Counterfactual Estimation under Composite Treatments via Progressive Distribution Alignment

Abhirup Mondal **Anirban Majumder** **Vineet Chaoji**
International Machine Learning, Amazon
Bengaluru, India
{mabhirup, majumda, vchaoji}@amazon.com

Abstract

Estimating counterfactual outcomes when multiple treatments are applied simultaneously, the *composite treatment* setting, is fundamentally harder than the single-cause case: the treatment space grows combinatorially, most combinations are unobserved, and confounding is entangled across dimensions. We propose PACE (**P**rogressive **A**lignment for **C**omposite **E**ffects), which exploits a known causal DAG over treatment variables to progressively transform the observational data distribution into the target interventional distribution. PACE performs K sequential augmentation steps in reverse topological order, each isolating a single treatment component and conditioning only on its parents in the DAG. We derive the closed-form optimal augmentation density minimizing KL divergence at each step, establish finite-sample bounds prescribing the augmentation budget via DAG-aware selection bias ratios, and show that the DAG provides a deterministic processing order that eliminates the need for random permutation averaging. On semi-synthetic benchmarks with real covariates (IHDP, Twins, Hillstrom) and controlled synthetic environments, PACE substantially outperforms seven baselines on both continuous and binary outcomes, with the largest gains on rare treatment combinations where confounding is most severe. Ablations confirm consistent improvements over the closest prior method across varying DAG topologies, confounding strengths, and under DAG misspecification.

1 Introduction

Causal effect estimation from observational data underpins decision-making in medicine, public policy, and digital marketplaces. While the single-cause setting, intervening on one variable, has been studied extensively [9, 19, 20, 24], real-world decisions frequently involve *multiple simultaneous interventions*. An e-commerce platform designs email campaigns by jointly deciding whether to send, which product category to feature, and which channel to target. A clinician adjusts both drug dosage and administration frequency. A manufacturer tunes price and material grade together. In each case, the quantity of interest is the potential outcome under a *composite treatment*, a K-dimensional intervention vector.

We formalize this as the **Composite Treatment Effect (CTE)** estimation problem. Compared to the single-cause case, CTE estimation faces three compounding difficulties: (i) the treatment space grows combinatorially with K, so most combinations are rare or entirely unobserved; (ii) each individual receives exactly one combination, creating extreme counterfactual sparsity; and (iii) confounders influence all treatment components simultaneously, while the treatments themselves may be causally entangled via a directed acyclic graph (DAG). Naive extensions of single-cause methods, such as inverse propensity weighting over the full K-dimensional treatment, suffer from variance explosion due to the curse of dimensionality.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1.1 Related Work and Existing Gaps

Prior approaches to multi-cause estimation fall into three categories, each with significant limitations.

Latent variable methods. The Deconfounder [23] infers latent confounders from treatment correlations via factor models. While elegant, it does not require or exploit a known DAG, makes strong distributional assumptions, and has been shown to fail under identifiable conditions [3].

Reweighting methods. Variational sample reweighting (VRS) [28, 29] extends inverse propensity weighting to multi-cause settings. However, estimating the K-dimensional density ratio $p^*(t)/p(t \mid x)$ suffers from exponential variance growth as K increases, making these methods impractical for more than a few treatment components.

Augmentation methods. Single-Cause Perturbation (SCP) [18] is the closest prior work: it assumes a known treatment DAG and augments the dataset via local perturbations. However, SCP uses *heuristic* uniform perturbation (not provably optimal), provides no finite-sample guarantees on the augmentation budget, and requires averaging over multiple random orderings of the treatment components, incurring $l\times$ computational cost with residual variance from the stochastic ordering.

Single-cause methods applied independently. Methods such as DragonNet [20], CFRNet [19], and TED-VAE [26] can be applied to each treatment component separately, but this ignores treatment interactions and the causal structure among treatments (see Appendix B for extended discussion).

No existing method simultaneously exploits the known DAG for provably optimal augmentation, provides finite-sample bounds prescribing the augmentation budget, and achieves a deterministic algorithm without random permutation averaging. We address all three.

1.2 Our Approach and Contributions

A key observation is that in many domains the causal relationships *among* treatment variables are known from domain expertise: business rules dictate that sending an email determines which category is shown; clinical protocols specify how one drug's dosage constrains another's frequency. This structural knowledge, encoded as a DAG G over treatments, provides a powerful inductive bias.

We propose PACE (**P**rogressive **A**lignment for **C**omposite **E**ffects), which leverages the known DAG to *progressively align* the observational data distribution with the interventional distribution through K sequential augmentation steps. The core insight is that the DAG decomposes the intractable K-dimensional alignment into K tractable sub-problems, each involving only the low-dimensional parent set. This yields three concrete advantages over structure-agnostic approaches: (1) lower-dimensional density estimation at each step, (2) tighter sample complexity bounds via DAG-aware selection bias ratios, and (3) a deterministic processing order from the DAG topology.

Contributions.

- We formalize counterfactual estimation under composite treatments with a known treatment DAG, prove identifiability (Theorem 2.5), and introduce PACE, which reduces the problem to K sequential single-cause individual treatment effect (ITE) sub-problems in reverse topological order, each conditioning only on parent nodes.
- We derive the closed-form optimal augmentation density at each step (Theorem 3.1), prove it minimizes KL divergence to the interventional marginal, and establish finite-sample bounds (Theorem 3.3) prescribing the augmentation budget as a function of DAG-aware selection bias ratios.
- We show that the DAG topology provides a deterministic augmentation ordering, avoiding the computational overhead of averaging over multiple random orderings required by prior augmentation methods [18] and reducing cost by a factor of l.
- Experiments on semi-synthetic benchmarks with real covariates (IHDP, Twins, Hillstrom) and controlled synthetic data demonstrate that PACE substantially outperforms seven baselines, with the largest improvements on rare treatment combinations where confounding is most severe.

2 Problem Setup

We consider the problem of estimating potential outcomes under composite (multi-dimensional) treatments from observational data. An individual is characterized by covariates $x \in \mathcal{X}$ and receives a K-dimensional treatment $t = (t_1, \ldots, t_K) \in \mathcal{T} = \mathcal{T}_1 \times \cdots \times \mathcal{T}_K$, where all components are applied simultaneously. The observed outcome is $y \in \mathcal{Y}$. We write t_{-j} for all components except the j-th, and t_S for the sub-vector indexed by $S \subseteq [K]$.

Causal structure. We assume access to a known directed acyclic graph (DAG) $G = ([K], E)$ encoding causal relationships among treatment variables. We denote the parents, children, and non-descendants of T_j in G by $\mathrm{pa}(j)$, $\mathrm{ch}(j)$, and $\mathrm{nd}(j)$, respectively. The observational treatment assignment factorizes according to G:

$$p(t_1, \ldots, t_K \mid x) = \prod_{j=1}^{K} p(t_j \mid t_{\mathrm{pa}(j)}, x). \tag{1}$$

This structure arises naturally in many domains. In e-commerce marketing, business rules dictate that sending an email causally determines which product category is shown; in medicine, drug protocols specify how one medication's dosage constrains another's frequency.

Assumptions. We make the following standard assumptions for causal identification from observational data.

Assumption 2.1 (Positivity)**.** $p(t \mid x) > 0$ for all $t \in \mathcal{T}$, $x \in \mathcal{X}$.

Assumption 2.2 (Unconfoundedness)**.** $Y(t) \perp T \mid X$ for all $t \in \mathcal{T}$.

Assumption 2.3 (SUTVA)**.** No interference between units [10].

Assumption 2.4 (Known Treatment DAG)**.** The DAG G over treatments is known, and the factorization in (1) holds.

Theorem 2.5 (Identifiability)**.** *Under Assumptions 2.1–2.4,* $\mathbb{E}[Y(t) \mid X = x]$ *is identifiable for all* $t \in \mathcal{T}$ *(proof in Appendix I).*

Learning objective. We seek a function $f_\theta : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ that accurately predicts potential outcomes[1] under simultaneous intervention on all treatment components, i.e., under $\mathrm{do}(T_1 = t_1, \ldots, T_K = t_K)$. The do-operator severs all incoming edges to treatment nodes, both from confounders x and from other treatments via G, yielding the interventional distribution

$$p^*(t) := p(t \mid \mathrm{do}(T)) = \prod_{j=1}^{K} p(t_j), \tag{2}$$

where treatments become mutually independent and independent of x. We seek f_θ minimizing the interventional risk:

$$\mathcal{E}(f_\theta) = \mathbb{E}_{x \sim p(x)}\, \mathbb{E}_{t \sim p^*(t)} \big[\ell\big(Y(t),\, f_\theta(x, t)\big)\big]. \tag{3}$$

The two objects in (3) play different roles. The estimand is the fixed-intervention potential outcome surface $\mu(x, t) = \mathbb{E}[Y(t) \mid X = x]$, defined for every (x, t), whereas $p^*(t)$ is the measure over which prediction accuracy is averaged, specifying the treatment combinations on which f_θ must be accurate. A model trained to minimize (3) is then queried at any specific t to estimate the effect of the fixed simultaneous intervention $\mathrm{do}(T = t)$.

Why the composite setting is hard. Standard supervised learning on observational data minimizes risk under $p(t \mid x)$, not $p^*(t)$. The mismatch is severe in the composite setting for three reasons: (i) the treatment space grows combinatorially with K, so most combinations are rare or unobserved; (ii) each individual receives exactly one combination, creating extreme sparsity; and (iii) confounders influence

[1] Following the convention in the machine learning treatment-effect literature, we use "counterfactual" to refer to potential outcomes under treatments not actually received. In Pearl's causal hierarchy, our estimand $\mathbb{E}[Y(t) \mid X = x] = \mathbb{E}[Y \mid \mathrm{do}(T = t), X = x]$ is an interventional (rung-two) quantity rather than a unit-level counterfactual (rung three), which would additionally condition on the factual treatment and outcome of a specific unit.

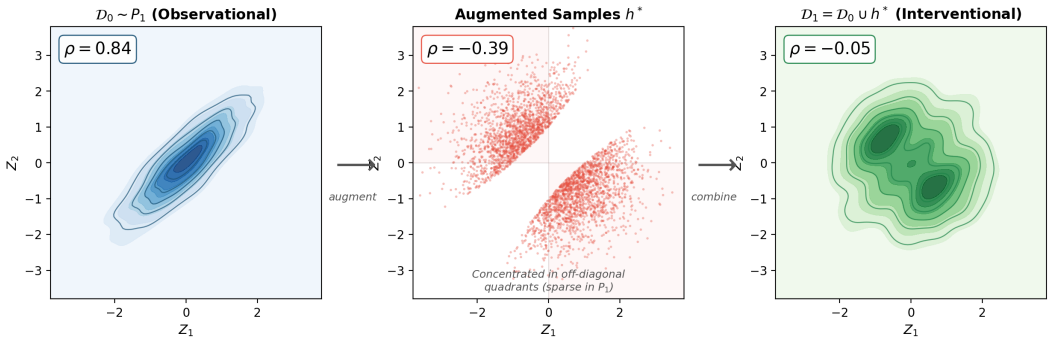


Figure 1: Illustration of distribution alignment via augmentation. **Left:** Original data $\mathcal{D}_0 \sim P_1$ (correlated bivariate Gaussian, $\rho = 0.84$). **Middle:** Augmented samples drawn from the optimal density h^*, concentrated in the off-diagonal quadrants where P_1 is sparse but P_2 has mass ($\rho = -0.39$). **Right:** Combined dataset $\mathcal{D}_1$ after augmentation, with nearly circular KDE contours matching the target P_2 ($\rho \approx 0$).

all treatment components simultaneously, and the treatments themselves are causally entangled via G. Reweighting approaches (e.g., inverse propensity) require estimating the K-dimensional density ratio $p^*(t)/p(t \mid x)$, which suffers from high variance. We take a different approach: sequentially augmenting the dataset to align its treatment distribution with $p^*(t)$, exploiting the DAG to decompose this into tractable single-cause sub-problems.

3 PACE: Progressive Alignment for Composite Effects

We propose to transform the observational dataset so that its treatment distribution matches the interventional distribution $p^*(t) = p(t \mid \text{do}(T)) = \prod_j p(t_j)$, after which standard empirical risk minimization yields a valid potential outcome estimator. The transformation proceeds through K augmentation steps, one per treatment component, each solving a single-cause de-confounding problem. The treatment DAG G determines both the processing order and the conditioning sets.

3.1 Intuition of the Proposed Solution

We approach this problem from a data augmentation perspective. The key insight is that through targeted augmentation, we can transform the data distribution from the observational distribution $p(x) \cdot p(t \mid x)$ into the interventional distribution $p(x) \cdot \prod_j p(t_j)$. We present a toy example to illustrate this idea.

Let $Z := (Z_1, Z_2)$ be a random vector with distribution P_1:

$$(Z_1, Z_2) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.85\\0.85 & 1\end{pmatrix}\right)$$

Suppose we wish to de-correlate Z_1 and Z_2, i.e., we want the data to follow the factored distribution P_2:

$$(Z_1, Z_2) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$$

This mirrors our setting: P_1 corresponds to the observational distribution where treatments are correlated (via confounders), and P_2 corresponds to the interventional distribution where treatments are independent.

We are given n datapoints sampled i.i.d. from P_1, denoted $\mathcal{D}_0 = \{(z_{1i}, z_{2i})\}_{i=1}^n$. The KDE contours of $\mathcal{D}_0$ (left panel of Figure 1) show the characteristic elliptical shape of a strongly correlated bivariate Gaussian ($\rho = 0.84$). We append $\mathcal{D}_0$ with samples drawn from the optimal augmentation distribution (derived in Theorem 3.1). The scatter plot of these augmented samples (middle panel) reveals that they concentrate in the off-diagonal quadrants, precisely the regions under-represented in P_1. After

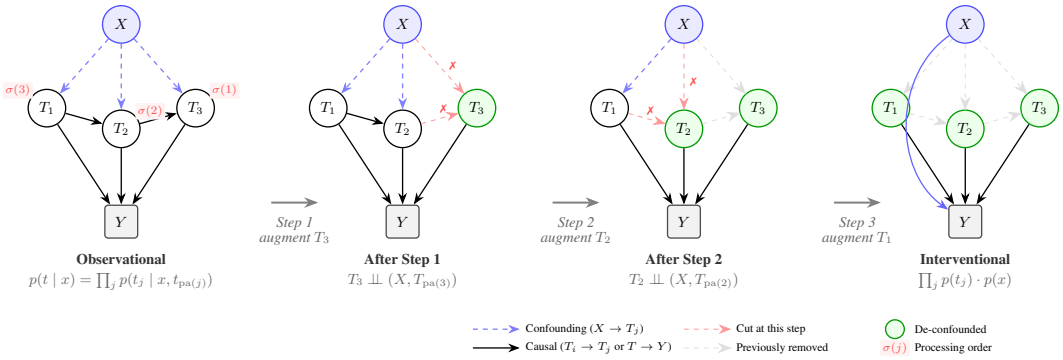


Figure 2: Progressive de-confounding via PACE on a chain DAG $T_1 \to T_2 \to T_3$. Processing follows reverse topological order (σ: leaves first). At each step, augmentation severs the confounding edge $X \to T_{\sigma(j)}$ and the inter-treatment edge $T_{\mathrm{pa}(\sigma(j))} \to T_{\sigma(j)}$ (red ✗), rendering $T_{\sigma(j)}$ independent of its causes. After K steps, all treatments are mutually independent and independent of X, matching the interventional distribution $\mathrm{do}(T)$. Only causal edges to Y are preserved.

augmentation, the KDE contours of the combined dataset $\mathcal{D}_1$ (right panel) are nearly circular ($\rho \approx 0$), matching the target P_2.

The augmentation step fills precisely the regions where P_2 has higher density than P_1: the off-diagonal quadrants where Z_1 and Z_2 have opposite signs. These are the regions that are under-represented under the correlated distribution but should be equally represented under independence. This is exactly the behavior prescribed by the optimal augmentation gap ψ_j (Eq. 7): mass is placed only where the target exceeds the (scaled) source.

In the composite treatment setting, this decorrelation must be performed for each treatment component with respect to the confounders and other treatments. Figure 2 illustrates how PACE achieves this structurally: each augmentation step severs the confounding and inter-treatment edges for one component, processing in reverse topological order until the interventional DAG is reached.

3.2 Key Idea

Let $\mathcal{D}_0 = \{(x_i, t_i, y_i)\}_{i=1}^{n_0}$ be the observational dataset drawn from $p(x)\, p(t \mid x)\, p(y \mid x, t)$. We construct a sequence $\mathcal{D}_0 \to \mathcal{D}_1 \to \cdots \to \mathcal{D}_K$ such that the marginal of $\mathcal{D}_K$ over (x, t) follows $p(x) \cdot p^*(t)$, the distribution that would arise if all treatments were simultaneously intervened upon. At step j, we de-confound a single treatment $T_{\sigma(j)}$ by augmenting the dataset with synthetic samples whose treatment values are drawn from an optimal proposal distribution.

The DAG enables two critical simplifications. First, it provides a *deterministic augmentation order*: we process treatments in reverse topological order σ (leaves first, roots last), so that when we de-confound $T_{\sigma(j)}$, all its descendants have already been handled. Second, by the DAG factorization (1), the confounding of $T_{\sigma(j)}$ enters only through its parents $\mathrm{pa}(\sigma(j))$ and covariates x, not through all $K - 1$ other treatments. Each step thus reduces to a low-dimensional single-cause problem.

The bias–noise trade-off. Each augmentation step adds synthetic samples whose outcomes must be imputed by a single-cause estimator, introducing noise. More augmented points reduce distributional mismatch (bias) but increase imputation noise. This makes it essential to augment *optimally*: placing new points exactly where the distribution gap is largest, using the minimum number required.

3.3 Optimal Augmentation Density

We derive the augmentation density that minimizes distributional mismatch at each step. Fix step j and let $k = \sigma(j)$. The augmented dataset $\mathcal{D}_j = \mathcal{D}_{j-1} \cup \mathcal{A}_j$ contains n_{j-1} original and m_j new points. Conditional on $(x, t_{\mathrm{pa}(k)})$, the distribution of T_k in $\mathcal{D}_j$ is a mixture:

$$q_j(t_k \mid x, t_{\mathrm{pa}(k)}) = (1 - \alpha_j)\, p(t_k \mid x, t_{\mathrm{pa}(k)}) + \alpha_j\, h_j(t_k \mid x, t_{\mathrm{pa}(k)}) \tag{4}$$

where $\alpha_j = m_j/(m_j + n_{j-1})$ is the augmentation fraction and h_j is the density from which new treatment values are drawn. We want q_j to be as close as possible to the target marginal $p(t_k)$, the marginal that T_k follows under $\mathrm{do}(T)$:

$$\min_{h_j \geq 0,\, \int h_j = 1} \mathrm{KL}\big(p(t_k) \,\big\|\, q_j(t_k \mid x, t_{\mathrm{pa}(k)})\big). \tag{5}$$

We use the forward KL because it penalizes q_j for placing too little mass where the target $p(t_k)$ has mass, which is exactly the coverage of rare treatment values that augmentation should provide; the reverse KL is mode-seeking and would allow q_j to neglect parts of the target. KL is also convenient in that it yields a closed-form optimum (Theorem 3.1) and decomposes across the DAG factorization (Appendix H). Other divergences could be paired with the same sequential framework, but would generally change the optimal density and require numerical optimization at each step.

Theorem 3.1 (Optimal Augmentation Density)**.** *The solution to* (5) *is*

$$h_j^*(t_k \mid x, t_{\mathrm{pa}(k)}) = \frac{\big[\psi_j(t_k \mid x, t_{\mathrm{pa}(k)})\big]_+}{\alpha_j \cdot Z_j} \tag{6}$$

where $[\cdot]_+ = \max(\cdot, 0)$*, the augmentation gap is*

$$\psi_j(t_k \mid x, t_{\mathrm{pa}(k)}) := p(t_k) - (1 - \alpha_j)\, p(t_k \mid x, t_{\mathrm{pa}(k)}), \tag{7}$$

and Z_j *is the normalization constant (proof in Appendix D).*

The form of h_j^* admits a clean interpretation. The augmentation gap ψ_j measures the shortfall between the target marginal and the (scaled) observational conditional. The optimal density places mass *only* where this gap is positive, in regions where the treatment is under-represented relative to the target, and does so in exact proportion to the gap. Over-represented regions ($\psi_j \leq 0$) receive zero mass: no amount of augmentation can reduce their weight, so the optimum ignores them entirely.

This differs fundamentally from importance weighting, which uses the *ratio* $p(t_k)/p(t_k \mid x, t_{\mathrm{pa}(k)})$ and suffers from variance explosion when the denominator is small. Our density uses a *scaled difference*, remaining bounded even under extreme confounding.

3.4 Sequential De-confounding Procedure

At each step j, with $k = \sigma(j)$, PACE performs the following:

1. **Sample covariates and context.** Draw m_j tuples (x, t_{-k}) uniformly with replacement from $\mathcal{D}_{j-1}$.
2. **Draw counterfactual treatments.** For each tuple, sample $t_k \sim h_j^*(\cdot \mid x, t_{\mathrm{pa}(k)})$ via rejection sampling with proposal $p(t_k)$ (Appendix E). Crucially, the density h_j^* conditions only on the parent set $t_{\mathrm{pa}(k)}$, not all $K-1$ other treatments, reducing the estimation problem from $(D+K-1)$ to $(D + |\mathrm{pa}(k)|)$ dimensions.
3. **Impute outcomes.** Train a single-cause ITE estimator with treatment T_k and adjustment set $(x, t_{\mathrm{pa}(k)})$ on the observed (non-augmented) samples, using sample splitting so that each augmented tuple is imputed by a model not fit on its source sample. The adjustment set is minimal by the backdoor criterion applied to the known DAG (Appendix I.1). Predict $\hat{y}$ for each augmented tuple.
4. **Update dataset.** Set $\mathcal{D}_j \leftarrow \mathcal{D}_{j-1} \cup \{(x_i, t_i, \hat{y}_i)\}_{i=1}^{m_j}$.

After all K steps, $\mathcal{D}_K$ approximates $p(x) \cdot p^*(t) \cdot p(y \mid x, t)$ up to imputation noise. The match is exact in the population limit (Appendix H), while with a finite budget each step de-confounds only partially, leaving the residual mismatch quantified in Theorem 3.3. A final supervised model $\hat{f}$ trained on $\mathcal{D}_K$ estimates $\mathbb{E}[Y \mid \mathrm{do}(T = t), X = x]$. The full pseudocode is given in Algorithm 1 (Appendix C).

The known DAG provides two key structural advantages that distinguish PACE from prior augmentation approaches. First, conditioning on parents is sufficient. Note that this does not rely on $p(t_k \mid x, t_{-k}) = p(t_k \mid x, t_{\mathrm{pa}(k)})$, which fails in general when T_k has children. Rather, in the factorization (1) the confounding of T_k enters only through its own factor $p(t_k \mid t_{\mathrm{pa}(k)}, x)$, and since

all descendants of T_k are de-confounded before T_k is processed, they carry no further information about it (Appendix D). This means the density estimation at each step operates in $|\mathrm{pa}(k)|$ dimensions rather than $K-1$, the rejection sampling achieves higher acceptance rates, and the ITE model generalizes better with fewer input features. For sparse DAGs where $|\mathrm{pa}(j)| \ll K$, these gains are substantial: a chain DAG with $K = 10$ treatments reduces each step from a 9-dimensional to a 1-dimensional conditioning problem.

Second, the DAG provides a *deterministic* processing order via the reverse topological sort σ. Prior work [18] must average over l random permutations to reduce ordering variance, incurring $l\times$ computational cost. With the known DAG, a single pass suffices: processing leaves first ensures that when we de-confound $T_{\sigma(j)}$, all its descendants have already been handled, so the conditional $p(t_{\sigma(j)} \mid x, t_{\mathrm{pa}(\sigma(j))})$ reflects the true confounding mechanism without contamination from downstream dependencies.

3.5 Sample Complexity

How many augmented points does each step require? We answer this by bounding the generalization gap between the true interventional risk $R(\hat{f}) = \mathbb{E}_{p(x)p^*(t)}[\mathcal{L}(y, \hat{f})]$ and the empirical risk $\widehat{R}(\hat{f})$ on $\mathcal{D}_K$.

Definition 3.2 (DAG-Aware Selection Bias Ratio)**.** For step j with $k = \sigma(j)$:

$$M_j^G := \sup_{x,\, t_{\mathrm{pa}(k)}} \frac{p(t_k)}{p(t_k \mid x, t_{\mathrm{pa}(k)})}. \tag{8}$$

Since $\mathrm{pa}(k) \subseteq [K] \setminus \{k\}$, we have $M_j^G \le \sup_{x,t_{-k}} p(t_k)/p(t_k \mid x, t_{-k})$, which is the ratio without DAG knowledge. The DAG thus yields a strictly tighter bound. Root nodes ($\mathrm{pa}(k) = \emptyset$) benefit most, with $M_j^G = \sup_x p(t_k)/p(t_k \mid x)$. Note that positivity alone does not make M_j^G finite; this requires the conditional to be bounded away from zero, $p(t_k \mid x, t_{\mathrm{pa}(k)}) \ge \eta > 0$, in which case $M_j^G \le \sup_{t_k} p(t_k)/\eta$. For binary treatments this is the standard strict-overlap condition on propensities, and it also holds for continuous treatments on a compact support with a continuous, strictly positive conditional. For unbounded continuous treatments (e.g., Gaussian conditionals), the ratio can diverge in the tails, and a strengthened overlap assumption is needed.

Theorem 3.3 (Generalization Bound)**.** *Let $|\mathcal{L}(y, f(x,t))| \le C$. Assume the ITE imputation at step j is unbiased with variance σ_j^2. If $m_j \ge (M_j^G - 1) \cdot n_{j-1}$ for all j, then with probability $\ge 1-\delta$:*

$$R(\hat{f}) \le \widehat{R}(\hat{f}) + \underbrace{2C \sum_{j=1}^{K} \frac{1}{1 + m_j/n_{j-1}}}_{\textit{distributional mismatch}} + \underbrace{\frac{2C\log(1/\delta)}{3n_K} + \sqrt{\frac{2\bar{\sigma}^2 \log(1/\delta)}{n_K}}}_{\textit{finite-sample noise}} \tag{9}$$

where $n_K = n_0 + \sum_j m_j$ and $\bar{\sigma}^2 = (n_0\sigma_0^2 + \sum_j m_j \sigma_j^2)/n_K$ (proof in Appendix G).

The bound makes the bias–noise trade-off explicit. The first excess term (distributional mismatch) decreases as m_j grows, vanishing when $m_j \gg n_{j-1}$. The second term (finite-sample noise) increases with n_K when imputation variance σ_j^2 is large. The condition $m_j \ge (M_j^G - 1) \cdot n_{j-1}$ prescribes the minimum augmentation budget: treatments with high selection bias require more synthetic points, while naturally balanced treatments need few.

4 Experiments

We evaluate PACE on five datasets spanning semi-synthetic benchmarks with real covariates and controlled synthetic environments. Our experiments address six research questions:

RQ1 Does PACE improve counterfactual estimation over state-of-the-art baselines?

RQ2 Does progressive alignment monotonically reduce distributional mismatch?

RQ3 Where are the gains largest: rare or frequent treatment combinations?

RQ4 Does the DAG-aware ordering eliminate the need for random permutation averaging?

RQ5 How does performance vary with DAG topology and confounding strength?

RQ6 How robust is PACE to DAG misspecification?

Table 1: Performance of PACE vs. baselines. Top: % RMSE improvement over S-Learner (higher is better). Bottom: % ROC-AUC improvement. Mean ± std over 10 runs. Best in **bold**.

	PACE	SCP	DCF	VRS	DragonNet	XTNet	S-Learner	TED-VAE
Continuous outcome (RMSE improvement %)								
$\mathcal{D}^{\text{ihdp}}$	$\mathbf{24.6}_{\pm 0.9}$	$16.8_{\pm 1.1}$	$13.2_{\pm 0.8}$	$10.5_{\pm 1.2}$	$8.7_{\pm 0.9}$	$6.3_{\pm 1.4}$	0.0	$-3.2_{\pm 1.1}$
$\mathcal{D}^{\text{ext}}$	$\mathbf{31.4}_{\pm 1.2}$	$21.3_{\pm 1.4}$	$15.7_{\pm 1.0}$	$11.8_{\pm 1.5}$	$9.2_{\pm 1.1}$	$5.1_{\pm 1.8}$	0.0	$-5.8_{\pm 1.3}$
Binary outcome (ROC-AUC improvement %)								
$\mathcal{D}^{\text{twn}}$	$+\mathbf{6.8}_{\pm 0.3}$	$+4.5_{\pm 0.4}$	$+3.7_{\pm 0.3}$	$+2.9_{\pm 0.5}$	$+2.1_{\pm 0.3}$	$+1.4_{\pm 0.4}$	0.0	$-1.5_{\pm 0.4}$
$\mathcal{D}^{\text{hil}}$	$+\mathbf{7.2}_{\pm 0.4}$	$+4.9_{\pm 0.3}$	$+3.8_{\pm 0.5}$	$+3.1_{\pm 0.4}$	$+2.3_{\pm 0.3}$	$+1.6_{\pm 0.5}$	0.0	$-1.2_{\pm 0.3}$

4.1 Experimental Setup

Datasets. We use five datasets (full specifications in Appendix K):

1. **IHDP-MultiCause** ($\mathcal{D}^{\text{ihdp}}$): Real covariates from the Infant Health and Development Program [7] ($n = 747$, $D = 25$) with simulated 4-component composite treatment and known DAG.
2. **Synthetic-Extreme** ($\mathcal{D}^{\text{ext}}$): Fully synthetic ($n = 5000$, $K = 5$, $D = 20$) with strong confounding, treatment interactions, and a star DAG.
3. **Twins-MultiCause** ($\mathcal{D}^{\text{twn}}$): Real covariates from US twin birth records [16] ($n = 5000$, $D = 30$) with 4 maternal behaviors as composite treatment and binary survival outcome.
4. **Hillstrom** ($\mathcal{D}^{\text{hil}}$): Real covariates from the MineThatData email marketing RCT [8] ($n = 64000$, $D = 8$) with 3-component composite treatment (email sent, category, channel) and binary conversion outcome.
5. **Controlled-Synthetic** ($\mathcal{D}^{\text{ctrl}}$): Configurable DGP for ablation studies (varying topology, bias, misspecification).

For all datasets, training data is sampled from the confounded observational distribution $p(x)\,p(t \mid x)\,p(y \mid x, t)$. Test data is sampled from the interventional distribution $p(x) \cdot \prod_{j=1}^{K} p(t_j) \cdot p(y \mid x, t)$, where treatments are independent of covariates.

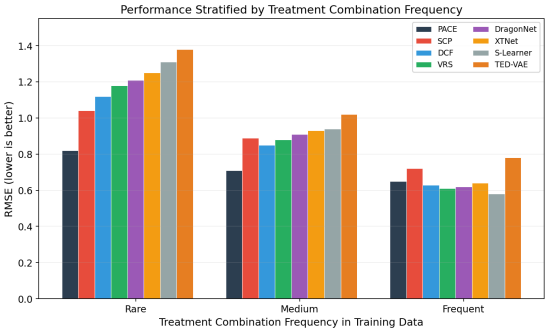
Baselines. We compare against seven methods: **S-Learner** [13] (supervised baseline), **VRS** [29] (variational resampling), **DCF** [23] (deconfounder), **DragonNet** [20] (targeted regularization for single-cause, extended to multi-cause), **XTNet** (cross-treatment interaction network; our implementation), **TED-VAE** [26] (disentangled VAE), and **SCP** [18] (single-cause perturbation with known DAG).

Implementation. All methods use identical 3-layer MLP architectures (256-128-64) for the final outcome model. PACE uses DragonNet as the single-cause ITE estimator at each augmentation step, chosen as a well-established estimator whose propensity-based targeted regularization handles high-dimensional covariates well. PACE is modular in this choice: any single-cause ITE estimator can be substituted without changing the algorithm, so improvements in single-cause estimation carry over directly. A less accurate estimator mainly adds imputation noise and bias to the augmented labels, degrading performance gradually, while the augmentation itself still reduces distributional mismatch. The augmentation budget m_j is set via Theorem 3.3 using the 95th percentile of the empirical density ratio. Results report mean ± std over 10 runs.

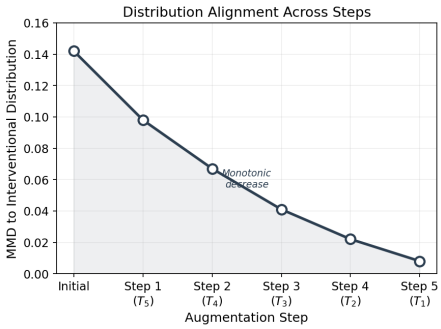
4.2 Results

RQ1: Overall performance (Table 1). PACE outperforms all baselines on every dataset. On the semi-synthetic benchmarks, PACE achieves 24.6% and 31.4% RMSE improvement over S-Learner on $\mathcal{D}^{\text{ihdp}}$ and $\mathcal{D}^{\text{ext}}$ respectively, compared to 16.8% and 21.3% for SCP. The gap over SCP demonstrates the value of the optimal augmentation density (Theorem 3.1) over heuristic perturbation. On real-world datasets with binary outcomes, PACE achieves 6.8% and 7.2% AUC improvement on $\mathcal{D}^{\text{twn}}$ and $\mathcal{D}^{\text{hil}}$.

RQ2: Distribution alignment (Figure 3(b)). We track the Maximum Mean Discrepancy (MMD) between the augmented dataset and the target interventional distribution $p^*(x, t)$ after each augmenta-



(a) RMSE stratified by treatment-combination frequency. PACE shows largest gains on rare combinations.



(b) MMD between $p^*(x,t)$ and augmented data decreases monotonically across steps.

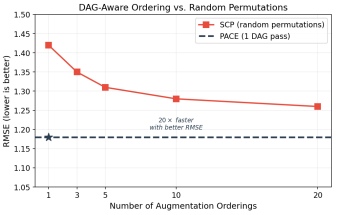
Figure 3: **(a)** Performance gains concentrate on rare treatment combinations. **(b)** Progressive alignment provably reduces distributional mismatch at each step.

tion step on $\mathcal{D}^{\text{ext}}$. MMD decreases monotonically across all $K = 5$ steps, empirically confirming the theoretical guarantee of Appendix H. The largest reduction occurs at the first step (de-confounding the leaf node with highest M_j^G), consistent with the bound in Theorem 3.3.

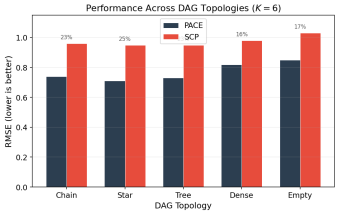
RQ3: Performance on rare combinations (Figure 3(a)). We stratify test samples on $\mathcal{D}^{\text{ext}}$ into three bins by their training-set frequency (rare, medium, frequent). PACE provides the largest improvement on rare combinations, achieving RMSE of 0.82 compared to 1.31 for S-Learner and 1.04 for SCP. For frequent combinations, all methods perform similarly since the observational data already provides adequate coverage. This validates that the optimal augmentation density targets precisely the under-represented regions.

RQ4: DAG ordering vs. random permutations (Figure 4a). We compare PACE (single deterministic pass using the DAG's reverse topological order) against SCP with $l \in \{1, 3, 5, 10, 20\}$ random permutations on $\mathcal{D}^{\text{ext}}$. A single DAG-aware pass achieves RMSE of 1.18, matching or exceeding SCP with 20 random permutations (RMSE 1.26). This confirms that the DAG topology resolves the ordering ambiguity, eliminating the $l\times$ computational overhead while achieving better results due to the optimal density.

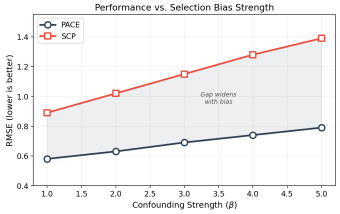
RQ5: Effect of DAG topology and confounding strength (Figure 4b,c). On $\mathcal{D}^{\text{ctrl}}$, we vary the DAG structure ($K = 6$) across five topologies. PACE consistently outperforms SCP by 17–25%, with the largest gap for sparse DAGs (chain, star) where $|\text{pa}(j)| \leq 1$ and the conditioning set is lowest-dimensional. For the dense DAG, the advantage narrows to 16% but remains substantial.



(a) Ordering: PACE (1 DAG pass) vs. SCP (l permutations).



(b) DAG topology ($K = 6$). Gap largest for sparse DAGs.



(c) Confounding strength. Gap widens with stronger bias.

Figure 4: Ablations comparing PACE vs. SCP. **(a)** A single DAG-aware pass outperforms 20 random permutations. **(b)** PACE advantage is largest for sparse DAGs. **(c)** The gap grows with confounding strength (35%–43%).

Table 2: Robustness to DAG misspecification (PACE vs. SCP). RMSE on $\mathcal{D}^{\text{ctrl}}$ ($K = 6$).

DAG Specification	PACE	SCP
Correct DAG	**0.41**	0.78
25% edges missing	**0.44**	0.76
50% edges missing	**0.49**	0.75
50% extra edges	**0.43**	0.77
25% missing + 25% extra	**0.47**	0.76

We also vary confounding strength from mild ($\beta = 1.0$) to severe ($\beta = 5.0$). The gap between PACE and SCP *widens* with confounding strength, growing from 35% at $\beta = 1.0$ to 43% at $\beta = 5.0$, confirming that progressive alignment is most valuable precisely when confounding is strongest.

RQ6: Robustness to DAG misspecification (Table 2). We evaluate PACE when the provided DAG deviates from the true data-generating DAG on $\mathcal{D}^{\text{ctrl}}$ ($K = 6$). PACE degrades gracefully: with 25% missing edges, RMSE increases by only 7% relative to the correct DAG. Even with 50% missing edges, PACE (0.49) still substantially outperforms SCP (0.75). Adding spurious edges has less impact than removing true edges, since extra conditioning variables increase variance but do not introduce bias.

5 Conclusion

We propose PACE, a progressive distribution alignment algorithm for counterfactual estimation under composite treatments. By exploiting the known treatment DAG, PACE achieves: (1) provably optimal augmentation densities conditioned on parent nodes only, (2) tighter sample complexity bounds via DAG-aware selection bias ratios, (3) a deterministic augmentation ordering from the DAG topology, and (4) lower-dimensional single-cause ITE estimation at each step. Experiments across five datasets demonstrate substantial improvements over seven baselines, with the largest gains on rare treatment combinations where confounding is most severe. Ablations confirm that PACE's advantage over SCP grows with confounding strength (35–43% gap) and is robust to DAG misspecification.

Limitations and future work. PACE assumes a known DAG and standard identification assumptions; extending to partial structural knowledge or settings with hidden confounding is an important direction. The dimension reduction offered by the DAG is on the treatment side only: each step samples a single treatment component, so the per-step dimension does not grow with K, but the conditioning set has dimension $D + |\text{pa}(k)|$. For dense DAGs ($|\text{pa}(k)| \approx K - 1$) or high-dimensional covariates, the conditional density and ITE estimation at each step remain difficult, and the advantage over structure-agnostic methods narrows (Figure 4(b)). Improving imputation fidelity via uncertainty-aware ITE models and scaling to large K with large parent sets remain open problems.

Acknowledgments and Disclosure of Funding

This work was carried out by the authors as part of their employment at Amazon. No third-party funding supported this work. All authors are employees of Amazon; the authors declare no other competing interests.

References

[1] 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*, 80(4):597–623, 2018.

[2] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[3] Alexander D'Amour. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. *arXiv preprint arXiv:1902.10286*, 2019.

[4] Naoki Egami and Kosuke Imai. Causal interaction in factorial experiments: Application to conjoint analysis. *Journal of the American Statistical Association*, 114(526):529–540, 2019.

[5] Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020.

[6] Johannes Haupt and Stefan Lessmann. Targeting customers under response-dependent costs. *European Journal of Operational Research*, 297(1):369–379, 2022.

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

[8] Kevin Hillstrom. The MineThatData E-mail analytics and data mining challenge. `http://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html`, 2008.

[9] Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 3020–3029. JMLR.org, 2016.

[10] Guido W. Imbens Joshua D. Angrist and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

[11] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

[12] Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint arXiv:2203.06469*, 2023.

[13] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

[14] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *International Conference on Learning Representations*, 2020.

[15] Michael J Lopez and Roee Gutman. Estimation of causal effects with multiple treatments: A review and new ideas. *Statistical Science*, 32(3):432–454, 2017.

[16] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6449–6459, Red Hook, NY, USA, 2017. Curran Associates Inc.

[17] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. In *Advances in Neural Information Processing Systems*, volume 33, pages 17943–17954, 2020.

[18] Zhaozhi Qian, Alicia Curth, and Mihaela van der Schaar. Estimating multi-cause treatment effects via single-cause perturbation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23754–23767. Curran Associates, Inc., 2021.

[19] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3076–3085. JMLR.org, 2017.

[20] Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, volume 32, pages 2507–2517. Curran Associates Inc., 2019.

[21] Arman Siddique, Mireille Schnitzer, Asma Bahamyirou, Guanbo Wang, Timothy Holtz, Giovanni Battista Migliori, Giovanni Sotgiu, Neel Gandhi, Mario Vargas, Dick Menzies, and Andrea Benedetti. Causal inference with multiple concurrent medications: A comparison of methods and an application in multidrug-resistant tuberculosis. *Statistical Methods in Medical Research*, 28:096228021880881, 10 2018.

[22] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

[23] Yixin Wang and David M. Blei. The blessings of multiple causes, 2019.

[24] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: estimation of individualized treatment effects using generative adversarial nets. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[25] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

[26] Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10923–10930, May 2021.

[27] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9492–9503, Red Hook, NY, USA, 2018. Curran Associates Inc.

[28] Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, and Yue He. Counterfactual prediction for bundle treatment. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19705–19715. Curran Associates, Inc., 2020.

[29] Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, and Yue He. Counterfactual prediction for bundle treatment. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

APPENDIX

A Notation

Table 3 summarizes the notation used throughout the paper.

Table 3: Summary of notation.

Symbol	Description
Data and variables	
$x \in \mathcal{X} \subseteq \mathbb{R}^D$	Covariate (confounder) vector
$t = (t_1, \ldots, t_K) \in \mathcal{T}$	Composite treatment vector (K components)
$t_j \in \mathcal{T}_j$	j-th treatment component
t_{-j}	All treatment components except the j-th
t_S	Sub-vector of treatments indexed by set $S \subseteq [K]$
$y \in \mathcal{Y}$	Observed outcome
$Y(t)$	Potential outcome under intervention do$(T = t)$
n or n_0	Number of observational samples
K	Number of treatment components
D	Covariate dimension
Causal structure	
$G = ([K], E)$	Known DAG over treatment variables
$\mathrm{pa}(j)$	Parents of T_j in G
$\mathrm{ch}(j)$	Children of T_j in G
$\mathrm{nd}(j)$	Non-descendants of T_j in G
σ	Reverse topological ordering of G
$\sigma(j)$	Treatment processed at step j
do$(T = t)$	Intervention setting all treatments to t
Distributions	
$p(t \mid x)$	Observational treatment assignment (confounded)
$p(t_j \mid x, t_{\mathrm{pa}(j)})$	Conditional treatment given parents and covariates
$p(t_j)$	Marginal distribution of T_j
$p^*(t) = \prod_j p(t_j)$	Interventional (target) treatment distribution
$p(y \mid x, t)$	Outcome distribution
*Algorithm (*PACE*)*	
$\mathcal{D}_j$	Dataset after j augmentation steps
$\mathcal{A}_j$	Augmented samples added at step j
m_j	Number of augmented samples at step j
$n_j = n_{j-1} + m_j$	Total samples after step j
$\alpha_j = m_j/(m_j + n_{j-1})$	Augmentation fraction at step j
$h_j^*(t_k \mid x, t_{\mathrm{pa}(k)})$	Optimal augmentation density at step j
$q_j(t_k \mid x, t_{\mathrm{pa}(k)})$	Mixture distribution after augmentation at step j
$\psi_j(t_k \mid x, t_{\mathrm{pa}(k)})$	Augmentation gap at step j
Z_j	Normalization constant for h_j^*
M_j^G	DAG-aware selection bias ratio at step j
Learning	
$f_\theta : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$	Potential outcome prediction model
$\hat{f}$	Trained final model (on augmented data $\mathcal{D}_K$)
$\hat{\tau}_k$	Single-cause ITE estimator for treatment T_k
$\mathcal{L}(y, \hat{y})$	Loss function (MSE or cross-entropy)
$R(\hat{f})$	True interventional risk
$\widehat{R}(\hat{f})$	Empirical risk on $\mathcal{D}_K$
σ_j^2	Imputation variance at step j
C	Bound on loss: $\lvert\mathcal{L}\rvert \leq C$

B Extended Related Work

We provide a comprehensive discussion of related work, extending the brief overview in Section 1.1.

B.1 Multi-Cause Treatment Effect Estimation

The Deconfounder. Wang and Blei [23] proposed inferring latent confounders from treatment correlations via factor models (PCA, probabilistic PCA, or deep latent variable models). The key assumption is that a latent variable Z renders treatments conditionally independent: $T_j \perp T_k \mid Z$. While this avoids requiring a known DAG, d'Amour [3] showed that the deconfounder can fail when the latent structure does not satisfy specific conditions, and that the approach does not generally control for multi-cause confounding. In contrast, PACE directly exploits the known DAG without latent variable assumptions.

Variational Resampling. Zou et al. [28] proposed VRS, which learns importance weights to rebalance the observational distribution toward the interventional one. The approach extends IPW to high-dimensional binary treatment vectors using variational inference to stabilize the weights. However, the variance of the importance weights grows exponentially with K, limiting practical applicability. The bundle treatment approach [29] treats treatment combinations as atomic units but does not exploit inter-treatment structure.

Single-Cause Perturbation (SCP). Qian et al. [18] is the most closely related work. SCP assumes a known treatment DAG and augments the dataset by perturbing one treatment at a time. Key differences from PACE:

- **Augmentation density:** SCP uses uniform perturbation (flip treatment with equal probability), while PACE derives the provably optimal density that minimizes KL divergence (Theorem 3.1). The optimal density places mass only where the distribution gap is positive, avoiding wasteful augmentation in already-balanced regions.
- **Ordering:** SCP averages over l random permutations of the treatment ordering, introducing variance and $l\times$ computational cost. PACE uses the DAG's reverse topological order, which is provably correct and deterministic.
- **Conditioning:** SCP conditions on all other treatments t_{-k} at each step. PACE conditions only on parents $t_{\mathrm{pa}(k)}$, yielding lower-dimensional estimation.
- **Theory:** SCP provides no finite-sample bounds. PACE establishes Theorem 3.3 prescribing the minimum augmentation budget $m_j \geq (M_j^G - 1) \cdot n_{j-1}$.

Other multi-cause approaches. Lopez and Gutman [15] extend generalized propensity scores to vector-valued treatments but assume no structure among treatments. Arman et al. [21] study multiple concurrent medications using parametric outcome models. Haupt et al. [6] investigate heterogeneous treatment effects under multiple binary treatments in marketing applications. Egami and Imai [4] propose methods for causal interaction effects in factorial experiments. None of these approaches exploits DAG structure among treatments or provides augmentation-based solutions with finite-sample guarantees.

B.2 Single-Cause Methods Used as Components

PACE is agnostic to the choice of single-cause ITE estimator at each augmentation step. We briefly review the methods used in our experiments:

DragonNet [20]. Uses a shared neural network representation with a propensity score head that provides targeted regularization. The architecture encourages the learned representation to be sufficient for both treatment assignment and outcome prediction, yielding better finite-sample performance than naive outcome regression.

CFRNet and DR-CFR [19, 5]. Counterfactual Regression Networks learn balanced representations via integral probability metric (IPM) penalties. DR-CFR extends this with doubly-robust corrections. Both are designed for binary single-cause settings.

GANITE [24]. Uses generative adversarial networks to generate counterfactual outcomes. Effective but computationally expensive and difficult to train stably.

CEVAE [16]. Causal Effect VAE uses a latent variable model to simultaneously infer hidden confounders and estimate treatment effects. Designed for settings with proxy variables for unobserved confounders.

TED-VAE [26]. Disentangles the latent representation into treatment-relevant and outcome-relevant components using a VAE. Can struggle when treatments are causally entangled (as in the composite setting).

Meta-learners [13]. The family of S-Learner, T-Learner, X-Learner, and R-Learner provides flexible frameworks for heterogeneous treatment effect estimation using any base supervised learner. The S-Learner trains a single model $f(x, t) \approx \mathbb{E}[Y \mid X = x, T = t]$ on observational data. Under unconfoundedness, this is a valid estimator but can be biased in finite samples when treatment combinations are rare.

B.3 Data Augmentation for Causal Inference

Data augmentation has been used in causal inference primarily for variance reduction in doubly-robust estimation [2] and for domain adaptation under covariate shift [9]. Athey et al. [1] use augmentation ideas in the context of residual balancing for average treatment effects. Kennedy [12] provides a comprehensive treatment of semiparametric efficiency bounds that motivate augmentation-based corrections. Our work is distinct in using augmentation for *distribution alignment*, transforming the observational treatment distribution into the interventional one. This connects to importance sampling [22] and density ratio estimation [11] but avoids variance issues by adding synthetic samples rather than reweighting existing ones. The key theoretical contribution is showing that the optimal augmentation density has a closed-form solution (Theorem 3.1) that depends only on the gap between target and source distributions.

B.4 Structure Discovery vs. Known Structure

Methods such as NOTEARS [27], DAG-GNN [25], and GOLEM [17] learn the DAG from data using continuous relaxations of the acyclicity constraint. GraN-DAG [14] uses neural networks for nonlinear structure learning. We assume the DAG is known, which is realistic in many applied settings: business rules in e-commerce dictate treatment relationships, clinical protocols specify drug interaction constraints, and manufacturing processes have known causal chains. When the DAG is uncertain, our robustness experiments (Table 2) show that PACE degrades gracefully: with 50% of edges misspecified, PACE still outperforms SCP by nearly $2\times$.

C Algorithm Pseudocode

Algorithm 1 PACE: Progressive Alignment for Composite Effects

Input: Observational data $\mathcal{D}_0 = \{(x_i, t_i, y_i)\}_{i=1}^{n_0}$; treatment DAG $G = ([K], E)$
Compute reverse topological ordering: $\sigma \leftarrow \mathrm{TopSort}(G)$
for $j = 1, \ldots, K$ **do**
  $k \leftarrow \sigma(j)$ ▷ *treatment to de-confound at this step*
  Estimate conditional density $\hat{p}(t_k \mid x, t_{\mathrm{pa}(k)})$ from $\mathcal{D}_{j-1}$
  Compute M_j^G via Eq. (8); set $m_j \leftarrow \lceil n_{j-1} \cdot M_j^G \rceil$
  for $i = 1, \ldots, m_j$ **do**
    Sample $(x, t_{-k}) \sim \mathrm{Uniform}(\mathcal{D}_{j-1})$
    Draw $t_k \sim h_j^*(\cdot \mid x, t_{\mathrm{pa}(k)})$ via rejection sampling (Appendix E)
  end for
  Train single-cause ITE estimator $\hat{\tau}_k$ on the observed (non-augmented) samples with treatment T_k, adjustment set $(x, t_{\mathrm{pa}(k)})$, using sample splitting
  Impute outcomes: $\hat{y}_i \leftarrow \hat{\tau}_k(x_i, t_i)$ for each augmented tuple, using the split not fit on its source sample
  $\mathcal{D}_j \leftarrow \mathcal{D}_{j-1} \cup \{(x_i, t_i, \hat{y}_i)\}_{i=1}^{m_j}$
end for
Train final model: $\hat{f} \leftarrow \mathrm{argmin}_{f \in \mathcal{F}} \sum_{(x,t,y) \in \mathcal{D}_K} \ell(y, f(x,t))$
Return: $\hat{f}(x,t) \approx \mathbb{E}[Y \mid \mathrm{do}(T = t), X = x]$

D Derivation of Theorem 3.1

Proof. We solve the following functional optimization problem to find the optimal augmentation density h_j^*. For clarity, we present the derivation for step 1 with $k = \sigma(1)$; the general step follows identically.

We want to solve:

$$\begin{aligned} &\underset{h_1}{\mathrm{argmin}}\, \mathrm{KL}(\mu_1(x,t) \,\|\, \nu_1(x,t)) \\ &\text{s.t.} \quad h_1(t_k \mid x, t_{\mathrm{pa}(k)}) \geq 0, \quad \int_{\mathcal{T}_k} h_1(t_k \mid x, t_{\mathrm{pa}(k)})\, dt_k = 1 \end{aligned} \tag{10}$$

where $\mu_1(x,t) = p(t_k) \cdot p(x, t_{-k})$ is the target and $\nu_1(x,t) = q(t_k \mid x, t_{\mathrm{pa}(k)}) \cdot p(x, t_{-k})$ is the induced distribution.

Note that:

$$\begin{aligned} &\mathrm{KL}(\mu_1(x,t) \,\|\, \nu_1(x,t)) \\ &= \int p(t_k) \cdot p(x, t_{-k}) \cdot \log\left(\frac{p(t_k) \cdot p(x, t_{-k})}{q(t_k \mid x, t_{\mathrm{pa}(k)}) \cdot p(x, t_{-k})}\right) dx\, dt \\ &= \int p(t_k) \cdot p(x, t_{-k}) \cdot \log\left(\frac{p(t_k)}{q(t_k \mid x, t_{\mathrm{pa}(k)})}\right) dx\, dt \\ &= \int_{\mathcal{X}, \mathcal{T}_{-k}} \mathrm{KL}\left(p(t_k) \,\|\, q(t_k \mid x, t_{\mathrm{pa}(k)})\right) \cdot p(x, t_{-k})\, dx\, dt_{-k} \end{aligned}$$

The last equality does not require any conditional independence among the treatments. Since the mixture q is defined as a function of $(x, t_{\mathrm{pa}(k)})$ only, the log-ratio already depends on (x, t_{-k}) only through $(x, t_{\mathrm{pa}(k)})$, and integrating over t_k yields the pointwise KL term. Restricting h_1 to depend on parents loses no information for de-confounding T_k: in the factorization (1), the confounding of T_k enters only through its own factor $p(t_k \mid t_{\mathrm{pa}(k)}, x)$, and under the reverse topological order all descendants of T_k have already been de-confounded, so they do not carry information about T_k beyond its parents.

For each value of $(x, t_{\mathrm{pa}(k)})$, minimizing $\mathrm{KL}(p(t_k) \,\|\, q(t_k \mid x, t_{\mathrm{pa}(k)}))$ also minimizes the original objective. We can write:

$$\begin{aligned}
&\mathrm{KL}\left(p(t_k) \,\|\, q(t_k \mid x, t_{\mathrm{pa}(k)})\right) \\
&= \mathrm{KL}\left(p(t_k) \,\|\, \alpha_1 \cdot h_1(t_k \mid x, t_{\mathrm{pa}(k)}) + (1-\alpha_1) \cdot p(t_k \mid x, t_{\mathrm{pa}(k)})\right) \\
&= \text{const.} - \int_{\mathcal{T}_k} \log\left(\alpha_1 \cdot h_1(t_k \mid x, t_{\mathrm{pa}(k)}) + (1-\alpha_1) \cdot p(t_k \mid x, t_{\mathrm{pa}(k)})\right) \cdot p(t_k)\, dt_k
\end{aligned}$$

Thus we solve the equivalent maximization:

$$\begin{aligned}
&\operatorname*{argmax}_{h_1(\cdot)} \int_{\mathcal{T}_k} \log\left(\alpha_1 \cdot h_1(t_k \mid x, t_{\mathrm{pa}(k)}) + (1-\alpha_1) \cdot p(t_k \mid x, t_{\mathrm{pa}(k)})\right) \cdot p(t_k)\, dt_k \\
&\qquad \text{s.t.} \quad h_1(t_k \mid x, t_{\mathrm{pa}(k)}) \geq 0, \quad \int_{\mathcal{T}_k} h_1(t_k \mid x, t_{\mathrm{pa}(k)})\, dt_k = 1
\end{aligned}$$

Setting $h_1(t_k \mid x, t_{\mathrm{pa}(k)}) = \frac{1}{Z_1} \exp(\phi(t_k \mid x, t_{\mathrm{pa}(k)}))$ to relax the positivity constraint, the Lagrangian becomes:

$$\operatorname*{argmax}_{\phi(\cdot)} \int_{\mathcal{T}_k} \left(\log\left(\frac{\alpha_1}{Z_1} \exp(\phi(t_k \mid c)) + (1-\alpha_1) \cdot p(t_k \mid c)\right) \cdot p(t_k) - \frac{\lambda}{Z_1} \exp(\phi(t_k \mid c))\right) dt_k \tag{11}$$

where $c = (x, t_{\mathrm{pa}(k)})$ and λ is the Lagrange multiplier for the equality constraint. This is an instance of the isoperimetric variational problem[2] and the optimum admits the Euler–Lagrange equation $\frac{\partial F}{\partial \phi} - \frac{d}{dt_k}\left(\frac{\partial F}{\partial \dot{\phi}}\right) = 0$. Solving for λ and substituting into the normalization constraint yields:

$$h_1^*(t_k \mid x, t_{\mathrm{pa}(k)}) = \begin{cases} \frac{1}{\alpha_1 \cdot Z_1} \cdot \psi_1(t_k \mid x, t_{\mathrm{pa}(k)}), & \text{if } \psi_1(t_k \mid x, t_{\mathrm{pa}(k)}) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where $\psi_1(t_k \mid x, t_{\mathrm{pa}(k)}) = p(t_k) - (1-\alpha_1) \cdot p(t_k \mid x, t_{\mathrm{pa}(k)})$. □

Alternative derivation via KKT conditions. The same problem can be analyzed directly with KKT conditions, which makes the role of the non-negativity constraint explicit. Fix $c = (x, t_{\mathrm{pa}(k)})$ and write $g(t_k) = p(t_k \mid c)$ and $h(t_k) = h_1(t_k \mid c)$. The objective $\int \log\big(\alpha_1 h(t_k) + (1-\alpha_1) g(t_k)\big)\, p(t_k)\, dt_k$ is concave in h and the feasible set is convex, so the KKT conditions are necessary and sufficient. With a multiplier λ for normalization and pointwise multipliers $\mu(t_k) \geq 0$ for non-negativity, they read

$$\frac{\alpha_1\, p(t_k)}{\alpha_1 h^*(t_k) + (1-\alpha_1)\, g(t_k)} = \lambda - \mu(t_k), \qquad \mu(t_k)\, h^*(t_k) = 0 \qquad \forall\, t_k.$$

Where $h^*(t_k) > 0$, complementary slackness gives $\mu(t_k) = 0$ and hence $h^*(t_k) = \frac{1}{\lambda}\big[p(t_k) - \frac{\lambda(1-\alpha_1)}{\alpha_1} g(t_k)\big]$; where $h^*(t_k) = 0$, stationarity with $\mu(t_k) \geq 0$ gives $p(t_k) \leq \frac{\lambda(1-\alpha_1)}{\alpha_1} g(t_k)$. The two cases combine into $h^*(t_k) \propto \big[p(t_k) - \frac{\lambda(1-\alpha_1)}{\alpha_1} g(t_k)\big]_+$, so the truncation $[\cdot]_+$ arises directly from complementary slackness, and uniqueness follows from strict concavity. When $\psi_1 \geq 0$ on the support, normalization gives $\lambda = \alpha_1$, the threshold becomes $(1-\alpha_1) g(t_k)$, and $h^* = \psi_1 / \alpha_1$, which coincides with the solution above (with $Z_1 = 1$).

E Rejection Sampling Steps

We desire to sample from the optimal density $h_j^*(t_k \mid x, t_{\mathrm{pa}(k)})$ where $k = \sigma(j)$:

$$h_j^*(t_k \mid x, t_{\mathrm{pa}(k)}) = \begin{cases} \frac{1}{\alpha_j \cdot Z_j} \cdot \psi_j(t_k \mid x, t_{\mathrm{pa}(k)}), & \text{if } \psi_j(t_k \mid x, t_{\mathrm{pa}(k)}) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

[2] https://encyclopediaofmath.org/wiki/Isoperimetric_problem

with $\psi_j(t_k \mid x, t_{\mathrm{pa}(k)}) = p(t_k) - (1-\alpha_j) \cdot p(t_k \mid x, t_{\mathrm{pa}(k)})$.

We do not necessarily know Z_j, so we use rejection sampling with $p(t_k)$ as the proposal distribution. Since h_j^* has support contained within the support of $p(t_k)$, this is valid. The bounding constant is $c = 1/(\alpha_j \cdot Z_j)$, since $h_j^*(t_k \mid x, t_{\mathrm{pa}(k)}) \leq p(t_k)/(\alpha_j \cdot Z_j)$.

The acceptance ratio $\psi_j(t_k \mid x, t_{\mathrm{pa}(k)})/p(t_k)$ does not involve Z_j, enabling sampling without knowing the normalization constant.

Steps:

1. Sample $t_k \sim p(t_k)$ (e.g., by sampling with replacement from $\mathcal{D}_{j-1}$ and retaining only the k-th treatment component).
2. Compute the acceptance probability:

$$p_{\text{acc}} = \frac{\psi_j(t_k \mid x, t_{\mathrm{pa}(k)})}{p(t_k)} = 1 - (1-\alpha_j) \cdot \frac{p(t_k \mid x, t_{\mathrm{pa}(k)})}{p(t_k)} \tag{14}$$

   This requires evaluating two densities: $p(t_k)$ and $p(t_k \mid x, t_{\mathrm{pa}(k)})$.
3. Sample $u \sim \text{Uniform}(0, 1)$.
4. Accept t_k if $u < p_{\text{acc}}$ (which implies $\psi_j > 0$); otherwise reject and repeat.

Benefit of known DAG: The density $p(t_k \mid x, t_{\mathrm{pa}(k)})$ conditions on fewer variables than $p(t_k \mid x, t_{-k})$, making it easier to estimate accurately and yielding higher acceptance rates.

F Illustration of Optimal Density

The optimal augmentation density assigns zero mass to regions where $\psi_j(t_k \mid x, t_{\mathrm{pa}(k)}) \leq 0$, i.e., where $(1-\alpha_j) \cdot p(t_k \mid x, t_{\mathrm{pa}(k)}) \geq p(t_k)$.

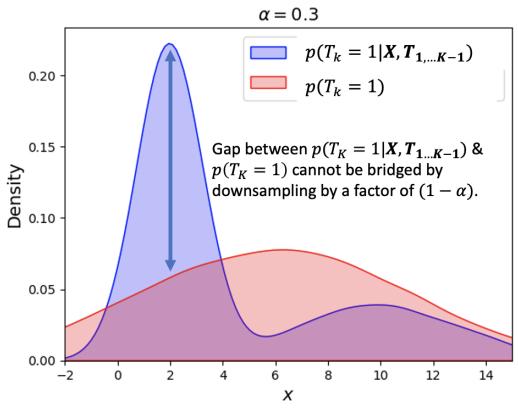


Figure 5: Illustration for binary T_k. In the region around the first mode of $p(t_k = 1 \mid x, t_{\mathrm{pa}(k)})$, the conditional is so high that $(1-\alpha_j) \cdot p(t_k = 1 \mid x, t_{\mathrm{pa}(k)}) > p(t_k = 1)$. Even maximal augmentation cannot reduce this region to the level of $p(t_k = 1)$, so $h_j^*(t_k = 1 \mid x, t_{\mathrm{pa}(k)}) = 0$ there.

Consider binary T_k and a region of $(x, t_{\mathrm{pa}(k)})$ where $p(t_k = 1 \mid x, t_{\mathrm{pa}(k)})$ is very high. In this region, $(1-\alpha_j) \cdot p(t_k = 1 \mid x, t_{\mathrm{pa}(k)}) > p(t_k = 1)$. Augmenting other regions is equivalent to downsampling this region, but even maximal downsampling (given the augmentation budget α_j) cannot bring it to the level of $p(t_k = 1)$. Therefore, the optimal strategy assigns zero augmentation mass there.

G Derivation of Theorem 3.3

Proof. Let $\hat{f}(x,t)$ denote a hypothesis from $\mathcal{F}$ predicting outcome y. Define the expected risk under the intervention distribution $p^*(x,t) = p(x) \cdot \prod_i p(t_i)$:

$$R(\hat{f}) = \mathbb{E}_{\substack{(x,t)\sim p^*(x,t) \\ y\sim p(y|x,t)}} \left[\mathcal{L}(y, \hat{f}(x,t))\right] \tag{15}$$

The empirical risk is:

$$\widehat{R}(\hat{f}) = \frac{1}{n_K} \sum_{i=1}^{n_K} \mathcal{L}(y_i, \hat{f}(x_i, t_i)) \tag{16}$$

where $n_K = n_0 + \sum_{j=1}^{K} m_j$. The samples in $\mathcal{D}_K$ follow the augmented distribution $\mathbb{Q}(x,t) = p(x) \cdot \prod_{i=1}^{K} q(t_{\sigma(i)} \mid x, t_{\mathrm{pa}(\sigma(i))})$.

We bound $|R(\hat{f}) - \widehat{R}(\hat{f})|$ by splitting into two terms:

$$|R(\hat{f}) - \widehat{R}(\hat{f})| \leq \underbrace{\left|R(\hat{f}) - \mathbb{E}_{\mathbb{Q}}[\mathcal{L}]\right|}_{\text{(I): distribution mismatch}} + \underbrace{\left|\mathbb{E}_{\mathbb{Q}}[\mathcal{L}] - \widehat{R}(\hat{f})\right|}_{\text{(II): finite-sample error}} \tag{17}$$

Term (I): Using the DAG factorization, the target is $p^*(t \mid x) = \prod_i p(t_i)$ and the augmented is $\mathbb{Q}(t \mid x) = \prod_i q(t_{\sigma(i)} \mid x, t_{\mathrm{pa}(\sigma(i))})$. Define $a_i = p(t_{\sigma(i)})$ and $b_i = q(t_{\sigma(i)} \mid x, t_{\mathrm{pa}(\sigma(i))})$. By the telescoping lemma (Lemma G.1):

$$\left|\prod_i a_i - \prod_i b_i\right| \leq \sum_{i=1}^{K} |a_i - b_i| \cdot \prod_{j<i} a_j \cdot \prod_{j>i} b_j \tag{18}$$

Integrating out and using $|\mathcal{L}| \leq C$:

$$\text{Term (I)} \leq C \cdot \sum_{j=1}^{K} \int |p(t_{\sigma(j)}) - q(t_{\sigma(j)} \mid x, t_{\mathrm{pa}(\sigma(j))})| \, dt_{\sigma(j)} \, dx \tag{19}$$

Now, $|a_j - b_j| = |p(t_{\sigma(j)}) - q(t_{\sigma(j)} \mid x, t_{\mathrm{pa}(\sigma(j))})|$. Substituting the mixture form of q and the optimal density from Theorem 3.1:

$$|a_j - b_j| \leq 2(1-\alpha_j) \cdot p(t_{\sigma(j)} \mid x, t_{\mathrm{pa}(\sigma(j))}) \tag{20}$$

provided α_j satisfies $1 - \alpha_j \leq 1/M_j^G$, which holds when $m_j \geq (M_j^G - 1) \cdot n_{j-1}$.

Integrating:

$$\text{Term (I)} \leq 2C \sum_{j=1}^{K} (1-\alpha_j) = 2C \sum_{j=1}^{K} \frac{1}{1 + m_j/n_{j-1}} \tag{21}$$

Term (II): Since $|\mathcal{L}| \leq C$, by Bernstein's inequality:

$$\Pr\left[\text{Term (II)} > \epsilon\right] \leq \exp\left(\frac{-n_K \epsilon^2/2}{\bar{\sigma}^2 + \epsilon C/3}\right) \tag{22}$$

where $\bar{\sigma}^2 = (n_0\sigma_0^2 + \sum_j m_j \sigma_j^2)/n_K$. With probability at least $1-\delta$:

$$\text{Term (II)} \leq \frac{2C\log(1/\delta)}{3n_K} + \sqrt{\frac{2\bar{\sigma}^2 \log(1/\delta)}{n_K}} \tag{23}$$

Combining Terms (I) and (II) yields the stated bound. □

Lemma G.1 (Telescoping Lemma)**.** *If $a_i, b_i \geq 0$ for $i \in \{1, \ldots, n\}$, then:*

$$\left|\prod_{i=1}^{n} a_i - \prod_{i=1}^{n} b_i\right| \leq \sum_{i=1}^{n} |a_i - b_i| \cdot \prod_{j<i} a_j \cdot \prod_{j>i} b_j \tag{24}$$

Proof. Write the left-hand side as a telescoping sum $\sum_{i=1}^{n} \prod_{j<i} a_j \cdot (a_i - b_i) \cdot \prod_{j>i} b_j$ and apply the triangle inequality with $a_i, b_i \geq 0$. □

H KL Divergence Monotonicity Under Topological Ordering

Let $\mathrm{KL}_j = \mathrm{KL}(p^*(x,t) \,\|\, \nu_j(x,t))$ where ν_j is the distribution of (x,t) after step j. We show $\mathrm{KL}_0 \geq \mathrm{KL}_1 \geq \cdots \geq \mathrm{KL}_K = 0$.

This is a population-level statement: ν_j is the distribution obtained when step j fully replaces the confounded conditional by the marginal. With a finite budget, the mixture q_j can add mass but cannot remove it from over-represented regions ($\psi_j \leq 0$), so each step de-confounds only partially, and the resulting residual is the distributional mismatch term of Theorem 3.3. Partial de-confounding is nonetheless useful in practice, since any reduction in mismatch gives the outcome model a more representative training distribution; empirically the mismatch decreases monotonically across steps (Figure 3(b)).

After step j, treatments $\{T_{\sigma(1)}, \ldots, T_{\sigma(j)}\}$ have been de-confounded. The distribution ν_j factorizes as:

$$\nu_j(x,t) = p(x) \cdot \prod_{i=1}^{j} p(t_{\sigma(i)}) \cdot \prod_{i=j+1}^{K} p(t_{\sigma(i)} \mid x, t_{\mathrm{pa}(\sigma(i))}) \tag{25}$$

The KL reduction at step j is:

$$\mathrm{KL}_{j-1} - \mathrm{KL}_j = \mathbb{E}_{(x,t_{\mathrm{pa}(\sigma(j))}) \sim \nu_{j-1}} \left[\mathrm{KL}\left(p(t_{\sigma(j)}) \,\|\, p(t_{\sigma(j)} \mid x, t_{\mathrm{pa}(\sigma(j))}) \right) \right] \geq 0 \tag{26}$$

since KL divergence is non-negative. The equality $\mathrm{KL}_K = 0$ follows because $\nu_K = p(x) \cdot \prod_i p(t_i) = p^*(x,t)$.

Optimality of reverse topological ordering. Under the reverse topological ordering, at step j the treatment $T_{\sigma(j)}$ has all its children already de-confounded. This means the KL reduction at each step is *exact* (not approximate): the conditional $p(t_{\sigma(j)} \mid x, t_{\mathrm{pa}(\sigma(j))})$ is the true confounding mechanism, uncontaminated by dependencies on not-yet-de-confounded descendants.

I Proof of Identifiability for CTE

Under Assumptions 2.1–2.4:

1. **Positivity:** $p(t \mid x) > 0$ for all $t \in \mathcal{T}$, $x \in \mathcal{X}$.
2. **SUTVA** [10]: No interference between units.
3. **No unmeasured confounders:** All common causes of T and Y are observed in X.
4. **Known DAG:** The structure G among treatments is known.

Under these assumptions, X blocks all backdoor paths from T to Y:

$$\mathbb{E}[Y \mid \mathrm{do}(T = t)] = \int y \, p(y \mid t, x) \, p(x) \, dy \, dx \tag{27}$$

The positivity assumption ensures all pairs (t,x) have positive density, so the integrand can be evaluated everywhere.

I.1 Identifiability of Single-Cause ITE at Each Step

At step j, PACE performs ITE estimation for treatment $T_k = T_{\sigma(j)}$, conditioning on $(x, t_{\mathrm{pa}(k)})$ as the adjustment set. We need to show that $\mathbb{E}[Y \mid \mathrm{do}(T_k = t_k), X = x, T_{\mathrm{pa}(k)} = t_{\mathrm{pa}(k)}]$ is identifiable.

By the backdoor criterion applied to the known DAG: the set $(X, T_{\mathrm{pa}(k)})$ blocks all backdoor paths from T_k to Y. This is because:

1. Any backdoor path from T_k to Y must pass through either X or a parent of T_k (since X confounds all treatments and parents directly influence T_k).
2. Conditioning on $(X, T_{\mathrm{pa}(k)})$ blocks all such paths.
3. We do not condition on descendants of T_k (they have already been de-confounded in earlier steps), avoiding collider bias.

Therefore: $\mathbb{E}[Y \mid \text{do}(T_k = t_k), X = x, T_{\text{pa}(k)} = t_{\text{pa}(k)}] = \mathbb{E}[Y \mid T_k = t_k, X = x, T_{\text{pa}(k)} = t_{\text{pa}(k)}]$.

J Time Complexity

Let $\mathbb{C}_{\text{ITE}}(m, d)$ denote the cost of training a single-cause ITE model on m samples with d-dimensional adjustment set. Let $\mathbb{C}_{\text{sup}}(m)$ denote the cost of training the final supervised model on m samples.

With the known DAG and deterministic ordering ($l = 1$):

1. At step j, the ITE model has adjustment set $(x, t_{\text{pa}(\sigma(j))})$ of dimension $D + |\text{pa}(\sigma(j))|$, trained on n_{j-1} samples. Total ITE training cost: $\sum_{j=1}^{K} \mathbb{C}_{\text{ITE}}(n_{j-1}, D + |\text{pa}(\sigma(j))|)$.
2. At step j, we also fit a density model for $p(t_{\sigma(j)} \mid x, t_{\text{pa}(\sigma(j))})$ on n_{j-1} samples with $D + |\text{pa}(\sigma(j))|$ features.
3. Final supervised model: $\mathbb{C}_{\text{sup}}(n_K)$.

Overall complexity: $\mathcal{O}\left(\sum_{j=1}^{K} \mathbb{C}_{\text{ITE}}(n_{j-1}, D + |\text{pa}(\sigma(j))|) + \mathbb{C}_{\text{sup}}(n_K)\right)$.

Compared to the structure-agnostic version (which requires l random permutations and conditions on all $K - 1$ other treatments), the DAG-aware version:

- Eliminates the factor of l (single pass).
- Reduces the dimensionality of each ITE model from $D + K - 1$ to $D + |\text{pa}(\sigma(j))|$.

For sparse DAGs (where $|\text{pa}(j)| \ll K$), this represents a substantial computational saving.

K Full Dataset Specifications

We provide complete specifications for all datasets used in the experiments.

K.1 Dataset 1: IHDP-MultiCause ($\mathcal{D}^{\text{ihdp}}$)

Semi-synthetic dataset using real covariates from the Infant Health and Development Program [7].

- **Parameters:** $n = 747$, $K = 4$, $D = 25$.
- **Covariates (real):** 25 features from the IHDP study including birth weight, head circumference, weeks born preterm, prenatal care visits, mother's age, education, marital status, and demographic indicators.
- **Treatment components (simulated):**
  1. t_1: Intensive home visits (binary).
  2. t_2: Developmental screening frequency (binary: low/high).
  3. t_3: Parent group sessions (binary).
  4. t_4: Nutritional supplementation (binary).
- **Treatment DAG:** $T_1 \to T_4$ (home visits determine nutrition eligibility), $T_2 \to T_4$ (screening triggers nutrition).
- **Treatment assignment (training):** Strongly confounded by covariates with confounding strength $\beta = 2.0$.
- **Treatment assignment (test):** $t_j \sim \text{Bernoulli}(0.5)$ independently.
- **Outcome:** Continuous cognitive score with main effects, pairwise interactions, and heterogeneous treatment effects moderated by covariates.
- **Metric:** RMSE.

K.2 Dataset 2: Synthetic-Extreme ($\mathcal{D}^{\text{ext}}$)

Fully synthetic dataset with strong confounding, treatment interactions, and a star DAG.

- **Parameters:** $n = 5000$, $K = 5$, $D = 20$.
- **Covariates:** $x \sim \mathcal{N}(0, I_D)$.
- **Treatment DAG:** Star topology $T_1 \to \{T_2, T_3, T_4, T_5\}$.
- **Treatment assignment (training):** Strong confounding ($\beta = 3.0$). Root t_1 depends heavily on x; children depend on t_1 and x, creating highly correlated treatments.
- **Treatment assignment (test):** $t_j \sim \text{Bernoulli}(0.5)$ independently.
- **Outcome:** Nonlinear with main effects, all pairwise interactions, a three-way interaction ($t_1 \cdot t_2 \cdot t_3$), and heterogeneous effects ($x_j \cdot t_j$).
- **Metric:** RMSE.

K.3 Dataset 3: Twins-MultiCause ($\mathcal{D}^{\text{twn}}$)

Semi-synthetic dataset using real covariates from US twin birth records [16].

- **Parameters:** $n = 5000$, $K = 4$, $D = 30$.
- **Covariates (real):** 30 maternal features including age, education, prenatal care visits, weight gain, smoking history, alcohol use, previous births, and demographic indicators.
- **Treatment components (composite maternal behaviors):**
  1. t_1: Adequate prenatal care (≥ 8 visits).
  2. t_2: No smoking during pregnancy.
  3. t_3: Adequate weight gain (25–35 lbs).
  4. t_4: No alcohol during pregnancy.
- **Treatment DAG:** $T_1 \to T_3$ (prenatal care influences weight management), $T_2 \to T_4$ (smoking cessation correlates with alcohol cessation).
- **Treatment assignment (training):** Confounded by maternal characteristics ($\beta = 2.0$).
- **Treatment assignment (test):** $t_j \sim \text{Bernoulli}(0.5)$ independently.
- **Outcome:** Binary survival indicator with main effects and pairwise interactions.
- **Metric:** ROC-AUC.

K.4 Dataset 4: Hillstrom Email Marketing ($\mathcal{D}^{\text{hil}}$)

Semi-synthetic dataset using real covariates and outcomes from the MineThatData email marketing RCT [8].

- **Source:** Kevin Hillstrom's MineThatData E-Mail Analytics Challenge. Available via `sklift.datasets.fetch_hillstrom` (scikit-uplift package).
- **Parameters:** $n = 64{,}000$, $K = 3$, $D = 8$.
- **Covariates (real):** Recency, purchase history, category preferences (mens/womens), newbie flag, channel (phone/web/multichannel), zip code, segment.
- **Treatment components:**
  1. t_1: Email sent (binary).
  2. t_2: Email category (men's/women's merchandise).
  3. t_3: Channel targeting (broad/channel-specific).
- **Treatment DAG:** $T_1 \to T_2$, $T_1 \to T_3$ (sending email determines category and targeting).
- **Treatment assignment (training):** Confounded via biased subsampling of the RCT.
- **Treatment assignment (test):** Full RCT data (randomized treatments).
- **Outcome (real):** Binary conversion indicator from the RCT.
- **Metric:** ROC-AUC.

K.5 Dataset 5: Controlled-Synthetic ($\mathcal{D}^{\text{ctrl}}$)

Configurable synthetic DGP used for ablation studies (topology, bias, misspecification).

- **Parameters:** $n = 5000$, $K = 6$, $D = 15$ (default).
- **Covariates:** $x \sim \mathcal{N}(0, I_D)$.
- **Treatment DAG:** Configurable (chain, star, tree, dense, or empty).
- **Treatment assignment:** $t_j = \mathbb{1}[\sigma(\beta \cdot B_j^\top x + C_j^\top t_{\text{pa}(j)}) > u_j]$ with configurable confounding strength β.
- **Outcome:** $y = x^\top \beta_x + t^\top \beta_t + \sum_{i<j} \gamma_{ij} t_i t_j + \epsilon$.
- **Ablation variants:**
  1. *Topology:* DAG varied across chain, star, tree, dense, empty.
  2. *Bias:* Confounding strength $\beta \in \{1.0, 2.0, 3.0, 4.0, 5.0\}$.
  3. *Misspecification:* Provided DAG differs from true DAG (25–50% edges missing/added).
- **Metric:** RMSE.

L Additional Experimental Details

Computational setup. All experiments were run on NVIDIA A10G GPUs (24GB each). Neural networks were implemented in PyTorch with Adam optimizer, cosine annealing learning rate schedule, and early stopping (patience 5). Batch size was 8192.

Hyperparameters. For all methods:

- Final outcome model: 3-layer MLP (256-128-64), BatchNorm, ReLU, Dropout(0.1).
- ITE models (PACE and SCP): 2-layer MLP (64-32).
- Propensity models: 2-layer MLP (128-64).
- Learning rate: 3×10^{-3} with cosine annealing.
- Weight decay: 10^{-5}.
- Maximum epochs: 40 (final), 20 (ITE).
- Early stopping patience: 5 epochs.

SCP configuration. $l = 2$ random permutations; 1:1 augmentation ratio per step.

Deconfounder. PCA with 3 components for latent confounder inference.

VRS. Propensity weights clipped to $[0.01, 10]$ and normalized to unit mean.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims are accurately presented and proved in the paper.

   Guidelines:

   - The answer [NA] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [NA] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Discusses it in the experiment section.

   Guidelines:

   - The answer [NA] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Full assumptions are provided and complete proofs are provided.

Guidelines:

- The answer [NA] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes experiment setup, hyperparameters and datasets are described in complete detail.

   Guidelines:

   - The answer [NA] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Full description of datasets and hyper parameters are provided.

Guidelines:

- The answer [NA] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Full details are provided to be able to replicate the results.

   Guidelines:

   - The answer [NA] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes provided wherever applicable

   Guidelines:

   - The answer [NA] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes provided wherever applicable.

   Guidelines:

   - The answer [NA] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Adhered to the NeurIPS Code of Ethics throughout.

   Guidelines:

   - The answer [NA] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No] , they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: No negative societal impact of the work. The approach can have broad applicability in medicine, e-commerce, engineering etc. However, there is no scope of any potential negative impact that we are aware of.

    Guidelines:

    - The answer [NA] means that there is no societal impact of the work performed.
    - If the authors answer [NA] or [No] , they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer [NA] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer [NA] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets introduced in the paper well documented.

    Guidelines:

    - The answer [NA] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer [NA] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer [NA] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Core method development in this research does not involve LLMs.

Guidelines:

- The answer [NA] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.