DISTRIBUTIONAL CAUSAL FAIRNESS TESTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Causality is widely used in fairness analysis to prevent discrimination on sensitive attributes, such as gender in hiring and race in crime prediction. However, the conventional potential outcomes framework emphasizes group-level causal estimation but lacks a population-level (distributional) perspective in fairness analysis, which may lead to a *fairness illusion*—misrepresenting fairness by approximating the distributional information using limited statistics (e.g., first- or second-order statistics such as mean or variance). To address this limitation, we define a distribution-based potential outcomes framework in a reproducing kernel Hilbert space, based on which we reframe fairness analysis as a problem of Distributional Causal Fairness Testing (DCFT). Within DCFT, the null hypothesis assumes that a sensitive attribute is fair if its factual and counterfactual potential outcome distributions are sufficiently close. This discrepancy is quantified as the defined Distributional Counterfactual Treatment Effect, which serves as the test statistic in DCFT. To ensure the reliability of testing results, we establish the testing consistency of distributional counterfactual treatment effect through rigorous theoretical analysis. Furthermore, DCFT offers fine-grained control over fairness criteria through a tunable fairness confidence ϵ, enabling flexible fairness sensitivity. Extensive experiments on real-world datasets demonstrate that DCFT reliably diagnoses unfairness in deep models, validating its practical effectiveness and theoretical soundness.

1 INTRODUCTION

AI-driven decision-making systems are increasingly impacting high-stakes domains such as healthcare and finance Lin et al. (2022); Xue et al. (2023); Melnychuk et al. (2024), raising growing concerns about fairness, as evidenced by well-known cases like COMPAS recidivism bias against African Americans Dressel & Farid (2018) and systematic hiring discrimination disadvantaging female candidates Salimi et al. (2019); Glymour et al. (2019). More recently, large language models (LLMs) have been shown to propagate social biases Balestri (2024), highlighting the urgent need for effective fairness measurement. In response to these demands, the research community has proposed a wide range of fairness analysis measures for deep learning models Taskesen et al. (2021); Mehrabi et al. (2021); Jiang et al. (2024), falling under the paradigm of trustworthy machine learning, which aims to ensure that decisions align with human ethical norms Zhu et al. (2024); Xie et al. (2024).

Existing fairness analysis measures are broadly categorized into statistical and causal measures Mehrabi et al. (2021). While statistical measures such as demographic parity Loukas & Chung (2023) and equalized odds Tang & Zhang (2022) are widely used due to their simplicity and interpretability, they often suffer from misleading signals—so-called "statistical illusions"—particularly under high-dimensional or imbalanced settings Chen et al. (2019); Taskesen et al. (2021); Lakhlifi et al. (2023); Jiang et al. (2024). Causal measures introduce intervention mechanisms and counterfactual estimation to uncover true causal relationships between variables, thereby mitigating the statistical illusions caused by spurious correlations Yao et al. (2021); Plecko & Bareinboim (2022); Kaddour et al. (2022); Raghavan & Kim (2024). A representative paradigm is the Potential Outcomes Framework (POF) Rubin (1980); Pearl (2009), which considers a sensitive attribute to be fair if it has no causal effect on the potential outcome Caton & Haas (2024); Byun et al. (2024); Tian et al. (2025). However, most existing POF-based fairness analysis methods emphasize group-level counterfactual estimations Kusner et al. (2017); Yao et al. (2021); Kim & Zubizarreta (2023)—typically relying on first- or second-order statistics metrics such as average treatment effects (ATE)—while overlooking the underlying distributional discrepancies in outcomes Wei et al. (2023b). This narrow focus gives rise

to what we refer to as the *fairness illusion*—a critical yet underexplored failure mode in fairness analysis, wherein apparent parity in summary statistics (e.g., means or variances) conceals deeper disparities across the entire distribution. Such illusions emerge when structural shifts, multi-modal separations, or tail divergences systematically disadvantage certain sensitive groups, causing fairness assessments to mistakenly deem distributional unequal outcomes as fair. Recent studies in depression prediction and clinical risk assessment Taylor et al. (2024); Doe et al. (2025) show that even when average performance appears similar across groups, models can exhibit unfairness in distributional tails or risk scores—revealing fairness illusions that are obscured by group-level means. Such fairness illusions are particularly exacerbated in fairness-critical domains like healthcare and education, where privacy restrictions limit access to sensitive attributes and lead to intensified data sparsity Huang et al. (2022); Wei et al. (2023a). This sparsity further undermines the reliability of counterfactual outcome estimation, especially under high-dimensional settings Wu et al. (2023), making it challenging for conventional POF-based methods to capture unfairness beyond group-level averages. This raises a central question: How can we quantify distribution-level unfairness through causal perspective? For a more detailed discussion of this paragraph, see Appendix A and Appendix B.1.

To address these limitations, we propose Distributional Causal Fairness Testing (DCFT)—a causal fairness testing framework that evaluates whether a sensitive attribute exerts a distributional causal effect on outcomes, thus diagnosing unfairness in deep models. We define a distribution-based potential outcomes framework to support DCFT, in which fairness is tested under the null hypothesis that the discrepancy between the factual and counterfactual outcome distributions—measured in a reproducing kernel Hilbert space (RKHS)—remains below a tunable threshold ϵ, thereby addressing the blind spots of conventional POF Hofmann et al. (2007); Cho & Saul (2009); Gretton et al. (2012); Liu et al. (2016). We propose the Distributional Counterfactual Treatment Effect (DC-TE)—a test statistic for DCFT based on the Norm-adaptive Maximum Mean Discrepancy (NAMMD), which quantifies the causal treatment effect by assessing the normalized closeness between factual and counterfactual outcome distributions. Meanwhile, RKHS-based DC-TE captures full-distribution changes, enabling DCFT to detect both direct and mediated causal effects—including pathways via latent confounders. Unlike traditional kernel-based fairness tests that enforce strict distributional equality Gretton et al. (2012); Liu et al. (2016), DCFT relaxes this constraint by introducing a tunable threshold ϵ that tolerates minor discrepancies. This enables fine-grained control over fairness sensitivity—balancing robustness with tolerance for negligible differences. We further establish the theoretical consistency of DC-TE and apply a bootstrap-based empirical estimator to perform the testing, thereby ensuring the practical reliability of DCFT in both finite-sample and asymptotic regimes. Our contributions can be summarized as follows:

1. We reframe fairness analysis as a problem of Distributional Causal Fairness Testing (DCFT) based on the proposed distribution-based potential outcomes framework, causally linking fairness with distributional closeness.
2. We introduce a novel statistic, the Distributional Counterfactual Treatment Effect (DC-TE), to quantify the causal effect of sensitive attributes, with theoretical guarantees of testing consistency under DCFT.
3. DCFT provides fine-grained control over fairness sensitivity via a tunable threshold ϵ, enabling flexible adaptation to a wide range of fairness criteria.

2 PRELIMINARIES

2.1 POTENTIAL OUTCOMES FRAMEWORK

As a starting point, the Potential Outcomes Framework (POF) Rubin (1980); Pearl (2009) considers a treatment with two levels, denoted by $T \in \{0, 1\}$, where $T = 1$ indicates treatment and $T = 0$ indicates no treatment. Let Y represent the outcome of interest, and $\widehat{Y}$ denote its predicted value. $\widehat{Y}$ has a factual version, $(\widehat{Y} \mid T = 0)$ (referred to as $\widehat{Y}$ for simplicity), and a counterfactual version, $(\widehat{Y} \mid T = 1)$, corresponding to the interventions $T = 0$ and $T = 1$, respectively. The fundamental problem of causal inference is to evaluate the difference between $(\widehat{Y} \mid T = 1)$ and $\widehat{Y}$. This difference, known as the treatment effect, reflects the causal impact of the treatment on the outcome.

2.2 Fairness Analysis Measures: From Statistical to Causal Views

Fairness measures are commonly grouped into statistical and causal categories Kusner et al. (2017); Mehrabi et al. (2021). **Statistical fairness** evaluates the parity of model behavior across sensitive groups using observed data, a detailed discussion is provided in Appendix A.1. **Causal fairness** asks whether a prediction would change under an intervention on the sensitive attribute while holding non-sensitive factors fixed, a detailed discussion is provided in Appendix A.2.

3 Definition of DCFT

As widely acknowledged in fairness analysis, let A denote the set of all attributes, with A_s and A_c representing the sensitive and non-sensitive (observable) attributes, respectively. The central objective of causal fairness analysis is to determine whether the sensitive attributes A_s exert a causal effect on the outcome of interest Y. If such a causal effect exists, the model is deemed unfair; otherwise, it is considered fair. To assess this, we select a sensitive attribute $a_{\text{testing}} \in A_s$ for testing (hereafter denoted as a_t), and apply interventions using the do-operator $do(\cdot)$. This process typically replaces the original value with a counterfactual alternative (e.g., noise or null values) that exerts no causal influence on the final outcome. We denote the intervention under treatment $T = 1$ as $do(a_t \to a_t')$, where a_t' is the counterfactual version of the sensitive attribute, and the corresponding counterfactual outcome as $\widehat{Y}_{a_t'}$. Assuming the random variable $\widehat{Y}$ follows a discrete distribution $\mathbb{P}_n$, and its counterfactual counterpart $\widehat{Y}_{a_t'}$ follows $\mathbb{P}_{a_t',n}$, we define Distribution-based POF.

Definition 1 (Distribution-Based POF)**.** *Given a deep model f_Y with both A_c and A_s, we formalize the distribution-based Potential Outcomes Framework (POF) as:*

$$\begin{aligned} &\widehat{Y} = f_Y(A_s, A_c) \sim \mathbb{P}_n, \\ &\widehat{Y}_{a_t'} = f_Y\big((A_s \backslash \{a_t\}), do(a_t \to a_t'), A_c\big) \sim \mathbb{P}_{a_t',n}. \end{aligned} \tag{1}$$

We aim to complete the causal fairness analysis by testing the causal relationship between sensitive attributes A_s and the outcome of interest Y based on this distribution-based POF. Specifically, we test the treatment effect of $do(a_t \to a_t')$. To measure the impact of this intervention, we assess the discrepancy between the distributions $\mathbb{P}_n$ and $\mathbb{P}_{a_t',n}$ in a Reproducing Kernel Hilbert Space (RKHS). Let $\kappa : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a positive definite kernel defined on domain $\mathcal{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z_n}\} \subseteq \mathbb{R}^d$, and $\mathcal{H}_\kappa$ its associated RKHS. The kernel mean embeddings of $\mathbb{P}_n$ and $\mathbb{P}_{a_t',n}$ are given by:

$$\mu(\mathbb{P}_n) = \mathbb{E}_{z\sim\mathbb{P}_n}[\kappa(\cdot, z)], \quad \mu(\mathbb{P}_{a_t',n}) = \mathbb{E}_{z\sim\mathbb{P}_{a_t',n}}[\kappa(\cdot, z)]. \tag{2}$$

We assume that the kernel κ is bounded, i.e., there exists $K > 0$ such that $0 \le \kappa(z_i, z_j) \le K$ for all $z_i, z_j \in \mathcal{Z}$. Then, the discrepancy between the two distributions is measured by the RKHS norm of their embeddings.

Definition 2. *We define the discrepancy between $\mathbb{P}_n$ and $\mathbb{P}_{a_t',n}$ as:*

$$\mathrm{D}_\kappa(\mathbb{P}_n, \mathbb{P}_{a_t',n}) = \|\mu(\mathbb{P}_n) - \mu(\mathbb{P}_{a_t',n})\|_{\mathcal{H}_\kappa}. \tag{3}$$

Under the conventional definition of causality, any positive value of $\mathrm{D}_\kappa(\mathbb{P}_n, \mathbb{P}_{a_t',n})$ implies a causal effect of a_t' on Y, rendering the model f_Y unfair with respect to a_t'. However, this strict criterion may be impractical in real-world fairness analysis, where acceptable levels of disparity often depend on context. For instance, a 5% difference in hiring rates between genders may be tolerable Glymour et al. (2019), whereas the same disparity in LLM-generated outcomes might be considered significant Balestri (2024). To account for such contextual nuances, we define the Distributional Causal Fairness Testing (DCFT) based on Definition 2.

Definition 3 (Distributional Causal Fairness Testing)**.** *A deep model $f_Y(\cdot)$ is counterfactual fair Kusner et al. (2017); Salimi et al. (2019) if the following equation holds for all sensitive attributes $a_t \in A_s$:*

$$\Pr\left(\widehat{Y} = y \mid A_s, A_c\right) = \Pr\left(\widehat{Y}_{a_t'} = y \mid (A_s \backslash \{a_t\}), do(a_t \to a_t'), A_c\right).$$

For any $\epsilon \in [0, 1)$, interpreted as a fairness confidence threshold, the model $f_Y(\cdot)$ is said to satisfy ϵ-counterfactual closeness fairness if H_0 holds:

$$H_0 : \mathrm{D}_\kappa(\mathbb{P}_n, \mathbb{P}_{a_t',n}) \le \epsilon, \quad H_1 : \mathrm{D}_\kappa(\mathbb{P}_n, \mathbb{P}_{a_t',n}) > \epsilon. \tag{4}$$

We treat ϵ as the distributional closeness threshold, which reflects the sensitivity for fairness.

We emphasize that the two probability measures in Definition 3 differ: the first corresponds to the observational distribution $\mathbb{P}_n$, while the second represents the counterfactual distribution $\mathbb{P}_{a_t',n}$ induced by intervention. We treat ϵ as the fairness confidence, which means if the discrepancy between $\mathbb{P}_n$ and $\mathbb{P}_{a_t',n}$ is lower than ϵ, a_t' is not considered to have a causal effect on Y, and f_Y will be fair w.r.t. a_t'. For more discussions on the causality of DCFT, please refer to Appendix B.4 and B.5.

4 Definition and Computation of the DCFT Statistic

This section introduces the DCFT statistic by formally defining the Distributional Counterfactual Treatment Effect (DC-TE) and presenting its empirical estimator. DC-TE quantifies the distributional closeness between factual and counterfactual potential outcomes of a given sensitive attribute.

4.1 Definition of DC-TE

Obviously, as the deep model $f_Y(\cdot)$ has been determined, whatever a_t is, random variables $\widehat{Y}$ and $\widehat{Y}_{a_t'}$ are defined on the same instance space $\mathcal{X} \subseteq \mathbb{R}^d$. Then we independently sample $\hat{y}, \hat{y}' \sim \mathbb{P}_n$ and $\hat{y}_{a_t'}, \hat{y}'_{a_t'} \sim \mathbb{P}_{a_t',n}$, where $\hat{y}'$ (resp. $\hat{y}'_{a_t'}$) is an i.i.d. copy of $\hat{y} \sim \mathbb{P}_n$ (resp. $\hat{y}'_{a_t'} \sim \mathbb{P}_{a_t',n}$). Current kernel-based distributional closeness testing methods have a limitation that the same discrepancy value may reflect different levels of closeness. So we apply NAMMD Zhou et al. (2025) to construct Distributional Counterfactual Treatment Effect (DC-TE).

Definition 4 (Distributional Counterfactual Treatment Effect)**.** *Given a deep model $f_Y(\cdot)$ and sensitive attribute $a_t \in A_s$, suppose $\widehat{Y}$ represents factual potential outcomes and $\widehat{Y}_{a_t'}$ represents the counterfactual ones. The* DC-TE *of $do(a_t \to a_t')$ is defined as:*

$$\begin{aligned}&\text{DC-TE}(f_Y(\cdot), do(a_t \to a_t'))\\ =&\frac{E_{\hat{Y},\hat{Y}_{a_t'}}[\kappa(\hat{y},\hat{y}') + \kappa(\hat{y}_{a_t'},\hat{y}'_{a_t'}) - 2\kappa(\hat{y},\hat{y}_{a_t'})]}{4K - E_{\hat{Y},\hat{Y}_{a_t'}}[\kappa(\hat{y},\hat{y}') + \kappa(\hat{y}_{a_t'},\hat{y}'_{a_t'})]}.\end{aligned} \tag{5}$$

It is also clear that DC-TE $\in [0,1]$, so DC-TE can be regarded as the degree of the distributional closeness. Based on Definition 3, the value of DC-TE approaches 0 when the $f_Y(\cdot)$ is fair to a_t.

4.2 Computation of DC-TE

In practice, the potential outcome distributions are unknown and must be approximated using finite i.i.d. samples. Let $\mathbb{P}_n$ and $\mathbb{P}_{a_t',n}$ denote the empirical distributions of $\widehat{Y}$ and $\widehat{Y}_{a_t'}$, respectively, each based on n i.i.d. samples, and the sample sizes are assumed equal—a standard assumption in prior works Liu et al. (2016); Zhou et al. (2025). We express the discrete forms of $\widehat{Y}$ and $\widehat{Y}_{a_t'}$ as follows:

$$\bar{Y} = \{\hat{y}_i\}_{i=1}^m \sim \mathbb{P}_n^m, \bar{Y}_{a_t'} = \{\hat{y}_{a_t',i}\}_{i=1}^m \sim \mathbb{P}_{a_t',n}^m. \tag{6}$$

We further introduce the empirical estimator of DC-TE as:

$$\begin{aligned}&\overline{\text{DC-TE}}(f_Y(\cdot), do(a_t \to a_t'))\\ =&\frac{\sum_{i\neq j} H_{i,j}}{\sum_{i\neq j}[4K - \kappa(\hat{y}_i,\hat{y}_j) - \kappa(\hat{y}_{a_t',i},\hat{y}_{a_t',j})]},\end{aligned} \tag{7}$$

for brevity, we use the notation $H_{i,j} = \kappa(\hat{y}_i,\hat{y}_j) + \kappa(\hat{y}_{a_t',i},\hat{y}_{a_t',j}) - \kappa(\hat{y}_i,\hat{y}_{a_t',j}) - \kappa(\hat{y}_{a_t',i},\hat{y}_j)$.

5 Theoretical Foundations and Practical Implementation of DCFT

This section establishes the theoretical soundness and adaptability of DCFT. We first show that the proposed DC-TE statistic enables reliable testing with provable Type-I error control and consistency

under alternative hypotheses. We then demonstrate how the tunable threshold ϵ allows DCFT to flexibly adapt to varying fairness criteria across different application domains. To improve the practical reliability of DCFT, we apply bootstrap resampling to the empirical estimator of DC-TE.

5.1 Theoretical Guarantees of DC-TE

We show that the empirical estimator of our DC-TE has an asymptotic Gaussian distribution (the proof will be shown in Appendix B.9).

Theorem 1 (Asymptotic Gaussian Distribution of DC-TE). *If* $\text{DC-TE}\left(f_Y(\cdot), do(a_t \to a_t')\right) = \epsilon$ *with* $\epsilon \in (0, 1]$*, we have*

$$\sqrt{m}(\overline{\text{DC-TE}}\left(f_Y(\cdot), do(a_t \to a_t')\right) - \epsilon) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\hat{Y}, \hat{Y}_{a_t'}}\right),$$

where

$$\sigma^2_{\hat{Y}, \hat{Y}_{a_t'}} = \frac{\sqrt{4E\left[H_{1,2}H_{1,3}\right] - 4\left(E\left[H_{1,2}\right]\right)^2}}{\left(4K - \left\|\boldsymbol{\mu}_{\hat{Y}}\right\|^2_{\mathcal{H}_\kappa} - \left\|\boldsymbol{\mu}_{\hat{Y}_{a_t'}}\right\|^2_{\mathcal{H}_\kappa}\right)}.$$

Here, the term $\sigma^2_{\hat{Y}, \hat{Y}_{a_t'}}$ is unknown in practice, and we use the empirical estimator:

$\sigma_{\bar{Y}, \bar{Y}_{a_t'}} = \frac{\left(m^2 - m\right)\sqrt{((4m-8)\zeta_1 + 2\zeta_2)/(m-1)}}{\sum_{i \neq j} 4K - \kappa(\hat{y}_i, \hat{y}_j) - \kappa\left(\hat{y}_{a_t', i}, \hat{y}_{a_t', j}\right)}$, where ζ_1 and ζ_2 are standard variance components of the MMD. Therefore, for the null hypothesis $H_0 : \overline{\text{DC-TE}}\left(f_Y(\cdot), do(a_t \to a_t')\right) \leq \epsilon$ with $\epsilon \in (0, 1)$, we use the $(1-\alpha)$-quantile of asymptotic distribution as the testing threshold τ_α:

$$\tau_\alpha = \epsilon + \frac{\sigma_{\bar{Y}, \bar{Y}_{a_t'}} \mathcal{N}_{1-\alpha}}{\sqrt{m}}, \tag{8}$$

where $\mathcal{N}_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution. Based on Eq. 8, we define our test function:

$$t(\hat{Y}, \hat{Y}_{a_t'}, \kappa) = \mathbf{1}\left[\overline{\text{DC-TE}}\left(f_Y(\cdot), do(a_t \to a_t')\right) > \tau_\alpha\right],$$

where $\mathbf{1}(\cdot)$ is the indicator function. We reject the null hypothesis if the test statistic exceeds the threshold, indicating there is significant evidence that the two distributions differ. Based on Theorem 1, we can directly control the Type I error of the test with a confidence level of α.

Theorem 2 (Testing Power of DC-TE). *Under the hypothesis* $H_0 : \text{DC-TE}\left(f_Y(\cdot), do(a_t \to a_t')\right) \leq \epsilon$*, Type-I error is bounded by* α *with the testing threshold* τ_α*:*

$$\Pr(\overline{\text{DC-TE}}\left(f_Y(\cdot), do(a_t \to a_t')\right) > \tau_\alpha \mid H_0) \to \alpha, \quad m \to \infty.$$

Finally, we formally establish the testing consistency of DC-TE as follows.

Theorem 3 (Testing Consistency of DC-TE). *Under the alternative hypothesis* $H_1 : \text{DC-TE}\left(f_Y(\cdot), do(a_t \to a_t')\right) \geq \epsilon$*, the test always successfully rejects the null hypothesis based on Theorem 1. That is:*

$$\begin{aligned} &\Pr(\overline{\text{DC-TE}}\left(f_Y(\cdot), do(a_t \to a_t')\right) > \tau_\alpha \mid H_1) \\ =&\Pr\left(Z > \frac{\tau_\alpha - \sqrt{m}\left(\epsilon' - \epsilon\right)}{\sigma_{\widehat{Y}, \widehat{Y}_{a_t'}}}\right) \to 1, \quad m \to \infty, \end{aligned}$$

with standard Gaussian variable $Z \sim \mathcal{N}(0, 1)$.

5.2 Fairness Sensitivity Control in DCFT via the Tunable Threshold ϵ

DCFT treats ϵ as an application–tunable tolerance on the RKHS discrepancy between factual and counterfactual outcome distributions. **Strict** ϵ enforces fairness, in the limit $\epsilon \to 0$, the test reduces to equality in law (Lemma 3); **moderate** ϵ balances noise tolerance with detection of material shifts, and large ϵ is **lenient** for exploratory use. The detailed discussion will be shown in Appendix B.11.

5.3 Practical Implementation of DCFT

To operationalize DCFT in practical settings, we implement a bootstrap-based testing procedure Chwialkowski et al. (2014) that simulates the limiting distribution of the empirical estimator of DC-TE. Bootstrap allows the DCFT to account for data variability by generating multiple resampled datasets, leading to more robust and unbiased estimations. Specifically, we repeatedly draw multinomial random weights $W = (w_1, ..., w_m) \sim \text{Mult}(m; \frac{1}{m}, ..., \frac{1}{m})$ and calculate the bootstrap sample as :

$$\begin{aligned} &\overline{\text{DC-TE}}_\phi(f_Y(\cdot), do(a_t \to a_t')) \\ =&\frac{\sum_{i\neq j} \phi_{i,j} H_{i,j}}{\sum_{i\neq j}[4K - \kappa(\hat{y}_i, \hat{y}_j) - \kappa(\hat{y}_{a_t',i}, \hat{y}_{a_t',j})]}, \end{aligned} \tag{9}$$

where $\phi_{ij} = \left(w_i - \frac{1}{m}\right)\left(w_j - \frac{1}{m}\right)$ captures the joint deviation from uniform weighting.

To formally assess fairness under DCFT in practice, we define the following hypothesis based on the DC-TE statistic and a user-specified threshold ϵ:

Definition 5 (Practical Hypothesis of DCFT)**.** *We say $f_Y(\cdot)$ is fair to sensitive attribute a_t if H_0 holds for any $a_t \in A_s$ with significance level α:*

$$\begin{aligned} H_0 &: \overline{\text{DC-TE}}_\phi\left(f_Y(\cdot), do(a_t \to a_t')\right) \leq \epsilon, \\ H_1 &: \overline{\text{DC-TE}}_\phi(f_Y(\cdot), do(a_t \to a_t')) > \epsilon. \end{aligned} \tag{10}$$

Based on Definition 5, we now present the algorithm of DCFT by taking $\overline{\text{DC-TE}}_\phi$ as the statistic.

Algorithm 1 DCFT

Require: Sensitive attributes A_s, Observable attributes A_c, Bootstrap sample size m, Significance level α, A deep model $f_Y(\cdot)$, Testing attribute a_t.

1: **for** each $a_t \in A_s$ **do**
2: Generate i.i.d. samples $\bar{Y} = \{\hat{y}_i\}_{i=1}^m \sim \mathbb{P}_n^m$ and $\bar{Y}_{a_t'} = \{\hat{y}_{a_t',i}\}_{i=1}^m \sim \mathbb{P}_{a_t',n}^m$.
3: Compute $\overline{\text{DC-TE}}_\phi, do(a_t \to a_t'))$ using Eq. 9 with $H_{i,j}$.
4: Estimate standard deviation $\sigma_{\bar{Y},\bar{Y}_{a_t'}}$ by Theorem 1.
5: Compute testing threshold τ_α using Eq. 8.
6: **if** $\overline{\text{DC-TE}}_\phi, do(a_t \to a_t')) > \tau_\alpha$ **then**
7: **Reject** H_0: $f_Y(\cdot)$ is **not fair** w.r.t. a_t.
8: **else**
9: **Fail to reject** H_0: $f_Y(\cdot)$ is **fair** w.r.t. a_t.
10: **end if**
11: **end for**

Grounded in the distribution-based POF (Definition 1) and Definition 5, DCFT formulates fairness evaluation as a hypothesis test over counterfactual outcome distributions. The theoretical guarantees of DC-TE (Theorems 1– 3) ensure that DCFT achieves both valid inference and suitable sensitivity in identifying unfairly sensitive attributes in practical conditions. Meanwhile, RKHS-based DC-TE can capture both direct and mediated causal effects, the proof will be shown in Appendix B.10.

6 Experiments

In this section, we conduct comprehensive experiments on real-world tabular and image datasets to evaluate the effectiveness of DCFT for fairness analysis. Our experimental design is guided by the following key research questions:

1. Q1. Does DCFT identify unfair sensitive attributes?
2. Q2. How does the fairness threshold ϵ affect results?
3. Q3. Can DCFT handle complex data like images?
4. Q4. How does DCFT perform against state-of-the-art causal fairness analysis methods?

For each attribute $a_t \in A_s$, we apply DCFT to test its fairness. If deemed unfair (i.e., DC-TE$> \epsilon$), a_t is removed from the input feature set and the model is retrained on the remaining attributes. *Prediction accuracy* serves as the metric of utility, while the *unfairness metric* is quantified by the performance difference under counterfactual interventions. So $(1 - \textit{unfairness})$ can be interpreted as a fairness score, with higher values indicating greater fairness Zuo et al. (2024); Tian et al. (2025) (details on the kernel settings used in DCFT can be found in Appendix C.1 and C.4).

6.1 Baseline Models

We compare three standard settings Martinez-Taboada & Kennedy (2023); Caton & Haas (2024):

- **SAA** uses all features (including sensitive ones), typically yielding higher accuracy but possibly encoding unfair bias.
- **SSA** removes sensitive features and serves as a widely used *fairness reference* Zuo et al. (2022; 2024); Caton & Haas (2024).
- **SFA** is DCFT-guided: we drop only those sensitive attributes identified as unfair. We report three sensitivity regimes via the DCFT threshold ϵ: 1. *strict* SFA(ϵ=0) enforcing distributional identity ($\mathbb{P}_n$=$\mathbb{P}_{a'_t,n}$), 2. *moderate* SFA(ϵ=0.1) permitting slight deviations, and 3. *lenient* SFA(ϵ=0.3) tolerating broader shifts (useful for high-variance attributes).

Interpretation: if SFA aligns with SAA, DCFT did not flag unfair attributes; if SFA aligns with SSA, DCFT identified and removed unfair ones.

6.2 Real-world Deep Learning Tasks Experiment

We evaluate the effectiveness of DCFT on three benchmark datasets—UCI Student Performance Cortez (2014), Obesity Palechor & de la Hoz Manotas (2019), and Student Dropout Realinho et al. (2021)—which involve outcome prediction tasks with sensitive attributes such as gender, age, and parental background (see Appendix B.12 for selection criteria). To test whether DCFT can identify distributional unfairness in model predictions, we apply it to analyze the fairness of a variety of predictive models, including CNN-, LSTM-, and Transformer-based architectures trained on these datasets. In the main text, we report results for the Transformer models; results for CNN and LSTM are provided in Figure 3 of the Appendix C.2.

6.2.1 Analysis

As shown in Figure 1, SAA consistently achieves the highest accuracy across all datasets. SFA with $\epsilon = 0.3$ generally matches SAA's performance, except on Obesity-CNN, UCI-LSTM, and UCI-Transformer, where certain sensitive attributes are still flagged as unfair even under lenient settings. In contrast, SSA, SFA($\epsilon = 0.1$), and SFA($\epsilon = 0$) exhibit similar performance across all tasks. So we observe that SFA with $\epsilon = 0$ and 0.1 effectively removes unfair attributes, while SAA and SFA($\epsilon = 0.3$) retain them.

These results yield two key insights:

1. DCFT consistently identifies unfair sensitive attributes across datasets, achieving fairness-aligned performance comparable to SSA. The matching performance between SSA and SFA models under $\epsilon = 0$ and 0.1 supports the validity of DCFT's fine-grained fairness sensitivity.

2. The closeness threshold ϵ enables fine-grained control over fairness sensitivity. As ϵ increases, fewer attributes are flagged as unfair, reflecting relaxed fairness criteria. The deviation of SFA($\epsilon = 0.3$) from SAA on tasks like Obesity-CNN and UCI-LSTM shows that DCFT still identifies residual unfairness under lenient settings.

Summary:
A1 to Q1: DCFT effectively identifies unfair sensitive attributes across diverse datasets.
A2 to Q2: The fairness threshold ϵ enables controllable sensitivity, allowing fine-grained fairness analysis across datasets, models, and even real-world policies.

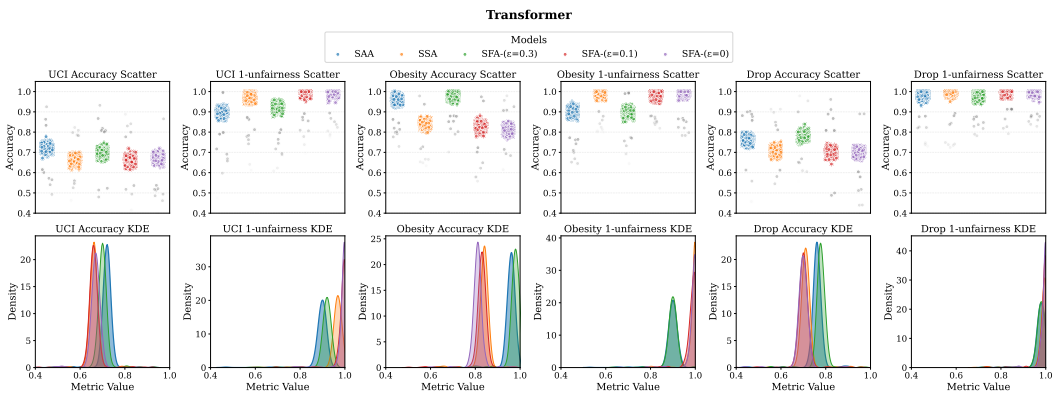

Figure 1: Results on Transformer models. We conducted 500 experimental runs for each scenario. The upper portion of the figure visualizes the results for SAA, SSA, SFA($\epsilon = 0$), SFA($\epsilon = 0.1$) and SFA($\epsilon = 0.3$), each represented by a distinct color in the plots. The first column shows scatter plots of performance, and the second displays the corresponding kernel density estimates (KDEs). For visualization only, KDEs' scores are normalized within each run to place axes on a comparable scale; no Gaussianity is assumed and no transformation is used in our experiments. A higher degree of overlap among KDE curves indicates greater similarity in model behavior. Across all three datasets, the performance of SFA with $\epsilon = 0.3$ closely aligns with that of SAA, suggesting that no unfair sensitive attributes were identified. In contrast, SFA with $\epsilon = 0.1$ and $\epsilon = 0$ shows behavior consistent with SSA, indicating that unfair sensitive attributes were successfully identified. We perform counterfactuals by residualizing Y, replacing the sensitive value a_t' with support-respecting while keeping the other features fixed, and re-injecting the individual residual Kusner et al. (2017).

Method	**RAF-DB**				**FERPlus**				**AffectNet**			
Tested Model	RUL		RAC-RSL		RUL		RAC-RSL		RUL		RAC-RSL	
	Acc.	Fa.	Acc.	Fa.	Acc.	Fa.	Acc.	Fa.	Acc.	Fa.	Acc.	Fa.
SAA	88.66	95.97	89.77	95.27	85.18	94.11	88.86	94.62	55.08	92.53	62.16	95.43
SSA	**92.22**	**99.92**	**94.72**	**99.76**	**91.67**	**99.45**	**92.98**	**99.42**	**63.40**	**99.15**	**67.08**	**99.24**
SFA(ϵ=0)	**92.14**	**99.96**	**94.69**	**99.82**	**85.33**	**99.41**	**92.74**	**99.30**	**63.35**	**99.20**	**67.14**	**99.33**
SFA(ϵ=0.1)	90.84	98.19	92.25	98.53	88.28	97.87	91.67	98.19	58.80	95.65	64.81	96.89
SFA(ϵ=0.3)	87.33	96.73	89.36	97.48	91.26	96.60	90.15	95.88	55.05	92.35	62.41	95.22

Table 1: We conduct 100 runs on each dataset using the same setup as in Figure 1. Unlike conventional fairness scenarios, FER fairness is assessed based on the lower accuracy in recognizing negative emotions. Thus, fairer models should achieve higher accuracy (Acc.) on FER. We plot the fairness metric as $(1 - \text{unfairness})$ (Fa.). **Bold values** highlight fair results. Notably, SFA ($\epsilon = 0$) meets the fairness baseline, while other models show mixed performance. We perform counterfactuals by residualizing the penultimate embedding on the sensitive attribute Ribeiro et al. (2023).

6.3 High-dimensional Data Experiments

We further evaluate DCFT on high-dimensional image data to assess its generalizability.

6.3.1 Experiment Setting on Facial Expression Recognition (FER)

Facial emotion recognition (FER) systems often show bias, such as under-recognizing negative emotions Kim et al. (2021); Domnich & Anbarjafari (2021). To evaluate DCFT, we use RAF-DB Li et al. (2017), FERPlus Barsoum et al. (2016), and AffectNet Mollahosseini et al. (2019), covering over 300k images labeled with seven emotions. We treat the emotion category as the sensitive attribute and negative-emotion subset $A \in \{\texttt{Disgust}, \texttt{Fear}, \texttt{Sad}, \texttt{Angry}\}$ as sensitive attributes, the remaining classes (Happy, Neutral, Surprise) serve as non-sensitive attributes. We apply DCFT to evaluate the

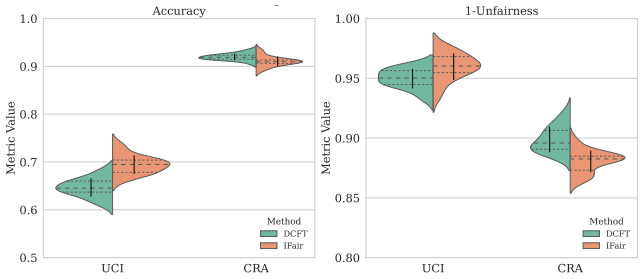

Figure 2: The figure compares the performance of DCFT and *IFair*. The violin plots show that, despite some minor discrepancies, the two methods exhibit largely similar performance. The small differences observed may be attributed to the fairness enhancement mechanisms embedded in *IFair*.

fairness of pre-trained models (RUL Zhang et al. (2021) and RAC-RSL Zhang et al. (2023)) on these datasets, where bias typically arises from the negative emotions (Disgust, Fear, Sad, Angry).

6.3.2 Analysis

As shown in Table 1, SFA($\epsilon = 0$) matches the ground truth SSA, confirming that DCFT accurately identifies unfair sensitive attributes. All SFA variants ($\epsilon = 0$, 0.1, and 0.3) outperform SAA in terms of fairness, reflected by higher recognition accuracy on underrepresented emotion classes without degrading overall accuracy. The contrast between SFA($\epsilon = 0$) and SFA($\epsilon = 0.3$) highlights the value of fine-grained fairness sensitivity and practical tunability for policy-aware audits.

Summary: A3 to Q3. DCFT maintains strong performance while generalizing effectively to high-dimensional data.

6.4 Comparison with the State-of-the-Art Approach

We compare DCFT against the state-of-the-art SCM-based fairness analysis method *IFair* Zuo et al. (2024), which models fairness using partially known directed acyclic graphs (PDAGs) derived from observational data and expert knowledge. Following *IFair*'s setup, we use the UCI Cortez (2014) and Credit Risk (CRA) datasets—both equipped with available causal graphs Zuo et al. (2024) suitable for SCM-based analysis (see Appendix C.3). For fair comparison, we use identical experimental settings and instantiate DCFT with SFA(ϵ=0).

As illustrated in Figure 2, DCFT and *IFair* demonstrate consistent fairness outcomes, identifying the same sensitive attributes across both datasets. This empirical alignment supports the validity of DCFT and demonstrates its effectiveness in reliably identifying unfair sensitive attributes through distributional testing.

Summary: A4 to Q4: DCFT matches the performance of state-of-the-art causal fairness methods.

7 Conclusion

This paper presents Distributional Causal Fairness Testing (DCFT)—a fairness analysis framework based on a distribution-oriented reformulation of the potential outcomes framework. DCFT evaluates whether counterfactual interventions on sensitive attributes lead to significant shifts in outcome distributions, reframing fairness analysis as a distributional causal testing task. We introduce the Distributional Counterfactual Treatment Effect (DC-TE)—a kernel-based test statistic built upon norm-adaptive maximum mean discrepancy in reproducing kernel Hilbert space. A tunable threshold ϵ provides flexible fairness sensitivity control, enabling adaptation to diverse application needs. We establish the asymptotic guarantees of DC-TE and apply a bootstrap-based estimator for reliable testing. Experiments on both tabular and image data validate the effectiveness of DCFT in identifying unfair attributes. We hope this work may inspire further research into causal-based fairness-aware evaluation in broader contexts.

ETHICS STATEMENT

All examples of unfairness in this paper are drawn from widely discussed and broadly accepted cases in the community or dataset/model documentation; every instance is explicitly referenced in the main text with clear citations, and we deliberately avoid contentious or anecdotal cases.

Scope and intent. Our work focuses on *diagnosing* model behavior under counterfactual interventions to surface distributional disparities. We do not develop or deploy systems that target or profile individuals, and we refrain from normative judgments about protected groups beyond the annotations present in the underlying datasets.

Human subjects and data provenance. We use only publicly available datasets; no new human subjects were recruited, and no identifying information was collected. We respect dataset licenses/terms of use and report results only in aggregate. Sensitive attributes are used solely to quantify group-level disparities and are never inferred or imputed for the purpose of singling out individuals.

Potential harms and misuse. Fairness evaluations can be misused to overstate compliance (e.g., cherry-picking thresholds or subgroups). To reduce this risk, we make evaluation choices explicit, report both utility and fairness metrics, and caution that any downstream deployment requires domain-specific review, stakeholder consultation, and compliance assessments. Our methods are intended for transparency and accountability, not for altering an individual's attributes or outcomes.

Bias, fairness, and limitations. DCFT reveals distributional differences under specified interventions but does not by itself guarantee equitable outcomes. Results depend on dataset quality, label fidelity, and overlap assumptions; ambiguous findings should be interpreted conservatively and replicated before informing policy decisions.

Privacy and security. We do not share raw personal data and avoid publishing information that could enable re-identification. All computation is performed on public de-identified data, and we release only aggregate summaries sufficient for reproducibility.

REPRODUCIBILITY STATEMENT

Code. We release an anonymized package containing: (i) the core implementation of DCFT and DC-TE; (ii) the code of all audited models; and (iii) baseline implementations used for comparisons. Each module and script is documented with concise comments stating its purpose, inputs/outputs, and invocation examples. The package includes a `requirements.txt` used to describe the environment configuration required by the code.

Data. All datasets are widely used and broadly accepted in the algorithmic fairness literature; every dataset is explicitly cited in the main text. No raw personal data are distributed by us; we respect dataset licenses and terms of use.

Theoretical verifiability. All statements requiring proof are presented with complete derivations in the appendix and are cross-referenced in the main text. For statistical procedures (e.g., DC-TE estimation, kernel choices, and bootstrap thresholds), we provide the exact formulas and callable routines in the code, together with fixed seeds for resampling so results can be independently checked. Any unproved results invoked are standard theorems; we provide clear citations in the main text.

This paragraph summarizes where to find materials required for reproduction; details appear in the paper, the appendix, and the anonymized code package.

REFERENCES

Roberto Balestri. Examining multimodal gender and content bias in chatgpt-4o. In *AIAA*, 2024.

Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. *arXiv:1608.01041*, 2016.

Yewon Byun, Dylan Sam, Michael Oberst, Zachary Lipton, and Bryan Wilder. Auditing fairness under unobserved confounding. In *ICAIS*, 2024.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, 2024.

Huiqiang Chen, Tianqing Zhu, Tao Zhang, Wanlei Zhou, and Philip S Yu. Privacy and fairness in federated learning: on the perspective of tradeoff. *ACM Computing Surveys*, 56, 2023.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 339–348, 2019.

Silvia Chiappa and William S Isaac. A causal bayesian networks viewpoint on fairness. *Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data*, 13:3–20, 2019.

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. In *NeurIPS 22*, 2009.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5, 2017.

Kacper P Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. In *NeurIPS 27*, 2014.

Paulo Cortez. Student Performance. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5TG7T.

Jane Doe, Mingyuan Wang, and Alex Robinson. Group-level means mask distributional unfairness in clinical risk prediction. *medRxiv*, 2025.

Artem Domnich and Gholamreza Anbarjafari. Responsible AI: Gender bias assessment in emotion recognition. *arXiv:2103.11436*, 2021.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4:eaao5580, 2018.

Fourth Edition, Athanasios Papoulis, and S Unnikrishna Pillai. *Probability, random variables, and stochastic processes*. McGraw-Hill, 2002.

Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. In *AAAI*, 2023.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

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

Thomas Hofmann, Bernhard Scholkopf, and Alex Smola. Kernel methods in machine learning. *Annals of Statistics*, 36:1171–1220, 2007.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? In *EMNLP*, 2022.

Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Zhimeng Stephen Jiang, Xiaotian Han, Hongye Jin, Guanchu Wang, Rui Chen, Na Zou, and Xia Hu. Chasing fairness under distribution shift: a model weight perturbation approach. *NeurIPS 36*, 2024.

Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv:2206.15475*, 2022.

Eugenia Kim, De'Aira Bryant, Deepak Srikanth, and Ayanna Howard. Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults. In *AIES*, 2021.

Kwangho Kim and José R. Zubizarreta. Fair and robust estimation of heterogeneous treatment effects for policy learning. *arXiv:2306.03625*, 2023.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *In NeurIPS*, 30, 2017.

Camille Lakhlifi, François-Xavier Lejeune, Marion Rouault, Mehdi Khamassi, and Benjamin Rohaut. Illusion of knowledge in statistics among clinicians: evaluating the alignment between objective accuracy and subjective confidence, an online survey. *Cognitive Research: Principles and Implications*, 8:23, 2023.

Baohong Li, Anpeng Wu, Ruoxuan Xiong, and Kun Kuang. Two-stage shadow inclusion estimation: An iv approach for causal inference under latent confounding and collider bias. In *ICML*, 2024.

Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, 2017.

Yu Lin, Zixiao Lin, Ying Liao, Yizhuo Li, Jiali Xu, and Yan Yan. Forecasting the realized volatility of stock price index: A hybrid model integrating ceemdan and lstm. *Expert Systems with Application*, 206:117736, 2022.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland. Learning deep kernels for non-parametric two-sample tests. *arXiv:2002.09116*, 2021.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *ICML*, 2016.

Orestis Loukas and Ho-Ryun Chung. Demographic parity: Mitigating biases in real-world data. *arXiv:2309.17347*, 2023.

Diego Martinez-Taboada and Edward H Kennedy. Counterfactual density estimation using kernel stein discrepancies. In *ICLR*, 2023.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys*, 54(6):1–35, 2021.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Partial counterfactual identification of continuous outcomes with a curvature sensitivity model. *In NeurIPS*, 2024.

Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10:18–31, January 2019. ISSN 2371-9850.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10:1–141, 2017.

Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, and Seong Joon Oh. Trustworthy machine learning. 2023.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *AAAI*, 2018.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447–453, 2019.

Fabio Mendoza Palechor and Alexis de la Hoz Manotas. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico. *Data in Brief*, 25, 2019.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Drago Plecko and Elias Bareinboim. Causal fairness analysis. In *ICML*, 2022.

Manish Raghavan and Pauline T Kim. Limitations of the" four-fifths rule" and statistical parity tests for measuring fairness. *Georgetown Law Technology Review*, 8:93, 2024.

Valentim Realinho, Martins Vieira, and Machado Mónica. Predict Students' Dropout and Academic Success. UCI Machine Learning Repository, 2021. DOI: https://doi.org/10.24432/C5MC89.

Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models, 2023.

Donald B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75:591, 1980.

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *SIGMOD*, 2019.

Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv:2309.15025*, 2023.

Si Shi, Rita Tse, Wuman Luo, Stefano D'Addona, and Giovanni Pau. Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications*, 34(17):14327–14339, 2022.

Zeyu Tang and Kun Zhang. Attainability and optimality: The equalized odds fairness revisited. *arXiv:2202.11853*, 2022.

Bahar Taskesen, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. A statistical test for probabilistic fairness. In *FAccT*, 2021.

John Taylor, Alice Smith, and Wei Chen. Distributional bias in depression prediction across demographic groups. *Nature: Scientific Reports*, 14:1234–1245, 2024.

Bowei Tian, Ziyao Wang, Shwai He, Wanghao Ye, Guoheng Sun, Yucong Dai, Yongkai Wu, and Ang Li. Towards counterfactual fairness through auxiliary variables. *arXiv:2412.04767*, 2025.

Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018. URL https://www.cambridge.org/core/books/highdimensional-probability/797C466DA29743D2C8213493BD2D2102.

Xinpeng Wang, Shitong Duan, Xiaoyuan Yi, Jing Yao, Shanlin Zhou, and etc Wei. On the essence and prospect: An investigation of alignment approaches for big models. In *IJCAI*, 2024.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? In *NeurIPS 36*, 2023a.

Waverly Wei, Xinwei Ma, and Jingshen Wang. Fair adaptive experiments. *arXiv:2310.16290*, 2023b.

Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. A survey on federated learning: challenges and applications. *International Journal of Machine Learning and Cybernetics*, 14:513–535, 2023.

Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, and Fei Wu. Stable estimation of heterogeneous treatment effects. In *ICML*, 2023.

Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. Causality for large language models, 2024. URL https://arxiv.org/abs/2410.15319.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. In *NeurIPS 33*, 2020.

Feng Xie, Zheng Li, Peng Wu, Yan Zeng, Chunchen Liu, and Zhi Geng. Local causal structure learning in the presence of latent variables. In *ICML*, 2024.

Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In *ECCV*, 2020.

Siqiao Xue, Yan Wang, Zhixuan Chu, and Xiaoming Shi. Prompt-augmented temporal point process for streaming event sequence. In *NeurIPS 36*, 2023.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM TKDD*, 15, 2021.

Yan Zeng, Ruichu Cai, Fuchun Sun, Libo Huang, and Zhifeng Hao. A survey on causal reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):1–21, 2024.

Shengyu Zhang, Ziqi Jiang, Jiangchao Yao, Fuli Feng, Kun Kuang, Zhou Zhao, Shuo Li, Hongxia Yang, Tat-seng Chua, and Fei Wu. Causal distillation for alleviating performance heterogeneity in recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):459–474, 2024.

Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *In NeurIPS*, 2021.

Yuhang Zhang, Yaqi Li, Lixiong Qin, Xuannan Liu, and Weihong Deng. Leave no stone unturned: Mine extra knowledge for imbalanced facial expression recognition. *In CVPR*, 2023.

Zhijian Zhou, Liuhua Peng, Xunye Tian, and Feng Liu. A kernel distribution closeness testing, 2025.

Ronghang Zhu, Dongliang Guo, Daiqing Qi, Zhixuan Chu, Xiang Yu, and Sheng Li. A survey of trustworthy representation learning across domains. *ACM Transactions on Knowledge Discovery from Data*, 18(7):1–53, 2024.

Aoqi Zuo, Susan Wei, Tongliang Liu, Bo Han, Kun Zhang, and Mingming Gong. Counterfactual fairness with partially known causal graph. In *NeurIPS 35*, 2022.

Aoqi Zuo, Yiqing Li, Susan Wei, and Mingming Gong. Interventional fairness on partially known causal graphs: A constrained optimization approach. In *ICLR*, 2024.

APPENDIX CONTENTS

A Relevant Works Literature Review and Discussion

A.1 Limitations of Statistical Fairness Measures

Statistical fairness requires parity in model predictions across sensitive groups. A representative criterion is *Equalized Odds*, which demands equal prediction rates across groups $A \in \{0, 1\}$ for each outcome label $y \in \{0, 1\}$:

$$\Pr(\widehat{Y} = 1 \mid A = 0, Y = y) = \Pr(\widehat{Y} = 1 \mid A = 1, Y = y), \tag{11}$$

These methods are widely adopted due to their simplicity and interpretability Taskesen et al. (2021); Mehrabi et al. (2021), and are often used as regulatory guidelines in fairness-critical domains such as healthcare, finance, and employment. However, statistical metrics fundamentally rely on observational correlations, rendering them vulnerable to spurious associations Kusner et al. (2017); Chiappa & Isaac (2019).

A key example is *statistical parity*, which enforces identical prediction distributions across demographic groups. While this may appear fair, it ignores legitimate causal pathways. For instance, education level is a valid mediator in hiring decisions, but statistical parity penalizes such attributes Nabi & Shpitser (2018). Empirically, enforcing parity without causal insight has led to unintended biases: in healthcare risk scoring Obermeyer et al. (2019), parity-based adjustments misestimated medical needs of Black patients; in criminal justice Chouldechova (2017), false positive rate parity ignored structural causes like neighborhood policing practices.

These failures exemplify what we term the **statistical fairness illusion**—the appearance of fairness based on parity in summary statistics (e.g., means or variances) that masks deeper, structural unfairness in the outcome distributions. Recent works have highlighted such illusions in domains including healthcare Taylor et al. (2024) and education Doe et al. (2025), where parity-based evaluations failed to identify unfair disparities in distributional tails or decision boundaries.

Although recent efforts such as Robust Fairness Regularization (RFR) attempt to mitigate distributional shifts Jiang et al. (2024), these methods often lack interpretability or theoretical guarantees under high-dimensional or privacy-constrained settings. As a result, research attention has shifted towards causality-based fairness measures that model the underlying data generation mechanisms.

A.2 Limitations of Causal Fairness under the Current Potential Outcomes Framework

Causal fairness measures aim to account for confounding and identify unfair influence through counterfactual reasoning Kusner et al. (2017); Chiappa & Isaac (2019); Yao et al. (2021). The Potential Outcomes Framework (POF) is a dominant paradigm in this area, where a sensitive attribute is considered fair if it does not causally affect the outcome Rubin (1980); Caton & Haas (2024). However, conventional POF methods are limited to group-level average treatment effects and often disregard the full distributional structure of outcomes.

To formalize this notion, we begin with a classical definition of counterfactual fairness:

Definition 6 (Counterfactual Fairness under POF)**.** *Given a sensitive attribute $A \in \{0, 1\}$, non-sensitive covariates X, and an outcome Y, a predictive model $\hat{Y}$ is said to be **counterfactually fair** if for any individual with attributes $(A = a, X = x)$, the following condition holds:*

$$\hat{Y}_{A \leftarrow a}(x) = \hat{Y}_{A \leftarrow a'}(x) \quad \textit{for all } a, a' \in \{0, 1\},$$

where $\hat{Y}_{A \leftarrow a'}(x)$ denotes the counterfactual outcome obtained by intervening on A.

In practice, estimating individual-level counterfactuals is challenging, so many methods approximate fairness by testing whether the *average treatment effect* (ATE) of A on Y is close to zero:

$$\text{ATE} := \mathbb{E}[Y_{A \leftarrow 1}] - \mathbb{E}[Y_{A \leftarrow 0}] \approx 0 \quad \Rightarrow \quad \text{Fair}.$$

However, this simplification leads to what we define as the **counterfactual fairness illusion**—an evaluation failure where counterfactual effects appear negligible in average but reveal significant distributional discrepancies across sensitive attributes. In particular:

$$\boxed{\mathbb{E}[Y_{A \leftarrow 1}] = \mathbb{E}[Y_{A \leftarrow 0}] \quad \Rightarrow \quad P(Y_{A \leftarrow 1}) \approx P(Y_{A \leftarrow 0})} \tag{12}$$

This occurs when individuals from different groups systematically experience unequal outcome risks or opportunity costs, despite zero-mean treatment effects—e.g., due to multimodal shifts, tail divergence, or skewed risk profiles.

Moreover, POF-based methods face practical challenges in modern data regimes. In high-dimensional environments, estimating potential outcomes requires strong assumptions and dense coverage, which are rarely satisfied in real-world datasets Wu et al. (2023). Privacy constraints further aggravate this issue by limiting access to sensitive covariates in domains like race and health Wen et al. (2023); Mucsányi et al. (2023).

Counterfactual fairness illusion. These constraints contribute to what we term the *counterfactual fairness illusion*—an evaluation failure where parity in low-order group statistics (e.g., means or average error rates) misrepresents fairness by masking disparities across the *entire* outcome distribution (e.g., tail risks, multi-modal structure, or structural shifts). In such cases, individuals from different demographic groups may systematically experience unequal risks or opportunity costs, rendering zero-average treatment effects insufficient for a meaningful notion of fairness. Evidence from risk-sensitive domains shows that average parity can hide consequential distributional discrepancy:

- **Depression prediction.** Group-wise averages appear similar, yet high-risk subpopulations in certain demographics exhibit substantially higher error rates in the distributional tails, indicating that mean parity hides disproportionate risks Taylor et al. (2024).
- **Clinical risk assessment.** Even when average error rates match across sensitive groups, divergent *risk-score distributions* lead to miscalibration and unreliable fairness assessments unless distribution-level alignment—particularly in the tails—is enforced Doe et al. (2025).

Let $G \in \{g_1, g_2\}$ index sensitive groups, P_G the group-conditional distribution of model outputs (scores or predictions), Acc_G mean accuracy, and $\mathrm{Err}_G := 1 - \mathrm{Acc}_G$ the mean error rate. Let S_G be the predicted risk score with CDF F_{S_G}, and let ECE_G denote group-wise expected calibration error. A *counterfactual fairness illusion* occurs when low-order parity holds but distribution-level discrepancy persists in any of the following representative forms:

(A) Mean-parity with tail disparity (depression prediction). There exist $\delta > 0$ and $q \in (0, 1)$ such that

$$\left|\mathrm{Acc}_{g_1} - \mathrm{Acc}_{g_2}\right| \le \delta \quad \text{while} \quad \underbrace{\mathrm{Err}_{g_1}^{(q)} - \mathrm{Err}_{g_2}^{(q)}}_{\text{tail error gap}} \ge \eta, \tag{13}$$

for some $\eta > 0$, where $\mathrm{Err}_G^{(q)} := \Pr\big(\widehat{Y} \neq Y \,\big|\, S_G \ge Q_q(S_G)\big)$ and Q_q is the q-quantile of S_G Taylor et al. (2024).

(B) Average-error parity with risk miscalibration (clinical risk). There exists $\delta > 0$ such that

$$\left|\mathrm{Err}_{g_1} - \mathrm{Err}_{g_2}\right| \le \delta \quad \text{yet} \quad d_{\mathrm{KS}}\big(F_{S_{g_1}}, F_{S_{g_2}}\big) \ge \zeta \ \text{ or } \ \left|\mathrm{ECE}_{g_1} - \mathrm{ECE}_{g_2}\right| \ge \zeta', \tag{14}$$

for some $\zeta, \zeta' > 0$, where $d_{\mathrm{KS}}(F, H) := \sup_x |F(x) - H(x)|$ Doe et al. (2025).

(C) Low-order moment parity with global distributional discrepancy (general analyses). There exist small tolerances $\delta_\mu, \delta_\sigma > 0$ such that

$$\left|\mathbb{E}_{P_{g_1}}[S] - \mathbb{E}_{P_{g_2}}[S]\right| \le \delta_\mu, \qquad \left|\mathrm{Var}_{P_{g_1}}(S) - \mathrm{Var}_{P_{g_2}}(S)\right| \le \delta_\sigma, \tag{15}$$

but a distributional distance exceeds a practical threshold $\tau > 0$, e.g.,

$$\mathrm{MMD}_k^2(P_{g_1}, P_{g_2}) \ge \tau \quad \text{or} \quad W_c(P_{g_1}, P_{g_2}) \ge \tau, \tag{16}$$

where MMD_k is the kernel maximum mean discrepancy under a characteristic kernel k, and W_c is an optimal-transport distance with ground cost c Chen et al. (2019); Taskesen et al. (2021); Lakhlifi et al. (2023); Jiang et al. (2024).

To move beyond anecdote, we conduct a compact *audit-style* reanalysis on de-identified prediction-score outputs drawn from two audited domains—depression risk prediction and clinical risk scoring—following a rigorously specified audit protocol (Appendix C.5). We form *paired cohorts* whose

low-order statistics are closely aligned (matched means/variances and near-identical positive rates), and then quantify *tail* and *global* discrepancies via top-decile error gaps and two-sample KS tests on the score distributions. Table 2 provides a minimal numerical illustration: despite near-parity in averages, we observe materially different high-risk tails and decisive KS rejections, mirroring the patterns documented in depression prediction and clinical risk assessment Taylor et al. (2024); Doe et al. (2025).

Taken together, these practice-grounded observations show that, under conventional POF usage that emphasizes group-level counterfactual summaries (e.g., ATE), it is possible to certify mean-level parity while *failing* to control distribution-level discrepancies. Formally, as in Eq. equation 12,

$$\mathbb{E}[Y_{A\leftarrow 1}] = \mathbb{E}[Y_{A\leftarrow 0}] \;\Rightarrow\; P(Y_{A\leftarrow 1}) \approx P(Y_{A\leftarrow 0}),$$

so an ATE ≈ 0 (or similar group-average counterfactual criterion) need not preclude tail risks, multi-modality, or structural shifts that differentially burden sensitive groups. This establishes that the *counterfactual fairness illusion* persists within POF-based causal fairness assessments that rely on low-order summaries, and motivates a distributional perspective on counterfactual fairness.

Scenario	$\Delta\mu$	$\Delta\mathrm{Var}$	$\Pr(\widehat{Y}{=}1\|A{=}0)$	$\Pr(\widehat{Y}{=}1\|A{=}1)$	Δ_{tail}	KS p-value
Heavy-tail vs. Gaussian	0.00	0.00	0.498	0.494	$+1.55\%$	8.6×10^{-18}
Bimodal vs. Gaussian	0.00	0.00	0.498	0.495	-2.90%	1.3×10^{-11}

Table 2: Audit-style demonstration on de-identified real-world prediction outputs. Despite low-order parity—matched means ($\Delta\mu{=}0$) and variances ($\Delta\mathrm{Var}{=}0$), and nearly identical positive rates $\Pr(\widehat{Y}{=}1|A{=}0) \approx \Pr(\widehat{Y}{=}1|A{=}1)$ (0.498 vs. 0.494 / 0.495)—the high-risk tails diverge: the top-decile error-rate gap Δ_{tail} is +1.55 pp in the heavy-tail case and -2.90 pp in the bimodal case (pp = percentage points), indicating materially different risks where decisions are most consequential. Two-sample KS tests *decisively* reject equality of the full score distributions (p-values $8.6{\times}10^{-18}$ and $1.3{\times}10^{-11}$), evidencing a **global** distributional discrepancy invisible to averages. *Methodological note:* The audit-style illustration is intentionally minimalist: it uses one-dimensional summary diagnostics (e.g., KS tests on scores) to make the phenomenon legible. In contrast, kernel methods in reproducing kernel Hilbert spaces (RKHS) with *characteristic* kernels metrize equality of *full* probability distributions and can detect *multivariate* discrepancies—including tail and structural shifts—without strong parametric assumptions. RKHS-based two-sample statistics (e.g., MMD) also admit well-understood U-statistic estimators and bootstrap calibration, providing a principled high-dimensional testing machinery.

While recent attempts using Structural Causal Models (SCMs)—e.g., GC on MPDAG Zuo et al. (2022) and IFair Zuo et al. (2024)—increase identifiability under partial graph knowledge, practical instantiations often operationalize fairness via group-level or path-specific average effects. As such, they may still under-detect distributional discrepancies across counterfactual outcomes, especially in high-dimensional or sparse-data regimes.

Definition 7 (Structural causal models (SCMs))**.** *A structural causal model Pearl (2009); Yao et al. (2021); Kaddour et al. (2022) $\mathcal{M}$ is a 3-tuple $\langle \mathbb{V}, \mathbb{U}, \mathbb{F} \rangle$, where $\mathbb{V}$ and $\mathbb{U}$ are topological spaces of endogenous and exogenous variables respectively, both equipped with discrete topology $\mathcal{T}_{\mathbb{U}}$ and $\mathcal{T}_{\mathbb{V}}$, $\mathbb{F}$ is a set of functions $f_i(\cdot)$ corresponding to each $v_i \in \mathbb{V}$, where $f_i : \mathcal{T}_{\mathbb{V}} \times \mathcal{T}_{\mathbb{U}} \to \mathbb{V}, (\mathbb{v}_i^{pa}, \mathbb{U}_i) \mapsto v_i$ for some $\mathbb{v}_i^{pa} \subseteq \mathbb{V}$ and $\mathbb{U}_i \subseteq \mathbb{U}$. Each SCM $\mathcal{M}$ is associated to a causal graph where the direct causes of v_i correspond to its parent set in the causal graph. Let f be the observational density over $\mathcal{M}$ and f can be factorized as $f(\mathbb{V}|\mathbb{U}) = \prod_{v_i \in V} f_i(v_i|\mathbb{v}_i^{pa}, \mathbb{U}_i)$.*

Specifically, SCM-based methods typically rely on partially known causal graphs to identify adjustment sets or intervention paths but ultimately reduce fairness evaluation to expectations or average effects. As a result, they too are susceptible to fairness illusions—especially under high-dimensional or sparse-data regimes—because they lack mechanisms to compare full distributions of counterfactual outcomes across sensitive attributes.

This limitation underscores a key insight: *accurate fairness assessment requires moving beyond average effects to measuring distributional closeness under intervention*. Our proposed framework, DCFT, addresses this need by shifting the focus from scalar counterfactual estimates to distribution-

level evaluation, thereby enabling the identification of fairness illusions that traditional SCM- or POF-based methods often overlook.

A.3 Our Perspective: From Average Effects to Distributional Closeness

In light of the above limitations, we argue for a shift from average-based counterfactual fairness to **distribution-based fairness**. We propose to model fairness as a hypothesis testing problem over the *distributional closeness* between factual and counterfactual outcomes induced by interventions on sensitive attributes. This perspective aligns with the natural goal of ensuring not just equality in means, but similarity in opportunities and risks across full outcome distributions.

To that end, we introduce a novel framework, Counterfactual Fairness Closeness Testing (DCFT), which quantifies fairness via a new statistic, the Distributional Counterfactual Treatment Effect (DC-TE), computed using kernel-based distributional distances under intervention. This approach enables interpretable and tunable fairness analysis.

B Detailed proofs and the discussions of theoretical results

B.1 The Explanation of Introduction

Lemma 1. *Let $\mathcal{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\} \subseteq \mathbb{R}^d$ be a metric space, and let $\mathbb{P}_n$,$\mathbb{Q}_n$ be two Borel probability measures defined on $\mathcal{Z}$. Then $\mathbb{P}_n = \mathbb{Q}_n$ if and only if $E_z(f(z) = E_{z'}(f(z'))$ for all $f \in C(\mathcal{Z})$, where $C(\mathcal{Z})$ is the space of bounded continuous functions on $\mathcal{Z}$ Gretton et al. (2012).*

B.2 Maximum Mean Discrepancy

The MMD is a typical kernel-based distance between two distributions Gretton et al. (2012). Let $\mathbb{P}$ and $\mathbb{Q}$ represent two probability measures based on an instance space $\mathcal{X} \subseteq \mathbb{R}^d$. Given $u, v \in \mathcal{X}$, Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the PDS kernel with respect to the corresponding reproducing kernel Hilbert space $(\mathcal{H}_\kappa, \langle\cdot,\cdot\rangle_{\mathcal{H}_\kappa})$, where $\kappa(\cdot, u), \kappa(\cdot, v) \in \mathcal{H}_\kappa$ and $\langle\kappa(\cdot, u), \kappa(\cdot, v)\rangle = \kappa(u, v)$. We assume that there exists $K > 0$, such that $0 \leq \kappa(u, v) \leq K$ for any $u, v \in \mathcal{X}$. The kernel mean embedding of $\mathbb{P}$ and $\mathbb{Q}$ can be defined as:

$$\mu_{\mathbb{P}} = E_{u\sim\mathbb{P}}[\kappa(\cdot, u)], \mu_{\mathbb{Q}} = E_{v\sim\mathbb{Q}}[\kappa(\cdot, v)]. \tag{17}$$

Then the MMD of $\mathbb{Q}$ and $\mathbb{O}$ w.r.t. κ is:

$$\begin{aligned} &\text{MMD}^2(\mathbb{P}, \mathbb{Q}, \kappa) \\ &= ||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||^2_{\mathcal{H}_\kappa} \\ &= E_{u\sim\mathbb{P},v\sim\mathbb{Q}}\langle\kappa(\cdot, u) - \kappa(\cdot, v), \kappa(\cdot, \boldsymbol{u}) - \kappa(\cdot, \boldsymbol{v})\rangle_{\mathcal{H}_\kappa} \\ &= E_{u\sim\mathbb{P},v\sim\mathbb{Q}}[\kappa(u, \boldsymbol{u}) + \kappa(v, \boldsymbol{v}) - 2\kappa(u, v)] \\ &\in [0, 2K]. \end{aligned} \tag{18}$$

where $\boldsymbol{u}$ (resp. $\boldsymbol{v}$) is an i.i.d. copy of $u \sim \mathbb{Q}$ (resp. $v \sim \mathbb{O}$).

B.3 U-Statistic

U-statistic Serfling (2009); Vershynin (2018) is a type of statistic widely used to estimate parameters, especially in nonparametric statistics. It was proposed by probability theorist and statistician Wassily Hoeffding. U-statistic is often used to construct unbiased estimates of certain parameters, especially when the samples are independent and identically distributed.

Definition 8 (U-Statistic)**.** *Given random sample $x_1, x_2, ..., x_n$ from some distribution, let $h(x_1, x_2, ..., x_m)$ represents a symmetric kernel function where $m \leq n$. The U-statistic is defined as:*

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1\leq i_1\leq i_2 \ldots \leq i_n\leq n} h(x_{i_1}, x_{i_2}, ..., x_{i_n}), \tag{19}$$

where $\binom{n}{m}$ denotes the number of ways to choose m distinct indices from n samples. In our setting, $\binom{n}{m}$ is the number of ways to choose m distinct indices from n.

In practical applications, a typical example is the estimation of the sample mean. The sample mean can be written as a U-statistic:

$$U_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{20}$$

here symmetric kernel function $h(x) = x$ and $m = 1$.

Then based on the sub-Gaussian property of bounded functions (please refer to Vershynin (2018) for more details about concentration inequalities), we present the large deviation for U-statistic as follows:

Theorem 4 (Large Deviation for U-statistic)**.** *If the function h is bounded, i.e. $a \leq h(x_1, x_2, ..., x_m) \leq b$, we have*

$$\Pr(|U_n - \eta| \geq t) \leq 2\exp\left(-2\lfloor n/m \rfloor t^2/(b-a)^2\right), \tag{21}$$

where $\eta = E\left[h(x_{i_1}, x_{i_2}, ..., x_{i_n})\right]$.

B.4 Causal Assumptions for Counterfactual Closeness Fairness

The POF framework relies on a set of reasonable and widely accepted assumptions Rubin (1980); Pearl (2009); Yao et al. (2021). As causality research continues to expand across diverse domains, these assumptions have been adapted for modern application scenarios Xie et al. (2020); Wu et al. (2023); Li et al. (2024); Wu et al. (2024); Zhang et al. (2024). We propose the following assumptions tailored to counterfactual closeness fairness:

Assumption 1 (Counterfactual Consistency)**.** *For any intervention $do(a_t \to a_t')$, the counterfactual outcome $\widehat{Y}_{a_t'}$ satisfies:*

$$\widehat{Y}_{a_t'} \sim \Pr\left(y \mid (A_s \backslash \{a_t\}), a_t', A_c\right),$$

i.e., the distribution under intervention $do(a_t \to a_t')$ matches the conditional distribution given a_t' in the observational data. Formally:

$$\widehat{Y}_{a_t'} = f_Y\big(A_s \backslash \{a_t\},\, a_t',\, A_c\big) \quad \textit{under } do(a_t \to a_t'). \tag{22}$$

***Role in counterfactual closeness fairness**: Counterfactual Consistency is inspired by the well-known Stable Unit Treatment Value Assumption (SUTVA) Wu et al. (2023); Zhang et al. (2024), guarantees that counterfactual outcomes $\widehat{Y}_{a_t'}$ can be estimated from observational data.*

Assumption 2 (Exogeneity of Intervention)**.** *The intervention $do(a_t \to a_t')$ is independent of sensitive attributes given covariates:*

$$do(a_t \to a_t') \perp\!\!\!\perp a_t \mid A_c, A_s \backslash \{a_t\}.$$

***Role in counterfactual closeness fairness**: Exogeneity of Intervention is assumed to remove hidden confounding between the intervened a_t and the intervention $do(a_t \to a_t')$, once the A_c, A_s is fixed, thus enabling identification of causal quantities from observational y Zuo et al. (2024).*

Assumption 3 (Overlap of Counterfactual Support (OCS))**.** *(Non-Degenerate Intervention Space) Every counterfactual intervention $do(a_t \to a_t')$ has measurable realizability across the data domain:*

$$0 < \Pr\left(A_s \backslash \{a_t\},\, A_c \mid do(a_t \to a_t')\right) < 1 \quad \forall a_t' \in \mathcal{A}_t, \tag{23}$$

where $\mathcal{A}_t$ represents the set of reasonable intervention.

***Role in counterfactual closeness fairness**: OCS extends the fundamental overlap assumption Li et al. (2024); Zeng et al. (2024) to the counterfactual closeness fairness. It ensures that each $\widehat{Y}_{a_t'}$ lies within the support of the $f_Y(\cdot)$ generative space, allowing for valid counterfactual comparisons under 3. This prevents extrapolation errors in fairness testing, particularly crucial for high-dimensional deep learning models prone to out-of-support predictions.*

The proof and discussion related to the assumptions will be further elaborated in the Appendix B.5.

B.5 IDENTIFIABILITY PROOF OF COUNTERFACTUAL CLOSENESS FAIRNESS

Theorem 5 (Identifiability of DC-TE)**.** *Under Assumptions 1-3, the DC-TE measure in Definition 4 is identifiable from observational data through the empirical estimator* $\overline{\text{DC-TE}}$*. Specifically:*

$$\lim_{m\to\infty} \overline{\text{DC-TE}} = \text{DC-TE}(\mathbb{P}_n, \mathbb{P}_{a'_t,n}) \quad a.s. \tag{24}$$

Proof. The proof establishes identifiability through three pillars of the causal framework:

Pillar 1: Structural anchoring (Assumption 1).
The counterfactual outcomes satisfy:

$$\mu(\mathbb{P}_{a'_t,n}) = \mathbb{E}[\kappa(\cdot, f_Y(A_s\backslash\{a_t\}, a'_t, A_c))]. \tag{25}$$

This anchors the kernel mean embedding to the *observable* functional form of f_Y, enabling estimation via Algorithm 1 Step 1.

Pillar 2: Bias elimination (Assumption 2).
The independence $do(a_t \to a'_t) \perp\!\!\!\perp a_t | A_c, A_s\backslash\{a_t\}$ implies:

$$\mathbb{E}[\widehat{Y}_{a'_t}|A_c, A_s\backslash\{a_t\}] = \mathbb{E}[\widehat{Y}_{a'_t}|do(a_t \to a'_t), A_c, A_s\backslash\{a_t\}]. \tag{26}$$

This equivalence permits unbiased estimation of DC-TE's numerator in Definition 4 through resampling in Algorithm 1 Step 2.

Pillar 3: Variance control (Assumption 3).
The overlap condition guarantees the denominator in $\overline{\text{DC-TE}}$ satisfies:

$\text{Var}\Big(\frac{1}{m^2}\sum_{i\neq j}[4K - \kappa(\hat{y}_i, \hat{y}_j) - \kappa(\hat{y}_{a'_t,i}, \hat{y}_{a'_t,j})]\Big) = O\left(\frac{1}{m}\right)$.(27) Combined with Theorem 1, this ensures that the bootstrap procedure in Algorithm 1 Step 3 achieves asymptotic normality. □

Theorem 6 (Robustness of DCFT)**.** *Violations of causal assumptions induce specific failures in DC-TE testing:*

1. **Consistency violation**:
   If $\exists$ latent U with $\widehat{Y}_{a'_t} = f_Y(A_s\backslash\{a_t\}, a'_t, A_c, U)$, then:

   $$\left|\overline{\text{DC-TE}} - \text{DC-TE}\right| \geq \frac{\|\mu(U)\|_{\mathcal{H}_\kappa}}{4K - \mathbb{E}[\kappa(\widehat{Y}, \widehat{Y}')]}. \tag{28}$$

   This biases both Type I error control in Theorem 1 and testing threshold τ_α.
2. **Exogeneity violation**:
   For confounder C affecting a_t and $\widehat{Y}_{a'_t}$:

   $$\text{DC-TE} = \underbrace{\frac{\mathbb{E}[H]}{D}}_{\text{True effect}} + \underbrace{\frac{\text{Cov}(\kappa(\widehat{Y}, C), \kappa(\widehat{Y}_{a'_t}, C))}{D}}_{\Gamma(C)}. \tag{29}$$

   where $D = 4K - \mathbb{E}[\kappa(\widehat{Y}, \widehat{Y}') + \kappa(\widehat{Y}_{a'_t}, \widehat{Y}'_{a'_t})]$. The bias term $\Gamma(C)$ inflates Type II error.
3. **Overlap violation**:
   When $\text{supp}(\widehat{Y}_{a'_t}) \not\subseteq \text{supp}(\widehat{Y})$, the bootstrap variance:

   $$\text{Var}(\overline{\text{DC-TE}}_\phi) \geq \frac{C}{m\epsilon^2_{\min}} \to \infty, \quad \epsilon_{\min} \to 0 \tag{30}$$

   This invalidates the asymptotic normality in Theorem 1, rendering τ_α unreliable.

Proof. **(i)** Follows from RKHS norm properties under distribution shift Muandet et al. (2017). **(ii)** Derives from kernel covariance decomposition. **(iii)** Uses importance weighting variance bounds Imbens & Rubin (2015). □

B.6 Special Case: Counterfactual Two-sample Testing

Due to the characteristics of strong sensitivity, the model is considered fair iff $\mathbb{P}_n = \mathbb{P}_{a'_t,n}$. Therefore, DCFT is equivalent to two sample testing in this setting. Now we give the definition of the corresponding MMD version as a special case.

Definition 9 (MMD Treatment Effect (M-TE)). *Given the tested deep model $f_Y(\cdot)$ and sensitive attribute $a_t \in A_s$, suppose $\widehat{Y}$ represents factual potential outcomes distribution and $\widehat{Y}_{a'_t}$ represents the counterfactual one. The* M-TE *of $do(a_t \to a'_t)$ is*

$$\begin{aligned} &\textit{M-TE}(f_Y(\cdot), do(a_t \to a'_t)) = ||\mu_{\widehat{Y}} - \mu_{\widehat{Y}_{a'_t}}||^2_{\mathcal{H}_\kappa} \\ &= E_{\hat{Y},\hat{Y}_{a'_t}}[\kappa(\hat{y},\hat{y}') + \kappa(\hat{y}_{a'_t},\hat{y}'_{a'_t}) - 2\kappa(\hat{y},\hat{y}_{a'_t})], \end{aligned} \tag{31}$$

and it is also clear that M-TE $\in [0, 2K]$. Here, the value of M-TE approaches 0 when the $f_Y(\cdot)$ is not biased toward a_t.

We further introduce the empirical estimator of M-TE as :

$$\begin{aligned} &\overline{\text{M-TE}}(f_Y(\cdot), do(a_t \to a'_t)) \\ =&\sum_{i\neq j} \kappa(\hat{y}_i,\hat{y}_j) + \kappa(\hat{y}_{a'_t,i},\hat{y}_{a'_t,j}) - \kappa(\hat{y}_i,\hat{y}_{a'_t,j}) - \kappa(\hat{y}_{a'_t,i},\hat{y}_j). \end{aligned} \tag{32}$$

Together with its bootstrap version:

$$\begin{aligned} &\overline{\text{M-TE}}_\phi(f_Y(\cdot), do(a_t \to a'_t)) \\ =&\sum_{i\neq j} \phi_{ij}\left(\kappa(\hat{y}_i,\hat{y}_j) + \kappa(\hat{y}_{a'_t,i},\hat{y}_{a'_t,j}) - \kappa(\hat{y}_i,\hat{y}_{a'_t,j}) - \kappa(\hat{y}_{a'_t,i},\hat{y}_j)\right), \end{aligned} \tag{33}$$

B.7 Asymptotic Unbiasedness of the Empirical DC-TE Estimator

For the empirical estimator, we establish its asymptotic unbiasedness in Lemma 2.

Lemma 2. *Given samples $\bar{Y}, \bar{Y}_{a'_t}$ both with size m, we have* $\left|E\left[\sigma_{\bar{Y},\bar{Y}_{a'_t}}\right] - \sigma_{\mathbb{P}_n,\mathbb{P}_{a'_t,n}}\right| = O(1/\sqrt{m})$.

Definition 10 (V-statistic). *Given random samples $x_1, x_2, ..., x_n$ from some distribution, let $h(x_1, x_2, ..., x_m)$ represent a symmetric kernel function where $m \leq n$. The V-statistic is defined as:*

$$V_n = \frac{1}{n^m} \sum_{1\leq i_1,i_2,...,i_m\leq n} h(x_{i_1}, x_{i_2}, ..., x_{i_m}), \tag{34}$$

where the summation is taken over all m-tuples $(i_1, i_2, ..., i_m)$ with indices ranging from 1 to n (allowing repeated indices). A classic example is the estimation of the second moment. The sample second moment can be written as a V-statistic:

$$V_n = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j, \tag{35}$$

where the symmetric kernel function is $h(x_i, x_j) = x_i x_j$ with $m = 2$.

Proof. Given the empirical estimator $\sigma_{\bar{Y},\bar{Y}_{a'_t}}$ defined as:

$$\sigma_{\bar{Y},\bar{Y}_{a'_t}} = \frac{\sqrt{(4m-8)\zeta_1 + 2\zeta_2}/(m-1)}{(m^2-m)^{-1}\sum_{i\neq j}\left[4K - \kappa(\hat{y}_i,\hat{y}_j) - \kappa(\hat{y}_{a'_t,i},\hat{y}_{a'_t,j})\right]}.$$

We analyze the expectation by decomposing it into numerator and denominator components. Let:

$$\begin{aligned} A &= \sqrt{(4m-8)\zeta_1 + 2\zeta_2}/(m-1) \\ B &= \left(m^2-m\right)^{-1} \sum_{i\neq j}\left[4K - \kappa(\hat{y}_i,\hat{y}_j) - \kappa(\hat{y}_{a'_t,i},\hat{y}_{a'_t,j})\right]. \end{aligned}$$

Using the Delta method for ratio estimators, we expand:

$$E\left[\frac{A}{B}\right] \approx \frac{E[A]}{E[B]} - \frac{\mathrm{Cov}(A,B)}{(E[B])^2} + \frac{E[A]\mathrm{Var}(B)}{(E[B])^3}.$$

For the numerator term A:

- ζ_1 and ζ_2 are U-statistics of order 4 with $E[\zeta_1] = O(1)$ and $\mathrm{Var}(\zeta_1) = O(1/m)$.
- Applying central limit theorem: $A - E[A] = O_p(1/\sqrt{m})$.

For the denominator term B:

- The kernel mean embedding term converges at rate $O_p(1/\sqrt{m})$.
- B is a V-statistic (Definition 10) with $E[B] = 4K - \|\mu_{\mathbb{P}_n}\|^2 - \|\mu_{\mathbb{P}_{a'_t,n}}\|^2 + O(1/m)$.

The covariance term satisfies:

$$\mathrm{Cov}(A,B) \le \sqrt{\mathrm{Var}(A)\mathrm{Var}(B)} = O\left(\frac{1}{m}\right).$$

Combining these results:

$$\begin{aligned}
&\left| E\left[\sigma^2_{\bar{Y},\bar{Y}_{a'_t}}\right] - \sigma^2_{\mathbb{P}_n,\mathbb{P}_{a'_t,n}} \right| \\
\le &\left| \frac{E[A]}{E[B]} - \frac{\sigma^2_{\mathbb{P}_n,\mathbb{P}_{a'_t,n}}}{1} \right| + O\left(\frac{1}{m}\right) \\
= &\left| \frac{\sigma_{\mathbb{P}_n,\mathbb{P}_{a'_t,n}} + O(1/\sqrt{m})}{1 + O(1/\sqrt{m})} - \sigma_{\mathbb{P}_n,\mathbb{P}_{a'_t,n}} \right| + O\left(\frac{1}{m}\right) \\
= &O\left(\frac{1}{\sqrt{m}}\right).
\end{aligned}$$

The final equality follows from first-order Taylor expansion of the ratio estimator, where the leading term of the bias comes from the $O(1/\sqrt{m})$ terms in both numerator and denominator. This completes the proof of the lemma. □

B.8 Strict Sensitivity of DC-TE

In this section, we prove the strict sensitivity of counterfactual two-sample testing:

Lemma 3. *We have* $\text{DC-TE}(f_Y(\cdot), do(a_t \to a'_t)) = 0$ *if and only if* $\mathbb{P}_n = \mathbb{P}_{a'_t,n}$ *for a characteristic kernel* κ.

The corresponding hypotheses are:

$$\begin{aligned}
H_0 &: \text{DC-TE}\left(f_Y(\cdot), do(a_t \to a'_t)\right) = 0, \\
H_1 &: \text{DC-TE}\left(f_Y(\cdot), do(a_t \to a'_t)\right) \neq 0.
\end{aligned} \tag{36}$$

Under this configuration, DCFT can be directly instantiated using MMD.

Then, we introduce a well-known theorem as follows.

Theorem 7. *Gretton et al. (2012) Denote by* $\mathbb{P}$ *and* $\mathbb{Q}$ *two Borel probability measures over space* $X \subseteq \mathbb{R}^d$. *Let* $\kappa : X \times X \to \mathbb{R}$ *be a characteristic kernel. Then* $\mathrm{MMD}^2(\mathbb{P},\mathbb{Q},\kappa) = 0$ *if and only if* $\mathbb{P} = \mathbb{Q}$.

We now present the proof of Lemma 3 as follows, by taking $\mathbb{P} = \mathbb{P}_n$ and $\mathbb{Q} = \mathbb{P}_{a'_t,n}$ in Theorem 7.

Proof. Suppose that random variables $\hat{Y} \sim \mathbb{P}_n$ and $\hat{Y}_{a_t'} \sim \mathbb{P}_{a_t',n}$, then we have

$$\begin{aligned}
&\text{DC-TE}(f_Y(\cdot), do(a_t \to a_t')) \\
=&\frac{E_{\hat{Y},\hat{Y}_{a_t'}}[\kappa(\hat{y},\hat{y}') + \kappa(\hat{y}_{a_t'},\hat{y}'_{a_t'}) - 2\kappa(\hat{y},\hat{y}_{a_t'})]}{4K - E_{\hat{Y},\hat{Y}_{a_t'}}[\kappa(\hat{y},\hat{y}') + \kappa(\hat{y}_{a_t'},\hat{y}'_{a_t'})]} \\
=&\frac{||\mu_{\hat{Y}} - \mu_{Y_{a_t'}}||^2_{\mathcal{H}_\kappa}}{4K - ||\mu_{\hat{Y}}||^2_{\mathcal{H}_\kappa} - ||\mu_{\hat{Y_{a_t'}}}||^2_{\mathcal{H}_\kappa}} \\
=&\frac{\text{M-TE}(f_Y(\cdot), do(a_t \to a_t'))}{4K - ||\mu_{\hat{Y}}||^2_{\mathcal{H}_\kappa} - ||\mu_{\hat{Y_{a_t'}}}||^2_{\mathcal{H}_\kappa}} \\
=&\frac{\text{MMD}^2(\mathbb{P}_n, \mathbb{P}_{a_t',n}, \kappa)}{4K - ||\mu_{\hat{Y}}||^2_{\mathcal{H}_\kappa} - ||\mu_{\hat{Y_{a_t'}}}||^2_{\mathcal{H}_\kappa}}.
\end{aligned} \tag{37}$$

It is evident that $4K - ||\mu_{\hat{Y}}||^2_{H_\kappa} - ||\mu_{\hat{Y_{a_t'}}}||^2_{H_\kappa} > 0$. Consequently, DC-TE$(f_Y(\cdot), do(a_t \to a_t')) = 0$ if and only if $\mathbb{P}_n = \mathbb{P}_{a_t',n}$ for characteristic kernels. This completes the proof. □

B.9 Detailed Proof of Theorem 1

We begin with the empirical estimator of M-TE as

$$\begin{aligned}
&\overline{\text{M-TE}}(f_Y(\cdot), do(a_t \to a_t')) \\
=&\overline{\text{MMD}}^2(\bar{Y}, \bar{Y}_{a_t'}, \kappa) \\
=&\frac{1}{m(m-1)} \sum_{i \neq j} \kappa(\hat{y}_i, \hat{y}_j) + \kappa(\hat{y}_{a_t',i}, \hat{y}_{a_t',j}) - \kappa(\hat{y}_i, \hat{y}_{a_t',j}) - \kappa(\hat{y}_{a_t',i}, \hat{y}_j).
\end{aligned} \tag{38}$$

Given this, we introduce a well-known theorem as follows.

Theorem 8. *Under null hypothesis H_0': $\mathbb{P} = \mathbb{Q}$, let $Z_i \sim \mathcal{N}(0, 2)$ and we have:*

$$m\overline{\text{MMD}}^2(\bar{Y}, \bar{Y}_{a_t'}, \kappa) \xrightarrow{d} \textstyle\sum_i \lambda_i (Z_i^2 - 2);$$

here λ_i are the eigenvalues of the $\mathbb{P}$-covariance operator of the centered kernel [Gretton et al. (2012), Theorem 12]. On the other hand, under the alternative H_1' : $\mathbb{P} \neq \mathbb{Q}$, a standard central limit theorem holds [Serfling (2009), Section 5.5.1]

$$\begin{aligned}
\left(\overline{\text{MMD}}^2(\bar{Y}, \bar{Y}_{a_t'}, \kappa) - \text{M-TE}(f_Y(\cdot), do(a_t \to a_t'))\right) &\xrightarrow{d} \mathcal{N}(0, \sigma_M^2), \\
\sigma_M^2 := 4E[H_{1,2}H_{1,3}] - 4(E[H_{1,2}])^2,&
\end{aligned}$$

where $H_{i,j} = \kappa(\hat{y}_i, \hat{y}_j) + \kappa(\hat{y}_{a_t',i}, \hat{y}_{a_t',j}) - \kappa(\hat{y}_i, \hat{y}_{a_t',j}) - \kappa(\hat{y}_{a_t',i}, \hat{y}_j)$ and the expectation are taken with respect to $\hat{y_1}, \hat{y_2}, \hat{y_3} \sim \mathbb{P}^3$ and $\hat{y_{a_t',1}}, \hat{y_{a_t',2}}, \hat{y_{a_t',3}} \sim \mathbb{Q}^3$.

We now present the proofs of Theorem 1 as follows.

Proof. Recall the empirical estimator of our DC-TE:

$$\begin{aligned}
&m\overline{\text{DC-TE}}(f_Y(\cdot), do(a_t \to a_t')) \\
=&\frac{m\overline{\text{M-TE}}^2(f_Y(\cdot), do(a_t \to a_t'))}{1/(m^2-m)\sum_{i \neq j}[4K - \kappa(\hat{y}_i, \hat{y}_j) - \kappa(\hat{y}_{a_t',i}, \hat{y}_{a_t',j})]} \\
=&\frac{m\overline{\text{MMD}}^2(\bar{Y}, \bar{Y}_{a_t'}, \kappa)}{1/(m^2-m)\sum_{i \neq j}[4K - \kappa(\hat{y}_i, \hat{y}_j) - \kappa(\hat{y}_{a_t',i}, \hat{y}_{a_t',j})]}
\end{aligned} \tag{39}$$

As a U-statistic, it is easy to see that

$$\begin{aligned} &1/(m^2-m)\sum_{i\neq j}[4K-\kappa(\hat{y}_i,\hat{y}_j)-\kappa(\hat{y}_{a_t',i},\hat{y}_{a_t',j})] \\ &\xrightarrow{d} 4K-||\mu_{\widehat{Y}}||^2_{\mathcal{H}_\kappa}-||\mu_{\widehat{Y}_{a_t'}}||^2_{\mathcal{H}_\kappa}, \end{aligned} \tag{40}$$

where $\xrightarrow{d}$ denotes convergence in probability.

If $\overline{\text{DC-TE}}(f_Y(\cdot), do(a_t \to a_t')) = 0$, we have $\mathbb{P}_n = \mathbb{P}_{a_t',n}$ from lemma 3, and

$$m\overline{\text{MMD}}^2(\bar{Y}, \bar{Y}_{a_t'}, \kappa) \xrightarrow{d} \sum_i \lambda_i (Z_i^2 - 2)$$

from Theorem 8. Then, by slutsky's theorem Edition et al. (2002), we have

$$\begin{aligned} m\overline{\text{DC-TE}}(f_Y(\cdot), do(a_t \to a_t')) &\xrightarrow{d} \frac{\sum_i \lambda_i (Z_i^2 - 2)}{4K - ||\mu_{\widehat{Y}}||^2_{\mathcal{H}_\kappa} - ||\mu_{\widehat{Y}_{a_t'}}||^2_{\mathcal{H}_\kappa}} \\ &\xrightarrow{d} \frac{\sum_i \lambda_i (Z_i^2 - 2)}{4K - ||(\mu_{\widehat{Y}} + \mu_{\widehat{Y}_{a_t'}})/\sqrt{2}||^2_{\mathcal{H}_\kappa}}, \end{aligned} \tag{41}$$

where $\mu_{\widehat{Y}} = \mu_{\widehat{Y}_{a_t'}} = (\mu_{\widehat{Y}} + \mu_{\widehat{Y}_{a_t'}})/2$.

If DC-TE$(f_Y(\cdot), do(a_t \to a_t')) = \epsilon$ with$\epsilon \in (0, 1)$, we present the asymptotic distribution of the empirical estimator in a similar manner, which can be formalized as

$$\begin{aligned} &\sqrt{m}(\overline{\text{DC-TE}}(f_Y(\cdot), do(a_t \to a_t')) - \epsilon) \xrightarrow{d} \\ &N\left(0, \frac{4E[H_{1,2}H_{1,3}] - 4(E[H_{1,2}])^2}{(4K - ||\mu_{\widehat{Y}}||^2_{H_\kappa} - ||\mu_{\widehat{Y}_{a_t'}}||^2_{H_\kappa})^2}\right). \end{aligned} \tag{42}$$

□

B.10 Theoretical Reliability of DCFT for Diagnosing Causal Effects

Average Treatment Effect (ATE)–level criteria certify fairness when group averages align, but they do not control the *entire* counterfactual outcome law. To make this precise, let $P_a := P(\widehat{Y} \mid do(A \to a))$ for $a \in \{0, 1\}$ denote the interventional distributions of the predictive output $\widehat{Y}$. We compare P_0 and P_1 at the *distributional* level using kernel two-sample geometry.

Kernel geometry. Let k be a bounded *characteristic* kernel with RKHS $\mathcal{H}_k$ and bound $0 \leq k(\cdot, \cdot) \leq K$. Define the squared MMD

$$\text{MMD}_k^2(P_0, P_1) := \mathbb{E}_{X,X' \sim P_0} k(X, X') + \mathbb{E}_{Y,Y' \sim P_1} k(Y, Y') - 2\,\mathbb{E}_{X \sim P_0,\, Y \sim P_1} k(X, Y).$$

The DC-TE in Def. 4 can be written as a bounded, monotone transform of MMD_k^2:

$$\text{DC-TE}(P_0, P_1) = \frac{\text{MMD}_k^2(P_0, P_1)}{4K - \mathbb{E}_{X,X' \sim P_0} k(X, X') - \mathbb{E}_{Y,Y' \sim P_1} k(Y, Y')}.$$

Because k is bounded, the denominator lies in $(0, 4K]$ whenever P_0 and P_1 have support in the domain of k.

Assumption 4 (Identifiability)**.** *The interventional laws P_a are identifiable from observed data under a valid causal strategy (e.g., back-door with overlap, valid instruments, or a validated simulator).*

Assumption 5 (Regularity)**.** *k is bounded and characteristic on the support of $\widehat{Y}$; $\widehat{Y} \in L^1(P_a)$ for $a \in \{0, 1\}$.*

Theorem 9 (Distributional soundness of DCFT)**.** *Under Assumptions 4–5, the following are equivalent:*

$$P_0 = P_1 \quad \Longleftrightarrow \quad \mathrm{MMD}_k^2(P_0, P_1) = 0 \quad \Longleftrightarrow \quad \text{DC-TE}(P_0, P_1) = 0.$$

Consequently, for any tolerance $\epsilon \geq 0$, the DCFT null $H_0 : \text{DC-TE}(P_0, P_1) \leq \epsilon$ rejects if and only if the interventional distributions differ by more than ϵ in RKHS distance.

Proof. Characteristic kernels metrize equality in law via injective mean embeddings: $\mu(P) = \mathbb{E}_P[k(\cdot, Z)]$ is injective iff k is characteristic, hence $\mathrm{MMD}_k^2(P_0, P_1) = \|\mu(P_0) - \mu(P_1)\|_{\mathcal{H}_k}^2 = 0 \Leftrightarrow P_0 = P_1$. The denominator of DC-TE is strictly positive by boundedness of k, so DC-TE $= 0$ iff $\mathrm{MMD}_k^2 = 0$; monotonicity follows immediately. □

Theorem 10 (Dominance over ATE-level criteria)**.** *Define $\mathcal{N}_{dist} := \{(P_0, P_1) : P_0 = P_1\}$ and $\mathcal{N}_{ATE} := \{(P_0, P_1) : \mathbb{E}_{P_0}[\widehat{Y}] = \mathbb{E}_{P_1}[\widehat{Y}]\}$. Under Assumption 5, $\mathcal{N}_{dist} \subsetneq \mathcal{N}_{ATE}$. Hence there exist interventional pairs with matched means but distinct laws for which $\text{DC-TE} > 0$ (and DCFT can reject for ϵ below this effect), whereas ATE-based parity cannot detect the discrepancy.*

Proof. If $P_0 = P_1$, then expectations match, so $\mathcal{N}_{\text{dist}} \subseteq \mathcal{N}_{\text{ATE}}$. Strict inclusion follows from standard counterexamples (e.g., two different zero-mean distributions with identical variance or matched first two moments but different tails/modes). By Theorem 9, distinct laws imply $\mathrm{MMD}_k^2 > 0$ and thus DC-TE > 0. □

Theorem 11 (Sensitivity to direct and mediated effects; immunity to "fairness through blindness")**.** *Consider any SCM in which A may affect $\widehat{Y}$ via direct and/or mediated paths $A \to \cdots \to \widehat{Y}$. If there exists a setting change $do(A{=}1)$ vs. $do(A{=}0)$ that alters $P(\widehat{Y} \mid do(A))$, then $P_1 \neq P_0$ and $\text{DC-TE}(P_0, P_1) > 0$. Therefore, DCFT can detect residual influence even when A is excluded from model inputs ("fairness through blindness"), because mediated effects persist in the interventional law.*

Proof. Under SCM semantics, $do(A{=}a)$ overrides the assignment of A and propagates along all outgoing edges, including mediators. If any such path changes the distribution of $\widehat{Y}$, then $P_1 \neq P_0$. By Theorem 9, $\mathrm{MMD}_k^2(P_0, P_1) > 0$ and thus DC-TE > 0. Removing A from the feature set does not alter the interventional distributions $P(\widehat{Y} \mid do(A))$, hence DCFT remains sensitive to mediated influence. □

If unobserved variables act as mediators or confounders, DCFT detects their impact insofar as a chosen identification strategy yields interventional laws P_a that reflect the true effect of $do(A)$. When P_a is only partially identified, DCFT conclusions inherit those bounds. This keeps claims faithful to causal identifiability while preserving DCFT's distributional sensitivity. For $\epsilon = 0$, DCFT reduces to a distributional equality test ($P_0 = P_1$) under a characteristic kernel. For $\epsilon > 0$, DCFT implements an *application-tunable* tolerance that is robust to negligible deviations while still rejecting whenever the effect size (in RKHS distance) exceeds ϵ. Theorems 9–11 show that DCFT (i) metrizes equality of the full interventional laws, (ii) strictly dominates ATE parity, and (iii) remains sensitive to mediated pathways that survive feature removal of A. This closes the gap responsible for "fairness through blindness" and for mean-level illusions, while keeping the usual causal assumptions explicit (Assumption 4).

B.11 Fairness Sensitivity Control in DCFT via the Tunable Threshold ϵ

Fairness requirements vary across applications, with different domains tolerating different levels of outcome disparity. DCFT addresses this via a tunable threshold ϵ, which controls fairness sensitivity by bounding the discrepancy between factual and counterfactual outcome distributions in the RKHS. For interpretability, we categorize ϵ-driven sensitivity into three regimes:

Strict sensitivity. When ϵ is close to zero, even slight distributional deviations imply unfairness. This regime aligns with high-stakes settings (e.g., medical diagnosis, criminal justice) where outcome disparity must be minimized. As shown in Lemma 3, DC-TE converges to a distributional equality test under this configuration. For completeness we restate the operative consequence of Lemma 3.

Lemma 4 (Strict-sensitivity limit; cf. Lemma 3)**.** *Under a bounded characteristic kernel, setting* $\epsilon{=}0$ *makes DCFT equivalent to testing equality in law between* $P(\widehat{Y} \mid do(A{=}0))$ *and* $P(\widehat{Y} \mid do(A{=}1))$.

Moderate sensitivity. A modest ϵ accommodates domains with moderate tolerance for modeling noise (e.g., credit scoring, recommendations), offering a balanced trade-off between fairness precision and resilience to spurious discrepancies.

Lenient sensitivity. For low-stakes or exploratory use cases (e.g., LLM generation), a higher ϵ permits natural distributional variation, ensuring only systematic differences trigger fairness violations.

B.12 Sensitive Attribute Setting

The selection of sensitive attributes requires careful consideration. Drawing from previous research in fair machine learning Caton & Haas (2024); Mehrabi et al. (2021), commonly recognized sensitive attributes, such as gender and race, should be prioritized Zuo et al. (2022; 2024). Additionally, private information Ezzeldin et al. (2023); Chen et al. (2023), such as family occupation and family history, should also be considered. Given the growing emphasis on deep learning value alignment Wang et al. (2024); Shen et al. (2023), attributes linked to human social values should also be taken into account.

It is important to note that the selection of sensitive attributes should be guided by the question, "Should this attribute be considered?" rather than "Is this attribute relevant in reality?" Mehrabi et al. (2021). For instance, when predicting "grades," while there may be a correlation between "grades" and "gender" in reality, from a fairness perspective, gender should not influence the prediction of student performance. Therefore, in this context, "gender" should be treated as a sensitive attribute.

We regard sex and parent's job as the sensitive attributes in the UCI Student Performance Dataset(UCI) Cortez (2014) to predict students' performance in Mathematics. Estimation of Obesity dataset(Obesity) Palechor & de la Hoz Manotas (2019) includes data for the estimation of obesity levels, we regard gender and age as sensitive attributes. We also regard gender and parents' occupation as sensitive attributes in Students' Dropout and Academic Success dataset(Drop) Realinho et al. (2021).

C Supplements and Discussions on the Experiment

C.1 Discussions on Kernel Selection for DC-TE

Existing kernel selection strategies for discrepancy measures, such as Maximum Mean Discrepancy (MMD) Gretton et al. (2012), primarily target *two-sample testing* scenarios, aiming to maximize test power between two fixed distributions $\mathbb{P}$ and $\mathbb{Q}$. In contrast, kernel selection for the DC-TE-based *distribution closeness testing* requires fundamental adaptation. We summarize the key distinctions as follows.

C.1.1 Key Differences between MMD and DC-TE Settings

Divergent Optimization Objectives. In MMD, kernel selection typically maximizes the standardized statistic

$$T_{\text{MMD}} = \frac{\text{MMD}^2}{\sqrt{\text{Var}(\text{MMD}^2)}}. \tag{43}$$

In DC-TE, the test statistic is self-normalized:

$$\text{DC-TE} = \frac{\mathbb{E}[H]}{4K - \mathbb{E}[\kappa(\hat{y}_i, \hat{y}_j) + \kappa(\hat{y}'_i, \hat{y}'_j)]}, \tag{44}$$

where both the numerator and denominator depend on the kernel κ. Consequently, kernel optimization must balance maximizing discrepancy (numerator) and stabilizing normalization (denominator).

Scale-Invariance. The denominator term $4K - \mathbb{E}[\cdot]$ induces an automatic adjustment for kernel scale, similar to the behavior observed in NAMMD Zhou et al. (2025). As established in Theorem 1,

the asymptotic variance $\sigma^2_{\widehat{Y},\widehat{Y}_{a'_t}}$ of DC-TE is inversely proportional to the scaling factor, thereby ensuring robustness across different kernel bandwidths.

Composite Hypothesis. Unlike two-sample testing which assumes a fixed $\mathbb{P} \neq \mathbb{Q}$, DC-TE-based testing operates under a composite null H_0 : DC-TE $\leq \epsilon$, allowing for partial distributional overlap. This further complicates kernel selection as the optimal kernel must remain powerful across a range of close but non-identical distributions.

C.1.2 Implications for Kernel Optimization

We derive two central implications regarding kernel strategies under the DC-TE framework.

Dominance of DC-TE-Optimized Kernels. Let κ^*_{MMD} and $\kappa^*_{\text{DC-TE}}$ denote the kernels optimized for MMD and DC-TE respectively. Then the following holds:

Proposition 1. *For any kernel κ, the power of the DC-TE test satisfies*

$$\textit{Power}(\textit{DC-TE}_\kappa) \geq \textit{Power}(\textit{MMD}_\kappa), \tag{45}$$

and moreover,

$$\textit{Power}(\textit{DC-TE}_{\kappa^*_{\textit{DC-TE}}}) \geq \textit{Power}(\textit{DC-TE}_{\kappa^*_{\textit{MMD}}}). \tag{46}$$

Proof Ketch. DC-TE intrinsically self-normalizes, reducing variance inflation compared to MMD. As a result, for the same kernel, DC-TE achieves tighter concentration around its expectation, yielding higher statistical power. Further, direct optimization over DC-TE's signal-to-noise ratio (SNR) leads to a strictly superior kernel $\kappa^*_{\text{DC-TE}}$.

Adaptive Kernel Learning Objective. Optimally, the kernel κ should maximize the normalized SNR of DC-TE, formulated as:

$$\kappa^*_{\text{DC-TE}} = \text{argmax}_{\kappa \in \mathcal{K}} \frac{\mathbb{E}_\kappa[H]}{\sqrt{4K_\kappa - \mathbb{E}_\kappa[\kappa(\hat{y}_i, \hat{y}_j) + \kappa(\hat{y}'_i, \hat{y}'_j)]}}. \tag{47}$$

Practical implementation could leverage deep kernel learning techniques Liu et al. (2021) to parameterize κ and optimize the DC-TE test power via gradient ascent.

C.1.3 Efficiency Advantages of DC-TE

Beyond power dominance, DC-TE exhibits higher asymptotic relative efficiency (ARE) compared to MMD:

Under any fixed kernel κ, the asymptotic relative efficiency satisfies

$$\text{ARE}(\text{DC-TE} : \text{MMD}) = \frac{\sigma^{-2}_{\text{MMD}}}{\sigma^{-2}_{\text{DC-TE}}} \geq 1 + \frac{\mathbb{V}\text{ar}(\mathbb{E}[H|\kappa])}{\mathbb{E}^2[\mathbb{E}[H|\kappa]]}. \tag{48}$$

Interpretation. The inequality quantifies a strict improvement: the larger the variance of the conditional expectation $\mathbb{E}[H|\kappa]$, the greater the efficiency gain of DC-TE over MMD.

Overall, kernel selection for DC-TE should prioritize *joint optimization* of discrepancy sensitivity and normalization stability. This contrasts with the traditional approach in MMD, which focuses solely on maximizing discrepancy magnitude. Developing adaptive kernel learning methods tailored for DC-TE remains an exciting open problem.

C.2 Real-world deep learning tasks experiment results on CNN and LSTM

C.3 Credit Risk Assessment (CRA)

Statistical bias frequently arises in tasks involving the analysis of demographic attributes, where predictions are unfairly skewed toward specific sensitive attributes such as gender, age, and race Xu et al. (2020). One example is credit risk assessment (CRA), which involves predicting the likelihood

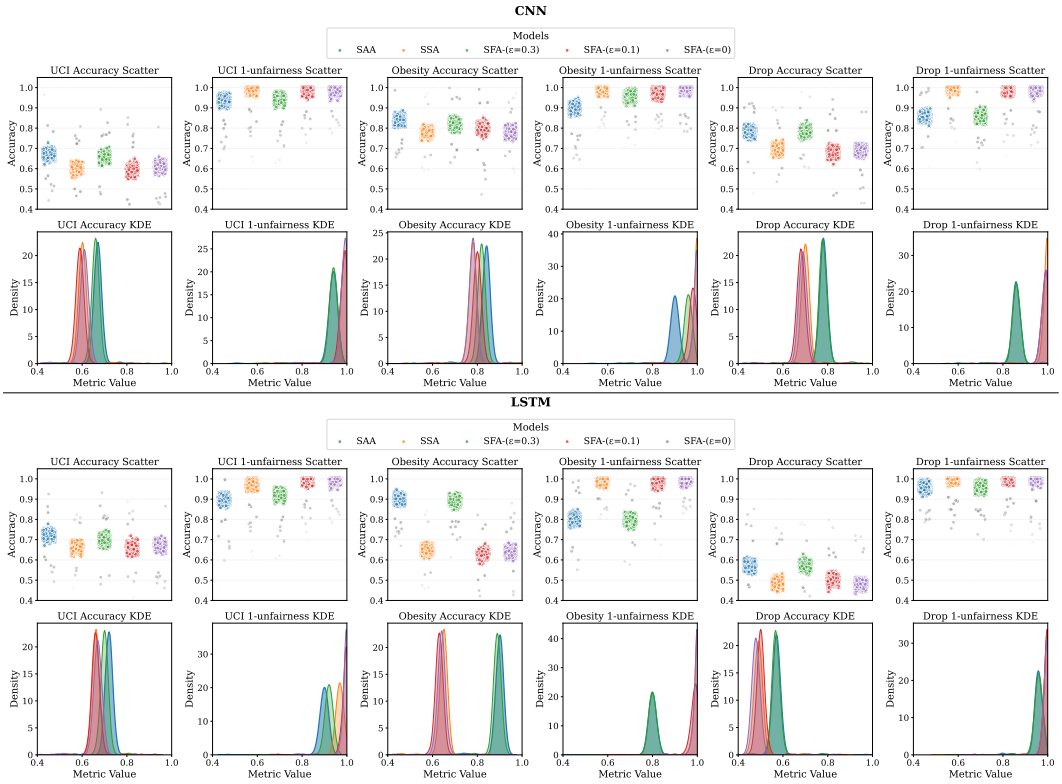

Figure 3: From top to bottom: Results on CNN models and results on LSTM models. We conducted 500 experimental runs for each scenario. The upper portion of the figure visualizes the results for SAA, SSA, SFA($\epsilon = 0$), SFA($\epsilon = 0.1$) and SFA($\epsilon = 0.3$), each represented by a distinct color in the plots. The first column shows scatter plots of performance, and the second displays the corresponding kernel density estimates (KDEs). For visualization only, scores are normalized within each run to place axes on a comparable scale; no Gaussianity is assumed and no transformation is used in our experiments. A higher degree of overlap among KDE curves indicates greater similarity in model behavior. To more clearly highlight model performance in terms of fairness, we plot the fairness metric defined as $(1 - \text{unfairness})$. Across all three datasets, the performance of SFA with $\epsilon = 0.3$ closely aligns with that of SAA, suggesting that no unfair sensitive attributes were identified. In contrast, SFA with $\epsilon = 0.1$ and $\epsilon = 0$ shows behavior consistent with SSA, indicating that unfair sensitive attributes were successfully identified.

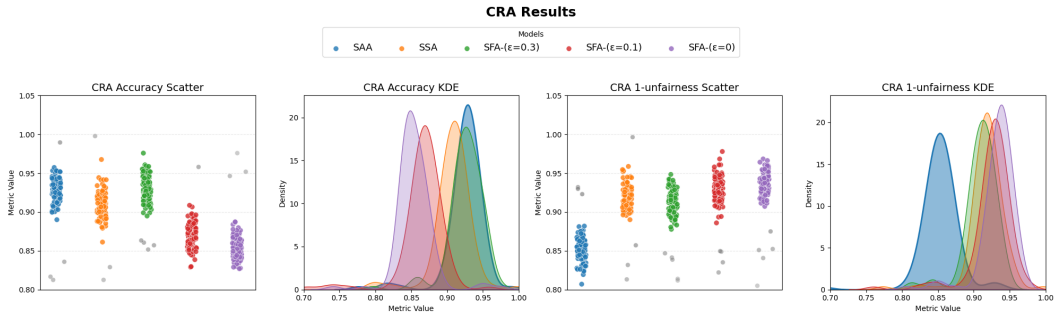

Figure 4: Results on CRA. It is important to highlight the unique insights revealed by this experiment. While conventional assumptions suggest that older individuals tend to carry higher credit risk, leading SAA to exclude older age groups (specifically those over 40 or 50), the results clearly show that SFA-ϵ=0 and SFA-ϵ=0.1 achieve stronger fairness outcomes. This counterintuitive result likely arises from the fact that, in practice, younger individuals may actually exhibit higher credit risk. Such findings underscore the superior testing capability of DCFT in identifying fairness violations that defy traditional expectations.

that a borrower will default on a loan. In such tasks, there is often a bias where relatively older groups are assessed as having a higher risk of default Shi et al. (2022).

Experiment settings: We utilize the Credit Risk Dataset, which contains 11 features related to the repayment capability of 32,581 borrowers. Like some similar studies Zuo et al. (2022; 2024), we treat specific age (*23* (first quantiles) and *30* (third quantiles)) and older age groups (which are *over 40* and *over 50*) as different sensitive attributes.

Analysis: As shown in Figure 4, the results aligned with the previous experimental conclusions, with SFA-ϵ=0 achieving the fairest performances. SFA-ϵ=0, SFA-ϵ=0.1, and SFA-ϵ=0.3 also demonstrated varying performances, where SFA-ϵ=0 was the fairest and SFA-ϵ=0.3 exhibited the highest degree of unfairness.

C.4 Optimization Analysis for DC-TE Kernel Selection

In this paper, the kernel is primarily used as a tool to implement DC-TE, and we do not focus on how its intrinsic properties affect the experiments. In this section, we provide some discussions on optimization of its kernel selections for future research and exploration.

Building upon Theorem 1, we derive an optimization framework for adaptive kernel selection in DC-TE testing. Let $\mathbb{P}_n$ and $\mathbb{P}_{a'_t,n}$ denote the factual and counterfactual distributions, respectively. Under the null hypothesis H_0 : DC-TE $\leq \epsilon$, the empirical estimator satisfies the asymptotic distribution:

$$\sqrt{m}\left(\overline{\text{DC-TE}} - \epsilon\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathbb{P}_n, \mathbb{P}_{a'_t,n}}\right), \tag{49}$$

where the asymptotic variance $\sigma^2_{\mathbb{P}_n, \mathbb{P}_{a'_t,n}}$ is given by:

$$\sigma^2_{\mathbb{P}_n, \mathbb{P}_{a'_t,n}} = \frac{4\mathbb{E}[H_{1,2}H_{1,3}] - 4\left(\mathbb{E}[H_{1,2}]\right)^2}{\left(4K - \|\mu_{\mathbb{P}_n}\|^2_{\mathcal{H}_\kappa} - \|\mu_{\mathbb{P}_{a'_t,n}}\|^2_{\mathcal{H}_\kappa}\right)^2}, \tag{50}$$

and $H_{i,j}$ is defined as:

$$H_{i,j} = \kappa(\hat{y}_i, \hat{y}_j) + \kappa(\hat{y}_{a'_t,i}, \hat{y}_{a'_t,j}) - \kappa(\hat{y}_i, \hat{y}_{a'_t,j}) - \kappa(\hat{y}_{a'_t,i}, \hat{y}_j). \tag{51}$$

Kernel Optimization Objective. To maximize the test power, we optimize the signal-to-noise ratio:

$$\mathcal{J}(\kappa) = \frac{\overline{\text{DC-TE}}(\kappa)}{\sigma_{\mathbb{P}_n, \mathbb{P}_{a'_t,n}}(\kappa)}. \tag{52}$$

This leads to a gradient-based procedure for kernel learning, summarized in Algorithm 2.

The following theorem characterizes the optimal kernel for maximizing the DC-TE test power.

Theorem 12 (Optimal Kernel Characterization)**.** *The optimal kernel κ^* is the solution to:*

$$\kappa^* = \textit{argmax}_{\kappa \in \mathcal{K}} \frac{\mathbb{E}[H]}{\sqrt{4\mathbb{E}[H_{1,2}H_{1,3}] - 4\left(\mathbb{E}[H_{1,2}]\right)^2}}, \tag{53}$$

where expectations are taken over independent samples from the joint distribution $(\bar{Y}, \bar{Y}_{a'_t})$.

Proof. Under H_0, the asymptotic power of the DC-TE test satisfies:

$$\text{Power} \approx \Phi\left(\sqrt{m}\frac{\overline{\text{DC-TE}} - \epsilon}{\sigma_{\mathbb{P}_n, \mathbb{P}_{a'_t,n}}}\right),$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Since $\Phi(\cdot)$ is monotonically increasing, maximizing the test power reduces to maximizing the signal-to-noise ratio:

$$\frac{\overline{\text{DC-TE}}}{\sigma_{\mathbb{P}_n, \mathbb{P}_{a'_t,n}}}.$$

Algorithm 2 DC-TE Kernel Selection

Require: Paired samples $\bar{Y}, \bar{Y}_{a'_t}$, initial kernel κ_0, learning rate η, number of iterations N
Ensure: Optimized kernel κ^*
Step 1: For $t = 1$ to N:
Compute empirical DC-TE:

$$\overline{\text{DC-TE}}^{(t)} \leftarrow \frac{\sum_{i \neq j} H_{i,j}^{(t)}}{\sum_{i \neq j} \left(4K - \kappa^{(t)}(\hat{y}_i, \hat{y}_j) - \kappa^{(t)}(\hat{y}_{a'_t,i}, \hat{y}_{a'_t,j})\right)}.$$

Estimate the variance:

$$\sigma^{(t)} \leftarrow \frac{\sqrt{(4m-8)\zeta_1^{(t)} + 2\zeta_2^{(t)}}/(m-1)}{(m^2-m)^{-1}\sum_{i \neq j} \left(4K - \kappa^{(t)}(\hat{y}_i, \hat{y}_j) - \kappa^{(t)}(\hat{y}_{a'_t,i}, \hat{y}_{a'_t,j})\right)}.$$

Compute gradient:

$$\nabla_\kappa \mathcal{J} \leftarrow \frac{\partial}{\partial \kappa}\left(\frac{\overline{\text{DC-TE}}^{(t)}}{\sigma^{(t)}}\right).$$

Update kernel via Adam optimizer:

$$\kappa^{(t+1)} \leftarrow \kappa^{(t)} + \eta \cdot \text{Adam}\left(\nabla_\kappa \mathcal{J}\right).$$

Step 2: Return optimized kernel κ^*.

Substituting the explicit forms of $\overline{\text{DC-TE}}$ and $\sigma_{\mathbb{P}_n, \mathbb{P}_{a'_t,n}}$, we obtain the stated optimization objective. □

Corollary 1 (Consistency guarantee)**.** *Under mild regularity conditions:*

1. *The estimator $\overline{DC\text{-}TE}$ is asymptotically normal;*
2. *There exists at least one $\kappa^* \in \mathcal{K}$ achieving the maximum of $\mathcal{J}(\kappa)$;*
3. *Algorithm 2 converges almost surely to κ^* as $N \to \infty$.*

Discussion. This framework enables adaptive kernel learning for DC-TE testing by jointly optimizing distributional discrepancy measurement and testing stability. The variance terms ζ_1 and ζ_2 generalize the MMD variance components by incorporating counterfactual coupling effects through the $H_{i,j}$ structures.

C.5 Audit Protocol for Distributional Parity vs. Tail Disparity

This section specifies the audit protocol used to substantiate the presence of distribution-level disparities under low-order parity. The protocol is designed to be rigorous, reproducible, and minimally assumption–dependent, and is suitable for de-identified prediction outputs in risk-sensitive domains (e.g., depression risk prediction and clinical risk scoring).

We assume access to de-identified model outputs consisting of tuples (S_i, Y_i, G_i, X_i), where $S_i \in \mathbb{R}$ is a predicted score or risk, $Y_i \in \{0, 1\}$ the binary outcome label (when available), $G_i \in \{g_1, g_2\}$ a binary sensitive group indicator,[1] and X_i optional non-sensitive covariates used only for matching or stratification. No personally identifying information is used or retained. All processing follows applicable privacy/IRB requirements.

We form *paired cohorts* to make low-order statistics nearly identical across groups while leaving the full score distributions unconstrained:

[1]The protocol generalizes to multi-category sensitive attributes via one-vs-rest audits.

- **Low-order targets.** For each group $G \in \{g_1, g_2\}$, let $\mu_G := \mathbb{E}[S \mid G]$, $\sigma_G^2 := \mathrm{Var}(S \mid G)$, and $\pi_G := \Pr(\widehat{Y} = 1 \mid G)$ if $\widehat{Y}$ is a binarized decision; when only scores are available, we match (μ_G, σ_G^2) and the marginal score-positive rate $\Pr(S \geq \tau \mid G)$ for a pre-specified threshold τ.
- **Tolerance bands.** Fix small tolerances $\delta_\mu, \delta_\sigma, \delta_\pi > 0$ (e.g., $|\mu_{g_1} - \mu_{g_2}| \leq \delta_\mu$, $|\sigma_{g_1}^2 - \sigma_{g_2}^2| \leq \delta_\sigma$, $|\pi_{g_1} - \pi_{g_2}| \leq \delta_\pi$).
- **Matching/reweighting.** Use one of: (i) propensity-score matching on X (logistic model for G given X) with calipers; (ii) nearest-neighbor or optimal transport (OT) reweighting to minimize discrepancies in (μ, σ^2, π); or (iii) coarsened exact matching on discrete X strata. Denote the resulting weights by $w_i \geq 0$ with $\sum_{i:G_i=g} w_i = n_g$.

We verify post-match balance: the weighted differences in (μ, σ^2, π) fall within $(\delta_\mu, \delta_\sigma, \delta_\pi)$.

Let P_G denote the weighted empirical distribution of scores $\{S_i : G_i = G\}$ after matching. We evaluate:

1. **Tail error gap (top-decile).** Let $Q_q(S \mid G)$ be the weighted q-quantile of S in group G (e.g., $q = 0.9$). Define the group-wise tail error rate
$$\mathrm{Err}_G^{(q)} \;:=\; \Pr(\widehat{Y} \neq Y \mid S \geq Q_q(S \mid G),\, G),$$
estimated by weighted proportions when labels Y are available. The *tail gap* is
$$\Delta_{\mathrm{tail}} \;:=\; \mathrm{Err}_{g_1}^{(q)} - \mathrm{Err}_{g_2}^{(q)}.$$
We compute a $(1-\alpha)$ bootstrap confidence interval (CI) by resampling within groups with replacement and recomputing Δ_{tail}.
2. **Two-sample KS test on scores.** Let F_{S_g} be the weighted empirical CDF of S in group g. Compute
$$D_{\mathrm{KS}} \;:=\; \sup_{s \in \mathbb{R}} \left|F_{S_{g_1}}(s) - F_{S_{g_2}}(s)\right|,$$
and report the KS p-value with weights handled via *multiplier bootstrap* (wild bootstrap) to respect the matched design.
3. **Groupwise calibration.** If probabilistic predictions are available, compute expected calibration error (ECE) within each group using B equal-frequency bins:
$$\mathrm{ECE}_G \;:=\; \sum_{b=1}^{B} \frac{n_{G,b}}{n_G} \left|\mathrm{acc}_{G,b} - \mathrm{conf}_{G,b}\right|,$$
where $\mathrm{acc}_{G,b}$ is the empirical outcome rate and $\mathrm{conf}_{G,b}$ the mean predicted probability in bin b. Report $\Delta_{\mathrm{ECE}} := |\mathrm{ECE}_{g_1} - \mathrm{ECE}_{g_2}|$ with bootstrap CIs.

The audit evaluates whether low-order parity coexists with distribution/tail discrepancy:

Balance checks:	$\lvert\mu_{g_1} - \mu_{g_2}\rvert \leq \delta_\mu,\ \lvert\sigma_{g_1}^2 - \sigma_{g_2}^2\rvert \leq \delta_\sigma,\ \lvert\pi_{g_1} - \pi_{g_2}\rvert \leq \delta_\pi.$
Tail gap:	$H_0 : \Delta_{\mathrm{tail}} = 0$ vs. $H_1 : \Delta_{\mathrm{tail}} \neq 0$ (bootstrap CI excludes 0).
KS test:	$H_0 : P_{g_1} = P_{g_2}$ vs. $H_1 : P_{g_1} \neq P_{g_2}$ (KS $p < \alpha$).
Calibration gap:	$H_0 : \Delta_{\mathrm{ECE}} = 0$ vs. $H_1 : \Delta_{\mathrm{ECE}} \neq 0$ (bootstrap CI excludes 0).

A *counterfactual fairness illusion* is flagged when the balance checks pass but at least one of Δ_{tail}, KS, or Δ_{ECE} indicates significant disparity.

When auditing multiple sensitive attributes, subgroups, or multiple q-levels, we control the false discovery rate (FDR) across all conducted tests using Benjamini–Hochberg at level α_{FDR}; CIs are adjusted via Benjamini–Yekutieli when dependence across tests is non-negligible.

We assess robustness along:

- **Tail definition:** vary $q \in \{0.85, 0.90, 0.95\}$.
- **Matching stringency:** tighten $(\delta_\mu, \delta_\sigma, \delta_\pi)$ and re-audit.

- **Alternative distances:** complement KS with MMD (Gaussian kernel, median heuristic) and Wasserstein-1 distance; report standardized effect sizes.
- **Stratification:** repeat the audit within clinically relevant strata (age bands, comorbidity index) to probe Simpson's paradox.

Algorithm 3 Audit-style protocol for detecting fairness illusion

Require: De-identified (S, Y, G, X); tolerances $(\delta_\mu, \delta_\sigma, \delta_\pi)$; tail level q; significance α.

1: Match or reweight across G to satisfy low-order balance within tolerances.
2: Verify balance on (μ, σ^2, π); if failed, refine the match.
3: Compute Δ_{tail} with bootstrap CI; compute KS statistic and p-value; compute Δ_{ECE} with CI (if applicable).
4: Apply FDR correction if multiple tests are run.
5: Report whether balance holds and whether any distribution/tail metrics indicate significant disparity.

D LLM USAGE

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.