# Multiple Treatments Causal Effects Estimation with Task Embeddings and Balanced Representation Learning

**Anonymous authors**
**Paper under double-blind review**

## Abstract

The simultaneous application of multiple treatments is increasingly common in many fields, such as healthcare and marketing. In such scenarios, it is important to estimate the single treatment effects and the interaction treatment effects that arise from treatment combinations. Previous studies have proposed using independent outcome networks with subnetworks for interactions, or combining task embedding networks that capture treatment similarity with variational autoencoders. However, these methods suffer from the lack of parameter sharing among related treatments, or the estimation of unnecessary latent variables reduces the accuracy of causal effect estimation. To address these issues, we propose a novel deep learning framework that incorporates a task embedding network and a representation learning network with the balancing penalty. The task embedding network enables parameter sharing across related treatment patterns because it encodes elements common to single effects and contributions specific to interaction effects. The representation learning network with the balancing penalty learns representations nonparametrically from observed covariates while reducing discrepancies in representation distributions across different treatment patterns. This process reduces the estimation variance caused by selection bias and avoids model misspecification. Simulation studies demonstrate that the proposed method outperforms existing baselines, and application to real-world marketing datasets confirms the practical implications and utility of our framework.

## 1 Introduction

Estimating the single and interaction effects of multiple, simultaneously applied treatments is a critical challenge in many fields, such as healthcare and marketing. For example, combination drug therapies can lead to complex interactions and unforeseen side effects that are not apparent from studying each drug in isolation (Gradman et al., 2010; Webster, 2016; Mokhtari et al., 2017). Similarly, since the total impact of concurrent marketing promotions often exceeds the simple sum of their single effects, accounting for these interactions is crucial for optimizing marketing strategy (Danaher et al., 2008; Lemon & Nowlis, 2002; Lesscher et al., 2021; Naik & Raman, 2003). These situations demonstrate a critical need for robust analytical frameworks that can precisely estimate both single and interaction treatment effects to guide effective decision-making.

Whereas existing methods have been proposed to estimate causal effects under multiple treatments, they suffer from critical structural limitations that degrade their performance. For example, Neural Counterfactual Relation Estimation (NCoRE) (Parbhoo et al., 2021) addresses treatment interaction by using separate outcome prediction networks for each treatment and additional interaction subnetworks that are activated only when multiple treatments are applied simultaneously. However, this reliance on separate networks prevents parameter sharing across similar treatments, which leads to unstable estimates, especially when samples for specific treatment patterns are limited. Similarly, Task Embedding–based Causal Effect Variational Autoencoder (TECE-VAE) (Saini et al., 2019), which combines a treatment similarity–aware task embedding network with a variational autoencoder, assumes the existence of latent covariates and treats observed ones as proxies. This approach, in turn, introduces a susceptibility to model misspecification because its forced estimation of latent variables can be detrimental when observed covariates are sufficient. Therefore, the

structural constraints of existing methods limit their ability to provide accurate and robust estimates of both single and interaction treatment effects.

In this study, we propose a novel deep learning framework called Causal Inference for Single and Interaction treatment effects Network (CISI-Net). CISI-Net integrates two core components: a task embedding network designed to capture treatment similarity and a representation learning network for reducing estimation variance.. First, the task embedding network learns to assign similar embedding vectors to similar treatments, and it places treatment patterns with similar causal effects closer in the embedding space. This structure enables the embedding vectors to encode components common to single effects and contributions specific to interaction effects. This alignment allows parameter sharing across related treatment patterns, which reduce the need for independent networks for each treatment pattern. Moreover, it also improves the accuracy of causal effect estimation by distinguishing components common to single effects from those specific to interaction effects. Second, the representation learning network with the balancing penalty learns representations nonparametrically from the observed covariates while adjusting the representation distributions to be aligned across different treatment patterns, which mitigates the estimation variance caused by selection bias. Unlike TECE-VAE, this data-driven flexibility removes the coercion to infer latent covariates, and thus reduces the risk of accuracy degradation caused by model misspecification.

To assess the effectiveness of the proposed framework, we conducted extensive evaluations using both three simulations and two real-world datasets. In three simulation datasets, our method consistently outperformed all baseline approaches in estimating both single and interaction treatment effects. Our real-world case study based on multiple marketing promotion datasets further demonstrates the practical implications and utility of our framework.

## 2 Related Work

This section reviews the main contributions and limitations of the deep learning–based methods for causal effect estimation. These methods can broadly be categorized into two groups. The first group consists of methods developed for single-treatment settings, which primarily aim to estimate the causal effect of one treatment applied in isolation. These methods can be adapted to multiple-treatment settings by treating each treatment pattern as a distinct treatment and are commonly used as baselines in multiple-treatment research (Parbhoo et al., 2021; Saini et al., 2019). The second group consists of methods developed for multiple-treatment settings. These methods include approaches that extend single-treatment methods for multiple treatments or are specifically designed to handle multiple treatments.

Existing methods in the first group, designed for single-treatment settings (Johansson et al., 2016; Künzel et al., 2019; Shi et al., 2019; Yang et al., 2024; Zhu et al., 2023), are by their very nature unable to model the interaction effects that arise from the joint application of multiple treatments. For example, Treatment-Agnostic Representation Network (TARNet) (Shalit et al., 2017) learns a shared representation that is independent of treatment and uses it to predict counterfactual outcomes. Counterfactual Regression (CFR) (Shalit et al., 2017) extends TARNet by introducing a balancing penalty based on Integral Probability Metrics (IPM) (Sriperumbudur et al., 2010), which aligns the representation distributions of the treated and control groups to bound the expected estimation error. Although these approaches are effective for estimating single treatment effects, they are fundamentally limited by their core assumption of treatment isolation, which makes them incapable of modeling interactions.

In the second group, existing methods for multiple treatments can be categorized according to their methodological paradigms. These methods can be divided into data augmentation–based approaches (Qian et al., 2021) and model architecture–driven approaches (Chauhan et al., 2025; Parbhoo et al., 2021; Saini et al., 2019; Tsuboi et al., 2024; Zou et al., 2020). This distinction is important because data augmentation–based approaches rely on additional identification assumptions, such as sequential ignorability, which differ from the assumptions underlying model architecture-driven methods. Even among model architecture–driven approaches, existing methods differ in their target settings, including the types of treatments, the number of outcomes, and the modeling assumptions. Table 1 summarizes these differences and clarifies the position of CISI-Net relative to other methods. Although dosage combination network (DC-Net) (Schweisthal et al., 2023) is architecturally similar to CISI-Net, it is specifically designed for continuous treatments and, there-

Table 1: Comparison of model architecture-driven approaches under multiple treatments.

| Method | Treatment type | Outcome type | Parameter sharing mechanism | Latent covariate assumption |
|---|---|---|---|---|
| NCoRE | Binary | Single | None | ✗ |
| TECE-VAE | Binary | Single | Task embedding + VAE | ✓ |
| H-Learner | Binary/Continuous | Multiple | Hypernetwork-based weight generation | ✗ |
| DC-Net | Continuous | Single | Dosage-combination network | ✗ |
| CISI-Net | Binary | Single | Task embedding + shared outcome network | ✗ |

fore cannot be directly applied to the binary treatment setting considered here. Unlike H-Learner (Chauhan et al., 2025), which is designed to model treatment-outcome tasks across composite treatments and multiple outcomes, CISI-Net targets different estimands, which are causal effects under a single-outcome setting. Accordingly, in what follows, we focus on previous studies that are most closely aligned with our setting of binary multiple-treatment effect estimation for a single outcome.

Within the class of model architecture-driven methods for single-outcome causal effect estimation, existing methods (Parbhoo et al., 2021; Saini et al., 2019; Tsuboi et al., 2024; Zou et al., 2020) face two major limitations. The first limitation is a structural inability to share parameters across related treatments, which leads to unstable estimation under data sparsity caused by the limited number of samples for specific treatment patterns. The extensions of single-treatment methods can model interaction effects by assigning independent outcome prediction networks to each treatment pattern (Parbhoo et al., 2021; Saini et al., 2019). NCoRE introduces interaction subnetworks that are activated only when multiple treatments are applied simultaneously, which enables explicit modeling of interaction treatment effects. These methods depend on having sufficient samples for each treatment pattern because each network is updated only with samples corresponding to its treatment pattern. Since users who receive multiple treatments are typically rare (Chu et al., 2022), networks corresponding to infrequent treatment patterns remain poorly trained. In addition, because outcome prediction networks and interaction subnetworks are constructed independently, these methods cannot use similarities among treatments, which further increases instability in estimation.

The second limitation is a susceptibility to model misspecification. Although TECE-VAE addresses the first limitation by combining a treatment similarity-aware task embedding network with a VAE, it relies on a VAE that assumes the presence of latent covariates. This method assumes the presence of latent covariates, which are a subset or the entirety of the true covariates that remain unobserved by the analyst, and treats observed covariates as proxies for these latent covariates. When the assumed latent-variable generative model does not align with the true data-generating process, the inferred latent variables may fail to represent covariate structure, which leads to model misspecification. A similar risk of model misspecification arises in other latent-variable–based approaches for multiple treatments (Ma et al., 2021), such as Variational Sample Reweighting (Zou et al., 2020) and Decofounder (Wang & Blei, 2021). Such assumptions about latent variables can reduce robustness and limit the practical applicability of these methods in real-world settings.

Our CISI-Net addresses these limitations through two core components. First, the task embedding network captures treatment similarities by mapping related treatment patterns to proximal points in an embedding space. This network offers two key advantages over existing methods. Because it provides a mechanism to disentangle single effects from interaction effects, it enables the explicit modeling of interaction treatment effects that single-treatment methods cannot capture. Furthermore, by bringing similar treatments closer together, it facilitates parameter sharing across different treatment patterns, which improves estimation stability and removes the need for separate networks for each combination. Second, the representation learning network with the balancing penalty directly learns balanced representations from observed covariates to control the estimation variance induced by distributional imbalances. Crucially, because this data-driven approach does not rely on the rigid latent-variables assumptions like TECE-VAE, this design minimizes the risk of accuracy degradation from model misspecification.

## 3 Causal Inference under Multiple Treatments

We formulate the causal inference problem under multiple treatments within the potential outcomes framework (Rubin, 2005). The goal is to estimate the single and interaction treatment effects of multiple binary

treatments on a continuous outcome using observed covariates. We consider $N$ independent units, indexed by $i = 1, \ldots, N$. For each unit $i$, we observe a covariate vector $\boldsymbol{x}_i \in \mathbb{R}^d$ drawn from the covariate space $\boldsymbol{X}$. There are $K$ binary treatments. For each unit $i$, let the random variable $\boldsymbol{T}_i$ denote the treatment assignment. Its realization, $\boldsymbol{t}_i \in \{0,1\}^K$, is the vector representing the specific treatment the unit actually received. Under any possible treatment vector $\boldsymbol{t} \in \{0,1\}^K$, the potential outcome for unit $i$ is a scalar value $Y_i(\boldsymbol{t}) \in \mathbb{R}$, which represents the outcome that would have been observed had unit $i$ received treatment $\boldsymbol{t}$. In practice, we only observe one of these potential outcomes for each unit. The observed outcome is denoted by $Y_i$, and the specific realized outcome is denoted by $y_i$. Therefore, for each unit $i$, the observed data consists of the triplet $(\boldsymbol{x}_i, \boldsymbol{t}_i, y_i)$.

To identify causal effects of interest from the observed data, we adopt the following three assumptions commonly used in observational studies (Imbens & Rubin, 2015).

**Assumption 1. (Stable Unit Treatment Value Assumption)** *(1) no interference, meaning that the outcome of one unit is unaffected by the treatment assignments of other units; and (2) consistency of treatment, meaning that the potential outcomes correspond to well-defined and unique treatments (i.e., $Y_i = Y_i(\boldsymbol{T}_i)$).*

**Assumption 2. (Ignorability.)** *For any treatment pattern, the potential outcome is independent of the assigned treatment $T$ given the observed covariates $X$. Formally, for all $\boldsymbol{t}$,*

$$Y(\boldsymbol{t}) \perp \boldsymbol{T} \mid \boldsymbol{X}$$

**Assumption 3. (Overlap.)** *Every unit has a non-zero probability of receiving any treatment pattern given its observed covariates. Formally, for all $\boldsymbol{t}$ and $\boldsymbol{x}$,*

$$0 < P(\boldsymbol{T} = \boldsymbol{t} \mid \boldsymbol{X} = \boldsymbol{x}) < 1$$

Our estimands of interest, the average single effect (ASE) and the average interaction effect (AIE), are defined using the conditional expected potential outcome $\mu(\boldsymbol{x}, \boldsymbol{t})$ given by

$$\mu(\boldsymbol{x}, \boldsymbol{t}) \coloneqq \mathbb{E}[Y(\boldsymbol{t}) \mid \boldsymbol{X} = \boldsymbol{x}]. \tag{1}$$

Under Assumptions 1-3, this quantity is identified from observed data. This key result is summarized as follows.

**Proposition 1.** *Under Assumptions 1-3, the conditional average potential outcome $\mu(\boldsymbol{x}, \boldsymbol{t})$ is identified and is equal to the conditional expectation of the observed outcome as follows:*

$$\mu(\boldsymbol{x}, \boldsymbol{t}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}]. \tag{2}$$

We now formally define our estimands of interest. First, the ASE for treatment $k$ is the average effect of applying only treatment $k$ compared to no treatment. Let $\boldsymbol{t}_{+k}$ be the one-hot vector for treatment $k$. The ASE is defined as

$$\tau_{\text{ASE}}(k) \coloneqq \mathbb{E}_{\boldsymbol{X}}\left[\mu(\boldsymbol{x}, \boldsymbol{t}_{+k}) - \mu(\boldsymbol{x}, \boldsymbol{0})\right]. \tag{3}$$

The ASE is mathematically equivalent to the average combination effect (Egami & Imai, 2019) under the special case where the treatment vector is one-hot, meaning that only a single treatment is applied. The ASE captures the causal effect of applying a single treatment.

Second, the AIE (Egami & Imai, 2019) quantifies the interaction effect for a combination of treatments $S$ (where $S \in \{S' \subseteq \{1, \ldots, K\} \mid |S'| \geq 2\}$) as follows:

$$\tau_{\text{AIE}}(S) \coloneqq \mathbb{E}_{\boldsymbol{X}}\left[\sum_{Q \subseteq S} (-1)^{|S|-|Q|}\, \mu\left(\boldsymbol{x}, \boldsymbol{t}_{(+Q)}\right)\right], \tag{4}$$

where $t_{(+Q)}$ is the treatment vector that sets the components indexed by $Q$ to one and all remaining components to zero. This definition corresponds to the special case where the baseline treatment vector is

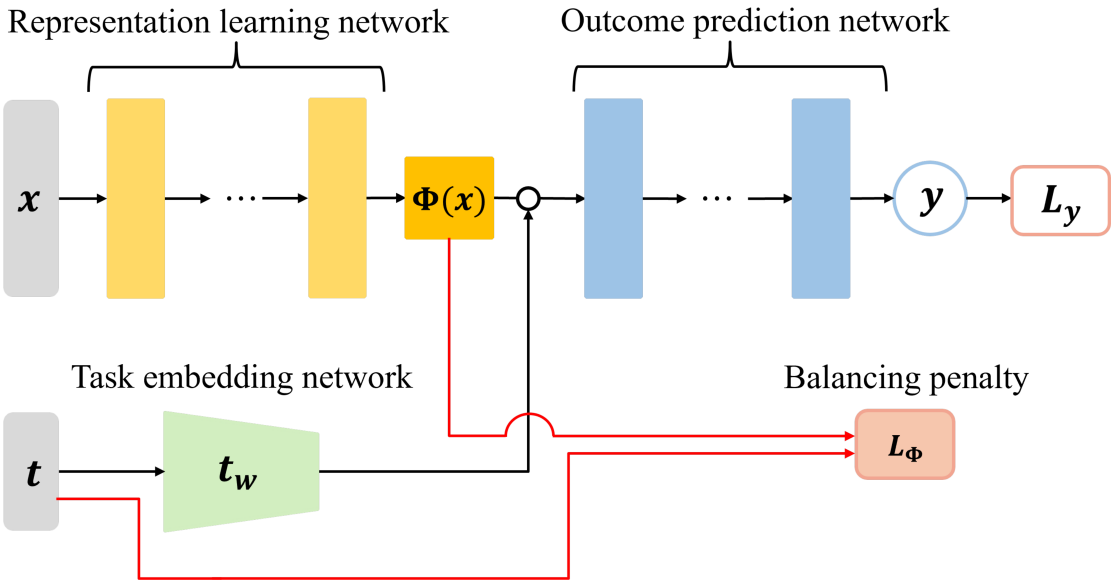



Figure 1: The architecture of CISI-Net consists of three components: the representation learning network (yellow), the task embedding network (green), and the outcome prediction network (blue). The latent representation $\Phi(\boldsymbol{x})$ is concatenated with the task embedding vector $t_w(\boldsymbol{t})$ to predict the outcome $y$. The model is trained with two loss terms: the prediction loss $L_y$ and the balancing penalty $L_\Phi$ (red).

the all-zero condition and each treatment element is binary, as discussed in prior work (Egami & Imai, 2019). The AIE defines interaction effects as differences that cannot be explained by simply adding single treatment effects. This additive-scale definition captures interaction effects in terms of absolute differences, without imposing any assumption that the outcome-generating process is additive. For example, when $K = 2$ and $S = \{1, 2\}$, $\tau_{\text{AIE}}(\{1, 2\}) = \mathbb{E}_{\boldsymbol{x}}[\mu(\boldsymbol{x}, (1, 1)) - \mu(\boldsymbol{x}, (1, 0)) - \mu(\boldsymbol{x}, (0, 1)) + \mu(\boldsymbol{x}, (0, 0))]$. $\tau_{\text{AIE}}(\{1, 2\})$ quantifies how much the treatment effect of $t = (1, 1)$ differs from what would be expected if the treatment effects of $(1, 0)$ and $(0, 1)$ were simply additive.

A direct consequence of Proposition 1 is that our main estimands are also identified. The proofs for Proposition 1 and Corollary 1 are provided in the Appendix A.

**Corollary 1.** *Under Assumptions 1-3, the ASE and AIE are identified.*

## 4 Causal Inference for Single and Interaction Treatment Effects Network

### 4.1 Model Architecture

The objective of CISI-Net is to learn $\mu(\boldsymbol{x}, \boldsymbol{t})$ and estimate the causal effects defined in Equations (3) and (4) for any possible treatment vector $\boldsymbol{t}$. Figure 1 illustrates the architecture of CISI-Net for estimating $\hat{\mu}(\boldsymbol{x}, \boldsymbol{t})$. The model consists of three main components: (i) a representation learning network with the balancing penalty, (ii) a task embedding network, and (iii) an outcome prediction network. These components are jointly optimized end-to-end, which ensures that all components are consistently aligned to learn $\mu(\boldsymbol{x}, \boldsymbol{t})$ and reliably estimate the causal effects.

First, the representation learning network with the balancing penalty maps the observed covariates $\boldsymbol{x} \in \mathbb{R}^d$ to a latent representation space which is suitable for estimating causal effects with reduced distributional imbalances caused by selection bias. Formally, it learns a function $\Phi : \mathbb{R}^d \to \mathbb{R}^p$, where $p$ is a hyperparameter indicating the dimensionality of the learned representations. A balancing penalty based on the IPM (Shalit et al., 2017) is applied to the learned representations to minimize the discrepancy between the distributions associated with different treatment patterns, which reduces distributional imbalances. This ensures that

counterfactual predictions do not suffer from high variance arising from differences in covariate distributions across treatment patterns.

Second, the task embedding network enables the model to treat outcome prediction across treatment patterns as related prediction tasks by capturing similarities among treatment patterns, rather than modeling each treatment pattern independently. Specifically, the task embedding network maps the binary treatment vector $\boldsymbol{t}$ into a $q$-dimensional continuous task embedding vector via a multi-layer perceptron $\mathrm{MLP}_w$ (LeCun et al., 2015), denoted as $t_w(\boldsymbol{t}) = \mathrm{MLP}_w(\boldsymbol{t}) \in \mathbb{R}^q$. This embedding captures similarities among treatments by positioning treatment patterns with related causal effects closer in the embedding space. As a result, the embedding itself encodes information that distinguishes components common to single treatments from contributions specific to interaction treatment effects. The embedding vector $t_w(\boldsymbol{t})$ is then passed to the outcome prediction network, where parameter sharing across similar treatments becomes possible.

Third, the outcome prediction network takes the concatenated vector of $\Phi(\boldsymbol{x})$ and $t_w(\boldsymbol{t})$ as input, which forms a $(p+q)$-dimensional representation, and outputs the predicted outcome $Y$ using a neural network $h : \mathbb{R}^{p+q} \to \mathbb{R}$. Unlike previous approaches that use separate networks for each treatment (Mondal et al., 2022; Parbhoo et al., 2021), CISI-Net uses a single shared prediction function $h$ for all treatment patterns. In the task embedding space, because treatment patterns with similar causal effects are positioned close to each other, CISI-Net allows the outcome prediction network to reuse parameters across related treatment patterns. Through this parameter sharing, CISI-Net achieves stable estimation by avoiding the construction of separate networks for each treatment pattern.

Our model estimates causal effects by predicting counterfactual outcomes based on the treatment vector as a direct input. Through training, the proposed model learns a function $\hat{\mu}(\boldsymbol{x}, \boldsymbol{t})$ composed of a representation network $\Phi$, a task embedding network $t_w$, and an outcome prediction network $h$. Instead of using the actual observed treatment vector $\boldsymbol{t}$, a counterfactual treatment vector $\boldsymbol{t}'$ is fed into $\hat{\mu}(\boldsymbol{x}, \boldsymbol{t})$, and the causal effect is estimated by Equations (3) and (4).

### 4.2 Objective Function

The objective function of CISI-Net provides a principled optimization strategy for simultaneously addressing two key challenges in causal effect estimation: maximizing outcome prediction accuracy and correcting distributional imbalances caused by selection bias. Optimizing only for the former can result in biased counterfactual predictions, whereas focusing solely on the latter can compromise the expressive capacity of the model. To address this balance among accuracy, bias correction, and overfitting, we design the loss function $L$ as the sum of three components:

$$L = L_y + \alpha L_\Phi(\Phi, \boldsymbol{t}) + \beta \|w\|_2, \tag{5}$$

where the first term $L_y$ represents the outcome prediction error, the second term $L_\Phi(\cdot, \cdot)$ is a balancing penalty that reduces distributional differences in the representation space across treatment patterns, and the third term is an L2 regularization term applied to the network weights. The coefficients $\alpha$ and $\beta$ are hyperparameters that control the strength of the corresponding components.

Minimizing the outcome prediction error $L_y$ is essential for an accurate estimation of causal effects. When the frequency of observed treatment patterns is imbalanced, certain treatment patterns may be underrepresented, which results in biased predictions. To address this issue, we introduce a correction based on the empirical frequency of each treatment pattern (Shalit et al., 2017):

$$w_i(\boldsymbol{t}_i) = \frac{1}{2} \left( \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}[\boldsymbol{t}_i = \boldsymbol{t}_j] \right)^{-1}, \tag{6}$$

$$L_y = \frac{1}{N} \sum_{i=1}^{N} w_i(\boldsymbol{t}_i)(y_i - \hat{y}_i)^2, \tag{7}$$

where $\hat{y}_i$ denotes the predicted outcome for unit $i$, and $w_i(\boldsymbol{t}_i)$ is the scaled inverse of the relative frequency of the treatment $\boldsymbol{t}_i$ in the observed data.

The second term $L_\Phi$ is designed to reduce estimation variance by aligning the representation distributions in all treatment patterns. This term is defined by computing the IPM between the distribution of learned representations of each pair of distinct treatment patterns and averaging these discrepancies. The use of IPM allows us to effectively capture non-linear discrepancies between probability distributions, and thus contributes to better generalization in causal effect estimation (Shalit et al., 2017; Shi et al., 2019):

$$L_\Phi = \frac{1}{\binom{|T|}{2}} \sum_{\{\boldsymbol{a},\boldsymbol{b}\}\in\binom{T}{2}} \mathrm{IPM}(S_{\boldsymbol{a}}, S_{\boldsymbol{b}}), \tag{8}$$

where $\Phi(x)$ is the representation network mapping the input covariates $x$ into a latent representation space, and we define the set of representations for a given treatment pattern $\boldsymbol{a}$ as $S_a \coloneqq \{\Phi(\boldsymbol{x}_i) \mid t_i = a\}$. The $\mathrm{IPM}(\cdot,\cdot)$ measures the discrepancy between the two distributions of two such sets. Here, $|T|$ denotes the number of distinct treatment patterns and $\binom{|T|}{2}$ is the total number of pairs of treatment patterns.

## 5 Simulation Experiment

This section describes the experimental setup, the hyperparameters for the proposed and baseline models, and the evaluation metrics used in the analysis. To evaluate the effectiveness of the proposed method, we conducted two types of experiments. First, we compared the proposed method with several existing methods using three simulation datasets to assess the estimation performance. Second, we conducted an ablation study using one of the three simulation datasets to investigate the contributions of the key components of the proposed method, specifically the task embedding network and the balancing penalty.

### 5.1 Simulation Datasets

We design three simulation scenarios to comprehensively evaluate our framework's performance under different causal directed acyclic graphs and outcome-generating functions. The first scenario, assuming full observability of true covariates, serves as a fundamental setting for comparing estimation accuracy against baselines in the presence of treatment interactions. The second scenario introduces latent covariates, which are true covariates but unobservable, to assess the robustness of each method in a more realistic setting. Finally, the third scenario features multiple treatments but no interaction effects. This setting is specifically designed to evaluate the robustness of CISI-Net against model misspecification by testing whether our model maintains performance even when the data lacks the interaction effects it is built to capture. Collectively, these scenarios allow for a thorough assessment of our model's accuracy, robustness to latent covariates, and resilience to model misspecification.

To clarify the words used in this section and beyond, we distinguish among three types of covariates: true covariates, latent covariates, and observed covariates. First, the true covariates refer to the complete set of variables that influence both the treatment and the outcome. When all true covariates are observed, the data satisfy the assumption of no unobserved covariates. Second, the latent covariates are a subset of or the entirety of the true covariates that are unobserved by the analyst. They represent unobserved covariates that may induce bias in the estimation of causal effects if not properly accounted for (Robins et al., 2000). Third, the observed covariates consist of variables that are the true covariates themselves or proxy variables that partially capture the information of latent covariates. The observed covariates represent the variables that are available to the analyst and can be directly used for causal effect estimation.

We vary the structure of the observed covariates $\boldsymbol{x}_{i,\mathrm{o}}$ and the true covariates $\boldsymbol{x}_{i,\mathrm{t}}$ by scenarios. In all scenarios, only the observed covariates $\boldsymbol{x}_{i,\mathrm{o}}$ are available to the model during training and evaluation. The true covariates $\boldsymbol{x}_{i,\mathrm{t}}$ are used exclusively for the data-generating process (i.e., to generate treatments and outcomes) and are treated as unobservable to the model at inference time.

In all scenarios, given the true covariates $\boldsymbol{x}_{i,\mathrm{t}}$, treatment assignment and outcome generation follow the same functional form, with the number of treatments fixed at $K = 3$ and the sample size set to $N = 50,000$. We introduce an indicator variable $H$ to mirror realistic conditions under which treatment assignment depends

on true covariates in a non-smooth and discontinuous (Ascarza, 2018).

$$H_i^{(1)} = \mathbb{I}\left(x_{i,\mathrm{t}}^{(1)} + x_{i,\mathrm{t}}^{(2)} > 1\right), \quad H_i^{(2)} = \mathbb{I}\left(x_{i,\mathrm{t}}^{(3)} + x_{i,\mathrm{t}}^{(4)} < -0.5\right),$$
$$t_i^{(k)} \sim \mathrm{Bern}\left(\sigma\left(\boldsymbol{w}_{t_k}^\top \boldsymbol{x}_{i,\mathrm{t}} - \lambda_1 H_i^{(1)} - \lambda_2 H_i^{(2)} - \delta\right)\right), \quad k = 1, 2, 3,$$
$$\boldsymbol{t_i} = (t_i^{(1)}, t_i^{(2)}, t_i^{(3)}),$$
$$Y_i \sim \mathcal{N}\big(f(\boldsymbol{x}_{i,\mathrm{t}}, \boldsymbol{t}_i, l)\,,\, 1^2\big), \quad l \in \{0, 1\},$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\sigma(\cdot)$ is the sigmoid function defined as $\sigma(x) = 1/(1 + \exp(-x))$, and $\delta$ and $\lambda$ are bias parameters that vary across simulation scenarios. The vector $\boldsymbol{w}_{t_k}$ is the weight vector whose elements are independently drawn from the uniform distribution $U(-0.5, 0.5)$. The outcome-generating function $f$ is defined as follows:

$$\begin{aligned} f(\boldsymbol{x}_{i,\mathrm{t}}, \boldsymbol{t_i}, l) = {} & \boldsymbol{w}_x^\top \boldsymbol{x}_{i,\mathrm{t}} + (x_{i,\mathrm{t}}^{(1)} + 1)t_i^{(1)} + 1.2(x_{i,\mathrm{t}}^{(2)} + 1)t_i^{(2)} \\ & + 0.8(x_{i,\mathrm{t}}^{(3)} + 1)t_i^{(3)} + l\Big\{(x_{i,\mathrm{t}}^{(4)} + 0.5)t_i^{(1)}t_i^{(2)} \\ & - 0.5(x_{i,\mathrm{t}}^{(5)} - 0.1)t_i^{(1)}t_i^{(3)} + 0.1(x_{i,\mathrm{t}}^{(6)} + 0.1)t_i^{(2)}t_i^{(3)} \\ & + 0.7(x_{i,\mathrm{t}}^{(7)} + 1)t_i^{(1)}t_i^{(2)}t_i^{(3)}\Big\} + 2, \end{aligned}$$

where vector $\boldsymbol{w}_x$ is the weight vector whose elements are independently drawn from the uniform distribution $U(-1, 1)$, and the parameter $l \in \{0, 1\}$ determines whether interaction effect terms are included $(l = 1)$ or excluded $(l = 0)$.

The first scenario assumes that true covariates are directly observable, as specified by the following equations:

$$x_{i,n}^{(j)} \sim N(c_n^{(j)}, 1^2), \quad x_{i,u}^{(j)} \sim U(-1, 1),$$
$$\boldsymbol{x}_{i,\mathrm{o}} = \boldsymbol{x}_{i,\mathrm{t}} = (\boldsymbol{x}_{i,n}, \boldsymbol{x}_{i,u}),$$

where $j \in \{1, \ldots, 15\}$ and $c_n^{(j)}$ are drawn from the uniform distribution $U(-1, 1)$. In simulation dataset 1, $l = 1$ , $\delta = 1$, $\lambda_1 = 1$ and $\lambda_2 = 1.2$.

The second scenario reflects the assumption adopted by deep generative models (Louizos et al., 2017; Saini et al., 2019) that latent covariates exist. Specifically, it considers real-world situations where latent covariates, such as the economic status or lifestyle of units, exist, and only proxy variables, which correspond to the observed covariates (e.g., residential area, occupation, purchase history) are observed. Mathematically, this data-generating process is defined as:

$$\boldsymbol{z}_i \sim N(\boldsymbol{0}, I_{10}), \quad x_{i,n}^{(j)} \sim N(\boldsymbol{w}_n^{(j)\top} \boldsymbol{z}_i, 1^2),$$
$$x_{i,u}^{(j)} \sim N(\boldsymbol{w}_u^{(j)\top} \boldsymbol{z}_i, 5^2), \quad x_{i,b}^{(j)} \sim \mathrm{Bern}(\sigma(\boldsymbol{w}_b^{(j)\top} \boldsymbol{z}_i)),$$
$$\boldsymbol{x}_{i,\mathrm{o}} = (\boldsymbol{x}_{i,n}, \boldsymbol{x}_{i,u}, \boldsymbol{x}_{i,b}), \quad \boldsymbol{x}_{i,\mathrm{t}} = \boldsymbol{z}_i,$$

where $j \in \{1, \ldots, 10\}$ and the vectors $\boldsymbol{w}_n^{(j)}$, $\boldsymbol{w}_u^{(j)}$, and $\boldsymbol{w}_b^{(j)}$ are weight vectors whose elements are drawn independently from $U(-1, 1)$. $I_d$ denotes the identity matrix $d \times d$, which is also used as the variance-covariance matrix. In simulation dataset 2, $l = 1$ , $\delta = 0.8$, $\lambda_1 = 0.1$ and $\lambda_2 = 0.1$.

The third scenario adapts the structure of true and observed covariates to match that in the first scenario and assumes multiple treatments exist, with no interaction effects present. In simulation dataset 3, the interaction treatment effect is eliminated by setting the interaction effect control parameter to $l = 0$, and it is set to $\delta = 1$, $\lambda_1 = 1$ and $\lambda_2 = 1.2$.

In summary, the three scenarios are designed to isolate different challenges. Scenario 1 evaluates estimation accuracy under full observability. Scenario 2 examines robustness when true covariates are latent and observed covariates are only proxy variables. Scenario 3 tests robustness to model misspecification in the absence of interaction effects. Additional results for datasets generated under different settings (e.g., homogeneous treatment effects and varying numbers of treatments) are provided in Appendix B.

Table 2: Comparison of the number of trainable parameters across methods.

| | TARNet | CFR-WASS | TECE-VAE | NCoRE | CISI-Net |
|---|---|---|---|---|---|
| #Params | 0.73M | 0.73M | 0.43M | 0.28M | 0.14M |

### 5.2 Implementation Details

Our proposed method consists of three neural networks, all of which are built using fully connected (FC) layers (LeCun et al., 2015) with 200 units per hidden layer and leaky ReLU activation (Xu et al., 2020). Three neural networks have two hidden layers. The task embedding network $t_w$ outputs a five-dimensional embedding vector. The balancing penalty coefficient $\alpha$ was set to 0.1, and the IPM used in the penalty was the Wasserstein distance (Sriperumbudur et al., 2010).

To enable a comprehensive comparison between methods that extend single treatment models to multiple treatment settings and those specifically designed for multiple treatments, we select four baseline methods: TARNet (Shalit et al., 2017), CFR with Wasserstein-based balancing (CFR-WASS)(Shalit et al., 2017), TECE-VAE (Saini et al., 2019), and NCoRE (Parbhoo et al., 2021). Although H-Learner (Chauhan et al., 2025) can be reduced to a single-outcome special case, we do not include it as a baseline because the original paper reports that its performance in this setting is comparable to existing single-outcome baselines. To adapt TARNet and CFR-WASS to the multi-treatment setting, we construct a separate outcome prediction network for each of the $2^K$ treatment patterns. In CFR-WASS, the balancing penalty coefficient $\alpha$ is fixed at one. In TECE-VAE, the latent dimension is set to 25, and the task embedding network has three hidden layers with 200 units and ELU activation (Clevert et al., 2016). The task embedding network $t_w$ produces a five-dimensional embedding vector. In NCoRE, each interaction subnetwork is implemented as two FC layers with 200 units per layer and ReLU activation (Nair & Hinton, 2010). Table 2 shows the number of trainable parameters for each method as a reference for model capacity.

All models are trained with Adam optimizer (Kingma & Ba, 2017), and used a learning rate of $10^{-4}$, a batch size of 256, and an L2 regularization of $10^{-5}$. Each dataset is divided into training sets of 70% and test sets of 30%, and all evaluations are carried out on the test set. Training a model with one dataset and evaluating causal effects on an unseen test set is a conventional estimation scheme in causal inference using machine learning (Liu et al., 2020; Okasa, 2022).

### 5.3 Evaluation Metrics

We evaluate the estimation performance of single and interaction treatment effects using absolute errors. These are defined analogously to the absolute ATE estimation error commonly used in single-treatment studies (Cheng et al., 2022; Johansson et al., 2016), and computed as the absolute difference between the true and estimated ASE or AIE. Specifically, we define $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$ as follows:

$$\epsilon_{\text{ASE}}(k) = \left|\tau_{\text{ASE}}(k) - \hat{\tau}_{\text{ASE}}(k)\right|, \tag{9}$$

$$\epsilon_{\text{AIE}}(S) = \left|\tau_{\text{AIE}}(S) - \hat{\tau}_{\text{AIE}}(S)\right|, \tag{10}$$

where $k \in \{1, \dots K\}$ indexes single treatments, and $S$ denotes a subset of treatments with $|S| \geq 2$. $\hat{\tau}_{\text{ASE}}(\cdot)$ and $\hat{\tau}_{\text{AIE}}(\cdot)$ denote the estimated ASE and AIE.

To ensure robustness against randomness in data generation, we generate 100 independent datasets for each scenario using different random seeds. For both the baseline comparison and the ablation study, we evaluate each model across these 100 datasets and report the average of the resulting $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$.

## 6 Simulation Results

Table 3 shows $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$ of the proposed and baseline methods across three types of simulation datasets. In all datasets, CISI-Net consistently achieved the lowest $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$, emphasizing its superior ability to accurately estimate both single and interaction treatment effects under diverse conditions. Notably, this

Table 3: Comparison of mean and standard deviation of $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$ across simulation datasets. Here, $k \in \{1, 2, 3\}$ indexes single treatments for ASE, and $S \subseteq \{1, 2, 3\}$ (with $|S| \geq 2$) denotes treatment combinations for AIE.

| | | $\epsilon_{\text{ASE}}$ | | | $\epsilon_{\text{AIE}}$ | | | |
|---|---|---|---|---|---|---|---|---|
| Sim. | Method | $k = 1$ | $k = 2$ | $k = 3$ | $S = \{1, 2\}$ | $S = \{2, 3\}$ | $S = \{1, 3\}$ | $S = \{1, 2, 3\}$ |
| 1 | TARNet | 0.12 ± 0.10 | 0.12 ± 0.07 | 0.13 ± 0.07 | 0.18 ± 0.15 | 0.16 ± 0.12 | 0.18 ± 0.14 | 0.34 ± 0.29 |
| | CFR-WASS | 0.13 ± 0.09 | 0.13 ± 0.08 | 0.14 ± 0.06 | 0.19 ± 0.12 | 0.17 ± 0.12 | 0.17 ± 0.13 | 0.33 ± 0.19 |
| | NCoRE | **0.10 ± 0.08** | 0.10 ± 0.08 | 0.10 ± 0.08 | 0.17 ± 0.12 | 0.16 ± 0.10 | 0.17 ± 0.11 | 0.23 ± 0.18 |
| | TECE-VAE | 0.13 ± 0.10 | 0.13 ± 0.10 | 0.12 ± 0.09 | **0.12 ± 0.08** | 0.14 ± 0.10 | 0.15 ± 0.11 | 0.21 ± 0.15 |
| | **CISI-Net** | **0.10 ± 0.07** | **0.09 ± 0.08** | **0.08 ± 0.07** | **0.12 ± 0.10** | **0.13 ± 0.10** | **0.12 ± 0.10** | **0.12 ± 0.10** |
| 2 | TARNet | 0.19 ± 0.14 | 0.27 ± 0.16 | 0.14 ± 0.12 | 0.27 ± 0.16 | **0.23 ± 0.15** | 0.20 ± 0.14 | 0.32 ± 0.26 |
| | CFR-WASS | 0.20 ± 0.14 | 0.26 ± 0.16 | 0.16 ± 0.13 | 0.28 ± 0.17 | **0.23 ± 0.17** | 0.22 ± 0.16 | 0.36 ± 0.28 |
| | NCoRE | 0.32 ± 0.27 | 0.35 ± 0.28 | 0.30 ± 0.23 | 0.31 ± 0.27 | 0.36 ± 0.29 | 0.36 ± 0.28 | 0.55 ± 0.36 |
| | TECE-VAE | 0.21 ± 0.17 | 0.24 ± 0.19 | 0.17 ± 0.13 | 0.29 ± 0.23 | 0.24 ± 0.14 | 0.19 ± 0.12 | 0.38 ± 0.19 |
| | **CISI-Net** | **0.16 ± 0.13** | **0.19 ± 0.13** | **0.12 ± 0.09** | **0.26 ± 0.18** | **0.23 ± 0.14** | **0.18 ± 0.12** | **0.21 ± 0.14** |
| 3 | TARNet | 011 ± 0.08 | 0.11 ± 0.72 | 0.10 ± 0.07 | 0.17 ± 0.12 | 0.14 ± 0.12 | 0.16 ± 0.12 | 0.31 ± 0.26 |
| | CFR-WASS | 011 ± 0.08 | 0.11 ± 0.09 | 0.10 ± 0.07 | 0.16 ± 0.13 | 0.14 ± 0.11 | 0.16 ± 0.12 | 0.28 ± 0.20 |
| | NCoRE | **0.10 ± 0.08** | 0.10 ± 0.07 | 0.09 ± 0.06 | 0.17 ± 0.13 | 0.16 ± 0.12 | 0.15 ± 0.13 | 0.19 ± 0.16 |
| | TECE-VAE | 0.12 ± 0.10 | 0.12 ± 0.09 | 0.10 ± 0.08 | 0.13 ± 0.11 | **0.12 ± 0.09** | **0.10 ± 0.09** | 0.11 ± 0.09 |
| | **CISI-Net** | **0.10 ± 0.06** | **0.09 ± 0.08** | **0.08 ± 0.06** | **0.12 ± 0.09** | **0.12 ± 0.08** | **0.10 ± 0.08** | **0.09 ± 0.07** |

performance was achieved despite CISI-Net having fewer trainable parameters than the baseline methods, as shown in Table 2. In simulation scenario 2, where the observed covariates are only proxy variables for latent covariates, CISI-Net maintained high estimation performance. These results demonstrate that even when only proxy variables rather than true covariates can be observed, CISI-Net can estimate causal effects with high performance without relying on latent variable estimation. Furthermore, in simulation scenario 3, where no actual interaction treatment effects exist, and the utility of the task embedding network is limited, CISI-Net still achieved the best accuracy. These results suggest that CISI-Net is flexible, because it captures interaction treatment effects when they exist and avoids degradation when they do not. This flexibility offers a significant practical advantage, because it frees analysts from the need to pre-screen for interaction treatment effects before applying CISI-Net. In this way, CISI-Net ensures stable performance across various data generation processes, whether interaction is present or absent, providing both accuracy and reliability compared to baselines.

When extended to multi-treatment settings, single-treatment architectures like TARNet and CFR-WASS are fundamentally limited by their inability to share information across similar treatment patterns. Although their original design enables relatively accurate estimation of ASE in some scenarios, these methods treat each treatment pattern independently and construct separate outcome prediction networks for each case. Because parameter updates occur only for samples that receive the corresponding treatment, they cannot use similarities among similar treatment patterns and thus miss opportunities for parameter sharing. This structural limitation explains why these methods, although competitive in estimating single treatment effects, are considerably less stable when estimating interaction effects. These results suggest that naively extending single-treatment architectures is insufficient for robust estimation under multiple concurrent treatments.

NCoRE, which possesses interaction subnetworks updated only in samples receiving multiple treatments, demonstrates the limitations of estimating interaction effects. Since such samples are typically rare, these subnetworks are poorly trained under data sparsity, which results in large $\epsilon_{\text{AIE}}$. This instability becomes more severe in simulation scenario 2, because the interaction subnetworks must be learned from imbalanced treatment combinations under a more challenging causal inference setting. Moreover, its architecture, which

Table 4: Ablation study on simulation dataset 1. Here, $k \in \{1, 2, 3\}$ indexes single treatments for ASE, and $S \subseteq \{1, 2, 3\}$ (with $|S| \geq 2$) denotes treatment combinations for AIE. TE indicates whether the task embedding network is used, and BP indicates whether the balancing penalty is applied. $\alpha$ is set to 0.1.

| | | $\epsilon_{\text{ASE}}$ | | | $\epsilon_{\text{AIE}}$ | | | |
|---|---|---|---|---|---|---|---|---|
| TE | BP | $k=1$ | $k=2$ | $k=3$ | $S=\{1,2\}$ | $S=\{2,3\}$ | $S=\{1,3\}$ | $S=\{1,2,3\}$ |
| ✗ | ✗ | **0.10 ± 0.07** | 0.11 ± 0.07 | 0.10 ± 0.05 | 0.18 ± 0.10 | 0.25 ± 0.14 | 0.25 ± 0.10 | 0.66 ± 0.04 |
| ✓ | ✗ | 0.11 ± 0.09 | 0.11 ± 0.08 | 0.09 ± 0.07 | 0.18 ± 0.15 | 0.19 ± 0.15 | 0.21 ± 0.14 | 0.14 ± 0.10 |
| ✗ | ✓ | **0.10 ± 0.08** | 0.10 ± 0.07 | 0.10 ± 0.06 | 0.16 ± 0.11 | 0.23 ± 0.14 | 0.25 ± 0.13 | 0.59 ± 0.07 |
| ✓ | ✓ | **0.10 ± 0.07** | **0.09 ± 0.08** | **0.08 ± 0.07** | **0.12 ± 0.10** | **0.13 ± 0.10** | **0.12 ± 0.10** | **0.12 ± 0.10** |

constructs outcome and interaction networks independently, prevents any parameter sharing across treatment patterns. This design misses a critical opportunity to use treatment similarities, which in turn leads to more unstable estimates. In simulation scenario 3, these interaction subnetworks become redundant, which in turn limits stable estimation when interaction is absent.

Although TECE-VAE captures interaction treatment effects, it underperforms compared to CISI-Net due to unnecessary latent recovery caused by model misspecification. In simulation scenario 1, where no latent covariates exist, the model is forced to infer unnecessary latent variables, which results in degraded accuracy in both ASE and AIE estimation. In simulation scenario 2, which is a favorable setting for TECE-VAE, TECE-VAE remains competitive for some $\epsilon_{\text{AIE}}$ involving two treatments but still underperforms CISI-Net in most metrics, especially in $\epsilon_{\text{AIE}}(\{1, 2, 3\})$. These results show that reliance on latent assumptions of TECE-VAE degrades robustness across diverse settings compared to CISI-Net.

Table 4 shows the results of an ablation study on simulation scenario 1 that evaluates the impact of the task embedding network and the balancing penalty on the performance of CISI-Net. These results demonstrate that whereas each component individually contributes to performance, combining the task embedding network with a balancing penalty yields the lowest estimation errors for all metrics. The balancing penalty, which reduces estimation variance, alone improves AIE estimation because units exposed to multiple treatments are subject to strong selection bias. The task embedding network encodes both components common to single treatments and contributions unique to interaction effects, and this encoding improves the accuracy of estimating single and interaction treatment effects. The interaction of each component enables the model to consistently achieve lower errors across both single and interaction effects compared to using either component in isolation. The additional results of the ablation study using other simulation datasets are presented in Appendix B.3.

Figure 2 demonstrates that the learned task embedding vectors by CISI-Net successfully reflect the structural similarity of treatment patterns. As the Jaccard similarity (Ji et al., 2013) between treatment vectors increases, the cosine similarity (Steck et al., 2024) between their corresponding embedding vectors also rises consistently. This result indicates that the task embedding network captures the overlap in treatment components and encodes their causal relevance. This architectural choice enables parameter sharing across related treatment patterns, which in turn supports the accurate estimation of both single and interaction effects.

Figure 3 shows $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$ for different values of the balancing penalty coefficient $\alpha$ of our CISI-Net. The results reveal that a small positive value for $\alpha$ consistently achieves the lowest $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$, emphasizing the importance of moderate distributional balance. This finding illustrates a fundamental trade-off inherent to balancing-based causal inference methods. When $\alpha = 0$, although the model prioritizes predictive accuracy, the lack of a balancing penalty introduces higher variance due to distributional imbalances, especially for AIE. Conversely, an excessively large $\alpha$ (e.g., $\alpha = 10$) degrades performance by forcing distributional consistency at the expense of the predictive power of the representation. This sensitivity is not a limitation unique to our CISI-Net and is a well-documented characteristic of models that employ balancing regularization (Johansson et al., 2022; Shalit et al., 2017). Therefore, developing a mechanism to adjust $\alpha$ based on data characteristics

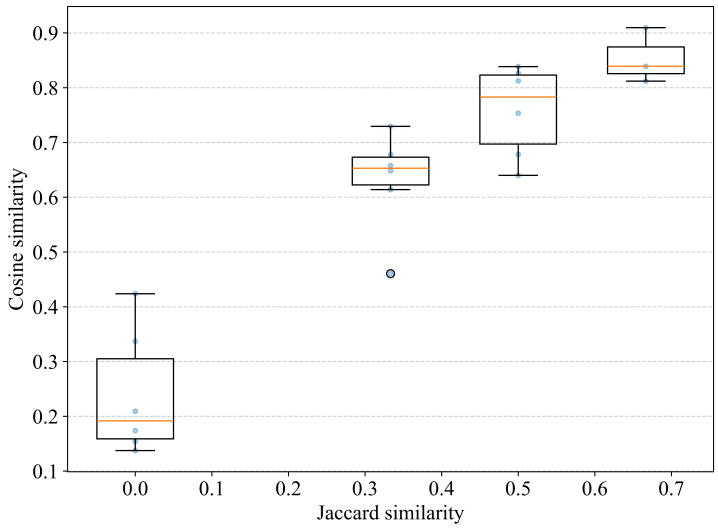



Figure 2: Relationship between treatment vector similarity and learned task embedding vector similarity in simulation dataset 1. For two treatment vectors $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$, their corresponding task embedding vectors are denoted as $t_w(\boldsymbol{t}_1)$ and $t_w(\boldsymbol{t}_2)$. The x-axis shows the Jaccard similarity between $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$, and the y-axis shows the cosine similarity between $t_w(\boldsymbol{t}_1)$ and $t_w(\boldsymbol{t}_2)$. Box plots summarize the distribution of cosine similarities for each Jaccard similarity value.

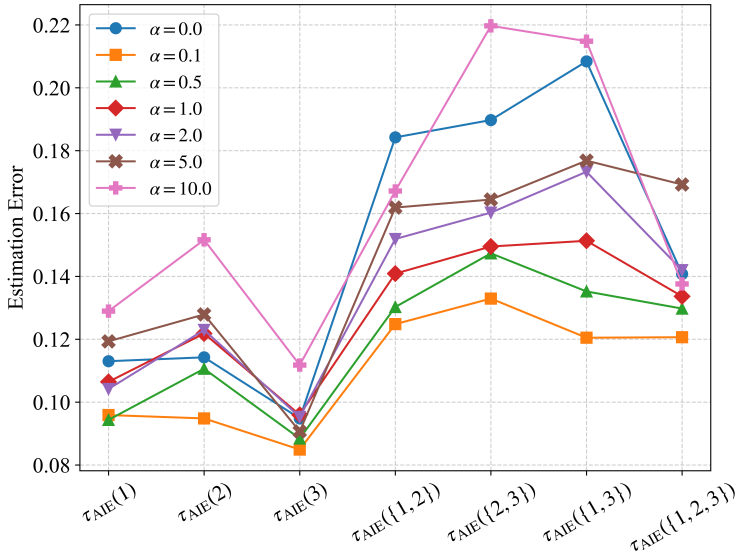



Figure 3: $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$ for different values of the balancing penalty coefficient $\alpha$ in CISI-Net.

adaptively presents a promising direction for future research, which could enhance the stability and versatility of this entire class of models.

Figure 4 shows $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$ of CISI-Net under different sample sizes. The results indicate that both $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$ decrease as the sample size increases, but $\epsilon_{\text{AIE}}$ remain high even at $N = 10000$. In the multiple-treatment setting, this strict sample requirement for estimating causal effects is a universal challenge because the number of available samples for each treatment pattern decreases as the number of treatment combinations increases (Chauhan et al., 2025; Qian et al., 2021). A possible future direction is to integrate complementary data augmentation methods, which follow a different paradigm, into CISI-Net. In particular,

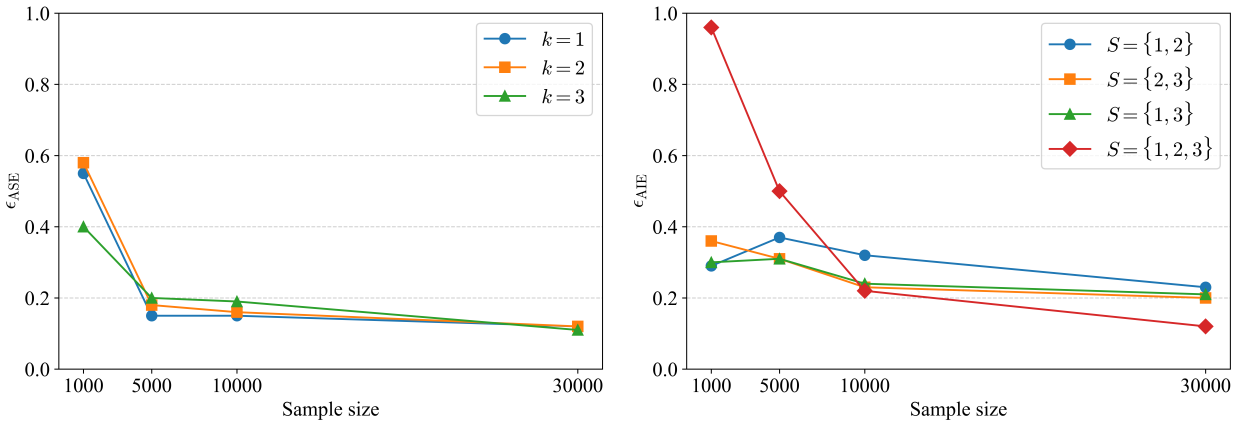



Figure 4: Estimation errors of CISI-Net under varying sample sizes on simulation dataset 1. The left panel shows $\epsilon_{\text{ASE}}$, and the right panel shows $\epsilon_{\text{AIE}}$. Each line represents the mean error computed over 100 random seeds.

CISI-Net can be extended with a data augmentation strategy that uses learned task embeddings to identify related treatment combinations and generate informative synthetic samples for estimating interaction effects for rare treatment patterns.

## 7 Application to Multiple Marketing Promotions

In this section, we apply CISI-Net to a real-world dataset derived from multiple marketing promotions. The marketing promotions were conducted by merchants affiliated with a mobile payment platform operated by a mobile payment provider in Japan. The objective of this analysis is to empirically demonstrate that CISI-Net can estimate both single and interaction treatment effects under observational data.

We use two real-world marketing datasets in which multiple marketing promotions were simultaneously conducted. The first dataset includes three marketing promotions: two promotions organized by the same merchant group (denoted as $\text{CP}_1$ and $\text{CP}_2$) and one online campaign conducted by another merchant group (denoted as $\text{CP}_3$). The observed sample proportions were approximately 2% for $\text{CP}_1$ only, 15% for $\text{CP}_2$ only, 42% for $\text{CP}_3$ only, 0.2–1% for combinations of two CPs, 0.2% for all three promotions simultaneously, and the remainder as the control group. The second dataset includes two marketing promotions conducted by two different merchants in the same industry (denoted as $\text{CP}_4$ and $\text{CP}_5$). The observed sample proportions in the second dataset were approximately 29% for $\text{CP}_4$ only, 4% for $\text{CP}_5$ only, 1% for both promotions, and the remainder as the control group. In all datasets, the control group consisted of users who were not exposed to any promotion during the treatment period and were randomly sampled from those who had at least one mobile payment transaction in the month before the promotions.

In all datasets, we defined the outcome as the total payment amount during the one month following campaign implementation, and empirically analyzed the causal effects of the two datasets.For numerical stability and confidentiality, the outcome $Y$ was standardized within each dataset by using its own overall mean and standard deviation. Accordingly, the results reported in this section are based on standardized outcomes. The original-scale causal effects can be recovered by applying the inverse transformation of the scaler to the model outputs. This preprocessing does not distort the estimation results because standardization is an affine transformation (Thakral & Tô, 2023). In particular, consistent interpretation between the standardized and original scales is ensured because the signs of the estimated causal effects are preserved. The observed covariates consisted of 71 variables, which included service usage histories and user demographic attributes.

The hyperparameters of the proposed model were fixed to the same values as those described in Section 5.2. A sensitivity analysis regarding the balancing-penalty weight $\alpha$ in Equation (5) is presented in Appendix C.

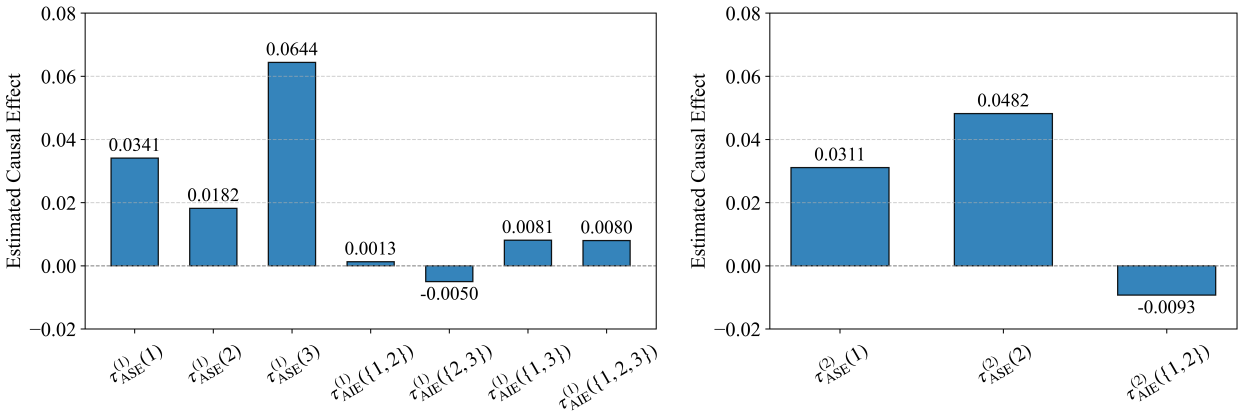



Figure 5: Estimated single and interaction treatment effects obtained from two real-world marketing promotion datasets. The left panel corresponds to dataset 1 ($CP_1$-$CP_3$), and the right panel corresponds to dataset 2 ($CP_4$ and $CP_5$). Here, $\tau^{(d)}_{\text{ASE}}(k)$ and $\tau^{(d)}_{\text{AIE}}(S)$ are the estimated effects for single and multiple treatments in dataset $d$. The outcome is standardized prior to estimation.

Each dataset is divided into training sets of 70% and test sets of 30%, and causal effects were estimated on the test set according to Equations (3) and (4).

Figure 5 shows the estimated causal effects obtained from two real-world marketing promotion datasets on a standardized outcome scale. In the first dataset (left panel), all single treatment effects were positive, with $CP_3$ achieving the largest single treatment effect among the three promotions. Regarding interaction treatment effects $\tau^{(1)}_{\text{AIE}}(\{1,2\})$ between two promotions conducted by the same merchant group, a positive effect was observed. On the other hand, the interaction treatment effects between promotions conducted by different merchant groups ($\tau^{(1)}_{\text{AIE}}(\{2,3\})$, $\tau^{(1)}_{\text{AIE}}(\{1,3\})$, and $\tau^{(1)}_{\text{AIE}}(\{1,2,3\})$) exhibited heterogeneous patterns. In the second dataset (right panel), both single treatment effects were positive, with $\tau^{(2)}_{\text{ASE}}(2)$ being relatively larger. The interaction effect $\tau^{(2)}_{\text{AIE}}(\{1,2\})$ was negative, which implies cannibalization between similar promotions conducted by different merchants.

The results from the first dataset demonstrate that the direction and magnitude of the interaction treatment effect vary depending on the combination of promotions. The synergy between $CP_1$ and $CP_2$ suggests that conducting multiple promotions within the same merchant group at the same time increases user engagement more effectively than conducting them separately at different times. This result is consistent with previous studies, which have shown that within-group promotions often produce synergies by reinforcing consumer touchpoints (Dorotic et al., 2021; Lesscher et al., 2021; Widdecke et al., 2023; Zantedeschi et al., 2017). In contrast, the heterogeneous patterns of interaction effects observed for cross-group promotions (e.g., $\tau^{(1)}_{\text{AIE}}(\{2,3\})$ and $\tau^{(1)}_{\text{AIE}}(\{1,3\})$) indicate that promotional effects may vary depending on the degree of market overlap between different merchant groups. These results are consistent with findings from partnership loyalty programs, where collaboration between distinct partners can result in either synergy or cannibalization depending on context (Dorotic et al., 2021).

The results from the second dataset suggest that significant cannibalization occurs when competing merchants conduct promotions simultaneously. These results demonstrate that when similar promotions are conducted by different merchants in the same industry at the same time, consumers may split their spending across merchants, which reduces the total promotional effect. Our analysis is consistent with previous research indicating that overlapping promotions from competing merchants tend to create cannibalization rather than synergy (Dorotic et al., 2021). Overall, these results show that CISI-Net reproduces empirically validated behavioral patterns in multi-promotion settings and enables the estimation of when and where positive or negative interactions emerge.

## 8 Conclusion

In this study, we propose the Causal Inference for Single and Interaction treatment effects Network (CISI-Net), a novel deep learning framework specifically designed to estimate both single and interaction effects in multi-treatment scenarios. CISI-Net estimates causal effects by combining a task embedding network with a representation learning network with the balancing penalty. CISI-Net encodes both elements common to single effects and contributions specific to interaction effects by using the task embedding network, thereby improving the stability and accuracy of causal effect estimation. Additionally, CISI-Net reduces estimation variance via the representation learning network with the balancing penalty.

The experimental results demonstrate the effectiveness of CISI-Net. First, the simulation studies show that our CISI-Net consistently outperforms existing baselines in estimating causal effects across a wide range of conditions, regardless of the presence of interaction effects. Second, these findings are validated by our real-world case study, in which CISI-Net successfully estimated both single and interaction effects from promotional data, and the results are consistent with prior marketing research. These results suggest that CISI-Net has great potential as a practical analytical tool for applications such as evaluating combined drug effects in medicine and optimizing complex promotional strategies in marketing.

Our work opens up several important avenues for future research. A first direction is to explore systematic strategies for selecting or adaptively adjusting the weight $\alpha$ of the IPM-based balancing penalty. A second direction is to investigate methodological extensions to address the strict sample requirements. This direction particularly indicates investigating methods for stably estimating interaction treatment effects, even under conditions where sample sizes per treatment pattern are severely limited, such as those involving simultaneous multiple treatments. Possible directions for future research include incorporating data augmentation–based approaches to stabilize parameter estimation. A third direction is to reduce the risk of model misspecification further, extend our framework to a doubly robust estimation setting (Kennedy, 2023) by integrating both propensity score modeling and outcome regression.

## References

Eva Ascarza. Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, 55(1):80–98, 2018. doi: 10.1509/jmr.16.0163. URL https://doi.org/10.1509/jmr.16.0163.

Vinod Kumar Chauhan, Lei Clifton, Gaurav Nigam, and David A. Clifton. Individualised treatment effects estimation with composite treatments and composite outcomes. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–7. IEEE, July 2025. doi: 10.1109/embc58623.2025.11252651. URL http://dx.doi.org/10.1109/EMBC58623.2025.11252651.

Lu Cheng, Ruocheng Guo, Raha Moraffah, Paras Sheth, K. Selçuk Candan, and Huan Liu. Evaluation methods and measures for causal learning algorithms. *IEEE Transactions on Artificial Intelligence*, 3(6): 924–943, 2022. doi: 10.1109/TAI.2022.3150264.

Ho Jin Cho, Jungmi Chae, Sang-Heon Yoon, and Dong-Sook Kim. Aging and the prevalence of polypharmacy and hyper-polypharmacy among older adults in south korea: A national retrospective study during 2010–2019. *Frontiers in Pharmacology*, Volume 13 - 2022, 2022. ISSN 1663-9812. doi: 10.3389/fphar.2022.866318. URL https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2022.866318.

Zhixuan Chu, Hui Ding, Guang Zeng, Yuchen Huang, Tan Yan, Yulin Kang, and Sheng Li. Hierarchical capsule prediction network for marketing campaigns effect. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pp. 3043–3051, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557099. URL https://doi.org/10.1145/3511808.3557099.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus), 2016. URL https://arxiv.org/abs/1511.07289.

Peter J. Danaher, André Bonfrer, and Sanjay Dhar. The effect of competitive advertising interference on sales for packaged goods. *Journal of Marketing Research*, 45(2):211–225, 2008. doi: 10.1509/jmkr.45.2.211. URL https://doi.org/10.1509/jmkr.45.2.211.

Matilda Dorotic, Dennis Fok, Peter C. Verhoef, and Tammo H. A. Bijmolt. Synergistic and cannibalization effects in a partnership loyalty program. *Journal of the Academy of Marketing Science*, 49(5):1021–1042, Sep 2021. ISSN 1552-7824. doi: 10.1007/s11747-020-00759-7. URL https://doi.org/10.1007/s11747-020-00759-7.

Naoki Egami and Kosuke Imai. Causal interaction in factorial experiments: Application to conjoint analysis. *Journal of the American Statistical Association*, 114(526):529–540, 2019. doi: 10.1080/01621459.2018.1476246. URL https://doi.org/10.1080/01621459.2018.1476246.

Boris Freidlin, Edward L. Korn, Robert Gray, and Alison Martin. Multi-arm clinical trials of new agents: Some design considerations. *Clinical Cancer Research*, 14(14):4368–4371, 07 2008. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-08-0325. URL https://doi.org/10.1158/1078-0432.CCR-08-0325.

Alan H. Gradman, Jan N. Basile, Barry L. Carter, and George L. Bakris. Combination therapy in hypertension. *Journal of the American Society of Hypertension*, 4(2):90–98, 2010. ISSN 1933-1711. doi: 10.1016/j.jash.2010.03.001. URL https://www.sciencedirect.com/science/article/pii/S1933171110000434.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Jianqiu Ji, Jianmin Li, Shuicheng Yan, Qi Tian, and Bo Zhang. Min-max hash for jaccard similarity. In *2013 IEEE 13th International Conference on Data Mining*, pp. 301–309, 2013. doi: 10.1109/ICDM.2013.119.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 3020–3029, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/johansson16.html.

Fredrik D. Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23(166):1–50, 2022. URL http://jmlr.org/papers/v23/19-511.html.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008 – 3049, 2023. doi: 10.1214/23-EJS2157. URL https://doi.org/10.1214/23-EJS2157.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1804597116.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL https://doi.org/10.1038/nature14539.

Katherine N. Lemon and Stephen M. Nowlis. Developing synergies between promotions and brands in different price-quality tiers. *Journal of Marketing Research*, 39(2):171–185, 2002. doi: 10.1509/jmkr.39.2.171.19089. URL https://doi.org/10.1509/jmkr.39.2.171.19089.

Lisan Lesscher, Lara Lobschat, and Peter C. Verhoef. Do offline and online go hand in hand? cross-channel and synergy effects of direct mailing and display advertising. *International Journal of Research in Marketing*, 38(3):678–697, 2021. ISSN 0167-8116. doi: https://doi.org/10.1016/j.ijresmar.2020.11.003. URL https://www.sciencedirect.com/science/article/pii/S0167811620301002.

Ruoqi Liu, Changchang Yin, and Ping Zhang. Estimating individual treatment effects with time-varying confounders. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 382–391, 2020. doi: 10.1109/ICDM50108.2020.00047.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/94b5bde6de888ddf9cde6748ad2523d1-Paper.pdf.

Jing Ma, Ruocheng Guo, Aidong Zhang, and Jundong Li. Multi-cause effect estimation with disentangled confounder representation. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021. doi: 10.24963/ijcai.2021/384. URL https://par.nsf.gov/biblio/10300529.

Reza Bayat Mokhtari, Tina S Homayouni, Narges Baluch, Evgeniya Morgatskaya, Sushil Kumar, Bikul Das, and Herman Yeger. Combination therapy in combating cancer. *Oncotarget*, 8(23):38022, 2017.

Abhirup Mondal, Anirban Majumder, and Vineet Chaoji. MEMENTO: Neural model for estimating individual treatment effects for multiple treatments. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pp. 3381–3390, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557125. URL https://doi.org/10.1145/3511808.3557125.

Karthik Muralidharan, Mauricio Romero, and Kaspar Wüthrich. Factorial designs, model selection, and (incorrect) inference in randomized experiments. *The Review of Economics and Statistics*, 107(3):589–604, 05 2025. ISSN 0034-6535. doi: 10.1162/rest_a_01317. URL https://doi.org/10.1162/rest_a_01317.

Prasad A. Naik and Kalyan Raman. Understanding the impact of synergy in multimedia communications. *Journal of Marketing Research*, 40(4):375–388, 2003. doi: 10.1509/jmkr.40.4.375.19385. URL https://doi.org/10.1509/jmkr.40.4.375.19385.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Gabriel Okasa. Meta-learners for estimation of causal effects: Finite sample cross-fit performance, 2022. URL https://arxiv.org/abs/2201.12692.

Sonali Parbhoo, Stefan Bauer, and Patrick Schwab. NCoRE: Neural counterfactual representation learning for combinations of treatments, 2021. URL https://arxiv.org/abs/2103.11175.

Mahesh K. B. Parmar, Friederike. M.-S. Barthel, Matthew Sydes, Ruth Langley, Rick Kaplan, Elizabeth Eisenhauer, Mark Brady, Nicholas James, Michael A. Bookman, Ann-Marie Swart, Wendi Qian, and Patrick Royston. Speeding up the evaluation of new agents in cancer. *JNCI: Journal of the National Cancer Institute*, 100(17):1204–1214, 09 2008. ISSN 0027-8874. doi: 10.1093/jnci/djn267. URL https://doi.org/10.1093/jnci/djn267.

Zhaozhi Qian, Alicia Curth, and Mihaela van der Schaar. Estimating multi-cause treatment effects via single-cause perturbation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 23754–23767. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c793b3be8f18731f2a4c627fb3c6c63d-Paper.pdf.

James M. Robins, Andrea Rotnitzky, and Daniel O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In M. Elizabeth Halloran and Donald Berry (eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pp. 1–94, New York, NY, 2000. Springer New York. ISBN 978-1-4612-1284-3.

Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005. doi: 10.1198/016214504000001880. URL https://doi.org/10.1198/016214504000001880.

Shiv Kumar Saini, Sunny Dhamnani, Aakash Srinivasan, Akil Arif Ibrahim, and Prithviraj Chavan. Multiple treatment effect estimation using deep generative model with task embedding. In *The World Wide Web Conference*, WWW '19, pp. 1601–1611, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313744. URL https://doi.org/10.1145/3308558.3313744.

Jonas Schweisthal, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Reliable off-policy learning for dosage combinations. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 67900–67924. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d69103d7895f4e2083f24b664003d386-Paper-Conference.pdf.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/shalit17a.html.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8fb5f8be2aa9d6c64a04e3ab9f63feee-Paper.pdf.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. Non-parametric estimation of integral probability metrics. In *2010 IEEE International Symposium on Information Theory*, pp. 1428–1432, 2010. doi: 10.1109/ISIT.2010.5513626.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 887–890, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701726. doi: 10.1145/3589335.3651526. URL https://doi.org/10.1145/3589335.3651526.

Neil Thakral and Linh T Tô. *When are estimates independent of measurement units?* Boston University-Department of Economics, 2023.

Yuki Tsuboi, Yuta Sakai, Ryotaro Shimizu, and Masayuki Goto. Multiple treatment effect estimation for business analytics using observational data. *Cogent Engineering*, 11(1):2300557, 2024. doi: 10.1080/23311916.2023.2300557. URL https://doi.org/10.1080/23311916.2023.2300557.

Yixin Wang and David Blei. A proxy variable view of shared confounding. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10697–10707. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wang21c.html.

Rachel M. Webster. Combination therapies in oncology. *Nature Reviews Drug Discovery*, 15(2):81–82, 2016. ISSN 1474-1784. doi: 10.1038/nrd.2016.3. URL https://doi.org/10.1038/nrd.2016.3.

Kai A. Widdecke, Wiebke I.Y. Keller, Karen Gedenk, and Barbara Deleersnyder. Drivers of the synergy between price cuts and store flyer advertising at supermarkets and discounters. *International Journal of Research in Marketing*, 40(2):455–474, 2023. ISSN 0167-8116. doi: 10.1016/j.ijresmar.2022.10.001. URL https://www.sciencedirect.com/science/article/pii/S0167811622000702.

Jin Xu, Zishan Li, Bowen Du, Miaomiao Zhang, and Jing Liu. Reluplex made more practical: Leaky ReLU. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1–7, 2020. doi: 10.1109/ISCC50000.2020.9219587.

Hao Yang, Zexu Sun, Hongteng Xu, and Xu Chen. Revisiting counterfactual regression through the lens of gromov-wasserstein information bottleneck, 2024. URL https://arxiv.org/abs/2405.15505.

Daniel Zantedeschi, Eleanor McDonnell Feit, and Eric T Bradlow. Measuring multichannel advertising response. *Management Science*, 63(8):2706–2728, 2017.

Fujin Zhu, Jie Lu, Adi Lin, Junyu Xuan, and Guangquan Zhang. Direct learning with multi-task neural networks for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering*, 35(3): 2457–2470, 2023. doi: 10.1109/TKDE.2021.3112591.

Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, and Yue He. Counterfactual prediction for bundle treatment. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19705–19715. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e430ad64df3de73e6be33bcb7f6d0dac-Paper.pdf.

# Appendix

## A Proofs of Identifiability

This appendix provides the proofs for the identifiability of the causal estimands introduced in Section 3. We first provide the proof of Proposition 1.

*Proof.* The proof proceeds by a sequence of equalities, starting from the definition of $\mu(\boldsymbol{x}, \boldsymbol{t})$:

$$\begin{aligned}
\mu(\boldsymbol{x}, \boldsymbol{t}) &:= \mathbb{E}[Y(\boldsymbol{t}) \mid \boldsymbol{X} = \boldsymbol{x}] && \text{(by Definition in (1))} \\
&= \mathbb{E}[Y(\boldsymbol{t}) \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] && \text{(by Assumption 2)} \\
&= \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] && \text{(by Assumption 1)}
\end{aligned}$$

Assumption 3 ensures this final quantity is well-defined. This completes the proof of identifiability for $\mu(\boldsymbol{x}, \boldsymbol{t})$. □

Second, we provide the proof of Corollary 1.

*Proof.* The ASE and AIE are defined as functions of the conditional average potential outcome $\mu(\boldsymbol{x}, \boldsymbol{t})$. Specifically, they are constructed using linear combinations and expectations over the distributions of $\boldsymbol{X}$. Since Proposition 1 establishes that $\mu(\boldsymbol{x}, \boldsymbol{t})$ is identifiable from observed data, any quantities derived from it through these operations are also identifiable. □

## B Additional Simulation Results

### B.1 Additional Simulation Results under Homogeneous Treatment Effects

This appendix reports experimental results using three simulation datasets with constant single and interaction treatment effects to evaluate the causal effect estimation performance of the proposed method. The first scenario assumes that all true covariates are directly observable. The second scenario assumes that only latent covariates are observed. The third scenario represents a more realistic hybrid setting where true covariates and proxies for latent covariates coexist within the observed covariates, which combines the structures of the first and second scenarios. This allows us to test the generalizability of our method in complex settings.

In all scenarios, given the true covariates $\boldsymbol{x}_{i,\mathrm{t}}$, treatment assignment and outcome generation follow the same functional form, with the number of treatments fixed at $K = 3$ and the sample size set to $N = 50,000$.

$$t_i^{(k)} \sim \mathrm{Bern}\left(\sigma\left(\boldsymbol{w}_{t_k}^\top \boldsymbol{x}_{i,\mathrm{t}}\right) - \delta\right), \quad k = 1, 2, 3,$$
$$\boldsymbol{t_i} = t_i^{(1)}, t_i^{(2)}, t_i^{(3)}),$$
$$Y_i \sim \mathcal{N}\big(f(\boldsymbol{x}_{i,\mathrm{t}}, \boldsymbol{t}_i)\,, 1^2\big), \quad l \in \{0, 1\},$$

where $\sigma(x) = 1/(1 + \exp(-x))$, and $\delta$ is a bias parameter that varies across simulation datasets. The vector $\boldsymbol{w}_{t_k}$ is the weight vector whose elements are independently drawn from the uniform distribution $U(-1, 1)$. The outcome-generating function $f$ is defined as follows:

$$f(\boldsymbol{x}_{i,\mathrm{t}}, \boldsymbol{t_i}) = \boldsymbol{w}_x^\top \boldsymbol{x}_{i,\mathrm{t}} + t_i^{(1)} + 1.2t_i^{(2)} + 0.8t_i^{(3)} + t_i^{(1)}t_i^{(2)} - 0.5t_i^{(1)}t_i^{(3)} + 0.1t_i^{(2)}t_i^{(3)} + 0.7t_i^{(1)}t_i^{(2)}t_i^{(3)} + 2.$$

We vary the structure of the observed covariates $\boldsymbol{x}_{i,\mathrm{o}}$ and the true covariates $\boldsymbol{x}_{i,\mathrm{t}}$ as follows:

**Simulation dataset 1A**

$$x_{i,n}^{(j)} \sim N(c_n^{(j)}, 1^2), \quad x_{i,u}^{(j)} \sim U(-10, 10), \quad x_{i,b}^{(j)} \sim \mathrm{Bern}(p_b^{(j)}),$$
$$\boldsymbol{x}_{i,\mathrm{obs}} = \boldsymbol{x}_{i,\mathrm{true}} = (\boldsymbol{x}_{i,n}, \boldsymbol{x}_{i,u}, \boldsymbol{x}_{i,b}),$$

where $j \in \{1, \ldots, 10\}$ and $c_n^{(j)}$ are drawn from the uniform distribution $U(-1, 1)$. Similarly, each $p_b^{(j)}$ is sampled from $U(0, 1)$. In simulation dataset 1A, $\delta = 0.3$.

**Simulation dataset 2A**

$$\boldsymbol{z}_i \sim N(\boldsymbol{0}, I_{10}), \quad x_{i,n}^{(j)} \sim N(\boldsymbol{w}_n^{(j)\top} \boldsymbol{z}_i, 1^2),$$
$$x_{i,u}^{(j)} \sim N(\boldsymbol{w}_u^{(j)\top} \boldsymbol{z}_i, 5^2), \quad x_{i,b}^{(j)} \sim \mathrm{Bern}(\sigma(\boldsymbol{w}_b^{(j)\top} \boldsymbol{z}_i)),$$
$$\boldsymbol{x}_{i,\mathrm{o}} = (\boldsymbol{x}_{i,n}, \boldsymbol{x}_{i,u}, \boldsymbol{x}_{i,b}), \quad \boldsymbol{x}_{i,\mathrm{t}} = \boldsymbol{z}_i,$$

where $j \in \{1, \ldots, 10\}$ and the vectors $\boldsymbol{w}_n^{(j)}$, $\boldsymbol{w}_u^{(j)}$, and $\boldsymbol{w}_b^{(j)}$ are weight vectors whose elements are drawn independently from $U(-1, 1)$. $I_d$ denotes the identity matrix $d \times d$, which is also used as the variance-covariance matrix. In simulation dataset 2A, $\delta = 0.2$.

**Simulation dataset 3A**

$$\boldsymbol{z}_i \sim N(\boldsymbol{0}, I_{15}), \quad x_{i,p}^{(l)} \sim N(\boldsymbol{w}_p^{(l)\top} \boldsymbol{z}_i, 1^2),$$
$$x_{i,n}^{(j)} \sim N(c_n^{(j)}, 1^2), \quad x_{i,u}^{(j)} \sim U(-10, 10), \quad x_{i,b}^{(j)} \sim \mathrm{Bern}(p_b^{(j)}),$$
$$\boldsymbol{x}_{i,\mathrm{o}} = (\boldsymbol{x}_{i,n}, \boldsymbol{x}_{i,u}, \boldsymbol{x}_{i,b}, \boldsymbol{x}_{i,p}), \quad \boldsymbol{x}_{i,\mathrm{t}} = (\boldsymbol{x}_{i,n}, \boldsymbol{x}_{i,u}, \boldsymbol{x}_{i,b}, \boldsymbol{z}_i),$$

where $l \in \{1, \ldots, 15\}$ and $j \in \{1, \ldots, 5\}$. Vectors $\boldsymbol{w}_p^{(l)}$ are weight vectors whose elements are drawn independently from $U(-1, 1)$. In simulation dataset 3A, $\delta = 0.1$.

Table 5 shows the $\epsilon_{\mathrm{ASE}}$ and $\epsilon_{\mathrm{AIE}}$ of the proposed and baseline methods in three simulation datasets. In all datasets, the proposed method achieves the lowest estimation errors. These results demonstrate CISI-Net's ability to accurately estimate single treatment and interaction treatment effects even when no heterogeneity exists in their effects.

### B.2 Scalability to a Larger Number of Treatments

We investigate whether the proposed method maintains its performance advantage over existing methods designed for multiple treatment settings as the number of treatments increases. With $K$ treatments, the total number of possible treatment patterns grows exponentially as $2^K$, and the number of possible interaction

Table 5: Comparison of mean and standard deviation of $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$ across simulation datasets. Here, $k \in \{1, 2, 3\}$ indexes single treatments for ASE, and $S \subseteq \{1, 2, 3\}$ (with $|S| \geq 2$) denotes treatment combinations for AIE.

| | | $\epsilon_{\text{ASE}}$ | | | $\epsilon_{\text{AIE}}$ | | | |
|---|---|---|---|---|---|---|---|---|
| Sim. | Method | $k = 1$ | $k = 2$ | $k = 3$ | $S = \{1, 2\}$ | $S = \{2, 3\}$ | $S = \{1, 3\}$ | $S = \{1, 2, 3\}$ |
| 1A | TARNet | $0.19 \pm 0.12$ | $0.17 \pm 0.13$ | $0.16 \pm 0.13$ | $0.27 \pm 0.18$ | $0.25 \pm 0.19$ | $0.24 \pm 0.17$ | $0.41 \pm 0.31$ |
| | CFR-WASS | $0.19 \pm 0.13$ | $0.17 \pm 0.13$ | $0.17 \pm 0.13$ | $0.28 \pm 0.23$ | $0.26 \pm 0.23$ | $0.26 \pm 0.19$ | $0.47 \pm 0.38$ |
| | NCoRE | $0.10 \pm 0.08$ | $0.09 \pm 0.10$ | $0.09 \pm 0.08$ | $0.15 \pm 0.13$ | $0.14 \pm 0.11$ | $0.16 \pm 0.14$ | $0.20 \pm 0.15$ |
| | TECE-VAE | $0.43 \pm 0.10$ | $0.47 \pm 0.10$ | $0.37 \pm 0.07$ | $0.56 \pm 0.20$ | $0.08 \pm 0.07$ | $0.41 \pm 0.11$ | $0.73 \pm 0.10$ |
| | CISI-Net | $\mathbf{0.04 \pm 0.03}$ | $\mathbf{0.04 \pm 0.03}$ | $\mathbf{0.04 \pm 0.03}$ | $\mathbf{0.05 \pm 0.04}$ | $\mathbf{0.04 \pm 0.03}$ | $\mathbf{0.05 \pm 0.03}$ | $\mathbf{0.06 \pm 0.04}$ |
| 2A | TARNet | $0.05 \pm 0.04$ | $0.06 \pm 0.04$ | $0.05 \pm 0.04$ | $0.09 \pm 0.07$ | $0.08 \pm 0.07$ | $0.08 \pm 0.06$ | $0.14 \pm 0.11$ |
| | CFR-WASS | $0.07 \pm 0.06$ | $0.07 \pm 0.06$ | $0.07 \pm 0.05$ | $0.11 \pm 0.09$ | $0.10 \pm 0.07$ | $0.16 \pm 0.07$ | $0.17 \pm 0.13$ |
| | NCoRE | $0.05 \pm 0.04$ | $0.05 \pm 0.04$ | $0.04 \pm 0.03$ | $0.09 \pm 0.07$ | $0.08 \pm 0.06$ | $0.08 \pm 0.06$ | $0.12 \pm 0.11$ |
| | TECE-VAE | $0.09 \pm 0.06$ | $0.11 \pm 0.07$ | $0.09 \pm 0.06$ | $0.12 \pm 0.05$ | $0.11 \pm 0.04$ | $0.09 \pm 0.04$ | $0.13 \pm 0.09$ |
| | CISI-Net | $\mathbf{0.04 \pm 0.03}$ | $\mathbf{0.05 \pm 0.03}$ | $\mathbf{0.04 \pm .03}$ | $\mathbf{0.04 \pm 0.03}$ | $\mathbf{0.03 \pm 0.02}$ | $\mathbf{0.03 \pm 0.02}$ | $\mathbf{0.04 \pm 0.03}$ |
| 3A | TARNet | $0.35 \pm 0.26$ | $0.31 \pm 0.25$ | $0.32 \pm 0.26$ | $0.24 \pm 0.19$ | $0.24 \pm 0.19$ | $0.24 \pm 0.19$ | $0.36 \pm 0.33$ |
| | CFR-WASS | $0.29 \pm 0.23$ | $0.27 \pm 0.22$ | $0.29 \pm 0.22$ | $0.23 \pm 0.19$ | $0.24 \pm 0.19$ | $0.24 \pm 0.18$ | $0.40 \pm 0.32$ |
| | NCoRE | $0.38 \pm 0.29$ | $0.35 \pm 0.26$ | $0.30 \pm 0.26$ | $0.28 \pm 0.23$ | $0.24 \pm 0.18$ | $0.23 \pm 0.18$ | $0.36 \pm 0.33$ |
| | TECE-VAE | $0.36 \pm 0.18$ | $0.43 \pm 0.18$ | $0.31 \pm 0.19$ | $0.17 \pm 0.13$ | $0.13 \pm 0.09$ | $0.29 \pm 0.11$ | $0.58 \pm 0.15$ |
| | CISI-Net | $\mathbf{0.18 \pm 0.14}$ | $\mathbf{0.17 \pm 0.13}$ | $\mathbf{0.18 \pm 0.15}$ | $\mathbf{0.11 \pm 0.09}$ | $\mathbf{0.08 \pm 0.06}$ | $\mathbf{0.08 \pm 0.07}$ | $\mathbf{0.11 \pm 0.10}$ |

effects increases accordingly. Scenarios requiring the estimation of a large number of causal effects frequently arise in real-world applications (Parmar et al., 2008; Freidlin et al., 2008). Therefore, we evaluate the practical utility of each method by assessing its accuracy of estimation across varying numbers of treatments.

To examine the scalability of each method, we conduct experiments under varying numbers of treatments $K \in \{4, 5, 8, 10\}$. In practice, the number of simultaneous treatments rarely exceeds ten (Cho et al., 2022; Freidlin et al., 2008; Muralidharan et al., 2025; Tsuboi et al., 2024). Therefore, our simulation setting with an upper bound of $K \leq 10$ represents a practical range that adequately covers the vast majority of multiple-treatment scenarios. For all settings, treatment assignment and outcome generation follow the same functional form, and the sample size is fixed at $N = 50,000$. The treatment vector, covariates, and outcomes are based on the following data generation process:

$$
\begin{gathered}
x_{i,n}^{(j)} \sim N(c_n^{(j)}, 1^2), \quad x_{i,u}^{(j)} \sim U(-1, 1), \quad \boldsymbol{x}_i = (\boldsymbol{x}_{i,n}, \boldsymbol{x}_{i,u}), \\
H_i^{(v)} = \mathbb{I}\left( x_i^{(v)} + x_i^{(v+1)} > 1 \right), \quad \boldsymbol{H}_i = (H_i^{(1)}, H_i^{(2)}), \\
t_i^{(k)} \sim \text{Bern}\left( \sigma\left( \boldsymbol{w}_{t_k}^\top \boldsymbol{x}_i - \boldsymbol{w}_H^\top \boldsymbol{H}_i \right) \right), \quad \boldsymbol{t}_i = (t_i^{(1)}, \ldots, t_i^{(K)}), \\
Y_i \sim \mathcal{N}\big( g(\boldsymbol{x}_i, \boldsymbol{t}_i)\, , 1^2 \big)\, ,
\end{gathered}
$$

where $j \in \{1, \ldots, 15\}$ denotes the covariate index, $v \in \{1, 2\}$, and $k \in \{1, \ldots, K\}$. $\mathbb{I}(\cdot)$ is the indicator function, $\sigma(\cdot)$ is the sigmoid function defined as $\sigma(x) = 1/(1 + \exp(-x))$. The vectors $\boldsymbol{w}_{t_k}$ and $\boldsymbol{w}_H$ are the weight vectors whose elements are independently drawn from the uniform distribution $U(-1, 1)$. We set $\boldsymbol{x}_{i,\text{o}} = \boldsymbol{x}_{i,\text{t}} = (\boldsymbol{x}_{i,n}, \boldsymbol{x}_{i,u})$. The outcome generating function $g$ is defined as follows:

$$
g(\boldsymbol{x}_i, \boldsymbol{t}_i) = \boldsymbol{w}_x^\top \boldsymbol{x}_i + \sum_{\substack{\emptyset \neq S \subseteq \{1, \ldots, K\} \\ |S| \leq K}} \tau_S(\boldsymbol{x}_i) \prod_{k \in S} t_i^{(k)} + 2, \quad \tau_S(\boldsymbol{x}_i) = w_S \left( 1 + x_i^{(j_S)} \right),
$$

where $j_S \in \{1, \ldots, 30\}$ denotes a covariate index randomly assigned to each subset $S$. The vector $\boldsymbol{w}_x$ is a weight vector whose elements are independently drawn from the uniform distribution $U(-1, 1)$, and each

Table 6: Comparison of the number of trainable parameters across methods under different numbers of treatments $K$. M denotes $10^6$.

| $K$ | TECE-VAE | NCoRE | CISI-Net |
|---|---|---|---|
| 4 | 0.425M | 0.518M | 0.137M |
| 5 | 0.426M | 0.842M | 0.137M |
| 8 | 0.429M | 2.298M | 0.138M |
| 10 | 0.430M | 3.670M | 0.138M |

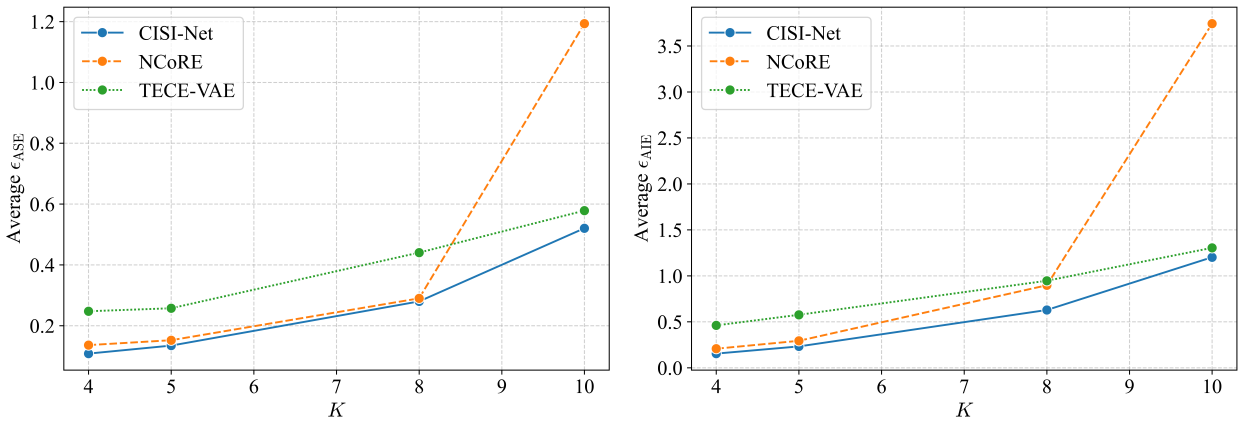



Figure 6: Average $\epsilon_{\text{ASE}}$ (left) and $\epsilon_{\text{AIE}}$ (right) under varying numbers of treatments $K \in \{4, 5, 8, 10\}$

$w_S$ is independently drawn from $U(-2, 2)$. This setup induces treatment assignment bias through $\boldsymbol{H}_i$ and heterogeneous treatment effects through $\tau_S(\boldsymbol{x}_i)$.

The hyperparameters specified in Section 5.2 are applied to CISI-Net and the multiple-treatment baseline methods NCoRE and TECE-VAE. Table 6 shows the number of trainable parameters for each method as a reference for model capacity.

Figure 6 shows the average $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$ across varying numbers of treatments $K$. Estimation errors are averaged over all single effects and interaction effects, respectively. CISI-Net consistently achieves lower estimation errors than the baseline methods for both ASE and AIE, and maintains its superior performance as $K$ increases. These results suggest that CISI-Net provides a stable and accurate estimate as the number of treatments increases, demonstrating its practical utility.

The baseline methods, NCoRE and TECE-VAE, underperform CISI-Net, suggesting limitations inherent in their respective architectures. Whereas NCoRE demonstrates competitive performance on both metrics for $K \leq 8$, its performance deteriorates markedly at $K = 10$. This behavior reflects a limitation of the NCoRE architecture, which constructs outcome and interaction networks independently and does not adequately promote parameter sharing across treatment patterns. As the number of treatment patterns increases, insufficient parameter sharing can result in an unstable estimation. Although TECE-VAE shows a more gradual increase in both $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$, its estimation error remains consistently higher than that of CISI-Net, especially for $K \leq 8$. Unlike NCoRE, TECE-VAE employs task embeddings and does not construct independent networks for each treatment pattern. As a result, performance degradation is mitigated as $K$ increases. Despite this advantage, reliance on latent-variable modeling under potential model misspecification may lead to unnecessary latent recovery, which can result in inferior estimation accuracy compared to CISI-Net.

Table 7: Additional results comparing CISI-Net with and without the balancing penalty (BP) on simulation datasets 2 and 3. Here, $k \in \{1, 2, 3\}$ indexes single treatments for ASE, and $S \subseteq \{1, 2, 3\}$ with $|S| \geq 2$ denotes treatment combinations for AIE. The balancing coefficient $\alpha$ is set to 0.1 when the balancing penalty is applied.

| | | $\epsilon_{\text{ASE}}$ | | | $\epsilon_{\text{AIE}}$ | | | |
|---|---|---|---|---|---|---|---|---|
| Sim. | Method | $k = 1$ | $k = 2$ | $k = 3$ | $S = \{1, 2\}$ | $S = \{2, 3\}$ | $S = \{1, 3\}$ | $S = \{1, 2, 3\}$ |
| 2 | CISI-Net w/o BP | 0.16 ± 0.13 | 0.20 ± 0.13 | 0.12 ± 0.10 | 0.27 ± 0.15 | 0.23 ± 0.14 | 0.19 ± 0.12 | 0.21 ± 0.13 |
| | CISI-Net | 0.16 ± 0.13 | 0.19 ± 0.13 | 0.12 ± 0.09 | 0.26 ± 0.18 | 0.23 ± 0.14 | 0.18 ± 0.12 | 0.21 ± 0.14 |
| 3 | CISI-Net w/o BP | 0.11 ± 0.07 | 0.11 ± 0.08 | 0.09 ± 0.07 | 0.18 ± 0.15 | 0.16 ± 0.11 | 0.17 ± 0.11 | 0.09 ± 0.06 |
| | CISI-Net | 0.10 ± 0.06 | 0.09 ± 0.08 | 0.08 ± 0.06 | 0.12 ± 0.09 | 0.12 ± 0.08 | 0.10 ± 0.08 | 0.09 ± 0.07 |

### B.3 Additional Ablation Results

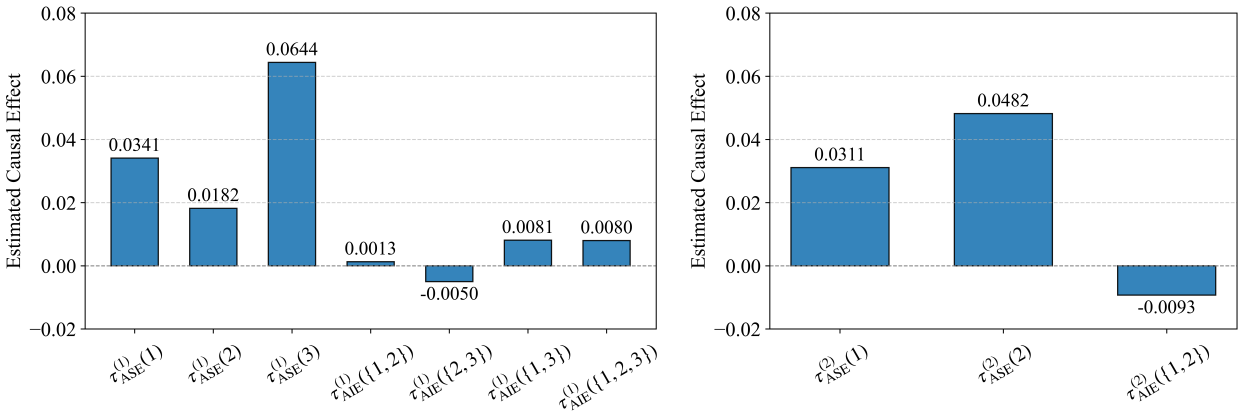
We conduct additional ablation studies using simulation datasets that differ from the one used in Section 6 to verify the dataset-independent utility of the balancing penalty in CISI-Net. Specifically, we use simulation datasets 2 and 3, which are defined in Section 5.1. To ensure robustness against randomness in data generation, we generate 100 independent datasets for each scenario using different random seeds and report the average $\epsilon_{\text{ASE}}$ and $\epsilon_{\text{AIE}}$ across these datasets for each model. Table 7 shows the results of the ablation study for each dataset.

Overall, the results demonstrate that applying the balancing penalty improves the accuracy of causal effect estimation across the two additional simulation datasets. In simulation dataset 2, the improvement by balancing penalty is relatively small. This may be because the observed covariates serve as proxies for latent covariates, which partially limits the benefits of balancing representation distributions due to the incomplete information contained in these proxies. Nevertheless, CISI-Net with the balancing penalty achieves equal or lower errors compared to CISI-Net without the balancing penalty across most metrics. These results suggest that the contribution of the balancing penalty is not limited to a specific dataset but supports stable causal effect estimation across different data-generating processes, regardless of the presence of interaction effects.

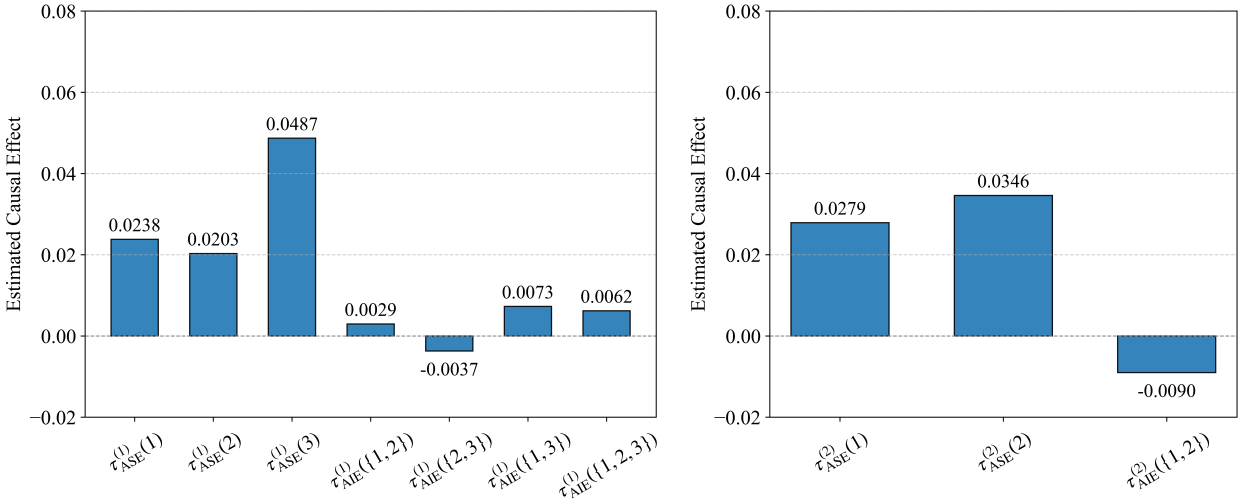
## C Sensitivity Analysis in Empirical Applications

This appendix reports how the estimated causal effects change when varying the balancing-penalty weight $\alpha$ in a real-world dataset derived from multiple marketing promotions. In real-world observational data, the data-generating process is unknown, and therefore, it is unclear whether the hyperparameter configuration that performs best in simulation settings is optimal for empirical analyses. Among the various hyperparameters in CISI-Net that control different components of the model, we focus on the balancing-penalty weight $\alpha$, which is a core component of the proposed method. By changing $\alpha$, we evaluate whether the qualitative conclusions of causal effect estimation depend excessively on a specific hyperparameter setting.

Figure 7 shows the estimated causal effects obtained when the balancing-penalty weight is set to $\alpha = 0.5$ and $\alpha = 1.0$. A comparison with the results reported in Figure 5 demonstrates that the magnitude of the estimated effects varies across these values. At the same time, the sign of all estimated effects remains consistent and does not reverse. This result demonstrates that the qualitative conclusions of the real-world analysis are stable across variations in the balancing-penalty weight.




(a) $\alpha = 0.5$




(b) $\alpha = 1.0$

Figure 7: Estimated single and interaction treatment effects obtained from two real-world marketing promotion datasets under different balancing-penalty weights. The top panel corresponds to $\alpha = 0.5$, and the bottom panel corresponds to $\alpha = 1.0$. In each panel, the left side shows results for dataset 1 ($\mathrm{CP}_1$–$\mathrm{CP}_3$), and the right side shows results for dataset 2 ($\mathrm{CP}_4$ and $\mathrm{CP}_5$). Here, $\tau^{(d)}_{\mathrm{ASE}}(k)$ and $\tau^{(d)}_{\mathrm{AIE}}(S)$ denote the estimated effects for single and multiple treatments in dataset $d$. The outcome is standardized prior to estimation.