ENHANCING CAUSAL EFFECT FOUNDATION MODEL WITH TASK-SPECIFIC MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal effect estimation aims to evaluate the effects of a treatment and plays a vital role in decision-making. Benefiting from in-context learning, tabular foundation models have shown promising performance for causal effect estimation without fine-tuning on task-specific context data. Despite their strong generalization ability to new tasks, foundation models cannot fully address the distribution shift between treatment groups caused by confounding bias, which limits the performance of causal effect estimation. One natural approach is to learn an encoder to explicitly align the representations of groups. However, enforcing representation alignment could lead to over-balancing, which means that predictive information for outcome prediction is partially removed. Although minimizing the prediction loss of the foundation model can provide guidance for learning the encoder to some extent, the complex architecture of the foundation model makes it difficult to effectively propagate task-relevant gradient signals from the output back to the encoder. To address this issue, we incorporate a lightweight task-specific predictor after the encoder to preserve outcome-relevant predictive information in the learned representations. By doing this, the foundation model is enhanced by the task-specific encoder and predictor trained on context data. Moreover, we design a gated aggregation mechanism to adaptively fuse the predictions of the foundation model and the task-specific model. We theoretically analyze the estimation error of our method, and conduct extensive experiments based on multiple backbones of foundation models. The experimental results on benchmark datasets demonstrate the effectiveness of our method and its general applicability across different backbones.

1 INTRODUCTION

Causal effect estimation from observational data plays an important role in decision-making and has been widely used in many real-world applications, such as public policy (Athey & Wager, 2021) and recommendation (Schnabel et al., 2016; Hays & Raghavan, 2025). The fundamental challenge of causal effect estimation is that for each individual, we can only observe the factual outcome under treatment or non-treatment, while the counterfactual outcome cannot be observed. To tackle this, machine learning-based methods have been proposed to learn a task-specific model for outcome prediction (Johansson et al., 2016; Louizos et al., 2017). By training a sophisticated model to minimize the factual prediction error, these methods have dominated causal effect estimation in the last decades (Yao et al., 2018; Shi et al., 2019).

Over the past few years, with the progress of in-context learning (Brown et al., 2020; Xie et al., 2021), tabular foundation models have shown promising performance for causal effect estimation (Hollmann et al., 2025; Qu et al., 2025). Different from conventional approaches that train a specific model for each task, a tabular foundation model for causal effect estimation is trained on a large collection of datasets and can be applied to new tasks without fine-tuning on task-specific context data (Robertson et al., 2026; Balazadeh Meresht et al., 2026). By learning a prior over causal effect estimation tasks during pretraining, these models can perform Bayesian inference conditioned on the context data of a new task, achieving competitive performance across diverse causal effect estimation tasks.

Although foundation models have demonstrated strong generalization ability to new tasks, they still suffer from distribution imbalance between control and treated groups induced by confounding bias, which can limit the performance of causal effect estimation. A related issue has also been recognized

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.

in existing studies (He & Shi, 2026; Loza et al., 2026), which find that label or feature shift can degrade the performance of foundation models. Directly fine-tuning a foundation model to tackle confounding bias is computationally expensive and may lead to overfitting, especially when the amount of context data is limited. Therefore, we instead seek to retain the generalization ability of foundation models while alleviating the distribution imbalance between treatment groups.

One natural approach is to learn an encoder to align the representations of treated and control groups before feeding them into the foundation model. Nevertheless, enforcing representation alignment may lead to an over-balancing issue (Yan et al., 2026), where outcome-relevant predictive information is partially removed from the learned representations. Although the prediction loss of the foundation model can be leveraged as supervised information to guide the encoder, the complex architecture of the foundation model makes it difficult to effectively propagate task-specific supervised signal from the output back to the encoder. To address this issue, we incorporate a lightweight task-specific predictor after the encoder to provide a more direct and effective supervision pathway for preserving outcome-relevant information in the learned representations. By doing this, the foundation model is enhanced with a task-specific encoder and predictor trained on context data, with the parameters of the foundation model frozen.

Moreover, we design a learnable gated aggregation mechanism to adaptively combine the predictions of the foundation model and the task-specific model, allowing the final model to leverage both the generalization ability of the foundation model and the task-specific information extracted from the context data. We further analyze the estimation error of the proposed method, and evaluate it with diverse foundation model backbones, including both general-purpose tabular foundation models, and causal effect foundation models. Extensive experiments on benchmark datasets demonstrate the effectiveness of our method.

The major contributions are summarized in the following:

- We propose to enhance causal effect foundation models by learning a task-specific encoder and predictor for reducing confounding bias and effect estimation error.
- We adaptively combine the predictions of causal effect foundation models and task-specific models by learnable gated aggregation, and analyze the estimation error of our method.
- We apply the proposed enhancement method to multiple foundation models and conduct extensive experiments to demonstrate its effectiveness and generality across different backbones.

2 Related Work

Treatment effect estimation from observational data requires addressing confounding induced by nonrandom treatment assignment. Existing task-specific approaches tackle this challenge through covariate adjustment and flexible outcome modeling (Rosenbaum & Rubin, 1983; Abadie & Imbens, 2006; Hill, 2011; Wager & Athey, 2018; Athey et al., 2019), as well as neural predictive and generative models (Louizos et al., 2017; Yoon et al., 2018; Shi et al., 2019; Curth & Van der Schaar, 2021). Among these approaches, representation balancing explicitly addresses the distributional discrepancy between treated and control groups, aiming to improve counterfactual generalization while retaining information for outcome prediction (Johansson et al., 2016; Shalit et al., 2017). Subsequent studies refine this principle through local similarity preservation, overlap considerations, and adversarial learning (Yao et al., 2018; Zhang et al., 2020; Lu et al., 2020). Optimal transport further provides a framework for distribution alignment, with recent methods improving transport based balancing and incorporating local proximity information (Li et al., 2021; Wang et al., 2023; 2025).

General purpose tabular models provide a means of transferring predictive knowledge across datasets through pretraining. These models adapt to new tasks through in context inference conditioned on observed examples (Hollmann et al., 2025; Qu et al., 2025; Zhang et al., 2025) or through hypernetworks that construct task-specific predictors (Bonet et al., 2024; Müller et al., 2025). Recent developments further extend their scalability and generalization across diverse tabular prediction tasks, including classification and regression (Qu et al., 2025; Zhang et al., 2025; Ma et al., 2026a). By reusing knowledge acquired during pretraining, these approaches reduce reliance on learning predictive relationships solely from the current dataset. However, their training objectives primarily target general prediction, and strong predictive performance alone does not ensure valid treatment effect

estimation, which additionally requires addressing confounding and the absence of counterfactual supervision.

Recent studies extend tabular foundation models to more complex settings, including large and temporal distribution shifts (Helli et al., 2024) and high dimensional datasets (Liu & Ye, 2025). Among them, some extend tabular foundation models to causal inference by incorporating causal structures and interventions into pretraining. These approaches learn transferable causal priors from simulated tasks to support interventional prediction and treatment effect estimation (Robertson et al., 2026; Balazadeh Meresht et al., 2026; Ma et al., 2026b), with related applications extending to causal fairness (Robertson et al., 2025). Such developments enable the reuse of causal knowledge across tasks. However, strong cross-task generalization does not necessarily imply robustness to the covariate distribution shift between treated and control groups in observational data. Recent empirical evidence shows that tabular foundation models exhibit degraded predictive performance under different types of distribution shifts (Loza et al., 2026). This limitation is particularly relevant to causal effect estimation, where the treated and control groups may follow different distributions due to confounding. We address this by combining a task-specific encoder and predictor with the frozen foundation model through gated aggregation and representation balancing for treatment effect estimation.

3 PRELIMINARY

We consider the standard Neyman–Rubin potential outcome framework for treatment effect estimation from observational data (Rubin, 1974; Splawa-Neyman et al., 1990). Let $\{(\mathbf{x}_i, t_i, y_i)\}_{i=1}^n$ denote an observational dataset, where $\mathbf{x}_i \in \mathcal{X}$ represents the covariates of the i-th sample, $t_i \in \{0, 1\}$ is the binary treatment indicator, with $t_i = 1$ denoting the treated group and $t_i = 0$ denoting the control group, $Y(1)$ and $Y(0)$ denote the potential outcomes under treatment and control.

For a given covariate vector $\mathbf{x} \in \mathcal{X}$, our objective is to estimate the conditional average treatment effect (CATE), defined as

$$\tau(\mathbf{x}) = \mathbb{E}[Y(1) - Y(0) \mid X = \mathbf{x}]\,. \tag{1}$$

For the identification of the CATE from observational data, we adopt the following standard assumptions.

Assumption 1 (Ignorability)**.** Conditional on the observed covariates X, the treatment assignment is independent of the potential outcomes: $\{Y(0), Y(1)\} \perp T \mid X$.

Assumption 2 (Positivity)**.** For any $\mathbf{x} \in \mathcal{X}$ with positive probability density: $0 < P(T = t \mid X = \mathbf{x}) < 1$.

Under these assumptions, $\tau(\mathbf{x})$ can be represented by the potential outcome functions $\mu_t(\mathbf{x})$ as follows:

$$\tau(\mathbf{x}) = \mathbb{E}[Y(1) - Y(0) \mid X = \mathbf{x}] = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}). \tag{2}$$

To estimate the treatment effect, we learn the potential outcome predictors $\hat{\mu}_0(\mathbf{x})$ and $\hat{\mu}_1(\mathbf{x})$. The estimated individual treatment effect is then given by

$$\hat{\tau}(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x}). \tag{3}$$

4 METHODOLOGY

Figure 1 illustrates our **t**ask-**e**nhanced **c**ausal effect **F**oundation **M**odel (tecFM), which enhances a frozen foundation model with task-specific learning for causal effect estimation. A trainable encoder maps the covariates into a representation space, where Wasserstein distance is used to reduce the distribution discrepancy between the treated and control groups. To preserve outcome-relevant information, a lightweight task-specific predictor provides direct supervision to the encoder, and its predictions are adaptively combined with those of the frozen foundation model through a learnable gate. The encoder, task-specific predictor, and gate are jointly optimized, and individual treatment effects are obtained from the difference between the estimated potential outcomes.

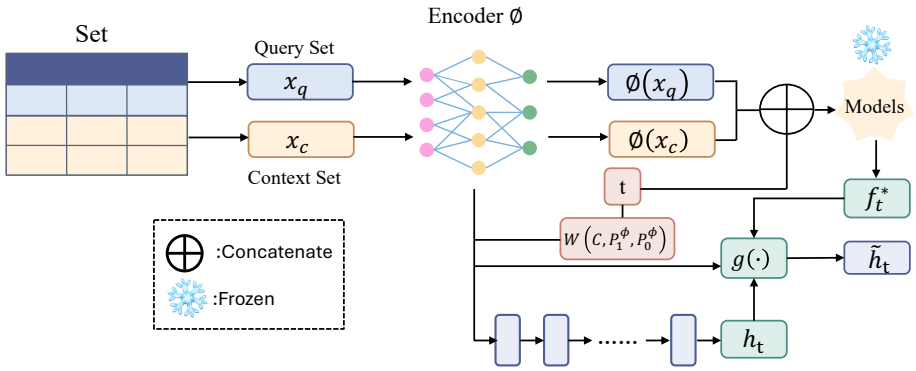


Figure 1: Overview of the proposed method.

4.1 Representation Balancing For Foundation Model

Tabular Foundation Models. We focus on tabular foundation models that make predictions through in-context learning. These models are pretrained on a large collection of tasks, each typically containing a context set and a query set (Qu et al., 2025; Balazadeh Meresht et al., 2026). The context set provides labeled examples, and the model learns to predict the outcomes of query samples using their inputs together with the context examples. Through this process, the model learns predictive knowledge that can be transferred across tasks. Their prediction objective is formulated as follows:

$$\Phi^* = \arg\min_{\Phi} \sum_{m=1}^{M} \sum_{j=1}^{N_m} -\log p(y_j^m | \mathbf{x}_j^m, D_m; \Phi) \tag{4}$$

where M is the number of pretraining tasks, N_m denotes the number of query samples in task m, D_m is the corresponding context dataset, Φ represents the shared parameters of the foundation model, and Φ^* denotes the pretrained model parameters. When applied to a new task, the model keeps the pretrained parameters fixed and performs predictions using the observational data from the current task as context, as follows:

$$f_t^* = \mathcal{G}(\mathcal{C}, \mathcal{Q}; \Phi^*), \tag{5}$$

where f_t^* denotes the prediction function under treatment t, $\mathcal{G}$ represents the pretrained foundation model, $\mathcal{C} = (\mathbf{x}_i, t_i, y_i)$ denotes the observational data of the current task used as context, and $\mathcal{Q} = (\mathbf{x}_i, t_i)$ denotes the query covariates for which potential outcomes are predicted.

Although tabular foundation models acquire strong generalization ability through large-scale pretraining, treatment assignment in observational data is typically nonrandom, resulting in different covariate distributions between the treated and control groups. Consequently, the context data provided to the foundation model for the current task may still exhibit substantial distribution discrepancy between the two groups. Existing studies have also shown that changes in input feature distributions can lead to performance degradation in tabular foundation models (Loza et al., 2026; He & Shi, 2026). This indicates that strong generalization ability on across tasks does not necessarily imply inherent robustness to distribution shift within new tasks.

To alleviate this issue, we introduce a trainable representation encoder ϕ before the frozen foundation model and map the raw covariates $\mathbf{x}_i$ into a new representation $\phi(\mathbf{x}_i)$. Accordingly, the context data and query input of the current task are transformed as

$$\mathcal{C}_\phi = \{(\phi(\mathbf{x}_i), t_i, y_i)\}_{i=1}^N, \ \mathcal{Q}_\phi = (\phi(\mathbf{x}_i), t_i). \tag{6}$$

The corresponding potential outcome prediction of the foundation model is then given by

$$f_t^* \circ \phi = \mathcal{G}\left(\mathcal{C}_\phi, \mathcal{Q}_\phi; \Phi^*\right). \tag{7}$$

Here, the pretrained parameters Φ^* remain fixed, while the encoder ϕ is optimized using the observational data of the current task. In this way, keeping the foundation model frozen avoids the computational overhead and potential overfitting associated with fine-tuning, while allowing the encoder to adapt the input representations to the current task. Let P_1^ϕ and P_0^ϕ denote the representation distributions of the treated and control groups, respectively. We reduce their distribution discrepancy using the Wasserstein distance and define the representation balancing loss as

$$\mathcal{W}(C, P_1^\phi, P_0^\phi) = \inf_{\gamma \in \Gamma(P_1^\phi, P_0^\phi)} \int C\left(\phi(\mathbf{x}), \phi(\mathbf{x}')\right) \, d\gamma\left(\phi(\mathbf{x}), \phi(\mathbf{x}')\right), \tag{8}$$

where $\Gamma(P_1^\phi, P_0^\phi)$ denotes the set of transport plans with marginals P_1^ϕ and P_0^ϕ, and $C\left(\phi(\mathbf{x}), \phi(\mathbf{x}')\right) = \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2$ is the transport cost between two representations. Combining the above components, the final objective is formulated as follows:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \left(f_t^*(\mathbf{x}_i) - y_i\right)^2 + \lambda_w \hat{\mathcal{W}}(C, P_1^\phi, P_0^\phi), \tag{9}$$

where $\hat{\mathcal{W}}(C, P_1^\phi, P_0^\phi)$ denotes the representation balancing loss computed between the empirical distributions of the treated and control representations, λ_w is a hyperparameter controlling the contribution of this balancing term, and $f_t^*(\mathbf{x}_i)$ denotes the frozen foundation model used to predict outcomes under treatment t.

4.2 Task-specific Prediction for Representation Learning

Representation alignment reduces the distribution discrepancy between treated and control groups. However, excessive alignment may also suppress information needed to predict potential outcomes, limiting estimation performance. Although the factual outcome loss of the foundation model encourages the encoder to retain such information, its gradient must pass through the complex architecture of the frozen foundation model before reaching the encoder. This indirect path may limit how effectively the encoder adapts to the current task. To provide more direct outcome supervision, we introduce a lightweight task-specific predictor on top of the encoder. Given a representation encoder ϕ and a task-specific outcome predictor h_t, the task-specific branch is optimized using the factual outcomes of the current task as follows:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \left(g f_t^*(\mathbf{x}_i) + (1-g) h_t(\phi(\mathbf{x}_i)) - y_i\right)^2 + \lambda_w \hat{\mathcal{W}}(C, P_1^\phi, P_0^\phi), \tag{10}$$

where g combines the two predictions using a fixed, predefined fusion weight, which can be implemented by simple average fusion with $g = 0.5$. In this way, the model can leverage both the generalization ability of the frozen foundation model and the task-specific information captured by the lightweight predictor.

To further improve the fusion of the two predictors, we replace the fixed fusion weight with a learnable gating function. Specifically, $g_t(\phi(\mathbf{x}_i)) \in (0, 1)$ denotes the adaptive gating weight for sample i under treatment condition t, and is jointly optimized with the representation encoder and the task-specific predictor. Accordingly, the gated potential outcome prediction is given by

$$\tilde{h}_t(\phi(\mathbf{x}_i)) = g_t(\phi(\mathbf{x}_i)) h_t(\phi(\mathbf{x}_i)) + \left(1 - g_t(\phi(\mathbf{x}_i))\right) f_t^*(\phi(\mathbf{x}_i)), \tag{11}$$

where $\tilde{h}_t(\phi(\mathbf{x}_i))$ denotes the potential outcome estimated by aggregating the predictions of the task-specific model and the foundation model. Therefore, the final objective $\mathcal{L}_y$ is formulated as follows:

$$\mathcal{L}_y = \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{h}_t(\phi(\mathbf{x}_i)) - y_i\right)^2 \tag{12}$$

The adaptive weight allows the contributions of the foundation model and the task-specific predictor to vary across samples and treatment conditions. During training, an early saturation of g_t toward either 0 or 1 may prevent the two predictors from being sufficiently exploited. We therefore introduce an entropy regularization term to discourage premature collapse, defined as

$$H(g_t(\phi(\mathbf{x}_i))) = -g_t(\phi(\mathbf{x}_i)) \log g_t(\phi(\mathbf{x}_i)) - \left(1 - g_t(\phi(\mathbf{x}_i))\right) \log\left(1 - g_t(\phi(\mathbf{x}_i))\right) \tag{13}$$

This term discourages premature saturation of the gate, allowing both predictors to contribute to the fused output while the task-specific predictor learns from the current task. The corresponding gated prediction loss is defined as

$$\mathcal{L}_g = -\frac{1}{2n}\sum_{i=1}^{n}\sum_{t\in\{0,1\}} H(g_t(\phi(\mathbf{x}_i)), \tag{14}$$

Combining the above components, the overall training objective of tecFM is formulated as

$$\mathcal{L} = \mathcal{L}_y + \lambda_g \mathcal{L}_g + \lambda_w \hat{\mathcal{W}}(C, P_1^{\phi}, P_0^{\phi}), \tag{15}$$

where $\mathcal{L}_y$ denotes the factual prediction loss of the gated prediction, $\mathcal{L}_g$ denotes the gating regularization term, and $\hat{\mathcal{W}}(C, P_1^{\phi}, P_0^{\phi})$ denotes the representation balancing loss between the treated and control groups. The coefficients λ_g and λ_w control the contributions of the corresponding objectives. During optimization, the parameters of the foundation model remain fixed, while the encoder, task-specific predictor, and gating network are jointly updated.

4.3 Theoretical Analysis

We analyze the corresponding errors of the task-specific model and the foundation model. For $t, t' \in \{0, 1\}$, we define the weighted potential outcome prediction errors of the task-specific and foundation models as

$$\mathcal{E}_{t,h}^{(t')} = \int_{\mathcal{X}} g_t(\mathbf{x}) \left(h_t(\mathbf{x}) - \mu_t(\mathbf{x})\right)^2 dP_{t'}(\mathbf{x}), \tag{16}$$

$$\mathcal{E}_{t,f}^{(t')} = \int_{\mathcal{X}} \left(1 - g_t(\mathbf{x})\right) \left(f_t^*(\mathbf{x}) - \mu_t(\mathbf{x})\right)^2 dP_{t'}(\mathbf{x}). \tag{17}$$

Here, h_t and f_t^* denote the potential outcome predictions of the task-specific model and the foundation model under treatment condition t, respectively, while μ_t denotes the corresponding true potential outcome function. The errors are evaluated over the covariate distribution $P_{t'}$ of group t'. Based on these definitions, we derive a bound relating the factual prediction errors of both models and the discrepancy between treatment groups to the individual treatment effect estimation error.

Theorem 1 (Individual Treatment Effect Estimation Error Bound)**.** *The individual treatment effect estimation error is jointly controlled by the factual outcome estimation errors and the Wasserstein discrepancy:*

$$\epsilon_{\text{PEHE}} \leq 2\left(\mathcal{E}_{1,h}^{(1)} + \mathcal{E}_{1,f}^{(1)} + \mathcal{E}_{0,h}^{(0)} + \mathcal{E}_{0,f}^{(0)}\right) + \mathcal{O}(\mathcal{W}(C, P_1, P_0))\,. \tag{18}$$

Theorem 1 shows that the PEHE is controlled by the factual outcome prediction errors of the foundation model and the task-specific model, together with the discrepancy between the treated and control distributions. The factual error of the foundation model is minimized by learning the encoder while keeping the foundation model fixed. The factual error of the task-specific model is minimized by learning both the encoder and its predictor. 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 distributions P_1 and P_0 are unavailable and can only be approximated using finite samples. Let $\widehat{P}_1$ and $\widehat{P}_0$ denote the corresponding empirical distributions. We therefore derive a finite sample bound that connects the population PEHE to the empirical Wasserstein discrepancy.

Theorem 2 (Finite Sample PEHE Bound)**.** *With high probability, the individual treatment effect estimation error satisfies*

$$\epsilon_{\text{PEHE}} \leq 2\left(\mathcal{E}_{1,h}^{(1)} + \mathcal{E}_{1,f}^{(1)} + \mathcal{E}_{0,h}^{(0)} + \mathcal{E}_{0,f}^{(0)}\right) + \mathcal{O}\Big(\mathcal{W}(C, \widehat{P}_1, \widehat{P}_0)\Big) + \mathcal{O}(\varepsilon_n)\,, \tag{19}$$

where ε_n denotes the finite sample discrepancy between the empirical and population transport quantities.

Theorem 2 connects the population treatment effect estimation error to the Wasserstein discrepancy computed from finite samples. As the sample size increases, the finite sample term ε_n decreases and the empirical discrepancy approaches its population counterpart.

The detailed theoretical proofs are provided in the Appendix A.

5 EXPERIMENTS

In this section, we first present the experimental setup, including the compared methods, datasets, and evaluation metrics. After that, we report the main results on the simulation and semi-synthetic datasets. Additional analyses, including ablation studies and sensitivity analyses, are provided in the Appendix C. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU with 24 GB of memory.

5.1 EXPERIMENTAL SETTINGS

Compared Methods. We compare our method with conventional prediction methods and representative causal inference methods. Abadie & Imbens (2006) introduced KNN matching, which estimates counterfactual outcomes using nearest neighbors from the opposite treatment group. MLP predicts potential outcomes using multilayer perceptrons, while Chipman et al. (2010); Hill (2011) used BART for potential outcome estimation. For representative causal inference methods, we conduct CFR (Shalit et al., 2017), DESCN (Zhong et al., 2022), CausalOT (Li et al., 2021), ESCFR (Wang et al., 2023), and ORIC (Yan et al., 2025) for comparison.

We further compare four recent tabular and causal effect foundation models. Among them, TabICL (Qu et al., 2025; 2026) is a state-of-the-art general-purpose tabular foundation model, DoPFN (Robertson et al., 2026), CausalFM (Ma et al., 2026b) and CausalPFN (Balazadeh Meresht et al., 2026) are tabular foundation models designed for causal effect estimation. For each model, we evaluate the standalone setting using the original covariates. In addition, we adopt these foundation models as the backbone to perform our method. The corresponding variants are named tecFM-TabICL, tecFM-DoPFN, tecFM-CausalFM, and tecFM-CausalPFN.

Evaluation Metrics. Following standard practice in individual treatment effect estimation (Hill, 2011), we use the Precision in Estimation of Heterogeneous Effect (PEHE) as the main evaluation metric. PEHE is defined as $\mathrm{PEHE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tau}(\mathbf{x}_i) - \tau(\mathbf{x}_i)\right)^2$, where $\hat{\tau}(\mathbf{x}_i) = \hat{\mu}_1(\mathbf{x}_i) - \hat{\mu}_0(\mathbf{x}_i)$ denotes the estimated individual treatment effect and $\tau(\mathbf{x}_i) = \mu_1(\mathbf{x}_i) - \mu_0(\mathbf{x}_i)$ denotes the ground-truth treatment effect. A smaller PEHE indicates more accurate estimation of heterogeneous treatment effects.

5.2 RESULTS ON SIMULATION AND SEMI-SYNTHETIC DATA

In this section, we present the experimental results on both simulation and semi-synthetic datasets. For more detailed information on the settings of the data sets, please refer to the appendix B.

Simulation data. We construct a simulation dataset with 5,000 samples and 30 covariates to evaluate treatment effect estimation under different levels of confounding. The parameter α controls the strength of confounding in the treatment assignment mechanism, with larger values indicating stronger treatment selection bias. For each value of α, we conduct 10 independent experiments. As shown in Table 1, tecFM achieves consistently strong performance across different values of α, outperforming conventional baselines such as KNN, MLP, and CFR. This demonstrates that the learned balanced representation preserves information relevant to potential outcome prediction. We further observe that the standalone tabular models exhibit substantial performance variation, whereas integrating them into our framework generally leads to lower estimation errors, particularly for TabICL, DoPFN, CausalFM, and CausalPFN. The task-specific model exploits task-relevant information from the context data, while the foundation model retains its strong generalization ability, and the gating mechanism adaptively combines their complementary predictions to improve potential outcome and treatment effect estimation.

Semi-synthetic data. We evaluate tecFM on two semi-synthetic benchmarks, IHDP and ACIC 2016. IHDP contains 747 samples with 25 covariates and includes 100 realizations, while ACIC 2016 contains 4,802 samples with 58 covariates and includes 10 realizations. As shown in Table 2, incorporating different foundation models into tecFM generally reduces the estimation error compared with directly applying the corresponding pretrained models. The improvements are particularly evident for TabICL, DoPFN, and CausalFM, showing that task-specific learning from the current context data can effectively complement the knowledge acquired during pretraining. On IHDP, tecFM-CausalFM achieves the lowest error under the Within setting, while CausalPFN performs

Table 1: Experimental results under different α values for the Out evaluation protocol in terms of $\sqrt{\epsilon_{\text{PEHE}}}$. Results are reported as mean $\pm$ standard deviation. The best and second-best results in each column are shown in bold and underlined, respectively.

Method	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1.0$	$\alpha = 1.2$
KNN	0.6531 ± 0.0107	0.6526 ± 0.0058	0.6635 ± 0.0075	0.6865 ± 0.0059	0.6930 ± 0.0081
MLP	0.2495 ± 0.0034	0.2472 ± 0.0049	0.2575 ± 0.0033	0.2538 ± 0.0035	0.2618 ± 0.0023
BART	0.4128 ± 0.0126	0.3867 ± 0.0123	0.4071 ± 0.0088	0.4237 ± 0.0143	0.4465 ± 0.0118
CFR	0.1934 ± 0.0067	0.1824 ± 0.0102	0.1688 ± 0.0065	0.1903 ± 0.0117	0.1876 ± 0.0098
DESCN	0.3784 ± 0.0176	0.3888 ± 0.0219	0.3858 ± 0.0163	0.4206 ± 0.0215	0.7413 ± 0.0324
CausalOT	1.0270 ± 0.0115	1.0192 ± 0.0115	1.0173 ± 0.0091	1.0109 ± 0.0079	0.7954 ± 0.0210
EsCFR	0.1525 ± 0.0089	0.1528 ± 0.0131	0.1516 ± 0.0069	0.1562 ± 0.0127	0.1892 ± 0.0105
ORIC	0.2775 ± 0.0045	0.2841 ± 0.0070	0.3028 ± 0.0079	0.2945 ± 0.0041	0.2250 ± 0.0047
TabICL	0.3682 ± 0.0217	0.3216 ± 0.0196	0.3923 ± 0.0176	0.4600 ± 0.0109	0.3846 ± 0.0190
DoPFN	1.0312 ± 0.0137	1.0971 ± 0.0229	1.0171 ± 0.0149	1.0802 ± 0.0141	1.0947 ± 0.0191
CausalFM	2.0637 ± 0.0089	2.1486 ± 0.0152	2.1120 ± 0.0131	2.2007 ± 0.0117	2.1837 ± 0.0089
CausalPFN	0.2473 ± 0.0085	0.2778 ± 0.0133	0.2481 ± 0.0072	0.2911 ± 0.0140	0.2705 ± 0.0075
tecFM-TabICL	$\mathbf{0.1207 \pm 0.0029}$	$\mathbf{0.1195 \pm 0.0030}$	$\underline{0.1272 \pm 0.0048}$	$\underline{0.1256 \pm 0.0067}$	0.1249 ± 0.0049
tecFM-DoPFN	0.1621 ± 0.0091	0.1591 ± 0.0140	0.1584 ± 0.0031	0.1480 ± 0.0039	0.2969 ± 0.0185
tecFM-CausalFM	0.1496 ± 0.0053	0.1375 ± 0.0048	0.1356 ± 0.0051	0.1371 ± 0.0050	0.2389 ± 0.0079
tecFM-CausalPFN	$\underline{0.1232 \pm 0.0031}$	$\underline{0.1212 \pm 0.0019}$	$\mathbf{0.1231 \pm 0.0026}$	$\mathbf{0.1233 \pm 0.0034}$	$\mathbf{0.1275 \pm 0.0036}$

Table 2: Experimental results on IHDP and ACIC 2016 in terms of $\sqrt{\epsilon_{\text{PEHE}}}$. Results are reported as mean $\pm$ standard deviation. The best and second-best results in each column are shown in bold and underlined, respectively.

Method	IHDP		ACIC 2016	
	Within	Out	Within	Out
KNN	2.0178 ± 0.2024	3.4326 ± 0.5231	4.3033 ± 0.2821	3.9666 ± 0.6879
MLP	5.0052 ± 0.8262	2.3936 ± 0.1946	2.7184 ± 0.3683	3.1164 ± 0.5006
BART	2.5649 ± 0.3925	2.7597 ± 0.4733	1.9755 ± 0.3005	2.2827 ± 0.7153
CFR	0.7613 ± 0.0448	0.8459 ± 0.0741	1.7712 ± 0.4280	2.0831 ± 0.6841
CausalOT	7.1214 ± 0.7746	7.2179 ± 0.8301	4.6471 ± 0.6229	4.4597 ± 0.6764
DESCN	7.5551 ± 0.9005	7.5850 ± 0.9236	3.7700 ± 0.4693	3.5767 ± 0.4995
EsCFR	1.1887 ± 0.1500	1.4076 ± 0.2199	1.3470 ± 0.2750	$\mathbf{1.1749 \pm 0.1772}$
ORIC	1.0702 ± 0.0532	1.4152 ± 0.1610	2.0592 ± 0.3799	2.0079 ± 0.4018
TabICL	1.0166 ± 0.0805	0.8644 ± 0.0985	3.0961 ± 0.3304	3.8793 ± 0.6463
DoPFN	6.6546 ± 0.9072	6.6775 ± 1.0078	4.8082 ± 0.4325	4.5355 ± 0.4661
CausalFM	8.6269 ± 0.7412	8.7163 ± 0.7900	5.2166 ± 0.4357	4.9885 ± 0.4567
CausalPFN	0.8699 ± 0.0799	$\mathbf{0.5795 \pm 0.0699}$	1.5847 ± 0.3061	1.5203 ± 0.2345
tecFM-TabICL	0.6735 ± 0.0348	0.7317 ± 0.0925	$\underline{1.2514 \pm 0.2067}$	1.3208 ± 0.3843
tecFM-DoPFN	0.8035 ± 0.0591	0.9128 ± 0.0998	1.8558 ± 0.3681	2.0626 ± 0.6795
tecFM-CausalFM	$\mathbf{0.5292 \pm 0.0270}$	0.6802 ± 0.0827	1.8404 ± 0.4156	2.0369 ± 0.5764
tecFM-CausalPFN	$\underline{0.5925 \pm 0.0255}$	$\underline{0.6109 \pm 0.0579}$	$\mathbf{1.1347 \pm 0.2205}$	$\underline{1.2656 \pm 0.5101}$

best under the Out setting. On ACIC 2016, tecFM-CausalPFN achieves the lowest error under the Within setting and also shows strong performance under the Out setting. These results support our motivation that task-specific information from the current data can complement the generalization ability of foundation models for causal effect estimation. The consistent improvements across different backbones further demonstrate the general applicability of tecFM.

5.3 Ablation Study

Ablation Study on Representation Learning. Table 3 reports the Out evaluation results across five confounding strengths, with 10 repetitions per setting. *w/o* ϕ is the original foundation model, *with* $\hat{\mathcal{W}}$ adds the encoder and Wasserstein balancing, and tecFM w/o $\hat{\mathcal{W}}$ removes balancing from the

Table 3: Ablation results under different α values for the Out evaluation protocol in terms of $\sqrt{\epsilon_{\text{PEHE}}}$. Results are reported as mean $\pm$ standard deviation. The best and second-best results in each column are shown in bold and underlined, respectively.

Method	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1.0$	$\alpha = 1.2$
TabICL-ϕ					
w/o ϕ	0.3682 ± 0.0217	0.3216 ± 0.0196	0.3923 ± 0.0176	0.4600 ± 0.0109	0.3846 ± 0.0190
with $\hat{\mathcal{W}}$	0.1896 ± 0.0050	0.1799 ± 0.0068	0.1855 ± 0.0072	0.1872 ± 0.0041	0.1818 ± 0.0052
tecFM w/o $\hat{\mathcal{W}}$	0.1488 ± 0.0028	0.1492 ± 0.0074	0.1486 ± 0.0030	0.1568 ± 0.0063	0.1731 ± 0.0050
tecFM	$\mathbf{0.1207 \pm 0.0029}$	$\mathbf{0.1195 \pm 0.0030}$	0.1272 ± 0.0048	$\underline{0.1256 \pm 0.0067}$	$\mathbf{0.1249 \pm 0.0049}$
DoPFN-ϕ					
w/o ϕ	1.0312 ± 0.0137	1.0971 ± 0.0229	1.0171 ± 0.0149	1.0802 ± 0.0141	1.0947 ± 0.0191
with $\hat{\mathcal{W}}$	0.2610 ± 0.0075	0.2542 ± 0.0090	0.2762 ± 0.0118	0.2631 ± 0.0105	0.2925 ± 0.0128
tecFM w/o $\hat{\mathcal{W}}$	0.1582 ± 0.0045	0.1698 ± 0.0039	0.1697 ± 0.0075	0.1668 ± 0.0064	0.2878 ± 0.0195
tecFM	0.1621 ± 0.0091	0.1591 ± 0.0140	0.1584 ± 0.0031	0.1480 ± 0.0039	0.2969 ± 0.0185
CausalFM-ϕ					
w/o ϕ	2.0637 ± 0.0089	2.1486 ± 0.0152	2.1120 ± 0.0131	2.2007 ± 0.0117	2.1837 ± 0.0089
with $\hat{\mathcal{W}}$	1.5544 ± 0.0144	1.5369 ± 0.0130	1.5494 ± 0.0092	1.5488 ± 0.0093	1.5864 ± 0.0086
tecFM w/o $\hat{\mathcal{W}}$	0.1503 ± 0.0055	0.1368 ± 0.0049	0.1349 ± 0.0051	0.1374 ± 0.0058	0.2411 ± 0.0092
tecFM	0.1496 ± 0.0053	0.1375 ± 0.0048	0.1356 ± 0.0051	0.1371 ± 0.0050	0.2389 ± 0.0079
CausalPFN-ϕ					
w/o ϕ	0.2473 ± 0.0085	0.2778 ± 0.0133	0.2481 ± 0.0072	0.2911 ± 0.0140	0.2705 ± 0.0075
with $\hat{\mathcal{W}}$	0.1477 ± 0.0044	0.1361 ± 0.0063	0.1365 ± 0.0057	0.1493 ± 0.0047	0.1460 ± 0.0064
tecFM w/o $\hat{\mathcal{W}}$	0.1256 ± 0.0059	0.1260 ± 0.0052	$\underline{0.1241 \pm 0.0052}$	0.1260 ± 0.0065	0.1376 ± 0.0031
tecFM	$\underline{0.1232 \pm 0.0031}$	$\underline{0.1212 \pm 0.0019}$	$\mathbf{0.1231 \pm 0.0026}$	$\mathbf{0.1233 \pm 0.0034}$	$\underline{0.1275 \pm 0.0036}$

complete framework. Adding the encoder and balancing improves the original model, but this variant generally has higher errors than the variants with the task specific predictor and gated aggregation, particularly for DoPFN and CausalFM. Wasserstein balancing brings further gains in several settings, although its effect varies across backbones and confounding strengths. These results suggest that representation alignment alone is insufficient for accurate outcome prediction and benefits from task-specific learning.

Ablation Study Regarding Aggregation. We further conduct ablation studies of the gating mechanism on semi-synthetic and simulation datasets to examine the performance of tabular foundation models under the mean aggregation strategy introduced above, with and without learned representations. Due to space limitations, detailed discussions and experimental results are provided in the Appendix C.2.

More experimental results can be found on Appendix C, including the influence of hyperparameters, other experimental results, and more detailed ablation experiments.

6 Conclusion

In this paper, we investigate how to enhance causal effect foundation models through task-specific model while keeping their parameters frozen. We learn an encoder to align the representations of treated and control groups through Wasserstein distance. To preserve information relevant to outcome prediction during this process, we introduce a lightweight task-specific predictor that provides direct supervision to the encoder. A learnable gate then adaptively combines the predictions of the foundation model and the task-specific predictor. We analyze treatment effect estimation error and conducte experiments with multiple foundation model backbones. Our findings highlight the potential of combining the generalization ability of foundation models with the task-specific adaptability, rather than viewing them as competing paradigms.

AI USE STATEMENT

In this work, we used generative AI tools to search for relevant literature and improve the manuscript's language. 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

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267, 2006.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. 2019.

Vahid Balazadeh Meresht, Hamidreza Kamkari, Valentin Thomas, Junwei Ma, Bingru Li, Jesse Cresswell, and Rahul Krishnan. Causalpfn: Amortized causal effect estimation via in-context learning. *Advances in Neural Information Processing Systems*, 38:154945–154984, 2026.

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

David Bonet, Daniel Mas Montserrat, Xavier Giró-i Nieto, and Alexander G Ioannidis. Hyperfast: Instant classification for tabular data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 11114–11123, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

Alicia Curth and Mihaela Van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International conference on artificial intelligence and statistics*, pp. 1810–1818. PMLR, 2021.

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.

Chris Hays and Manish Raghavan. Double machine learning for causal inference under shared-state interference. *arXiv preprint arXiv:2504.08836*, 2025.

Rundong He and Jieming Shi. Prior-free tabular test-time adaptation. In *International Conference on Learning Representations*, volume 2026, pp. 69478–69494, 2026.

Kai Helli, David Schnurr, Noah Hollmann, Samuel Müller, and Frank Hutter. Drift-resilient tabpfn: In-context learning temporal distribution shifts on tabular data. *Advances in Neural Information Processing Systems*, 37:98742–98781, 2024.

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

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

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

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.

Si-Yang Liu and Han-Jia Ye. Tabpfn unleashed: A scalable and effective solution to tabular classification problems. *arXiv preprint arXiv:2502.02527*, 2025.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Malena Loza, David Chushig-Muzo, Eva Milara, Luis Bote-Curiel, Luis Estrada-Petrocelli, and Felipe Grijalva. Empirical evaluation of out-of-distribution performance of tabular foundation models. *arXiv preprint arXiv:2607.26000*, 2026.

Danni Lu, Chenyang Tao, Junya Chen, Fan Li, Feng Guo, and Lawrence Carin. Reconsidering generative objectives for counterfactual reasoning. *Advances in Neural Information Processing Systems*, 33:21539–21553, 2020.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Jesse Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L Caterini, and Maks Volkovs. Tabdpt: Scaling tabular foundation models on real data. *Advances in Neural Information Processing Systems*, 38:172692–172722, 2026a.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. In *International Conference on Learning Representations*, volume 2026, pp. 79065–79098, 2026b.

Andreas Müller, Carlo Curino, and Raghu Ramakrishnan. Mothernet: Fast training and inference via hyper-network transformers. In *International Conference on Learning Representations*, volume 2025, pp. 76666–76686, 2025.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *International Conference on Machine Learning*, 2025.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. 2026.

Jake Robertson, Noah Hollmann, Samuel Müller, Noor Awad, and Frank Hutter. FairPFN: A tabular foundation model for causal fairness. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51787–51808. PMLR, 13–19 Jul 2025.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-pfn: In-context learning for causal effect estimation. *Advances in Neural Information Processing Systems*, 38:174811–174848, 2026.

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.

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.

Jerzy Splawa-Neyman, Dorota M Dabrowska, and Terrence P Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pp. 465–472, 1990.

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

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

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.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

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.

Yuguang Yan, Haolin Yang, Zecong Chen, Weilin Chen, Ruichu Cai, and Zhifeng Hao. Adjusting prediction model through wasserstein geodesic for causal inference. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 80822–80843, 2026.

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.

Xingxuan Zhang, Gang Ren, Han Yu, Hao Yuan, Hui Wang, Jiansheng Li, Jiayun Wu, Lang Mo, Li Mao, Mingchao Hao, et al. Limix: Unleashing structured-data modeling capability for generalist intelligence. *arXiv preprint arXiv:2509.03505*, 2025.

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.

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 PROOFS OF THEOREMS

Let P_1 and P_0 denote the representation distributions of the treated and control groups, respectively. For any integrable function $m : \mathcal{X} \to \mathbb{R}$, we define the cross group balancing error induced by m as

$$\xi(m) = \left| \int_{\mathcal{X}} m(\mathbf{x})\, dP_1(\mathbf{x}) - \int_{\mathcal{X}} m(\mathbf{x})\, dP_0(\mathbf{x}) \right|. \tag{20}$$

The following result shows that the balancing error is controlled by the Wasserstein discrepancy between the treated and control representations.

Lemma 3 (Balancing Error Bound)**.** *Assume that m satisfies $|m(\mathbf{x}) - m(\mathbf{x}')| \le C(\mathbf{x}, \mathbf{x}'), \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Then,*

$$\xi(m) \le \mathcal{W}(C, P_1, P_0). \tag{21}$$

Lemma 3 shows that the Wasserstein discrepancy between the treated and control representations provides an upper bound on the cross group balancing error. Therefore, reducing $\mathcal{W}(C, P_1, P_0)$ directly limits the error induced by representation imbalance. In the following, we apply this result to the potential outcome prediction loss and analyze its effect on treatment effect estimation.

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

$$\xi(m) = \left\| E_{x \sim P_1(x)} m(x) - E_{x \sim P_0(x)} m(x) \right\| = \left\| \int_{\mathcal{X}} m(x) dP_1(x) - \int_{\mathcal{X}} m(x) dP_0(x) \right\|. \tag{22}$$

Under assumption of Theorem 1, the function m is dominated by the transport cost C, e, $|m(x) - m(x')| \le C(x, x')$. Therefore, the discrepancy in Eq. (22) is upper bounded by its worst-case value over all such admissible functions:

$$\xi(m) \le \sup_{|m(x) - m(x')| \le C(x, x')} \left\{ \int_{\mathcal{X}} m(x) dP_1(x) - \int_{\mathcal{X}} m(x) dP_0(x) \right\}. \tag{23}$$

By the Kantorovich duality of optimal transport (Villani, 2021), the above dual problem is upper bounded by the corresponding primal transport problem:

$$\sup_{|m(x) - m(x')| \le C(x, x')} \left\{ \int_{\mathcal{X}} m(x) dP_1(x) - \int_{\mathcal{X}} m(x) dP_0(x) \right\} \le \inf_{\pi \in \Pi(P_1, P_0)} \int_{\mathcal{X} \times \mathcal{X}} C(x, x') d\pi(x, x') \tag{24}$$

$$= \mathcal{W}(C, P_1, P_0). \tag{25}$$

Combining the above inequalities gives

$$\xi(m) \le \mathcal{W}(C, P_1, P_0). \tag{26}$$

□

A.1 PROOF OF THEOREM 1

Theorem 1. *The individual treatment effect estimation error is jointly controlled by the factual outcome estimation errors and the Wasserstein discrepancy:*

$$\epsilon_{\text{PEHE}} \le 2 \left(\mathcal{E}_{1,h}^{(1)} + \mathcal{E}_{1,f}^{(1)} + \mathcal{E}_{0,h}^{(0)} + \mathcal{E}_{0,f}^{(0)} \right) + \mathcal{O}(\mathcal{W}(C, P_1, P_0)). \tag{27}$$

Proof.

$$\begin{aligned}
\epsilon_{\text{PEHE}} &= \int_{\mathcal{X}} \left(\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x}) \right)^2 p(\mathbf{x})\, d\mathbf{x} \\
&= \int_{\mathcal{X}} \left[\left(\tilde{h}_1(\mathbf{x}) - \mu_1(\mathbf{x}) \right) - \left(\tilde{h}_0(\mathbf{x}) - \mu_0(\mathbf{x}) \right) \right]^2 p(\mathbf{x})\, d\mathbf{x} \\
&\le 2 \int_{\mathcal{X}} \left(\tilde{h}_1(\mathbf{x}) - \mu_1(\mathbf{x}) \right)^2 p(\mathbf{x})\, d\mathbf{x} + 2 \int_{\mathcal{X}} \left(\tilde{h}_0(\mathbf{x}) - \mu_0(\mathbf{x}) \right)^2 p(\mathbf{x})\, d\mathbf{x} \\
&= 2 \int_{\mathcal{X}} \left(\tilde{h}_1(\mathbf{x}) - \mu_1(\mathbf{x}) \right)^2 p(\mathbf{x} \mid T = 1)\, d\mathbf{x} + 2 \int_{\mathcal{X}} \left(\tilde{h}_1(\mathbf{x}) - \mu_1(\mathbf{x}) \right)^2 p(\mathbf{x} \mid T = 0)\, d\mathbf{x} \\
&\quad + 2 \int_{\mathcal{X}} \left(\tilde{h}_0(\mathbf{x}) - \mu_0(\mathbf{x}) \right)^2 p(\mathbf{x} \mid T = 1)\, d\mathbf{x} + 2 \int_{\mathcal{X}} \left(\tilde{h}_0(\mathbf{x}) - \mu_0(\mathbf{x}) \right)^2 p(\mathbf{x} \mid T = 0)\, d\mathbf{x}.
\end{aligned}$$

For each treatment condition t, the squared prediction error of the gated model can be expanded as

$$\begin{aligned}
&\left(\tilde{h}_t(\mathbf{x}) - \mu_t(\mathbf{x})\right)^2 \\
&= \left[g_t(\mathbf{x})h_t(\mathbf{x}) + (1 - g_t(\mathbf{x}))\, f_t^*(\mathbf{x}) - \mu_t(\mathbf{x})\right]^2 \\
&= \left[g_t(\mathbf{x})\,(h_t(\mathbf{x}) - \mu_t(\mathbf{x})) + (1 - g_t(\mathbf{x}))\,(f_t^*(\mathbf{x}) - \mu_t(\mathbf{x}))\right]^2 \\
&= g_t^2(\mathbf{x})\,(h_t(\mathbf{x}) - \mu_t(\mathbf{x}))^2 \\
&\quad + (1 - g_t(\mathbf{x}))^2\,(f_t^*(\mathbf{x}) - \mu_t(\mathbf{x}))^2 \\
&\quad + 2g_t(\mathbf{x})\,(1 - g_t(\mathbf{x}))\,(h_t(\mathbf{x}) - \mu_t(\mathbf{x}))\,(f_t^*(\mathbf{x}) - \mu_t(\mathbf{x})) \\
&= g_t(\mathbf{x})\,(h_t(\mathbf{x}) - \mu_t(\mathbf{x}))^2 \\
&\quad + (1 - g_t(\mathbf{x}))\,(f_t^*(\mathbf{x}) - \mu_t(\mathbf{x}))^2 \\
&\quad - g_t(\mathbf{x})\,(1 - g_t(\mathbf{x}))\,(h_t(\mathbf{x}) - f_t^*(\mathbf{x}))^2 .
\end{aligned} \tag{28}$$

We may define

$$\Delta_t^{(t')} = \int_{\mathcal{X}} g_t(\mathbf{x})\,(1 - g_t(\mathbf{x}))\,(h_t(\mathbf{x}) - f_t^*(\mathbf{x}))^2\; dP_{t'}(\mathbf{x}). \tag{29}$$

Substituting Eq. equation 28 into the four integrals gives

$$\begin{aligned}
\epsilon_{\text{PEHE}} \le\; & 2\left(\mathcal{E}_{1,h}^{(1)} + \mathcal{E}_{1,f}^{(1)} - \Delta_1^{(1)}\right) + 2\left(\mathcal{E}_{1,h}^{(0)} + \mathcal{E}_{1,f}^{(0)} - \Delta_1^{(0)}\right) \\
& + 2\left(\mathcal{E}_{0,h}^{(1)} + \mathcal{E}_{0,f}^{(1)} - \Delta_0^{(1)}\right) + 2\left(\mathcal{E}_{0,h}^{(0)} + \mathcal{E}_{0,f}^{(0)} - \Delta_0^{(0)}\right).
\end{aligned} \tag{30}$$

□

Moreover, since $g_t(\mathbf{x}) \in (0, 1)$, we have $\Delta_t^{(t')} \ge 0$ for any $t, t' \in \{0, 1\}$. Under the Theorem 1, it is easy to prove that $\mathcal{E}_{0,h}^{(1)} \le \mathcal{W}(C, P_1, P_0)$, so that:

$$\epsilon_{\text{PEHE}} \le 2\left(\mathcal{E}_{1,h}^{(1)} + \mathcal{E}_{1,f}^{(1)} + \mathcal{E}_{0,h}^{(0)} + \mathcal{E}_{0,f}^{(0)}\right) + \mathcal{O}(\mathcal{W}(C, P_1, P_0)). \tag{31}$$

A.2 Proof of Theorem 2

To prove Theorem 2, we first introduce two lemmas. Lemma 4 provides a concentration bound for empirical measures, while Lemma 5 relates the population Wasserstein discrepancy to its empirical counterpart.

Lemma 4. *((Bolley et al., 2007), Theorem 1.1). Let μ be a probability measure on $\mathbb{R}^d$ satisfying the $T_1(\zeta)$ inequality, and let*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i} \tag{32}$$

be its empirical measure. For any $d' > d$ and $\zeta' < \zeta$, there exists a constant n_0 such that, for any $\epsilon > 0$ and $n \ge n_0 \max(\epsilon^{-(d'+2)}, 1)$,

$$\mathbb{P}\left[\mathcal{W}(\mu, \hat{\mu}) > \epsilon\right] \le \exp\left(-\frac{\zeta' n \epsilon^2}{2}\right). \tag{33}$$

Lemma 5. *For any $\delta \in (0, 1)$, with probability at least $1 - 2\delta$,*

$$\mathcal{W}(C, P_1, P_0) \le \mathcal{W}(C, \hat{P}_1, \hat{P}_0) + 2\sqrt{\frac{2\log(1/\delta)}{\zeta' n}}. \tag{34}$$

Proof. By the triangle inequality of the Wasserstein metric,

$$\begin{aligned}
\mathcal{W}(C, P_1, P_0) &\le \mathcal{W}(C, P_1, \hat{P}_1) + \mathcal{W}(C, \hat{P}_1, P_0) \\
&\le \mathcal{W}(C, P_1, \hat{P}_1) + \mathcal{W}(C, \hat{P}_1, \hat{P}_0) + \mathcal{W}(C, \hat{P}_0, P_0).
\end{aligned} \tag{35}$$

By Lemma 4, with probability at least $1-\delta$,

$$\mathcal{W}(C, P_1, \hat{P}_1) \leq \sqrt{\frac{2\log(1/\delta)}{\zeta' n}}, \tag{36}$$

$$\mathcal{W}(C, \hat{P}_0, P_0) \leq \sqrt{\frac{2\log(1/\delta)}{\zeta' n}}. \tag{37}$$

Applying the union bound, both inequalities hold simultaneously with probability at least $1-2\delta$. Substituting them into Eq. equation 35 gives

$$\mathcal{W}(C, P_1, P_0) \leq \mathcal{W}(C, \hat{P}_1, \hat{P}_0) + 2\sqrt{\frac{2\log(1/\delta)}{\zeta' n}}. \tag{38}$$

□

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

$$\epsilon_{\text{PEHE}} \leq 2\left(\mathcal{E}^{(1)}_{1,h} + \mathcal{E}^{(1)}_{1,f} + \mathcal{E}^{(0)}_{0,h} + \mathcal{E}^{(0)}_{0,f}\right) + \mathcal{O}\Big(\mathcal{W}(C, \hat{P}_1, \hat{P}_0)\Big) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{39}$$

Proof. By Theorem 1, the PEHE satisfies

$$\epsilon_{\text{PEHE}} \leq 2\left(\mathcal{E}^{(1)}_{1,h} + \mathcal{E}^{(1)}_{1,f} + \mathcal{E}^{(0)}_{0,h} + \mathcal{E}^{(0)}_{0,f}\right) + \mathcal{O}(\mathcal{W}(C, P_1, P_0)). \tag{40}$$

By Lemma 5, with probability at least $1-2\delta$,

$$\mathcal{W}(C, P_1, P_0) \leq \mathcal{W}(C, \hat{P}_1, \hat{P}_0) + 2\sqrt{\frac{2\log(1/\delta)}{\zeta' n}}. \tag{41}$$

Substituting this result into the PEHE bound gives

$$\begin{aligned}\epsilon_{\text{PEHE}} \leq\ & 2\left(\mathcal{E}^{(1)}_{1,h} + \mathcal{E}^{(1)}_{1,f} + \mathcal{E}^{(0)}_{0,h} + \mathcal{E}^{(0)}_{0,f}\right) \\ & + \mathcal{O}\Big(\mathcal{W}(C, \hat{P}_1, \hat{P}_0)\Big) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right).\end{aligned} \tag{42}$$

Therefore, the population PEHE is bounded by the gated factual prediction errors, the empirical Wasserstein discrepancy between the treated and control groups, and a finite sample estimation term. □

B Dataset Setting

Simulation. We construct a simulation dataset with 5,000 samples and 30 covariates to evaluate treatment effect estimation under different levels of confounding. The covariates are generated from a correlated multivariate Gaussian distribution $X_i \sim \mathcal{N}(0, \Sigma)$, where $\Sigma_{jk} = 0.3^{|j-k|}$. Treatment assignment is determined by the nonlinear score $s(X) = 0.8X_1 - 0.6X_2 + 0.5X_3 + 0.4X_4X_5 - 0.3X_6^2 + 0.2\sin(X_7)$, with propensity score $e(X) = \text{clip}(\sigma(\alpha s(X)), 0.03, 0.97)$, where $\alpha \in \{0.4, 0.6, 0.8, 1.0, 1.2\}$ controls the confounding strength. The untreated potential outcome is generated by $\mu_0(X) = 1 + 0.5X_1 - 0.3X_2 + 0.2X_3^2 + \sin(X_4) + 0.3X_5X_6 - 0.2X_9$, while the individual treatment effect is defined as $\tau(X) = 1 + 0.5X_1 - 0.4X_3 + 0.3X_7^2 + 0.2\sin(X_8) + 0.1X_{10}X_{11}$, and $\mu_1(X) = \mu_0(X) + \tau(X)$. The potential outcomes are then given by $Y(0) = \mu_0(X) + \epsilon_0$ and $Y(1) = \mu_1(X) + \epsilon_1$, where $\epsilon_0, \epsilon_1 \sim \mathcal{N}(0, 0.1^2)$. For each value of α, we generate 10 realizations to evaluate the stability of different methods as confounding increases and overlap decreases.

ACIC 2016. ACIC 2016 is a widely used benchmark for causal effect estimation(Dorie et al., 2019). Its covariates are derived from real observational data, while treatment assignments and potential

Table 4: Experimental results under different α values for the Within evaluation protocol in terms of $\sqrt{\epsilon_{\text{PEHE}}}$. Results are reported as mean $\pm$ standard deviation. The best and second-best results in each column are shown in bold and underlined, respectively.

Method	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1.0$	$\alpha = 1.2$
KNN	0.8205 ± 0.0034	0.8388 ± 0.0028	0.8513 ± 0.0041	0.8608 ± 0.0035	0.8662 ± 0.0045
MLP	0.2349 ± 0.0012	0.2346 ± 0.0017	0.2406 ± 0.0027	0.2382 ± 0.0026	0.2447 ± 0.0031
BART	0.4056 ± 0.0068	0.3993 ± 0.0049	0.4183 ± 0.0112	0.4268 ± 0.0088	0.4275 ± 0.0063
CFR	0.1905 ± 0.0075	0.1769 ± 0.0099	0.1683 ± 0.0056	0.1869 ± 0.0102	0.1877 ± 0.0096
DESCN	0.3742 ± 0.0161	0.3838 ± 0.0149	0.3922 ± 0.0196	0.4133 ± 0.0198	0.7362 ± 0.0139
CausalOT	1.0227 ± 0.0066	1.0181 ± 0.0046	1.0137 ± 0.0084	1.0119 ± 0.0050	0.7972 ± 0.0072
EsCFR	0.1527 ± 0.0085	0.1544 ± 0.0144	0.1540 ± 0.0077	0.1574 ± 0.0129	0.1805 ± 0.0087
ORIC	0.2830 ± 0.0044	0.2930 ± 0.0067	0.3002 ± 0.0068	0.2938 ± 0.0029	0.2116 ± 0.0034
TabICL	0.1403 ± 0.0012	0.1407 ± 0.0016	0.1458 ± 0.0010	0.1522 ± 0.0032	$\underline{0.1441 \pm 0.0009}$
DoPFN	1.0450 ± 0.0025	1.0807 ± 0.0032	1.0317 ± 0.0030	1.0550 ± 0.0029	1.0758 ± 0.0022
CausalFM	2.0821 ± 0.0043	2.1589 ± 0.0036	2.0951 ± 0.0029	2.2034 ± 0.0036	2.1744 ± 0.0034
CausalPFN	0.1946 ± 0.0014	0.2333 ± 0.0030	0.2140 ± 0.0023	0.2332 ± 0.0009	0.2195 ± 0.0027
tecFM-TabICL	$\underline{0.1248 \pm 0.0032}$	$\underline{0.1244 \pm 0.0024}$	$\underline{0.1310 \pm 0.0048}$	$\underline{0.1294 \pm 0.0066}$	0.1259 ± 0.0030
tecFM-DoPFN	0.1669 ± 0.0082	0.1620 ± 0.0134	0.1633 ± 0.0031	0.1541 ± 0.0041	0.2907 ± 0.0151
tecFM-CausalFM	0.1508 ± 0.0051	0.1380 ± 0.0051	0.1360 ± 0.0059	0.1374 ± 0.0053	0.2363 ± 0.0071
tecFM-CausalPFN	$\mathbf{0.1204 \pm 0.0024}$	$\mathbf{0.1233 \pm 0.0027}$	$\mathbf{0.1228 \pm 0.0026}$	$\mathbf{0.1219 \pm 0.0017}$	$\mathbf{0.1241 \pm 0.0027}$

outcomes are generated according to predefined mechanisms. The dataset contains $4{,}802$ samples and 58 covariates, including continuous and discrete variables with complex dependence structures. The treatment assignment and outcome generation processes involve nonlinear relationships, feature interactions, and varying levels of selection bias, making the dataset suitable for evaluating treatment effect estimation under complex confounding. Since the potential outcome generation mechanism is known, the ground truth individual treatment effects are available for direct evaluation using metrics such as PEHE.

IHDP. The IHDP dataset is derived from the Infant Health and Development Program and is one of the most commonly used benchmarks for individual treatment effect estimation (Hill, 2011). After standard preprocessing, it contains 747 samples and 25 covariates describing characteristics of infants and their mothers. Following the commonly adopted experimental setting, the original covariates and treatment assignments are retained, while potential outcomes are generated using predefined outcome functions, allowing the ground truth individual treatment effects to be obtained. The covariate distributions of the treated and control groups exhibit noticeable differences, making IHDP suitable for evaluating representation learning, distribution balancing, and treatment effect estimation under selection bias and limited sample size.

C More Results

In this section, we will introduce various ablation experiments we conducted on semi-synthetic datasets and simulation datasets.

C.1 Representation ablation study

We further report the representation ablation results under the Within evaluation protocol in Table 5 and Table 6, complementing the Out results discussed in the main text. The experimental settings and ablation variants remain unchanged. Overall, the Within results exhibit a trend consistent with the Out evaluation. Using representation balancing alone yields model dependent improvements, indicating that simply aligning the treated and control representations is not always sufficient for effective potential outcome prediction. In contrast, incorporating the task-specific predictor and gated aggregation consistently improves the performance of most foundation model backbones by exploiting task relevant information from the current data. Further combining task-specific learning with representation balancing generally leads to the strongest or competitive performance across different datasets and confounding strengths. In particular, the complete tecFM achieves stable performance across the simulation settings with varying α, while also providing clear improvements

Table 5: Ablation results under different α values for the Within evaluation protocol in terms of $\sqrt{\epsilon_{\text{PEHE}}}$. Results are reported as mean $\pm$ standard deviation. For each foundation model, results that outperform the corresponding foundation model are shown in bold.

Method	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1.0$	$\alpha = 1.2$
TabICL-ϕ					
w/o ϕ	0.1403 ± 0.0012	0.1407 ± 0.0016	0.1458 ± 0.0010	0.1522 ± 0.0032	0.1441 ± 0.0009
wtih $\hat{\mathcal{W}}$	0.2198 ± 0.0057	0.2099 ± 0.0100	0.2176 ± 0.0081	0.2162 ± 0.0060	0.2070 ± 0.0045
tecFM w/o $\hat{\mathcal{W}}$	0.1512 ± 0.0023	0.1602 ± 0.0043	0.1473 ± 0.0029	0.1586 ± 0.0051	0.1731 ± 0.0043
tecFM	$\mathbf{0.1248 \pm 0.0032}$	$\mathbf{0.1244 \pm 0.0024}$	$\mathbf{0.1310 \pm 0.0048}$	$\mathbf{0.1294 \pm 0.0066}$	$\mathbf{0.1259 \pm 0.0030}$
DoPFN-ϕ					
w/o ϕ	1.0450 ± 0.0025	1.0807 ± 0.0032	1.0317 ± 0.0030	1.0550 ± 0.0029	1.0758 ± 0.0022
with $\hat{\mathcal{W}}$	0.2635 ± 0.0059	0.2611 ± 0.0052	0.2810 ± 0.0106	0.2622 ± 0.0080	0.2809 ± 0.0087
tecFM w/o $\hat{\mathcal{W}}$	$\mathbf{0.1589 \pm 0.0031}$	0.1701 ± 0.0055	0.1721 ± 0.0069	0.1678 ± 0.0061	0.2820 ± 0.0174
tecFM	0.1669 ± 0.0082	$\mathbf{0.1620 \pm 0.0134}$	$\mathbf{0.1633 \pm 0.0031}$	$\mathbf{0.1541 \pm 0.0041}$	$\mathbf{0.2907 \pm 0.0151}$
CausalFM-ϕ					
w/o ϕ	2.0821 ± 0.0043	2.1589 ± 0.0036	2.0951 ± 0.0029	2.2034 ± 0.0036	2.1744 ± 0.0034
with $\hat{\mathcal{W}}$	1.5523 ± 0.0046	1.5586 ± 0.0038	1.5557 ± 0.0031	1.5642 ± 0.0041	1.5515 ± 0.0044
tecFM w/o $\hat{\mathcal{W}}$	0.1519 ± 0.0055	$\mathbf{0.1373 \pm 0.0050}$	$\mathbf{0.1350 \pm 0.0058}$	0.1374 ± 0.0057	$\mathbf{0.2360 \pm 0.0063}$
tecFM	$\mathbf{0.1508 \pm 0.0051}$	0.1380 ± 0.0051	0.1360 ± 0.0059	$\mathbf{0.1374 \pm 0.0053}$	0.2363 ± 0.0071
CausalPFN-ϕ					
w/o ϕ	0.1946 ± 0.0014	0.2333 ± 0.0030	0.2140 ± 0.0023	0.2332 ± 0.0009	0.2195 ± 0.0027
with $\hat{\mathcal{W}}$	0.1534 ± 0.0046	0.1430 ± 0.0043	0.1470 ± 0.0056	0.1529 ± 0.0048	0.1444 ± 0.0042
tecFM w/o $\hat{\mathcal{W}}$	0.1264 ± 0.0055	0.1271 ± 0.0047	0.1306 ± 0.0051	0.1228 ± 0.0053	0.1342 ± 0.0026
tecFM	$\mathbf{0.1204 \pm 0.0024}$	$\mathbf{0.1233 \pm 0.0027}$	$\mathbf{0.1228 \pm 0.0026}$	$\mathbf{0.1219 \pm 0.0017}$	$\mathbf{0.1241 \pm 0.0027}$

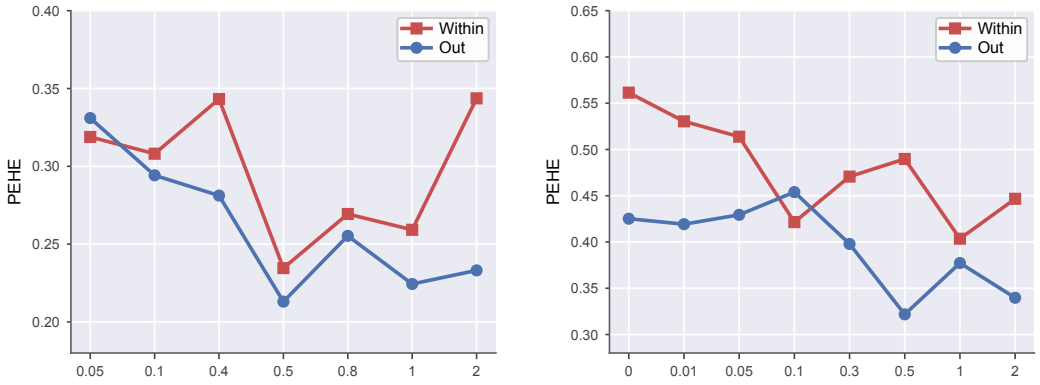


Figure 2: Sensitivity analysis of λ_w and λ_g on IHDP based on a single experimental run. The left and right panels show the sensitivity to the Wasserstein balancing coefficient λ_w and the gating regularization coefficient λ_g, respectively. Red squares and blue circles denote the Within and Out results, respectively.

on IHDP and ACIC 2016 for most backbone models. Although representation balancing does not improve every configuration individually, its combination with task-specific learning produces more robust gains, suggesting that the two components play complementary roles in improving potential outcome prediction. These results further confirm that the benefits of tecFM are not specific to a particular evaluation protocol and remain consistent across different data settings.

Table 6: Ablation results on IHDP and ACIC 2016 in terms of $\sqrt{\epsilon_{\text{PEHE}}}$. Results are reported as mean $\pm$ standard deviation. For each foundation model, the best result in each column is shown in bold.

Method	IHDP		ACIC 2016	
	Within	Out	Within	Out
TabICL-ϕ				
w/o ϕ	1.0166 ± 0.0805	0.8644 ± 0.0985	3.0961 ± 0.3304	3.8793 ± 0.6463
with $\hat{\mathcal{W}}$	1.4235 ± 0.0943	1.1276 ± 0.1418	1.4096 ± 0.2120	1.4643 ± 0.5342
tecFM w/o $\hat{\mathcal{W}}$	0.8961 ± 0.0846	0.9472 ± 0.1301	1.5471 ± 0.2769	1.6494 ± 0.4343
tecFM	$\mathbf{0.6735 \pm 0.0348}$	$\mathbf{0.7317 \pm 0.0925}$	$\mathbf{1.2514 \pm 0.2067}$	$\mathbf{1.3208 \pm 0.3843}$
DoPFN-ϕ				
w/o ϕ	6.6546 ± 0.9072	6.6775 ± 1.0078	4.8082 ± 0.4325	4.5355 ± 0.4661
with $\hat{\mathcal{W}}$	3.5741 ± 0.4139	3.1004 ± 0.4032	2.2135 ± 0.3552	2.3926 ± 0.6869
tecFM w/o $\hat{\mathcal{W}}$	0.8524 ± 0.0560	1.0040 ± 0.1309	1.9873 ± 0.4268	2.2095 ± 0.6851
tecFM	$\mathbf{0.8035 \pm 0.0591}$	$\mathbf{0.9128 \pm 0.0998}$	$\mathbf{1.8558 \pm 0.3681}$	$\mathbf{2.0626 \pm 0.6795}$
CausalFM-ϕ				
w/o ϕ	8.6269 ± 0.7412	8.7163 ± 0.7900	5.2166 ± 0.4357	4.9885 ± 0.4567
with $\hat{\mathcal{W}}$	7.8979 ± 0.7602	7.9818 ± 0.8112	5.2287 ± 0.4174	5.4254 ± 0.6218
tecFM w/o $\hat{\mathcal{W}}$	0.5783 ± 0.0273	0.7435 ± 0.0849	1.8383 ± 0.4352	2.0864 ± 0.6443
tecFM	$\mathbf{0.4694 \pm 0.0241}$	$\mathbf{0.6239 \pm 0.0796}$	$\mathbf{1.7949 \pm 0.4124}$	$\mathbf{1.9841 \pm 0.5401}$
CausalPFN-ϕ				
w/o ϕ	0.8699 ± 0.0799	$\mathbf{0.5795 \pm 0.0699}$	1.5847 ± 0.3061	1.5203 ± 0.2345
with $\hat{\mathcal{W}}$	0.8210 ± 0.0444	0.7194 ± 0.0401	1.2534 ± 0.2633	1.3348 ± 0.5273
tecFM w/o $\hat{\mathcal{W}}$	0.6577 ± 0.0233	0.6904 ± 0.0596	1.3825 ± 0.2953	1.5165 ± 0.5487
tecFM	$\mathbf{0.5925 \pm 0.0255}$	0.6109 ± 0.0579	$\mathbf{1.1347 \pm 0.2205}$	$\mathbf{1.2656 \pm 0.5101}$

C.2 Ablation Study Regarding Aggregation

Table 7 compares prediction aggregation strategies for TabICL and CausalPFN on IHDP and ACIC 2016. TabICL is a general purpose tabular foundation model, while CausalPFN is designed for causal effect estimation. In *w/o* ϕ, the foundation model and task-specific predictor are combined by averaging their predictions without representation learning. In *with* $\hat{\mathcal{W}}$, the encoder learns balanced representations, while the predictions are still averaged. The complete tecFM uses balanced representations and learns a gate that varies the contribution of each predictor across individuals and treatment conditions. On both datasets, balanced mean aggregation improves over *w/o* ϕ, and tecFM achieves the lowest errors under both the Within and Out protocols. These results show the benefits of combining representation balancing with adaptive aggregation.

Table 8 extends this comparison to simulation data under different confounding strengths. For TabICL, tecFM consistently improves over the mean aggregation variants in both the In and Out settings. For CausalPFN, adaptive gating also reduces the error in most settings, while balanced mean aggregation performs slightly better at the strongest confounding level. These results suggest that the relative contributions of the two predictors can vary with the data and treatment condition.

C.3 Sensitivity analysis

In this section, we examine the sensitivity of tecFM to the Wasserstein balancing coefficient λ_w and the gating regularization coefficient λ_g using a single experimental run on IHDP. In the left panel, introducing Wasserstein balancing reduces both Within and Out errors, with both reaching their lowest values around $\lambda_w = 0.5$. Larger values do not bring consistent improvements, suggesting that

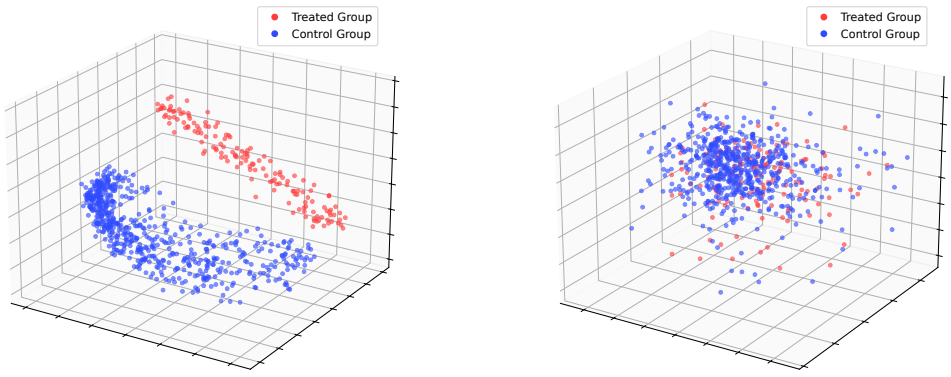
Table 7: Ablation study of different prediction aggregation strategies on IHDP and ACIC 2016 in terms of $\sqrt{\epsilon_{\text{PEHE}}}$. Results are reported as mean $\pm$ standard deviation.

Method	IHDP		ACIC 2016	
	Within	Out	Within	Out
TabICL				
w/o ϕ	2.1354 ± 0.2183	1.2767 ± 0.1476	1.5047 ± 0.2778	1.5281 ± 0.5849
with $\hat{\mathcal{W}}$	2.1227 ± 0.2344	1.2597 ± 0.1571	1.4686 ± 0.2705	1.4907 ± 0.5505
tecFM	$\mathbf{0.6735 \pm 0.0348}$	$\mathbf{0.7317 \pm 0.0925}$	$\mathbf{1.2514 \pm 0.2067}$	$\mathbf{1.3208 \pm 0.3843}$
CausalPFN				
w/o ϕ	1.0654 ± 0.0436	1.0254 ± 0.0597	1.9238 ± 0.2955	2.0523 ± 0.5030
with $\hat{\mathcal{W}}$	0.6709 ± 0.0304	0.7318 ± 0.0695	1.3062 ± 0.2752	1.3542 ± 0.4442
tecFM	$\mathbf{0.5925 \pm 0.0255}$	$\mathbf{0.6109 \pm 0.0579}$	$\mathbf{1.1347 \pm 0.2205}$	$\mathbf{1.2656 \pm 0.5101}$

Table 8: Ablation study of different prediction aggregation strategies on the simulation dataset under different α values in terms of $\sqrt{\epsilon_{\text{PEHE}}}$. Results are reported as mean $\pm$ standard deviation. For each backbone, results that outperform the corresponding original foundation model are shown in bold.

Method	$\alpha = 0.4$		$\alpha = 0.6$		$\alpha = 0.8$		$\alpha = 1.0$		$\alpha = 1.2$	
	Within	Out	Within	Out	Within	Out	Within	Out	Within	Out
TabICL										
w/o ϕ	0.2420 ± 0.0058	0.2605 ± 0.0084	0.2574 ± 0.0096	0.2576 ± 0.0091	0.2493 ± 0.0063	0.2625 ± 0.0063	0.2513 ± 0.0088	0.2455 ± 0.0067	0.2151 ± 0.0055	0.2100 ± 0.0052
with $\hat{\mathcal{W}}$	0.2443 ± 0.0037	0.2619 ± 0.0073	0.2532 ± 0.0064	0.2563 ± 0.0091	0.2534 ± 0.0093	0.2578 ± 0.0087	0.1940 ± 0.0123	0.1949 ± 0.0119	0.1469 ± 0.0050	0.1462 ± 0.0049
tecFM	$\mathbf{0.1248 \pm 0.0032}$	$\mathbf{0.1207 \pm 0.0029}$	$\mathbf{0.1244 \pm 0.0024}$	$\mathbf{0.1195 \pm 0.0030}$	$\mathbf{0.1310 \pm 0.0048}$	$\mathbf{0.1272 \pm 0.0048}$	$\mathbf{0.1294 \pm 0.0066}$	$\mathbf{0.1256 \pm 0.0067}$	$\mathbf{0.1259 \pm 0.0030}$	$\mathbf{0.1249 \pm 0.0049}$
CausalPFN										
w/o ϕ	0.1483 ± 0.0045	0.1501 ± 0.0076	0.1478 ± 0.0044	0.1431 ± 0.0049	0.1546 ± 0.0043	0.1545 ± 0.0036	0.1498 ± 0.0024	0.1480 ± 0.0040	0.2475 ± 0.0050	0.2435 ± 0.0034
with $\hat{\mathcal{W}}$	0.1316 ± 0.0067	0.1303 ± 0.0058	0.1387 ± 0.0073	0.1365 ± 0.0068	0.1390 ± 0.0078	0.1319 ± 0.0074	0.1301 ± 0.0058	0.1238 ± 0.0069	0.1199 ± 0.0034	0.1214 ± 0.0047
tecFM	$\mathbf{0.1204 \pm 0.0024}$	$\mathbf{0.1232 \pm 0.0031}$	$\mathbf{0.1233 \pm 0.0027}$	$\mathbf{0.1212 \pm 0.0019}$	$\mathbf{0.1228 \pm 0.0026}$	$\mathbf{0.1231 \pm 0.0026}$	$\mathbf{0.1219 \pm 0.0017}$	$\mathbf{0.1233 \pm 0.0034}$	$\mathbf{0.1238 \pm 0.0020}$	$\mathbf{0.1275 \pm 0.0036}$

excessive emphasis on distribution alignment may be unhelpful for outcome prediction. In the right panel, performance also varies nonmonotonically with λ_g. A small value gives the lowest Out error, while a moderate value yields the lowest Within error and a relatively low Out error.



(a) Without balancing (b) Wasserstein balancing

Figure 3: Visualization of the transformed representations produced by CausalPFN under different balancing settings.

C.4 Visualization Results

We further visualize the representations produced by CausalPFN on the first realization of ACIC 2016 before and after applying Wasserstein balancing. As shown in Figure 3, without balancing, the treated and control groups remain distinguishable after being processed by the tabular foundation model. After incorporating the proposed representation balancing mechanism, the discrepancy between the two groups is reduced, leading to more aligned representations. This visualization provides empirical

evidence that our method can learn more balanced representations for the treated and control groups before potential outcome prediction.