CAUSAL STRUCTURE-GUIDED OPTIMAL TRANSPORT FOR TREATMENT EFFECT ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Treatment effect estimation seeks to evaluate the effect of a treatment based on non-randomized trial data. To reduce the confounding bias between the control and treated groups, existing methods usually align the distributions of the groups by learning balanced representations from covariates. Despite their success, these methods simply model the covariates as a flat vector while ignoring the intrinsic causal structure among variables. To address this limitation, we exploit the causal structure among covariates to learn balanced representations for causal effect estimation. Specifically, we incorporate the causal structure into the framework of optimal transport, which has demonstrated a powerful ability for distribution alignment. We first discover the causal structure and establish the structural equations of covariates, and introduce the residuals into optimal transport, where the residuals capture individual information for heterogeneous treatment effect estimation. We further leverage the causal structure to refine the cost function in optimal transport, deriving a causal structure-aware transport plan. Theoretically, we reveal that the estimation error of the effect can be bounded by the proposed causal structure-guided optimal transport cost. Empirically, we conduct experiments on synthetic and real-world datasets to demonstrate the effectiveness of our proposed method.

1 INTRODUCTION

Treatment effect estimation from observational data has been widely applied in various real-world scenarios (Athey & Imbens, 2016; Atan et al., 2018; Athey & Wager, 2019; Sato et al., 2020; Ding et al., 2022), including recommendation (Schnabel et al., 2016), healthcare (Bica et al., 2020), and advertising (Bottou et al., 2013). Under the Rubin-Neyman potential outcome framework (Rubin, 1974), one can train potential outcome prediction models on the control and treated groups, respectively, and then estimate the effect by comparing the predicted potential outcomes under treatment and no treatment (Künzel et al., 2019). One major issue is confounding bias arising from the different distributions of confounders in the two groups, resulting in limited generalization performance of the outcome prediction models (Melnychuk et al., 2024; Yan et al., 2025).

To mitigate confounding bias, existing methods typically reduce the distributional imbalance between the control and treated groups to obtain balanced representations (Johansson et al., 2016; Shalit et al., 2017), where the imbalance can be measured by the Maximum Mean Discrepancy (Gretton et al., 2012) or the Hilbert–Schmidt Independence Criterion (Gretton et al., 2005; 2007). Recently, the Wasserstein discrepancy derived from optimal transport has attracted increasing attention because of its efficacy for distribution alignment (Wang et al., 2023). By finding an optimal transport plan between two distributions that minimizes the total transport cost, optimal transport exploits the underlying geometric structure of the data to quantify the distribution imbalance, making it suitable for balanced representation learning in causal effect estimation (Peyré & Cuturi, 2019).

However, these optimal transport-based methods simply model covariates as unstructured vectors and overlook their intrinsic causal structure, which plays a crucial role in characterizing covariate dependencies and inducing reliable cross-group alignment. To address this limitation, we incorporate the causal structure of covariates into optimal transport to propose a balanced representation learning method, termed as Causal Structure-Guided Optimal Transport (CSOT). The structure is leveraged in the underlying transport cost, and the derived structure-guided Wasserstein discrep-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ancy is minimized to align the representation distributions of the control and treated groups, thereby improving the reduction of the confounding bias.

Specifically, we first employ causal discovery to recover the structure among covariates and establish a structural equation for each covariate. Based on the learned structural relationships, we extract empirical residuals that characterize individual-specific information unexplained by the structure, which is essential for heterogeneous treatment effect estimation. To better leverage such individual information, we incorporate the residuals into representation learning and the calculation of the underlying transport cost. Furthermore, we refine the transport cost according to the uncovered causal graph, which yields a causal structure-aware optimal transport plan used in balanced representation learning. We further prove that the proposed optimal transport discrepancy upper bounds the cross-group balancing error and the effect estimation error, which establishes a theoretical foundation for our proposed method. Extensive experiments on synthetic and real-world benchmark datasets demonstrate the effectiveness and robustness of the proposed method across different causal discovery methods.

We summarize the principal contributions as follows.

- We propose a causal structure-guided optimal transport model that incorporates causal structure among covariates into balanced representation learning for treatment effect estimation.
- We theoretically establish upper bounds on the cross-group balancing error and the effect estimation error in terms of our derived structure-guided optimal transport discrepancy.
- We conduct extensive experiments on synthetic and real-world benchmarks using multiple causal discovery methods to demonstrate the effectiveness and robustness of the proposed method.

2 Related Work

Causal Effect Estimation Non-random treatment assignment can induce systematic differences in the covariate distributions of the treated and control groups, thereby introducing confounding bias (Rubin, 1974; Rosenbaum & Rubin, 1983). Existing approaches mainly rely on the propensity score to reduce the bias (Robins et al., 2000; Bang & Robins, 2005), or learn balanced representations to address imbalance by aligning the distributions of the treated and control groups in the embedding space (Shalit et al., 2017; Yoon et al., 2018; Shi et al., 2019; Melnychuk et al., 2024; Nagalapatti et al., 2024). Several representative studies further improve balancing by controlling global distributional discrepancy or preserving local sample structure. Shalit et al. (2017) learned balanced representations by minimizing the discrepancy between treated and control distributions. Yao et al. (2018) proposed SITE to balance treatment distributions while preserving local similarity among samples.

Recently, optimal transport has been applied in balanced representation learning to measure the distribution imbalance of the control and treated groups (Li et al., 2021; Wang et al., 2023; Yan et al., 2025). Optimal transport studies how to transport masses from one distribution to another distribution with a minimized cost (Monge, 1781; Kantorovitch, 1958; Villani, 2021), which can be leveraged to measure the distribution imbalance of the control and treated groups. Wang et al. (2023) develop an enhanced optimal transport discrepancy and incorporate factual-outcome information to mitigate mini-batch distribution estimation bias. More recently, Yan et al. (2025) align treatment-conditional distributions with the overall covariate distribution using transport costs defined in the learned representation space, thereby reducing confounding bias without data splitting. Despite these advances, most existing methods treat covariates as an unstructured feature vector and do not explicitly model the structural dependencies among covariates. Unlike these methods, our method incorporates causal structure into the transport cost, enabling cross-group alignment to account for both representation similarity and structural information.

Causal Discovery Causal discovery aims to recover the causal structure among variables from observational data. Existing causal discovery methods can be broadly categorized into constraint-based approaches (Spirtes et al., 2000), functional causal models (Shimizu et al., 2006; 2011), and continuous optimization methods (Zheng et al., 2018; Ng et al., 2020). Spirtes et al. (2000) aim to

represent a classical constraint-based approach that removes and directs edges through conditional independence tests under the causal Markov and faithfulness assumptions. GOLEM subsequently adopts a likelihood-based objective with soft sparsity and acyclicity regularization for scalable linear DAG learning (Ng et al., 2020). In this study, we apply causal discovery methods to obtain structure among covariates as auxiliary information, and then introduce the structural information into the potential outcome framework, which is implemented by designing a structure-guided transport cost to learn balanced representations.

3 PRELIMINARY

We adopt the Rubin-Neyman potential outcome framework (Rubin, 1974; Rosenbaum & Rubin, 1983). Consider an observational dataset $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^{N}$, where (x_i, t_i, y_i) is an observed realization of (X, T, Y) for the i-th unit and N denotes the number of units. $X \in \mathbb{R}^d$ is a d-dimensional covariate vector, $T \in \{0, 1\}$ is a binary treatment variable, and $Y \in \mathbb{R}$ is the outcome variable. The observed outcome Y is the potential outcome $Y(t)$ corresponding to the actually received treatment t.

For a given covariate values $X = x_i$, our objective is to estimate the individual treatment effect (ITE) as follows:

$$\tau_i = y_i(1) - y_i(0). \tag{1}$$

The specific definitions of each assumption are provided in the Appendix A.

The causal structure among these covariates is represented by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{X_1, \ldots, X_d\}$ is the node set, and $\mathcal{E}$ is the edge set, where d denotes the number of covariates. A directed edge $X_k \to X_j$ indicates that X_k is a direct parent of X_j.

Let $A \in \mathbb{R}^{d \times d}$ denote the weighted adjacency matrix of $\mathcal{G}$, where $A_{kj} \neq 0$ indicates the existence of the directed edge $X_k \to X_j$. The parent set of X_j is defined as

$$\mathrm{Pa}(j) = \{k : A_{kj} \neq 0\}. \tag{2}$$

4 METHODOLOGY

Figure 1 illustrates the overview of our proposed method CSOT. We first employ a causal discovery method to learn structural dependencies among covariates, and then establish structural equations to extract regression residuals for capturing individual information. The residuals are not only leveraged to augment inputs but also incorporated into the optimal transport model for balanced representation learning. Moreover, the discovered causal structure is also exploited to guide the transport cost according to the importance of the features. We describe the technical details in the following.

4.1 RESIDUAL-AWARE REPRESENTATION LEARNING

Existing counterfactual representation learning methods typically regard covariates as a vector without considering structural dependencies among covariates. To address this, we discover the causal structure from data and leverage it to learn an augmented representation. Specifically, under an additive structural equation model, each covariate is generated according to

$$X_j = f_j\left(X_{\mathrm{Pa}(j)}\right) + \varepsilon_j, \qquad j = 1, \ldots, d, \tag{3}$$

where $f_j(\cdot)$ characterizes the structural relationship between X_j and its parents, and ε_j captures the individual information that cannot be explained by the structural relationship (Hoyer et al., 2008; Peters et al., 2014). Since the true causal structure among covariates is unknown, we apply a causal discovery method to the observed covariate matrix X to estimate its weighted adjacency matrix $\widehat{A}$ of the causal graph. Following (Yu et al., 2019; Bello et al., 2022), given a threshold δ, the estimated parent set of X_j is defined as

$$\widehat{\mathrm{Pa}}(j) = \left\{k : \left|\widehat{A}_{kj}\right| > \delta\right\}. \tag{4}$$

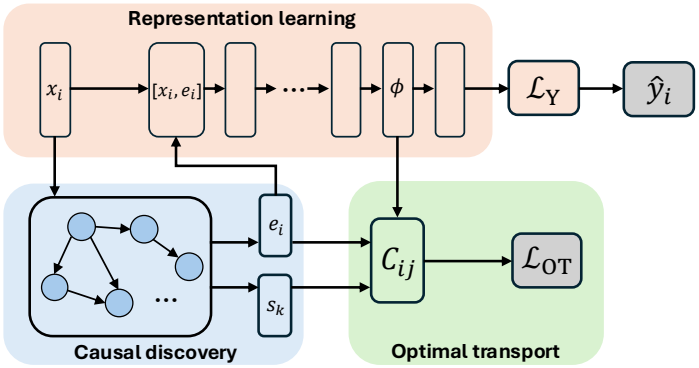


Figure 1: Overview of the proposed CSOT method.

Based on this sparsified causal graph, for the covariate j of the unit i, the function $f_j(\cdot)$ in Eq. (3) is implemented by

$$\widehat{f}_j\left(x_{i,\widehat{\mathrm{Pa}}(j)}\right) = x_{i,\widehat{\mathrm{Pa}}(j)}^{\top}\widehat{w}_j, \tag{5}$$

where $x_{i,\widehat{\mathrm{Pa}}(j)}$ are the parents of the covariate j of the unit i, and the parameters $\widehat{w}_j$ are obtained by solving the following ridge regression problem

$$\widehat{w}_j = \arg\min_{w} \sum_{i \in \mathcal{I}_{\mathrm{tr}}} \left(x_{ij} - x_{i,\widehat{\mathrm{Pa}}(j)}^{\top} w\right)^2 + \lambda_f \left\|w\right\|_2^2, \tag{6}$$

in which x_{ij} is the covariate j of the unit i, $\mathcal{I}_{\mathrm{tr}}$ denotes the set of training samples, and λ_f is the regularization coefficient. Since all covariates are standardized using statistics computed from the training set to obtain zero mean, the structural regression is fitted without an intercept.

After obtaining the estimated structural equation model, the empirical residual of the covariate j in the unit i is calculated by the following

$$e_{ij} = x_{ij} - \widehat{f}_j\left(x_{i,\widehat{\mathrm{Pa}}(j)}\right). \tag{7}$$

The residual e_{ij} captures the individual variation of the unit i in the j-th covariate that cannot be explained by its estimated parent variables. By aggregating the residuals across all covariates, we obtain the structural residual vector of the i-th sample:

$$e_i = [e_{i1}, e_{i2}, \ldots, e_{id}]. \tag{8}$$

In this formulation, the original covariate vector x_i characterizes the overall observed state of the sample, whereas the structural residual vector e_i captures its conditional deviations from the learned local structural relationships. Based on these two complementary views, we construct the structure-augmented input as

$$u_i = [x_i, e_i], \tag{9}$$

and obtain the residual-aware representation through an encoder:

$$z_i = \Phi(u_i), \tag{10}$$

where $\Phi(\cdot)$ is a network. This design introduces a structure-aware inductive bias by explicitly incorporating deviations unexplained by the learned local relationships. The joint input of covariates and residuals provides complementary observed and structure-conditioned information for subsequent cross-group alignment.

4.2 Structure-Guided Optimal Transport

Due to the distribution imbalance between the control and treated groups caused by the confounders, the potential outcome prediction model suffers from limited generalization ability (Johansson et al.,

2016; Shalit et al., 2017). To alleviate this, existing methods usually learn balanced representations by minimizing the Wasserstein discrepancy derived from optimal transport (Wang et al., 2023; Yan et al., 2025), while the causal structural information is omitted. Based on the above discovered causal structure among covariates, we propose a structure-guided optimal transport model to enhance balanced representation learning.

Optimal transport seeks to find an optimal coupling between two probability distributions with a minimal transport cost (Villani et al., 2009; Peyré & Cuturi, 2019). Formally, let $\mathcal{X}$ be the feature space, $\alpha, \beta \in \mathcal{M}(\mathcal{X})$ be two probability measures, and $\Pi(\alpha, \beta)$ be the set of the joint distributions. Given a nonnegative ground cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, optimal transport derives the Wasserstein discrepancy as follows

$$\mathcal{W}(c, \alpha, \beta) = \inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} c(a, b)\, \mathrm{d}\pi(a, b). \tag{11}$$

The Wasserstein discrepancy measures the imbalance degree between two distributions, which can be leveraged to learn balanced representations for causal effect estimation (Yan et al., 2024; Wang et al., 2025).

In practice, the treated and control distributions are unknown and only finite observational samples are available. Therefore, we establish optimal transport on empirical distributions of the groups. Let $\{x_{1,i}\}_{i=1}^{n_1}$ and $\{x_{0,j}\}_{j=1}^{n_0}$ denote the treated and control samples, where n_1 and n_0 denote the numbers of treated and control units, respectively. Their empirical distributions are defined as

$$\widehat{P}_1 = \sum_{i=1}^{n_1} p_{1,i}\delta_{x_{1,i}}, \qquad \widehat{P}_0 = \sum_{j=1}^{n_0} p_{0,j}\delta_{x_{0,j}}, \tag{12}$$

where $\delta_{x_{1,i}}$ and $\delta_{x_{0,j}}$ denote the Dirac measures concentrated at the treated unit $x_{1,i}$ and the control unit $x_{0,j}$, respectively. Moreover, $p_{1,i}$ and $p_{0,j}$ are their empirical probability weights, satisfying $\sum_{i=1}^{n_1} p_{1,i} = 1$ and $\sum_{j=1}^{n_0} p_{0,j} = 1$. The transport plan between the two groups is represented by a non-negative matrix π, where π_{ij} denotes the amount of mass transported from the treated sample $x_{1,i}$ to the control sample $x_{0,j}$. The feasible transport plans are constrained by

$$\Pi(p_1, p_0) = \left\{\pi \geq 0 : \pi\mathbf{1} = p_1,\ \pi^\top\mathbf{1} = p_0\right\}, \tag{13}$$

where p_1 and p_0 are the probability weight vectors of the treated and control groups.

Given the treated unit i and control unit j, and their representations $z_{1,i} = \Phi([x_{1,i}, e_{1,i}])$ and $z_{0,j} = \Phi([x_{0,j}, e_{0,j}])$, classic entropic optimal transport finds the optimal plan by solving the following problem.

$$\min_{\pi \in \Pi(\hat{P}_1, \hat{P}_0)} \sum_{i,j} \pi_{ij} \|z_{1,i} - z_{0,j}\|_2^2 + \gamma\Omega(\pi), \tag{14}$$

where $\Omega(\pi) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_0} \pi_{ij} \log \pi_{ij}$ denotes the negative entropy and $\gamma > 0$ denotes the entropic regularization coefficient.

To incorporate causal structural information into optimal transport, we introduce structural residual discrepancies between the treated unit i and the control unit j into the transport cost by calculating the distance $\sum_{k=1}^{d}(e_{1,ik} - e_{0,jk})^2$. Furthermore, considering that different variables exhibit heterogeneous structural roles in the learned causal graph, we derive a residual weight s_k to adaptively aggregate residual discrepancies across covariates. Formally, let $B = \left|\widehat{A}_X\right|$ be the absolute value of the adjacency matrix. The multi-hop structural influence of covariate X_k is defined as

$$\widetilde{s}_k = \sum_{j=1}^{d}\sum_{h=1}^{H} \left(B^h\right)_{kj}, \tag{15}$$

where H is the maximum propagation depth. We set $H = 2$ in all experiments for simplicity and efficacy, and calculate the influence coefficient as

$$\widetilde{s}_k = \sum_{j=1}^{d} \left[B_{kj} + \left(B^2\right)_{kj}\right]. \tag{16}$$

Here, B_{kj} measures the direct structural influence of X_k on X_j, whereas $(B^2)_{kj}$ aggregates its two-hop influence through intermediate covariates. The contribution of each directed path is determined by the product of the absolute edge weights along that path.

After that, we normalize the coefficient by the following

$$s_k = \frac{\widetilde{s}_k}{\frac{1}{d}\sum_{\ell=1}^{d} \widetilde{s}_\ell + \varepsilon}, \tag{17}$$

where ℓ indexes all covariates, and $\varepsilon > 0$ is a small constant introduced to avoid a zero denominator and ensure numerical stability. This normalization makes the average structural weight approximately equal to one, thereby preventing the weighting procedure from arbitrarily changing the scale of the residual transport cost.

Based on the above influence coefficients and the residuals from the causal structure, we refine the transport cost to propose a causal structure-guided optimal transport model for balanced representation learning. Specifically, for the i-th treated unit and j-th control unit, their learned representations and structural residuals are denoted by $(z_{1,i}, e_{1,i})$ and $(z_{0,j}, e_{0,j})$. We define the structure-guided transport cost between the two units as the following squared distance

$$C_{ij}^2 = \|z_{1,i} - z_{0,j}\|_2^2 + \lambda_e \sum_{k=1}^{d} s_k \left(e_{1,ik} - e_{0,jk}\right)^2, \tag{18}$$

where λ_e is the trade-off hyperparameter. And the structure-guided balancing loss is defined as

$$\mathcal{L}_{\mathrm{OT}} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_0} \pi_{ij}^{\star} C_{ij}^2 \tag{19}$$

$$\text{s.t.} \quad \pi^{\star} = \arg\min_{\pi \in \Pi(\widehat{P}_1, \widehat{P}_0)} \sum_{i=1}^{n_1}\sum_{j=1}^{n_0} \pi_{ij} C_{ij}^2 + \gamma\Omega(\pi). \tag{20}$$

Such a structure-guided optimal transport model takes the distance between representations and residuals into consideration, and the influence coefficients from the causal structure are leveraged to guide the transport. Consequently, the proposed mechanism incorporates causal structural information into representation balancing, improves the comparability between treated and control units, and provides a more reliable balanced representation for subsequent counterfactual outcome prediction.

4.3 Outcome Prediction and Overall Objective

Given the residual-aware representation z_i, we employ two treatment-specific outcome prediction models $h_0(\cdot)$ and $h_1(\cdot)$ to estimate the conditional potential outcomes under the control and treated conditions, and then minimize the following factual loss function by comparing the factual outcomes and the predictions:

$$\mathcal{L}_Y = \frac{1}{N}\sum_{i=1}^{N} \left[y_i - t_i h_1(z_i) - (1 - t_i) h_0(z_i)\right]^2. \tag{21}$$

To learn balanced representations for potential outcome prediction, we combine the factual prediction loss with the causal structure-guided optimal transport loss by the following

$$\mathcal{L} = \mathcal{L}_Y + \lambda_{\mathrm{OT}} \mathcal{L}_{\mathrm{OT}}. \tag{22}$$

This joint objective encourages the learned representation to retain outcome-predictive information while reducing the structure-aware distributional discrepancy between the treated and control groups.

After training, the individual treatment effect is estimated as

$$\widehat{\tau}(x_i) = h_1(z_i) - h_0(z_i). \tag{23}$$

4.4 Theoretical Analysis

Let P_1 and P_0 denote the covariate distributions of the treated and control groups, respectively. For any integrable measurable function $m : \mathcal{X} \to \mathbb{R}$, we quantify the cross-group imbalance induced by m as

$$\begin{aligned}\xi(m) &= |\mathbb{E}_{X\sim P_1}[m(X)] - \mathbb{E}_{X\sim P_0}[m(X)]| \\ &= \left| \int_{\mathcal{X}} m(x)\,\mathrm{d}P_1(x) - \int_{\mathcal{X}} m(x)\,\mathrm{d}P_0(x) \right|.\end{aligned} \tag{24}$$

The following result shows that this cross-group imbalance is controlled by the proposed causal structure-guided Wasserstein discrepancy.

Theorem 1 (Causal Structure-Guided Balancing Error Bound)**.** *Let $C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ denote the causal structure-guided ground cost. Assume that m satisfies $|m(x) - m(x')| \le C(x, x'), \forall x, x' \in \mathcal{X}$.We have the following result*

$$\xi(m) \le \mathcal{W}(C, P_1, P_0). \tag{25}$$

Theorem 1 shows that the proposed structure-aware transport discrepancy provides an upper bound on the cross-group balancing error.

Next, we analyze the individual effect estimation error in terms of the structure-guided optimal transport. For $t, t' \in \{0, 1\}$, define the outcome estimation error under treatment condition t, evaluated over the covariate distribution of group t', as

$$\mathcal{E}_t^{(t')} = \int_{\mathcal{X}} \left(h_t(x) - g_t(x)\right)^2 p(x \mid T = t')\mathrm{d}x, \tag{26}$$

where h_t and g_t denote the estimated and true potential outcome functions. Based on this definition, we derive the following bound that connects the factual outcome estimation errors and cross-group discrepancy to the individual treatment effect estimation error.

Theorem 2 (Individual Treatment Effect Estimation Error Bound)**.** *The individual treatment effect estimation error is jointly controlled by the factual outcome estimation errors and the causal structure-guided Wasserstein discrepancy:*

$$\epsilon_{\mathrm{PEHE}} \le \mathcal{E}_1^{(1)} + \mathcal{E}_0^{(0)} + \mathcal{O}\left(\mathcal{W}(C, P_1, P_0)\right). \tag{27}$$

Theorem 2 shows that the PEHE is controlled by the factual outcome prediction errors and the proposed causal structure-guided transport discrepancy. Here, $\mathcal{O}(\mathcal{W}(C, P_1, P_0))$ denotes a term bounded by a constant multiple of $\mathcal{W}(C, P_1, P_0)$.

In practice, the population covariate distributions are unavailable and can only be approximated using finite empirical samples. We therefore establish a high-probability bound connecting the population and empirical causal structure-guided transport discrepancies, and derive the resulting finite-sample PEHE bound.

Theorem 3 (Finite-Sample PEHE Bound)**.** *For any $\delta \in (0, 1)$, with probability at least $1 - 2\delta$, we have the following result*

$$\epsilon_{\mathrm{PEHE}} \le \mathcal{E}_1^{(1)} + \mathcal{E}_0^{(0)} + \mathcal{O}(\mathcal{W}(C, \hat{P}_1, \hat{P}_0)) + \mathcal{O}\left(1/\sqrt{\delta n}\right). \tag{28}$$

where ϵ_{PEHE} denotes the mean squared error in estimating individual treatment effects. Theorem 3 connects the population treatment effect estimation error to the structure-aware optimal transport discrepancy computed from finite samples. The additional term depends on both the empirical transport loss and the discrepancy between the empirical and population transport costs. As the treated and control sample sizes increase, the concentration term vanishes, and the empirical discrepancy approaches its population counterpart.

The proofs of all theoretical results are provided in Appendix B.

5 Experiments

5.1 Experimental Setup

We compare CSOT with representative treatment effect estimation methods, including traditional methods: OLS, KNN, PSM (Rosenbaum & Rubin, 1983), T-Learner (Künzel et al., 2019),

Table 1: Results on IHDP and ACIC 2016 in terms of $\sqrt{\epsilon_{\text{PEHE}}}$. Results are reported as mean $\pm$ standard deviation. Lower values indicate better performance. The best and second-best results in each column are shown in bold and underlined, respectively.

Method	IHDP		ACIC 2016	
	Within	Out	Within	Out
OLS	3.692 ± 6.894	3.739 ± 4.718	2.360 ± 1.447	2.211 ± 1.398
KNN	2.873 ± 3.634	3.728 ± 5.270	4.098 ± 0.718	4.190 ± 0.982
PSM	5.312 ± 7.705	5.534 ± 8.553	3.611 ± 1.748	3.436 ± 1.825
T-Learner	10.786 ± 10.769	10.791 ± 10.851	2.683 ± 1.095	2.490 ± 1.098
BART	3.171 ± 4.873	3.606 ± 6.023	2.133 ± 0.572	2.023 ± 0.537
MLP	2.257 ± 2.637	2.369 ± 2.962	2.627 ± 0.831	2.538 ± 0.832
BNN	5.548 ± 8.453	5.660 ± 8.967	3.990 ± 1.610	3.799 ± 1.743
CFR-Wass	0.761 ± 0.045	0.846 ± 0.074	1.982 ± 1.427	1.819 ± 1.398
GANITE	2.724 ± 0.188	2.819 ± 0.239	3.977 ± 0.498	3.663 ± 0.540
SITE	0.816 ± 0.044	1.076 ± 0.076	3.658 ± 0.607	1.426 ± 0.227
DragonNet	7.952 ± 7.589	8.035 ± 8.096	5.316 ± 1.333	5.156 ± 1.533
DKLite	5.510 ± 8.079	5.677 ± 8.875	3.679 ± 1.777	3.468 ± 1.841
DESCN	7.555 ± 9.005	7.585 ± 9.236	3.770 ± 1.484	3.577 ± 1.579
CausalOT	7.121 ± 7.746	7.218 ± 8.301	4.647 ± 1.970	4.460 ± 2.139
CFR-Pro	0.632 ± 0.062	0.759 ± 0.099	1.781 ± 0.416	1.586 ± 0.391
CE-RCFR	1.152 ± 0.158	1.114 ± 0.143	1.752 ± 0.375	1.442 ± 0.240
ESCFR	1.356 ± 1.885	1.531 ± 2.483	1.552 ± 0.990	1.362 ± 0.823
CSOT-Direct	$\underline{0.556 \pm 0.043}$	$\underline{0.614 \pm 0.060}$	$\mathbf{1.277 \pm 0.284}$	$\mathbf{0.990 \pm 0.172}$
CSOT-ICA	0.570 ± 0.045	0.627 ± 0.068	$\underline{1.295 \pm 0.288}$	$\underline{1.031 \pm 0.179}$
CSOT-NOTEARS	0.573 ± 0.043	0.620 ± 0.063	1.352 ± 0.306	1.034 ± 0.176
CSOT-PC	0.558 ± 0.049	0.624 ± 0.069	1.366 ± 0.291	1.094 ± 0.181
CSOT-GOLEM	$\mathbf{0.552 \pm 0.043}$	$\mathbf{0.611 \pm 0.060}$	1.340 ± 0.272	1.091 ± 0.180

Table 2: Out-of-sample results on the synthetic dataset under different confounding levels α in terms of $\sqrt{\epsilon_{\text{PEHE}}}$. Results are reported as mean $\pm$ standard deviation. Lower values are better. The best result is shown in bold, and the second-best result is underlined.

Method	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1.0$
OLS	0.3644 ± 0.0159	0.3552 ± 0.0126	0.3495 ± 0.0129	0.3622 ± 0.0141	0.3582 ± 0.0075
KNN	0.6476 ± 0.0142	0.6439 ± 0.0083	0.6445 ± 0.0176	0.6511 ± 0.0152	0.6490 ± 0.0142
PSM	0.4019 ± 0.0250	0.3965 ± 0.0172	0.3865 ± 0.0184	0.3962 ± 0.0153	0.3912 ± 0.0121
T-Learner	0.6011 ± 0.0284	0.6042 ± 0.0312	0.6086 ± 0.0292	0.6050 ± 0.0302	0.6348 ± 0.0544
BART	0.3938 ± 0.0150	0.4012 ± 0.0159	0.3963 ± 0.0150	0.4202 ± 0.0186	0.4215 ± 0.0108
MLP	0.3871 ± 0.0071	0.3951 ± 0.0110	0.3812 ± 0.0142	0.3884 ± 0.0182	0.3865 ± 0.0158
BNN	0.2277 ± 0.0151	0.2317 ± 0.0133	0.2382 ± 0.0137	0.2532 ± 0.0113	0.2596 ± 0.0137
CFR-Wass	0.3418 ± 0.0058	0.3387 ± 0.0045	0.3374 ± 0.0038	0.3393 ± 0.0035	0.3407 ± 0.0032
GANITE	0.7005 ± 0.0445	0.6117 ± 0.0335	0.7816 ± 0.0559	0.6547 ± 0.0545	0.8099 ± 0.0485
SITE	0.2621 ± 0.0066	0.2613 ± 0.0083	0.2746 ± 0.0163	0.2638 ± 0.0120	0.2843 ± 0.0173
DragonNet	2.0530 ± 0.0085	2.0563 ± 0.0131	2.0535 ± 0.0125	2.0535 ± 0.0110	2.0494 ± 0.0099
DKLite	0.3682 ± 0.0101	0.3752 ± 0.0113	0.3876 ± 0.0112	0.4046 ± 0.0123	0.4276 ± 0.0110
DESCN	0.3746 ± 0.0153	0.3784 ± 0.0176	0.3888 ± 0.0219	0.3858 ± 0.0163	0.4206 ± 0.0215
CausalOT	1.0290 ± 0.0076	1.0270 ± 0.0115	1.0192 ± 0.0115	1.0173 ± 0.0091	1.0109 ± 0.0079
CFR-Pro	0.1206 ± 0.0020	0.1252 ± 0.0032	0.1237 ± 0.0030	0.1336 ± 0.0073	0.1336 ± 0.0061
CE-RCFR	0.1388 ± 0.0058	0.1512 ± 0.0102	0.1451 ± 0.0062	0.1412 ± 0.0095	0.1483 ± 0.0059
ESCFR	0.1235 ± 0.0105	0.1220 ± 0.0094	0.1234 ± 0.0101	0.1222 ± 0.0052	0.1277 ± 0.0109
CSOT-Direct	0.1024 ± 0.0022	0.1055 ± 0.0026	0.1063 ± 0.0027	0.1073 ± 0.0028	0.1107 ± 0.0038
CSOT-ICA	$\underline{0.1022 \pm 0.0018}$	0.1038 ± 0.0022	0.1058 ± 0.0029	$\underline{0.1067 \pm 0.0031}$	$\underline{0.1091 \pm 0.0037}$
CSOT-NOTEARS	$\mathbf{0.1008 \pm 0.0016}$	$\mathbf{0.1023 \pm 0.0024}$	$\mathbf{0.1045 \pm 0.0029}$	$\mathbf{0.1054 \pm 0.0030}$	0.1106 ± 0.0036
CSOT-PC	0.1024 ± 0.0018	0.1047 ± 0.0023	0.1056 ± 0.0025	0.1067 ± 0.0033	$\mathbf{0.1083 \pm 0.0032}$
CSOT-GOLEM	0.1026 ± 0.0016	$\underline{0.1037 \pm 0.0022}$	$\underline{0.1049 \pm 0.0028}$	0.1089 ± 0.0043	0.1097 ± 0.0038

BART (Chipman et al., 2010), and MLP; representation-based methods BNN (Johansson et al., 2016), CFR-Wass (Shalit et al., 2017), GANITE (Yoon et al., 2018), SITE (Yao et al., 2018), DragonNet (Shi et al., 2019), DKLite (Zhang et al., 2020), and DESCN (Zhong et al., 2022); and recent counterfactual and optimal transport-based methods CausalOT (Li et al., 2021), CFR-Pro (Wang et al., 2025), CE-RCFR (Wang et al., 2024) and ESCFR (Wang et al., 2023). Detailed descriptions and implementation settings are provided in Appendix C.

Evaluation Metric. To evaluate the model's ability to estimate heterogeneous individual treatment effects, we adopt the Precision in Estimation of Heterogeneous Effect (PEHE) as the primary evaluation metric ϵ_{PEHE}, defined as $\epsilon_{\text{PEHE}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i - \tau_i)^2$, where $\hat{\tau}_i = \hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)$ denotes the predicted individual treatment effect and $\tau_i = \mu_1(x_i) - \mu_0(x_i)$ denotes the ground-truth individual treatment effect.

Datasets. We evaluate CSOT on three benchmark datasets. The Infant Health and Development Program (IHDP) dataset (Hill, 2011) examines the effect of specialized childcare programs on infant development and includes 747 subjects with 25 covariates. The 2016 Atlantic Causal Inference Conference (ACIC 2016) dataset (Dorie et al., 2019) contains 4,802 samples with 58 covariates. It provides a challenging setting with complex covariate relationships and heterogeneous treatment effects. The synthetic dataset consists of 5,000 samples with 30 covariates and is designed to assess treatment effect estimation under varying degrees of confounding. Specifically, the parameter $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ controls the strength of the treatment-selection mechanism, with larger values inducing greater covariate imbalance between the treated and control groups. Detailed data-generation procedures and experimental protocols are provided in Appendix C.

5.2 Results and Discussions

Tables 1 and 2 report the performance of CSOT and the baseline methods on the real-world datasets and synthetic datasets. The IHDP results are averaged over 100 realizations, whereas the ACIC 2016 and synthetic results are each averaged over 10 realizations. We employ five causal structure discovery methods configured under linear assumptions, namely DirectLiNGAM (Shimizu et al., 2011), ICA-LiNGAM (Shimizu et al., 2006), NOTEARS (Zheng et al., 2018), PC (Spirtes et al., 2000), and GOLEM (Ng et al., 2020). The corresponding variants are denoted as CSOT-Direct, CSOT-ICA, CSOT-NOTEARS, CSOT-PC, and CSOT-GOLEM, respectively, and are evaluated under the same experimental setting. The results show that incorporating structural residual information into representation learning and distribution balancing provides a more fine-grained criterion for aligning treated and control units, thereby improving counterfactual outcome prediction and individual treatment effect estimation. On the synthetic data, as α increases, covariate imbalance becomes more severe. CSOT maintains competitive performance because its causal structure-guided optimal transport enables more reliable cross-group alignment. Although different causal discovery methods may produce different graphs, our method consistently achieves strong performance, indicating its robustness to variations in the discovered causal structure.

Compared with traditional methods such as KNN, PSM, and BART, which generally have limited nonlinear representation capabilities, representation-based methods achieve better individual treatment effect estimation performance on several datasets. However, methods such as CFR-Wass, SITE, and DKLite primarily rely on distribution alignment in the latent representation space and generally overlook the structural dependencies among covariates. CSOT further learns a structural dependency graph over the covariates and extracts structural residuals, enabling the learned representations to preserve both the semantic and structural characteristics of the observational data. Compared with optimal transport-based balancing methods such as CausalOT and ESCFR, which define the transport cost solely based on distances in the learned representation space without exploiting structural information beyond feature embeddings, CSOT explicitly incorporates structural residual distances into the optimal transport cost, leading to more structure-aware cross-group alignment and more reliable distribution balancing.

6 Conclusion

In this paper, we incorporate causal structure among covariates into optimal transport for treatment effect estimation. The regression residuals of the structural equations are leveraged to learn comprehensive representations, and the structure is exploited to guide optimal transport for reducing the balancing error between the representations of the control and treated groups. We theoretically analyze the generalization bound of our method, and conduct experiments on synthetic and real-world datasets to demonstrate the effectiveness of our method. We not only propose a treatment effect estimation method, but also explore how to introduce causal structure into the potential outcome framework to estimate causal effect.

AI USE STATEMENT

In this work, we used generative AI tools for literature search and manuscript polishing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Onur Atan, James Jordon, and Mihaela Van der Schaar. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2):37–51, 2019.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.

Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International conference on machine learning*, pp. 884–895. PMLR, 2020.

François Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *arXiv preprint math/0503123*, 2007.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14 (11), 2013.

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. 2010.

Sihao Ding, Peng Wu, Fuli Feng, Yitong Wang, Xiangnan He, Yong Liao, and Yongdong Zhang. Addressing unmeasured confounder for recommendation with sensitivity analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 305–315, 2022.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

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

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

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

Leonid Kantorovitch. On the translocation of masses. *Management science*, 5(1):1–4, 1958.

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.

Qian Li, Zhichao Wang, Shaowu Liu, Gang Li, and Guandong Xu. Causal optimal transport for treatment effect estimation. *IEEE transactions on neural networks and learning systems*, 34(8): 4083–4095, 2021.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. In *International Conference on Learning Representations*, volume 2024, pp. 39402–39425, 2024.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pp. 666–704, 1781.

Lokesh Nagalapatti, Pranava Singhal, Avishek Ghosh, and Sunita Sarawagi. Pairnet: training with observed pairs to estimate individual treatment effect. *arXiv preprint arXiv:2406.03864*, 2024.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. 2014.

Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

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

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Masahiro Sato, Sho Takemori, Janmajay Singh, and Tomoko Ohkuma. Unbiased learning for the causal effect of recommendation. In *Proceedings of the 14th ACM conference on recommender systems*, pp. 378–387, 2020.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pp. 1670–1679. PMLR, 2016.

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

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Fan Wang, Chaochao Chen, Weiming Liu, Tianhao Fan, Xinting Liao, Yanchao Tan, Lianyong Qi, and Xiaolin Zheng. Ce-rcfr: Robust counterfactual regression for consensus-enabled treatment effect estimation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3013–3023, 2024.

Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. Optimal transport for treatment effect estimation. *Advances in Neural Information Processing Systems*, 36:5404–5418, 2023.

Hao Wang, Zhichao Chen, Zhaoran Liu, Xu Chen, Haoxuan Li, and Zhouchen Lin. Proximity matters: Local proximity enhanced balancing for treatment effect estimation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 2927–2937, 2025.

Yuguang Yan, Hao Zhou, Zeqin Yang, Weilin Chen, Ruichu Cai, and Zhifeng Hao. Reducing balancing error for causal inference via optimal transport. In *Forty-first International Conference on Machine Learning*, 2024.

Yuguang Yan, Zongyu Li, Haolin Yang, Zeqin Yang, Hao Zhou, Ruichu Cai, and Zhifeng Hao. Reducing confounding bias without data splitting for causal inference via optimal transport. In *Forty-second International Conference on Machine Learning*, 2025.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in neural information processing systems*, 31, 2018.

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

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International conference on machine learning*, pp. 7154–7163. PMLR, 2019.

Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10923–10930, 2021.

Yao Zhang, Alexis Bellot, and Mihaela Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pp. 1005–1014. PMLR, 2020.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Kailiang Zhong, Fengtong Xiao, Yan Ren, Yaorong Liang, Wenqing Yao, Xiaofeng Yang, and Ling Cen. Descn: Deep entire space cross networks for individual treatment effect estimation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4612–4620, 2022.

A STANDARD ASSUMPTION

To ensure that the conditional potential outcome functions are identifiable from observational data, we consider the problem under three standard assumptions:

Assumption 1: Stable Unit Treatment Value Assumption The potential outcome of each unit is unaffected by the treatment assignments of other units, and each treatment condition has only one well-defined version.

Assumption 2: Unconfoundedness Conditioned on the covariates, the treatment assignment is independent of the potential outcomes:

$$\{Y(0), Y(1)\} \perp\!\!\!\perp T \mid X. \tag{29}$$

Assumption 3: Overlap For every x with positive probability density, the treatment assignment is not deterministic:

$$0 < p(T = 1 \mid X = x) < 1. \tag{30}$$

Under these assumptions, the conditional potential outcome functions can be identified from the observed distribution:

$$\begin{aligned} \tau(x) &= \mathbb{E}[Y(1) - Y(0) \mid X = x] \\ &= g(1, x) - g(0, x) = g_1(x) - g_0(x) \end{aligned} \tag{31}$$

where $g_1(x) = g(1, x)$ and $g_0(x) = g(0, x)$ denote the conditional potential outcome functions under treatment and control, respectively.

B PROOFS OF THEORETICAL RESULTS

This appendix provides the proofs of the theoretical results established in Section 4.4. Throughout the proofs, P_1 and P_0 denote the covariate distributions of the treated and control populations, respectively.

B.1 PROOF OF THEOREM 1

Theorem 1 *For any measurable real-valued function m, define its balancing error between the treated and control populations as*

$$\begin{aligned} \xi(m) &= |\mathbb{E}_{X \sim P_1}[m(X)] - \mathbb{E}_{X \sim P_0}[m(X)]| \\ &= \left| \int_{\mathcal{X}} m(x) \, \mathrm{d}P_1(x) - \int_{\mathcal{X}} m(x) \, \mathrm{d}P_0(x) \right|. \end{aligned} \tag{32}$$

Assume that m satisfies

$$|m(x) - m(x')| \leq C_s(x, x'). \tag{33}$$

Then,

$$\xi(m) \leq \mathcal{W}(C_s, P_1, P_0). \tag{34}$$

Proof. Firstly, for the balancing error between the treated and control populations on m:

$$\xi(m) = \left\| \mathbb{E}_{x \sim P_1(x)} m(x) - \mathbb{E}_{x \sim P_0(x)} m(x) \right\| = \left\| \int_{\mathcal{X}} m(x) dP_1(x) - \int_{\mathcal{X}} m(x) dP_0(x) \right\|. \tag{35}$$

Under the assumption of Theorem 1, the function m is dominated by the transportation cost C_s, *i.e.*, $|m(x) - m(x')| \leq C_s(x, x')$. Therefore, the discrepancy in Eq. (35) is upper bounded by its worst-case value over all such admissible functions:

$$\xi(m) \leq \sup_{|m(x) - m(x')| \leq C_s(x, x')} \left\{ \int_{\mathcal{X}} m(x) dP_1(x) - \int_{\mathcal{X}} m(x) dP_0(x) \right\}. \tag{36}$$

By the Kantorovich duality of optimal transport, the above dual problem is upper bounded by the corresponding primal transportation problem:

$$\sup_{|m(x)-m(x')|\leq C_s(x,x')} \left\{ \int_{\mathcal{X}} m(x) dP_1(x) - \int_{\mathcal{X}} m(x) dP_0(x) \right\} \leq \inf_{\pi \in \Pi(P_1,P_0)} \int_{\mathcal{X}\times\mathcal{X}} C_s(x,x') d\pi(x,x') \tag{37}$$

$$= \mathcal{W}(C_s, P_1, P_0). \tag{38}$$

Combining the above inequalities gives

$$\xi(m) \leq \mathcal{W}(C_s, P_1, P_0). \tag{39}$$

□

B.2 Proof of Theorem 2

Theorem 2 *The individual treatment effect estimation error is jointly controlled by the factual outcome estimation errors and the causal-structure-guided Wasserstein discrepancy:*

$$\epsilon_{\text{PEHE}} \leq \mathcal{E}_1^{(1)} + \mathcal{E}_0^{(0)} + \mathcal{O}\left(\mathcal{W}(C_s, P_1, P_0)\right). \tag{40}$$

Proof.

$$\begin{aligned}
\epsilon_{PEHE} = \int_{\mathcal{X}} (\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x} &= \int_{\mathcal{X}} ((h_1(\mathbf{x}) - h_0(\mathbf{x})) - (f_1(\mathbf{x}) - f_0(\mathbf{x})))^2 p(\mathbf{x}) d\mathbf{x} \\
&= \int_{\mathcal{X}} ((h_1(\mathbf{x}) - f_1(\mathbf{x})) + (f_0(\mathbf{x}) - h_0(\mathbf{x})))^2 p(\mathbf{x}) d\mathbf{x} \\
&\leq \int_{\mathcal{X}} ((h_1(\mathbf{x}) - f_1(\mathbf{x}))^2 + (f_0(\mathbf{x}) - h_0(\mathbf{x}))^2) p(\mathbf{x}) d\mathbf{x} \\
&= \int_{\mathcal{X}} ((h_1(\mathbf{x}) - f_1(\mathbf{x}))^2 p(\mathbf{x}|t=1) d\mathbf{x} \\
&\quad + \int_{\mathcal{X}} ((h_1(\mathbf{x}) - f_1(\mathbf{x}))^2 p(\mathbf{x}|t=0) d\mathbf{x} \\
&\quad + \int_{\mathcal{X}} ((h_0(\mathbf{x}) - f_0(\mathbf{x}))^2 p(\mathbf{x}|t=1) d\mathbf{x} \\
&\quad + \int_{\mathcal{X}} ((h_0(\mathbf{x}) - f_0(\mathbf{x}))^2 p(\mathbf{x}|t=0) d\mathbf{x} \\
&= \mathcal{E}_1^{(1)} + \mathcal{E}_0^{(1)} + \mathcal{E}_1^{(0)} + \mathcal{E}_0^{(0)}.
\end{aligned} \tag{41}$$

where $\mathcal{E}_1^{(1)}$ and $\mathcal{E}_0^{(0)}$ denote the within-group outcome estimation errors for the treated and control groups, respectively, while $\mathcal{E}_0^{(1)}$ and $\mathcal{E}_1^{(0)}$ denote the corresponding cross-group prediction errors. Under the Theorem 1, it is easy to prove that $\mathcal{E}_0^{(1)} \leq \mathcal{W}(C_s, P_1, P_0)$. so that:

$$\epsilon_{\text{PEHE}} \leq \mathcal{E}_1^{(1)} + \mathcal{E}_0^{(0)} + \mathcal{O}\left(\mathcal{W}(C_s, P_1, P_0)\right). \tag{42}$$

□

B.3 Proof of Theorem 3

To prove Theorem 3, we present the following lemma and prove it in the subsequent sections.

Lemma 4. *For any $\delta \in (0,1)$, with probability at least $1-2\delta$,*

$$\mathcal{W}(C_s, P_1, P_0) \leq \mathcal{W}(C_s, \hat{P}_1, \hat{P}_0) + \mathcal{O}\left(1/\sqrt{\delta n}\right). \tag{43}$$

Now, we remind Theorem 3 and prove it.

Theorem 3 (Finite-Sample PEHE Bound). *For any $\delta \in (0,1)$, with probability at least $1-2\delta$,*

$$\epsilon_{\text{PEHE}} \leq \mathcal{E}_1^{(1)} + \mathcal{E}_0^{(0)} + \mathcal{O}(\mathcal{W}(C_s, \hat{P}_1, \hat{P}_0)) + \mathcal{O}\left(1/\sqrt{\delta n}\right). \tag{44}$$

Proof. Combining Theorem 2 and Lemma 4, it is easy to obtain:

$$\begin{aligned}\epsilon_{\text{PEHE}} &\leq \mathcal{E}_1^{(1)} + \mathcal{E}_0^{(0)} + \mathcal{O}\left(\mathcal{W}(C_s, P_1, P_0)\right) \\ &\leq \mathcal{E}_1^{(1)} + \mathcal{E}_0^{(0)} + \mathcal{O}\big(\mathcal{W}(C_s, \hat{P}_1, \hat{P}_0) + 1/\sqrt{\delta n}\big) \\ &= \mathcal{E}_1^{(1)} + \mathcal{E}_0^{(0)} + \mathcal{E}_1^{(1)} + \mathcal{E}_0^{(0)} + \mathcal{O}(\mathcal{W}(C_s, \hat{P}_1, \hat{P}_0)) + \mathcal{O}\left(1/\sqrt{\delta n}\right). \end{aligned} \tag{45}$$

□

B.4 Proof of Lemma 4

Lemma 4 *For any $\delta \in (0,1)$, with probability at least $1-2\delta$,*

$$\mathcal{W}(C_s, P_1, P_0) \leq \mathcal{W}(C_s, \hat{P}_1, \hat{P}_0) + \mathcal{O}\left(1/\sqrt{\delta n}\right). \tag{46}$$

Proof. With the triangular inequality of the Wasserstein metric,we have:

$$\begin{aligned}\mathcal{W}(C_s, P_1, P_0) &\leq \mathcal{W}(C_s, P_1, \hat{P}_1) + \mathcal{W}(C_s, \hat{P}_1, P_0) \\ &\leq \mathcal{W}(C_s, P_1, \hat{P}_1) + \mathcal{W}(C_s, \hat{P}_1, \hat{P}_0) + \mathcal{W}(C_s, \hat{P}_0, P_0)\end{aligned} \tag{47}$$

Next, we present Lemma 5, which establishes the convergence of the empirical measure $\widehat{\mu}$ to the population measure μ in the Wasserstein distance. This result enables us to derive a finite-sample generalization bound using empirical rather than population measures.

Lemma 5. *((Bolley et al., 2007), Theorem 1.1). Let μ be a probability measure in $\mathbb{R}^d$ satisfying $T_1(\zeta)$ inequality, and $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be its associated empirical measure with n units. Then for any $d' > d$ and $\zeta' < \zeta$, there exists some constant n_0 depending on d' and some square exponential moment of μ such that for any $\epsilon > 0$ and $n \geq n_0 \max(\epsilon^{-(d'+2)}, 1)$*

$$\mathbb{P}\left[\mathcal{W}_1(\mu, \hat{\mu}) > \epsilon\right] \leq \exp\left(-\frac{\zeta' n \epsilon^2}{2}\right), \tag{48}$$

where d', ζ' can be calculated explicitly.

The Hoeffding inequality in Lemma 5 gives the following inequality which holds with the probability at least $1-\delta$:

$$\begin{aligned}\mathcal{W}(C_s, P_1, \hat{P}_1) &\leq \sqrt{2\log\left(\frac{1}{\delta}\right)/\zeta' n}, \\ \mathcal{W}(C_s, \hat{P}_0, P_0) &\leq \sqrt{2\log\left(\frac{1}{\delta}\right)/\zeta' n}.\end{aligned} \tag{49}$$

Combining Eq. (47) and Eq. (49) together, we have:

$$\begin{aligned}\mathcal{W}(C_s, P_1, P_0) &\leqslant \sqrt{2\log\left(\frac{1}{\delta}\right)/\zeta' n} + \sqrt{2\log\left(\frac{1}{\delta}\right)/\zeta' n} + \mathcal{W}(C_s, \hat{P}_1, \hat{P}_0) \\ &= \mathcal{W}(C_s, \hat{P}_1, \hat{P}_0) + 2\sqrt{2\log\left(\frac{1}{\delta}\right)/\zeta' n} \\ &:= \mathcal{W}(C_s, \hat{P}_1, \hat{P}_0) + \mathcal{O}\left(1/\sqrt{\delta n}\right),\end{aligned} \tag{50}$$

which finishes the proof.

□

C EXPERIMENTAL DETAILS

C.1 EXPERIMENTAL SETUP

To comprehensively evaluate the performance of SROT-CFR for individual treatment effect estimation, we select representative baselines from statistical learning, neural networks, counterfactual representation learning, and optimal transport. All methods use the same data splits, evaluation metrics, and repeated experimental settings, and we consistently report out-of-sample results. (i) Traditional methods: **OLS** estimates potential outcomes using linear regression; **KNN** predicts counterfactual outcomes based on local nearest neighbors; **PSM** Rosenbaum & Rubin (1983) matches similar treated and control units according to propensity scores; **T-Learner** Künzel et al. (2019) trains separate outcome models for the two treatment conditions; **BART** Chipman et al. (2010) employ Bayesian additive regression trees and Bayesian neural networks, respectively, to model complex nonlinear potential outcome functions. **MLP** serves as a basic neural baseline that directly learns the mapping from covariates and treatment assignments to outcomes; (ii) Representation balanced methods: **BNN**Johansson et al. (2016) and **CFR**Shalit et al. (2017) reduces the discrepancy between treated and control representations using MMD or Wasserstein based regularization; **GANITE**Yoon et al. (2018) uses generative adversarial networks to impute missing counterfactual outcomes; **SITE**Yao et al. (2018) further preserves local sample similarities while performing global distribution balancing; **DragonNet**Shi et al. (2019) jointly estimates potential outcomes and propensity scores and improves treatment effect estimation through targeted regularization; **DKLite**Zhang et al. (2020) employs deep kernel learning to improve counterfactual generalization in regions with sufficient overlap; distributional regularization. and **DESCN**Zhong et al. (2022) adopts multitask learning to jointly model treatment propensity, observed outcomes, and individual treatment effects, thereby mitigating distributional shifts between treatment groups. (iii) Optimal Transport based methods: **CausalOT**Li et al. (2021) formulates matching between treated and control units as an optimal transport problem; **CFR-Pro**Wang et al. (2025) introduces more fine-grained sample relationships or local constraints into counterfactual representation learning; **CE-RCFR**Wang et al. (2024) improves robustness to sampling variation and treatment-group imbalance through robust representation learning and **ESCFR**Wang et al. (2023) alleviates mini-batch distribution estimation bias through an enhanced optimal transport discrepancy and Sinkhorn optimization; **ORIC**Yan et al. (2025), as a recent treatment effect estimation method, is also included under the same experimental protocol.

C.2 DATASETS

Simulation Data.

To evaluate the performance of the proposed method under complex covariate structures and different confounding levels, we construct a synthetic dataset containing $5{,}000$ samples and 30 covariates. Specifically, X_1–X_{24} are generated from three structural modules involving linear, nonlinear, and interaction relationships:

$$X_j = f_j\big(X_{\mathrm{Pa}(j)}\big) + \epsilon_j, \qquad \epsilon_j \sim \mathcal{N}(0, 0.4^2), \tag{51}$$

while X_{25}–$X_{30} \sim \mathcal{N}(0,1)$ are independent covariates. The treatment assignment is generated according to

$$e(X) = \sigma(\alpha q(X)), \qquad T \sim \mathrm{Bernoulli}(e(X)), \tag{52}$$

where $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ controls the confounding strength. The potential outcomes are defined as

$$\mu_1(X) = \mu_0(X) + \tau(X), \tag{53}$$

$$Y(t) = \mu_t(X) + \epsilon_Y, \qquad \epsilon_Y \sim \mathcal{N}(0, 0.5^2), \tag{54}$$

where both $\mu_0(X)$ and the heterogeneous treatment effect $\tau(X)$ contain linear, nonlinear, and interaction terms. For each confounding level, we generate ten datasets using different random seeds, resulting in a total of 50 datasets. Treatment assignments and potential outcomes are generated solely from the original covariates X, while structural residuals are learned during model training and are not involved in the data-generating process.

real-world dataset Following Zhang et al. (2021); Wang et al. (2023), we evaluate SROT-CFR on two widely used semi-synthetic benchmark datasets, **IHDP** Hill (2011) and **ACIC 2016** Dorie

Table 3: In-sample results on the synthetic dataset under different confounding levels α in terms of $\sqrt{\epsilon_{\text{PEHE}}}$. Results are reported as mean $\pm$ standard deviation, with lower values indicating better performance. The best result is shown in bold, and the second-best result is underlined. Rankings are based on the unrounded mean values.

Method	$\boldsymbol{\alpha = 0.2}$	$\boldsymbol{\alpha = 0.4}$	$\boldsymbol{\alpha = 0.6}$	$\boldsymbol{\alpha = 0.8}$	$\boldsymbol{\alpha = 1.0}$
OLS	0.3582 ± 0.0105	0.3506 ± 0.0102	0.3473 ± 0.0104	0.3520 ± 0.0111	0.3505 ± 0.0053
KNN	0.6209 ± 0.0093	0.6249 ± 0.0094	0.6261 ± 0.0115	0.6276 ± 0.0098	0.6324 ± 0.0085
PSM	0.3983 ± 0.0246	0.3963 ± 0.0161	0.3905 ± 0.0158	0.3901 ± 0.0139	0.3838 ± 0.0109
T-Learner	0.5952 ± 0.0242	0.6054 ± 0.0338	0.6091 ± 0.0272	0.6086 ± 0.0263	0.6342 ± 0.0491
BART	0.3890 ± 0.0115	0.3958 ± 0.0148	0.3936 ± 0.0158	0.4074 ± 0.0201	0.4054 ± 0.0178
MLP	0.4319 ± 0.0101	0.4358 ± 0.0098	0.4292 ± 0.0084	0.4312 ± 0.0124	0.4289 ± 0.0118
BNN	0.2281 ± 0.0148	0.2354 ± 0.0126	0.2412 ± 0.0129	0.2471 ± 0.0125	0.2560 ± 0.0143
CFR-Wass	0.3440 ± 0.0051	0.3411 ± 0.0034	0.3403 ± 0.0026	0.3420 ± 0.0025	0.3432 ± 0.0019
GANITE	0.9091 ± 0.0534	0.7380 ± 0.0353	0.9020 ± 0.0602	0.8387 ± 0.0474	0.8576 ± 0.0423
SITE	0.2609 ± 0.0082	0.2623 ± 0.0074	0.2617 ± 0.0152	0.2591 ± 0.0111	0.2869 ± 0.0171
DragonNet	2.0523 ± 0.0059	2.0510 ± 0.0043	2.0514 ± 0.0036	2.0511 ± 0.0063	2.0520 ± 0.0041
DKLite	0.3696 ± 0.0044	0.3796 ± 0.0130	0.3924 ± 0.0047	0.3996 ± 0.0084	0.4199 ± 0.0127
DESCN	0.3677 ± 0.0095	0.3742 ± 0.0161	0.3838 ± 0.0149	0.3922 ± 0.0196	0.4133 ± 0.0198
CausalOT	1.0286 ± 0.0084	1.0227 ± 0.0066	1.0181 ± 0.0046	1.0137 ± 0.0084	1.0119 ± 0.0050
CFR-Pro	0.1194 ± 0.0020	0.1236 ± 0.0030	0.1209 ± 0.0038	0.1312 ± 0.0079	0.1310 ± 0.0060
CE-RCFR	0.1385 ± 0.0060	0.1510 ± 0.0094	0.1445 ± 0.0061	0.1420 ± 0.0093	0.1471 ± 0.0057
ESCFR	0.1236 ± 0.0119	0.1219 ± 0.0087	0.1249 ± 0.0098	0.1234 ± 0.0050	0.1278 ± 0.0115
CSOT-Direct	0.1141 ± 0.0039	0.1159 ± 0.0045	0.1237 ± 0.0079	0.1239 ± 0.0049	0.1208 ± 0.0055
CSOT-ICA	0.1017 ± 0.0022	0.1026 ± 0.0023	0.1047 ± 0.0033	0.1054 ± 0.0037	$\underline{0.1077 \pm 0.0045}$
CSOT-NOTEARS	$\mathbf{0.0995 \pm 0.0019}$	$\mathbf{0.1010 \pm 0.0024}$	$\mathbf{0.1034 \pm 0.0033}$	$\mathbf{0.1040 \pm 0.0037}$	0.1089 ± 0.0044
CSOT-PC	$\underline{0.1016 \pm 0.0022}$	0.1036 ± 0.0026	0.1047 ± 0.0028	$\underline{0.1053 \pm 0.0040}$	$\mathbf{0.1066 \pm 0.0039}$
CSOT-GOLEM	0.1020 ± 0.0021	$\underline{0.1026 \pm 0.0024}$	$\underline{0.1038 \pm 0.0032}$	0.1069 ± 0.0048	0.1081 ± 0.0044

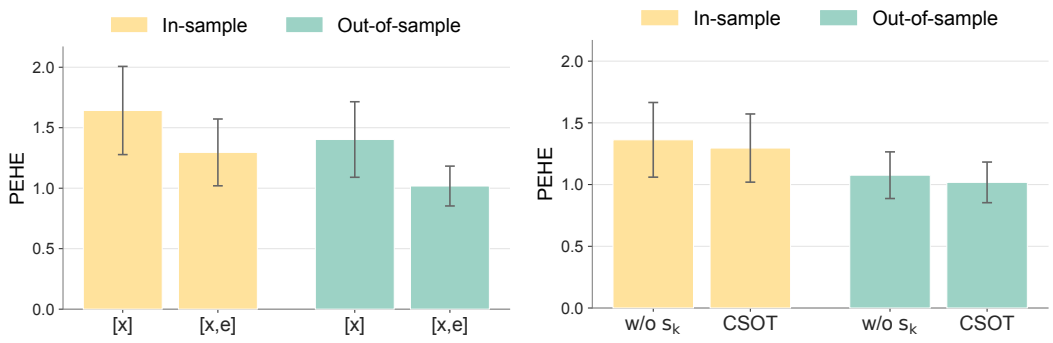
et al. (2019). IHDP originates from a study of interventions for premature infants and contains 747 samples with 25 covariates. Its semi-synthetic construction preserves the real covariates and treatment assignments while generating potential outcomes according to predefined response functions, thereby providing ground-truth individual treatment effects for quantitative evaluation. ACIC 2016 is based on real data from a perinatal study and contains $4{,}802$ samples with 58 covariates. Treatment assignments and potential outcomes are simulated over the real covariate distribution under settings with varying degrees of confounding, overlap, outcome-function complexity, and treatment-effect heterogeneity. Overall, these semi-synthetic datasets preserve the complex covariate distributions of real-world data while providing known treatment effects through simulated treatment assignments or potential outcomes, enabling quantitative evaluation under realistic observational settings.

C.3 Ablation Study and Sensitivity Analysis

Ablation Study. Figure 2 presents the ablation results on ACIC 2016 over 10 independent realizations. The bars and error bars report the mean PEHE and standard deviation, respectively. Figure 2a evaluates the effect of incorporating structural information into the input. Here, $[X]$ directly feeds the original covariates into the encoder, whereas $[X, e]$ additionally incorporates the estimated structural residual vector e. Compared with $[X]$, $[X, e]$ achieves lower mean PEHE and smaller standard deviation under both in-sample and out-of-sample settings. This result indicates that the structural residuals provide complementary information about deviations from the estimated covariate relationships, thereby improving the accuracy and stability of the learned representation.

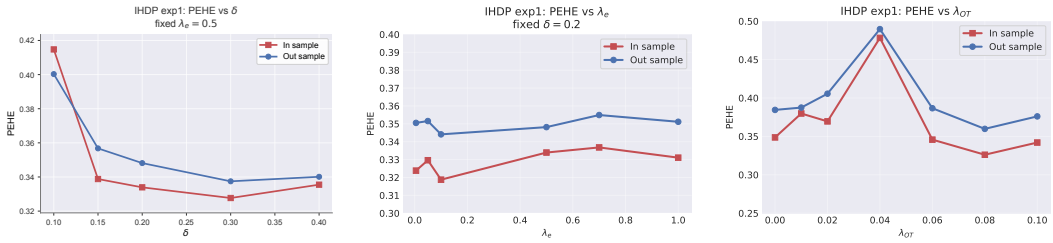
Figure 2b evaluates the contribution of the graph derived weights s_k to the transport cost C. The variant w/o s_k uses $[X, e]$ as the input and retains the structural discrepancy in the transport cost but assigns uniform importance to all covariates. CSOT uses the normalized weights s_k derived from the estimated covariate graph. Ours achieves lower mean PEHE and smaller standard deviation than w/o s_k under both evaluation settings. This result demonstrates that s_k enables the transport cost to account for the different structural roles of the covariates, leading to more reliable alignment between the treated and control groups.

Sensitivity Analysis. Figure 3 presents a sensitivity analysis on the first realization of the IHDP dataset. DirectLiNGAM is used to estimate the covariate structure, and each hyperparameter is varied separately while the remaining settings are fixed. We vary the edge-selection threshold δ,



(a) Structure augmented input (b) Transport cost

Figure 2: Ablation analysis of CSOT on ACIC 2016 in terms of ϵ_{PEHE}.



(a) δ (b) λ_e (c) λ_{OT}

Figure 3: Hyperparameter sensitivity analysis of CSOT on IHDP with respect to δ, λ_e, and λ_{OT}.

the structural-residual coefficient λ_e, and the optimal transport coefficient λ_{OT} individually while keeping the remaining settings fixed. Specifically, λ_e is fixed at 0.5 when varying δ, whereas δ is fixed at 0.2 when varying λ_e. The results show that increasing δ initially reduces both in-sample and out-of-sample PEHE, with the best performance obtained around $\delta = 0.3$, indicating that filtering weak or potentially spurious edges improves the estimated covariate structure. For λ_e, relatively small values perform better, suggesting that structural residuals provide useful aligning information. For λ_{OT}, moderate values provide a favorable trade-off between factual prediction and cross-group alignment. These findings demonstrate that appropriately balancing structural filtering, structural guidance, and distribution alignment is important for accurate and stable treatment effect estimation.