MULTICFR: COUNTERFACTUAL REGRESSION FOR MULTIMODAL TREATMENT EFFECT ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing individual treatment effect (ITE) estimation methods mainly focus on tabular covariates. However, in practice, tabular data often coexist with text and image modalities, particularly in healthcare. Leveraging such multimodal data can capture richer confounding information and enable more effective confounding adjustment. Only a few methods transfer multimodal fusion techniques to ITE estimation, but overlook explicit tabular semantics and the trade-off between distribution balancing and factual outcome prediction. To address these issues, we propose **MultiCFR**, a counterfactual regression method for multimodal ITE estimation with two key designs. (1) Tabular semantic anchoring uses explicit tabular semantics to guide relevant information extraction from text tokens and image patches. (2) Dual-branch representation learning separates the fused representation into balanced and predictive branches to mitigate confounding bias while preserving useful information for factual outcome prediction. Theoretically, we derive a generalization-error bound that characterizes the trade-off between distribution balancing and predictive information preservation. Empirically, we construct two semi-synthetic multimodal datasets for standardized evaluation, MIMIC-TIT and IHDP-TIT, where MultiCFR consistently outperforms existing baselines with lower ITE estimation errors.

1 INTRODUCTION

Estimating Individual Treatment Effects (ITE) from observational data is important for personalized decision-making in diverse domains such as healthcare and e-commerce (Li et al., 2024; Liu et al., 2021; Feuerriegel et al., 2024; Bica & Van der Schaar, 2022), by quantifying the causal impact of treatments. Existing ITE estimation methods mainly rely on structured tabular covariates to characterize individuals and adjust for confounding (Shalit et al., 2017; Huang et al., 2024; Guo et al., 2026; Cao et al., 2026). However, confounding information may not only be present in tabular data but also in text and images, which capture complementary information about the underlying confounders (Schulte et al., 2025; Deshpande et al., 2022). Especially in healthcare, patient information is inherently multimodal, including not only structured electronic health records (EHRs) but also rich unstructured data. For example, tumor size observed in CT scans and tumor stage appearing in radiology reports influence both treatment selection and health outcomes, acting as multimodal confounders. Ignoring such information can lead to incomplete individual representations and insufficient confounding adjustment, ultimately limiting the accuracy of ITE estimation.

Despite the necessity of accounting for multimodal confounding, only a few methods have explored non-tabular data for ITE estimation (Wu et al., 2024; Schulte et al., 2025; Spindler et al., 2026). These methods mainly employ separate modality encoders (text, images, tabular) to produce modality-specific embeddings, and then integrate them through cross-modal fusion. However, this fusion strategy fails to exploit the explicit semantic information in tabular data to guide the representation learning of unstructured modalities. For example, a pneumonia diagnosis in EHRs can guide the extraction of related abnormalities from chest X-rays and relevant descriptions from clinical reports. Besides, these methods do not explicitly consider the trade-off between representation balancing and outcome prediction. This is important because multimodal data may contain both treatment-related and outcome-predictive information, and indiscriminately balancing the entire representation may suppress useful predictive information and degrade treatment effect estimation.

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.

To address these limitations, we propose **MultiCFR**, a counterfactual regression method that leverages tabular semantics to guide multimodal representation learning and balances representations while preserving useful information for factual outcome prediction. First, we introduce **tabular semantic anchoring**, which leverages the explicit semantics of structured tabular features to construct semantic anchors. These anchors guide the model to identify and extract relevant information from text tokens and image patches, rather than treating all multimodal features equally during fusion. Second, we perform a **dual-branch representation learning** strategy to balance treatment-group distributions while preserving useful predictive information. An anchor-conditioned soft gate allocates the fused multimodal representation into a balance-regularized branch and a complementary predictive branch. Treatment-group balancing is applied only to the former, while both branches are jointly used for potential-outcome prediction. This design reduces distributional discrepancy in the balance-regularized branch without enforcing balance on the entire representation. **Theoretically**, we derive a generalization-error bound that relates ITE estimation error to factual prediction error and treatment-group discrepancy, revealing the trade-off between distribution balancing and predictive information preservation. **Empirically**, we generate two semi-synthetic multimodal benchmark datasets. MIMIC-TIT integrates patient-level clinical data, radiology reports, and chest X-rays from MIMIC-IV and MIMIC-CXR with semi-synthetic treatment effects, while IHDP-TIT extends IHDP by retaining selected tabular covariates and converting the remaining covariates into corresponding image and text modalities.

Our main contributions are summarized as follows:

- We propose MultiCFR, where tabular semantic anchoring guides relevant information extraction from text tokens and image patches, and dual-branch representation learning coordinates distribution balancing and factual outcome prediction.
- We derive a generalization error bound for ITE estimation that characterizes the trade-off between balancing distributions and preserving useful information for factual outcome prediction.
- We construct two semi-synthetic multimodal datasets for standardized evaluation: MIMIC-TIT and IHDP-TIT. MultiCFR reduces $\sqrt{\epsilon_{\text{PEHE}}}$ by 22.3% and 14.4% over the strongest baselines on the two datasets, respectively.

2 Related Work

In this section, we review two lines of research closely related to our work: tabular treatment effect estimation and multimodal treatment effect estimation.

Tabular Treatment Effect Estimation. Traditional methods commonly use propensity scores or sample reweighting, including propensity score matching, double machine learning, and entropy balancing (Li et al., 2016; Jung et al., 2021; Athey et al., 2018; De Bartolomeis et al., 2025). More recently, representation-learning approaches have sought to mitigate confounding bias and can be broadly categorized into covariate-balancing and adversarial-learning methods (Johansson et al., 2022). Covariate-balancing methods learn representations that reduce distribution discrepancies between treatment groups. CFRNet (Shalit et al., 2017) introduced integral probability metrics, such as MMD and the Wasserstein distance, to balance treated and control distributions. Subsequent studies have extended this paradigm through propensity-based matching, joint treatment–outcome modeling, and sample-level alignment (Huang et al., 2024; Zhong et al., 2022; Yao et al., 2018; Kazemi & Ester, 2024; Zhang et al., 2020). Besides, adversarial-learning methods employ discriminators to align latent treatment-group distributions or facilitate counterfactual outcome estimation. Representative approaches include GANITE, SCIGAN, ABCEI, and CBRE (Yoon et al., 2018; Bica et al., 2020; Du et al., 2021; Zhou et al., 2022). Nevertheless, these methods primarily operate on structured covariates and therefore cannot exploit complementary information embedded in images and text. In many settings, such information is not adequately captured by tabular data but can be crucial for ITE estimation.

Multimodal Treatment Effect Estimation. With the increasing availability of multimodal data, several recent studies have begun to estimate treatment effects under multimodal confounding. DoubleML-Deep employed modality-specific encoders and cross-modal attention to learn a joint

representation for treatment effect estimation within a double machine learning framework (Spindler et al., 2026). Deshpande et al. (2022) proposed DMSE, which modeled images, text, and other observations as proxies for latent confounders through deep structural equations and product-of-experts inference. Feng et al. (2026) proposed MixingInsights, which combined structured variables and text representations for average causal effect estimation. In clinical applications, Bi-AdaIN conditioned imaging representations on tabular variables and reduced their dependence on treatment assignment (Wei et al., 2026), while ONTRAM combined clinical variables and medical images within an interpretable ordinal-outcome model (Herzog et al., 2026). Despite these advances, existing methods neither fully exploit explicit tabular semantics nor consider the trade-off between distribution balancing and factual outcome prediction.

3 PRELIMINARIES

Following the Neyman-Rubin potential outcome framework (Rubin, 2005; Shalit et al., 2017), we formally define the problem setup. Let $A \in \{0, 1\}$ denote the treatment indicator, where $A = 1$ indicates that a unit receives the treatment (e.g., medication) and $A = 0$ indicates that it serves as a control (e.g., placebo). The outcome space is denoted by $\mathcal{Y} \subseteq \mathbb{R}$, and each unit is associated with two potential outcomes $Y(0), Y(1) \in \mathcal{Y}$, corresponding to the outcomes under control and treatment, respectively. In the standard causal inference literature, pretreatment covariates are typically tabular and can be represented as $X^{\text{tab}} = (X_1, \ldots, X_p) \in \mathbb{R}^p$, where $p \in \mathbb{N}$. We extend this setting to multimodal treatment effect estimation. For RGB images, the image covariates can be represented as $X^{\text{img}} \in \mathbb{R}^{3\times h\times w}$, where $h, w \in \mathbb{N}$. For textual covariates, X^{txt} can be represented as an input matrix containing token, segmentation, and position information, such that $X^{\text{txt}} \in \mathbb{R}^{3\times S}$, where $S \in \mathbb{N}$ denotes the sequence length. When all modalities are jointly used as pretreatment covariates, we denote the multimodal covariates by $\mathcal{X} = \left(X^{\text{tab}}, X^{\text{txt}}, X^{\text{img}}\right)$, where X^{tab}, X^{txt}, and X^{img} denote the tabular, textual, and image inputs, respectively. For a unit assigned to treatment $A = a$, only the factual outcome $Y(a)$ is observed, while the counterfactual outcome $Y(1 - a)$ remains unobserved. The observed outcome is therefore given by $Y = (1-A)Y(0)+AY(1)$. Our goal is to estimate the ITE and evaluate its accuracy using the metric *Precision in Estimation of Heterogeneous Effect* (PEHE). Relevant definitions and identification assumptions are provided in Appendix A.

4 METHOD

We introduce **MultiCFR**, a novel counterfactual regression method for ITE estimation under multimodal confounding. MultiCFR consists of two key components: (1) tabular semantic anchoring for leveraging the feature semantics of tabular covariates to guide relevant information extraction from unstructured modalities, and (2) dual-branch representation learning for balancing treatment-group distributions while retaining complementary predictive information. Figure 1 illustrates the overall framework of MultiCFR.

4.1 MULTIMODAL REPRESENTATION ENCODING

MultiCFR first encodes tabular, textual, and visual inputs into modality-specific token sequences. An individual-specific tabular semantic anchor is then used as a query to retrieve relevant text tokens and image patches, which further refine the fused multimodal representation.

Tabular tokenization. Let x_{ij}^{tab} denote the standardized value of the jth tabular covariate and $m_{ij} \in \{0, 1\}$ its missingness indicator. We construct one token for each covariate:

$$\mathbf{h}_{ij}^{\text{tab}} = x_{ij}^{\text{tab}}\mathbf{v}_j + m_{ij}\mathbf{u}_j + \mathbf{e}_j, \qquad j = 1, \ldots, p, \tag{1}$$

where $\mathbf{v}_j, \mathbf{u}_j, \mathbf{e}_j \in \mathbb{R}^d$ are learnable value, missingness, and feature-identity embeddings, respectively. The resulting tabular token matrix is denoted by $\mathbf{H}_i^{\text{tab}} \in \mathbb{R}^{p\times d}$.

Multimodal encoding and fusion. Text and image inputs are processed by modality-specific encoders and projected into the same d-dimensional representation space, yielding token-level text features $\mathbf{H}_i^{\text{txt}}$ and patch-level image features $\mathbf{H}_i^{\text{img}}$. The three modality-specific token sequences

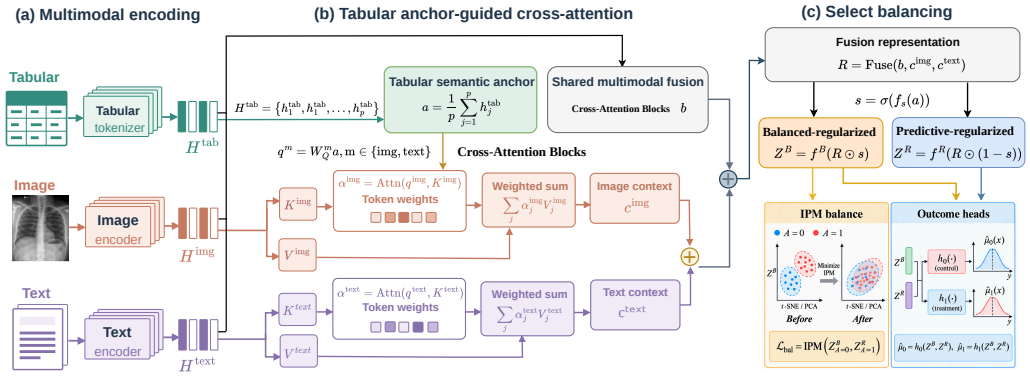

Figure 1: Overview of the MultiCFR framework. Multimodal encoders extract tabular, image, and text representations. The tabular representation forms an individual-specific semantic anchor to guide cross-modal information retrieval, while the fused representation is separated into balance-regularized and predictive branches for selective balancing and potential-outcome prediction.

are then integrated through cross-modal attention and attention pooling to obtain a base multimodal representation:

$$\mathbf{b}_i = \mathrm{Fusion}\left(\mathbf{H}_i^{\mathrm{tab}}, \mathbf{H}_i^{\mathrm{txt}}, \mathbf{H}_i^{\mathrm{img}}\right) \in \mathbb{R}^d. \tag{2}$$

Detailed modality-specific representation learning and cross-modal fusion operations are provided in Appendix B.1.

4.2 Tabular Semantic Anchoring

Standard multimodal fusion treats the three modalities equally during fusion, although their semantic structures are different. Tabular covariates have explicit and stable feature identities, whereas relevant information in text and images is distributed across many tokens or patches. MultiCFR therefore uses the structured tabular representation as an individual-specific semantic reference to retrieve relevant information from unstructured modalities.

We aggregate the tabular tokens into the semantic anchor

$$\boldsymbol{\alpha}_i = \frac{1}{p}\sum_{j=1}^{p} \mathbf{h}_{ij}^{\mathrm{tab}} \in \mathbb{R}^d. \tag{3}$$

Unlike a sample-invariant query shared by all individuals, $\boldsymbol{\alpha}_i$ varies across individuals because it depends on their observed tabular values and missingness patterns. It therefore carries both feature semantics and individual-specific pretreatment information.

For each unstructured modality $m \in \{\mathrm{txt}, \mathrm{img}\}$, the anchor is transformed into a query and used to retrieve a modality-specific context:

$$\mathbf{q}_i^m = q_m(\boldsymbol{\alpha}_i), \tag{4}$$
$$\widetilde{\mathbf{c}}_i^m = \mathrm{MHA}_m(\mathbf{q}_i^m, \mathrm{LN}(\mathbf{H}_i^m), \mathrm{LN}(\mathbf{H}_i^m)), \tag{5}$$
$$\mathbf{c}_i^m = o_m(\widetilde{\mathbf{c}}_i^m). \tag{6}$$

Equivalently, for one attention head, the retrieval weights can be written as

$$\boldsymbol{\pi}_i^m = \mathrm{softmax}\left(\frac{\mathbf{q}_i^m(\mathbf{K}_i^m)^\top}{\sqrt{d_k}}\right), \qquad \widetilde{\mathbf{c}}_i^m = \boldsymbol{\pi}_i^m \mathbf{V}_i^m, \tag{7}$$

where invalid text positions are masked before normalization and o_m is a layer-normalization–linear–GELU output projection. Thus, different patients may retrieve different textual tokens or image regions even when the same modality encoder is used.

We further compute an anchor-conditioned scalar gate to adaptively control the contribution of each retrieved modality-specific context in multimodal fusion:

$$g_i^m = \sigma(f_m([\boldsymbol{\alpha}_i; \mathbf{c}_i^m])), \qquad m \in \{\text{txt}, \text{img}\}, \tag{8}$$

where $[\cdot;\cdot]$ denotes concatenation. The gated text and image contexts are projected and added to the base multimodal representation, followed by layer normalization:

$$\mathbf{r}_i = \mathrm{LN}\Big(\mathbf{b}_i + g_i^{\text{txt}} \mathbf{W}_\Delta^{\text{txt}} \mathbf{c}_i^{\text{txt}} + g_i^{\text{img}} \mathbf{W}_\Delta^{\text{img}} \mathbf{c}_i^{\text{img}}\Big). \tag{9}$$

In this way, $\mathbf{b}_i$ provides a summary of the three modalities, while the gated contexts incorporate individual-specific text and image information selected by the tabular anchor.

4.3 Dual-Branch Representation Learning

Applying a balancing penalty to the entire multimodal representation may suppress information that is useful for factual outcome prediction. To better control the trade-off between representation balancing and predictive information preservation, MultiCFR softly allocates the anchor-refined representation into two complementary branches. Specifically, we compute a feature-wise allocation gate from the individual-specific semantic anchor:

$$\mathbf{s}_i = \sigma(\mathbf{W}_s \,\mathrm{LN}(\boldsymbol{\alpha}_i) + \mathbf{b}_s), \qquad \mathbf{s}_i \in (0,1)^d. \tag{10}$$

Based on this gate, we construct a balance-regularized representation $\mathbf{z}_i^B$ and a complementary predictive representation $\mathbf{z}_i^R$:

$$\mathbf{z}_i^B = f_B(\mathbf{r}_i \odot \mathbf{s}_i), \tag{11}$$

$$\mathbf{z}_i^R = f_R(\mathbf{r}_i \odot (\mathbf{1} - \mathbf{s}_i)), \tag{12}$$

where $\odot$ denotes element-wise multiplication, and f_B and f_R are separate nonlinear projections. The complementary masks softly control how the anchor-refined multimodal information is allocated to the two branches.

The two branches are trained with different regularization objectives. Specifically, $\mathbf{z}^B$ is optimized by both factual outcome prediction and treatment-group balancing, while $\mathbf{z}^R$ is optimized only through factual outcome prediction. In this way, MultiCFR avoids imposing the balancing constraint on the entire multimodal representation, allowing complementary predictive information to be retained for potential-outcome estimation.

4.4 Balancing and Counterfactual Regression

Given the dual-branch representations, MultiCFR applies distribution balancing only to the balance-regularized branch $\mathbf{z}^B$. Specifically, we define the dimension-normalized representation $w_i = z_i^B/\sqrt{d}$ and minimize the maximum mean discrepancy (MMD) (Gretton et al., 2012; Shalit et al., 2017) between the treated and control distributions:

$$\mathcal{L}_{\text{bal}} = \mathrm{MMD}\left(\{w_i : A_i = 1\}, \{w_i : A_i = 0\}\right). \tag{13}$$

The two branches are jointly used for potential-outcome prediction:

$$\widehat{\mu}_a(\mathcal{X}_i) = h_a([\mathbf{z}_i^B; \mathbf{z}_i^R]), \qquad a \in \{0,1\}. \tag{14}$$

Since only the factual outcome is observed, the prediction loss is

$$\mathcal{L}_{\text{f}} = \frac{1}{N}\sum_{i=1}^{N} \left(Y_i - \widehat{\mu}_{A_i}(\mathcal{X}_i)\right)^2. \tag{15}$$

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{f}} + \lambda_{\text{bal}}\mathcal{L}_{\text{bal}} + \lambda_{\text{reg}}\|\Theta\|_2^2, \tag{16}$$

where λ_{bal} and λ_{reg} control the balancing and regularization strengths, respectively. After training, the ITE is estimated as $\widehat{\tau}_i = \widehat{\mu}_1(\mathcal{X}_i) - \widehat{\mu}_0(\mathcal{X}_i)$. The complete training procedure is provided in Appendix B.2.

5 THEORETICAL ANALYSIS

We analyze the trade-off between distribution balancing and preserving information useful for factual outcome prediction. Following the IPM-based analysis of counterfactual regression (Shalit et al., 2017), we adapt the argument to MultiCFR, where balancing is imposed only on one branch and the full dual-branch representation is used for outcome prediction. Let

$$\mathbf{Z} = [\mathbf{Z}^B; \mathbf{Z}^R], \qquad \mathbf{W} = \mathbf{Z}^B/\sqrt{d}, \tag{17}$$

where $\mathbf{Z}^B$ is balance-regularized and $\mathbf{Z}^R$ provides a complementary predictive pathway. The potential-outcome heads satisfy

$$\widehat{\mu}_a(\mathcal{X}) = h_a(\mathbf{Z}), \qquad \widehat{\tau}(\mathcal{X}) = h_1(\mathbf{Z}) - h_0(\mathbf{Z}). \tag{18}$$

Consistent with the training objective in Section 4.4, MultiCFR minimizes

$$\widehat{\mathcal{L}} = \widehat{\mathcal{L}}_{\mathrm{f}} + \lambda_{\mathrm{bal}}\widehat{\mathrm{MMD}} + \lambda_{\mathrm{reg}}\|\Theta\|_2^2, \tag{19}$$

where the MMD is computed only between the treated and control distributions of $\mathbf{W}$. The L_2 term controls model complexity and is therefore not a causal-discrepancy term in the PEHE bound below.

Balancing–prediction trade-off in ITE generalization.

Lemma 1. *For unrestricted measurable outcome heads and a fixed learned representation* $\mathbf{Z}$*,*

$$\inf_{h_0,h_1} \mathbb{E}\Big[\big(\widehat{\tau}(\mathcal{X}) - \tau(\mathcal{X})\big)^2\Big] = \mathbb{E}[\mathrm{Var}(\tau(\mathcal{X}) \mid \mathbf{Z})]\,. \tag{20}$$

Moreover, since $\mathbf{Z}^B$ *is a component of* $\mathbf{Z}$*,*

$$\mathbb{E}\big[\mathrm{Var}(\tau(\mathcal{X}) \mid \mathbf{Z}^B)\big] - \mathbb{E}[\mathrm{Var}(\tau(\mathcal{X}) \mid \mathbf{Z})] = \mathbb{E}\Big[\big(\mathbb{E}[\tau(\mathcal{X}) \mid \mathbf{Z}] - \mathbb{E}[\tau(\mathcal{X}) \mid \mathbf{Z}^B]\big)^2\Big] \geq 0. \tag{21}$$

Lemma 1 shows that retaining the complementary branch can reduce the irreducible ITE error associated with information loss in the learned representation.

Let q_a^B denote the distribution of $\mathbf{W}$ conditional on $A = a$, and let $\pi = \Pr(A = 1)$. Define

$$F_a = \mathbb{E}\big[(Y - h_a(\mathbf{Z}))^2 \mid A = a\big]\,, \tag{22}$$

$$e_a(\S) = \big(h_a(\mathbf{Z}(\S)) - \mu_a(\S)\big)^2\,, \tag{23}$$

$$\ell_a^{\mathrm{F}}(w) = \mathbb{E}[e_a(\mathcal{X}) \mid \mathbf{W} = w, A = a], \tag{24}$$

$$\ell_a^{\mathrm{CF}}(w) = \mathbb{E}[e_a(\mathcal{X}) \mid \mathbf{W} = w, A = 1 - a], \tag{25}$$

$$\Delta_a^B = \mathbb{E}_{\mathbf{W}\sim q_{1-a}^B}\left[\left|\ell_a^{\mathrm{CF}}(\mathbf{W}) - \ell_a^{\mathrm{F}}(\mathbf{W})\right|\right]. \tag{26}$$

Let k be a positive-definite kernel with RKHS $\mathcal{H}_k$. Assume $\ell_a^{\mathrm{F}} \in \mathcal{H}_k$ and $\|\ell_a^{\mathrm{F}}\|_{\mathcal{H}_k} \leq B_a$. Write $d_B = \mathrm{MMD}(q_0^B, q_1^B)$.

Theorem 2. *Let* $\widehat{\tau}(\mathcal{X}) = h_1(\mathbf{Z}) - h_0(\mathbf{Z})$ *denote the estimated ITE based on the dual-branch representation* $\mathbf{Z} = [\mathbf{Z}^B; \mathbf{Z}^R]$*, and let* $d_B = \mathrm{MMD}(q_0^B, q_1^B)$ *denote the treatment-group discrepancy of the balance-regularized branch. For* $a \in \{0, 1\}$*, let* F_a *denote the factual prediction error,* Δ_a^B *the residual discrepancy, and* B_a *the corresponding RKHS norm bound. Under consistency, multimodal ignorability, overlap, and the preceding RKHS condition, the PEHE is bounded by*

$$\epsilon_{\mathrm{PEHE}} \leq 2(F_0 + F_1) + 2\big[\pi B_0 + (1 - \pi)B_1\big]d_B + 2\big[\pi\Delta_0^B + (1 - \pi)\Delta_1^B\big]. \tag{27}$$

Eq. 27 theoretically supports our motivation that stronger distribution balancing does not necessarily lead to more accurate ITE estimation. Although reducing the MMD of W decreases the distribution-discrepancy term, a smaller discrepancy does not necessarily yield a lower PEHE when factual prediction error increases. This motivates the dual-branch design of MultiCFR, which balances $\mathbf{Z}^B$ while retaining $\mathbf{Z}^R$ for potential-outcome prediction. This trade-off is further validated empirically in Section 6.3, where a smaller representation discrepancy does not always correspond to a lower ITE estimation error. Detailed proofs and a conditional analysis of tabular semantic anchoring are provided in Appendix C.

6 EXPERIMENTS

To evaluate the effectiveness of MultiCFR under multimodal confounding, we conduct comprehensive experiments on MIMIC-TIT and IHDP-TIT including motivation validation, performance evaluation, component analysis, and modality analysis.

6.1 DATASETS

MIMIC-TIT: We construct MIMIC-TIT, a semi-synthetic tabular–image–text dataset with 7396 samples from MIMIC-IV (Johnson et al., 2023) and MIMIC-CXR (Johnson et al., 2019). Each sample combines 18 structured variables, a frontal chest radiograph, and a radiology report. The multimodal covariates are drawn from real-world data, while treatment assignments and potential outcomes are generated from pretreatment multimodal data. Details are provided in Appendix D.1.

IHDP-TIT: We extend the widely used semi-synthetic IHDP benchmark (Hill, 2011) to a multimodal setting and construct IHDP-TIT, while preserving the original 747 subjects, treatments, potential outcomes, and 100 simulation replications. Of the 25 original covariates, 9 remain in tabular form, while 8 are transformed into textual descriptions and the remaining 8 into visual representations. Further construction details are provided in Appendix D.2.

6.2 BASELINES AND METRICS

We compare MultiCFR with five representative multimodal causal effect estimation methods: DoubleML-Deep (Spindler et al., 2026), DMSE (Deshpande et al., 2022), Bi-AdaIN (Wei et al., 2026), ONTRAM (Herzog et al., 2026), and MixingInsights (Feng et al., 2026). For a fair comparison, all methods are adapted to our benchmark datasets and evaluated using the same modality inputs and data splits. Following previous studies (Shi et al., 2019; Zhang et al., 2025; Guo et al., 2026), we evaluate treatment effect estimation using the root Precision in Estimation of Heterogeneous Effect ($\sqrt{\epsilon_{\text{PEHE}}}$) and the absolute error in Average Treatment Effect (ϵ_{ATE}). We additionally report the factual Root Mean Squared Error ($\text{RMSE}_{Factual}$) to evaluate factual outcome prediction. Detailed metric definitions are provided in Appendix D.3.

6.3 EXPERIMENTAL RESULTS

We evaluate MultiCFR from four perspectives. First, **motivation validation** examines the effectiveness of tabular-guided information extraction and dual-branch representation learning. Second, **performance evaluation** assesses the overall effectiveness of MultiCFR against existing methods. Third, **component analysis** investigates the contributions of semantic anchoring and dual-branch representation learning. Finally, **modality analysis** evaluates the contributions and complementarity of different modalities. Additional experimental results are provided in the Appendix E.1.

Motivation Validation. (1) **Tabular semantic anchoring.** Figure 2 illustrates how the individual-specific tabular anchor actively guides information extraction from unstructured modalities. For the chest X-ray, the anchor-conditioned patch attention assigns higher weights to the central thoracic region, indicating that tabular semantics influence visual feature extraction rather than being simply concatenated with image features. For the radiology report, the anchor-conditioned token attention emphasizes words that are more relevant to the individual's clinical state.(2) **Dual-branch representation learning.** Figure 3 shows that stronger distribution balance does not necessarily lead to more accurate ITE estimation. Bi-AdaIN achieves the smallest MMD (0.0013) but a higher PEHE, whereas MultiCFR obtains the lowest PEHE (0.338) with a larger MMD (0.0109). This supports balancing only part of the representation while preserving useful information for factual outcome prediction.

Performance Evaluation. We report the multimodal ITE estimation results in Table 1. Key findings include: (1) MultiCFR achieves the best overall performance on both datasets across ITE estimation, ATE estimation, and factual outcome prediction. (2) On MIMIC-TIT, MultiCFR reduces the out-of-sample $\sqrt{\epsilon_{\text{PEHE}}}$, ϵ_{ATE}, and $\text{RMSE}_{\text{Factual}}$ by 22.3%, 34.1%, and 6.1%, respectively, compared with the strongest baseline. The out-of-sample $\sqrt{\epsilon_{\text{PEHE}}}$ increases by only 1.8% relative to

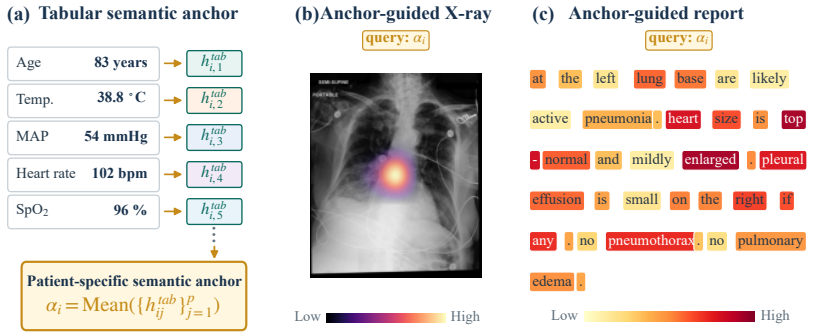

Figure 2: Qualitative visualization of tabular semantic anchoring. (a) Individual-specific anchor constructed from tabular covariates. (b) Anchor-to-patch attention on the chest X-ray. (c) Anchor-to-token attention on the radiology report.

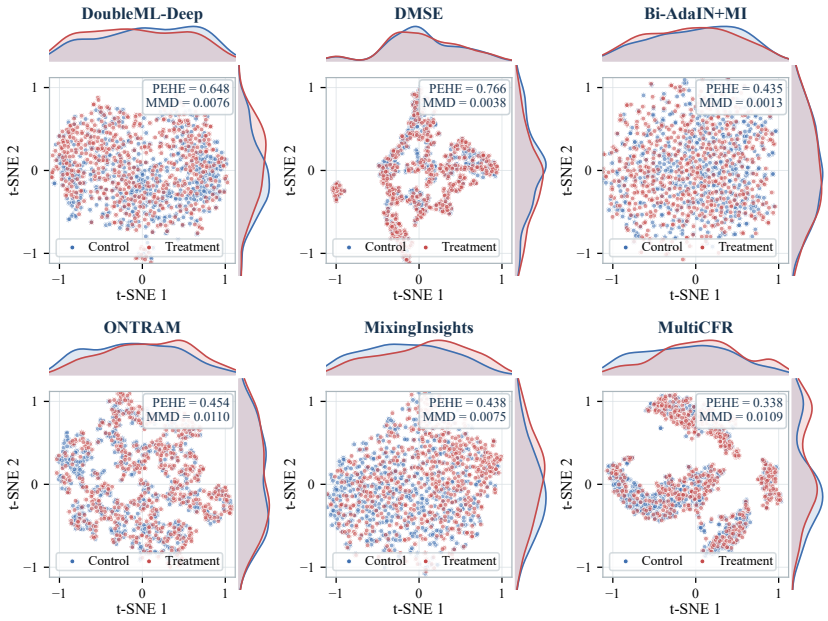

Figure 3: Visualization of treatment–control representation balance and ITE estimation performance across different methods.

the within-sample result, indicating stable generalization. (3) On IHDP-TIT, MultiCFR reduces the out-of-sample $\sqrt{\epsilon_{\text{PEHE}}}$ by 14.4% over MixingInsights and ϵ_{ATE} by 38.8% over DoubleML-Deep, while also achieving the lowest factual RMSE.

Components Analysis. Table 4 in Appendix E.1 evaluates the contribution of individual-specific anchoring and selective balancing on MIMIC-TIT under the out-of-sample setting. Removing the anchor increases $\sqrt{\epsilon_{\text{PEHE}}}$ from 0.338 to 0.354 and ϵ_{ATE} from 0.120 to 0.182, demonstrating the contribution of semantic anchoring to ITE estimation. Selective balancing consistently outperforms both no balancing and global balancing. Compared with no balancing, it reduces $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE} by 10.6% and 38.1%, respectively, while reducing the ATE error by 28.1% over global balancing. These results quantitatively verify the effectiveness of the two key components in MultiCFR.

Modality Analysis. Table 2 compares different modality combinations on MIMIC-TIT and IHDP-TIT. (1) Compared with the tabular-only setting, the full multimodal model reduces $\sqrt{\epsilon_{\text{PEHE}}}$, ϵ_{ATE}, and $\text{RMSE}_{\text{Factual}}$ by 35.2%, 72.1%, and 13.1% on MIMIC-TIT, and by 18.4%, 14.9%, and 18.0%

Table 1: ITE estimation performance on MIMIC-TIT and IHDP-TIT datasets. Results are reported as mean (std). Best results are shown in **bold**.

Method	$\sqrt{\epsilon_{PEHE}^{within}}$	ϵ_{ATE}^{within}	$\text{RMSE}_{Factual}^{within}$	$\sqrt{\epsilon_{PEHE}^{out-of}}$	ϵ_{ATE}^{out-of}	$\text{RMSE}_{Factual}^{out-of}$
			MIMIC-TIT			
DoubleML-Deep	0.520 (0.024)	0.413 (0.030)	1.407 (0.008)	0.648 (0.029)	0.563 (0.033)	1.420 (0.006)
DMSE	0.779 (0.043)	0.695 (0.064)	2.166 (0.380)	0.766 (0.040)	0.678 (0.062)	2.164 (0.387)
Bi-AdaIN	0.335 (0.013)	0.281 (0.024)	1.742 (0.005)	0.435 (0.010)	0.182 (0.018)	1.694 (0.013)
ONTRAM	0.466 (0.026)	0.334 (0.010)	1.321 (0.035)	0.454 (0.030)	0.317 (0.021)	1.283 (0.010)
MixingInsights	0.450 (0.007)	0.303 (0.006)	1.330 (0.001)	0.438 (0.012)	0.288 (0.011)	1.276 (0.001)
MultiCFR	**0.332 (0.031)**	**0.119 (0.020)**	**1.090 (0.036)**	**0.338 (0.032)**	**0.120 (0.021)**	**1.198 (0.008)**
			IHDP-TIT			
DoubleML-Deep	5.555 (0.815)	0.365 (0.051)	5.251 (0.665)	5.712 (0.889)	0.703 (0.131)	5.446 (0.772)
DMSE	5.625 (0.816)	0.858 (0.116)	5.445 (0.701)	5.768 (0.881)	0.880 (0.118)	5.649 (0.806)
Bi-AdaIN	7.825 (0.763)	4.478 (0.168)	4.701 (0.546)	7.913 (0.816)	4.246 (0.140)	6.119 (0.783)
ONTRAM	7.337 (0.772)	3.404 (0.163)	3.890 (0.435)	7.462 (0.844)	3.395 (0.184)	4.431 (0.564)
MixingInsights	3.013 (0.384)	0.644 (0.025)	2.624 (0.320)	3.646 (0.527)	0.782 (0.064)	3.521 (0.490)
MultiCFR	**2.207 (0.324)**	**0.263 (0.046)**	**2.139 (0.245)**	**3.121 (0.545)**	**0.430 (0.077)**	**3.253 (0.483)**

Table 2: Out-of-sample ITE estimation performance with different modality combinations on MIMIC-TIT and IHDP-TIT. Results are reported as mean (std). Best results are shown in **bold**.

Method	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\text{RMSE}_{Factual}$
	MIMIC-TIT		
Tabular	0.522 (0.032)	0.430 (0.026)	1.379 (0.003)
Tabular + Text	0.409 (0.061)	0.251 (0.043)	1.210 (0.008)
Tabular + Image	0.601 (0.079)	0.504 (0.084)	1.398 (0.003)
Tabular + Text + Image (Ours)	**0.338 (0.032)**	**0.120 (0.021)**	**1.198 (0.008)**
	IHDP-TIT		
Tabular	3.824 (0.616)	0.505 (0.083)	3.965 (0.570)
Tabular + Text	3.591 (0.591)	0.534 (0.115)	3.760 (0.547)
Tabular + Image	3.521 (0.576)	0.492 (0.074)	3.632 (0.522)
Tabular + Text + Image (Ours)	**3.121 (0.545)**	**0.430 (0.077)**	**3.253 (0.483)**

on IHDP-TIT, respectively. (2) The contribution of individual modalities varies across datasets: text provides greater gains on MIMIC-TIT, whereas images are more beneficial on IHDP-TIT. Nevertheless, jointly using tabular, text, and image data achieves the best performance on both datasets, demonstrating the benefit of leveraging complementary information beyond tabular covariates.

7 Conclusion

In this paper, we proposed MultiCFR, a novel counterfactual regression framework for ITE estimation under multimodal confounding. MultiCFR combines tabular semantic anchoring with dual-branch representation learning to guide information extraction from unstructured modalities and achieve selective balancing without discarding useful outcome-predictive information. We further provided a theoretical analysis that characterizes the trade-off between balancing distributions and preserving useful information for factual outcome prediction. To support standardized evaluation, we constructed two multimodal benchmark datasets, MIMIC-TIT and IHDP-TIT. Comprehensive experiments, including motivation validation, performance evaluation, component analysis, and modality analysis, demonstrate the effectiveness of the proposed designs and show that MultiCFR achieves more accurate ITE estimation than existing methods.

AI USE STATEMENT

We used ChatGPT in a limited way, primarily for language editing and improving clarity of presentation. LLMs are not used as essential, original, or non-standard components of the research. We take full responsibility for the final content, including all text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):597–623, 2018.

Ioana Bica and Mihaela Van der Schaar. Transfer learning on heterogeneous feature spaces for treatment effects estimation. In *Advances in Neural Information Processing Systems*, pp. 37184–37198, 2022.

Ioana Bica, James Jordon, and Mihaela Van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 16434–16445, 2020.

Fuyuan Cao, Jiaxuan Zhang, and Xiaoli Li. Pite: Multi-prototype alignment for individual treatment effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 19871–19879, 2026.

Xinyan Chen and Jianfei Yang. X-fi: A modality-invariant foundation model for multimodal human sensing. In *International Conference on Learning Representations*, pp. 97378–97397, 2025.

Piersilvio De Bartolomeis, Julia Kostin, Javier Abad, Yixin Wang, and Fanny Yang. Doubly robust identification of treatment effects from multiple environments. In *International Conference on Learning Representations*, pp. 17801–17831, 2025.

Shachi Deshpande, Kaiwen Wang, Dhruv Sreenivas, Zheng Li, and Volodymyr Kuleshov. Deep multi-modal structural equations for causal effect estimation with unstructured proxies. In *Advances in Neural Information Processing Systems*, pp. 10931–10944, 2022.

Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *Data Mining and Knowledge Discovery*, 35(4):1713–1738, 2021.

Lu Feng, Adi Lin, and Zhiyong Luo. Mixinginsights: A framework for causal inference with confounder representation learning from mixed structured and textual data. *Cognitive Computation*, 18(1):40, 2026.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela Van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

Jiayi Guo, Haoxuan Li, Ye Tian, and Peng Wu. A relative error-based evaluation framework of heterogeneous treatment effect estimators. In *International Conference on Learning Representations*, pp. 58694–58717, 2026.

Lisa Herzog, Pascal Bühler, Ezequiel de la Rosa, Beate Sick, and Susanne Wegener. Outcome prediction and individualized treatment effect estimation in patients with large vessel occlusion stroke. In *Image Analysis in Stroke Diagnosis and Interventions*, pp. 73–82, 2026.

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

Yiyan Huang, Siyi Wang, Cheuk Hang Leung, Qi Wu, Dongdong Wang, and Zhixiang Huang. Dignet: Learning decomposed patterns in representation balancing for treatment effect estimation. *Transactions on Machine Learning Research*, 2024.

Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23(166):1–50, 2022.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating identifiable causal effects through double machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12113–12122, 2021.

Amirreza Kazemi and Martin Ester. Adversarially balanced representation for continuous treatment effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13085–13093, 2024.

Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3768–3774, 2016.

Yijun Li, Cheuk Hang Leung, Xiangqian Sun, Chaoqun Wang, Yiyan Huang, Xing Yan, Qi Wu, Dongdong Wang, and Zhixiang Huang. The causal impact of credit lines on spending distributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 180–187, 2024.

Ruoqi Liu, Lai Wei, and Ping Zhang. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. *Nature Machine Intelligence*, 3(1):68–75, 2021.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Rickmer Schulte, David Rügamer, and Thomas Nagler. Adjustment for confounding using pretrained representations. In *Proceedings of the International Conference on Machine Learning*, pp. 53557–53580, 2025.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the International Conference on Machine Learning*, pp. 3076–3085, 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.

Martin Spindler, Philipp Bach, Victor Chernozhukov, Sven Klaassen, Jan Teichert-Kluge, and Suhas Vijaykumar. Valid inference for treatment effects under multimodal confounding. In *Proceedings of the Conference on Causal Learning and Reasoning*, pp. 222–247, 2026.

Siddharth Srivastava and Gaurav Sharma. Omnivec2-a novel transformer based network for large scale multimodal and multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27402–27414, 2024.

Yawen Wei, Zhen Li, Jonghye Woo, Jinsong Ouyang, Georges El Fakhri, and Xiaofeng Liu. Individualized treatment effect inference of head and neck cancer with multimodal data. *APSIPA Transactions on Signal and Information Processing*, 15(2):3–20, 2026.

Yinjun Wu, Mayank Keoliya, Kan Chen, Neelay Velingker, Ziyang Li, Emily J Getzen, Qi Long, Mayur Naik, Ravi B Parikh, and Eric Wong. DISCRET: Synthesizing faithful explanations for treatment effect estimation. In *Proceedings of the International Conference on Machine Learning*, pp. 53597–53618, 2024.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pp. 2638–2648, 2018.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *International Conference on Learning Representations*, pp. 98891–98913, 2025.

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.

Jiaxuan Zhang, Emadeldeen Eldele, Fuyuan Cao, Yang Wang, Xiaoli Li, and Jiye Liang. Counterfactual contrastive learning with normalizing flows for robust treatment effect estimation. In *Proceedings of the International Conference on Machine Learning*, pp. 74819–74839, 2025.

Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1005–1014, 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 ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4612–4620, 2022.

Guanglin Zhou, Lina Yao, Xiwei Xu, Chen Wang, and Liming Zhu. Cycle-balanced representation learning for counterfactual inference. In *Proceedings of the 2022 SIAM International Conference on Data Mining*, pp. 442–450, 2022.

A Definitions and Assumptions

We introduce the following definitions and identification assumptions for multimodal individual treatment effect estimation.

Let

$$\mathcal{X} = \left(X^{\mathrm{tab}}, X^{\mathrm{txt}}, X^{\mathrm{img}}\right)$$

denote the pretreatment multimodal covariates, $A \in \{0,1\}$ the treatment assignment, and $Y(a)$ the potential outcome under treatment a.

Definition 3 (Individual Treatment Effect)**.** *The individual treatment effect is defined as*

$$\tau(x) = \mathbb{E}\left[Y(1) - Y(0) \mid \mathcal{X} = x\right]. \tag{28}$$

Definition 4 (Estimated Individual Treatment Effect)**.** *Let $f : \mathcal{X} \times \{0,1\} \to \mathcal{Y}$ be an outcome prediction function. The estimated individual treatment effect is*

$$\widehat{\tau}_f(x) = f(x,1) - f(x,0). \tag{29}$$

Definition 5 (PEHE)**.** *The expected Precision in Estimation of Heterogeneous Effect is defined as*

$$\epsilon_{\mathrm{PEHE}}(f) = \mathbb{E}_{\mathcal{X}}\left[\left(\widehat{\tau}_f(\mathcal{X}) - \tau(\mathcal{X})\right)^2\right]. \tag{30}$$

Following the standard identification conditions in causal inference, we adopt the following assumptions for multimodal treatment effect estimation.

Assumption 6 (Consistency)**.** *For an individual receiving treatment $A = a$, the observed outcome equals the corresponding potential outcome:*

$$Y = Y(a).$$

Assumption 7 (Multimodal Ignorability)**.** *The potential outcomes are independent of treatment assignment conditional on the jointly observed pretreatment multimodal covariates:*

$$\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid \mathcal{X}.$$

Assumption 8 (Overlap)**.** *For any x in the support of $\mathcal{X}$,*

$$0 < P(A = 1 \mid \mathcal{X} = x) < 1.$$

B Additional Technical Details

This section provides additional technical details of MultiCFR. We first present the complete multimodal representation encoding process and then summarize the overall training and inference procedure.

B.1 Multimodal Representation Encoding

This section provides the complete implementation details of the multimodal representation encoding module described in Section 4.1. MultiCFR first converts tabular, textual, and visual inputs into token sequences in a shared representation space, and then integrates them through cross-modal attention.

Tabular tokenization. Let x_{ij}^{tab} denote the standardized value of the jth tabular covariate and $m_{ij} \in \{0,1\}$ its missingness indicator. We construct one token for each covariate:

$$\mathbf{h}_{ij}^{\mathrm{tab}} = x_{ij}^{\mathrm{tab}}\mathbf{v}_j + m_{ij}\mathbf{u}_j + \mathbf{e}_j, \qquad j = 1, \ldots, p, \tag{31}$$

where $\mathbf{v}_j, \mathbf{u}_j, \mathbf{e}_j \in \mathbb{R}^d$ are learnable value, missingness, and feature-identity embeddings, respectively. The resulting tabular token matrix is denoted by

$$\mathbf{H}_i^{\mathrm{tab}} = [\mathbf{h}_{i1}^{\mathrm{tab}}; \ldots; \mathbf{h}_{ip}^{\mathrm{tab}}] \in \mathbb{R}^{p \times d}. \tag{32}$$

Text and image encoding. Text and image inputs are processed by modality-specific encoders:

$$\widetilde{\mathbf{H}}_i^{\mathrm{txt}} = E_{\mathrm{txt}}(\mathbf{x}_i^{\mathrm{txt}}), \qquad \widetilde{\mathbf{H}}_i^{\mathrm{img}} = E_{\mathrm{img}}(\mathbf{x}_i^{\mathrm{img}}). \tag{33}$$

The resulting token-level text features and patch-level image features are projected into the shared d-dimensional representation space:

$$\mathbf{H}_i^m = \widetilde{\mathbf{H}}_i^m \mathbf{W}_m + \mathbf{b}_m, \qquad m \in \{\mathrm{txt}, \mathrm{img}\}. \tag{34}$$

Padding positions in the text sequence are masked during attention.

Shared cross-modal fusion. Before semantic anchoring, we apply cross-modal attention (Chen & Yang, 2025; Ye et al., 2025; Srivastava & Sharma, 2024) to capture interactions among heterogeneous modalities. For two distinct modalities m and n, the update from n to m is

$$\mathbf{U}_i^{m\leftarrow n} = \mathrm{MHA}_{m\leftarrow n}(\mathbf{Q}_m(\mathrm{LN}(\mathbf{H}_i^m)), \mathbf{K}_n(\mathrm{LN}(\mathbf{H}_i^n)), \mathbf{V}_n(\mathrm{LN}(\mathbf{H}_i^n))), \tag{35}$$

where MHA denotes multi-head attention.

The cross-modal features are then refined through residual connections and feed-forward networks:

$$\mathbf{G}_i^m = \mathrm{LN}\left(\mathbf{H}_i^m + \sum_{n\neq m} \mathbf{W}_o^m \mathbf{U}_i^{m\leftarrow n}\right), \tag{36}$$

$$\overline{\mathbf{H}}_i^m = \mathrm{LN}(\mathbf{G}_i^m + \mathrm{FFN}_m(\mathbf{G}_i^m)). \tag{37}$$

Finally, attention pooling summarizes the refined multimodal tokens into a base multimodal representation:

$$\mathbf{b}_i = \mathrm{AttnPool}\Big([\overline{\mathbf{H}}_i^{\mathrm{tab}}; \overline{\mathbf{H}}_i^{\mathrm{txt}}; \overline{\mathbf{H}}_i^{\mathrm{img}}]\Big) \in \mathbb{R}^d. \tag{38}$$

Modality embeddings are added before pooling to preserve token source information.

B.2 Training and Inference Procedure

Having specified the multimodal representation encoding process, we next summarize how it is integrated with tabular semantic anchoring, dual-branch representation learning, and selective balancing in the complete MultiCFR framework. Algorithm 1 presents the overall training and inference procedure.

C Theoretical Analysis

This appendix provides the detailed theoretical analysis underlying Section 5. We follow the definitions and identification assumptions in Appendix A. Following the general IPM-based analysis of counterfactual regression (Shalit et al., 2017), we derive a PEHE bound adapted to the selective balancing mechanism of MultiCFR. We further provide a conditional analysis of tabular semantic anchoring.

C.1 Theoretical Setup and Training Objective

Define the conditional potential-outcome functions as

$$\mu_a(x) = \mathbb{E}[Y(a) \mid \mathcal{X} = x], \qquad a \in \{0, 1\}. \tag{39}$$

By Definition 3,

$$\tau(x) = \mu_1(x) - \mu_0(x). \tag{40}$$

Dual-branch representation. Let $\mathbf{r} = r(\mathcal{X})$ denote the anchor-refined multimodal representation and $\mathbf{s} = s(X^{\mathrm{tab}}) \in (0, 1)^d$ the anchor-conditioned allocation gate. MultiCFR constructs

$$\mathbf{Z}^B = f_B(\mathbf{r} \odot \mathbf{s}), \qquad \mathbf{Z}^R = f_R(\mathbf{r} \odot (\mathbf{1} - \mathbf{s})), \qquad \mathbf{Z} = [\mathbf{Z}^B; \mathbf{Z}^R]. \tag{41}$$

Algorithm 1 Training and Inference of MultiCFR

Require: Training data $\mathcal{D} = \{(\mathbf{x}_i^{\mathrm{tab}}, \mathbf{x}_i^{\mathrm{txt}}, \mathbf{x}_i^{\mathrm{img}}, A_i, Y_i)\}_{i=1}^N$; balance weight λ_{bal}; regularization weight λ_{reg}; RBF bandwidths Σ; number of epochs E
Ensure: Learned parameters Θ and ITE estimator $\widehat{\tau}(\cdot)$

Initialize trainable parameters Θ
Freeze pretrained text and image encoders and cache their features
for $e = 1, \ldots, E$ **do**
  for each minibatch $\mathcal{B} \subset \mathcal{D}$ containing both treatment groups **do**
  Multimodal encoding and fusion
    for all $i \in \mathcal{B}$ **do**
      Construct tabular tokens:
      $\mathbf{h}_{ij}^{\mathrm{tab}} \leftarrow x_{ij}^{\mathrm{tab}} \mathbf{v}_j + m_{ij} \mathbf{u}_j + \mathbf{e}_j$
      Obtain text and image token representations $\mathbf{H}_i^{\mathrm{txt}}$ and $\mathbf{H}_i^{\mathrm{img}}$
      Compute the base multimodal representation:
      $\mathbf{b}_i \leftarrow \mathrm{CrossModalFusion}(\mathbf{H}_i^{\mathrm{tab}}, \mathbf{H}_i^{\mathrm{txt}}, \mathbf{H}_i^{\mathrm{img}})$
  Tabular semantic anchoring
      Construct the individual-specific semantic anchor:
      $\boldsymbol{\alpha}_i \leftarrow \frac{1}{p} \sum_{j=1}^p \mathbf{h}_{ij}^{\mathrm{tab}}$
      for $m \in \{\mathrm{txt}, \mathrm{img}\}$ **do**
        Retrieve anchor-relevant information:
        $\mathbf{c}_i^m \leftarrow \mathrm{AnchorAttention}(q_m(\boldsymbol{\alpha}_i), \mathbf{H}_i^m)$
        Compute the modality gate:
        $g_i^m \leftarrow \sigma(f_m([\boldsymbol{\alpha}_i; \mathbf{c}_i^m]))$
      end for
      Refine the multimodal representation:
      $\mathbf{r}_i \leftarrow \mathrm{LN}\Big(\mathbf{b}_i + \sum_{m \in \{\mathrm{txt}, \mathrm{img}\}} g_i^m \mathbf{W}_\Delta^m \mathbf{c}_i^m\Big)$
  Dual-branch representation learning
      Compute the soft allocation gate:
      $\mathbf{s}_i \leftarrow \sigma(\mathbf{W}_s \, \mathrm{LN}(\boldsymbol{\alpha}_i) + \mathbf{b}_s)$
      Construct the two complementary branches:
      $\mathbf{z}_i^B \leftarrow f_B(\mathbf{r}_i \odot \mathbf{s}_i), \quad \mathbf{z}_i^R \leftarrow f_R(\mathbf{r}_i \odot (\mathbf{1} - \mathbf{s}_i))$
      Predict potential outcomes:
      $\widehat{\mu}_{a,i} \leftarrow h_a([\mathbf{z}_i^B; \mathbf{z}_i^R]), \quad a \in \{0, 1\}$
    end for
  Selective balancing and optimization
    Compute the factual regression loss:
    $\mathcal{L}_{\mathrm{f}} \leftarrow \dfrac{1}{|\mathcal{B}|} \displaystyle\sum_{i \in \mathcal{B}} (Y_i - \widehat{\mu}_{A_i,i})^2$
    Compute the balancing loss:
    $\mathcal{L}_{\mathrm{bal}} \leftarrow \mathrm{MMD}\left(\left\{\dfrac{\mathbf{z}_i^B}{\sqrt{d}} : A_i = 1\right\}, \left\{\dfrac{\mathbf{z}_i^B}{\sqrt{d}} : A_i = 0\right\}; \Sigma\right)$
    Compute the overall objective:
    $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{f}} + \lambda_{\mathrm{bal}} \mathcal{L}_{\mathrm{bal}} + \lambda_{\mathrm{reg}} \|\Theta\|_2^2$
    Update Θ by minimizing $\mathcal{L}$
  end for
  Evaluate the validation loss and retain the best checkpoint
end for
  Inference
for all test individuals i **do**
  Predict $\widehat{\mu}_0(\mathcal{X}_i)$ and $\widehat{\mu}_1(\mathcal{X}_i)$
  Estimate the ITE:
  $\widehat{\tau}(\mathcal{X}_i) \leftarrow \widehat{\mu}_1(\mathcal{X}_i) - \widehat{\mu}_0(\mathcal{X}_i)$
end for
return Θ and $\widehat{\tau}(\cdot)$

The balance-regularized branch $\mathbf{Z}^B$ is directly constrained across treatment groups, whereas $\mathbf{Z}^R$ preserves complementary predictive information. Both branches are used for potential-outcome prediction:

$$\widehat{\mu}_a(\mathcal{X}) = h_a(\mathbf{Z}), \qquad \widehat{\tau}(\mathcal{X}) = h_1(\mathbf{Z}) - h_0(\mathbf{Z}). \tag{42}$$

Following the implementation, define the dimension-normalized balanced representation

$$\mathbf{W} = \frac{\mathbf{Z}^B}{\sqrt{d}}. \tag{43}$$

Let q_a^B denote the distribution of $\mathbf{W}$ conditional on $A = a$, and let $\pi = P(A = 1)$.

Definition 9 (Maximum Mean Discrepancy)**.** *Let k be a positive-definite kernel with RKHS $\mathcal{H}_k$. For two distributions P and Q, the maximum mean discrepancy is*

$$\mathrm{MMD}(P, Q) = \sup_{\|f\|_{\mathcal{H}_k} \le 1} \left| \mathbb{E}_P[f(U)] - \mathbb{E}_Q[f(U)] \right|. \tag{44}$$

We denote

$$d_B = \mathrm{MMD}(q_0^B, q_1^B). \tag{45}$$

Training objective. For a training sample of size N, the factual regression loss is

$$\widehat{\mathcal{L}}_{\mathrm{f}} = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \widehat{\mu}_{A_i}(\mathcal{X}_i) \right)^2. \tag{46}$$

The complete training objective is

$$\widehat{\mathcal{L}} = \widehat{\mathcal{L}}_{\mathrm{f}} + \lambda_{\mathrm{bal}} \widehat{\mathrm{MMD}}(\widehat{q}_0^B, \widehat{q}_1^B) + \lambda_{\mathrm{reg}} \|\Theta\|_2^2, \tag{47}$$

where λ_{bal} and λ_{reg} control the balancing and model regularization strengths, respectively. The L_2 term controls model complexity and does not appear as a causal-discrepancy term in the population PEHE bound derived below.

C.2 Predictive Information Preservation

We first characterize the value of retaining the complementary predictive branch.

Lemma 10. *For a fixed representation $\mathbf{Z}$ and unrestricted measurable outcome heads,*

$$\inf_{h_0, h_1} \mathbb{E}\Big[\big(\widehat{\tau}(\mathcal{X}) - \tau(\mathcal{X}) \big)^2 \Big] = \mathbb{E}[\mathrm{Var}(\tau(\mathcal{X}) \mid \mathbf{Z})]. \tag{48}$$

Moreover,

$$\mathbb{E}\big[\mathrm{Var}(\tau(\mathcal{X}) \mid \mathbf{Z}^B)\big] - \mathbb{E}[\mathrm{Var}(\tau(\mathcal{X}) \mid \mathbf{Z})] = \mathbb{E}\Big[\big(\mathbb{E}[\tau(\mathcal{X}) \mid \mathbf{Z}] - \mathbb{E}[\tau(\mathcal{X}) \mid \mathbf{Z}^B] \big)^2 \Big] \ge 0. \tag{49}$$

Proof. For any measurable predictor $g(\mathbf{Z})$, the conditional expectation $\mathbb{E}[\tau(\mathcal{X}) \mid \mathbf{Z}]$ minimizes squared error. Since any measurable $g(\mathbf{Z})$ can be represented as the difference of two unrestricted heads (e.g., $h_1 = g$ and $h_0 = 0$),

$$\inf_{h_0, h_1} \mathbb{E}\Big[\big(h_1(\mathbf{Z}) - h_0(\mathbf{Z}) - \tau(\mathcal{X}) \big)^2 \Big] = \inf_{g} \mathbb{E}\big[(g(\mathbf{Z}) - \tau(\mathcal{X}))^2 \big] = \mathbb{E}[\mathrm{Var}(\tau(\mathcal{X}) \mid \mathbf{Z})]. \tag{50}$$

Because $\mathbf{Z}^B$ is a deterministic component of $\mathbf{Z} = [\mathbf{Z}^B; \mathbf{Z}^R]$, we have $\sigma(\mathbf{Z}^B) \subseteq \sigma(\mathbf{Z})$. Applying the law of total variance conditionally on $\mathbf{Z}^B$ gives

$$\mathrm{Var}(\tau \mid \mathbf{Z}^B) = \mathbb{E}[\mathrm{Var}(\tau \mid \mathbf{Z}) \mid \mathbf{Z}^B] + \mathrm{Var}(\mathbb{E}[\tau \mid \mathbf{Z}] \mid \mathbf{Z}^B). \tag{51}$$

Taking expectations yields Eq. 49. □

Lemma 10 states that, for fixed learned branches, retaining $\mathbf{Z}^R$ cannot increase the irreducible ITE error of an unrestricted predictor and can strictly reduce it whenever $\mathbf{Z}^R$ contains treatment-effect information not measurable from $\mathbf{Z}^B$ alone.

C.3 SELECTIVE-BALANCING PEHE BOUND

For $a \in \{0,1\}$, define the response-surface error

$$e_a(\S) = \left(h_a(\mathbf{Z}(\S)) - \mu_a(\S)\right)^2. \tag{52}$$

We distinguish its conditional expectation in the factual group and in the opposite treatment group:

$$\ell_a^{\mathrm{F}}(w) = \mathbb{E}[e_a(\mathcal{X}) \mid \mathbf{W} = w, A = a], \tag{53}$$

$$\ell_a^{\mathrm{CF}}(w) = \mathbb{E}[e_a(\mathcal{X}) \mid \mathbf{W} = w, A = 1-a]. \tag{54}$$

Define the residual conditional loss shift

$$\Delta_a^B = \mathbb{E}_{\mathbf{W}\sim q_{1-a}^B}\left[\left|\ell_a^{\mathrm{CF}}(\mathbf{W}) - \ell_a^{\mathrm{F}}(\mathbf{W})\right|\right]. \tag{55}$$

This term captures outcome-prediction differences between treatment groups that remain after conditioning on the branch directly regularized by MMD. It may include information carried by $\mathbf{Z}^R$ as well as other information in $\mathcal{X}$ not summarized by $\mathbf{W}$.

Define

$$G_a = \mathbb{E}[e_a(\mathcal{X}) \mid A = a], \tag{56}$$

$$C_a = \mathbb{E}[e_a(\mathcal{X}) \mid A = 1-a], \tag{57}$$

$$F_a = \mathbb{E}\left[(Y - h_a(\mathbf{Z}))^2 \mid A = a\right]. \tag{58}$$

Assumption 11 (RKHS regularity)**.** *For each $a \in \{0,1\}$, the factual conditional loss function satisfies*

$$\ell_a^{\mathrm{F}} \in \mathcal{H}_k, \qquad \|\ell_a^{\mathrm{F}}\|_{\mathcal{H}_k} \le B_a < \infty. \tag{59}$$

Lemma 12. *Under Assumption 11,*

$$C_a \le G_a + \Delta_a^B + B_a d_B, \qquad a \in \{0,1\}. \tag{60}$$

Proof. By the definitions above,

$$C_a - G_a = \mathbb{E}_{q_{1-a}^B}[\ell_a^{\mathrm{CF}}(\mathbf{W})] - \mathbb{E}_{q_a^B}[\ell_a^{\mathrm{F}}(\mathbf{W})] \tag{61}$$

$$= \mathbb{E}_{q_{1-a}^B}\left[\ell_a^{\mathrm{CF}}(\mathbf{W}) - \ell_a^{\mathrm{F}}(\mathbf{W})\right] + \mathbb{E}_{q_{1-a}^B}[\ell_a^{\mathrm{F}}(\mathbf{W})] - \mathbb{E}_{q_a^B}[\ell_a^{\mathrm{F}}(\mathbf{W})]. \tag{62}$$

The first term is upper bounded by Δ_a^B. By the RKHS definition of MMD and Assumption 11,

$$\left|\mathbb{E}_{q_{1-a}^B}[\ell_a^{\mathrm{F}}(\mathbf{W})] - \mathbb{E}_{q_a^B}[\ell_a^{\mathrm{F}}(\mathbf{W})]\right| \le B_a \,\mathrm{MMD}(q_0^B, q_1^B) = B_a d_B. \tag{63}$$

Combining the two inequalities proves Eq. 60. □

Lemma 13. *Under consistency and multimodal ignorability,*

$$F_a = G_a + \mathbb{E}[\mathrm{Var}(Y(a) \mid \mathcal{X}) \mid A = a] \ge G_a. \tag{64}$$

Proof. By consistency, $Y = Y(a)$ for units with $A = a$. By multimodal ignorability, $\mathbb{E}[Y(a) \mid \mathcal{X}, A = a] = \mu_a(\mathcal{X})$. Hence

$$F_a = \mathbb{E}\left[(Y(a) - h_a(\mathbf{Z}))^2 \mid A = a\right] \tag{65}$$

$$= \mathbb{E}\left[(\mu_a(\mathcal{X}) - h_a(\mathbf{Z}))^2 \mid A = a\right] + \mathbb{E}[\mathrm{Var}(Y(a) \mid \mathcal{X}) \mid A = a] \tag{66}$$

$$= G_a + \mathbb{E}[\mathrm{Var}(Y(a) \mid \mathcal{X}) \mid A = a]. \tag{67}$$

The variance term is nonnegative. □

Theorem 14. *Let $\widehat{\tau}(\mathcal{X}) = h_1(\mathbf{Z}) - h_0(\mathbf{Z})$ denote the estimated ITE based on the dual-branch representation $\mathbf{Z} = [\mathbf{Z}^B; \mathbf{Z}^R]$, and let $d_B = \mathrm{MMD}(q_0^B, q_1^B)$ denote the treatment-group discrepancy of the balance-regularized branch. For $a \in \{0,1\}$, let F_a denote the factual prediction error, Δ_a^B the residual discrepancy, and B_a the corresponding RKHS norm bound. Under consistency, multimodal ignorability, overlap, and the preceding RKHS condition, the PEHE is bounded by*

$$\epsilon_{\mathrm{PEHE}} \le 2(F_0 + F_1) + 2\left[\pi B_0 + (1-\pi)B_1\right] d_B + 2\left[\pi \Delta_0^B + (1-\pi)\Delta_1^B\right]. \tag{68}$$

Proof. From Eq. 42,

$$\widehat{\tau}(\mathcal{X}) - \tau(\mathcal{X}) = (h_1(\mathbf{Z}) - \mu_1(\mathcal{X})) - (h_0(\mathbf{Z}) - \mu_0(\mathcal{X})). \tag{69}$$

Using $(u-v)^2 \le 2u^2 + 2v^2$ gives

$$\epsilon_{\text{PEHE}} \le 2 \sum_{a=0}^{1} \mathbb{E}[e_a(\mathcal{X})]. \tag{70}$$

For $a = 0$,

$$\mathbb{E}[e_0(\mathcal{X})] = (1-\pi)G_0 + \pi C_0 \le G_0 + \pi(\Delta_0^B + B_0 d_B). \tag{71}$$

where the inequality follows from Lemma 12. Similarly,

$$\mathbb{E}[e_1(\mathcal{X})] \le G_1 + (1-\pi)(\Delta_1^B + B_1 d_B). \tag{72}$$

Therefore,

$$\sum_{a=0}^{1} \mathbb{E}[e_a(\mathcal{X})] \le G_0 + G_1 + \big[\pi B_0 + (1-\pi)B_1\big] d_B + \pi\Delta_0^B + (1-\pi)\Delta_1^B. \tag{73}$$

Finally, Lemma 13 gives $G_a \le F_a$. □

Implication. Theorem 14 separates the PEHE upper bound into three components: factual prediction error, treatment-group discrepancy in the branch regularized by MMD, and a residual conditional loss shift. Consequently, a smaller MMD does not by itself imply a smaller PEHE bound. A decrease in d_B can be offset by an increase in factual risk or in Δ_a^B. Combined with Lemma 10, this motivates retaining $\mathbf{Z}^R$ for prediction rather than forcing the entire multimodal representation to satisfy the same balancing constraint. This is a comparison of bound components and does not imply unconditional dominance over global balancing.

C.4 Conditional Analysis of Tabular Semantic Anchoring

The following result formalizes when an individual-specific tabular anchor can concentrate retrieval on relevant positions in an unstructured modality. The statement is conditional: it does not assume that a positive attention margin is automatically achieved during training.

For modality $m \in \{\text{img}, \text{txt}\}$, let N_m be the number of image patches or text tokens. Let k_{mj} and v_{mj} denote the key and value at position j, and let $q_m(\boldsymbol{\alpha})$ be the query generated from the tabular semantic anchor. Define

$$s_{mj} = \frac{\langle q_m(\boldsymbol{\alpha}), k_{mj}\rangle}{\sqrt{d_k}}, \qquad \pi_{mj} = \frac{\exp(s_{mj})}{\sum_{\ell=1}^{N_m}\exp(s_{m\ell})}, \qquad c_m = \sum_{j=1}^{N_m} \alpha_{mj} v_{mj}. \tag{74}$$

Let $S_m \subset \{1, \dots, N_m\}$ be a nonempty proper subset of positions relevant to the downstream outcome prediction task. Define the attention-renormalized relevant context

$$c_m^\star = \frac{1}{\sum_{j\in S_m}\pi_{mj}} \sum_{j\in S_m} \pi_{mj} v_{mj}. \tag{75}$$

Assumption 15 (Anchor relevance margin)**.** *For modality m, suppose $\|v_{mj}\| \le V_m$ for all j and the anchor-induced scores satisfy*

$$\min_{j\in S_m} s_{mj} - \max_{j\notin S_m} s_{mj} \ge \gamma_m > 0. \tag{76}$$

Lemma 16. *Under Assumption 15, the total attention mass assigned to positions outside S_m,*

$$\rho_m = \sum_{j\notin S_m} \pi_{mj}, \tag{77}$$

satisfies

$$\rho_m \le \frac{N_m - |S_m|}{|S_m|} e^{-\gamma_m}, \qquad \|c_m - c_m^\star\| \le 2V_m\rho_m. \tag{78}$$

Proof. Let $s_m^+ = \min_{j\in S_m} s_{mj}$ and $s_m^- = \max_{j\notin S_m} s_{mj}$. Assumption 15 implies $s_m^- \le s_m^+ - \gamma_m$. Therefore,

$$\sum_{j\notin S_m} e^{s_{mj}} \le (N_m - |S_m|)e^{s_m^-} \le (N_m - |S_m|)e^{s_m^+ - \gamma_m}, \tag{79}$$

$$\sum_{j\in S_m} e^{s_{mj}} \ge |S_m| e^{s_m^+}. \tag{80}$$

Hence

$$\rho_m \le \frac{\sum_{j\notin S_m} e^{s_{mj}}}{\sum_{j\in S_m} e^{s_{mj}}} \le \frac{N_m - |S_m|}{|S_m|} e^{-\gamma_m}. \tag{81}$$

Writing the full context as $c_m = (1-\rho_m)c_m^\star + \rho_m c_m^{\mathrm{irr}}$, where c_m^{irr} is the normalized weighted average over positions outside S_m, gives

$$\|c_m - c_m^\star\| = \rho_m \|c_m^{\mathrm{irr}} - c_m^\star\| \le 2V_m\rho_m. \tag{82}$$

This proves Eq. 78. □

Theorem 17. *Keep the generic multimodal pathway and tabular anchor fixed. Let $\mathbf{Z}^\star$ denote the dual-branch representation obtained by replacing the retrieved image and text contexts with $c_m^\star$. Assume the downstream mapping G from the two retrieved contexts to $\mathbf{Z}$ satisfies*

$$\|G(c_{\mathrm{img}}, c_{\mathrm{txt}}) - G(c'_{\mathrm{img}}, c'_{\mathrm{txt}})\| \le L_G \left(\|c_{\mathrm{img}} - c'_{\mathrm{img}}\| + \|c_{\mathrm{txt}} - c'_{\mathrm{txt}}\|\right). \tag{83}$$

Then

$$\|\mathbf{Z} - \mathbf{Z}^\star\| \le \xi := 2L_G \sum_{m\in\{\mathrm{img},\mathrm{txt}\}} V_m \frac{N_m - |S_m|}{|S_m|} e^{-\gamma_m}. \tag{84}$$

If h_a is L_a^h-Lipschitz, then

$$\sqrt{\mathbb{E}[(h_a(\mathbf{Z}) - \mu_a(\mathcal{X}))^2]} \le \sqrt{\mathbb{E}[(h_a(\mathbf{Z}^\star) - \mu_a(\mathcal{X}))^2]} + L_a^h\xi. \tag{85}$$

Proof. Applying Lemma 16 to the image and text contexts and then using Eq. 83 yields

$$\|\mathbf{Z} - \mathbf{Z}^\star\| \le L_G \sum_{m\in\{\mathrm{img},\mathrm{txt}\}} \|c_m - c_m^\star\| \tag{86}$$

$$\le 2L_G \sum_{m\in\{\mathrm{img},\mathrm{txt}\}} V_m \frac{N_m - |S_m|}{|S_m|} e^{-\gamma_m} = \xi. \tag{87}$$

If h_a is L_a^h-Lipschitz, then $|h_a(\mathbf{Z}) - h_a(\mathbf{Z}^\star)| \le L_a^h\xi$. By Minkowski's inequality,

$$\sqrt{\mathbb{E}[(h_a(\mathbf{Z}) - \mu_a(\mathcal{X}))^2]} \tag{88}$$

$$\le \sqrt{\mathbb{E}[(h_a(\mathbf{Z}^\star) - \mu_a(\mathcal{X}))^2]} + \sqrt{\mathbb{E}[(h_a(\mathbf{Z}) - h_a(\mathbf{Z}^\star))^2]} \tag{89}$$

$$\le \sqrt{\mathbb{E}[(h_a(\mathbf{Z}^\star) - \mu_a(\mathcal{X}))^2]} + L_a^h\xi. \tag{90}$$

This proves Eq. 85. □

Lemma 18. *Suppose additionally that $\mu_a(\mathcal{X}) = g_a(\mathbf{Z}^\star)$ for $a \in \{0,1\}$, where g_a is L_a^g-Lipschitz. Then*

$$\mathbb{E}[\mathrm{Var}(\tau(\mathcal{X}) \mid \mathbf{Z})] \le (L_0^g + L_1^g)^2\xi^2. \tag{91}$$

Proof. Under the stated condition, $\tau(\mathcal{X}) = g_1(\mathbf{Z}^\star) - g_0(\mathbf{Z}^\star)$. Consider the $\mathbf{Z}$-measurable predictor $g_1(\mathbf{Z}) - g_0(\mathbf{Z})$. By Lipschitz continuity and Eq. 84,

$$\left|\tau(\mathcal{X}) - \big(g_1(\mathbf{Z}) - g_0(\mathbf{Z})\big)\right| \le |g_1(\mathbf{Z}^\star) - g_1(\mathbf{Z})| + |g_0(\mathbf{Z}^\star) - g_0(\mathbf{Z})| \le (L_1^g + L_0^g)\xi. \tag{92}$$

The conditional expectation is the optimal squared-error predictor from $\mathbf{Z}$, so Lemma 10 gives Eq. 91. □

Scope of the anchoring result. Lemma 16 and Theorem 17 are conditional guarantees. They show that if the tabular anchor induces a positive score margin between relevant and irrelevant positions, the attention mass on irrelevant content decays exponentially in that margin and the resulting representation approaches its relevant-context counterpart. They do not guarantee that training always produces $\gamma_m > 0$, nor do they by themselves upper-bound the residual Δ_a^B in Theorem 14.

D SEMI-SYNTHETIC DATASET CONSTRUCTION

D.1 MIMIC-TIT DATASET

We constructed a patient-level multimodal benchmark from MIMIC-IV v3.1 and MIMIC-CXR v2.0.0. Unlike semi-synthetic benchmarks that combine independently sampled tabular, textual, and image datasets, all modalities in our cohort are naturally linked to the same patient and clinical episode. For patient i, one observation is represented as

$$\mathcal{O}_i = \left(X_i^{\mathrm{tab}}, X_i^{\mathrm{txt}}, X_i^{\mathrm{img}}, A_i, Y_i\right). \tag{93}$$

Here, X_i^{tab} denotes structured electronic health record (EHR) variables, X_i^{txt} is a radiology report, X_i^{img} is a frontal chest radiograph, and A_i and Y_i are the semi-synthetic treatment and observed outcome, respectively. The observed modalities are real, whereas treatment assignments, potential outcomes, and individual treatment effects are generated by a prespecified data-generating process (DGP).

We matched MIMIC-CXR studies to MIMIC-IV admissions and ICU stays using `subject_id`, `hadm_id`, and `stay_id`. Adult patients with an ICU stay of at least 24 hours and at least one AP/PA radiograph within the first 24 hours were retained. For multiple eligible episodes, the earliest ICU stay and frontal study were selected.

The radiograph acquisition time t_i^x was used as the index time, and structured measurements were collected from

$$\mathcal{W}_i = [a_i^x - 24\ \mathrm{h},\ a_i^x]. \tag{94}$$

After excluding patients without valid dynamic EHR measurements and prespecified extreme-value cases, the final cohort contained 7396 patients.

D.1.1 STRUCTURED VARIABLES AND PREPROCESSING

The structured modality contains 18 numeric variables, including demographic, comorbidity, vital-sign, and laboratory measurements, together with 12 missingness indicators. Missing values were imputed using training-set medians, and numeric features were standardized as

$$\widetilde{x}_{ij} = \frac{x_{ij} - \widehat{\mu}_{j,\mathrm{tr}}}{\widehat{\sigma}_{j,\mathrm{tr}}}. \tag{95}$$

The data were split into 5183 training, 1103 validation, and 1110 test patients.

D.1.2 SEMI-SYNTHETIC TREATMENT AND OUTCOMES

We generated treatment assignments and potential outcomes from the real multimodal observations so that the ground-truth ITE is known. Let

$$\mathcal{S}(v_i) = \frac{v_i - \widehat{\mu}_{v,\mathrm{tr}}}{\widehat{\sigma}_{v,\mathrm{tr}}}. \tag{96}$$

A structured confounding component was defined as

$$C_i^{\mathrm{shared}} = \mathcal{S}\big[\mathcal{S}(\mathrm{WBC}_i) - \mathcal{S}(\mathrm{SpO}_{2,i}) + \mathcal{S}(\mathrm{RR}_i) + \mathcal{S}(\mathrm{Lactate}_i)\big]. \tag{97}$$

To incorporate complementary unstructured information, we used the CheXpert Lung Opacity label only in the oracle DGP. Its residual component after regressing on C_i^{shared} defines

$$C_i^{\mathrm{private}} = \mathcal{S}(r_{L,i}), \tag{98}$$

and the total confounding score is

$$C_i^{(\rho)} = \mathcal{S}\Big(C_i^{\mathrm{shared}} + \rho C_i^{\mathrm{private}}\Big). \tag{99}$$

Similarly, an outcome-prognostic score was constructed from age, Charlson comorbidity score, and the CheXpert Cardiomegaly label, followed by residualization with respect to $C_i^{(\rho)}$ to obtain $P_i^{(\rho)}$.

Treatment assignment was generated as

$$e_i^{(\gamma,\rho)} = \text{sigmoid}\Big(\gamma C_i^{(\rho)}\Big), \qquad A_i = \mathbb{I}\Big\{U_i < e_i^{(\gamma,\rho)}\Big\}, \tag{100}$$

where $U_i \sim \text{Uniform}(0,1)$.

The conditional potential outcomes were defined as

$$\begin{aligned} \mu_{0,i}^{(\beta,\rho)} &= C_i^{(\rho)} + \beta P_i^{(\rho)}, \\ \tau_i^{(\rho)} &= 0.5 + 0.5\tanh\Big(P_i^{(\rho)}\Big), \\ \mu_{1,i}^{(\beta,\rho)} &= \mu_{0,i}^{(\beta,\rho)} + \tau_i^{(\rho)}. \end{aligned} \tag{101}$$

Using shared noise $\varepsilon_i \sim \mathcal{N}(0,1)$,

$$Y_i(0) = \mu_{0,i}^{(\beta,\rho)} + \varepsilon_i, \qquad Y_i(1) = \mu_{1,i}^{(\beta,\rho)} + \varepsilon_i, \tag{102}$$

so that

$$Y_i(1) - Y_i(0) = \tau_i^{(\rho)}. \tag{103}$$

The primary benchmark uses $(\rho,\gamma,\beta) = (1,1,1)$.

D.1.3 Model Inputs

The models receive only the structured EHR variables, missingness indicators, raw radiology reports, and chest radiographs. CheXpert labels, latent DGP variables, propensity scores, potential outcomes, and true ITEs are used only for benchmark construction or evaluation and are never provided as model inputs.

D.2 IHDP-TIT Dataset Construction

We constructed IHDP-TIT from the standard semi-synthetic Infant Health and Development Program (IHDP) benchmark (Hill, 2011). The original benchmark contains 747 infants, 25 pre-treatment covariates, and 100 simulation replications. Each replication provides treatment assignments, factual and counterfactual outcomes, and ground-truth conditional potential outcomes, with 672 training and 75 test samples.

IHDP-TIT changes only the representation of the pre-treatment covariates while preserving the original causal estimation task. For subject i in replication r, the observed data are represented as

$$\mathcal{O}_i^{(r)} = \Big(X_i^{\text{tab}}, X_i^{\text{txt}}, X_i^{\text{img}}, A_i^{(r)}, Y_i^{\text{F},(r)}\Big), \tag{104}$$

where X_i^{tab}, X_i^{txt}, and X_i^{img} denote the tabular, textual, and visual modalities, respectively. Unlike MIMIC-TIT, IHDP-TIT does not introduce a new treatment or outcome data-generating process; all causal quantities are inherited directly from the original IHDP benchmark.

The 25 covariates were partitioned into three mutually exclusive sets,

$$\mathcal{I}_{\text{tab}} \dot\cup \mathcal{I}_{\text{txt}} \dot\cup \mathcal{I}_{\text{img}} = \{1,\dots,25\}, \tag{105}$$

so that no original covariate was discarded or duplicated across modalities. Following the IHDP-Mini setting used in DMSE Deshpande et al. (2022), nine covariates were retained as tabular variables. The remaining 16 covariates were divided equally between the text and image modalities.

D.2.1 Tabular, text and image modalities

The tabular modality directly retains the nine selected covariates:

$$X_i^{\text{tab}} = X_{i,\mathcal{I}_{\text{tab}}}. \tag{106}$$

These variables include infant characteristics, maternal background, prenatal exposure, and a study-center indicator.

Table 3: IHDP-TIT Covariate Description.

Modality	Number	Original IHDP variables
Tabular	9	`bw`, `b.head`, `preterm`, `birth.o`, `sex`, `b.marr`, `mom.lths`, `cig`, and `har`
Text	8	`momage`, `mom.hs`, `mom.scoll`, `first`, `booze`, `drugs`, `work.dur`, and `prenatal`
Image	8	`nnhealth`, `twin`, `ark`, `ein`, `mia`, `pen`, `tex`, and `was`

The text modality was generated deterministically from the eight text-source covariates:

$$X_i^{\mathrm{txt}} = g_{\mathrm{txt}}\left(X_{i,\mathcal{I}_{\mathrm{txt}}}; v_i\right), \tag{107}$$

where g_{txt} converts the structured attributes into a short prenatal-history description and $v_i \in \{1,2,3,4\}$ denotes one of four sentence templates. The generated text describes maternal age, education, parity, employment during pregnancy, prenatal care, alcohol use, and drug use.

All allocated attributes are included without omission, contradiction, or label flipping. The template is selected deterministically using the subject identifier, introducing moderate linguistic variation without changing the underlying information. Treatment assignments, outcomes, and potential outcomes are never used during text generation.

The image modality consists of deterministic 224×224 RGB synthetic visual proxies:

$$X_i^{\mathrm{img}} = g_{\mathrm{img}}\left(X_{i,\mathcal{I}_{\mathrm{img}}}; \xi_i\right), \tag{108}$$

where ξ_i is a subject-specific deterministic rendering seed. The neonatal-health variable is normalized as

$$h_i = \mathrm{clip}\left(\frac{\mathrm{nnhealth}_i - \mathrm{nnhealth}_{\min}}{\mathrm{nnhealth}_{\max} - \mathrm{nnhealth}_{\min}}, 0, 1\right). \tag{109}$$

The normalized value h_i controls the amplitude and frequency of a waveform and the number of active status segments in the image. The `twin` variable determines whether one or two infant silhouettes are displayed, while six study-center indicators determine the background color and corner-marker style. The absence of these six indicators produces the reference rendering style.

Small deterministic variations in position and waveform phase are introduced to avoid identical images while remaining independent of treatments and outcomes. No numerical values, variable names, treatment labels, outcome labels, or machine-readable codes are printed in the images. These images should therefore be interpreted as controlled synthetic visual proxies rather than real clinical images.

D.2.2 Treatment and Outcomes

For every replication r, IHDP-TIT exactly preserves the original treatment assignment $T_i^{(r)}$, factual outcome $Y_i^{\mathrm{F},(r)}$, counterfactual outcome $Y_i^{\mathrm{CF},(r)}$, and conditional potential outcomes $\mu_{0,i}^{(r)}$ and $\mu_{1,i}^{(r)}$. The ground-truth ITE is consequently defined as

$$\tau_i^{(r)} = \mu_{1,i}^{(r)} - \mu_{0,i}^{(r)}. \tag{110}$$

The original 672/75 training–test split is also retained for each of the 100 replications.

Because subject ordering may differ across replication archives, we constructed a stable subject identifier from each infant's complete 25-dimensional covariate vector. This identifier aligns the tabular record, textual description, and image for the same subject across all replications:

$$\mathrm{subject_id}_i = \mathrm{text_id}_i = \mathrm{image_id}_i. \tag{111}$$

D.2.3 Model Inputs

Each model receives only the nine tabular covariates and the corresponding text and image modalities. The full 25-dimensional covariate vectors are retained in a separate oracle file solely for auditing and full-tabular reference experiments; they are not accessible to the multimodal models.

Similarly, counterfactual outcomes, conditional potential outcomes, and true ITEs are used only for evaluation.

Every subject has all three modalities. We do not introduce random modality missingness, text omission, image occlusion, or target-correlated perturbations. Validation confirms that the three feature partitions are disjoint and collectively cover all 25 covariates, all 747 subjects have aligned tabular, text, and image records, and all treatment and outcome arrays are exactly preserved across the 100 replications.

D.3 METRICS

We evaluate treatment effect estimation from both individual- and population-level perspectives. Specifically, we use the root Precision in Estimation of Heterogeneous Effect ($\sqrt{\epsilon_{\text{PEHE}}}$) to measure the accuracy of individual-level treatment effect estimation, and the absolute error in Average Treatment Effect (ϵ_{ATE}) to assess the accuracy of population-level effect estimation. To additionally evaluate the model's ability to predict observed outcomes, we report the factual Root Mean Squared Error ($\text{RMSE}_{Factual}$). Let $\tau_i = \mu_1(\mathbf{x}_i) - \mu_0(\mathbf{x}_i)$ and $\hat{\tau}_i = \hat{\mu}_1(\mathbf{x}_i) - \hat{\mu}_0(\mathbf{x}_i)$ denote the true and estimated ITEs for individual i, respectively. The metrics are defined as

$$\sqrt{\epsilon_{\text{PEHE}}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\tau}_i - \tau_i\right)^2}, \tag{112}$$

where $\sqrt{\epsilon_{\text{PEHE}}}$ quantifies the discrepancy between the estimated and true treatment effects across individuals.

$$\epsilon_{\text{ATE}} = \left|\frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i - \frac{1}{N}\sum_{i=1}^{N}\tau_i\right|, \tag{113}$$

where ϵ_{ATE} measures the deviation between the estimated and true average treatment effects.

The factual RMSE is defined as

$$\text{RMSE}_{Factual} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_{a_i}(\mathbf{x}_i) - y_i^{\text{F}}\right)^2}, \tag{114}$$

where y_i^{F} denotes the observed factual outcome and $\hat{\mu}_{a_i}(\mathbf{x}_i)$ is the corresponding predicted factual outcome. Thus, $\text{RMSE}_{Factual}$ reflects the predictive accuracy of the learned potential-outcome functions on observed outcomes. Lower values indicate better performance for all three metrics.

E ADDITIONAL EXPERIMENTS.

E.1 ADDITIONAL EXPERIMENTS ON MIMIC-TIT DATASET.

Table 4: Key component analysis on MIMIC-TIT. Results are reported as mean (std). Best results are shown in **bold**.

Setting	**Anchor**	**Balancing**	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\text{RMSE}_{Factual}$
w/o Anchor	✗	Confounder	0.354 (0.040)	0.182 (0.051)	1.244 (0.013)
No Balance	✓	None	0.378 (0.050)	0.194 (0.067)	1.216 (0.007)
Global Balance	✓	all	0.346 (0.030)	0.167 (0.045)	1.217 (0.009)
MultiCFR (Ours)	✓	Confounder	**0.338 (0.032)**	**0.120 (0.021)**	**1.198 (0.008)**

×: Not used; ✓: Used.

Figure 4 further examines patient-level treatment effect estimation. Compared with No Balance, MultiCFR produces ITE predictions that are more concentrated around the ideal line $\hat{\tau} = \tau$, indicating better calibration between predicted and true effects. This improvement is also reflected in

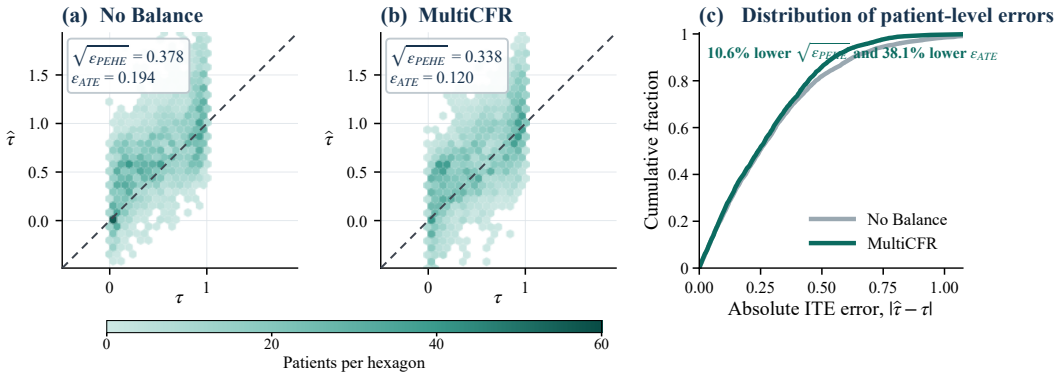

Figure 4: ITE calibration and patient-level error comparison.

the error distribution, where MultiCFR yields smaller patient-level absolute ITE errors for a larger fraction of samples. Overall, MultiCFR reduces $\sqrt{\epsilon_{\text{PEHE}}}$ from 0.378 to 0.338 and ϵ_{ATE} from 0.194 to 0.120, corresponding to relative reductions of 10.6% and 38.1%, respectively.

E.2 Additional Experiments on IHDP-TIT dataset.

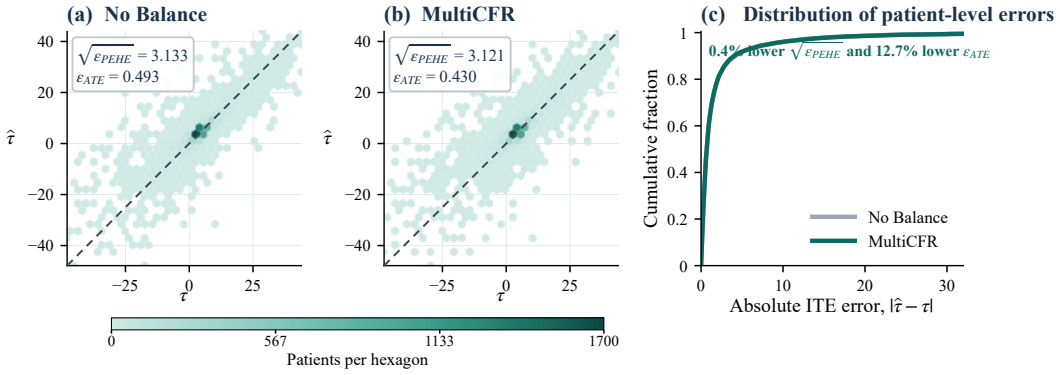

Figure 5: ITE calibration and patient-level error comparison on IHDP-TIT dataset.

Figure 5 compares MultiCFR with its no-balancing variant at the patient level. MultiCFR produces ITE predictions that are more closely aligned with the ideal line $\hat{\tau} = \tau$ and reduces $\sqrt{\epsilon_{\text{PEHE}}}$ from 3.133 to 3.121 and ϵ_{ATE} from 0.493 to 0.430, corresponding to reductions of 0.4% and 12.7%, respectively. Together, the two analyses show that neither excessive balancing nor the absence of balancing is desirable, supporting our selective balancing strategy that reduces treatment-group discrepancy while preserving predictive information.

Figure 6 further examines the relationship between representation balance and treatment-effect estimation on IHDP-TIT. A smaller distribution discrepancy does not necessarily lead to a lower PEHE. For example, Bi-AdaIN achieves the smallest MMD (0.0048) but a much larger PEHE (7.913), whereas MultiCFR obtains the lowest PEHE (3.121) with a moderate MMD (0.2019). Compared with MixingInsights, MultiCFR further reduces PEHE from 3.646 to 3.121 despite having a slightly larger MMD. These results indicate that aggressively pursuing distribution overlap may suppress information useful for treatment-effect estimation.

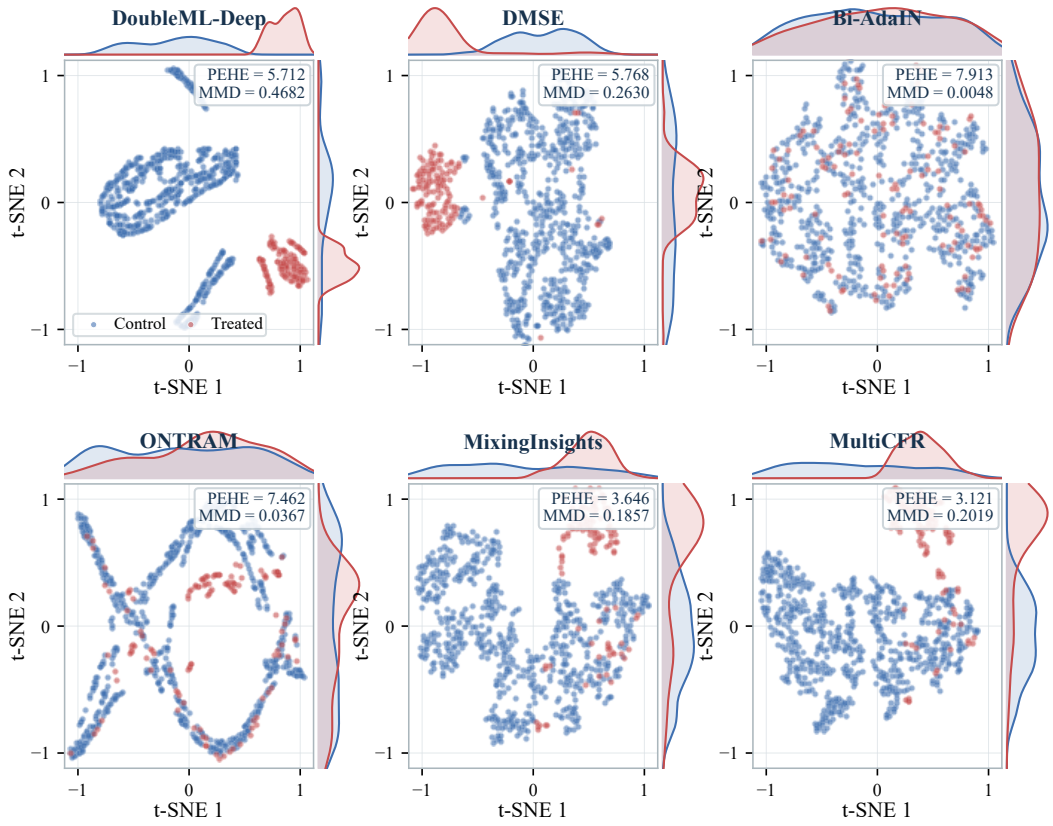

Figure 6: Visualization of treatment–control representation balance and ITE estimation performance across different methods on IHDP-TIT dataset.