DSCF: Dual-Source Counterfactual Fusion for High-Dimensional Combinatorial Interventions

Jitong Dou[1], **Lingrui Luo**[3], **Bing Zhu**[3], **Hengliang Luo**[3], **Mingjun Zhong**[2], **Yurong Cheng**[1]
[1]Beijing Institute of Technology
[2]University of Aberdeen
[3]Meituan
doujitong@bit.edu.cn, {luolingrui, zhubing04, luohengliang}@meituan.com, yrcheng@bit.edu.cn, mingjun.zhong@abdn.ac.uk

Abstract

Estimating counterfactual outcomes from observational data is critical for informed decision-making in domains such as personalized marketing, healthcare, and online platforms. In these contexts, decision processes frequently involve high-dimensional combinatorial interventions, including bundled channel allocation or product set recommendations. For such scenarios, both causal assessment of historical strategies and optimization of novel interventions necessitate models capable of extrapolating to intervention combinations that are underrepresented or entirely absent in observational data. Specifically, in digital marketing, companies often need to evaluate new combinations of channels or target emerging user segments that have not been previously exposed. This challenge is exacerbated by inherent biases in observational datasets, stemming from prior allocation policies and targeting mechanisms, which further aggravate coverage sparsity and compromise off-support counterfactual inference. In this work, we propose Dual-Source Counterfactual Fusion (DSCF), a scalable framework that enables accurate counterfactual prediction under high-dimensional combinatorial interventions, with improved robustness to confounding bias. DSCF jointly models observational data and proxy counterfactual samples through a dual-head mixture-of-experts architecture and domain-guided fusion. This design effectively integrates bias reduction and information diversity while enabling adaptive generalization to counterfactual inputs. Extensive experiments on both synthetic and semi-synthetic datasets demonstrate the effectiveness and robustness of DSCF across diverse scenarios.

Introduction

Understanding the joint effects of multiple interdependent interventions is increasingly critical to decision-making in domains such as online platforms and digital marketing (Yao et al. 2022). Figure 1 illustrates a typical digital marketing scenario: advertisers determine exposure strategies based on user characteristics (e.g., demographics and behavioral signals), which in turn influence the set of marketing channels a user is exposed to. These channel combinations, together with user intent, influence downstream business outcomes such as conversion rate and long-term retention. To enable

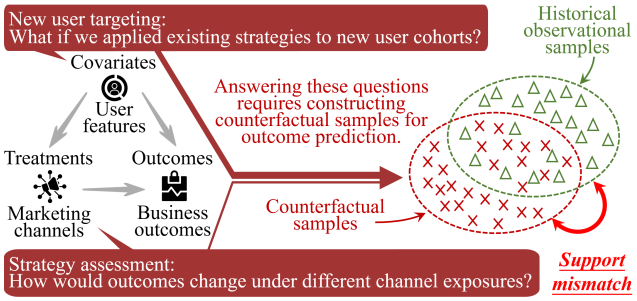

Figure 1: Illustration of key counterfactual questions and support mismatch in high-dimensional intervention settings.

fine-grained business analysis and future strategic optimization, a core counterfactual question is: how would these outcomes change if we assigned the user a different combination of channels? Modeling each channel in isolation fails to capture the synergistic or antagonistic effects that arise from their co-occurrence, necessitating a shift from single-intervention analysis to combinatorial counterfactual modeling.

However, historical observational data only documents outcomes for a limited and biased subset of intervention combinations, specifically those deployed under prior allocation policies that systematically prioritize safe, high-performing strategies aligned with historically targeted user segments. This policy-driven selection mechanism results in systematic selection bias and coverage sparsity. Furthermore, the inherent low-rank structure of user populations exacerbates this challenge: demographically or behaviorally similar users tend to receive homogenized treatments, leaving extensive regions of the combinatorial intervention space underexplored. Under such conditions, direct empirical learning from observational data is insufficient to support forward-looking strategy development (e.g., targeting novel user cohorts, reconfiguring campaign bundles, or simulating budget reallocations). Without the capacity for out-of-support generalization, such models are inherently incapable of providing actionable insights for future decision-making.

This need for off-support counterfactual generalization under high-dimensional, biased conditions presents a critical yet underexplored challenge, particularly in internet ap-

plications such as high-value action discovery (M-Squared 2025), multi-touch attribution (Ren et al. 2018; Arava et al. 2018; Yao et al. 2022), and ad optimization (Shi et al. 2024). Existing methods (Wang et al. 2024) are effective only for low-dimensional combinatorial interventions and fall short in industrial-scale applications with high-dimensional interventions, due to poor off-support generalization, restrictive assumptions, or prohibitive resource consumption. To mitigate these limitations, we propose **Dual-Source Counterfactual Fusion (DSCF)**, a scalable framework for accurate and robust counterfactual prediction under support-sparse, high-dimensional combinatorial interventions. DSCF seeks to combine the low-bias nature of proxy counterfactual samples (obtained via matching) with the complementary, richer information content of observational data. It jointly learns from both domains and incorporate a domain classifier to enable input-dependent fusion. Throughout the pipeline, DSCF imposes minimal assumptions on data distribution and variable type, rendering it suitable for industrial-scale applications.

We evaluate DSCF on both synthetic and semi-synthetic benchmarks. On synthetic data, it consistently achieves improvements across diverse experimental configurations. On semi-synthetic datasets constructed from real-world user logs, it reduces RMSE and MAE by 32.1% and 48.3%, respectively, compared to the state-of-the-art methods.

Related Work

Classical ITE extensions. Traditional ITE methods, such as sample reweighting (Arbour, Dimmery, and Sondhi 2021; Chesnaye et al. 2022), matching (Stuart 2010; Schwab, Linhardt, and Karlen 2018; Wu et al. 2023), and representation learning (Shalit, Johansson, and Sontag 2017; Shi, Blei, and Veitch 2019), aim to adjust for confounding by aligning covariate distributions across treatment groups. Some efforts extend these methods to combinatorial settings by treating covariates and interventions as a joint feature space and applying above adjustment to the joint distribution. However, the exponentially large treatment space leads to severe support sparsity and renders direct adjustment ineffective. Reweighting-based methods (Zou et al. 2020) merely rescale sample weights within the observed support and cannot extrapolate beyond it (Cortes, Mansour, and Mohri 2010), while matching tends to oversample a small subset of observational samples, reducing diversity and increasing variance. Representation learning methods (Tanimoto et al. 2021) often assume invariant treatment effects across domains, which is an unrealistic premise in highly context-dependent tasks. Moreover, the learned representations tend to become non-invertible under sparse coverage(Johansson, Sontag, and Ranganath 2019), resulting in irreversible information loss and degraded estimation quality.

Advanced modeling paradigms. Recent methods tailored to combinatorial interventions include counterfactual data augmentation (Qian, Curth, and van der Schaar 2021), low-rank modeling (Agarwal, Agarwal, and Vijaykumar 2023), and meta-learning (Chauhan et al. 2025). While effective in constrained settings, these approaches often rely on strong assumptions or incur significant resource overhead, limiting their scalability to high-dimensional, support-sparse regimes. Data augmentation methods require expanding the training set by a factor of the intervention dimension and fitting a separate predictor for each intervention component, incurring substantial computational costs and storage overhead. Low-rank models capture only coarse structures and fail to represent high-order interactions prevalent in complex systems. While meta-learning approaches offer adaptability, their reliance on nested optimization and heavily parameterized architectures hinders scalability and applicability in high-dimensional settings.

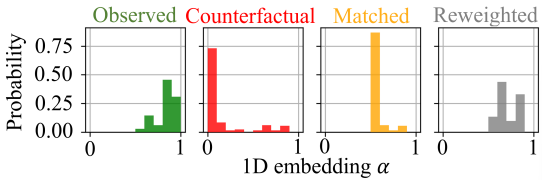

Figure 2: Empirical distributions of observed, counterfactual, matched, and reweighted sample features over a latent one-dimensional space.

Problem Statement

In combinatorial counterfactual prediction, the goal is to predict outcomes under different combinations of interventions and contexts, based on observational data. The observational data is denoted as $D_{obs} = \{(\mathbf{x}_i, \mathbf{t}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents covariates (e.g., user demographics), $\mathbf{t}_i \in \{0,1\}^p$ represents treatment assignments, and $y_i \in \mathbb{R}$ is the observed outcome (e.g., conversion rate). Each element of $\mathbf{t}_i$, referred to as a *cause*, indicates the presence or absence of a specific intervention (such as whether a particular marketing channel was accessed). The full vector $\mathbf{t}_i$, consisting of all causes, represents the *treatment*, the joint assignment of all binary interventions applied to a given unit. We aim to learn a hypothesis $f_\theta : \mathbb{X} \times \mathbb{T} \to \mathbb{R}$ which predicts the outcome y based on both covariate $\mathbf{x}$ and treatment $\mathbf{t}$. We use binary treatments for clarity, although the method could be applied to more general intervention types, including dense or categorical inputs.

Although the exact form of the counterfactual distribution may vary across applications, eliminating confounding between covariates and causes remains a universal objective. As a practical approximation, we adopt a factorized form that assumes independence between covariates and causes. Specifically, we aim to minimize the expected loss under a factorized counterfactual distribution $P(\mathbf{X}) \prod_{i=1}^p P(T^i)$: $\mathbb{E}_{P(\mathbf{X}) \prod_{i=1}^p P(T^i)} \left[\mathcal{L}(f_\theta(\mathbf{X}, \mathbf{T}), y(\mathbf{X}, \mathbf{T}))\right]$, where $\mathcal{L}(\cdot, \cdot)$ is the error function and $y(\cdot, \cdot)$ denotes the true outcome (Zou et al. 2020). Given the combinatorial explosion of the intervention space, the inherent bias in observational data, and practical requirements for generalizing to off-support counterfactual scenarios, we do not make the positivity assumption (Rosenbaum and Rubin 1983; Pearl 2010), which requires every treatment to have a non-zero probability of being observed.

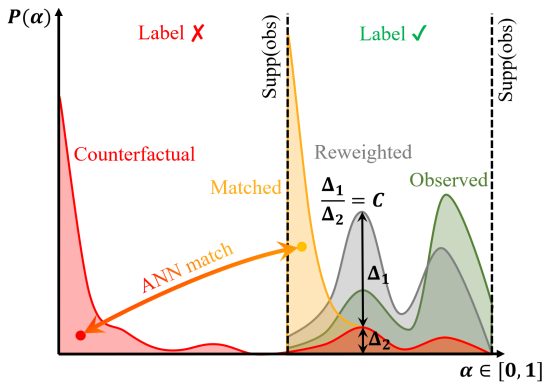

Figure 3: Schematic illustration of how reweighting and matching respond to support mismatch over the same space.

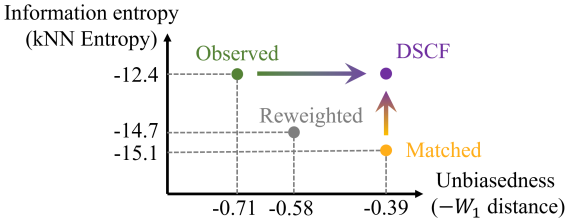

Figure 4: Illustration of the trade-off between unbiasedness and information richness across training distributions.

Method

In this section, we first illustrate the distributional heterogeneity among observational, reweighted, and matched data through a motivating example. This insight motivates the design of DSCF, which jointly learns from observational and matched data to harness their complementary strengths. We then introduce its core components and deployment details.

Limitations under Support Constraints

While matching and reweighting are widely used for counterfactual prediction under combinatorial interventions, they exhibit notable limitations when the positivity assumption fails. We begin with an illustrative example in Figure 2. Here, high-dimensional feature vectors $(\mathbf{x}, \mathbf{t})$ from real-world user logs are projected onto a one-dimensional latent axis $\alpha \in [0, 1]$, partitioned into 10 equal-width bins. The y-axis reflects the sample proportion in each bin. Observational samples (drawn from the logged dataset directly) and counterfactual samples (constructed by independently permuting the treatment columns of the observational data) exhibit clear support mismatch: the former are concentrated in $[0.5, 1.0]$, while the latter are primarily located in $[0.0, 0.5]$.

We next apply reweighting and matching to adjust the observational training distribution toward the target counterfactual distribution. As shown in Figure 3, reweighting rescales the weights of in-support observational samples based on their counterfactual likelihood but cannot, by design, extrapolate to out-of-support regions. Matching, by contrast, more closely approximates the counterfactual distribution, but often over-concentrates in narrow regions—for example, collapsing all probability mass from $[0.0, 0.5]$ onto the single point $\alpha = 0.5$.

To further assess these trade-offs, we evaluate observational, matched, and reweighted training distributions along two axes: **unbiasedness** (measured by the negative Wasserstein-1 distance to the counterfactual distribution) and **information richness** (measured by kNN entropy). As shown in Figure 4, matching achieves the lowest bias but suffers from substantial entropy loss due to oversampling a narrow subset of observational units—a direct consequence of positivity violations. In contrast, observational data, while biased, retains higher entropy, indicating greater information diversity. Given that the potential outcome mapping $(\mathbf{x}, \mathbf{t}) \mapsto y$ is assumed invariant across domains, our proposed method, DSCF, seeks to address this trade-off by jointly learning from both sources. It combines the bias reduction offered by matching with the informational diversity of observational data to enable robust and accurate counterfactual prediction.

Dual-Source Counterfactual Fusion

DSCF implements this dual-source strategy via a dual-head Multi-gate Mixture-of-Experts (MMoE) (Ma et al. 2018) architecture, with a domain classifier adaptively fusing the two heads at inference time. An overview is shown in Figure5.

Proxy counterfactual dataset construction. To obtain a training subset that is minimally biased with respect to the target counterfactual distribution, we construct a proxy counterfactual dataset via matching, aiming to approximate the joint distribution over $\mathbb{X} \times \mathbb{T}$ as closely as possible. We first independently shuffle each column of the observational treatment matrix and concatenate the result with the original covariates, yielding feature vectors $\{(\bar{\mathbf{x}}_i, \bar{\mathbf{t}}_i)\}_{i=1}^n \sim P(\mathbf{X}) \prod_{i=1}^p P(T^i)$. We then perform approximate nearest neighbor (ANN) search, using the synthetic inputs $(\bar{\mathbf{x}}, \bar{\mathbf{t}})$ to query the observational dataset D_{obs} in the joint feature space $\mathbb{X} \times \mathbb{T}$. For each query, we include the entire matched observational sample $(\mathbf{x}', \mathbf{t}', y')$ in the proxy dataset $D_{\text{pcf}} = \{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{t}}_i, \tilde{y}_i)\}_{i=1}^n$, so that the mapping $(\mathbf{x}, \mathbf{t}) \mapsto y$ remains faithful to the original data-generating process. We employ industrial-grade ANN engines such as FAISS (Johnson, Douze, and Jégou 2019) for scalable retrieval.

Dual-head joint learning with MMoE. We adopt MMoE as our main prediction module, consisting of a shared set of K expert networks $\{E_k(\cdot)\}_{k=1}^K$ and two task-specific gating networks, G^{obs} and G^{pcf}. The former facilitates cross-domain knowledge sharing and supervision, while the latter assigns each input to a soft combination of experts to enable adaptive specialization. Let $\mathbf{z}_{\text{obs}} = [\mathbf{x}; \mathbf{t}]$ and $\mathbf{z}_{\text{pcf}} = [\tilde{\mathbf{x}}; \tilde{\mathbf{t}}]$ denote the inputs from observational and proxy data (i.e., proxy counterfactual data), respectively. Each gating network produces softmax-normalized expert weights:

$$\mathbf{g}^{\text{obs}} = \sigma(G^{\text{obs}}(\mathbf{z}_{\text{obs}})), \quad \mathbf{g}^{\text{pcf}} = \sigma(G^{\text{pcf}}(\mathbf{z}_{\text{pcf}})),$$

where $\sigma(\cdot)$ denotes the softmax activation. These weights generate task-specific representations as weighted combina-

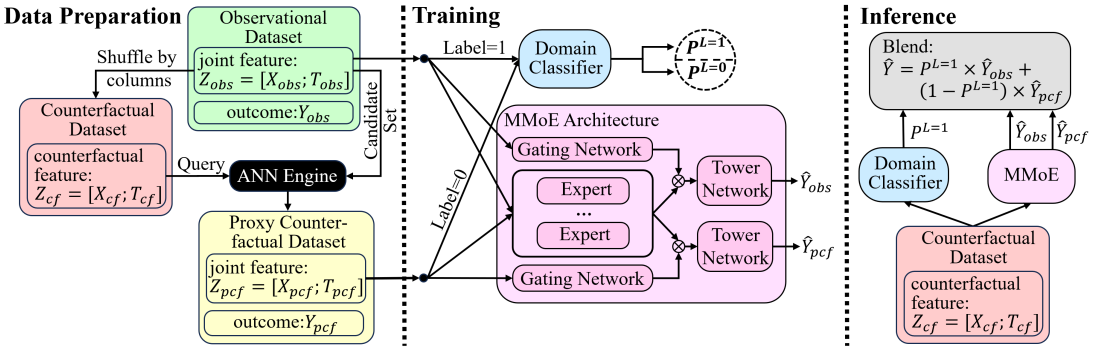

Figure 5: DSCF framework overview: data preparation, training, and inference. During inference, new samples drawn from the same counterfactual distribution as in the data preparation stage are fed into the model for prediction.

tions of expert outputs:

$$\mathbf{h}^{\text{obs}} = \sum_{k=1}^{K} g_k^{\text{obs}} \cdot E_k(\mathbf{z}_{\text{obs}}), \quad \mathbf{h}^{\text{pcf}} = \sum_{k=1}^{K} g_k^{\text{pcf}} \cdot E_k(\mathbf{z}_{\text{pcf}}).$$

Final predictions are produced by task-specific output towers:

$$\hat{y}^{\text{obs}} = o_{\text{obs}}(\mathbf{h}^{\text{obs}}), \quad \hat{y}^{\text{pcf}} = o_{\text{pcf}}(\mathbf{h}^{\text{pcf}}).$$

During training, $\hat{y}^{\text{obs}}$ and $\hat{y}^{\text{pcf}}$ are supervised using samples from D_{obs} and D_{pcf}, respectively.

Domain-guided prediction fusion. Given that observational and proxy counterfactual data originate from different regions of the true counterfactual distribution, we introduce a domain classifier $g_{\text{cls}}(\cdot)$ to fuse the supervision signals provided by both prediction heads. Specifically, $g_{\text{cls}}(\cdot)$ takes the combined input $\mathbf{z} = [\mathbf{x}; \mathbf{t}]$ and outputs a confidence score: $\alpha = \sigma(g_{\text{cls}}(\mathbf{z}))$, where $\sigma(\cdot)$ denotes the sigmoid function and $\alpha \in [0, 1]$ represents the probability that the input comes from the observational domain. We assign domain labels $L \in \{0, 1\}$, with $L = 1$ for observational samples and $L = 0$ for proxy counterfactuals. The classifier is trained independently using balanced pairs $(\mathbf{z}_{\text{obs}}, 1)$ and $(\mathbf{z}_{\text{pcf}}, 0)$, ensuring no prior bias. We deliberately decouple domain classification from the MMoE framework to prevent interference from the inductive biases of the prediction heads through gradient propagation, and to improve training stability.

Training and inference. The MMoE-based prediction module is optimized to minimize the expected supervised loss over both data sources:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{(\mathbf{x},\mathbf{t},y)\sim D_{\text{obs}}} \left[\mathcal{L}(\hat{f}_{\text{obs}}(\mathbf{x}, \mathbf{t}), y) \right] + \mathbb{E}_{(\tilde{\mathbf{x}},\tilde{\mathbf{t}},\tilde{y})\sim D_{\text{pcf}}} \left[\mathcal{L}(\hat{f}_{\text{pcf}}(\tilde{\mathbf{x}}, \tilde{\mathbf{t}}), \tilde{y}) \right],$$

where $\hat{f}_{\text{obs}}$ and $\hat{f}_{\text{pcf}}$ denote the two prediction routes within the MMoE:

$$\hat{f}_{\text{obs}}(\mathbf{x}, \mathbf{t}) := o_{\text{obs}} \left(\sum_{k=1}^{K} g_k^{\text{obs}} \cdot E_k([\mathbf{x}; \mathbf{t}]) \right),$$

$$\hat{f}_{\text{pcf}}(\tilde{\mathbf{x}}, \tilde{\mathbf{t}}) := o_{\text{pcf}} \left(\sum_{k=1}^{K} g_k^{\text{pcf}} \cdot E_k([\tilde{\mathbf{x}}; \tilde{\mathbf{t}}]) \right),$$

with E_k denoting the shared experts and $g_k^{\text{obs}}, g_k^{\text{pcf}}$ the task-specific gating weights. The domain classifier is trained independently using binary cross-entropy: $\mathcal{L}_{\text{cls}} = \mathbb{E}_{(\mathbf{z},L)} \left[\mathcal{L}_{\text{CE}}(\sigma(g_{\text{cls}}(\mathbf{z})), L) \right]$, where $\mathcal{L}_{\text{CE}}$ denotes the binary cross-entropy loss, $\mathbf{z} = [\mathbf{x}; \mathbf{t}]$, and $L \in \{0, 1\}$ indicates the domain label. At inference, final predictions are fused using the domain affinity:

$$\hat{f}_{\text{DSCF}}(\mathbf{x}, \mathbf{t}) = \alpha(\mathbf{x}, \mathbf{t}) \cdot \hat{f}_{\text{obs}}(\mathbf{x}, \mathbf{t}) + (1 - \alpha(\mathbf{x}, \mathbf{t})) \cdot \hat{f}_{\text{pcf}}(\mathbf{x}, \mathbf{t}),$$

where $\alpha(\mathbf{x}, \mathbf{t}) := \sigma(g_{\text{cls}}([\mathbf{x}; \mathbf{t}]))$ denotes the learned domain affinity.

Theoretical Justification

We provide a theoretical justification for the DSCF framework by analyzing its expected risk under the true counterfactual distribution P_{cf}. Let $\hat{f}_{\text{DSCF}}$ be the final fused prediction and $\mathcal{L} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ a pointwise loss function.

Assumptions. We assume the loss function $\mathcal{L}(y, \hat{y})$ is L_ℓ-Lipschitz in $\hat{y}$ and bounded by B_ℓ. The true outcome function $y(\mathbf{x}, \mathbf{t})$ is L_y-Lipschitz, and both prediction heads $\hat{f}_{\text{obs}}$, $\hat{f}_{\text{pcf}}$ are L_f-Lipschitz and bounded in output by B. The domain classifier $\alpha(\mathbf{x}, \mathbf{t}) := \sigma(g_{\text{cls}}([\mathbf{x}; \mathbf{t}]))$ has classification error at most ε_{cls} over a balanced mixture of D_{obs} and D_{pcf}.

Theorem 1 (Counterfactual Risk Bound for DSCF) *Let P_{cf} denote the true counterfactual distribution and P_{pcf} the proxy counterfactual distribution constructed via approximate matching. Let $\mathcal{L}_{\text{obs}}(\mathbf{x}, \mathbf{t}) := \mathcal{L}(\hat{f}_{\text{obs}}(\mathbf{x}, \mathbf{t}), y(\mathbf{x}, \mathbf{t}))$ and $\mathcal{L}_{\text{pcf}}(\mathbf{x}, \mathbf{t}) := \mathcal{L}(\hat{f}_{\text{pcf}}(\mathbf{x}, \mathbf{t}), y(\mathbf{x}, \mathbf{t}))$. Then the expected counterfactual risk satisfies:*

$$\mathbb{E}_{(\mathbf{x},\mathbf{t})\sim P_{\text{cf}}} \left[\mathcal{L}\big(\hat{f}_{\text{DSCF}}(\mathbf{x}, \mathbf{t}), y(\mathbf{x}, \mathbf{t})\big) \right] \leq \underbrace{\mathbb{E}_{(\mathbf{x},\mathbf{t})\sim P_{\text{pcf}}} \left[\min \left\{ \mathcal{L}_{\text{obs}}(\mathbf{x}, \mathbf{t}),\ \mathcal{L}_{\text{pcf}}(\mathbf{x}, \mathbf{t}) \right\} \right]}_{\textit{oracle prediction}} + \underbrace{\varepsilon_{\text{proxy}}}_{\textit{proxy bias}} + \underbrace{B_{\mathcal{L}} \cdot \varepsilon_{\text{cls}}}_{\textit{fusion penalty}}.$$

where $\varepsilon_{\text{proxy}} := L_\ell(L_y + L_f) \cdot \varepsilon_{\text{ANN}}$, with ε_{ANN} denoting the maximal distance between a target counterfactual input and its matched proxy neighbor. The fusion penalty term is bounded by $B_{\mathcal{L}} \cdot \varepsilon_{\text{cls}}$, where $B_{\mathcal{L}} = 2L_\ell B$.

Each term in the bound reflects a key design component of DSCF:

- *Oracle prediction:* reflects the benefit of enlarging the hypothesis space. Adding an extra head increases the chance that at least one yields lower pointwise risk for a given input, offering a tighter ideal risk baseline than using a single domain alone. Additional gains from cross-domain knowledge sharing are demonstrated empirically.
- *Proxy bias:* measures the discrepancy between P_{cf} and P_{pcf}, bounded when ANN matching yields geometrically close neighbors, even under positivity violations.
- *Fusion penalty:* captures the extra risk due to imperfect fusion; smaller $\varepsilon_{\mathrm{cls}}$ implies a tighter overall bound.

Remark 1 (Distributional Advantage over Reweighting) *Proxy matching achieves a strictly smaller 1-Wasserstein distance to the true counterfactual distribution than permutation weighting under typical positivity violation. This advantage stems from proxy matching's ability to directly minimize transport cost by approximating off-support mass through nearest neighbors. In contrast, reweighting normalizes within-support density without perceiving unseen regions, leading to larger distributional discrepancy. Consequently, proxy matching induces a tighter risk bound under standard Lipschitz conditions.*

Experiments

We evaluate the proposed DSCF framework on both synthetic and semi-synthetic datasets to assess its effectiveness in counterfactual prediction under high-dimensional combinatorial interventions. We compare DSCF against representative baselines and analyze the impact of its core components through detailed result analysis and ablation studies.

Experiment Setup

Baselines. We compare against the following baselines. **S-Learner** and **$\mathbf{NN_{pcf}}$** are supervised models trained on observational and proxy counterfactual data, respectively. **PW** (Arbour, Dimmery, and Sondhi 2021) and **VSR** (Zou et al. 2020) reweight observational samples to match the counterfactual joint distribution over covariates and interventions. **RMNet** (Tanimoto et al. 2021) learns domain-invariant representations. **H-Learner** (Chauhan et al. 2025) is a meta-learning approach for multi-intervention, multi-outcome settings, adapted here to single-outcome prediction. **DSCF-Sep** is a variant of our model that disables joint training and directly fuses S-Learner and $\mathrm{NN_{pcf}}$ using a domain classifier. We exclude **SCP** (Qian, Curth, and van der Schaar 2021), which requires p separate models and p-fold data augmentation, making it impractical for high-dimensional combinatorial interventions. **Synthetic Combinations (SC)** (Agarwal, Agarwal, and Vijaykumar 2023) is also excluded, as its idealized setting does not align with our problem context, resulting in poor performance.

Implementation details. For fair comparison, the predictive models used in S-Learner, $\mathrm{NN_{pcf}}$, PW, VSR, and RMNet all adopt a 4-layer MLP with 128 hidden units per layer. In DSCF, each expert shares the first two layers of this architecture, with the number of experts set to 5. For H-Learner, the hypernetwork width is set to $4p$ to ensure sufficient capacity. All models are trained with the same number of epochs, learning rate, optimizer, and batch size.

Evaluation metrics. We evaluate all models on a held-out test set drawn from the factorized counterfactual distribution $P(\mathbf{X})\prod_{i=1}^{p}P(T^i)$. Metrics include Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), averaged over 5 random seeds.

Synthetic Experiment on Fully Controlled Data

Dataset. We construct a synthetic dataset to evaluate counterfactual prediction under high-dimensional combinatorial interventions with fully controlled ground truth and realistic data characteristics. Each sample consists of a covariate vector $\mathbf{x} \in \mathbb{R}^d$, a binary treatment vector $\mathbf{t} \in \{0,1\}^p$, and a real-valued outcome $y \in \mathbb{R}$.

Covariates. To mimic the structure of real-world user data (e.g., long-tailed, low-rank, and nonlinear), we first sample a latent vector $\mathbf{z} \in \mathbb{R}^r$ from a 50-component Gaussian Mixture Model (GMM), where $\mathbf{z} \sim \sum_{k=1}^{50}\pi_k \cdot \mathcal{N}(\mu_k, \Sigma_k)$ and $\pi_k \propto 1/k^\alpha$ with $\alpha = 1.5$, such that the cluster weights follow a Zipf distribution. The latent vector is mapped to the covariate space via a linear projection and nonlinear transformation: $\mathbf{x} = \mathbf{W}_{\mathrm{up}}\mathbf{z} + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 0.01^2\mathbf{I})$. For nonlinearity, the first third of features undergo $\tanh$, and the second third use exponential transformation.

Treatments. Each treatment dimension is assigned independently using a confounded logistic model: $P(T^j = 1 \mid \mathbf{x}) = \sigma(\gamma \cdot \mathbf{x}^\top\beta^{(j)} + \eta_j)$, where $\eta_j \sim \mathcal{N}(0, 0.1^2)$ and γ controls the confounding strength.

Outcomes. The outcome combines additive effects, interaction terms, and nonlinearities: $y = \mathbf{x}^\top\beta_x + \mathbf{t}^\top\beta_t + \sum_{(i,j)\in\mathcal{I}}\left(2 \cdot t_i t_j \cdot \alpha_{ij} + 0.3 \cdot t_i \cdot \mathbf{x}^\top\gamma_{ij}\right) + 0.2 \cdot \phi(\mathbf{x}, \mathbf{t}) + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 0.1^2)$ and $\phi(\mathbf{x}, \mathbf{t}) = \sin(\mathbf{x}^\top w_x) + \exp(-|\mathbf{t}^\top w_t|)$. The interaction set $\mathcal{I}$ includes $\lceil c \cdot p/2 \rceil$ random treatment pairs, with $c = 5$ controlling outcome complexity.

We vary the number of treatments $p \in \{10, 20, 30\}$ and the confounding strength $\gamma \in \{0.1, 0.3, 0.5, 1.0\}$. Each training set contains $p \times 10{,}000$ samples, and the corresponding test set is drawn from the factorized counterfactual distribution with the same sample size.

Results. Table 1 reports RMSE and MAE of all methods under varying numbers of intervention components and confounding strengths. Overall, DSCF achieves the best performance on 20 out of 24 metrics, and ranks second on the remaining 4, consistently outperforming all baselines.

Under low confounding ($\gamma = 0.1$), H-Learner slightly outperforms DSCF at $p = 30$, but as γ increases, it becomes unstable—its RMSE at $p = 20$ under $\gamma = 0.3$ and $\gamma = 1.0$ exceeds 15 and 40 respectively. In contrast, DSCF remains robust and consistently outperforms its variants and all baselines. Notably, $\mathrm{NN_{pcf}}$ often surpasses more sophisticated alternatives such as PW, VSR, and RMNet. This observation supports our hypothesis that reweighting

Method	$\gamma = 0.1$		$\gamma = 0.3$		$\gamma = 0.5$		$\gamma = 1.0$	
	RMSE $\pm\ \sigma$	MAE $\pm\ \sigma$	RMSE $\pm\ \sigma$	MAE $\pm\ \sigma$	RMSE $\pm\ \sigma$	MAE $\pm\ \sigma$	RMSE $\pm\ \sigma$	MAE $\pm\ \sigma$
$p = 10$								
S-Learner	$3.952_{\pm 0.160}$	$1.434_{\pm 0.040}$	$5.499_{\pm 0.013}$	$1.834_{\pm 0.008}$	$6.164_{\pm 0.057}$	$2.197_{\pm 0.035}$	$6.265_{\pm 0.042}$	$2.574_{\pm 0.010}$
NN_{pcf}	$3.776_{\pm 0.172}$	$1.385_{\pm 0.037}$	$5.330_{\pm 0.020}$	$1.632_{\pm 0.032}$	$5.682_{\pm 0.022}$	$1.776_{\pm 0.003}$	$6.130_{\pm 0.055}$	$2.109_{\pm 0.007}$
PW	$4.339_{\pm 0.098}$	$1.485_{\pm 0.015}$	$5.641_{\pm 0.040}$	$1.823_{\pm 0.017}$	$5.979_{\pm 0.076}$	$2.079_{\pm 0.023}$	$6.359_{\pm 0.130}$	$2.652_{\pm 0.061}$
VSR	$4.072_{\pm 0.054}$	$1.455_{\pm 0.004}$	$5.625_{\pm 0.077}$	$1.864_{\pm 0.008}$	$6.009_{\pm 0.092}$	$2.130_{\pm 0.018}$	$6.301_{\pm 0.031}$	$2.662_{\pm 0.009}$
RMNet	$3.869_{\pm 0.077}$	$1.437_{\pm 0.011}$	$5.444_{\pm 0.086}$	$1.827_{\pm 0.037}$	$\underline{5.657}_{\pm 0.037}$	$2.133_{\pm 0.027}$	$6.227_{\pm 0.017}$	$2.622_{\pm 0.012}$
H-Learner	$3.766_{\pm 0.027}$	$\underline{1.121}_{\pm 0.012}$	$5.743_{\pm 0.244}$	$1.663_{\pm 0.086}$	$6.532_{\pm 0.023}$	$1.856_{\pm 0.013}$	$\underline{5.621}_{\pm 0.015}$	$2.165_{\pm 0.011}$
DSCF-Sep (ours)	$\underline{3.761}_{\pm 0.131}$	$1.229_{\pm 0.024}$	$\underline{5.329}_{\pm 0.014}$	$\underline{1.554}_{\pm 0.018}$	$5.790_{\pm 0.017}$	$\underline{1.771}_{\pm 0.011}$	$6.082_{\pm 0.036}$	$\underline{2.097}_{\pm 0.001}$
DSCF (ours)	$\mathbf{3.100}_{\pm 0.026}$	$\mathbf{0.832}_{\pm 0.017}$	$\mathbf{4.469}_{\pm 0.152}$	$\mathbf{1.132}_{\pm 0.003}$	$\mathbf{5.112}_{\pm 0.200}$	$\mathbf{1.350}_{\pm 0.048}$	$\mathbf{5.207}_{\pm 0.066}$	$\mathbf{1.640}_{\pm 0.042}$
$p = 20$								
S-Learner	$4.459_{\pm 0.094}$	$2.081_{\pm 0.037}$	$6.423_{\pm 0.277}$	$2.476_{\pm 0.079}$	$5.588_{\pm 0.067}$	$2.568_{\pm 0.074}$	$7.849_{\pm 0.168}$	$3.494_{\pm 0.008}$
NN_{pcf}	$4.580_{\pm 0.057}$	$2.036_{\pm 0.031}$	$6.178_{\pm 0.044}$	$2.134_{\pm 0.018}$	$5.642_{\pm 0.101}$	$2.482_{\pm 0.109}$	$7.393_{\pm 0.216}$	$3.083_{\pm 0.061}$
PW	$4.892_{\pm 0.154}$	$1.998_{\pm 0.003}$	$6.114_{\pm 0.133}$	$2.393_{\pm 0.008}$	$5.541_{\pm 0.356}$	$2.692_{\pm 0.134}$	$7.654_{\pm 0.242}$	$3.302_{\pm 0.037}$
VSR	$4.361_{\pm 0.092}$	$2.016_{\pm 0.055}$	$6.286_{\pm 0.248}$	$2.348_{\pm 0.028}$	$\underline{5.470}_{\pm 0.020}$	$2.609_{\pm 0.032}$	$7.638_{\pm 0.126}$	$3.250_{\pm 0.043}$
RMNet	$4.827_{\pm 0.132}$	$2.085_{\pm 0.027}$	$6.129_{\pm 0.205}$	$2.235_{\pm 0.015}$	$5.488_{\pm 0.119}$	$2.597_{\pm 0.036}$	$7.742_{\pm 0.361}$	$3.472_{\pm 0.126}$
H-Learner	$\mathbf{3.469}_{\pm 0.025}$	$\underline{1.542}_{\pm 0.016}$	$15.165_{\pm 2.051}$	$\underline{1.995}_{\pm 0.003}$	$6.756_{\pm 0.108}$	$\underline{2.237}_{\pm 0.027}$	$41.361_{\pm 4.743}$	$3.180_{\pm 0.012}$
DSCF-Sep (ours)	$4.280_{\pm 0.061}$	$1.756_{\pm 0.021}$	$\underline{6.125}_{\pm 0.125}$	$2.010_{\pm 0.024}$	$5.507_{\pm 0.084}$	$2.383_{\pm 0.088}$	$\underline{7.362}_{\pm 0.130}$	$\underline{3.066}_{\pm 0.058}$
DSCF (ours)	$\underline{3.725}_{\pm 0.153}$	$\mathbf{1.234}_{\pm 0.058}$	$\mathbf{5.462}_{\pm 0.160}$	$\mathbf{1.401}_{\pm 0.118}$	$\mathbf{4.673}_{\pm 0.105}$	$\mathbf{1.674}_{\pm 0.055}$	$\mathbf{6.640}_{\pm 0.185}$	$\mathbf{2.202}_{\pm 0.193}$
$p = 30$								
S-Learner	$1.742_{\pm 0.019}$	$1.273_{\pm 0.015}$	$4.295_{\pm 0.087}$	$2.713_{\pm 0.041}$	$9.196_{\pm 0.047}$	$5.436_{\pm 0.028}$	$11.751_{\pm 0.117}$	$7.532_{\pm 0.065}$
NN_{pcf}	$1.712_{\pm 0.013}$	$1.258_{\pm 0.009}$	$2.956_{\pm 0.007}$	$1.985_{\pm 0.020}$	$6.466_{\pm 0.031}$	$4.108_{\pm 0.036}$	$9.385_{\pm 0.014}$	$6.002_{\pm 0.021}$
PW	$1.841_{\pm 0.012}$	$1.310_{\pm 0.008}$	$3.948_{\pm 0.027}$	$2.473_{\pm 0.027}$	$7.814_{\pm 0.088}$	$4.600_{\pm 0.041}$	$11.226_{\pm 0.032}$	$7.343_{\pm 0.016}$
VSR	$1.794_{\pm 0.036}$	$1.300_{\pm 0.028}$	$4.683_{\pm 0.006}$	$2.947_{\pm 0.015}$	$7.256_{\pm 0.046}$	$4.494_{\pm 0.029}$	$11.369_{\pm 0.040}$	$7.326_{\pm 0.023}$
RMNet	$1.887_{\pm 0.054}$	$1.361_{\pm 0.030}$	$3.967_{\pm 0.040}$	$2.568_{\pm 0.023}$	$8.074_{\pm 0.109}$	$4.788_{\pm 0.056}$	$11.227_{\pm 0.047}$	$7.263_{\pm 0.033}$
H-Learner	$\mathbf{0.820}_{\pm 0.036}$	$\mathbf{0.599}_{\pm 0.025}$	$\underline{2.633}_{\pm 0.026}$	$\underline{1.701}_{\pm 0.007}$	$\mathbf{4.806}_{\pm 0.060}$	$\underline{3.115}_{\pm 0.044}$	$\underline{7.268}_{\pm 0.016}$	$\underline{5.057}_{\pm 0.013}$
DSCF-Sep (ours)	$1.527_{\pm 0.013}$	$1.078_{\pm 0.003}$	$2.931_{\pm 0.007}$	$1.943_{\pm 0.019}$	$6.463_{\pm 0.031}$	$4.098_{\pm 0.036}$	$9.384_{\pm 0.014}$	$6.001_{\pm 0.021}$
DSCF (ours)	$\underline{0.983}_{\pm 0.045}$	$\underline{0.705}_{\pm 0.021}$	$\mathbf{2.458}_{\pm 0.265}$	$\mathbf{1.517}_{\pm 0.130}$	$\underline{5.070}_{\pm 0.242}$	$\mathbf{2.919}_{\pm 0.068}$	$\mathbf{6.823}_{\pm 0.218}$	$\mathbf{4.389}_{\pm 0.126}$

Table 1: Prediction errors (RMSE and MAE) on synthetic datasets. Best results are bolded, second-best are underlined.

Method	$n = 0.1$M		$n = 8$M	
	RMSE $\pm\ \sigma$	MAE $\pm\ \sigma$	RMSE $\pm\ \sigma$	MAE $\pm\ \sigma$
S-Learner	$0.945_{\pm 0.009}$	$0.559_{\pm 0.003}$	$0.604_{\pm 0.003}$	$0.366_{\pm 0.003}$
NN_{pcf}	$0.870_{\pm 0.001}$	$0.527_{\pm 0.002}$	$0.519_{\pm 0.003}$	$0.288_{\pm 0.002}$
PW	$1.122_{\pm 0.008}$	$0.586_{\pm 0.002}$	$0.568_{\pm 0.002}$	$0.346_{\pm 0.002}$
VSR	$0.874_{\pm 0.003}$	$0.530_{\pm 0.002}$	$0.576_{\pm 0.006}$	$0.355_{\pm 0.004}$
RMNet	$0.936_{\pm 0.005}$	$0.550_{\pm 0.002}$	$0.581_{\pm 0.004}$	$0.353_{\pm 0.004}$
H-Learner	$3.890_{\pm 0.344}$	$0.577_{\pm 0.021}$	$0.732_{\pm 0.012}$	$0.455_{\pm 0.004}$
DSCF-Sep (ours)	$\underline{0.839}_{\pm 0.001}$	$\underline{0.509}_{\pm 0.001}$	$\underline{0.504}_{\pm 0.004}$	$\underline{0.279}_{\pm 0.002}$
DSCF (ours)	$\mathbf{0.668}_{\pm 0.001}$	$\mathbf{0.398}_{\pm 0.001}$	$\mathbf{0.353}_{\pm 0.002}$	$\mathbf{0.149}_{\pm 0.002}$

Table 2: Evaluation on the semi-synthetic dataset at two data scales ($n = 0.1$M / 8M) using RMSE and MAE.

and representation-based methods become unreliable under positivity violations, where observational support is insufficient. In such cases, nearest neighbor retrieval from the observational dataset—guided by permuted counterfactual queries—offers a simple yet effective way to perceive information beyond the observational support. This empirical pattern directly motivates our proxy construction strategy.

Semi-Synthetic Experiment on Real-World Data

Dataset. To evaluate DSCF on real-world data, we construct a semi-synthetic dataset based on user logs from a large-scale short-video platform. The covariates include user demographics, while the interventions consist of 50 user-content interaction variables encompassing binary indicators, categorical features, and continuous scores (e.g., play duration, clicks, shares). The prediction target is the change in the user's monthly lifetime.

To fully leverage real-world business data while retaining full control over the outcome generation process, we adopt a two-stage procedure—comprising (1) parameter estimation and (2) sample generation—to construct semi-synthetic data that: (i) preserves the empirical input distribution, (ii) captures realistic outcome variability (e.g., long-tailed behavior), (iii) avoids model-induced bias, and (iv) maintains sufficient functional complexity. Specifically, we first apply a fixed, non-trainable two-layer MLP with Kaiming initialization to the concatenated input $[\mathbf{x}; \mathbf{t}]$, producing latent representations $\mathbf{z} = f([\mathbf{x}; \mathbf{t}]) \in \mathbb{R}^k$. We then estimate a symmetric matrix $\mathbf{M} \in \mathbb{R}^{k \times k}$ by minimizing the squared error between the quadratic form $\mathbf{z}^\top \mathbf{M} \mathbf{z}$ and the observed outcome y across the observational dataset. After fixing the parameters $f(\cdot)$ and $\mathbf{M}$, we compute synthetic outcomes for both original and permuted input pairs as $y = \mathbf{z}^\top \mathbf{M} \mathbf{z} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1^2)$.

To evaluate model performance under different data regimes, we use the full dataset of 8 million samples for training and testing, and additionally report results on a low-resource subset containing 0.1 million samples.

Results. Table 2 reports the performance of all methods on the semi-synthetic dataset. DSCF consistently outperforms

id	Reg Data		Reg Model	Cls Data			Output	$\gamma = 0.1$		$\gamma = 0.3$		$\gamma = 0.5$		$\gamma = 1.0$	
	D_{obs}	D_{pcf}		D_{obs}	D_{pcf}	D_{cf}		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
(1)	✓		MLP					3.374	1.596	5.405	2.341	6.982	3.400	8.622	4.534
(2)		✓	MLP					3.356	1.560	4.821	1.917	5.930	2.789	7.636	3.731
(3)	✓	✓	MLP×2	✓	✓		Reweighted	3.189	1.354	4.795	1.836	5.920	2.751	7.610	3.722
(4)	✓		MMoE					2.988	1.155	4.604	1.725	5.695	2.475	7.516	3.688
(5)		✓	MMoE					2.782	1.097	4.473	1.559	5.431	2.253	6.995	3.211
(6)	✓	✓	MMoE				Obs Head	2.658	0.999	4.387	1.489	5.214	2.172	6.462	2.993
(7)	✓	✓	MMoE				Pcf Head	2.610	0.968	4.155	1.372	4.973	1.994	6.290	2.751
(8)	✓	✓	MMoE				avg	**2.602**	0.924	4.197	1.372	5.005	2.012	6.248	2.779
(9)	✓	✓	MMoE	✓		✓	Reweighted	2.603	0.931	4.159	1.358	4.962	1.986	6.236	2.747
(10)	✓	✓	MMoE-lite	✓	✓		Reweighted	2.664	**0.881**	4.318	1.402	5.430	2.175	6.714	3.025
(11)	✓	✓	MMoE	✓	✓		Reweighted	2.603	0.924	**4.130**	**1.350**	**4.952**	**1.981**	**6.223**	**2.743**

Table 3: Ablation results across varying confounding strengths γ, with RMSE and MAE averaged over $p \in \{10, 20, 30\}$.

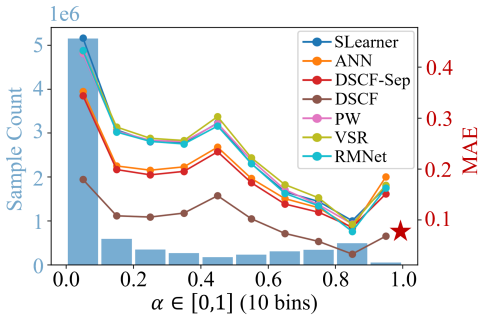

Figure 6: Model performance (MAE) across different values of domain affinity α, with sample counts shown in blue bars.

all baselines across both evaluation metrics and data scales. The performance gap between DSCF and the second-best method (NN_{pcf}) widens with increased training data: under the low-resource setting ($n = 0.1$M), DSCF reduces RMSE and MAE by 23.3% and 24.3%, respectively; under the high-resource setting ($n = 8$M), the improvements grow to 32.1% (RMSE) and 48.3% (MAE). These results highlight DSCF's superior scalability and its enhanced ability to exploit large-scale data.

Figure 6 reports the MAE of different methods across test samples grouped by their domain affinity scores α, as predicted by the domain classifier. The vast majority of samples fall into the counterfactual region ($\alpha \in [0.0, 0.1]$), indicating a pronounced distributional shift under high-dimensional combinatorial interventions, where most target configurations are rarely or never observed.

The methods exhibit a clear three-tier performance hierarchy. Traditional approaches (S-Learner, PW, RMNet, VSR) behave similarly across bins and collapse in low-α regions, highlighting their inability to handle severe positivity violations. Proxy-based methods (NN_{pcf}, DSCF-Sep) perform better in counterfactual regions by exploiting matched samples and capturing off-support structure. At the top tier, DSCF achieves the lowest MAE across all bins. Its consistent margin over proxy-based methods indicates that joint training enables complementary information flow between domains, while the proxy branch provides unbiased supervision that further regularizes the observational head, yielding gains even in high-α regions.

Ablation Study

Table 3 presents ablations to isolate the contributions of DSCF's three key components:

Proxy data. Rows (1)(2) and (4)(5) reveal that training on D_{pcf} alone already surpasses D_{obs} alone across all γ, confirming the strong signal supplied by our proxy construction.

Joint training. Comparing the single-source MMoE variants (4)(5) with their joint-training counterparts (6)(7) shows that adding the second data source consistently improves performance under the same architecture. This demonstrates that the gains arise from leveraging complementary training distributions, rather than from structural complexity alone.

Domain-guided fusion. Comparing rows (1)(2)(3) and (8)(9)(11) demonstrates that adaptive weighting consistently reduces both RMSE and MAE. The advantage widens when α is small, i.e., when observational samples still occupy a non-negligible fraction of the counterfactual domain. This highlights the fusion module's ability to exploit complementary signals in partially overlapping data.

Model capacity control. Row (10) reduces every hidden dimension of the expert, gate, and tower networks in the prediction module by half, yielding a parameter count comparable to the MLP baselines. Despite this, it still achieves superior performance, confirming that our gains stem from architectural design rather than over-parameterization.

Conclusion

We present DSCF, a principled and scalable framework for counterfactual prediction under high-dimensional combinatorial interventions. By jointly leveraging observational data and proxy counterfactual samples through a dual-head MMoE architecture and domain-guided fusion, DSCF integrates bias reduction and information richness without relying on strong structural assumptions. Theoretically, we establish a novel risk bound under the true counterfactual distribution, decomposing the estimation error into oracle prediction, proxy bias, and fusion penalty, each of which is tightly aligned with a corresponding model component. Empirically, DSCF consistently outperforms existing methods across synthetic and semi-synthetic benchmarks, demonstrating superior robustness, scalability, and generalization.

Acknowledgments

Jitong Dou and Yurong Cheng are supported by the NSFC (Grant Nos. 62472027, U21A20516, 62225203, 62427808) and by the Beijing Natural Science Foundation (L241010).

References

Agarwal, A.; Agarwal, A.; and Vijaykumar, S. 2023. Synthetic combinations: A causal inference framework for combinatorial interventions. *Advances in Neural Information Processing Systems*, 36: 19195–19216.

Arava, S. K.; Dong, C.; Yan, Z.; Pani, A.; et al. 2018. Deep neural net with attention for multi-channel multi-touch attribution. *arXiv preprint arXiv:1809.02230*.

Arbour, D.; Dimmery, D.; and Sondhi, A. 2021. Permutation weighting. In *International Conference on Machine Learning*, 331–341. PMLR.

Chauhan, V. K.; Clifton, L.; Nigam, G.; and Clifton, D. A. 2025. Individualised Treatment Effects Estimation with Composite Treatments and Composite Outcomes. *arXiv preprint arXiv:2502.08282*.

Chesnaye, N. C.; Stel, V. S.; Tripepi, G.; Dekker, F. W.; Fu, E. L.; Zoccali, C.; and Jager, K. J. 2022. An introduction to inverse probability of treatment weighting in observational research. *Clinical kidney journal*, 15(1): 14–20.

Cortes, C.; Mansour, Y.; and Mohri, M. 2010. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23.

Johansson, F. D.; Sontag, D.; and Ranganath, R. 2019. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 527–536. PMLR.

Johnson, J.; Douze, M.; and Jégou, H. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1930–1939.

M-Squared. 2025. How High Value Actions (HVAs) are Reshaping Marketing Mix Models. https://msquared.club/blogs/attribution-today/. "High Value Actions are meaningful digital behaviors that signal consumer interest, intent, or future purchase likelihood.".

Pearl, J. 2010. Causal inference. *Causality: objectives and assessment*, 39–58.

Qian, Z.; Curth, A.; and van der Schaar, M. 2021. Estimating multi-cause treatment effects via single-cause perturbation. *Advances in Neural Information Processing Systems*, 34: 23754–23767.

Ren, K.; Fang, Y.; Zhang, W.; Liu, S.; Li, J.; Zhang, Y.; Yu, Y.; and Wang, J. 2018. Learning multi-touch conversion attribution with dual-attention mechanisms for online advertising. In *Proceedings of the 27th acm international conference on information and knowledge management*, 1433–1442.

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

Schwab, P.; Linhardt, L.; and Karlen, W. 2018. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, 3076–3085. PMLR.

Shi, C.; Blei, D.; and Veitch, V. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32.

Shi, W.; Fu, C.; Xu, Q.; Chen, S.; Zhang, J.; Zhu, Q.; Hua, Z.; and Yang, S. 2024. Ads Supply Personalization via Doubly Robust Learning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 4874–4881.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1): 1.

Tanimoto, A.; Sakai, T.; Takenouchi, T.; and Kashima, H. 2021. Regret minimization for causal inference on large treatment space. In *International Conference on Artificial Intelligence and Statistics*, 946–954. PMLR.

Wang, Y.; Li, H.; Zhu, M.; Wu, A.; Xiong, R.; Wu, F.; and Kuang, K. 2024. Causal Inference with Complex Treatments: A Survey. *arXiv preprint arXiv:2407.14022*.

Wu, A.; Kuang, K.; Xiong, R.; Li, B.; and Wu, F. 2023. Stable estimation of heterogeneous treatment effects. In *International Conference on Machine Learning*, 37496–37510. PMLR.

Yao, D.; Gong, C.; Zhang, L.; Chen, S.; and Bi, J. 2022. CausalMTA: Eliminating the user confounding bias for causal multi-touch attribution. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4342–4352.

Zou, H.; Cui, P.; Li, B.; Shen, Z.; Ma, J.; Yang, H.; and He, Y. 2020. Counterfactual prediction for bundle treatment. *Advances in Neural Information Processing Systems*, 33: 19705–19715.