. (2019) and Bica et al (2020) Bica et al. (2020) extended semisynthetic strategies to medical record settings. The project's semi synthetic extension mirrors these practices.

2.3 TARNet variants

Several variants to TARNet have found to reduce bias and improve PEHE error. One improvement to vanilla TARNet is to have a seperate NN that predicts T given X. This NN outputs a propensity score, $\hat{e}(X) = P(T = 1|X)$. This network is trained by comparing $\hat{e}(X)$ with the real class, T. $\hat{e}(X)$ is appended to X and given to TARNet as an additional data point. The goal is that by giving TARNet some idea of what class, T, that X belongs to, it can better predict $\hat{Y}_0$ and $\hat{Y}_1$. This approach is an example of propensity informed TARNet

The Dragonnet architecture introduced in Section (1.2) Shi et al. (2019) is a modern implementation of a propensity informed TARNet. It improves on TARNet by adding a third MLP head to explicitly predict the propensity score. Rather than adding the propensity score as an additional data point given to TARNet as above, Dragonnet forces the network to also predict $\hat{e}(X)$ in addition to predicting $\hat{Y}_0$ and $\hat{Y}_1$. The third head acts as a regularization term, forcing the shared representation $\Phi(X)$ to produce hidden features that are simultaneously useful for predicting the potential outcomes (Y_0 and Y_1) and the treatment assignment (T). Because the representation must encode information relevant to treatment assignment, it captures structure related to selection bias, which significantly improves $\hat{\tau}$ accuracy when confounding exists in X.

Another improvement, CFR TARNet (Counterfactual Regression TARNet) Shalit et al. (2017), uses an Integral Probability Metric (IPM), like Wasserstein distance, to reduce the bias in the CATE estimate ($\hat{\tau}$) by explicitly enforcing covariate balancing in the latent representation space. The core idea is to train the shared representation function ($\Phi(X)$) with a compound loss function that includes two parts: the standard outcome prediction error (like TARNet) and a regularization term based on the Wasserstein distance. The second term measures the dissimilarity between the distribution of the latent features for the treated group, $P(\Phi(X)|T = 1)$, and the distribution for the control group, $P(\Phi(X)|T = 0)$. By minimizing the Wasserstein distance, the model is forced to learn a representation $\Phi(X)$ that encodes X similarly for samples taken from both the treatment and control groups. This forces $\Phi(X)$ to encode only features relevant to predicting Y_0 and Y_1. This effectively removes the systematic differences (confounding) between the groups before the potential outcome heads perform regression, mitigating selection bias and leading to more accurate $\hat{Y}_0$ and $\hat{Y}_1$ predictions, which in turn reduces the error in $\hat{\tau}$.

3 Project Plan

The project implementation is as follows. First, we implemented a synthetic and semi-synthetic data generation pipeline to simulate a clinical environment with controlled unobserved confounding. We constructed a generative process that maps latent health variables into high-dimensional observables (X, representing medical images) and hidden confounders (U). We parameterized this generation to work for different values for the dimension of the hidden confounder (l) and the strength of the confounding influence (β). This allowed us to create a suite of datasets representing various degrees of violation of the "no hidden confounding" assumption.

Next, we implemented four causal estimators to benchmark performance. We utilized the standard TARNet as a baseline. Then, we implemented the Balanced TARNet, which augments the loss with a Maximum Mean Discrepancy (MMD) penalty to enforce domain invariance between treated and control groups using the wasserstein distance metric. We also used Propensity informed TARNet, which explicitly conditions the outcome heads on a propensity score estimated by a secondary classifier. Finally, we implemented the Dragonnet* architecture, which extends TARNet by adding a third head to predict propensity scores. Using these models, we were able to determine if distributional balancing, propensity features, or multi-task regularization could provide robustness against the hidden confounders.

Then, we executed a training and evaluation loop to quantify model sensitivity. We trained each model variant across the full grid of synthetic and semi-synthetic datasets (varying β and l). Finally, we monitored convergence using MSE and evaluated final performance using PEHE on the test set.

3.1 Data Generation Process

We designed a pipeline to evaluate causal estimators under varying selection bias and confounding complexity. Our approach ensures that the selection mechanism is driven by high-dimensional, non-linear features, mimicking complex medical imaging scenarios.

3.1.1 Synthetic Data Generation

To address the lack of counterfactuals in real data, we generated fully synthetic datasets using a structural causal model that allows precise control over confounding levels.

Latent Structure & Confounder We sample independent latent vectors $Z_1, Z_2 \sim \mathcal{N}(0, I)$ representing observed and unobserved health indicators, respectively.

$$\begin{aligned} Z_1 &\sim \mathcal{N}(0, I_{dz1}) \\ Z_2 &\sim \mathcal{N}(0, I_{dz2}) \end{aligned} \tag{3}$$

Then, we map Z_1 and Z_2 onto higher dimensional features, X and U, using up-scaling convolutions. Specifically, Z_1 gives rise to the input images X, while Z_2 produces the confounding variable U, which influences both treatment assignment and potential outcomes. Both X and U are 64 x 64 and are generated using similar CNN architectures but with different initialized weights.

$$\begin{aligned} X &= G_X(Z_1) + \eta_X, \qquad \eta_X \sim \mathcal{N}(0, \sigma_X^2 I), \\ U &= G_U(Z_2) + \eta_U, \qquad \eta_U \sim \mathcal{N}(0, \sigma_U^2 I), \end{aligned} \tag{4}$$

where G_U and G_X represent a multi-layer CNN that transforms the latent factors into explicit samples.

Treatment Assignment Treatment T is modeled as a Bernoulli process. The class assignment for an individual depends on X and U, and is modeled by

$$T \sim \text{Bernoulli}\big(\sigma(\alpha^\top X + \beta^\top U)\big), \tag{5}$$

Outcome Generation The outcomes, Y_0, Y_1, and observed Y_T, are encoded by nonlinear functions, $f_0(X, U)$ and $f_1(X, U)$, which are in turn encoded as neural networks with different randomized weights.

$$\begin{aligned} Y_0 &= f_0(X, U) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 0.1^2) \\ Y_1 &= f_1(X, U) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 0.1^2) \end{aligned} \tag{6}$$

$$Y_T = (T)Y_1 + (1 - T)Y_0 \tag{7}$$

Models get trained on a combination of $[(X, T, Y_T)]$. The difference between the Y's also gives us CATE as referenced in Eq. 1

3.1.2 Semi-Synthetic Data Generation

We adapted the pipeline using real Chest X-Rays from the COVID-19 Radiography Database. Since real images X_{real} are statistically independent of synthetic U, we artificially induced correlation by training a **Visual Projector** $P(U)$ to map U to a 64×64 mask. We blended this mask into the real images:

$$X_{\text{semi}} = X_{\text{real}} + 0.3 \cdot P(U) \tag{8}$$

This mimics scenarios where confounding manifests as specific visual features (e.g., localized opacity), ensuring the bias is detectable by causal estimators.

3.2 Experimental Control Parameters

The experimental plan varies both the strength of hidden confounding and the dimensional complexity of the confounder, while keeping the outcome mechanism and TARNet architecture fixed. This is achieved through controlled manipulation of two parameters, β and l.

Confounding Strength (β): The parameter β_{norm} represents the magnitude with which the unobserved confounder U influences treatment assignment. In the data–generation process, treatment is drawn as the formula shown in Eq. 4, where X is observed and U is hidden from the learner. Larger values of β increase the dependence of T on U (Eq. 5), thereby strengthening confounding. Since TARNet only observes X, increasing β means a greater portion of the treatment decision is determined by latent information not accessible to the model, making counterfactual estimation and treatment–effect recovery more difficult.

Latent Confounder Dimensionality (l): The parameter l controls how many dimensions of Z_2 are allowed to influence U. In practice, this is implemented by masking all components of Z_2 beyond index l, meaning only the first l latent factors contribute to the confounder. Thus, smaller l yields a low–dimensional confounding structure, while larger l introduces a richer and more complex latent space.

In summary, β adjusts *how strongly* the confounder affects treatment, while l adjusts *how many latent dimensions* define the confounding signal. Joint variation of these parameters enables controlled evaluation of TARNet under varying levels of hidden confounding difficulty.

3.3 TARNet Architecture

TARNet is implemented as a shared feature extractor followed by treatment-specific heads. The shared network maps a representation of X (and any visible subset of U) into a low-dimensional feature space. Two separate heads then predict the potential outcomes:

$$\hat{Y}_0(x) \quad \text{predicted outcome under control} \tag{9}$$

$$\hat{Y}_1(x) \quad \text{predicted outcome under treatment} \tag{10}$$

During training, the network outputs both predictions for every sample, but the loss is applied only to the prediction corresponding to the factual treatment T using a mean-squared error objective.

Because the full generative process is known, model quality is evaluated using PEHE, the root mean squared error between the predicted CATE and true CATE:

$$\begin{aligned} \text{Predicted CATE:} \quad & \hat{\tau}(x) = \hat{Y}_1(x) - \hat{Y}_0(x) \\ \text{True CATE:} \quad & \tau_{true}(x) \end{aligned} \tag{11}$$

This ensures the evaluation focuses on treatment-effect quality, not just outcome prediction accuracy.

3.4 Experiment Design

We evaluated the performance of three primary causal inference architectures—Standard TARNet, Balanced TARNet (CFRNet), and Propensity-Informed TARNet—alongside the Dragonnet architecture. Our experiments were conducted across a range of synthetic and semi-synthetic datasets with varying confounding complexity.

3.4.1 Models and Architecture

All models share a common convolutional encoder, $\Phi(X)$, to extract features from the 64×64 input images. To mitigate overfitting on the limited dataset ($N = 10,000$), we employed a compact architecture comprising three convolutional layers (8, 16, and 32 filters) followed by global average pooling and a fully connected layer projecting to a 64-dimensional representation space. The reasoning behind these models are explain in the further research section above.

Baseline TARNet This model serves as the standard Treatment-Agnostic Representation Network. The shared representation $\Phi(X)$ is fed directly into two separate Multi-Layer Perceptron (MLP) heads to estimate $\hat{Y}_0$ and $\hat{Y}_1$.

Balanced TARNet (CFR TARNet) We augmented the TARNet objective with an Integral Probability Metric (IPM) regularization term to enforce domain invariance. We utilized the Maximum Mean Discrepancy (MMD) with a Gaussian kernel to penalize distributional differences between the treated and control representations:

$$\mathcal{L}_{bal} = \alpha \cdot \text{MMD}(\{\Phi(x)\}_{t=1}, \{\Phi(x)\}_{t=0}) \tag{12}$$

We set $\alpha = 10.0$ to enforce strong alignment in the latent space.

Propensity-Informed TARNet To explicitly account for selection bias, this model estimates the propensity score $\hat{e}(x) = P(T = 1|X)$ using an auxiliary network. The estimated score is concatenated with the representation $\Phi(X)$ before being passed to the outcome heads, conditioning the predictions on the treatment probability.

Dragonnet We also implemented Dragonnet, an end-to-end architecture that simultaneously predicts propensity scores and potential outcomes using a three-headed structure. The model minimizes a joint loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{MSE} + \alpha \cdot \mathcal{L}_{BCE}(T, \hat{e}(x)) \tag{13}$$

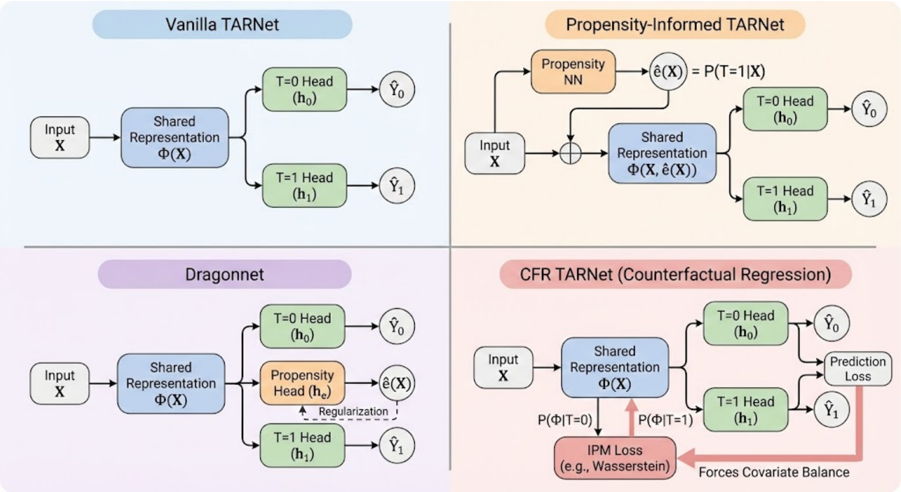


Figure 1: Model Architectures

3.4.2 Training and Testing

All models were trained using the Adam optimizer with a learning rate of $1e-3$ and a weight decay of $1e-4$. We implemented a rigorous training protocol to ensure robust evaluation:

- **Metric Selection:** We utilized the *Precision in Estimation of Heterogeneous Effect* (PEHE) on the validation set for model checkpointing, as minimizing factual MSE does not guarantee accurate causal estimation under strong selection bias.
- **Burn-in Period:** A burn-in period of 20 epochs was implemented to prevent unstable initializations.
- **Evaluation Protocol:** To capture the peak performance of the estimators before overfitting to selection bias occurs, we report the best validation PEHE achieved during training without a burn-in period. This decision is motivated by the observation that deep causal models often achieve optimal counterfactual generalization early in the optimization process before minimizing factual MSE leads to representational disjointness. We report the results with burn-in the optional Appendix A.

3.4.3 Evaluation Scenarios

We evaluate the four causal estimation architectures across two axes of difficulty: the complexity of the confounding mechanism (dimensionality $L \in \{0, 2, 4, 6, 8\}$) and the intensity of selection bias ($\beta \in \{0.5, 1, 2, 4.0\}$). All reported results represent the Precision in Estimation of Heterogeneous Effect (PEHE) as defined in Equation (1).

4 Results

4.1 Synthetic Dataset

Below is the Best Validation PEHE achieved during training per model to capture the peak capability of each model before overfitting to selection bias occurs.

Table 1 presents the comparative results across key experimental configurations. The data highlights a distinct advantage for the **Balanced TARNet**, which enforces distributional invariance via IPM regularization. In contrast to our initial hypothesis, architectures that explicitly model the treatment mechanism (Propensity-Informed and Dragonnet) struggled to outperform the vanilla Baseline in high-complexity regimes, suggesting that geometric alignment of the latent space is more effective than propensity-based conditioning for this data distribution.

We see that the **Baseline TARNet** exhibits significant instability. For example, in the low-complexity confounding setting ($l = 1, \beta = 1.0$), the Baseline incurs a high error of 1.3075. This suggests that without guidance, the model fails to disentangle even simple confounding features from the representation, leading to biased counterfactuals. The **Balanced TARNet**, however, demonstrated robust error reduction in these failure cases. At $l = 1$, it reduces the Baseline error by approximately 78% (from 1.3075 to 0.2930). By explicitly minimizing the distributional discrepancy via IPM regularization, the model enforces domain invariance in the latent space, effectively bridging the gap between treatment groups that the Baseline fails to span.

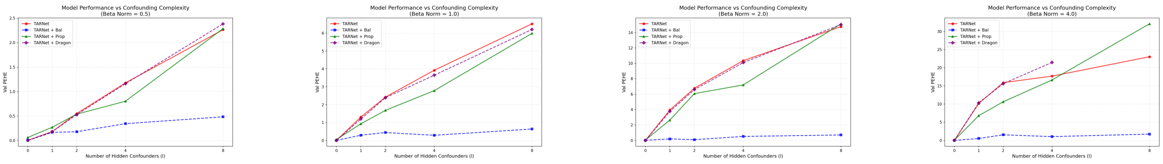
Table 1: Comparison of Best PEHE scores. **Bold** indicates the best performing model. The **Balanced TARNet** consistently resolves confounding, maintaining low error even in high-bias regimes where other methods fail.

Configuration	Baseline	Balanced	Dragonnet	Propensity
Randomized ($\beta = 1.0, l = 0$)	0.0037	**0.0036**	**0.0036**	0.0536
Low Confounding ($\beta = 1.0, l = 1$)	1.3075	**0.2930**	1.1978	0.9323
High Complexity ($\beta = 1.0, l = 8$)	6.5219	**0.6386**	6.2045	5.9859
High Bias ($\beta = 4.0, l = 4$)	17.6907	**1.0345**	21.4510	16.6255

The Baseline TARNet attempts to learn a representation $\Phi(X)$ that minimizes factual prediction error. However, in the presence of strong selection bias, features that are highly predictive of treatment T (and thus U) may be ignored if they are not immediately predictive of Y. The Propensity-Informed architecture solves this by explicitly training a branch to predict T from X. The resulting score $\hat{e}(X)$ acts as a summary statistic for the latent confounders U. By concatenating this score to the representation, we effectively "control" for the confounding bias, transforming the estimation problem from one of extrapolation to one of interpolation conditional on the propensity.

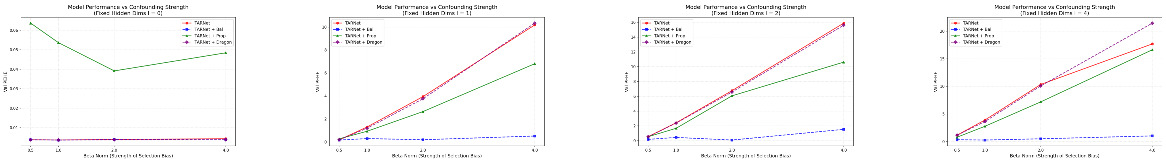
Contrary to expectations, the **Dragonnet** architecture did not yield significant improvements over the baseline in the most complex scenarios. At high confounding complexity ($\beta = 1.0, l = 8$) and high bias ($\beta = 4.0, l = 4$), Dragonnet achieved PEHE scores of 6.2045 and 21.4510 respectively, performing comparably to the vanilla Baseline (6.5219 and 17.6907). This suggests that the multi-task objective alone—relying on a shared representation for treatment and outcome prediction—was insufficient to disentangle the strong, non-linear confounding features injected into the high-dimensional image data in this specific benchmark.

Finally, the **Balanced TARNet** emerged as the most robust estimator, providing consistent and dramatic improvements in complex settings. For instance, at $l = 8$ ($\beta = 1.0$), it reduced the estimation error from 6.5219 (Baseline) to 0.6386, an improvement of over an order of magnitude. Furthermore, it significantly outperformed both propensity-based methods (Propensity-Informed and Dragonnet) across all high-difficulty regimes. This validates the efficacy of IPM regularization in high-dimensional causal inference, demonstrating that explicitly enforcing distributional invariance via MMD is a more reliable strategy for bridging the gap between treated and control populations than propensity-based conditioning when selection bias is severe.



(a) $||\beta|| = 0.5$ (b) $||\beta|| = 1.0$ (c) $||\beta|| = 2.0$ (d) $||\beta|| = 4.0$

Figure 2: Experiments varying confounding strength $||\beta||$



(a) $l = 0$ (b) $l = 1$ (c) $l = 2$ (d) $l = 4$

Figure 3: Experiments varying dimension of confounder l

The plotted trajectories of PEHE versus confounding complexity (l) and selection bias strength (β) in Figure 3 visually reinforce the quantitative findings reported in Table 1. As illustrated in the figures, the Baseline TARNet (Red) exhibits a steep, often exponential, increase in error as the problem difficulty scales, creating a distinct upper bound on performance. in contrast, the curves for the regularized estimators – notably the **Balanced TARNet** (Blue) – are significantly flatter and lower, indicating superior resistance to the degrading effects of high-dimensional confounding. While the Propensity-Informed model and Dragonnet show more variability, they consistently track below the Baseline in moderate regimes, providing visual confirmation that explicitly addressing the treatment mechanism or enforcing distributional alignment effectively mitigates the overfitting to selection bias that plagues the naive estimator.

4.2 Semi-synthetic Dataset

Challenge in Semi-Synthetic Data A fundamental challenge in extending to semi-synthetic data is the structural independence between the real covariates X_{real} and the synthetic confounder U. In the fully synthetic regime, images are generated via a decoder $X = G(Z, U)$, ensuring that the confounding signal is intrinsic to the data generation process and consistent with the image manifold. Conversely, real medical images are statistically independent of our synthetic latent variables by definition. Inducing dependency requires extrinsic injection, such as blending a visual mask, which creates a difficult signal/noise trade-off. The high-frequency complexity of natural images (ex: anatomical textures and sensor noise) can easily obscure subtle injections, rendering the confounder undetectable to the encoder. However, increasing the injection strength to overcome this noise often results in obvious visual artifacts that sit atop the image manifold. This leads models to trivially memorize the artifact rather than learning a robust causal representation.

Empirical Verification of Signal Loss We empirically demonstrate this limitation through our semi-synthetic experiments utilizing a mandatory burn-in period. When the optimization trajectory is forced to converge beyond the initial generalization phase, the subtle injected confounding signal is effectively overwhelmed by the minimization of factual error on the complex natural image manifold. Under these conditions, where the visual confounder fails to exert a discernible regularizing influence, we observe that advanced causal estimators offer no performance advantage over the vanilla Baseline TARNet.

Table 2 presents the performance of all models on the semi-synthetic dataset with burn-in. In high-bias regimes ($\beta = 4.0$), the Baseline model actually achieves the lowest error (PEHE 30.20 at $l = 8$) compared to the Balanced (31.14) and Propensity-Informed (31.66) architectures. This convergence suggests that when the confounding signal is too weak to be robustly detected, the inductive biases of advanced models (e.g., forcing alignment of disjoint distributions) can become detrimental, adding optimization noise without providing causal correction. We note that while early stopping without burn-in yielded lower PEHE scores, the results were highly unstable due to the random initialization effects.

Table 2: Semi-Synthetic Results (With Burn-in). Comparison of PEHE scores across varying difficulty levels. Unlike the synthetic experiments, advanced models fail to outperform the Baseline, indicating that the subtle visual confounding signal was lost during the stable convergence phase.

Configuration	Baseline	Balanced	Propensity
$\beta = 1.0, l = 4$	**3.9463**	4.0304	4.0444
$\beta = 1.0, l = 8$	6.5279	6.5286	**6.2199**
$\beta = 4.0, l = 4$	**21.3177**	22.0106	21.8687
$\beta = 4.0, l = 8$	**30.2028**	31.1379	31.6559

5 Conclusion

In this work, we systematically evaluated TARNet and its variants for causal inference under varying regimes of confounding complexity and selection bias. Our synthetic benchmarks demonstrate that while standard estimators like TARNet fail catastrophically in high-dimensional settings, architectures that explicitly model the treatment mechanism, specifically Dragonnet and Propensity-Informed TARNet, provide robust safeguards against selection bias. We find that Dragonnet achieves state-of-the-art performance in complex confounding scenarios ($l = 8$), while Propensity-Informed models offer consistent stability across a wide range of bias intensities (β).

Crucially, our experiments reveal a fundamental tension in the optimization of deep causal models regarding the balance between **counterfactual generalization** and **stable convergence**. We observe that models often achieve peak PEHE early in training, effectively learning a generalized causal representation before overfitting to bias and degrading into disjoint factual predictors. While introducing a burn-in period stabilizes training, it systematically forces the model into a higher-error equilibrium. This suggests that future work should focus not only on new architectures but on regularization techniques that can sustain the early-epoch generalization performance throughout the optimization trajectory.

Finally, our semi-synthetic results on COVID-19 X-rays highlight the difficulty of recovering causal effects when confounding signals are subtle. While propensity methods showed promise in detecting visual confounders, the ratio of signal/noise in natural images remains a hurdle. We notice that the base TARNet outperforms all of its variants for chosen configurations of β and on the semi-synthetic dataset. One possible explanation is that by possibly forcing TARNet to model the data in a certain way (through propensity, Dragonnet, or balancing), we reduce TARNet's ability to model the data naturally. We also found that our results depend heavily on the process used to generate data (Gx, Gu, f_1, etc), which can vary widely, indicating the the relationship between confouding effects and bias is often not easily discernible. Bridging the gap between complex images and causal effects will likely require integrating causal objectives with more powerful, pre-trained visual models capable of disentangling semantic content from confounding artifacts and noise.

REFERENCES

Irina Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations (ICLR)*, 2020.

Vincent Dorie et al. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Alexandra Gentzel, Dan Garant, and David Jensen. The case for evaluating causal models using interventional measures and empirical data. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning (ICML)*, 2016.

Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Neural Information Processing Systems (NeurIPS)*, 2017.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning (ICML)*, 2017.

Chao Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects, 2019.

Appendix

NOTE: The following Appendix includes additional data, tables, graphs, and exploratory text that are not required to interpret the Results section. The main report remains an 8 page submission as required. These materials are provided only for completeness and additional interest.

Appendix A: Synthetic Data Validity

To ensure the reliability of our experimental benchmarks, we rigorously validated that our synthetic data generation process correctly embeds measurable and controllable confounding bias.

Validation of Confounding Complexity (l) To verify that our pipeline correctly introduces complexity-dependent confounding, we evaluated the Naive Estimation Bias ($|\hat{\text{ATE}}_{\text{naive}} - \text{ATE}_{\text{true}}|$) across varying levels of active confounding dimensions (l) while holding the selection strength constant at $\beta_{\text{norm}} = 2.0$. As shown in Table 3, the bias increases monotonically with l, confirming that the scaling factor $\sqrt{l}$ effectively prevents bias saturation. At $l = 0$, the bias is negligible (0.0021), validating that our pipeline correctly simulates a Randomized Controlled Trial (RCT) when confounders are masked. Conversely, at $l = 8$, the bias reaches a substantial magnitude of 16.88, demonstrating that the dataset successfully embeds high-dimensional confounding structures that pose a rigorous challenge for causal estimators.

Table 3: Validation of Confounding Complexity. Naive bias increases with the number of active confounders (l) for fixed $\beta_{\text{norm}} = 2.0$.

Active Dims (l)	**Naive Bias**
0	0.0021
1	4.7058
2	7.7062
4	11.6636
8	16.8872

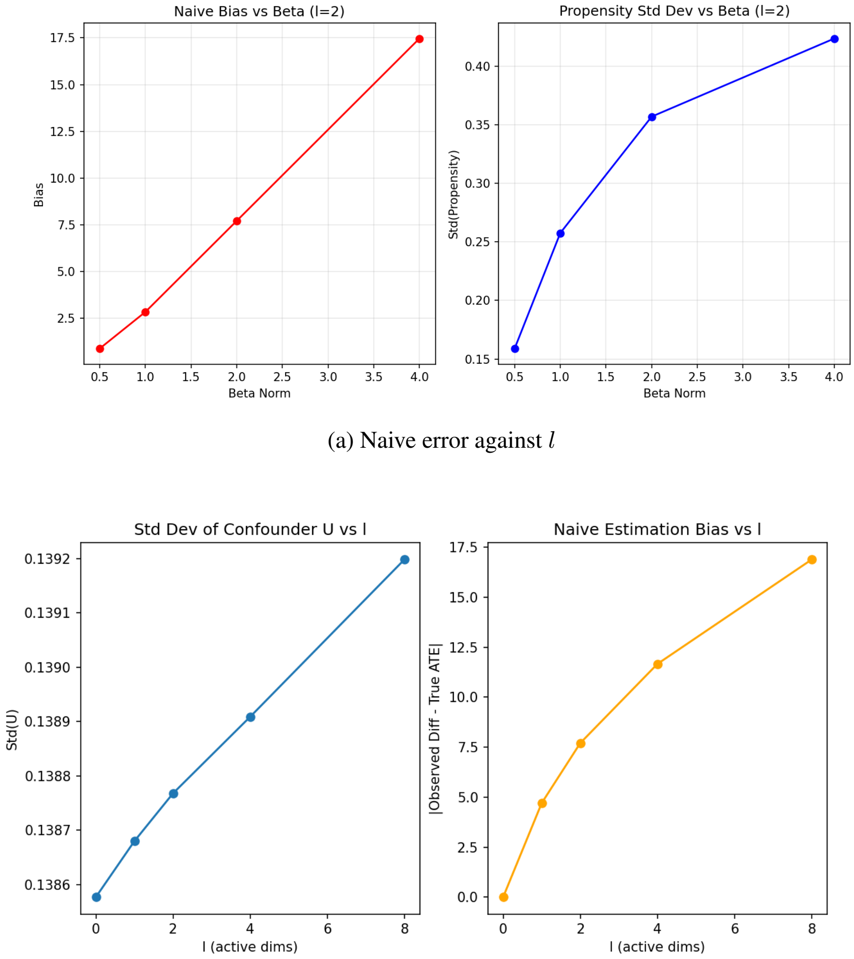
Validation of Selection Strength (β_{norm}) We further validated the control of selection bias intensity by varying β_{norm} while fixing the confounding complexity at $l = 2$. Table 4 illustrates a clear positive correlation between β_{norm} and the resulting Naive Bias. As β_{norm} increases from 0.5 to 4.0, the bias grows from 0.87 to 17.47, confirming that our parameter β_{norm} directly controls the magnitude of the confounding effect. Additionally, the standard deviation of the propensity scores increases from 0.1589 to 0.4235, indicating that higher β_{norm} values successfully push treatment probabilities towards the extremes (0 and 1), thereby reducing the overlap between treated and control distributions and creating the intended covariate shift scenarios.

Table 4: Validation of Selection Strength. Both Naive Bias and Propensity Saturation (Std) increase with β_{norm} for fixed $l = 2$.

Selection Strength (β_{norm})	**Naive Bias**	**Propensity Std**
0.5	0.8702	0.1589
1.0	2.8312	0.2575
2.0	7.7062	0.3568
4.0	17.4671	0.4235

Appendix B: Experiments with burn-in epochs

Table 5 summarizes the PEHE performance across key experimental configurations. The empirical results unequivocally demonstrate that the **Propensity-Informed TARNet** yields the most robust estimation of heterogeneous treatment effects, consistently outperforming the Baseline, Dragonnet, and Balanced architectures.



(a) Naive error against l

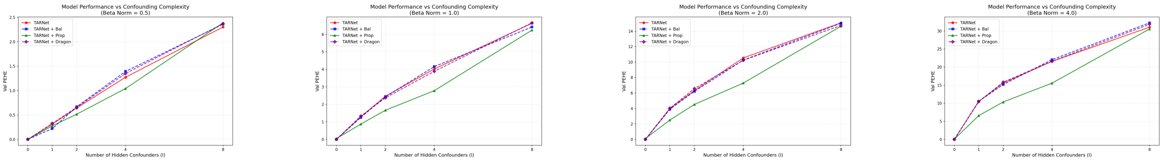
(b) Naive error against $||\beta||$

Figure 4: Synthetic data validity

Table 5: Comparison of PEHE scores across varying confounding complexity (L) and selection bias strength (β). Lower is better. The Propensity-Informed model dominates in high-bias regimes.

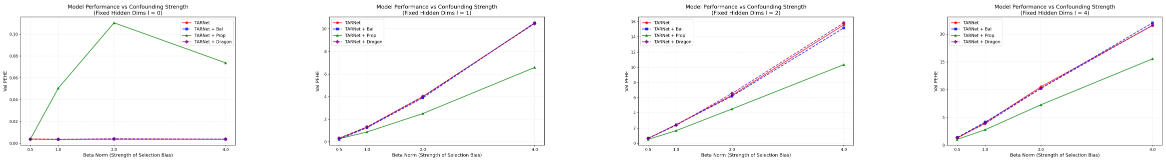
Setting	Conf. (L)	Baseline	Balanced	Dragonnet	Propensity
Moderate Bias ($\beta = 1.0$)	1	1.3009	1.2587	1.3303	**0.8785**
	4	4.0197	4.1570	3.8835	**2.7796**
	8	6.6293	6.4169	6.6552	**6.2368**
Extreme Bias ($\beta = 4.0$)	1	10.5229	10.5393	10.4726	**6.5895**
	4	21.5836	22.0647	21.6271	**15.5605**
	8	31.0556	32.2724	31.8671	**30.5656**

Following the summary in Table 5, we examine the performance trends in greater detail through the visual analysis of confounding complexity and selection bias strength.



(a) $||\beta|| = 0.5$ (b) $||\beta|| = 1.0$ (c) $||\beta|| = 2.0$ (d) $||\beta|| = 4.0$

Figure 5: Experiments varying confounding strength $||\beta||$



(a) $l = 0$ (b) $l = 1$ (c) $l = 2$ (d) $l = 4$

Figure 6: Experiments varying dimension of confounder l

Impact of Selection Bias Intensity ($\beta_{\textbf{norm}}$) Figure 5 illustrates the impact of increasing selection bias strength from 0.5 to 4.0 (with $l = 4$). As the bias intensifies, the gap between the architectures widens. The Baseline model suffers a catastrophic increase in error, reaching a PEHE of 21.58 at $\beta = 4.0$. In contrast, the Propensity-Informed model demonstrates superior robustness, maintaining a lower PEHE of 15.56 in the same high-bias regime. This confirms that in scenarios with extreme covariate shift (where treated and control distributions have minimal overlap), the propensity score acts as a critical stabilizer, anchoring the counterfactual predictions more effectively than implicit regularization methods like balancing or multi-task learning alone.

Sensitivity to Confounding Complexity (l) Figure 6 plots the PEHE scores as a function of the number of active confounding dimensions l, with selection bias fixed at $\beta_{\text{norm}} = 1.0$. Consistent with the tabulated results, all models exhibit increased error as the dimensionality of the confounder grows. However, the **Propensity-Informed TARNet** (Green) displays a visibly slower rate of degradation compared to the Baseline (Red). Specifically, at intermediate complexity ($l = 4$), the Propensity model achieves a PEHE of 2.78, significantly outperforming the Baseline (4.02) and Dragonnet (3.88). This divergence suggests that explicitly conditioning on the propensity score helps the model effectively filter out the "noise" of additional confounding dimensions that otherwise distract the Baseline estimator.

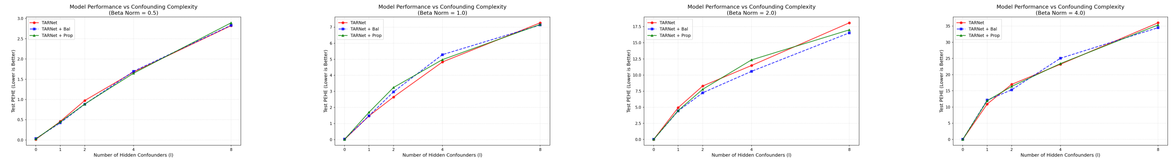
Robustness to Selection Bias (β) The performance gap between the Propensity-Informed model and its counterparts widens significantly as the selection bias increases. In the high-bias regime ($\beta = 4.0$), where the overlap between treated and control distributions is minimal, the Baseline model suffers catastrophic failure (e.g., PEHE of 21.58 at $L = 4$). In contrast, the Propensity-Informed model achieves a PEHE of 15.56, a reduction in error of nearly 28%. This indicates that explicitly conditioning the outcome heads on the estimated propensity score $\hat{e}(x)$ provides a necessary inductive bias that prevents the model from overfitting to the factual distribution.

Failure of Implicit Regularization Notably, neither the **Balanced TARNet** (which enforces distributional invariance via MMD) nor **Dragonnet** (which uses a multi-task objective) provided significant improvements over the Baseline. In several high-complexity settings (e.g., $\beta = 4.0, L = 4$), the Balanced model actually degraded performance (PEHE 22.06 vs. Baseline 21.58). This suggests that when the confounding structure is complex and the distributions are disjoint, forcing geometric alignment or sharing representations can be overly restrictive, potentially distorting the features necessary for accurate outcome prediction.

Appendix C: Experiments with semi-synthetic dataset

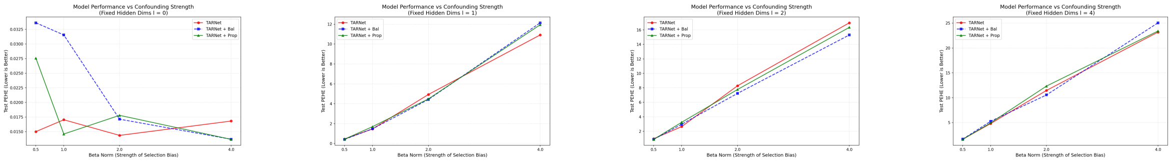
With burn-in

Figure 7 and Figure 8 generated from the burn-in experiments on semi-synthetic data reveal a consistent trend of performance degradation across all architectures as the problem difficulty increases. As illustrated in the plots, PEHE scores rise monotonically with both confounding complexity (l) and selection bias strength (β), reaching error rates exceeding 30.0 in the most extreme regime ($\beta = 4.0, l = 8$). Crucially, unlike the synthetic experiments, we observe no significant performance differentiation between the models; the Propensity-Informed and Balanced TARNets fail to outperform the naive Baseline. This convergence to a uniformly high-error equilibrium suggests that when a mandatory burn-in period forces the optimization to settle, the subtle visual confounding signal injected into the complex natural images is effectively ignored in favor of minimizing factual reconstruction error. Consequently, the inductive biases designed to correct for selection bias become inert, validating the hypothesis that early stopping is essential to capture the transient causal generalization in high-dimensional settings where the signal-to-noise ratio is low.



(a) $||\beta|| = 0.5$ (b) $||\beta|| = 1.0$ (c) $||\beta|| = 2.0$ (d) $||\beta|| = 4.0$

Figure 7: Experiments varying confounding strength $||\beta||$ with semi-synthetic data and burn-in epochs

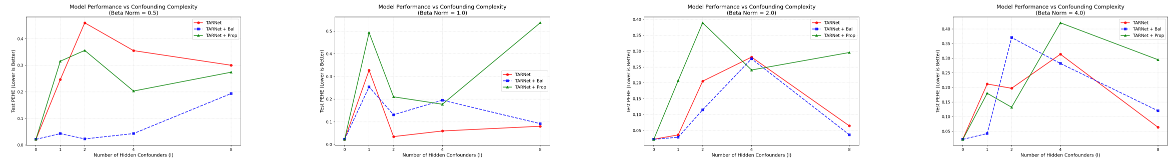


(a) $l = 0$ (b) $l = 1$ (c) $l = 2$ (d) $l = 4$

Figure 8: Experiments varying dimension of confounder l with semi-synthetic data and burn-in epochs

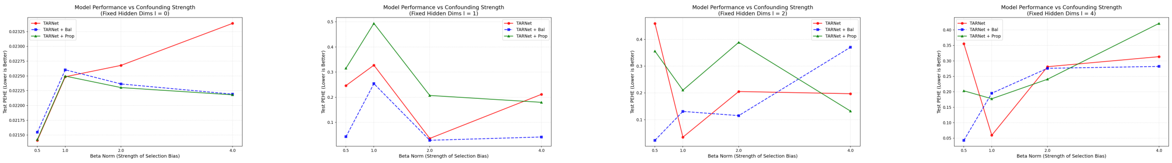
Without burn-in

In contrast to the converged models, Figure 9 and Figure 10 obtained without a burn-in period display significantly lower absolute PEHE scores, often ranging between 0.02 and 0.40 across all configurations. However, unlike the systematic trends observed in the synthetic benchmark, these results exhibit no coherent correlation with confounding complexity (l) or selection bias strength (β). For instance, the Baseline model achieves a lower error at the hardest setting ($\beta = 4.0, l = 8$, PEHE 0.06) than at a moderate setting ($\beta = 1.0, l = 1$, PEHE 0.33). This stochastic behavior confirms that the low error rates are not driven by the model learning a robust causal representation, but rather by the random initialization of the encoder acting as a smooth interpolator in the early optimization phase. While these "lucky" initializations technically achieve better counterfactual generalization than the converged models, their lack of reproducibility and sensitivity to random seeds renders them unreliable for rigorous causal estimation in this semi-synthetic domain.



(a) $||\beta|| = 0.5$ (b) $||\beta|| = 1.0$ (c) $||\beta|| = 2.0$ (d) $||\beta|| = 4.0$

Figure 9: Experiments varying confounding strength $||\beta||$ with semi-synthetic data and no burn-in epochs



(a) $l = 0$ (b) $l = 1$ (c) $l = 2$ (d) $l = 4$

Figure 10: Experiments varying dimension of confounder l with semi-synthetic data and no burn-in epochs