I See, Therefore I Do: Estimating Causal Effects for Image Treatments

Abhinav Thorat[*]
Sony Research
India
abhinav.thorat@sony.com

Ravi Kolla[*]
Sony Research
India
ravi.kolla@sony.com

Niranjan Pedanekar
Sony Research
India
niranjan.pedanekar@sony.com

Abstract

Causal effect estimation under observational studies is challenging due to the lack of ground truth data and treatment assignment bias. Though various methods exist in literature for addressing this problem, most of them ignore multi-dimensional treatment information by considering it as scalar. Recently, certain works have demonstrated the utility of this rich yet complex treatment information into the estimation process, resulting in better causal effect estimation. However, these works have been demonstrated on either graphs or textual treatments. There is a notable gap in existing literature in addressing higher dimensional data such as images that has a wide variety of applications. In this work, we propose a model named **NICE** (**N**etwork for **I**mage treatments **C**ausal effect **E**stimation), for estimating individual causal effects when treatments are images. NICE demonstrates an effective way to use the rich multidimensional information present in image treatments that helps in obtaining improved causal effect estimates. We then provide theoretical guarantees of NICE performance by deriving an upper bound on **PEHE** (**P**recision in the **E**stimation of **H**eterogeneous **E**ffects). To evaluate the empirical performance of NICE, we propose a novel semi-synthetic data simulation framework that generates Potential Outcomes (POs) when images serve as treatments. Empirical results on these datasets, under various setups including the *zero-shot* case, demonstrate that NICE significantly outperforms baselines.

Keywords

Causal Inference, Individual treatment effect estimation

ACM Reference Format:
Abhinav Thorat, Ravi Kolla, and Niranjan Pedanekar. 2018. I See, Therefore I Do: Estimating Causal Effects for Image Treatments. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Causal effect estimation in observational studies aims to understand the impact of specific treatments on particular outcomes using observed data [12, 14]. It remains a critical and challenging problem in the field of causal inference. It has widespread applications to domains such as healthcare [22], economics [20], social sciences [9], education [6], entertainment [25] and e-commerce [2]. In this work, we specifically study **I**ndividual **T**reatment **E**ffect (ITE) estimation for image treatments. ITE focuses on estimating the impact of a treatment at an individual user level, as opposed to average causal effects, which apply to entire populations or sub-populations. ITE helps in personalizing treatments based on user attributes. Some of the use-cases of ITE across various domains including personalization of content, product, and investment plan recommendations in the entertainment, e-commerce, and finance industries, respectively.

In causal effects estimation literature, majority of the works represent treatments in a one-hot encoding format or as categorical in nature. But, often these treatments can be multi-dimensional such as images, text and graphs, and contain rich auxiliary information. If made available, they can potentially be used to improve the causal effects estimation. This raises a question whether the causal effects estimates can be improved by utilizing treatments' auxiliary information in the estimation process. To that end, there are a limited number of works [4, 10, 11] in the literature that try to address the above question. These works have demonstrated ways of effective utilization of treatments' auxiliary information in the ITE estimation and showcased improved results. However, all these works have considered either graph or textual treatments, in their respective experiments. In this work, we study the ITE estimation problem for image treatments and demonstrate an effective way to utilize their auxiliary information, resulting in improved ITE estimates.

We now provide motivation for our problem, ITE estimation for image treatments, with the following prevalent use-cases in our daily lives. Consider an OTT (Over-The-Top) or a video streaming application that displays its contents using thumbnails to the users. Suppose each content is present in multiple thumbnail variants and the goal is to personalize thumbnails for users to maximize the user engagement. This problem can be posed as an ITE estimation problem where thumbnails and user preferences to them correspond to treatments and their effects respectively.

Consider another application in e-commerce that sells products by displaying product images on their platform. Each product typically features multiple images (photos) captured from various angles and under different lighting conditions. Suppose, if the goal is to personalize product display images to maximize click through rates then it can be approached using our framework by considering product images as treatments and estimating users' preferences as treatment effects.

We now briefly talk about the key challenges in our work. The first and foremost challenge is the lack of existing datasets for the ITE estimation of image treatments. It necessitated us to simulate a new dataset that required extensive research and experimentation in

[*] Both authors contributed equally to this research.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

terms of the appropriate image data and mathematical formulation of their induced effects on users. Next, even with the few existing works [4, 10, 11] in the literature that utilize treatments' auxiliary information none of them show empirical results with images as treatments. Finally, our setup containing multiple treatments under observational studies increases the possibility of confounding bias in the data due to the imbalanced assignment of treatments to users.

We now outline the salient contributions of our work that address the aforementioned challenges.

- First, we begin by noting that our problem setup, which involves images as treatments, itself is novel. The broader area of incorporating treatments' auxiliary information into ITE estimation is also a relatively recent development, with GraphITE [4] being the first work to appear in year 2021. Since then, only a limited number of works have explored this area.
- Second, we propose a semi-synthetic data simulation setup that generates Potential Outcomes (POs) for the case of multiple image treatments.
- Third, we propose a novel architecture, NICE, for ITE estimation of image treatments with a combination of MSE and **M**aximum **M**ean **D**iscrepancy (MMD) losses.
- Fourth, we provide theoretical guarantees for a broad class of algorithms, including NICE, all of which share the same properties, by deriving an upper bound on an error estimate, specifically PEHE.
- Next, we showcase the superior performance of NICE against baselines on PEHE across various treatment assignment bias conditions in numerical experiments.
- Finally, we also conduct experiments under zero-shot scenarios where models are evaluated on unseen treatments during training. Under these settings too, we observe that NICE outperforms baselines by a significant margin.

2 Literature Survey

In this work, we deal with the ITE estimation at individual user level as opposed to the predominantly studied Average Treatment Effects (ATE) [13, 18], estimated at the whole population level in the literature. Specifically, we study ITE estimation under multiple image treatments setup by utilizing treatments' auxiliary information in the estimation hence we restrict ourselves contrasting our work with only ITE estimation under multiple treatments and works that utilized treatments' auxiliary information in the estimation. Most of the works that involve multiple treatments [3, 15, 16, 23, 24] do not consider the rich treatment information in the estimation and merely represent them as scalars using one hot encoding.

There are few existing works [4, 10, 11] that utilized treatments' auxiliary information in the ITE estimation under multiple treatments setup. The authors in [4] considered the problem of ITE estimation for multiple graph treatments, different from our setup of image treatments, and demonstrated ways to utilize the graph treatments information to obtain improved causal effect estimates. The work in [10] deals with the ITE estimation for structured treatments such as graphs, images and text by incorporating their information in the estimation process. However, their algorithm, SIN, is evaluated only on the datasets with graph treatments in their experiments. SIN's performance on image treatments datasets has not been studied yet, making SIN one of the baselines for our work. The following latest work [11] also utilizes treatments' auxiliary information for ITE estimation. However, it primarily addresses zero-shot tasks, where the model estimates the causal effects of treatments not seen during training.

3 Problem Formulation

In this section, we briefly outline the problem considered in this work. Let $k \in \mathbb{N}$ and $n \in \mathbb{N}$ denote the number of available treatments and the number of instances or users. We use i, x and t for referencing users, their covariates and treatments respectively. Let $x_i \in \mathcal{X} \subset \mathbb{R}^d$, $t_i \in \{1, 2, \cdots, k\}$, denote covariates, index of assigned treatment of user-i respectively. We use $I_t \in \mathcal{I} \subset \mathbb{R}^m$ to denote the actual image corresponding to the treatment-t.

We follow the Rubin-Neyman [14] POs framework for introducing the problem. Let $Y_{i,t} \in \mathcal{Y}$ denotes the POs of user-i when treatment-t is applied. Since t_i is the treatment given to user-i, Y_{i,t_i} denotes the observed (factual) outcome of user-i. For brevity purposes, we write Y_{i,t_i} as $Y_{i,t}$. Further, whenever the user is understood from the context then we write $Y_{i,t}$ as just Y_t. Given user-i with covariates x_i, and a pair of treatments a, b, define the ITE, using the notation $\tau^{a,b}(x_i)$, as below:

$$\tau^{a,b}(x_i) = \mathbb{E}\left[Y_{i,a} \mid x = x_i\right] - \mathbb{E}\left[Y_{i,b} \mid x = x_i\right]. \tag{1}$$

In our setup, we assume that the treatments are images and are available to the model. In the following, we provide the technical formulation of our problem statement. Given n observations, $\{x_i, I_{t_i}, t_i, Y_{i,t_i}\}_{i=1}^n$, of users with covariates x_i, their assigned image treatments, I_{t_i}, treatment indices, t_i, and the corresponding observed POs, Y_{i,t_i} our goal is to estimate ITEs, given in Equation (1), for all pairs of treatments.

We quantify a model's performance using the standard metric in the literature named PEHE, defined as follows [16]:

$$\epsilon_{\text{PEHE}} = \frac{1}{\binom{k}{2}} \sum_{a=1}^{k} \sum_{b=1}^{a-1} \left[\frac{1}{n} \sum_{i=1}^{n} (\hat{\tau}^{a,b}(x_i) - \tau^{a,b}(x_i))^2 \right], \tag{2}$$

where $\hat{\tau}(\cdot)$ represents the estimated ITEs produced by the model.

4 Proposed Model

The NICE framework, given in Algorithm 1, designed for ITE estimation with image-based treatments, relies on the following three standard assumptions [7] in the literature: Unconfoundedness (Conditional Independence), Positivity (Overlap) and Stable Unit Treatment Value Assumption (SUTVA). Our proposed model, NICE, addresses ITE estimation by utilizing treatments' auxiliary information, specifically images. Figure 1 illustrates the detailed architecture of the NICE model, comprising three key steps mentioned below.

A. Generating representations for both covariates and treatments simultaneously then concatenating them.
B. Employing individual treatment head networks to generate counterfactual estimates.
C. Computing a regularization loss to mitigate the treatment assignment bias along with regression loss to ensure accurate predictions.

Detailed explanations of our model architecture are as follows.

Algorithm 1: NICE Training

```
Input: Observational data: $\mathcal{D} = \{(x_i, I_{t_i}, t_i, Y_{i,t})\}_{i=1}^{n} \sim \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}$, and hyper parameters $\alpha \geq 0$ and $\beta \geq 0$.
Output: An outcome prediction model: $f(\Phi, \Psi, \Pi)$,
where $\Pi = (\Pi_1, \Pi_2, \cdots, \Pi_k)$
Initialize parameters: $\Phi$ : MLP, $\Psi$ : MLP, $\Pi$ : MLPs
while not converged do
    Sample a mini-batch
    $B = \{(x_{i_o}, I_{t_{i_o}}, Y_{i_o,t_{i_o}})\}_{o=1}^{B} \subset \mathcal{D}_{\text{train}}$
    Mini-batch approximation of Regression Loss
    $\mathcal{L}_1 = \frac{1}{|B|} \sum_{o=1}^{|B|} (\hat{Y}_{i_o,t_{i_o}} - Y_{i_o,t_{i_o}})^2$
    Mini-batch approximation of the Treatment Regularization Loss
    $\mathcal{L}_2 = \frac{1}{\binom{k}{2}} \sum_{a=1}^{k} \sum_{b=1}^{a-1} \text{MMD}(\{\Phi\}_{t=a}, \{\Phi\}_{t=b})$
    Update Functions:
    $f(\Phi, \Psi, \Pi) \leftarrow f(\Phi, \Psi, \Pi) - \lambda.\nabla(f(\Phi, \Psi, \Pi))$
    Minimize $\alpha \cdot \mathcal{L}_1 + \beta \cdot \mathcal{L}_2$ using SGD
end while
```

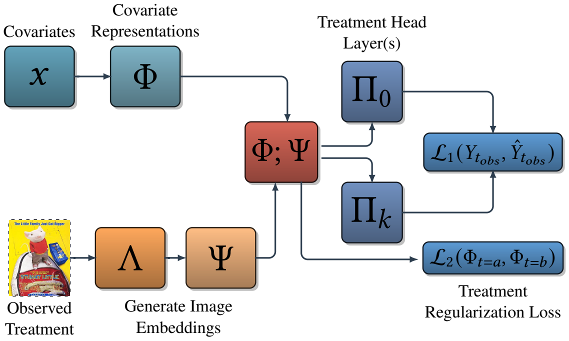
Covariates
Covariate Representations
Treatment Head Layer(s)
x
Φ
Π_0
$\Phi; \Psi$
$\mathcal{L}_1(Y_{t_{obs}}, \hat{Y}_{t_{obs}})$
Π_k
Λ
Ψ
$\mathcal{L}_2(\Phi_{t=a}, \Phi_{t=b})$
Observed Treatment
Generate Image Embeddings
Treatment Regularization Loss

Figure 1: NICE : Network for Image treatments Causal effect Estimation

4.1 Learning Representation of User Covariates and Observed Image Treatments

For consistent estimation of causal effects, in line with Assumption (1) of [17], we choose two functions in the NICE architecture for representing covariates and image treatments. To that end, we employ two distinct Fully Connected (FC) networks to learn representations of covariates, $x \in \mathcal{X}$, and observed image treatments $I_t \in \mathcal{I}$, capturing their low-dimensional embeddings. Specifically, we define two functions, $\Phi : \mathcal{X} \to \mathbb{R}^{d_1}$ and $\Psi : \Lambda(\mathcal{I}) \to \mathbb{R}^{d_2}$, to extract representations for covariates and treatment images.

The utility of learning covariate representations to enhance causal effect estimation has been previously demonstrated in the literature [17]. Similarly, learning treatment representations has been explored, particularly in graph-based contexts, as in Graphite [4], SIN [10] and CaML [11] algorithms. In our approach, we first use an existing image embedding model, denoted by Λ, for obtaining image embeddings. Then, these image embeddings are fed to a representation network, Ψ. Currently, we considered two popular well studied models in the literature such as ResNet [5] and VGG [19] as candidates for Λ in the NICE model. In particular, Λ is used solely to infer image treatment embeddings and is not a trainable component in the NICE model.

Note that, the standard multiple treatment setting utilizes one-hot encoding of discrete treatments. However, this approach fails to leverage the rich structural information inherent to image treatments and consequently suffers in causal effect estimates. Further, we concatenate the covariates and treatment representations to create a joint embedding, which is then utilized by the treatment head networks for ITE estimation.

4.2 Treatment Head Networks

In the second part of our model, we leverage concatenated embeddings of user covariates and treatments representations as a unified input to distinct treatment head networks corresponding to each treatment category. Given k available treatments, we train k number of FC networks to learn the functions for each individual treatment, aimed at estimating the POs. We denote these treatment head networks as Π_t for $t \in \{1, 2 \cdots, k\}$. Mathematically, for a user$-i$ with covariates x_i and observed treatment $t_{\text{obs}} = t$, Π_t is defined as:

$$\Pi_t\left(\Phi(x_i), \Psi(\Lambda(I_t))\right) = w_t \sigma\left(W_t^l \cdots \sigma\left(W_t^1\left(\Phi(x_i), \Psi(\Lambda(I_t))\right)\right)\right),$$

where W_t^l and w_t represent the weights of the l-th FC layer and the regression layer in the network head-t, respectively. The neural network bias terms follow the same rule and are omitted here for simplicity. With both components of the model described, our model's prediction of the PO for treatment t given instance i is defined as: $\hat{Y}_{i,t} = \Pi_t(\Phi(x_i), \Psi(\Lambda(I_t)))$.

4.3 Loss Function

We optimize NICE using a weighted combination of regression loss ($\mathcal{L}_1$) and treatment regularization loss ($\mathcal{L}_2$), with the total loss defined as $\mathcal{L} = \alpha\mathcal{L}_1 + \beta\mathcal{L}_2$, where α and β are tunable hyperparameters. We detail each loss component below.

To achieve high predictive accuracy on observed outcomes, we employ the traditional mean square error loss. Given that each treatment group exhibits a unique distribution, optimizing the regression loss enables us to capture the approximate means for each treatment group. Specifically, we optimize the head network corresponding to the observed treatment t. The regression loss function $\mathcal{L}_1$ is defined as: $\mathcal{L}_1 = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_{i,t} - Y_i)^2$.

We use the treatment regularization loss computed using covariate representations, to address treatment assignment bias — a crucial step in our model. Considering that the covariate distributions across treatments may differ, our goal is to achieve a balanced representation that accounts for treatment assignment bias. To that end, special cases of Integral Probability Metrics (IPM) have been utilized in the literature to obtain a balanced representation [8]. We employ a special case of IPM, specifically MMD loss [21], to obtain a balanced representation of covariates across all treatments. In particular, the treatment regularization loss computes the average MMD distance between covariate representations across all treatment pair combinations, whose mathematical formulation is given as: $\mathcal{L}_2 = \frac{1}{\binom{k}{2}} \sum_{a=1}^{k} \sum_{b=1}^{a-1} \text{MMD}(\{\Phi\}_{t=a}, \{\Phi\}_{t=b})$, where $\{\Phi\}_{t=a}$ denotes covariate representation for respective treatment subgroup.

This approach aims to optimize covariate representations, thereby mitigating the confounding effects.

5 Theoretical guarantees

In this section, we consider an arbitrary family of algorithms that share the same structure as NICE and derive an upper bound on PEHE. We begin by deriving bounds for the case of binary treatment i.e., $t = 0$ (control) or 1 (test). In our case of image treatments, $t = 0$ can be interpreted as providing some default image treatment to the user. We assume that given t, the corresponding I_t is deterministic and known. We assume there exists a joint distribution $p\left(x, t, Y_1, Y_0\right)$ that satisfies unconfoundedness, positivity, and SUTVA. Let $p^{t=1}(x) \triangleq p\left(x \mid t=1\right)$ and $p^{t=0}(x) \triangleq p\left(x \mid t=0\right)$ denote the covariate distributions under each treatment. Note that our theoretical results follow the framework established by [17].

Assumption 5.1. The representation functions $\Phi : \mathcal{X} \to \mathcal{R}_X$ and $\Psi : \mathcal{I} \to \mathcal{R}_I$ are twice differentiable and invertible.

Definition 5.2. Define $p_\Phi^{t=1}(r_x) \triangleq p_\Phi\left(r_x \mid t=1\right)$ and $p_\Phi^{t=0}(r_x) \triangleq p_\Phi\left(r_x \mid t=0\right)$ are the conditional probability distributions induced by the function Φ on representation space $\mathcal{R}_X$. Note that, as Φ is an invertible function, the induced probability distributions can be directly derived using change of variables formula.

Let $\Pi : \mathcal{R}_X \times (\mathcal{R}_I; \{0,1\}) \to \mathcal{Y}$[1] be a hypothesis function that maps $\Phi(x)$, $\Psi(I_t)$, and t to a scalar prediction. We decompose Π into Π_0 and Π_1, corresponding to the treatment assignment $t \in \{0,1\}$. Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ denote a loss function. We define two losses: one on the factual and another on the counterfactual data.

Definition 5.3. Let the expected loss for user x, the assigned treatment index t and the corresponding actual treatment I_t be defined as: $l_{\Pi,\Phi,\Psi}\left(x, I_t, t\right) = \int_{\mathcal{Y}} L\left(Y_t, \Pi\left(\Phi(x), \Psi(I_t), t\right)\right) p\left(Y_t \mid x\right) dY_t$. Then, define expected factual and counterfactual losses as follows:

$$\epsilon_{\mathrm{F}}\left(\Pi, \Phi, \Psi\right) = \int_{\mathcal{X}\times\{0,1\}} l_{\Pi,\Phi,\Psi}\left(x, I_t, t\right) p(x,t) dx dt$$

$$\epsilon_{\mathrm{CF}}\left(\Pi, \Phi, \Psi\right) = \int_{\mathcal{X}\times\{0,1\}} l_{\Pi,\Phi,\Psi}\left(x, I_t, t\right) p(x,1-t) dx dt$$

We now define the above factual loss for the treated and control groups separately below, and the same can be defined for the counterfactual loss as well.

$$\epsilon_{\mathrm{F}}^{t=1}\left(\Pi, \Phi, \Psi\right) = \int_{\mathcal{X}} l_{\Pi,\Phi,\Psi}\left(x, I_1, 1\right) p^{t=1}(x) dx$$

$$\epsilon_{\mathrm{F}}^{t=0}\left(\Pi, \Phi, \Psi\right) = \int_{\mathcal{X}} l_{\Pi,\Phi,\Psi}\left(x, I_0, 0\right) p^{t=0}(x) dx$$

Let $f : \mathcal{X} \times (\mathcal{I}; \{0,1\}) \to \mathcal{Y}$ be a hypothesis function. For instance, we can have $f\left(x, I_t, t\right) = \Pi\left(\Phi(x), \Psi(I_t), t\right)$.

Definition 5.4. For a given hypothesis function f, define expected PEHE as follows:

$$\epsilon_{\text{PEHE}}(f) = \int_{\mathcal{X}} (f(x, I_1, 1) - f(x, I_0, 0) - \mathbb{E}\left[Y_1 - Y_0 \mid x\right])^2 p(x) dx.$$

Note that the PEHE, as given in Equation (2) restricted to binary treatment case, serves as an unbiased estimator for the above in the finite sample case.

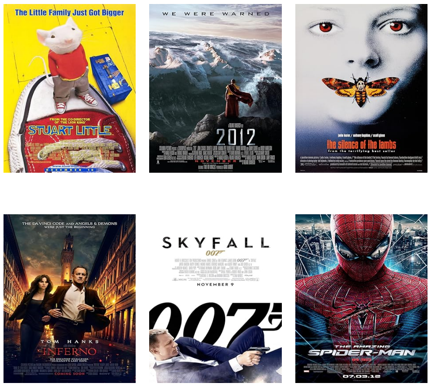

Figure 2: Few example posters considered as treatments

Definition 5.5. Denote $\sigma^2_{Y_t}(p(x,t))$ as the variance of Y_t w.r.to the distribution $p(x,t)$ given below:
$\sigma^2_{Y_t}(p(x,t)) = \int_{\mathcal{X}\times\mathcal{Y}} \left(Y_t - \mathbb{E}[Y_t \mid x]\right)^2 p\left(Y_t \mid x\right) p(x,t) dY_t dx$.
We define $\sigma^2_{Y_t} = \min\{\sigma^2_{Y_t}(p(x,t)), \sigma^2_{Y_t}(p(x,1-t))\}$ and $\sigma^2_Y = \min\{\sigma^2_{Y_0}, \sigma^2_{Y_1}\}$.

We now present our technical results, which provide an upper bound on the PEHE; proofs are provided in the Appendix.

Theorem 5.6. *Let $\Phi : \mathcal{X} \to \mathbb{R}_X$ and $\Psi : \mathcal{I} \to \mathcal{R}_I$ are twice differentible and invertible functions. Let Π be a hypothesis function. Let $\mathcal{G}$ denote a family of functions $g : \mathcal{R}_X \times (\mathcal{R}_I; \{0,1\})$. Assume the loss function L used to define $l_{\Pi,\Phi,\Psi}$ is the squared loss function. Further, assume that there exists a constant $D_{\Phi,\Psi} > 0$ s.t. the loss function $l(\cdot)$ satisfies the following: $\frac{1}{D_{\Phi,\Psi}} l_{\Pi,\Phi,\Psi}\left(\Phi^{-1}(r_x), \Psi^{-1}(r_{I_t}), t\right) \in \mathcal{G}$ for $t \in \{0,1\}$. Then, we have*

$$\epsilon_{\text{PEHE}}(\Pi, \Phi, \Psi) \le 2(\epsilon_F^{t=0}\left(\Pi, \Phi, \Psi\right) + \epsilon_F^{t=1}\left(\Pi, \Phi, \Psi\right) + D_{\Phi,\Psi}\mathtt{IPM}_{\mathcal{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right) - 2\sigma^2_Y), \quad (3)$$

where $\mathtt{IPM}_{\mathcal{G}}(p,q)$ be the IPM induced by $\mathcal{G}$ between probability distributions p and q, $\epsilon_F^{t=0}(\cdot)$ and $\epsilon_F^{t=1}(\cdot)$ are given in Definition 5.3 and σ^2_Y is given in Definition 5.5.

Corollary 5.7. *Consider the set up of k number of treatments, and under the conditions of Theorem 5.6, we have the following:*

$$\epsilon_{\text{PEHE}}(\Pi, \Phi, \Psi) \le \underbrace{\frac{2}{k}\sum_{a=1}^{k} \epsilon_F^{t=a}(\Pi, \Phi, \Psi) +}_{\text{MSE loss}}$$

$$\underbrace{\frac{4}{k(k-1)}\sum_{a=1}^{k}\sum_{b=1}^{a-1}(D_{\Phi,\Psi}\mathtt{IPM}_{\mathcal{G}}\left(p_\Phi^{t=a}, p_\Phi^{t=b}\right)}_{\text{Average IPM loss}} - 2\min\{\sigma^2_{Y_a}, \sigma^2_{Y_b}\}).$$

Remark 5.8. Note that the loss function used in NICE is inspired by the above result, with the aim of tightening the bound. This is achieved by minimizing a weighted sum of the MSE loss and the average IPM (MMD here) loss across all treatments pairs.

6 Data Simulation

Evaluating causal effects is challenging due to the lack of ground truth for counterfactuals in observational data. Prior work often

[1] Here, $(\mathcal{R}_I; \{0,1\})$ denotes the concatenation of the treatment index t to $\mathcal{R}_I$ along axis 0.

addresses this by creating semi-synthetic datasets, where covariates and treatments are real but POs are synthetically generated. However, to the best of our knowledge, there are no existing datasets available for our problem that involve images as treatments. Therefore, we focus on generating new semi-synthetic datasets to evaluate NICE algorithm against the baselines. Specifically, in our case, treatment images correspond to real world data, user covariates and POs are synthetically generated.

To simulate a realistic dataset under our setting, we design a setup inspired by personalization tasks in domains such as movie recommendations. Each user is represented as a 512-dimensional embedding, encapsulating the user's preference for specific attributes related to movies. The treatments in our setup are represented as images, specifically movie posters, which are also embedded into a 512-dimensional space to enable a direct mapping between the visual features of the treatments and the users' preferences. Treatment assignment is performed based on the alignment between user preferences and poster attributes. For example, a user with a strong preference for action and science fiction genres is more likely to be assigned a poster featuring futuristic visuals and intense action elements. The observed outcome corresponds to the user's engagement level, such as their likelihood of selecting or interacting with the poster, quantified as a continuous variable.

We generate our dataset using PosterLens [1] that contains posters of various movies and their respective ResNet [5] embeddings. We first randomly draw 20,000 posters' ResNet embeddings, of size 512, from the PosterLens dataset. These embeddings are used as a proxy for users' covariates [2]. For a given user with covariates, our goal is to estimate their opinion on the shown poster, a real valued scalar. Our hypothesis is that users like the posters similar to their preferences. Note that, here, treatments are the posters shown to users and their opinions are considered as a proxy for the POs (treatment effects). Few example posters considered as treatments are given in Figure 2. We generate multiple datasets based on the number of available treatments, that can be 4, 8 and 16.

To generate POs for the k number of treatments setup we first generate $(k+1)$ centroids as follows. Either randomly select $(k+1)$ ResNet embeddings from the 20,000 embeddings selected above, or train a KMeans clustering algorithm on the 20,000 embeddings with $(k+1)$ clusters as an input and take the resultant centroids. We use z_i to denote centroids. We use $Y_{i,t}$ to denote final POs for user-i and treatment-t. Our final POs are product of two quantities $\tilde{Y}_{i,t}$ and $d_{i,t}$ that are generated as follows.

- Generation of $\tilde{Y}_{i,t}$: For each treatment-t, generate μ_t and σ_t as follows: $\mu_t \sim \mathcal{N}(0.45, 0.15)$ and $\sigma_t \sim \mathcal{N}(0.1, 0.05)$[3]. Then, $\tilde{Y}_{i,t}$ is an i.i.d sample drawn from $\mathcal{N}(\mu_t, \sigma_t)$. Observe that the distribution of $\tilde{Y}_{i,t}$ is solely dependent on the treatment-t.
- Generation of $d_{i,t}$: It tries to measure the preference of user with covariates x_i to a treatment represented using its ResNet embedding which is defined as: $d_{i,t} = x_i^T z_t + x_i^T z_{k+1}\ \forall\, 1 \le i \le n \,\&\, 1 \le t \le k$.

Given the above, the final POs, denoted by $Y_{i,t}$ for any $1 \le i \le n$ and $1 \le t \le k$, are defined as

$$Y_{i,t} = c\tilde{Y}_{i,t} d_{i,t} = c\tilde{Y}_{i,t}\left[x_i^T z_t + x_i^T z_{k+1}\right], \tag{4}$$

where $c > 0$ is a fixed constant and we keep it as 5 in the experiments.

Table 1: Performance comparison of NICE vs baselines across various values of k. Here, $\kappa_a = 10$, $1 \le a \le k$ is chosen.

Method	$k = 4$	$k = 8$	$k = 16$
TarNet	128.2 ± 24.1	137.7 ± 26.7	152.8 ± 15.7
GraphITE	128.3 ± 24.5	135.3 ± 22.1	141.1 ± 12.8
SIN	127.7 ± 24.7	134.7 ± 22.1	139.4 ± 13.1
CaML	127.7 ± 24.7	138.0 ± 20.6	139.7 ± 13.0
NICE-ResNet	**91.9 ± 6.7**	**104.5 ± 12.7**	114.1 ± 6.2
NICE-VGG	97.3 ± 6.4	104.6 ± 12.8	**112.0 ± 5.6**

Table 2: Performance comparison of NICE in a zero-shot setting against baselines, to assess ITE estimation on unseen treatments.

Method	$k = 4$	$k = 8$	$k = 16$
TarNet	131.8 ± 29.2	150.7 ± 22.5	155.0 ± 23.7
GraphITE	128.3 ± 25.2	136.7 ± 19.3	145.1 ± 23.6
SIN	128.3 ± 26.5	133.2 ± 19.4	137.9 ± 13.9
CaML	120.7 ± 24.5	133.4 ± 17.8	135.9 ± 14.0
NICE-ResNet	**90.0 ± 6.5**	101.6 ± 13.7	110.4±8.9
NICE-VGG	95.3 ± 6.5	**101.1 ± 16.0**	**107.3 ± 10.1**

We briefly describe the observed treatment generation process, following the approaches in [8, 15, 16]. Let $p_{i,t}$ denote the probability of treatment-t assigned to user-i, defined as:

$$p_{i,t} = \mathtt{softmax}\left(\kappa_i Y_{i,t}\right), \tag{5}$$

where $\kappa = [\kappa_1, \kappa_2, \cdots, \kappa_k] > 0$, is a set of parameters that controls the treatment assignment bias. In other words, choosing $\kappa_a \gg \kappa_b$ for $b \neq a$ makes the treatment assignment distribution skewed toward treatment$-a$. For a given user-i, we randomly assign a treatment with the above probabilities, $p_{i,t}$, and call it as the observed treatment for that user.

7 Experiments

In this section, we provide details of the experiments conducted to evaluate the performance of NICE against various baselines. We begin by outlining the baselines considered in the experiments, followed by a comparison of NICE performance with these baselines across different scenarios, including zero-shot tasks. In all our experiments, we use the datasets as described in the Data Simulation section and all models are evaluated using the square root of PEHE metric defined in Equation (2). All model parameters are provided in the Appendix due to space limitations.

7.1 Baseline Methods

We compare our NICE with adaptations of existing methods to our problem, that leverage treatment attributes for estimating ITEs. To that end, we include modified versions of GraphITE [4], CaML [11], and SIN [10] in the baselines, as these algorithms incorporate treatments' auxiliary information to enhance ITE estimation. In particular, since GraphITE, SIN, and CaML use GCN for representing

[2] Due to the lack of datasets for image treatments, poster embeddings are considered (as a proxy) for users' covariates.

[3] We ensure that $\sigma_t > 0$ by regenerating samples whenever a non-positive sample is produced.

graph treatments, we adapted these methods by replacing graph representations with image representations, specifically using ResNet embeddings of the image treatments in our case, while maintaining the rest of the architecture. We also include an additional baseline that does not use treatment information, to effectively demonstrate the benefits of using treatment's auxiliary information in the ITE estimation. As NICE primarily operates on treatment head networks, we consider TARNet [17] as another baseline that does not use treatment information in the estimation.

GraphITE utilizes graphs as treatments to improve causal effect estimation with the HSIC criterion, reducing bias introduced by the treatment representation space. In the performance comparison Table 1, we include results of our experiments for GraphITE with the HSIC criterion. Similarly, we compare the performance of our algorithm with the SIN algorithm, which primarily relies on Robinson decomposition to include a quasi-convergence guarantees for estimators. CaML uses a meta-learning approach to estimate pseudo outcomes of estimators. One common modification required for our experiment was to replace the graph representation network in these algorithms, as they primarily utilize graphs as treatments, with an image representation network to evaluate NICE with these methods.

7.2 NICE Performance Assessment

We conducted a comprehensive evaluation of NICE across various experimental settings to assess its ITE estimation capabilities. The experimental evaluation focused on testing the hypothesis that integrating treatments' auxiliary information, particularly images, enhances the accuracy of ITE estimates. To validate this hypothesis, we employed semi-synthetic datasets as described in the Section 6. Our results demonstrate that NICE consistently outperforms the baselines in ITE estimation, particularly when dealing with 4, 8, and 16 treatment groups. Results shown in all tables are means and standard deviations of $\sqrt{\epsilon_{\text{PEHE}}}$ values computed across 10 iterations. We use bold face to indicate the best results in the tables. As the number of treatments, k, increases, the complexity of the problem escalates. Notably, the performance gap between NICE and the baseline methods widens as k increases, as illustrated in Table 1.

Table 3: Performance comparison of NICE vs baselines under varying assignment bias $\kappa_a = 10,\ 1 \leq a < k$ and $\kappa_k = 50$.

Method	$k = 4$	$k = 8$	$k = 16$
TarNet	136.2 ± 31.8	150.9 ± 27.5	152.9 ± 18.9
GraphITE	129.6 ± 24.6	141.3 ± 22.8	145.8 ± 13.7
SIN	127.7 ± 24.7	134.7 ± 22.1	139.7 ± 12.8
CaML	129.1 ± 24.0	137.3 ± 21.4	139.3 ± ± 13.9
NICE-ResNet	**91.8 ± 6.6**	**105.1 ± 13.2**	129.2 ± 22.2
NICE-VGG	100.8 ± 12.8	112.2 ± 25.0	**127.3 ± 22.7**

Table 4: Performance comparison of NICE vs baselines under varying assignment bias $\kappa_a = 10,\ 1 \leq a < k\ \&\ \kappa_k = 100$.

Method	$k = 4$	$k = 8$	$k = 16$
TarNet	131.9 ± 25.8	139.6 ± 22.7	149.0 ± 18.6
GraphITE	129.3 ± 24.3	143.0 ± 21.0	148.8 ± 13.1
SIN	127.7 ± 24.7	134.7 ± 22.1	139.5 ± 12.7
CaML	127.7 ± 24.7	134.2 ± 21.8	**135.4 ± 8.4**
NICE-ResNet	97.6 ± 14.3	**105.4 ± 13.7**	137.1 ± 21.7
NICE-VGG	**95.3 ± 6.4**	113.3 ± 16.5	135.4 ± 21.8

7.2.1 *Treatment-embedding agnostic setting.*

We now evaluate the performance of NICE in a treatment-embedding agnostic setup using a semi-synthetic dataset generated with ResNet embeddings. To assess the robustness of our approach, we also test NICE using VGG-based image embeddings [19] to account for potential dataset biases that might arise from our data generation process. In experiments involving variations in the number of treatments, we observe that NICE, when utilizing VGG embeddings, often surpasses the baseline methods and the ResNet-based NICE implementation, as shown in experiment evaluation Tables 1-4. These results underscore the model's effectiveness in causal estimation across different treatment representation embeddings.

7.2.2 *Zero-shot setting.*

Baselines that we compare NICE with also claim to have zero-shot capabilities for ITE estimation on unseen treatments during training of the models. We evaluate both NICE and these baselines for their zero-shot abilities in the context of images as treatment setups. For this evaluation, we use a modified version of the PEHE metric, referred to as the rooted Zero-Shot PEHE metric ($\epsilon^{\text{ZS}}_{\text{PEHE}}$), which is defined as follows: $\epsilon^{\text{ZS}}_{\text{PEHE}} = \frac{1}{k-1} \sum_{\substack{a=1 \\ a \neq z}}^{k} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\hat{\tau}^{a,z}(x_i) - \tau^{a,z}(x_i) \right)^2 \right]$, where z is the zero-shot treatment whose samples are not seen by the model during training. Observe that in the above equation, PEHE is computed using only treatment pairs that include the zero-shot treatment, as our goal is to evaluate the ITE estimation capabilities of the model in zero-shot scenarios. As illustrated in Table 2, NICE consistently outperforms baselines in zero-shot ITE estimation for the image treatments.

7.2.3 *High treatment assignment bias setting.*

In real-world scenarios, treatment assignments can be highly skewed based on user covariates, which significantly amplifies the assignment bias. To assess the performance of NICE under such conditions, we simulate scenarios by increasing the treatment assignment bias κ_a for a specific treatment, inducing a skewed assignment toward treatment$-a$. Specifically, we consider two scenarios with $\kappa_a = 10$ for $1 \leq a < k$ and $\kappa_k = 50$ and 100. As shown in Tables 3 and 4, NICE consistently outperforms baselines, demonstrating robustness under highly skewed treatment assignment scenarios.

8 Conclusion

In this study, we propose NICE, a novel framework designed to estimate ITEs when images are served as treatments. To validate the efficacy of NICE, we propose a unique semi-synthetic data simulation technique that generates POs for image treatments. NICE leverages image treatments' auxiliary information to estimate POs in scenarios involving multiple treatments. Notably, NICE demonstrates zero-shot causal effect estimation capabilities, enabling it to infer causal outcomes for novel treatments. Experimental results show that NICE consistently outperforms various baselines across different setups, achieving the best performance on the PEHE. For future work, we plan to explore the scalability of NICE to more complex datasets, as well as its applicability to real-world scenarios beyond semi-synthetic simulations. Additionally, we aim to enhance the framework's interpretability and extend its capabilities to handle more diverse and complex treatment types, such as video or multimodal data.

References

[1] Sasha Aptlin. 2021. PosterLens 25M dataset. *Kaggle* doi: 10.34740/KAGGLE/DS/1321802 (2021).

[2] David Chan, Rong Ge, Ori Gershony, Tim Hesterberg, and Diane Lambert. 2010. Evaluating online ad campaigns in a pipeline: causal models at scale. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 7–16.

[3] Ruocheng Guo, Jundong Li, and Huan Liu. 2020. Learning individual causal effects from networked observational data. In *Proceedings of the 13th international conference on web search and data mining*. 232–240.

[4] Shonosuke Harada and Hisashi Kashima. 2021. Graphite: Estimating individual effects of graph-structured treatments. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 659–668.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[6] Guanglei Hong and Stephen W Raudenbush. 2005. Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational evaluation and policy analysis* 27, 3 (2005), 205–224.

[7] Guido W Imbens and Donald B Rubin. 2015. *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

[8] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*. PMLR, 3020–3029.

[9] Kareem L Jordan. 2012. Juvenile transfer and recidivism: A propensity score matching approach. *Journal of Crime and Justice* 35, 1 (2012), 53–67.

[10] Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. 2021. Causal effect inference for structured treatments. *Advances in Neural Information Processing Systems* 34 (2021), 24841–24854.

[11] Hamed Nilforoshan, Michael Moor, Yusuf Roohani, Yining Chen, Anja Šurina, Michihiro Yasunaga, Sara Oblak, and Jure Leskovec. 2023. Zero-shot causal learning. *Advances in Neural Information Processing Systems* 36 (2023), 6862–6901.

[12] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press, Cambridge, UK.

[13] Judea Pearl. 2017. Detecting Latent Heterogeneity. *Sociological Methods & Research* 46, 3 (2017), 370–389.

[14] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

[15] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. 2020. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5612–5619.

[16] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. 2018. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656* (2018).

[17] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*. PMLR, 3076–3085.

[18] Ilya Shpitser and Judea Pearl. 2012. Identification of conditional interventional distributions. *arXiv preprint arXiv:1206.6876* (2012).

[19] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[20] Jeffrey A Smith and Petra E Todd. 2001. Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review* 91, 2 (2001), 112–118.

[21] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. 2012. On the empirical estimation of integral probability metrics. (2012).

[22] Thérèse A Stukel, Elliott S Fisher, David E Wennberg, David A Alter, Daniel J Gottlieb, and Marian J Vermeulen. 2007. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *Jama* 297, 3 (2007), 278–285.

[23] Abhinav Thorat, Ravi Kolla, Niranjan Pedanekar, and Naoyuki Onoe. 2023. Estimation of individual causal effects in network setup for multiple treatments. *arXiv preprint arXiv:2312.11573* (2023).

[24] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International conference on learning representations*.

[25] Yinan Yu, Hailiang Chen, Chih-Hung Peng, and Patrick YK Chau. 2022. The causal effect of subscription video streaming on DVD sales: Evidence from a natural experiment. *Decision Support Systems* 157 (2022), 113767.

Appendix

We begin by proving the following lemma that is required to establish Theorem 5.6. Note that, the proof of Theorem 5.6 follows along the lines of proof of Theorem 1 in [17].

Lemma 8.1. *Let $\Phi : \mathcal{X} \to \mathbb{R}_X$ and $\Psi : \mathcal{I} \to \mathcal{R}_I$ be twice differentiable and invertible functions. Let Π be an hypothesis function. Let $\mathcal{G}$ denote a family of functions $g : \mathcal{R}_X \times (\mathcal{R}_I; \{0,1\})$. Let $D_{\Phi,\Psi} > 0$ be a constant such that the loss function $l(\cdot)$ satisfies the following: $\frac{1}{D_{\Phi,\Psi}} l_{\Pi,\Phi,\Psi}\left(\Phi^{-1}(r_x), \Psi^{-1}(r_{I_t}), t\right) \in \mathcal{G}$ for $t \in \{0,1\}$. Then, we have*

$$\epsilon_{CF}\left(\Pi, \Phi, \Psi\right) \leq (1-u)\epsilon_F^{t=1}(\Pi, \Phi, \Psi) + u\epsilon_F^{t=0}(\Pi, \Phi, \Psi) + D_{\Phi,\Psi}\mathtt{IPM}_{\mathcal{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right),$$

where $\mathrm{IPM}_G(p,q)$ be the Integral Probability Metric induced by $\mathcal{G}$ between probability distributions p and q, $u := p(t=1)$, $\epsilon_{CF}()$, $\epsilon_F^{t=1}()$ and $\epsilon_F^{t=0}()$ are defined in Definition 5.3.

Proof.

$$\begin{aligned}
&\epsilon_{\mathrm{CF}}\left(\Pi, \Psi, \Phi\right) - (1-u)\epsilon_{\mathrm{F}}^{t=1}\left(\Pi, \Psi, \Phi\right) + u\epsilon_{\mathrm{F}}^{t=0}\left(\Pi, \Psi, \Phi\right) \\
&= (1-u)\left[\epsilon_{\mathrm{CF}}^{t=1}\left(\Pi, \Psi, \Phi\right) - \epsilon_{\mathrm{F}}^{t=1}\left(\Pi, \Psi, \Phi\right)\right] + u\left[\epsilon_{\mathrm{CF}}^{t=0}\left(\Pi, \Psi, \Phi\right) - \epsilon_{\mathrm{F}}^{t=0}\left(\Pi, \Psi, \Phi\right)\right] \qquad (6) \\
&= (1-u)\left[\int_{\mathcal{X}} l_{\Pi,\Psi,\Phi}\left(x, I_1, 1\right)\left(p^{t=0}(x) - p^{t=1}(x)\right)dx\right] + u\left[\int_{\mathcal{X}} l_{\Pi,\Psi,\Phi}\left(x, I_0, 0\right)\left(p^{t=1}(x) - p^{t=0}(x)\right)dx\right] \qquad (7) \\
&= (1-u)\left[\int_{\mathcal{R}_X} l_{\Pi,\Psi,\Phi}\left(\Phi^{-1}(r_x), \Psi^{-1}(r_{I_1}), 1\right)\left(p_\Phi^{t=0}(r_x) - p_\Phi^{t=1}(r_x)\right)dr_x\right] \\
&\qquad + u\left[\int_{\mathcal{R}_X} l_{\Pi,\Psi,\Phi}\left(\Phi^{-1}(r_x), \Psi^{-1}(r_{I_0}), 0\right)\left(p_\Phi^{t=1}(r_x) - p_\Phi^{t=0}(r_x)\right)dr_x\right] \qquad (8) \\
&= D_{\Phi,\Psi}(1-u)\left[\int_{\mathcal{R}_X} \frac{1}{D_{\Phi,\Psi}} l_{\Pi,\Psi,\Phi}\left(\Phi^{-1}(r_x), \Psi^{-1}(r_{I_1}), 1\right)\left(p_\Phi^{t=0}(r_x) - p_\Phi^{t=1}(r_x)\right)dr_x\right] \\
&\qquad + D_{\Phi,\Psi}u\left[\int_{\mathcal{R}_X} \frac{1}{D_{\Phi,\Psi}} l_{\Pi,\Psi,\Phi}\left(\Phi^{-1}(r_x), \Psi^{-1}(r_{I_0}), 0\right)\left(p_\Phi^{t=1}(r_x) - p_\Phi^{t=0}(r_x)\right)dr_x\right] \\
&\leq (1-u)D_{\Phi,\Psi}\sup_{g\in\mathcal{G}}\left|\int_{\mathcal{R}_X} g(r_x)\left(p_\Phi^{t=0}(r_x) - p_\Phi^{t=1}(r_x)\right)dr_x\right| + uD_{\Phi,\Psi}\sup_{g\in\mathcal{G}}\left|\int_{\mathcal{R}_X} g(r_x)\left(p_\Phi^{t=1}(r_x) - p_\Phi^{t=0}(r_x)\right)dr_x\right| \qquad (9) \\
&\leq D_{\Phi,\Psi}\mathtt{IPM}_{\mathcal{G}}\left(p_\Phi^{t=0}, p_\Phi^{t=1}\right) \qquad (10)
\end{aligned}$$

where (6) due to Lemma A3 in [17], (7) due to Definition 5.3, (8) due to Φ and Ψ are invertible functions, (9) due to a condition in the Lemma and (10) due to the definition of IPM. □

We now turn our attention to prove Theorem 5.6.

Proof of Theorem 5.6. Note that we can write $\epsilon_{\text{PEHE}}(f) = \epsilon_{\text{PEHE}}(\Pi, \Phi, \Psi)$ for some $f(x, I_t, t) = \Pi(\Phi(x), \Psi(I_t), t)$.

$$\begin{aligned}
\epsilon_{PEHE}(f) &= \int_{\mathcal{X}} ((f(x, I_1, 1) - f(x, I_0, 0)) - (m_1(x) - m_0(x)))^2 p(x) dx \\
&= \int_{\mathcal{X}} ((f(x, I_1, 1) - m_1(x)) + (m_0(x) - f(x, I_0, 0)))^2 p(x) dx && (\because \text{Rearranging the terms}) \\
&\leq 2 \int_{\mathcal{X}} ((f(x, I_1, 1) - m_1(x))^2 + (m_0(x) - f(x, I_0, 0))^2) p(x) dx && (\because (x+y)^2 \leq 2(x^2 + y^2)) \\
&= 2 \int_{\mathcal{X}} ((f(x, I_1, 1) - m_1(x))^2 p(x, t=1) dx + 2 \int_{\mathcal{X}} (m_0(x) - f(x, I_0, 0))^2) p(x, t=0) dx \\
&\qquad + 2 \int_{\mathcal{X}} ((f(x, I_1, 1) - m_1(x))^2 p(x, t=0) dx + 2 \int_{\mathcal{X}} (m_0(x) - f(x, I_0, 0))^2) p(x, t=1) dx \\
& && (\because p(x) = p(x, t=0) + p(x, t=1)) \\
&= 2 \int_{\mathcal{X}} (f(x, I_t, t) - m_t(x))^2 p(x, t) dx + 2 \int_{\mathcal{X}} (f(x, I_t, t) - m_t(x))^2 p(x, 1-t) dx \\
&\leq 2 \left(\epsilon_{\text{F}}(f) - \sigma_Y^2\right) + 2 \left(\epsilon_{\text{CF}}(f) - \sigma_Y^2\right) && (\because \text{Due to Lemma A5 in [17]}) \\
&\leq 2 \left(\epsilon_{\text{F}}(f) - \sigma_Y^2\right) + 2 \left((1-u)\epsilon_F^{t=1}(f) + u\epsilon_F^{t=0}(f) + D_{\Phi,\Psi} \text{IPM}_{\mathcal{G}} \left(p_\Phi^{t=1}, p_\Phi^{t=0}\right) - \sigma_Y^2\right) \\
& && (\because \text{Due to Lemma 8.1 }) \\
&= 2 \left(u\epsilon_{\text{F}}^{t=1}(f) + (1-u)\epsilon_{\text{F}}^{t=0}(f) - \sigma_Y^2\right) + 2 \left((1-u)\epsilon_F^{t=1}(f) + u\epsilon_F^{t=0}(f) + D_{\Phi,\Psi} \text{IPM}_{\mathcal{G}} \left(p_\Phi^{t=1}, p_\Phi^{t=0}\right) - \sigma_Y^2\right) \\
& && (\because \text{Due to Lemma A3 in [17]}) \\
&= 2(\epsilon_{\text{F}}^{t=0}(f) + \epsilon_{\text{F}}^{t=1}(f) + D_{\Phi,\Psi} \mathtt{IPM} \left(p_\Phi^{t=1}, p_\Phi^{t=0}\right) - 2\sigma_Y^2).
\end{aligned}$$

□

Proof of Corollary 5.7. It directly follows from the Theorem 5.6. It can be achieved by constructing sub-experiments with all possible combinations of pair of treatments and considering one of them as treated and the other as control. Then, by invoking the Theorem 5.6 for each sub-experiment establishes the desired result. □

Model Parameters

Hyperparameter	**Search Range**
Number of layers (Φ)	{4, 6, 8}
Number of nodes (Φ)	{200, 400, 600}
Number of layers (Ψ)	{4, 6, 8}
Number of nodes (Ψ)	{200, 400, 600}
Number of layers (Π_t)	{4, 6, 8}
Number of nodes (Π_t)	{200, 400, 600}
Alpha α	{0.5, 1.0}
Beta β	{0.5}
Batch Size	{256, 512}
Learning Rate	{0.1, 0.01}
Learning rate Decay	$\{1\mathrm{e}^{-1}\}$
Learning Scheduler Step	{10, 15}
Weight Decay	$\{1\mathrm{e}^{-4}\}$
Dropout	{0.1}
Activation	{Tanh, ELU}

Table 5: Hyperparameter search range for NICE (our proposed method) and baselines on semi-synthetic datasets.

We briefly outline the experimental setup for optimizing NICE and baseline algorithms. For covariate representation, we use a fully connected (FC) network with Tanh and ELU activation functions. Similarly, for treatment representation and causal estimator head networks, we employ FC networks with variations in the number of nodes and layers. In dataset simulation, we generate $20,000$ instances in each experiment. The data is split into training, validation, and test sets, and performance is assessed using the PEHE metric by comparing predicted POs against the ground truth for all instances. To achieve optimal performance for NICE-ResNet, NICE-VGG, and baseline algorithms, we implement techniques such as early stopping, learning rate scheduling, weight decay, and dropout.