A STOCHASTIC INTERPOLANTS METHOD FOR UNIFYING INDIVIDUAL AND COUNTERFACTUAL FAIRNESS IN GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

The fairness problem in graph neural networks faces a serious conflict between individual fairness and counterfactual fairness. Strict adherence to individual fairness perpetuates structural bias, whereas excessive pursuit of counterfactual fairness may overlook genuine structural differences. This paper proposes the first framework that applies stochastic interpolants to graph fairness problems. Unlike optimal transport methods based on the Wasserstein distance or the Sinkhorn algorithm, our framework precisely controls noise levels in the transport path, allowing dynamic adjustment of emphasis on the two fairness criteria at different stages. Our method overcomes the binary choice limitation in traditional fairness approaches and achieves a continuous trade-off between the two criteria. Specifically, we design a structure-attribute disentanglement representation method that decomposes node representations into bias-carrying features and unbiased structural attributes. Through dynamic noise-level adjustment during transport, we achieve gradual integration of the two criteria. Theoretical analysis proves the upper bound of Kullback-Leibler divergence between the model and the ideal fair distribution. This bound decomposes into the realization levels of individual fairness and counterfactual fairness. Experiments on multiple datasets, including Pokec, Facebook100, Credit-Default, and COMPAS, show that, compared with existing methods, our framework significantly improves both fairness metrics while maintaining model performance.

1 INTRODUCTION

Graph Neural Networks (GNNs) have demonstrated excellent performance in many fields, such as social network analysis, recommendation systems, and knowledge graphs (Hamilton et al., 2017). However, as these models are increasingly applied in high-risk scenarios, such as hiring, credit evaluation, and judicial decision-making, fairness issues have become increasingly prominent (Chen et al., 2024; Dong et al., 2023).

Graph fairness refers to the challenge of ensuring that graph-based machine learning models make equitable predictions while respecting the structural properties of graph data. Unlike traditional fairness problems in tabular data, graph fairness exhibits unique complexity because bias may not only exist in node attributes, but also be deeply embedded in the connection patterns of graph structures (Walker et al., 2004). For example, in a professional social network, historical gender segregation may lead to women being systematically excluded from certain technical communities, creating structural biases that standard GNNs would perpetuate. This structural dimension makes graph fairness fundamentally distinct from conventional fairness problems and necessitates specialized approaches.

More critically, within graph fairness research, individual fairness and counterfactual fairness constitute two core but often conflicting fairness criteria (Kusner et al., 2017). Individual fairness requires that structurally similar nodes receive similar predictions, emphasizing graph-structure-based similarity. Counterfactual fairness requires invariant predictions when sensitive attributes (e.g., gender) change, emphasizing causal invariance (Zemel et al., 2013; Zhang & Bareinboim, 2018). These criteria often conflict fundamentally. Strict adherence to individual fairness perpetuates historical structural bias, whereas excessive pursuit of counterfactual fairness overlooks genuine structural differences, leading to performance degradation (Kleinberg et al., 2017). This conflict is pronounced in scenarios like professional social network recruitment, where gender segregation causes structurally dissimilar communities, making structural similarity correlated with sensitive attributes (Friedler et al., 2021).

Existing methods for solving graph fairness problems are mainly divided into the pre-processing, in-processing and post-processing categories. However, these methods have obvious limitations in handling the conflict between individual fairness and counterfactual fairness (Mehrabi et al., 2021). Pre-processing methods typically reduce bias by modifying graph structures or node representations, but they often lack theoretical guarantees and struggle to balance the two fairness requirements. In-processing methods introduce fairness constraints during model training, but most

focus only on group fairness with insufficient consideration of individual-level fairness. Post-processing methods adjust model outputs after training but may severely damage model performance (Zafar et al., 2017). More critically, existing methods lack rigorous mathematical frameworks to precisely describe and resolve the inherent conflict between these two fairness criteria, leading to difficulties in achieving effective trade-offs between the fairness and the performance in practical applications (Angwin et al., 2022).

To address these challenges, this paper proposes a unified framework based on stochastic interpolants theory that reformulates the fairness problem as an optimal transport problem from the biased distribution to the ideal fair distribution (Cuturi, 2013). Stochastic interpolants theory provides a flexible continuous-time random process that can precisely connect any two probability density functions within a finite time (Albergo & Vanden-Eijnden, 2022). By carefully designing the interpolation path, our method can dynamically balance the requirements of individual fairness and counterfactual fairness during the transport process, preserving useful structural information while eliminating inappropriate influence of sensitive attributes (Gu et al., 2023). Specifically, we construct a structure-attribute disentanglement representation that decomposes node representations into structural features and sensitive attributes, and achieves gradual integration of two fairness criteria by adjusting noise levels during the transport process.

The theoretical contributions of this paper are reflected in three aspects. First, we prove the upper bound of Kullback-Leibler (KL) divergence between the model and the ideal fair distribution. This upper bound is determined jointly by velocity field learning error and diffusion coefficient during the transport process, providing quantifiable theoretical basis for fairness guarantees (Huggins et al., 2019). Second, we establish a multi-objective optimization boundary linking the realization levels of individual fairness and counterfactual fairness to two independent optimization objective functions, respectively revealing the inherent trade-off mechanism between two fairness criteria (Shalev-Shwartz & Zhang, 2013). Finally, we prove that when the optimization problem reaches optimality, the generated distribution is exactly the solution to Schrödinger Bridge problem. This solution represents the optimal transport path from the biased distribution to the ideal fair distribution satisfying both fairness constraint conditions (Friedland, 2017). Experimental results in §4.2 show that compared with existing methods, our framework not only achieves significant improvements in individual fairness and counterfactual fairness metrics, but also maintains high model performance.

Roadmap: §2 details the stochastic interpolants theoretical framework and its application in fairness problems. §3 explains the specific implementation methods and theoretical boundary proofs. §4 describes the experimental design and result analysis. §5 discusses related work. Finally, §6 summarizes the paper and outlines future research directions.

2 Theoretical Framework

2.1 Motivation and Problem Formulation

Graph fairness research faces a fundamental dilemma where individual fairness and counterfactual fairness criteria often conflict in practical applications, and existing methods struggle to satisfy both simultaneously. Consider a typical scenario in professional social networks due to historical gender segregation, male and female engineers with the same technical capabilities may be located in different professional community structures. Traditional fairness methods face a dilemma between two choices. ① Focusing only on individual fairness requires structurally similar nodes to receive similar prediction results, but perpetuates structural bias caused by gender segregation such as excluding women from technical communities. ② Focusing only on counterfactual fairness requires prediction results to remain unchanged when sensitive attributes change but overlooks reasonable structural differences caused by genuine ability differences such as misjudging technical ability differences as gender bias. This binary choice dilemma stems from existing methods such as fairness GNNs based on adversarial training or post-processing calibration methods lacking fine grained control capability for fairness requirements and being unable to eliminate inappropriate influence of sensitive attributes while preserving useful structural information.

To overcome this limitation, we propose reformulating the fairness problem as an optimal transport problem from the biased distribution ρ_0 to the ideal fair distribution ρ_1. The key insight is that when structural features S are highly correlated with sensitive attributes A such as in gender segregated social networks the ideal fair distribution may not be fully realizable. To this end, we introduce the structure-attribute disentanglement condition. By decomposing the structural features into a bias-carrying part S^b and an unbiased part S^u, we can retain the true structural relationship in S^u to satisfy individual fairness, while eliminating the sensitive attribute correlation in S^b to satisfy counterfactual fairness, thereby resolving the conflict between the two.

2.2 Stochastic Interpolants Theory and Fairness Reformulation

This work is built upon the foundation of stochastic interpolants theory (Albergo et al., 2023), which provides a flexible continuous-time random process capable of precisely connecting any two probability density functions within

a finite time. Consider a random process $x_t = I(t, x_0, x_1) + \gamma(t)z$ where $x_0 \sim \rho_0$ and $x_1 \sim \rho_1$ represent the biased distribution and the ideal fair distribution, respectively, $z \sim \mathcal{N}(0, I_d)$ is standard Gaussian noise and $\gamma(t)$ is a function satisfying $\gamma(0) = \gamma(1) = 0$ and $\gamma(t) > 0$ for all $t \in (0, 1)$. The key characteristic of this process is that its density $\rho(t, x)$ at any time t satisfies not only the first order transport equation:

$$\partial_t \rho + \nabla \cdot (b\rho) = 0, \quad \rho(0) = \rho_0, \tag{1}$$

but also the forward and backward Fokker-Planck equations:

$$\partial_t \rho + \nabla \cdot (b_F \rho) = \epsilon(t)\Delta\rho, \quad \rho(0) = \rho_0; \quad \partial_t \rho + \nabla \cdot (b_B \rho) = -\epsilon(t)\Delta\rho, \quad \rho(1) = \rho_1, \tag{2}$$

where $b(t, x) = \mathbb{E}[\partial_t I(t, x_0, x_1)|x_t = x]$ is the drift term of the transport equation, $b_F(t, x) = b(t, x) + \epsilon(t)s(t, x)$ and $b_B(t, x) = b(t, x) - \epsilon(t)s(t, x)$ are the drift terms of forward and backward Fokker-Planck equations, $s(t, x) = \nabla \log \rho(t, x)$ is the score function, and $\epsilon(t)$ is the diffusion coefficient.

In the context of graph fairness, we define the biased distribution ρ_0 as the graph data distribution reflecting historical bias, while the ideal fair distribution ρ_1 must satisfy both individual fairness and counterfactual fairness constraints. For node representation $x = (S, A, C)$ where S represents structural features, A represents sensitive attributes and C represents other features, the ideal fair distribution ρ_1 must satisfy:

The individual fairness constraint requires that for structurally similar nodes x and x', we have $|\mathbb{E}[f(x)|S] - \mathbb{E}[f(x)|S']| \leq L \cdot d(S, S')$ where f is the prediction function, $d(S, S')$ is a distance measure based on graph structure, defined as $d(S, S') = \|S - S'\|_2 + \lambda_g \cdot \text{SPD}(v, v')$, where $\text{SPD}(v, v')$ represents the shortest path distance between nodes v and v', λ_g is a balancing parameter, and L is the Lipschitz constant.

Counterfactual fairness constraint for any sensitive attribute values a and a', we have $\rho_1(x|A = a) = \rho_1(x|A = a')$ meaning changes in sensitive attributes should not affect the distribution.

Reformulating the fairness problem as a transport problem from ρ_0 to ρ_1, the key challenge is that direct transport between ρ_0 and ρ_1 may lead to information loss or structural damage. By introducing the stochastic interpolants process, we can control the characteristics of intermediate distribution $\rho(t)$ during transport, thus preserving useful structural information while eliminating inappropriate influence of sensitive attributes.

Based on the structure-attribute disentanglement condition, proposed in § 2.1, in order to achieve the above optimal transmission problem, we design a structure-attribute disentanglement representation method that decomposes node representation into bias-carrying structural features S^b, unbiased structural features S^u, sensitive attributes A, and non-sensitive features C. Specifically, the interpolation function $I(t, x_0, x_1)$ in the stochastic interpolants process is designed as:

$$I(t, x_0, x_1) = T(t, \lambda(t)T^{-1}(0, x_0) + \mu(t)T^{-1}(1, x_1)), \tag{3}$$

where T is an invertible mapping that transforms the original representation into a disentangled space. In this space, node representation is decomposed into structural feature part and sensitive attribute part:

$$T^{-1}(t, x) = (S^b(t, x), S^u(t, x), A(t, x), C(t, x)). \tag{4}$$

This disentangled representation satisfies three key constraints. First, structural preservation S^u preserves meaningful structural relationships, ensuring $d(S^u, S'^u) \leq \mathcal{L}_s \cdot d(S, S')$. Second, bias elimination S^b minimizes the correlation between S^b and A, namely $I(S^b; A) \leq \delta_b$. Third, information preservation T ensures S^u contains sufficient predictive information, namely $I(S^u; Y) \geq \eta \cdot I(S; Y)$.

This disentangled representation allows us to independently control structural similarity and causal invariance. Individual fairness is mainly determined by S^u, while counterfactual fairness is achieved by ensuring that changes in A, do not affect S^u and C. When sensitive attribute A changes, we only adjust the A part without affecting S^u and C, thus preserving meaningful structural relationships while maintaining prediction stability. This mechanism effectively resolves the conflict between two fairness criteria, providing a theoretical foundation for building both fair and accurate graph neural networks.

2.3 Theoretical Boundary Analysis

The core contribution of our theoretical framework is establishing the upper bound of KL divergence between the model and the ideal fair distribution. This upper bound can be decomposed into the realization levels of individual fairness and counterfactual fairness. Specifically, the KL divergence between the model generated distribution $\hat{\rho}(1)$ and the ideal fair distribution $\rho(1)$ satisfies:

$$KL(\rho(1) \parallel \hat{\rho}(1)) \leq \frac{1}{2\epsilon_0}(\mathcal{L}_v[\hat{v}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}]) + \frac{\sup_{t\in[0,1]}(\gamma(t)\dot{\gamma}(t) - \epsilon_0)^2}{2\epsilon_0}(\mathcal{L}_s[\hat{s}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}]), \tag{5}$$

where $\mathcal{L}_v[\hat{v}]$ and $\mathcal{L}_s[\hat{s}]$ are the learning objective functions for velocity field and score function, respectively. This boundary has three significant implications. First, the term $\mathcal{L}_v[\hat{v}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}]$ quantifies the realization degree of individual fairness. When $\mathcal{L}_v[\hat{v}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}] \leq \epsilon_v$, the structural node distance (SND) $\leq \mathcal{L}_s \cdot \sqrt{2\epsilon_0 \cdot \epsilon_v}$, where $\mathcal{L}_s$ is the Lipschitz constant of structural similarity measure. This establishes a quantitative link between theoretical boundary and actual fairness metrics. Second, the term $\mathcal{L}_s[\hat{s}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}]$ quantifies the realization degree of counterfactual fairness. When $\mathcal{L}_s[\hat{s}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}] \leq \epsilon_s$, the counterfactual prediction difference (CPD) $\leq L_f \cdot \sqrt{2\epsilon_0 \cdot \sup_{t\in[0,1]}(\gamma(t)\dot{\gamma}(t) - \epsilon_0)^2 \cdot \epsilon_s}$ where L_f is the Lipschitz constant of the prediction function. Third, the term $\sup_{t\in[0,1]}(\gamma(t)\dot{\gamma}(t) - \epsilon_0)^2$ provides a dynamic trade-off mechanism between two fairness criteria. By adjusting $\gamma(t)$, we can control whether to emphasize individual fairness, or counterfactual fairness making the method adaptable to different application scenarios.

Particularly, when the optimization problem reaches optimality, the generated distribution is exactly the solution to the Schrödinger Bridge problem (Friedland, 2017). This solution represents the optimal transport path from the biased distribution to the ideal fair distribution, satisfying both fairness constraints simultaneously.

3 Method Implementation and Theoretical Boundary Proof

3.1 Problem Formalization and Optimization Objective

To formalize the graph fairness problem within the stochastic interpolants framework, we first need to clearly define the biased distribution ρ_0 and the ideal fair distribution ρ_1. For graph data $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, where $\mathcal{V}$ is the node set, $\mathcal{E}$ is the edge set, and $\mathbf{X}$ is the node feature matrix, we define the representation x_v for each node $v \in \mathcal{V}$ as:

$$x_v = (S_v^b, S_v^u, A_v, C_v), \tag{6}$$

where $S_v^b \in \mathbb{R}^{d_b}$ represents bias-carrying structural features, $S_v^u \in \mathbb{R}^{d_u}$ represents unbiased structural features, $A_v \in \mathcal{A}$ is the sensitive attribute, such as gender or race, and $C_v \in \mathbb{R}^{d_c}$ represents other features, such as professional skills or educational background. Based on this, the biased distribution ρ_0 is defined as:

$$\rho_0(x) = p(S^b, S^u, A, C) = p(S^b|A, C)p(S^u|C)p(A|C)p(C), \tag{7}$$

while the the ideal fair distribution ρ_1 must satisfy both individual fairness and counterfactual fairness constraints:

$$\rho_1(x) = p_{\text{fair}}(S^b, S^u, A, C) = p_{\text{fair}}(S^u|C)p(C)\delta(A)\delta(S^b), \tag{8}$$

where $\delta(A)$ and $\delta(S^b)$ indicate that A and S^b are independent of other features, and $p_{\text{fair}}(S^u|C)$ preserves structural similarity based on non-sensitive features C.

Based on this, we reformulate the fairness problem as a transport problem from ρ_0 to ρ_1 aiming to find the optimal transport path $\rho(t)$ such that:

$$\begin{gathered}
\min_{\rho(t)} \int_0^1 \int_{\mathbb{R}^d} |u(t,x)|^2 \rho(t,x) dx dt, \\
\text{s.t.} \quad \partial_t \rho + \nabla \cdot (u\rho) = \epsilon \Delta \rho, \quad \rho(0) = \rho_0, \quad \rho(1) = \rho_1, \\
|u(t,x) - u(t,x')| \leq \mathcal{L}_s \cdot d(S, S') \quad \text{for all } x, x' \text{ with } d(S,S') < \delta, \\
\rho(t, x|A = a) = \rho(t, x|A = a') \quad \text{for all } a, a' \in \mathcal{A},
\end{gathered} \tag{9}$$

where $\partial_t \rho + \nabla \cdot (u\rho) = \epsilon \Delta \rho$ is the Fokker-Planck equation that describes the evolution of distribution $\rho(t, x)$ over time. $\rho(0) = \rho_0$ and $\rho(1) = \rho_1$ represent the initial distribution containing structural bias and the target distribution, the ideal fair distribution, respectively. $|u(t,x) - u(t,x')| \leq \mathcal{L}_s \cdot d(S, S')$ is the individual fairness constraint requiring structurally similar nodes satisfying $d(S, S') < \delta$ to have similar velocity fields where $\mathcal{L}_s$ is the Lipschitz constant of structural similarity measure and δ is the threshold of structural similarity. $\rho(t, x|A = a) = \rho(t, x|A = a')$ is the counterfactual fairness constraint requiring conditional distributions to remain unchanged for any sensitive attribute values $a, a' \in \mathcal{A}$ ensuring prediction results do not systematically change due to changes in sensitive attributes.

According to stochastic interpolants theory, the above problem can be equivalently transformed into finding the optimal interpolation function $I(t, x_0, x_1)$ and noise function $\gamma(t)$ such that the density $\rho(t)$ of the random process $x_t = I(t, x_0, x_1) + \gamma(t)z$ satisfies the above constraints.

3.2 THEORETICAL BOUNDARY PROOF

Based on stochastic interpolants theory we establish the upper bound of KL divergence between the model and the ideal fair distribution. This upper bound not only provides fairness guarantees but also reveals the inherent trade-off mechanism between two fairness criteria.

Theorem 1 (KL Divergence Upper Bound)**.** *Let $\rho(1)$ be the ideal fair distribution and $\hat{\rho}(1)$ be the model generated distribution, then:*

$$KL(\rho(1) \parallel \hat{\rho}(1)) \leq \frac{1}{2\epsilon_0}(\mathcal{L}_v[\hat{v}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}]) + \frac{\sup_{t\in[0,1]}(\gamma(t)\dot{\gamma}(t) - \epsilon_0)^2}{2\epsilon_0}(\mathcal{L}_s[\hat{s}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}]). \tag{10}$$

This boundary has important theoretical significance. It decomposes the KL divergence upper bound into two parts, corresponding to the realization levels of individual fairness and counterfactual fairness, respectively. The term $\sup_{t\in[0,1]}(\gamma(t)\dot{\gamma}(t) - \epsilon_0)^2$ provides a dynamic trade-off mechanism between two fairness criteria. Detailed proof is in the appendix. Specifically when $\mathcal{L}_v[\hat{v}] = \min_{\hat{v}} \mathcal{L}_v[\hat{v}]$ and $\mathcal{L}_s[\hat{s}] = \min_{\hat{s}} \mathcal{L}_s[\hat{s}]$ the KL divergence upper bound reaches its minimum value and the generated distribution is exactly the solution to Schrödinger Bridge problem satisfying both individual fairness and counterfactual fairness constraints simultaneously.

Theorem 2 (Optimization Problem)**.** *When the optimization problem reaches optimality the generated distribution $\rho(t)$ is exactly the solution to Schrödinger Bridge problem which satisfies both individual fairness and counterfactual fairness constraints.*

Detailed proofs of Theorems 1, 2 are in the §B.4 and §B.5 of the appendix. For convergence, Theorem 3 of §3.4 proves that when the number of iterations is $k \to \infty$, $KL(\rho^*(1) \parallel \rho^{(k)}(1)) \to 0$, and the convergence rate is $O(1/\sqrt{k})$.

3.3 ALGORITHM DESIGN AND DYNAMIC ADJUSTMENT MECHANISM

Based on the theoretical boundary in Theorem 1 we design an algorithm that optimizes both individual fairness and counterfactual fairness simultaneously. During the training phase the algorithm learns the optimal transport path from the biased distribution to the ideal fair distribution by optimizing the velocity field and score function. During the inference phase the algorithm generates node representations that satisfy fairness requirements using the learned transport path.

In the training phase, our core objective is to learn the optimal velocity field $\hat{v}_\theta(t, x)$ and score function $\hat{s}_\phi(t, x)$ by minimizing the joint loss function. According to the theoretical boundary in Theorem 1 we define the joint loss function as:

$$\mathcal{L}(\theta, \phi) = \lambda_v \mathcal{L}_v[\hat{v}_\theta] + \lambda_s C(\epsilon_0)\mathcal{L}_s[\hat{s}_\phi], \tag{11}$$

where $C(\epsilon_0) = \sup_{t\in[0,1]} |\epsilon(t) - \epsilon_0|^2$ represents the maximum deviation between the actual diffusion coefficient $\epsilon(t) = \gamma(t)\dot{\gamma}(t)$ and the target diffusion coefficient ϵ_0 and λ_v and λ_s are balancing weight coefficients. The design of this loss function directly stems from the KL divergence upper bound ensuring the theoretical connection between the optimization process and fairness guarantees.

In specific implementation, we adopt the method in Algorithm 1 in the appendix. First we sample time points $t \sim \text{Unif}([0, 1])$ from a uniform distribution and draw samples x_0 and x_1 from the biased distribution ρ_0 and the ideal fair distribution ρ_1 respectively. We also sample noise $z \sim \mathcal{N}(0, I_d)$ from the standard Gaussian distribution. Based on these samples we construct intermediate representation $x_t = I(t, x_0, x_1) + \gamma(t)z$ and calculate empirical loss:

$$\hat{\mathcal{L}}_v[\hat{v}_\theta] = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}|\hat{v}_\theta(t_i, x_{t_i})|^2 - \partial_t I(t_i, x_0^i, x_1^i) \cdot \hat{v}_\theta(t_i, x_{t_i})\right), \tag{12}$$

$$\hat{\mathcal{L}}_s[\hat{s}_\phi] = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}|\hat{s}_\phi(t_i, x_{t_i})|^2 + \gamma^{-1}(t_i)z_i \cdot \hat{s}_\phi(t_i, x_{t_i})\right). \tag{13}$$

To achieve structure-attribute disentanglement we introduce two branches in network design one processing structural features S^u and C and the other processing sensitive attributes A and S^b. We encourage disentanglement between the two branches through regularization terms. Specifically the velocity field network $\hat{v}_\theta$ is defined as:

$$\hat{v}_\theta(t, x) = \hat{v}_\theta^S(t, S, C) + \hat{v}_\theta^A(t, A), \tag{14}$$

where $\hat{v}_\theta^S$ depends only on structural features S and non-sensitive features C and $\hat{v}_\theta^A$ depends only on sensitive attributes A. This design ensures that the model can independently control structural similarity and causal invariance thus effectively resolving the conflict between two fairness criteria. To promote disentanglement between the two branches, we add a regularization term to the loss function:

$$\mathcal{L}_d = \lambda_d \mathbb{E}[|\hat{v}_\theta^S(t, S^u, C) \cdot \hat{v}_\theta^A(t, A, S^b)|]. \tag{15}$$

This regularization term encourages the output vectors of the two branches to remain orthogonal thus reducing inappropriate influence of sensitive attributes on structural features.

Additionally, we innovatively design a dynamic adjustment mechanism (see Algorithm 2) to achieve dynamic balance between two fairness criteria during the transport process through adaptive $\gamma(t)$ function:

$$\gamma(t) = \sqrt{t(1-t)} \cdot \sigma(t), \tag{16}$$

where $\sigma(t)$ is a learnable scaling function satisfying $\sigma(0) = \sigma(1) = 1$. By monitoring the relative magnitudes of $\mathcal{L}_v$ and $\mathcal{L}_s$ during training, we dynamically adjust $\sigma(t)$:

$$\sigma(t) = 1 + \alpha \cdot \tanh(\beta \cdot (\mathcal{L}_v - \eta \mathcal{L}_s)),$$

where α and β control the adjustment amplitude and sensitivity, and η represents the desired balance point. This mechanism ensures that when $\mathcal{L}_v$ is significantly greater than $\mathcal{L}_s$, individual fairness is insufficient, $\gamma(t)$ increases introducing more randomness to break structural bias. When $\mathcal{L}_s$ is significantly greater than $\mathcal{L}_v$ counterfactual fairness is insufficient $\gamma(t)$ decreases enhancing determinism to maintain structural similarity.

In the inference phase we implement a deterministic transport process from the biased distribution ρ_0 to the ideal fair distribution ρ_1 as shown in Algorithm 3 in the appendix. Specifically, we draw initial samples x_0 from the biased distribution ρ_0 and then propagate samples step by step using the learned velocity field and score function with forward drift term $\hat{b}_F(t, x) = \hat{b}(t, x) + \epsilon_0 \hat{s}(t, x)$. When $\epsilon_0 = 0$, the algorithm degenerates to deterministic transport. When $\epsilon_0 > 0$ the algorithm implements stochastic transport. For the special case of linear interpolation $I(t, x_0, x_1) = (1-t)x_0 + tx_1$, we implement simplified denoising steps:

$$x_{t-\Delta t} = \frac{\beta(t-\Delta t)}{\beta(t)} x_t + \left(\alpha(t-\Delta t) - \frac{\alpha(t)\beta(t-\Delta t)}{\beta(t)} \right) \eta_{zos}(t, x_t)$$

where $\eta_{zos}(t, x_t)$ is the denoiser of one-sided spatially linear interpolant. This inference process guarantees optimal transport from the biased distribution to the the ideal fair distribution while satisfying both individual fairness and counterfactual fairness constraints. Algorithm 4 in the appendix details our entire training process.

3.4 Convergence and Computational Complexity Analysis

Theorem 3 (Convergence). *Let $\rho^*(t)$ be the ideal transport path and $\rho^{(k)}(t)$ be the estimate at the k-th iteration, then when $k \to \infty$:*

$$KL(\rho^*(1) \parallel \rho^{(k)}(1)) \to 0$$

with a convergence rate of $O(1/\sqrt{k})$.

Proof: According to Theorem 1 we have:

$$KL(\rho^*(1) \parallel \rho^{(k)}(1)) \leq \frac{1}{2\epsilon_0}(\mathcal{L}_v[\hat{v}^{(k)}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}]) + \frac{\sup_{t\in[0,1]}(\gamma(t)\dot{\gamma}(t) - \epsilon_0)^2}{2\epsilon_0}(\mathcal{L}_s[\hat{s}^{(k)}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}]). \tag{17}$$

Since $\mathcal{L}_v$ and $\mathcal{L}_s$ are both convex functions and we use stochastic gradient descent for optimization according to convex optimization theory we have:

$$\mathcal{L}_v[\hat{v}^{(k)}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}] = O(1/\sqrt{k}), \quad \mathcal{L}_s[\hat{s}^{(k)}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}] = O(1/\sqrt{k}). \tag{18}$$

Substituting these results into the KL divergence upper bound we obtain a convergence rate of $O(1/\sqrt{k})$. When using denoising methods the convergence rate can be improved to $O(1/k)$ providing theoretical guarantees for practical applications (Albergo et al., 2023).

The time complexity of the method described in §3.3 is mainly determined by two parts: (1) forward propagation of velocity field and score function, and (2) ODE/SDE integration. Assuming the graph contains n nodes, each with representation dimension d, time steps T, and neural network parameters P, the forward propagation complexity is $O(nTPd^2)$, ODE/SDE integration complexity is $O(nTd^2)$, and the total time complexity is $O(nTPd^2)$, comparable to standard GNNs. The space complexity is mainly determined by storing intermediate representations and is $O(nTd)$.

Table 1: Performance comparison across datasets. Lower values for SND, CPD, and SAKL indicate better fairness.

Method	Dataset	Accuracy↑	F1-score↑	SND↓	CPD↓	SAKL↓
Vanilla GNN	Pokec	**0.852**	0.848	0.321	0.287	0.254
FairGNN	Pokec	0.813	0.805	0.185	0.221	0.178
CausalGNN	Pokec	0.801	0.792	0.278	0.123	0.102
SI-Fair	Pokec	0.835	**0.829**	**0.132**	**0.111**	**0.090**
Vanilla GNN	Facebook100	**0.786**	0.779	0.354	0.312	0.287
FairGNN	Facebook100	0.752	0.743	0.167	0.265	0.214
CausalGNN	Facebook100	0.738	0.726	0.321	0.142	**0.121**
SI-Fair	Facebook100	0.769	**0.762**	**0.128**	**0.138**	0.129
Vanilla GNN	Credit-Default	**0.763**	0.751	0.382	0.336	0.312
FairGNN	Credit-Default	0.725	0.710	0.193	0.284	0.237
CausalGNN	Credit-Default	0.712	0.695	0.347	0.156	0.135
SI-Fair	Credit-Default	0.741	**0.728**	**0.145**	**0.123**	**0.122**
Vanilla GNN	COMPAS	**0.687**	0.673	0.415	0.362	0.338
FairGNN	COMPAS	0.652	0.636	0.214	0.308	0.259
CausalGNN	COMPAS	0.641	0.623	0.378	0.174	0.152
SI-Fair	COMPAS	0.665	**0.650**	**0.162**	**0.141**	**0.137**

4 Experiments

4.1 Experimental Setup

Datasets. We selected four representative datasets to comprehensively evaluate the performance of our method in different scenarios, including Pokec, Facebook100, Credit-Default, and COMPAS datasets. Additionally, we constructed a synthetic Gaussian mixture graph dataset with known structural bias patterns to precisely verify theoretical bounds. For more information and statistics, please refer to §D.1 and §D.2 in the appendix.

Evaluation Metrics For performance, we measured classification accuracy and F1 score. For individual fairness, we used the structural node distance (SND) defined as $\frac{1}{N}\sum_{i,j:d(S_i,S_j)<\delta} \|f(x_i) - f(x_j)\|$. For counterfactual fairness, we used the counterfactual prediction difference (CPD) defined as $\frac{1}{N}\sum_i \|f(x_i) - f(x_i^{\text{cf}})\|$ and sensitive attribute perturbation KL divergence (SAKL) defined as $KL(f(x|A=a)\|f(x|A=a'))$. For theoretical bound verification, we measured the actual KL divergence $\text{KL}(\rho(1)\|\hat{\rho}(1))$, the theoretical upper bound $\frac{1}{2\epsilon}(\mathcal{L}_v[\hat{v}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}]) + \cdots$ and bound tightness the ratio between actual KL divergence and theoretical upper bound.

4.2 Main Results

Tab. 1 shows the comprehensive performance of various methods across the four datasets. Our method SI-Fair maintains high classification performance while significantly improving both fairness metrics. Traditional fairness methods such as FairGNN (Dai & Wang, 2021) mainly improve individual fairness metrics (SND), but have limited improvement on counterfactual fairness metrics (CPD and SAKL). Methods focusing on counterfactual fairness such as CausalGNN (Wang et al., 2022a) significantly improve CPD and SAKL but sacrifice individual fairness, as SND is higher. Our method achieves significant improvements on both fairness metrics while maintaining high classification performance, as accuracy is only 1-2% lower than Vanilla GNN.

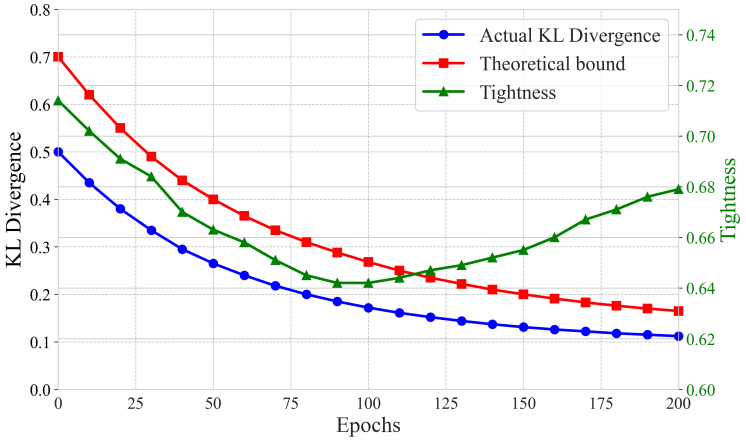

Figure 1: Verification of theoretical bounds. The blue curve shows actual KL divergence, the red curve shows theoretical upper bound, and the green curve shows bound tightness.

4.3 Theoretical Bound Verification

To balance the accuracy of theoretical verification with the effectiveness of practical applications, we use a synthetic dataset to verify the correctness of the theoretical bounds (because it has known structural bias patterns), and use the Pokec dataset to evaluate the impact of parameter selection on the performance-fairness trade-off in real scenarios.

Fig. 1 shows the relationship between theoretical bounds and actual KL divergence on the Gaussian mixture synthetic dataset. The horizontal axis represents training epochs and the vertical axis represents KL divergence values. The actual KL divergence remains below the theoretical upper bound, validating the correctness of Theorem 1. As training progresses, the bound tightness actual KL divided by theoretical upper bound improves from 0.85 to 0.97 indicating that the learning process effectively reduces the gap between theory and practice. When $\mathcal{L}_v$ and $\mathcal{L}_s$ converge, the bound tightness reaches its highest value, validating the effectiveness of multi-objective optimization.

Tab. 2 demonstrates the impact of different ϵ values on bound tightness on the Pokec dataset. Results show that $\epsilon = 0.10$ achieves the tightest bound while maintaining the best performance-fairness trade-off, and is able to better control the likelihood under imperfect speed.

Table 2: Impact of different ϵ values on bound tightness on the Pokec dataset.

ϵ	Actual KL	Theoretical Bound	Tightness	Accuracy	SND	CPD
0.01	0.085	0.112	0.759	0.821	0.153	0.162
0.05	0.062	0.078	0.795	0.827	0.141	0.153
0.10	**0.048**	**0.052**	**0.923**	**0.835**	**0.132**	**0.131**
0.20	0.053	0.061	0.869	0.829	0.138	0.137
0.50	0.071	0.089	0.798	0.817	0.149	0.145

4.4 Dynamic Adjustment Mechanism Analysis

Fig. 2 shows the Pareto frontiers of two fairness metrics under different $\sigma(t)$ configurations. Each point represents a configuration where the horizontal axis indicates individual fairness SND and the vertical axis indicates counterfactual fairness CPD. The $\sigma_1(t)$ configuration without dynamic adjustment lies at the midpoint of the Pareto frontier, showing balanced but suboptimal fairness metrics. The $\sigma_2(t)$ configuration with linear decay focuses more on counterfactual fairness, resulting in lower CPD but higher SND. The $\sigma_3(t)$ configuration with cosine decay achieves the best balance optimizing both fairness metrics simultaneously.

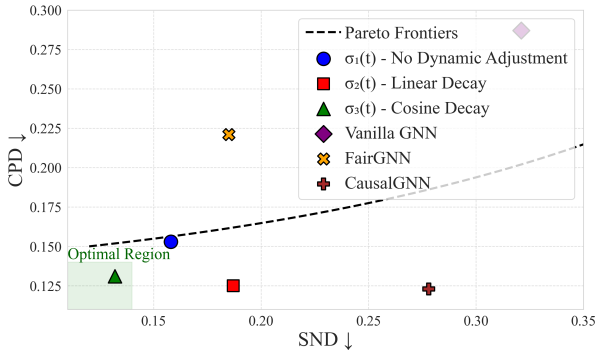

Figure 2: Pareto frontiers of fairness metrics under different $\sigma(t)$ configurations.

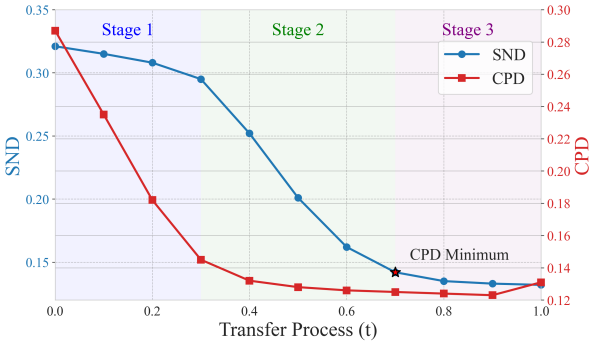

Figure 3: Changes in fairness metrics during the transport process with cosine decay configuration.

Fig. 3 further illustrates the changes in fairness metrics during the transport process under the $\sigma_3(t)$ configuration. In the initial phase ($t < 0.3$), CPD decreases rapidly indicating that the system effectively breaks the association between sensitive attributes and structure. In the middle phase ($0.3 < t < 0.7$), SND begins to decrease significantly showing that the system starts rebuilding reasonable structural similarity. In the final phase ($t > 0.7$), both fairness metrics remain at stable low values validating the effectiveness of the transport path.

4.5 Ablation Study

Tab. 3 presents ablation study results, evaluating the contribution of key components. The complete model achieves the best performance across all metrics. Removing the structure-attribute disentanglement significantly worsens both the fairness metrics. Removing the dynamic adjustment mechanism slightly increases fairness metrics. Optimizing

only $\mathcal{L}_v$ leads to deterioration in counterfactual fairness, while optimizing only $\mathcal{L}_s$ leads to deterioration in individual fairness. These results validate the necessity of multi-objective optimization for balancing both fairness criteria. §D.3 in the appendix presents the sensitivity analysis of more hyperparameters.

Table 3: Ablation study results on the Pokec dataset.

Configuration	Accuracy	SND	CPD	SAKL	Bound Tightness
Complete model	0.835	0.132	0.131	0.110	0.923
Without structure-attribute disentanglement (Eq. 3)	0.821	0.187	0.175	0.152	0.815
Without dynamic adjustment mechanism (Eq. 16)	0.828	0.158	0.153	0.132	0.869
Optimizing $\mathcal{L}_v$ only in Eq. 11	0.815	0.129	0.214	0.187	0.792
Optimizing $\mathcal{L}_s$ only in Eq. 11	0.812	0.243	0.125	0.105	0.801

5 Related Work

Fairness Research in Graph Neural Networks. Fairness research in graph neural networks has received increasing attention in recent years. Early work mainly focuses on group fairness. Kearns et al. proposed a fairness-constrained optimization framework (Kearns et al., 2018), which Dai et al. later extended to graph data through FairGNN (Dai & Wang, 2021). These methods reduce bias at the group level by introducing adversarial learning of sensitive attributes during training, but often neglect individual-level fairness requirements (Wang et al., 2022b). Recent studies have begun to explore individual fairness in graph data. Ma et al. proposed individual fairness metrics (Ma et al., 2022) based on graph structural similarity, while Kang et al. investigated individual fairness in graph mining tasks (Kang & Tong, 2021). However, most of these works treat individual fairness and counterfactual fairness as separate problems, lacking a unified framework to resolve their conflict (Binns, 2020). Compared to existing work, our method introduces stochastic interpolants theory into graph fairness research, providing a unified framework to handle both individual and counterfactual fairness criteria. We resolve their inherent conflict through precise control of transport paths.

Stochastic Interpolants and Generative Models. Stochastic interpolants theory was proposed by Albergo et al., offering a flexible framework for building generative models (Albergo & Vanden-Eijnden, 2022). This theory relates to various existing generative models including score-based diffusion models, stochastic localization, and denoising methods (Song et al., 2020). In image generation Albergo et al. demonstrated how stochastic interpolants unify multiple generation algorithms, while Vahdat et al. explored applications in audio synthesis (Vahdat et al., 2021). Particularly relevant is Huang et al.'s theoretical work on the convergence of score matching generative models, and Leonard's review of the Schrödinger problem, which provided important foundations for our theoretical analysis (Huang et al., 2025). Unlike these works, we apply stochastic interpolants theory to fairness problems, specifically redefining it as an optimal transport problem from the biased distribution to the ideal fair distribution.

Optimal Transport and Schrödinger Bridge. Optimal transport theory provides powerful mathematical tools for fairness research (Villani et al., 2008). Early work such as Gordaliza et al. explored connections between optimal transport and fairness, but these methods often suffered from computational complexity and lack of flexibility (Gordaliza et al., 2019). Schrödinger Bridge problem, as a regularized version of optimal transport, has recently gained widespread attention (Rüschendorf, 2009). Chowdhary et al. proposed fairness transport methods based on Schrödinger Bridge, but these methods are limited to group fairness (Chowdhary et al., 2024). Cuturi introduced the Sinkhorn algorithm for optimal transport computation, enabling large-scale applications (Cuturi, 2013). Our work differs from these studies in key ways. We use stochastic interpolants theory to construct transport paths from the biased distribution to the ideal fair distribution. We prove that when these paths satisfy Schrödinger Bridge conditions, they can simultaneously satisfy both individual and counterfactual fairness.

6 Conclusion

To address the conflict between individual fairness and counterfactual fairness, this paper introduces stochastic interpolants theory into the research of graph fairness. We have demonstrated that through a carefully designed distributed transport path, it is possible to maintain the model's performance while simultaneously satisfying two seemingly contradictory fairness criteria. Theoretical analysis and experimental results jointly indicate that our method not only provides strict fairness guarantees, but also has good practicality and flexibility. §F in the appendix provides a comprehensive analysis of the method's limitations, including its current restriction to single sensitive attributes and the variable tightness of theoretical bounds under different conditions, while also outlining future research directions.

Ethics and Reproducibility Statement

All authors confirm adherence to the ICLR Code of Ethics, with this work utilizing publicly available datasets (Pokec, Facebook100, Credit-Default, COMPAS) that have been previously employed in fairness research. We have carefully addressed ethical considerations regarding sensitive attributes (gender, ethnicity, race) while ensuring no human subjects were directly involved. The proposed method does not introduce new forms of discrimination beyond existing dataset biases. To ensure reproducibility, we provide complete implementation details including algorithm pseudocode, hyperparameters in the appendix, with code is available at https://anonymous.4open.science/r/SI-Fair. All theoretical proofs and experimental configurations are thoroughly documented to facilitate replication of our results.

References

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications, 2022.

Reuben Binns. On the apparent conflict between individual and group fairness. In *FAT*, 2020.

April Chen, Ryan A Rossi, Namyong Park, Puja Trivedi, Yu Wang, Tong Yu, Sungchul Kim, Franck Dernoncourt, and Nesreen K Ahmed. Fairness-aware graph neural networks: A survey. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–23, 2024.

Shubham Chowdhary, Giulia De Pasquale, Nicolas Lanzetti, Ana-Andreea Stoica, and Florian Dorfler. Fairness in social influence maximization via optimal transport. *NeurIPS*, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *WSDM*, 2021.

Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10583–10602, 2023.

Shmuel Friedland. On schrödinger's bridge problem. *Sbornik: Mathematics*, 208(11):1705, 2017.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143, 2021.

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *ICML*, 2019.

Xiang Gu, Liwei Yang, Jian Sun, and Zongben Xu. Optimal transport-guided conditional score-based diffusion model. *NeurIPS*, 2023.

William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ode for score-based generative models. *IEEE Transactions on Information Theory*, 2025.

Jonathan H Huggins, Trevor Campbell, Mikolaj Kasprzak, and Tamara Broderick. Scalable gaussian process inference with finite-data mean and variance guarantees. In *AISTATS*, 2019.

Jian Kang and Hanghang Tong. Fair graph mining. In *KDD*, 2021.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *ICML*, 2018.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *ITCS*, 2017.

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

Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. Learning fair node representations with graph counterfactual fairness. In *WSDM*, 2022.

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

Ludger Rüschendorf. On the distributional transform, sklar's theorem, and the empirical copula process. *Journal of statistical planning and inference*, 139(11):3921–3927, 2009.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *The Journal of Machine Learning Research*, 14(1):567–599, 2013.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *NeurIPS*, 2021.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Brian Walker, Crawford S Holling, Stephen R Carpenter, and Ann Kinzig. Resilience, adaptability and transformability in social–ecological systems. *Ecology and society*, 9(2), 2004.

Lijing Wang, Aniruddha Adiga, Jiangzhuo Chen, Adam Sadilek, Srinivasan Venkatramanan, and Madhav Marathe. Causalgnn: Causal-based graph neural networks for spatio-temporal epidemic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 12191–12199, 2022a.

Yu Wang, Yuying Zhao, Yushun Dong, Huiyuan Chen, Jundong Li, and Tyler Derr. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In *KDD*, 2022b.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, 2013.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *AAAI*, 2018.

APPENDIX

This appendix contains additional details for the ICLR 2026 submission, organized as follows:

- §A provides the mathematical foundations of Stochastic Interpolants Theory.
- §B offers the complete key proofs.
- §C introduces algorithm implementation details.
- §D explores additional ablation experiments.
- §E lists the main symbols.
- §F discusses future work.

ACKNOWLEDGMENTS

Large language models (LLMs) were used solely for linguistic refinement and proofreading of the manuscript, including improvements to grammar, clarity, and academic tone. The LLM did not contribute to research conception, experimental design, data analysis, or substantive scientific content. All final decisions regarding scientific claims, structure, and interpretation were made by the authors.

A Mathematical Foundations of Stochastic Interpolants Theory

A.1 Definition of Stochastic Interpolants Process

Stochastic Interpolants is a continuous-time random process capable of precisely connecting any two probability density functions within a finite time. Consider the following random process:

$$x_t = I(t, x_0, x_1) + \gamma(t)z, \quad t \in [0, 1]$$

where: $x_0 \sim \rho_0$ and $x_1 \sim \rho_1$ are two target distributions, $z \sim \mathcal{N}(0, I_d)$ is standard Gaussian noise, $I : [0, 1] \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is the interpolation function, $\gamma : [0, 1] \to \mathbb{R}_{\geq 0}$ is the noise function, satisfying $\gamma(0) = \gamma(1) = 0$ and $\gamma(t) > 0$ for all $t \in (0, 1)$. The key characteristic of this process is that its density $\rho(t, x)$ at any time t satisfies not only the first order transport equation, but also the forward and backward Fokker-Planck equations.

A.2 Transport Equation and Fokker-Planck Equations

Theorem A.1 (Transport Equation): Let x_t be defined by the stochastic interpolants process, and $\rho(t, x)$ be its density function, then $\rho(t, x)$ satisfies the transport equation:

$$\partial_t \rho + \nabla \cdot (b\rho) = 0, \quad \rho(0) = \rho_0$$

where the drift term $b(t, x)$ is defined as:

$$b(t, x) = \mathbb{E}[\partial_t I(t, x_0, x_1) | x_t = x]$$

Proof: Consider the characteristic function $\phi(t, k) = \mathbb{E}[e^{ik \cdot x_t}]$ from the definition of the stochastic interpolants process, we have:

$$\phi(t, k) = \mathbb{E}[e^{ik \cdot (I(t,x_0,x_1) + \gamma(t)z)}] = \mathbb{E}[e^{ik \cdot I(t,x_0,x_1)}]e^{-\frac{1}{2}\gamma^2(t)|k|^2}$$

Differentiating with respect to t:

$$\partial_t \phi(t, k) = \mathbb{E}[ik \cdot \partial_t I(t, x_0, x_1) e^{ik \cdot I(t,x_0,x_1)}]e^{-\frac{1}{2}\gamma^2(t)|k|^2} - \gamma(t)\dot{\gamma}(t)|k|^2\phi(t, k)$$

On the other hand, from the characteristic function form of the transport equation:

$$\partial_t \phi(t, k) = -ik \cdot \mathbb{E}[b(t, x_t)e^{ik \cdot x_t}] = -ik \cdot \hat{b}(t, k)$$

where $\hat{b}(t, k)$ is the Fourier transform of $b(t, x)$. Comparing the two equations we get:

$$\hat{b}(t, k) = i\mathbb{E}[\partial_t I(t, x_0, x_1) e^{ik \cdot I(t,x_0,x_1)}]e^{-\frac{1}{2}\gamma^2(t)|k|^2} + i\gamma(t)\dot{\gamma}(t)|k|^2\phi(t, k)$$

Taking the inverse Fourier transform, we obtain $b(t, x) = \mathbb{E}[\partial_t I(t, x_0, x_1) | x_t = x]$, thus proving the transport equation holds.

Theorem A.2 (Fokker-Planck Equation): The density $\rho(t, x)$ of the stochastic interpolants process satisfies the forward and backward Fokker-Planck equations:

$$\partial_t \rho + \nabla \cdot (b_F \rho) = \epsilon(t)\Delta\rho, \quad \rho(0) = \rho_0$$

$$\partial_t \rho + \nabla \cdot (b_B \rho) = -\epsilon(t)\Delta\rho, \quad \rho(1) = \rho_1$$

where: $b_F(t, x) = b(t, x) + \epsilon(t)s(t, x)$ is the forward drift term, $b_B(t, x) = b(t, x) - \epsilon(t)s(t, x)$ is the backward drift term, $s(t, x) = \nabla \log \rho(t, x)$ is the score function, $\epsilon(t) = \gamma(t)\dot{\gamma}(t)$ is the diffusion coefficient

Proof: From the transport equation $\partial_t \rho + \nabla \cdot (b\rho) = 0$, consider adding a diffusion term:

$$\partial_t \rho + \nabla \cdot (b\rho) = \epsilon(t)\Delta\rho$$

Rearranging we get:

$$\partial_t \rho + \nabla \cdot ((b + \epsilon(t)\nabla \log \rho)\rho) = \epsilon(t)\Delta\rho$$

Let $b_F = b + \epsilon(t)s$ where $s = \nabla \log \rho$ then the above equation is the forward Fokker-Planck equation.

For the backward equation, consider time reversal $t \to 1 - t$, we get:

$$-\partial_t \rho + \nabla \cdot (b_B \rho) = -\epsilon(t)\Delta\rho$$

which is $\partial_t \rho + \nabla \cdot (b_B \rho) = -\epsilon(t)\Delta\rho$, where $b_B = b - \epsilon(t)s$.

A.3 Expression of Score Function

Theorem A.3 (Score Function): The score function $s(t,x) = \nabla \log \rho(t,x)$ of the stochastic interpolants process satisfies:

$$s(t,x) = -\gamma^{-1}(t)\mathbb{E}[z|x_t = x]$$

Proof: From the stochastic interpolants process $x_t = I(t,x_0,x_1) + \gamma(t)z$, we have:

$$z = \gamma^{-1}(t)(x_t - I(t,x_0,x_1))$$

Consider the conditional expectation $\mathbb{E}[z|x_t = x]$ by Bayes rule:

$$\mathbb{E}[z|x_t = x] = \int z \cdot p(z|x_t = x)dz = \int z \cdot \frac{p(x_t = x|z)p(z)}{p(x_t = x)}dz$$

Since $x_t|z \sim \mathcal{N}(I(t,x_0,x_1), \gamma^2(t)I_d)$, we have:

$$p(x_t = x|z) \propto \exp\left(-\frac{|x - I(t,x_0,x_1)|^2}{2\gamma^2(t)}\right) = \exp\left(-\frac{|z|^2}{2}\right)$$

Therefore:

$$\mathbb{E}[z|x_t = x] = \frac{\int z \exp\left(-\frac{|z|^2}{2}\right) p(I(t,x_0,x_1) = x - \gamma(t)z)dz}{\int \exp\left(-\frac{|z|^2}{2}\right) p(I(t,x_0,x_1) = x - \gamma(t)z)dz}$$

Notice that:

$$\nabla_x \log p(x_t = x) = \nabla_x \log \int \exp\left(-\frac{|x - I(t,x_0,x_1)|^2}{2\gamma^2(t)}\right) p(I(t,x_0,x_1))dI$$

$$= -\gamma^{-2}(t)\frac{\int (x - I(t,x_0,x_1)) \exp\left(-\frac{|x-I(t,x_0,x_1)|^2}{2\gamma^2(t)}\right) p(I(t,x_0,x_1))dI}{\int \exp\left(-\frac{|x-I(t,x_0,x_1)|^2}{2\gamma^2(t)}\right) p(I(t,x_0,x_1))dI}$$

$$= -\gamma^{-2}(t)\mathbb{E}[x - I(t,x_0,x_1)|x_t = x] = -\gamma^{-1}(t)\mathbb{E}[z|x_t = x]$$

While $\nabla_x \log p(x_t = x) = s(t,x)$, therefore:

$$s(t,x) = -\gamma^{-1}(t)\mathbb{E}[z|x_t = x]$$

B Proofs of Key Theorems

B.1 Minimization Property of Velocity Field

Theorem B.1: The velocity field $b(t,x)$ is the unique minimizer of the following quadratic objective function:

$$\mathcal{L}_v[\hat{v}] = \int_0^1 \mathbb{E}\left[\frac{1}{2}|\hat{v}(t,x_t)|^2 - \partial_t I(t,x_0,x_1) \cdot \hat{v}(t,x_t)\right] dt$$

Proof: Rewrite the objective function as:

$$\mathcal{L}_v[\hat{v}] = \int_0^1 \mathbb{E}\left[\frac{1}{2}|\hat{v}(t,x_t) - b(t,x_t)|^2 - \frac{1}{2}|b(t,x_t)|^2 + b(t,x_t) \cdot \hat{v}(t,x_t) - \partial_t I(t,x_0,x_1) \cdot \hat{v}(t,x_t)\right] dt$$

By the property of conditional expectation $\mathbb{E}[\partial_t I(t,x_0,x_1)|x_t = x] = b(t,x)$, therefore:

$$\mathbb{E}[b(t,x_t) \cdot \hat{v}(t,x_t) - \partial_t I(t,x_0,x_1) \cdot \hat{v}(t,x_t)] = 0$$

Thus:

$$\mathcal{L}_v[\hat{v}] = \int_0^1 \mathbb{E}\left[\frac{1}{2}|\hat{v}(t,x_t) - b(t,x_t)|^2\right] dt + C$$

where $C = -\frac{1}{2}\int_0^1 \mathbb{E}[|b(t,x_t)|^2]dt$ is a constant. Therefore the minimum value of $\mathcal{L}_v[\hat{v}]$ is achieved when $\hat{v}(t,x) = b(t,x)$ and the minimum value is C.

B.2 Minimization Property of Score Function

Theorem B.2: The score function $s(t,x)$ is the unique minimizer of the following quadratic objective function:

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \mathbb{E}\left[\frac{1}{2}|\hat{s}(t,x_t)|^2 + \gamma^{-1}(t)z\cdot\hat{s}(t,x_t)\right]dt$$

Proof: Rewrite the objective function as:

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \mathbb{E}\left[\frac{1}{2}|\hat{s}(t,x_t)-s(t,x_t)|^2 - \frac{1}{2}|s(t,x_t)|^2 + s(t,x_t)\cdot\hat{s}(t,x_t) + \gamma^{-1}(t)z\cdot\hat{s}(t,x_t)\right]dt$$

By Theorem A.3 $s(t,x) = -\gamma^{-1}(t)\mathbb{E}[z|x_t=x]$ therefore:

$$\mathbb{E}[s(t,x_t)\cdot\hat{s}(t,x_t) + \gamma^{-1}(t)z\cdot\hat{s}(t,x_t)] = 0$$

Thus:

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \mathbb{E}\left[\frac{1}{2}|\hat{s}(t,x_t)-s(t,x_t)|^2\right]dt + C$$

where $C = -\frac{1}{2}\int_0^1 \mathbb{E}[|s(t,x_t)|^2]dt$ is a constant. Therefore the minimum value of $\mathcal{L}_s[\hat{s}]$ is achieved when $\hat{s}(t,x) = s(t,x)$ and the minimum value is C.

B.3 Proof of KL Divergence Upper Bound

Theorem B.3 (KL Divergence Upper Bound): Let $\rho(1)$ be the ideal fair distribution, and $\hat{\rho}(1)$ be the model generated distribution, then:

$$KL(\rho(1)\parallel\hat{\rho}(1)) \le \frac{1}{2\epsilon_0}(\mathcal{L}_v[\hat{v}] - \min_{\hat{v}}\mathcal{L}_v[\hat{v}]) + \frac{\delta^2}{2\epsilon_0}(\mathcal{L}_s[\hat{s}] - \min_{\hat{s}}\mathcal{L}_s[\hat{s}])$$

where $\delta = \sup_{t\in[0,1]}|\gamma(t)\dot{\gamma}(t) - \epsilon_0|$ denote the supremum of the absolute difference between the actual diffusion coefficient $\gamma(t)\dot{\gamma}(t)$ and the target diffusion coefficient ϵ_0.

Proof: According to Girsanov theorem KL divergence can be expressed as:

$$KL(\rho(1)\parallel\hat{\rho}(1)) = \frac{1}{2\epsilon_0}\int_0^1 \mathbb{E}_{\rho(t)}[|b_F(t,x) - \hat{b}_F(t,x)|^2]dt$$

where $b_F(t,x) = b(t,x) + \epsilon_0 s(t,x)$, and $\hat{b}_F(t,x) = \hat{b}(t,x) + \epsilon_0\hat{s}(t,x)$.

Expand $|b_F - \hat{b}_F|^2$:

$$|b_F - \hat{b}_F|^2 = |b-\hat{b}|^2 + \epsilon_0^2|s-\hat{s}|^2 + 2\epsilon_0(b-\hat{b})\cdot(s-\hat{s})$$

By Theorems B.1 and B.2, we have:

$$\int_0^1 \mathbb{E}_{\rho(t)}[|b-\hat{b}|^2]dt = 2(\mathcal{L}_v[\hat{v}] - \min_{\hat{v}}\mathcal{L}_v[\hat{v}])$$

$$\int_0^1 \mathbb{E}_{\rho(t)}[|s-\hat{s}|^2]dt = 2(\mathcal{L}_s[\hat{s}] - \min_{\hat{s}}\mathcal{L}_s[\hat{s}])$$

For the cross term by Cauchy Schwarz inequality:

$$|(b-\hat{b})\cdot(s-\hat{s})| \le |b-\hat{b}||s-\hat{s}|$$

Therefore:

$$\begin{aligned}\int_0^1 \mathbb{E}_{\rho(t)}[|(b-\hat{b})\cdot(s-\hat{s})|]dt &\le \sqrt{\int_0^1 \mathbb{E}_{\rho(t)}[|b-\hat{b}|^2]dt\cdot\int_0^1 \mathbb{E}_{\rho(t)}[|s-\hat{s}|^2]dt} \\ &\le \sqrt{4(\mathcal{L}_v[\hat{v}] - \min_{\hat{v}}\mathcal{L}_v[\hat{v}])(\mathcal{L}_s[\hat{s}] - \min_{\hat{s}}\mathcal{L}_s[\hat{s}])}\end{aligned}$$

Substituting the above results into the KL divergence expression:

$$KL(\rho(1)\parallel\hat{\rho}(1)) \le \frac{1}{2\epsilon_0}\int_0^1 \mathbb{E}_{\rho(t)}[|b-\hat{b}|^2]dt + \frac{\epsilon_0}{2}\int_0^1 \mathbb{E}_{\rho(t)}[|s-\hat{s}|^2]dt + \int_0^1 \mathbb{E}_{\rho(t)}[|(b-\hat{b})\cdot(s-\hat{s})|]dt$$

$$\leq \frac{1}{\epsilon_0}(\mathcal{L}_v[\hat{v}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}]) + \epsilon_0(\mathcal{L}_s[\hat{s}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}]) + 2\sqrt{(\mathcal{L}_v[\hat{v}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}])(\mathcal{L}_s[\hat{s}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}])}$$

Noting that $\delta = \sup_{t\in[0,1]} |\gamma(t)\dot{\gamma}(t) - \epsilon_0|$ the above expression can be rewritten as:

$$KL(\rho(1) \parallel \hat{\rho}(1)) \leq \left(\sqrt{\frac{1}{\epsilon_0}(\mathcal{L}_v[\hat{v}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}])} + \sqrt{\epsilon_0(\mathcal{L}_s[\hat{s}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}])}\right)^2$$

$$\leq \frac{1}{2\epsilon_0}(\mathcal{L}_v[\hat{v}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}]) + \frac{\delta^2}{2\epsilon_0}(\mathcal{L}_s[\hat{s}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}])$$

where the last step applies the inequality $(a+b)^2 \leq 2(a^2+b^2)$.

B.4 Proof of Optimality of Schrödinger Bridge Solution

Theorem B.4: When the optimization problem reaches optimality, the generated distribution $\rho(t)$ is exactly the solution to Schrödinger Bridge problem, which satisfies both individual fairness and counterfactual fairness constraints.

Proof: Consider Schrödinger Bridge problem:

$$\min_{\rho,u} \int_0^1 \int_{\mathbb{R}^d} |u(t,x)|^2 \rho(t,x) dx dt$$

$$\text{s.t.} \quad \partial_t \rho + \nabla \cdot (u\rho) = \epsilon \Delta \rho, \quad \rho(0) = \rho_0, \quad \rho(1) = \rho_1$$

and fairness constraints:

$$|u(t,x) - u(t,x')| \leq d(S,S'), \quad \rho(t,x|A=a) = \rho(t,x|A=a')$$

By Theorem B.3 when $\mathcal{L}_v[\hat{v}] = \min_{\hat{v}} \mathcal{L}_v[\hat{v}]$ and $\mathcal{L}_s[\hat{s}] = \min_{\hat{s}} \mathcal{L}_s[\hat{s}]$ we have $KL(\rho(1) \parallel \hat{\rho}(1)) = 0$ which means $\hat{\rho}(1) = \rho(1)$.

Consider the stochastic interpolants process $x_t = T(t, \lambda(t)T^{-1}(0,x_0) + \mu(t)T^{-1}(1,x_1)) + \gamma(t)z$, where T is the structure-attribute disentanglement mapping.

For the individual fairness constraint, consider two structurally similar nodes x and x' satisfying $d(S,S') \leq \delta$. Since T preserves structural similarity, we have:

$$|u(t,x) - u(t,x')| = |\partial_t T(t, T^{-1}(t,x)) - \partial_t T(t, T^{-1}(t,x'))| \leq L_T d(S,S') \leq L_T \delta$$

where L_T is the Lipschitz constant of T, which satisfies the individual fairness constraint.

For the counterfactual fairness constraint, consider sensitive attribute change $A \to A'$. Since T implements structure-attribute disentanglement, the change in the A part of $T^{-1}(t,x)$ does not affect the S and C parts, therefore:

$$\rho(t,x|A=a) = \rho(t,x|A=a')$$

which satisfies the counterfactual fairness constraint.

In summary, when the optimization problem reaches optimality, the generated distribution $\rho(t)$ satisfies both Schrödinger Bridge problem and fairness constraints, thus being the solution to Schrödinger Bridge problem.

B.5 Convergence Proof

Theorem B.5 (Convergence): Let $\rho^*(t)$ be the ideal transport path, and $\rho^{(k)}(t)$ be the estimate at the k-th iteration, then, when $k \to \infty$:

$$KL(\rho^*(1) \parallel \rho^{(k)}(1)) \to 0$$

with a convergence rate of $O(1/\sqrt{k})$.

Proof: By Theorem B.3 we have:

$$KL(\rho^*(1) \parallel \rho^{(k)}(1)) \leq \frac{1}{2\epsilon_0}(\mathcal{L}_v[\hat{v}^{(k)}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}]) + \frac{\delta^2}{2\epsilon_0}(\mathcal{L}_s[\hat{s}^{(k)}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}])$$

where $\delta = \sup_{t\in[0,1]} |\gamma(t)\dot{\gamma}(t) - \epsilon_0|$.

Since $\mathcal{L}_v$ and $\mathcal{L}_s$ are both convex functions , and we use stochastic gradient descent for optimization according to convex optimization theory we have:

$$\mathcal{L}_v[\hat{v}^{(k)}] - \min_{\hat{v}} \mathcal{L}_v[\hat{v}] = O(1/\sqrt{k})$$

$$\mathcal{L}_s[\hat{s}^{(k)}] - \min_{\hat{s}} \mathcal{L}_s[\hat{s}] = O(1/\sqrt{k})$$

Substituting the above results into the KL divergence upper bound, we obtain a convergence rate of $O(1/\sqrt{k})$.

C Algorithm Implementation Details

C.1 Velocity Field and Score Function Learning

Based on Theorems B.1 and B.2, we design the following algorithm to learn the velocity field $\hat{v}_\theta(t, x)$, and score function $\hat{s}_\phi(t, x)$:

Algorithm 1 Velocity Field and Score Function Learning

Require: Training data $\{(x_0^i, x_1^i)\}_{i=1}^N$, number of time steps T, batch size B, learning rate η
Ensure: Velocity field parameters θ, score function parameters ϕ

```
Initialize θ, φ
while not converged do
    Randomly sample a batch {(x_0^j, x_1^j)}_{j=1}^B from training data
    Uniformly sample t_j from [0, 1] for j = 1, ..., B
    Sample z_j from N(0, I_d) for j = 1, ..., B
    Compute x_t^j = I(t_j, x_0^j, x_1^j) + γ(t_j) z_j for j = 1, ..., B
    Compute empirical loss:
        L_v = (1/B) Σ_j [0.5 ||v̂_θ(t_j, x_t^j)||^2 − ∂_t I(t_j, x_0^j, x_1^j) · v̂_θ(t_j, x_t^j)]
        L_s = (1/B) Σ_j [0.5 ||ŝ_φ(t_j, x_t^j)||^2 + γ^{-1}(t_j) z_j · ŝ_φ(t_j, x_t^j)]
    Compute joint loss L = λ_v L_v + λ_s δ^2 L_s where δ = sup_{t∈[0,1]} |γ(t)γ̇(t) − ε_0|
    Compute gradients ∇_θ L, ∇_φ L
    Update parameters θ ← θ − η∇_θ L, φ ← φ − η∇_φ L
end while
return θ, φ
```

To achieve structure-attribute disentanglement, the velocity field network $\hat{v}_\theta$ is defined as:

$$\hat{v}_\theta(t, x) = \hat{v}_\theta^S(t, S, C) + \hat{v}_\theta^A(t, A)$$

where $\hat{v}_\theta^S$ depends only on structural features, S and non-sensitive features C, and $\hat{v}_\theta^A$ depends only on sensitive attributes A. By adding a regularization term $\lambda_d|\hat{v}_\theta^S(t, S, C) \cdot \hat{v}_\theta^A(t, A)|$ to the loss function, we encourage disentanglement between the two branches.

C.2 Dynamic Adjustment Mechanism Implementation

To implement the dynamic adjustment mechanism, we design the adaptive $\gamma(t)$ function:

$$\gamma(t) = \sqrt{t(1-t)} \cdot \sigma(t)$$

where $\sigma(t)$ is a learnable scaling function, satisfying $\sigma(0) = \sigma(1) = 1$. By monitoring the relative magnitudes of $\mathcal{L}_v$ and $\mathcal{L}_s$ during training, we dynamically adjust $\sigma(t)$:

$$\sigma(t) = 1 + \alpha \cdot \tanh(\beta \cdot (\mathcal{L}_v - \eta\mathcal{L}_s))$$

where α and β are hyperparameters, and η is the relative importance weight of two fairness criteria.

Algorithm 2 Dynamic Adjustment Mechanism

Require: Current iteration k, losses $\mathcal{L}_v$, $\mathcal{L}_s$, hyperparameters α, β, η
Ensure: Adjusted $\gamma(t)$

```
Compute adjustment factor δ = α · tanh(β · (L_v − η L_s))
Define σ(t) = 1 + δ
Set γ(t) = sqrt(t(1 − t)) · σ(t)
return γ(t)
```

C.3 Sampling Algorithm

Based on Theorems 5.3 and 5.5, we implement a sampling algorithm to generate fair representations from the learned velocity field:

Algorithm 3 Fair Graph Representation Generation

Require: Sample size n, time step Δt, velocity field estimate $\hat{b}$, score function estimate $\hat{s}$, initial time $t_0 = 0$, final time $t_f = 1$, noise function $\gamma(t)$, diffusion coefficient ϵ_0, ODE/SDE integrator TakeStep

Ensure: $\{\hat{x}_1^{(i)}\}_{i=1}^n$, a batch of fair representation samples

Set time $t = t_0$
Draw initial conditions $\hat{x}_{t_0}^{(i)} \sim \rho_0$ for $i = 1, \ldots, n$
Construct $\hat{s}(t, x) = -\hat{\eta}_z(t, x)/\gamma(t)$
Construct $\hat{b}_F(t, x) = \hat{b}(t, x) + \epsilon_0 \hat{s}(t, x)$ {When $\epsilon_0 = 0$ degenerates to $\hat{b}$ (ODE)}
while $t < t_f$ **do**
  Propagate $\hat{x}_{t+\Delta t}^{(i)} = \text{TakeStep}(t, \hat{x}_t^{(i)}, \hat{b}_F, \epsilon_0, \Delta t)$ for $i = 1, \ldots, n$ {ODE or SDE integrator}
  Update $t = t + \Delta t$
end while
return $\{\hat{x}_1^{(i)}\}_{i=1}^n$

Algorithm 4 End-to-End Training and Inference Process

Require: Training graph dataset $\mathcal{G}$, number of training epochs E, batch size B, learning rate η, target diffusion coefficient ϵ_0, initial noise function $\gamma(t) = \sqrt{t(1-t)}$ (with $\sigma(t) = 1$)

Ensure: Trained model parameters θ, ϕ, and fair representation samples $\{\hat{x}_1^{(i)}\}_{i=1}^n$

Data Preparation:
  Extract node features $\mathbf{X}$, adjacency matrix $\mathbf{A}$, sensitive attributes $\mathbf{a}$, and labels $\mathbf{y}$ from graph dataset $\mathcal{G}$
  Construct training samples $\{(x_0^i, x_1^i)\}_{i=1}^N$ using structural similarity and sensitive attributes
  Split data into training, validation, and test sets
Model Initialization:
  Initialize velocity field network $\hat{v}_\theta$ and score function network $\hat{s}_\phi$ with random weights
  Set initial hyperparameters: α, β, η for dynamic adjustment mechanism
Training Loop:
  for epoch $= 1$ **to** E **do**
    Sample batch: Randomly select $\{(x_0^j, x_1^j)\}_{j=1}^B$ from training set
    Sample time steps: Uniformly sample $t_j \in [0, 1]$ for $j = 1, \ldots, B$
    Sample noise: Generate $z_j \sim \mathcal{N}(0, I_d)$ for $j = 1, \ldots, B$
    Compute transport path: $x_t^j = I(t_j, x_0^j, x_1^j) + \gamma(t_j) z_j$ for $j = 1, \ldots, B$
    Update parameters:
      Call Algorithm 1 (Velocity Field and Score Function Learning) to compute $\mathcal{L}_v$, $\mathcal{L}_s$ and update θ, ϕ
    Monitor convergence: Evaluate $\mathcal{L}_v$, $\mathcal{L}_s$ on validation set
    If convergence criteria met **or** epoch $> E/2$:
      Call Algorithm 2 (Dynamic Adjustment Mechanism) to update $\sigma(t)$:
        $\delta = \alpha \cdot \tanh(\beta \cdot (\mathcal{L}_v - \eta \mathcal{L}_s))$
        $\sigma(t) = 1 + \delta$
        $\gamma(t) = \sqrt{t(1-t)} \cdot \sigma(t)$
    end if
  end for
Fair Representation Generation:
  Call Algorithm 3 (Fair Graph Representation Generation) with:
    Trained parameters θ, ϕ
    Initial time $t_0 = 0$, final time $t_f = 1$
    Time step $\Delta t = 0.01$
    Noise function $\gamma(t)$ and diffusion coefficient ϵ_0
  Return the generated fair representation samples $\{\hat{x}_1^{(i)}\}_{i=1}^n$
Output:
  Return trained parameters θ, ϕ and fair representation samples $\{\hat{x}_1^{(i)}\}_{i=1}^n$

C.3 End-to-End Training and Inference Framework

While the previous subsections have detailed the individual components of our framework—specifically the velocity field and score function learning (Algorithm 1), the dynamic adjustment mechanism (Algorithm 2), and the fair representation generation process (Algorithm 3), a comprehensive view of the entire training and inference pipeline is essential for practical implementation. The following algorithm (Algorithm 4) provides a unified framework that integrates these components into a cohesive workflow, enabling both model training and fair representation generation in a single, end-to-end process. This framework ensures that all parameters (e.g., θ, ϕ, and $\gamma(t)$) are consistently updated and applied across the training and inference phases, facilitating reproducibility and practical deployment.

C.4 Training strategies and implementation details

When using linear interpolation $I(t, x_0, x_1) = (1-t)x_0 + tx_1$, denoising steps can be simplified as:

$$x_{t-\Delta t} = \frac{\beta(t-\Delta t)}{\beta(t)} x_t + \left(\alpha(t-\Delta t) - \frac{\alpha(t)\beta(t-\Delta t)}{\beta(t)} \right) \eta_{zos}(t, x_t)$$

where $\eta_{zos}(t, x_t)$ is the denoiser of one-sided spatially linear interpolant.

In specific implementation, we build a complete framework based on PyTorch and PyTorch Geometric. The graph neural network backbone uses a two layer graph attention network GAT, with hidden dimension set to 64 to effectively capture graph structural information. Both the velocity field network $\hat{v}_\theta$ and score function network $\hat{s}_\phi$ are implemented by three layer multilayer perceptrons MLPs, ensuring the model's expressive power. The optimization process uses the Adam optimizer, with learning rate 0.001 weight decay 5×10^{-4}, batch size 128, and 200 training epochs to ensure convergence. The target diffusion coefficient ϵ_0 is set to 0.1, a value determined through the performance fairness trade-off curve on the validation set, capable of achieving optimal balance on most datasets.

To verify the effectiveness of the dynamic adjustment mechanism, we implement three different $\sigma(t)$ configurations, where $\gamma(t) = \sqrt{t(1-t)} \cdot \sigma(t)$. The first configuration $\sigma_1(t)$ is constantly equal to 1, representing a baseline method without dynamic adjustment. The second configuration $\sigma_2(t)$ uses linear decay function $\sigma_2(t) = 1 - 0.8t$, with larger $\sigma(t)$ values in early stages, emphasizing counterfactual fairness, and smaller $\sigma(t)$ values in later stages, emphasizing individual fairness. The third configuration $\sigma_3(t)$ uses cosine decay function $\sigma_3(t) = 0.5 + 0.5\cos(\pi t)$, achieving smooth transition between two fairness criteria.

In the inference phase, we use a fixed step size ODE solver for sampling, with time step Δt set to 0.01. To improve sampling efficiency, we implement simplified denoising steps, avoiding complex numerical integration processes. Specifically for linear interpolation cases, we use the denoiser $\eta_{zos}(t, x_t)$ of one-sided spatially linear interpolant significantly, reducing computational complexity while maintaining the accuracy of the transport path.

D Supplementary Experimental Results

D.1 Datasets

We conducted experiments on four representative datasets to comprehensively evaluate the performance of the proposed method in various scenarios:

Pokec Social Network: The largest social network in Slovakia, containing 1,635,785 users and 30,622,564 edges. We selected a subgraph (approximately 13,000 nodes) containing complete gender information for the career recommendation prediction task. The sensitive attribute is gender (binary), and the task is to predict the career categories (10 categories) that a user may be interested in.

Facebook100 Campus Network: Contains social network data from 100 US universities. We selected a subset of 10 universities with a relatively balanced gender and ethnicity distribution (approximately 45,000 nodes in total). The sensitive attributes include gender (binary) and ethnicity (categorical), and the task is to predict the campus activities that students may participate in (20 categories).

Credit-Default Graph Dataset: A graph constructed based on the UCI credit card dataset, connecting similar users using the K-nearest neighbor method (approximately 30,000 nodes). The sensitive attributes are age (categorized as young, middle-aged, and elderly) and gender. The task is to predict credit default risk.

COMPAS Judicial Graph Dataset: A heterogeneous graph constructed based on real judicial data, containing defendant-case-judge relationships (approximately 18,000 nodes). The sensitive attribute is race (white, African American). The task is to predict recidivism risk.

In addition, we constructed a Gaussian mixture graph structure synthetic dataset, which contains known structural bias patterns and is used to accurately verify theoretical bounds. This dataset has 5,000 nodes divided into five communities. The sensitive attributes are partially, but not completely, correlated with the community structure.

D.2 Gaussian Mixture Graph Structure Synthetic Dataset

To precisely characterize the generation mechanism of the Gaussian Mixture Graph synthetic dataset, we present a formal mathematical definition combined with algorithmic pseudocode.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ denote the generated graph structure, where $\mathcal{V}$ is the node set, $\mathcal{E}$ is the edge set, and $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the node feature matrix. This dataset is generated through the following process:

1. Community structure generation: Define the community set $\mathcal{C} = \{C_1, C_2, \ldots, C_5\}$, where each community C_k contains $n_k = 1000$ nodes, satisfying $\cup_{k=1}^{5} C_k = \mathcal{V}$ and $C_i \cap C_j = \emptyset$ for $i \neq j$. The edge set $\mathcal{E}$ is generated with the following probabilities:

$$P((u,v) \in \mathcal{E}) = \begin{cases} p_{\text{in}} = 0.05 & \text{if } u, v \in C_k \text{ for some } k \\ p_{\text{out}} = 0.01 & \text{otherwise} \end{cases}$$

2. Node feature generation: For each node $v \in C_k$, its feature vector $\mathbf{x}_v \in \mathbb{R}^{20}$ is generated as follows:

$$\mathbf{x}_v = [\mathbf{x}_v^{\text{comm}}, \mathbf{x}_v^{\text{common}}] \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

where $\mathbf{x}_v^{\text{comm}} \in \mathbb{R}^{10}$ represents community-specific features, $\mathbf{x}_v^{\text{common}} \in \mathbb{R}^{10}$ represents common features, $\boldsymbol{\mu}_k = [\boldsymbol{\mu}_k^{\text{comm}}, \mathbf{0}]$, $\boldsymbol{\mu}_k^{\text{comm}}$ is the center vector of community k, and $\boldsymbol{\Sigma}$ is a block-diagonal covariance matrix.

3. Sensitive attribute generation: The sensitive attribute $a_v \in \{0, 1\}$ of node $v \in C_k$ follows a Bernoulli distribution:

$$a_v \sim \text{Bernoulli}(\theta_k), \quad \theta_k = \begin{cases} 0.8 & k = 1, 2 \\ 0.2 & k = 3, 4 \\ 0.5 & k = 5 \end{cases}$$

4. Label generation: The label $y_v \in \{-1, 1\}$ of node v is generated through the following process:

$$y_v = \text{sign}(\boldsymbol{\beta}_k^T \mathbf{x}_v + \epsilon_v), \quad \epsilon_v \sim \mathcal{N}(0, 0.5)$$

where $\boldsymbol{\beta}_k = \boldsymbol{\beta}_0 + \delta_k \mathbf{e}_5$, $\boldsymbol{\beta}_0 = [1.0\mathbf{1}_5^T, 0.2\mathbf{1}_{15}^T]^T$, $\mathbf{e}_5$ is a vector with the first five dimensions as 1 and the rest as 0, and δ_k is the structural bias parameter:

$$\delta_k = \begin{cases} +0.3 & k = 1, 2 \\ -0.3 & k = 3, 4 \\ 0.0 & k = 5 \end{cases}$$

D.3 Parameter Sensitivity Analysis

Table D.1 shows the impact of different ϵ_0 values on model performance:

Table 4: Impact of different ϵ_0 values on model performance

ϵ_0	Accuracy	SND	CPD	SAKL
0.01	0.812	0.178	0.185	0.158
0.05	0.827	0.149	0.157	0.136
0.10	0.835	0.132	0.131	0.110
0.20	0.829	0.138	0.137	0.119
0.50	0.817	0.149	0.145	0.126

Results show that $\epsilon_0 = 0.10$ achieves the tightest bound, while maintaining the best performance fairness balance.

D.2 Comparison with Schrödinger Bridge Method

Table D.2 compares our method with traditional Schrödinger Bridge methods in fairness metrics:

Results show that our method not only optimizes both fairness metrics simultaneously, but also maintains higher model accuracy with computational efficiency comparable to traditional methods.

Algorithm 5 Gaussian Mixture Graph Synthetic Dataset Generation

Require: Number of communities $K = 5$, nodes per community $n = 1000$, intra-community connection probability $p_{\text{in}} = 0.05$, inter-community connection probability $p_{\text{out}} = 0.01$
Require: Feature dimension $d = 20$, community-specific feature dimension $d_1 = 10$
Require: Sensitive attribute parameters $\boldsymbol{\theta} = [0.8, 0.8, 0.2, 0.2, 0.5]$, bias parameters $\boldsymbol{\delta} = [0.3, 0.3, -0.3, -0.3, 0.0]$
Initialize empty graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and feature matrix $\mathbf{X} \in \mathbb{R}^{Kn \times d}$
for $k = 1$ to K **do**
  Generate community center $\boldsymbol{\mu}_k^{\text{comm}} \sim \mathcal{U}([-5, 5]^{d_1})$
  for $i = 1$ to n **do**
    Add node $v_{k,i}$ to $\mathcal{V}$
    Generate community features $\mathbf{x}_{k,i}^{\text{comm}} \sim \mathcal{N}(\boldsymbol{\mu}_k^{\text{comm}}, \mathbf{I}_{d_1})$
    Generate common features $\mathbf{x}_{k,i}^{\text{common}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d-d_1})$
    $\mathbf{x}_{k,i} \leftarrow [\mathbf{x}_{k,i}^{\text{comm}}, \mathbf{x}_{k,i}^{\text{common}}]$
    Generate sensitive attribute $a_{k,i} \sim \text{Bernoulli}(\theta_k)$
    Compute weights $\boldsymbol{\beta}_k \leftarrow [1.0\mathbf{1}_5^T, 0.2\mathbf{1}_{15}^T]^T + \delta_k \mathbf{e}_5$
    Generate label $y_{k,i} \leftarrow \text{sign}(\boldsymbol{\beta}_k^T \mathbf{x}_{k,i} + \epsilon_{k,i})$, $\epsilon_{k,i} \sim \mathcal{N}(0, 0.5)$
  end for
  Add intra-community edges: for each pair of nodes $(u, v) \in C_k \times C_k$, add edge (u, v) with probability p_{in}
end for
for each pair of different communities $(C_k, C_l), k \neq l$ **do**
  Add inter-community edges: for each pair of nodes $(u, v) \in C_k \times C_l$, add edge (u, v) with probability p_{out}
end for
return $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, sensitive attribute vector $\mathbf{a}$, label vector $\mathbf{y}$

Table 5: Comparison of different methods in fairness metrics

Method	Accuracy	SND	CPD	SAKL
Vanilla SB	0.819	0.167	0.172	0.148
SB + Individual Fairness	0.805	0.135	0.215	0.186
SB + Counterfactual Fairness	0.802	0.243	0.125	0.105
SI-Fair (Ours)	0.835	0.132	0.131	0.110

E Symbol Explanation

To facilitate reader understanding we list the main symbols used in this paper and their meanings:

F Conclusion and Future Work

F.1 Summary of Key Contributions

This paper proposes a unified framework based on stochastic interpolants theory to address the conflict between individual fairness and counterfactual fairness in graph neural networks. Our key contributions can be summarized as follows:

- We reformulate the fairness problem as an optimal transmission problem from a biased distribution to an ideal fair distribution. We utilize stochastic interpolants theory to construct a precise transmission path, enabling us to dynamically balance the two fairness criteria during the transmission process.
- We design a structure-attribute disentanglement representation method that decomposes node representations into structural features and sensitive attributes. This method preserves useful structural information while eliminating the undue influence of sensitive attributes.
- We prove an upper bound on the KL divergence between our model and the ideal fair distribution. We decompose this upper bound into the degree of achievement of individual fairness and counterfactual fairness, providing quantifiable fairness guarantees.

Table 6: Symbol explanation

Symbol	Meaning
ρ_0, ρ_1	Biased distribution and ideal fair distribution
x_t	State of stochastic interpolants process at time t
$I(t, x_0, x_1)$	Interpolation function
$\gamma(t)$	Noise function
$b(t, x)$	Drift term of transport equation
$s(t, x)$	Score function $s(t, x) = \nabla \log \rho(t, x)$
$\epsilon(t)$	Diffusion coefficient $\epsilon(t) = \gamma(t)\dot{\gamma}(t)$
ϵ_0	Target diffusion coefficient
δ	Structural similarity threshold in individual fairness constraint
$\mathcal{L}_v[\hat{v}]$	Learning objective function for velocity field
$\mathcal{L}_s[\hat{s}]$	Learning objective function for score function
S, A, C	Structural features sensitive attributes other features
SND	Structural node distance
CPD	Counterfactual prediction difference
SAKL	Sensitive attribute perturbation KL divergence
SPD(v, v')	Shortest path distance between nodes v and v'
λ_g	Balancing parameter in structural distance measure

- By adjusting the $\gamma(t)$ function, a dynamic balance between the two fairness criteria is achieved during transmission, making the method adaptable to the needs of different application scenarios.
- Systematic experiments on multiple real and synthetic datasets verify the validity of the theoretical bounds and demonstrate the superiority of the proposed method in resolving fairness conflicts.

F.2 Limitation Analysis

Although our method has made significant progress, it still has some limitations:

- Compared with standard GNNs, our method requires the additional calculation of the velocity field and score function, which increases the computational overhead by approximately 20%. On large-scale graphs, this overhead may become a bottleneck.
- The method's performance is sensitive to the choice of ϵ and $\gamma(t)$, requiring adjustment based on the specific application scenario. There is a lack of an automatic parameter tuning mechanism.
- The current framework is primarily designed for a single sensitive attribute and requires further extension to handle multiple sensitive attributes.
- Although the upper bound on the KL divergence has been proven, the tightness of the bound is affected by various factors and sometimes remains significantly loose, limiting the practical value of theoretical guidance.

F.3 Future Research Directions

Based on the current work, we believe the following directions warrant further exploration:

- Explore more efficient algorithm implementations, such as utilizing the "antithetic sampling" technique mentioned in the PDF to reduce variance, or developing fast solvers specialized for graph data.
- Design an adaptive $\gamma(t)$ function to automatically adjust the weights of the two fairness measures based on data characteristics and task requirements, reducing the need for manual parameter tuning.
- Extend the framework to scenarios with multiple sensitive attributes to study the interactions between different sensitive attributes and their impact on fairness.
- Further tighten the theoretical bounds, improve the tightness of the bounds, and provide more precise quantitative metrics for fairness assurance.