ENFORCING CONDITIONAL INDEPENDENCE IN FLOW MATCHING VIA DIVERGENCE-DIFFERENCE REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study label-conditional fairness by encouraging a prediction score to be independent of sensitive attributes given the target label. We propose **CI-CFM**, a conditional flow-matching predictor with a compatible conditional-density regularizer. A normalized auxiliary model estimates group-conditional score densities, and their mixture defines the label-conditional reference. Evaluating both densities at the deployed score yields a differentiable log-ratio penalty, with density fitting and predictor updates alternated without adversarial training. The population construction recovers conditional mutual information when the component densities and mixing law match the true distributions; error bounds characterize departures from this identity. Across ten synthetic seeds, regularizing the deployed score rather than a single stochastic endpoint reduces mean residual HSIC by 38.7% without increasing mean prediction error. On Adult, CI-CFM lowers the mean equalized-odds gap from 0.1206 to 0.0628 and EDDI from 0.0663 to 0.0605, with small decreases in mean AUROC and F1. Across four MIMIC-III/IV settings, it achieves the highest AUROC and lowest EO and EDDI point estimates among the compared systems. CIFAR-10 and CelebA experiments retain comparable sample quality under an independent-nuisance null control. Code is anonymously available at `https://anonymous.4open.science/r/CICFM-0B67/`.

1 INTRODUCTION

Fair prediction requires controlling how model outputs vary across sensitive groups while retaining information needed for the task. We study this requirement through label-conditional independence:

$$S_\theta(X) \perp Z \mid Y_{\text{true}},$$

where $S_\theta(X)$ is the deployed prediction score, Z denotes a sensitive attribute, and Y_{true} is the target label. This condition requires the score distribution to remain invariant across sensitive groups conditional on the true label. In binary classification, it implies equalized odds (EO) for decisions formed using a common threshold (Hardt et al., 2016). For regression, it provides a continuous-output criterion for conditional group invariance.

Conditional independence can be encouraged through adversarial learning, kernel regularization, and variational or contrastive information objectives (Zhang et al., 2018; Pogodin et al., 2023; Poole et al., 2019). A density-based approach starts from conditional mutual information, whose log-density ratio compares the group-conditional score distribution with its label-conditional marginal. Turning this population target into a training objective involves two concrete design choices: how to relate the fitted conditional and marginal densities, and which prediction variable to regularize. The latter is particularly relevant when a stochastic predictor is deployed by averaging several generated outputs.

We introduce **CI-CFM**, a conditional flow-matching framework with a compatible conditional-density regularizer. Building on the population divergence characterization of Ahuja et al. (2021), we fit normalized group-conditional score densities and construct their label-conditional marginal by mixing the same components. The mixing law is estimated from training data, with exact summation for categorical attributes and a sampled mixture for continuous attributes. Evaluating the component and mixture densities at the same score yields a differentiable plug-in regularizer. The auxiliary density

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

model is fitted by likelihood, and the predictor receives gradients through the score while the density parameters are held fixed.

The second design choice is to apply this regularizer directly to the deployed prediction score. Conditional flow matching produces predictions by integrating a learned vector field from sampled initial states (Lipman et al., 2023; Tong et al., 2024). For prediction tasks, CI-CFM averages four terminal outputs from a fixed antithetic noise bank and uses the same score construction during training and evaluation. The regularizer therefore acts on the output used to assess predictive performance and fairness. Together, density compatibility and deployment alignment give the training objective a consistent component-to-mixture construction and a clearly defined prediction target.

Controlled experiments demonstrate the value of this design. Across ten synthetic seeds, deployment alignment reduces cross-fitted residual HSIC, our primary conditional-dependence diagnostic, by **38.7%** relative to an otherwise matched variant that regularizes a single stochastic endpoint, while preserving predictive accuracy. CI-CFM also achieves lower primary HSIC than adversarial, CIRCE, and conditional-InfoNCE regularizers on all ten seeds under matched generator-update budgets. A separate non-Gaussian stress test retains dependence reductions under a misspecified mixing law.

Real-data experiments assess the method in tabular and clinical prediction. On Adult, CI-CFM reduces the mean EO gap from 0.1206 to 0.0628 relative to unregularized CFM. Across the four evaluated MIMIC-III/IV dataset–task settings, it achieves the highest AUROC and lowest EO and EDDI point estimates among the compared systems. CIFAR-10 and CelebA experiments extend the evaluation to feature-space generation, measuring sample quality under an independent-nuisance null control.

Our contributions are threefold. **(1) Compatible conditional-density regularization.** We construct the label-conditional marginal from the same normalized group-conditional components, yielding a component-to-mixture objective for categorical and continuous sensitive attributes. **(2) Deployment-aligned flow prediction.** We regularize the fixed aggregated score used at inference and propagate its gradients through the terminal flow predictions. **(3). Controlled evaluation of the mechanism and its applications.** Matched baselines and alignment ablations identify gains in conditional-dependence diagnostics, while stress tests and real-data studies characterize predictive utility, fairness, and computational cost.

2 Related Work

Fairness-aware prediction and generative modeling. Fairness-aware prediction addresses group disparities through post-processing, constrained optimization, and representation learning. Post-processing adjusts decision rules to satisfy criteria such as equalized odds (EO), while reductions and constrained learning incorporate fairness requirements into predictor training (Hardt et al., 2016; Woodworth et al., 2017; Agarwal et al., 2018; Cotter et al., 2019). Representation-based approaches control sensitive information in learned features through explicit objectives or adversarial training (Zemel et al., 2013; Edwards & Storkey, 2015; Beutel et al., 2017; Zhang et al., 2018). FairGAN and related approaches extend adversarial fairness objectives to synthetic data generation (Xu et al., 2018; Sattigeri et al., 2019). These approaches act on different objects: decisions, representations, or generated data. Our prediction setting targets conditional independence of the deployed score, $S_\theta \perp Z \mid Y_{\mathrm{true}}$, so that the regularized quantity is the score used at inference.

Enforcing conditional independence. Kernel methods characterize dependence through feature-space discrepancies, with conditional extensions accounting for a conditioning variable (Gretton et al., 2005; Fukumizu et al., 2007; Zhang et al., 2011). CIRCE develops a non-adversarial conditional-invariance regularizer using an offline conditional-mean embedding regression and provides consistency guarantees under its assumptions (Pogodin et al., 2023). Information-theoretic approaches estimate mutual information through variational or contrastive objectives (Barber & Agakov, 2004; Belghazi et al., 2018; Hjelm et al., 2019); their estimation properties and representation-learning behavior have been studied in Poole et al. (2019); Tschannen et al. (2020); McAllester & Stratos (2020). More directly, Ahuja et al. (2021) characterize CI through divergences between joint and conditionally factorized reference distributions and develop an adversarial implementation. Hwa et al. (2024) impose CI in learned representations, including diffusion-autoencoder latents, using

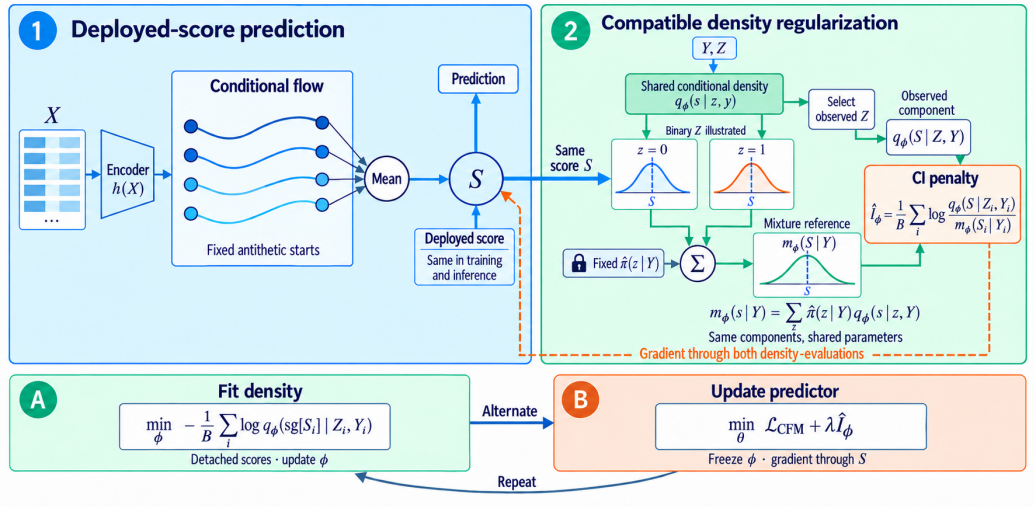

Figure 1: **CI-CFM framework.** The conditional flow predictor averages terminal states from fixed initial conditions to produce the deployed score S. A shared conditional density model $q_\phi(s \mid z, y)$ is fitted to detached scores. For discrete Z, mixing its group-conditional components with fixed training-set weights yields $m_\phi(s \mid y)$. The observed-group component and the mixture are evaluated at the same score to form the log-ratio regularizer. Training alternates between fitting the density model and updating the predictor with ϕ fixed, while gradients pass through both density evaluations into S. The auxiliary density model is not needed at inference.

divergence-based objectives and dynamic sampling. Building on this CI perspective, we construct a likelihood-ratio surrogate from normalized conditional densities and their shared mixture marginal.

Flows and flow matching. Normalizing flows model densities through invertible transformations and change-of-variables likelihoods (Rezende & Mohamed, 2015; Dinh et al., 2017; Papamakarios et al., 2017; Kingma & Dhariwal, 2018). Continuous normalizing flows parameterize these transformations through ordinary differential equations (Chen et al., 2018; Grathwohl et al., 2019). Flow matching (FM) and conditional flow matching (CFM) train vector fields by regression along prescribed probability paths, avoiding numerical trajectory integration in the standard training objective (Lipman et al., 2023; Tong et al., 2024; Liu et al., 2023a; Albergo et al., 2025). The CFM objective learns transport dynamics; conditional-independence requirements enter through an additional objective. CI-CFM uses CFM as the predictive backbone and normalized density models to construct that regularizer.

Positioning of CI-CFM. CI-CFM centers on two design choices: compatible density construction and deployed-score alignment. In the prediction setting, a conditional density models $S_\theta \mid Z, Y_{\text{true}}$, and its reference density is formed by mixing the same components over the label-conditional sensitive-attribute distribution estimated from training data. Evaluating both densities at the same deployed score yields the component-to-mixture log ratio used for prediction regularization. This construction replaces adversarial distribution matching with conditional-density fitting and requires no pairwise kernel evaluations in the predictor-side penalty. Its distinguishing feature is the coupling of the density components, their marginal, and the score being regularized. Appendix A.5 compares the corresponding system-level designs.

3 Method

CI-CFM augments a conditional flow-matching predictor with a density-based conditional-independence regularizer. Given covariates X, the predictor produces a score S. An auxiliary density model estimates how this score is distributed within each label and sensitive group. We compare that conditional density with a label-conditional reference formed by mixing the same components. Their log ratio provides the regularizer used to train the predictor. Figure 1 illustrates the method, and Algorithm 1 summarizes its training procedure.

3.1 Problem Formulation and Theoretical Motivation

Let $\mathcal{D} = \{(X_i, Y_i, Z_i)\}_{i=1}^N$ be i.i.d. observations, where X denotes covariates, Y the ground-truth label, and Z a sensitive attribute. We learn a predictor G_θ with scalar score $S = G_\theta(X)$ satisfying

$$S \perp Z \mid Y. \tag{1}$$

This condition requires the score distribution to agree across sensitive groups within each label. For binary classification, it implies equalized odds under any common deterministic score threshold. We describe the prediction setting here; the feature-generation setting is specified in Appendix C.3.

Conditional mutual information provides a population measure of this dependence:

$$I_\theta(S; Z \mid Y) = \mathbb{E}\left[\log \frac{p_\theta(S \mid Z, Y)}{p_\theta(S \mid Y)}\right]. \tag{2}$$

Here p_θ is the distribution induced by the data and the predictor. The following KL specialization of Ahuja et al. (2021) formalizes this population target.

Theorem 3.1 (Divergence characterization of conditional independence)**.** *Let P_θ be the joint law of (S, Z, Y), and let Q_θ be the conditionally factorized reference with density*

$$q_\theta(s, z, y) = p_\theta(s, y)\nu(z \mid y), \tag{3}$$

where $\nu(z \mid y)$ is a reference conditional density. Write P_{ZY} and Q_{ZY} for their (Z, Y) marginals. Assume that each pair of laws in the KL terms below is mutually absolutely continuous and that both KL divergences are finite. Then

$$\xi_\theta := D_{\mathrm{KL}}(P_\theta \| Q_\theta) - D_{\mathrm{KL}}(P_{ZY} \| Q_{ZY}) = I_\theta(S; Z \mid Y). \tag{4}$$

Consequently, $\xi_\theta \geq 0$, with equality if and only if $S \perp Z \mid Y$.

Proof. The KL chain rule separates the first term into CMI and the second KL term. The full derivation is given in Appendix A.1.

Equation (2) suggests a direct training strategy: estimate the two conditional score densities and use their log ratio as a regularizer. Our key construction is to obtain the denominator by averaging the numerator's conditional components over sensitive groups. When the components and mixing law match their population counterparts, this construction recovers the CMI density ratio. For the exact mixture, Proposition A.3 bounds the difference between the population log-ratio objective and CMI in terms of conditional-density and mixing-law errors. Its assumptions and proof are given in Appendix A.2. This motivates fitting the conditional density alongside the predictor.

3.2 Conditional Flow Matching Backbone and Parameterization

We parameterize G_θ using independent conditional flow matching (Lipman et al., 2023; Tong et al., 2024). For a training pair (X, Y), draw $a_0 \sim \mathcal{N}(0, 1)$ and $t \sim \mathrm{Unif}[0, 1]$, and define

$$a_t = (1 - t)a_0 + tY, \qquad u_t = Y - a_0.$$

An encoder maps X to a context vector $h = \mathrm{Enc}_\theta(X)$. The vector field $v_\theta(a_t, h, t)$ is trained by

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{(X,Y),a_0,t}\left[\|v_\theta(a_t, h, t) - u_t\|_2^2\right].$$

The encoder and vector-field architecture are chosen for the input modality.

To make a prediction, we integrate the learned ODE

$$\frac{da_t}{dt} = v_\theta(a_t, h, t)$$

from fixed Gaussian initial states and average the terminal states. This average defines the score $S = G_\theta(X)$. The CI regularizer is evaluated on this same score during training, so it acts on the quantity used at inference rather than on a separate stochastic endpoint.

3.3 FAIRNESS-CONSTRAINED TRAINING FRAMEWORK

We train the predictor with the combined objective

$$\mathcal{L}_{\text{gen}}(\theta) = \mathcal{L}_{\text{CFM}}(\theta) + \lambda \widehat{I}_\phi(\theta), \tag{5}$$

where $\widehat{I}_\phi$ is the minibatch density-ratio surrogate defined below and λ controls the regularization strength. Both losses are evaluated on minibatches during optimization.

Auxiliary conditional density. We fit a normalized density $q_\phi(s \mid z, y)$ to the predictor's scores. It estimates the score distribution for a given sensitive group and label. For discrete Z, estimate group probabilities $\widehat{\pi}(z \mid y)$ from the training partition and construct

$$m_\phi(s \mid y) = \sum_{z \in \mathcal{Z}} \widehat{\pi}(z \mid y)\, q_\phi(s \mid z, y). \tag{6}$$

The reference m_ϕ therefore averages the group-conditional densities at the same label. It is derived from q_ϕ, not fitted by a second density head. This is the compatible-mixture construction.

For continuous Z, fit $\widehat{p}_\eta(z \mid y)$ to training-set pairs (Z, Y) and replace the sum by

$$m_\phi(s \mid y) = \int q_\phi(s \mid z, y) \widehat{p}_\eta(z \mid y)\, dz.$$

The mixing law is fitted before predictor training and then held fixed; its parameters are suppressed in m_ϕ. We approximate this integral using conditional samples obtained from a fixed base-noise bank and the fitted mixing model. The samples depend on y, but not on s, so the resulting finite mixture remains normalized. The draws depend on y, not on s, so the numerical approximation is also a normalized mixture. Continuous labels are handled through the conditional models.

Minibatch CI regularizer. For scores $S_i = G_\theta(X_i)$, compute

$$\widehat{I}_\phi(\theta) = \frac{1}{B} \sum_{i=1}^{B} \left[\log q_\phi(S_i \mid Z_i, Y_i) - \log m_\phi(S_i \mid Y_i) \right], \tag{7}$$

using the finite-mixture approximation when Z is continuous. Both densities are evaluated at the same prediction score. If the fitted score distributions agree across groups within a label, each component equals their mixture and the log ratio is zero. The penalty thus targets within-label group dependence while the CFM loss preserves the prediction task.

Optimization and gradient flow. Each iteration first updates the density model by maximum likelihood on detached scores:

$$\mathcal{L}_{\text{den}}(\phi) = -\frac{1}{B} \sum_{i=1}^{B} \log q_\phi(\text{sg}[S_i] \mid Z_i, Y_i),$$

where sg stops gradients into the predictor. We then freeze the auxiliary parameters and update θ using (5). The regularizer's gradient passes through both density evaluations into S_i and the prediction ODE solves. Repeating these updates lets the density model track the evolving score distribution. The auxiliary model is trained by likelihood rather than an adversarial objective, and is not needed at inference.

Implementation. The prediction score averages four terminal states from two fixed antithetic pairs of Gaussian initial states. We use the same states and averaging rule in training, validation, and testing. Local synthetic and tabular experiments use eight integration steps; the CFM regression loss uses separately sampled initial states. Continuous-mixture evaluation uses 32 fixed draws with antithetic pairing in the mixing model's base-noise space, separate from the prediction draws. The local binary implementation uses a conditional rational-quadratic spline density and Laplace-smoothed group probabilities. Local confirmatory experiments fix $\lambda = 0.2$, with binary decision thresholds selected on validation data.

Remark 3.2 (Population target and training surrogate). The expected log ratio equals CMI when the conditional densities and mixing law match their population counterparts. During training, predictor updates differentiate the fitted log ratio with the density parameters held fixed. Propositions A.3 and A.2 quantify the effects of density estimation and numerical mixture approximation, respectively.

Table 1: Main synthetic benchmark: comparison within the CFM backbone (ten seeds). The single-endpoint variant changes the training regularization target, not the deployed prediction rule. Values are mean ± sample standard deviation across seeds. Arrows indicate lower is better. Bold values mark the lowest mean in each dependence column among the six CFM-based methods shown. HSIC values are multiplied by 10^3.

Method	RMSE ↓	HSIC ↓ ($\times 10^3$)	$\lvert\rho\rvert$ ↓	BMG ↓	MMD^2 ↓
CFM	0.423 ± 0.014	20.09 ± 6.13	0.625 ± 0.073	0.537 ± 0.048	0.375 ± 0.068
CFM + single-endpoint CI	0.422 ± 0.013	17.81 ± 5.87	0.594 ± 0.077	0.511 ± 0.046	0.348 ± 0.066
CFM + adversary	0.421 ± 0.013	17.96 ± 5.72	0.597 ± 0.075	0.516 ± 0.049	0.350 ± 0.068
CFM + CIRCE	0.423 ± 0.014	19.97 ± 6.09	0.623 ± 0.073	0.536 ± 0.048	0.374 ± 0.067
CFM + InfoNCE	0.423 ± 0.014	19.97 ± 6.07	0.623 ± 0.073	0.536 ± 0.048	0.374 ± 0.067
CI-CFM (ours)	0.420 ± 0.013	$\mathbf{10.92 \pm 5.54}$	$\mathbf{0.477 \pm 0.099}$	$\mathbf{0.432 \pm 0.062}$	$\mathbf{0.256 \pm 0.069}$

4 Simulation Studies

4.1 Experimental Setup

We evaluate how conditional-dependence regularization and deployed-score alignment affect a shared CFM predictor on two continuous-variable regression benchmarks. We report prediction accuracy and conditional-dependence diagnostics. In the main benchmark, a latent signal generates a correlated sensitive attribute Z, a noisy target Y, and four covariates containing noisy measurements of the signal and Z. Ten independently trained seeds use disjoint training, validation, and test sets of sizes 1,024, 1,024, and 2,048. A separate five-seed stress test introduces Laplace noise, a bimodal sensitive attribute, signal–attribute interactions, and a heavy-tailed nuisance coordinate. Its split sizes are 2,048/1,024/2,048, and the fitted mixing model remains affine-Gaussian. Methods and regularization coefficients were fixed before test evaluation. Data generation and training details are given in Appendix B.3.

Baselines and ablation. We compare CI-CFM with unregularized CFM and CFM regularized by a conditional adversary, CIRCE, or conditional InfoNCE. These flow methods use the same fixed four-endpoint prediction score and matched generator-update budgets. A single-endpoint CI variant uses the same compatible density construction but regularizes one stochastic endpoint during training. All regularized methods use the prespecified coefficient $\lambda = 0.2$.

Metrics and reporting. RMSE measures prediction error. Cross-fitted residual HSIC is the primary conditional-dependence diagnostic; absolute linear partial correlation ($|\rho|$), binned mean gap (BMG), and within-bin MMD^2 provide complementary checks. Lower values are preferable for each metric, and none is an exact CMI estimate. Tables report absolute test values as mean ± sample standard deviation across seeds. Paired comparisons use 95% confidence intervals from 10,000 paired-seed bootstrap resamples. Metric definitions appear in Appendix B.2.

4.2 Results

Comparison with baselines. Table 1 compares six CFM-based methods on the main benchmark. CI-CFM has the lowest mean RMSE and all four dependence diagnostics among the flow methods. Its mean RMSE is 0.420, compared with 0.423 for CFM, while residual HSIC decreases from 0.02009 to 0.01092. The adversarial, CIRCE, and InfoNCE baselines have mean HSIC values of 0.01796, 0.01997, and 0.01997, respectively. CI-CFM has lower HSIC than each of these regularizers on all ten paired seeds.

Effect of deployed-score alignment. The single-endpoint variant has mean HSIC 0.01781, versus 0.01092 for CI-CFM: a 38.7% reduction from regularizing the deployed score. The paired difference is -0.00689, with a 95% interval of $[-0.00931, -0.00441]$ from the prespecified bootstrap procedure; nine of ten seeds improve. Mean RMSE decreases from 0.422 to 0.420, and all three secondary dependence diagnostics also decrease. Relative to unregularized CFM, the ratio of mean RMSEs is

Table 2: Non-Gaussian stress test (five seeds), comparing the same six CFM-based methods and a misspecified affine-Gaussian mixing model. Values are mean ± sample standard deviation across seeds. Arrows indicate lower is better. Bold values mark the lowest mean in each dependence column among the six CFM-based methods shown. HSIC values are multiplied by 10^3.

Method	RMSE ↓	HSIC ↓ ($\times 10^3$)	$\lvert\rho\rvert$ ↓	BMG ↓	MMD^2 ↓
CFM	0.443 ± 0.005	2.56 ± 1.28	0.248 ± 0.040	0.187 ± 0.031	0.054 ± 0.013
CFM + single-endpoint CI	0.443 ± 0.004	2.07 ± 1.27	0.222 ± 0.048	0.171 ± 0.035	0.046 ± 0.013
CFM + adversary	0.443 ± 0.005	1.70 ± 0.99	0.207 ± 0.036	0.161 ± 0.029	0.041 ± 0.011
CFM + CIRCE	0.443 ± 0.005	2.55 ± 1.28	0.248 ± 0.040	0.187 ± 0.031	0.054 ± 0.013
CFM + InfoNCE	0.443 ± 0.005	2.48 ± 1.27	0.245 ± 0.040	0.185 ± 0.031	0.053 ± 0.012
CI-CFM (ours)	0.445 ± 0.004	$\mathbf{1.34 \pm 1.05}$	$\mathbf{0.180 \pm 0.051}$	$\mathbf{0.147 \pm 0.031}$	$\mathbf{0.035 \pm 0.013}$

Table 3: Adult test metrics: mean ± sample standard deviation over five paired seeds. Arrows indicate the preferred direction; bold marks the lower mean for each fairness metric. Full metrics and the single-endpoint ablation appear in Table 13.

Method	AUROC ↑	F1 ↑	EO gap ↓	EDDI ↓
CFM	0.9024 ± 0.0019	0.6610 ± 0.0109	0.1206 ± 0.0354	0.0663 ± 0.0012
CI-CFM	0.9004 ± 0.0050	0.6569 ± 0.0068	$\mathbf{0.0628 \pm 0.0086}$	$\mathbf{0.0605 \pm 0.0021}$

0.9921 (paired-seed bootstrap 95% CI: [0.9865, 0.9972]). Thus the alignment benefit is observed in dependence diagnostics without a mean prediction-error penalty in this benchmark.

Implementation and computational details appear in Appendix B.4.

Non-Gaussian stress test. Under mixing-model misspecification, CI-CFM achieves the lowest mean values of all four dependence diagnostics among the flow methods (Table 2). Relative to the single-endpoint variant, mean HSIC decreases from 0.00207 to 0.00134, a 35.1% reduction, with improvements on all five seeds. The paired HSIC difference is -0.000725 (95% bootstrap CI: $[-0.001025, -0.000425]$). This reduction accompanies a small increase in mean RMSE, from 0.443 for unregularized CFM to 0.445 for CI-CFM; the ratio of mean RMSEs is 1.0040 (95% paired-seed bootstrap CI: [1.0014, 1.0065]). Appendix B.5 provides additional paired comparisons and ablations. On both synthetic benchmarks, CI-regularized deterministic predictors yield lower dependence-diagnostic values but higher RMSE than CI-CFM (Appendix B.6).

5 Real-data Evaluation

We evaluate CI-CFM on tabular income prediction, clinical risk prediction, and feature-space image generation. Adult provides a controlled comparison with the same CFM predictor; MIMIC-III/IV compares complete prediction systems across four clinical settings. CIFAR-10 and CelebA assess sample quality under an independent-nuisance control. The prediction tasks regularize the fixed four-endpoint mean score, whereas the vision experiments regularize generated Inception features. Dataset-specific protocols appear in Appendix C.

5.1 Tabular prediction: Adult

We compare CFM and CI-CFM on income prediction with sex as the sensitive attribute, using five paired seeds, 60/20/20 train/validation/test splits, and training-only preprocessing. CI-CFM uses $\lambda = 0.2$; each method selects one global threshold by minimum validation error.

CI-CFM lowers EO gap from 0.1206 to 0.0628 (Table 3); the paired reduction is 0.05776 (95% bootstrap CI: [0.03417, 0.07898]). EO, TPR, FPR, and demographic-parity gaps decrease on all five seeds. Mean EDDI decreases from 0.0663 to 0.0605. AUROC and F1 means decrease slightly;

Table 4: **Performance comparison on MIMIC-III and MIMIC-IV for in-hospital mortality and readmission.** We compare backbone-only predictors, fairness-aware baselines, an adapted diffusion-style fairness baseline, and CI-CFM under the same EHR protocol. All results (%) are reported with 95% confidence intervals. Bold and underlining indicate first- and second-ranked point estimates in the indicated metric direction. Avg. Rank averages ranks across the five metrics within each dataset–task setting, using the displayed point estimates and average ranks for ties; neither ranks nor highlighting imply statistical superiority.

Model	In-Hospital Mortality						Readmission					
	AUROC (↑)	AUPR (↑)	F1 (↑)	EO (↓)	EDDI (↓)	Avg. Rank (↓)	AUROC (↑)	AUPR (↑)	F1 (↑)	EO (↓)	EDDI (↓)	Avg. Rank (↓)
Dataset 1: MIMIC-III												
Transformer Vaswani et al. (2017)	80.49 (78.58, 82.48)	38.54 (33.95, 43.61)	41.09 (37.64, 44.56)	10.09 (8.77, 14.54)	4.13 (3.56, 5.91)	7.6	70.26 (67.85, 72.49)	38.87 (35.09, 43.15)	40.43 (37.51, 43.36)	8.97 (8.55, 13.51)	6.05 (4.67, 8.68)	8.6
LSTM (Graves, 2012)	82.41 (80.55, 84.38)	42.34 (37.66, 47.69)	43.76 (40.56, 47.19)	8.77 (8.53, 13.81)	4.76 (3.98, 6.46)	5.2	72.39 (70.17, 74.71)	39.35 (35.67, 43.51)	41.58 (38.84, 44.5)	8.12 (7.97, 12.95)	6.15 (4.62, 8.57)	6.6
RNN Elman (1990)	81.39 (79.31, 83.50)	43.38 (38.45, 48.06)	43.26 (39.75, 46.98)	8.32 (8.08, 13.56)	4.06 (3.47, 6.02)	4.4	71.60 (69.41, 73.81)	38.93 (35.15, 42.91)	40.88 (37.94, 43.57)	9.62 (8.77, 14.29)	4.63 (3.98, 7.57)	7.6
CNN (LeCun et al., 1998)	82.28 (80.39, 84.17)	**44.19 (39.37, 49.07)**	44.03 (40.35, 47.47)	8.48 (8.33, 13.19)	4.58 (3.69, 6.11)	4.2	74.15 (72.14, 76.30)	42.55 (38.50, 46.78)	43.18 (40.42, 46.00)	9.46 (8.32, 13.75)	5.57 (4.32, 8.29)	4.2
FairEHR-CLP (Wang et al., 2024)	79.70 (77.63, 81.86)	35.83 (31.61, 40.29)	40.35 (36.79, 43.75)	8.46 (8.30, 13.63)	4.32 (3.46, 6.31)	8.3	73.72 (71.41, 75.98)	38.96 (35.05, 43.30)	41.65 (39.09, 43.97)	8.59 (8.28, 14.11)	6.10 (5.26, 9.32)	6.0
FLMD (Liu et al., 2023b)	81.77 (79.74, 83.72)	41.72 (37.13, 46.91)	43.51 (39.87, 47.12)	9.73 (8.95, 14.54)	4.32 (3.80, 5.83)	6.1	73.27 (71.13, 75.48)	41.08 (37.26, 45.40)	42.18 (39.41, 44.84)	7.87 (7.28, 13.26)	5.72 (4.53, 8.17)	4.2
FFVAE (Creager et al., 2019)	82.19 (80.27, 84.06)	41.06 (36.23, 45.85)	41.73 (38.08, 45.26)	7.86 (6.32, 11.08)	3.35 (2.84, 5.42)	4.8	71.84 (69.48, 73.96)	37.50 (33.83, 41.71)	40.57 (37.92, 42.95)	8.05 (7.56, 12.98)	4.53 (3.47, 7.37)	6.6
FarconVAE (Oh et al., 2022)	82.27 (80.21, 84.10)	40.80 (35.96, 45.81)	42.93 (39.17, 46.56)	7.38 (5.10, 15.42)	2.40 (1.50, 4.05)	4.2	73.70 (71.52, 75.83)	39.56 (35.80, 43.81)	42.05 (38.91, 45.14)	8.99 (6.11, 16.04)	3.82 (2.48, 7.00)	4.6
Diffusion-style (Gandikota et al., 2023)	79.60 (77.70, 81.67)	37.24 (32.82, 42.45)	40.93 (37.62, 44.22)	12.22 (10.99, 15.90)	4.23 (3.78, 6.49)	8.8	72.18 (69.91, 74.28)	38.28 (34.58, 42.56)	42.56 (39.43, 45.44)	8.06 (7.81, 12.78)	5.20 (4.14, 7.60)	5.6
CI-CFM (Ours)	**82.73 (80.60, 84.70)**	43.33 (38.60, 48.41)	**44.15 (40.73, 47.87)**	**7.31 (6.99, 11.87)**	**2.22 (2.09, 4.86)**	**1.4**	**74.20 (71.99, 76.41)**	**42.57 (38.74, 46.52)**	**44.15 (41.38, 46.77)**	**6.43 (6.05, 11.54)**	**3.5 (3.17, 6.61)**	**1.0**
Dataset 2: MIMIC-IV												
Transformer Vaswani et al. (2017)	82.35 (80.61, 84.16)	40.98 (36.80, 45.71)	42.94 (39.56, 46.23)	5.11 (4.73, 9.03)	3.02 (2.54, 4.22)	7.6	71.96 (69.92, 73.95)	39.87 (36.55, 43.65)	42.05 (39.22, 45.00)	6.20 (5.43, 8.99)	3.20 (2.70, 5.01)	8.4
LSTM (Graves, 2012)	82.93 (81.08, 84.83)	45.56 (40.89, 49.95)	43.50 (40.08, 46.69)	5.27 (5.22, 8.55)	3.20 (2.71, 4.26)	5.8	73.09 (71.10, 75.00)	43.63 (39.94, 46.85)	42.98 (40.35, 45.39)	7.52 (6.18, 10.63)	3.69 (3.05, 5.66)	6.4
RNN Elman (1990)	82.39 (80.46, 84.08)	43.65 (39.37, 48.06)	44.18 (40.52, 47.79)	5.69 (4.88, 9.29)	2.53 (2.03, 3.52)	6.2	72.66 (70.75, 74.60)	42.32 (38.75, 45.87)	43.20 (40.88, 45.65)	6.00 (5.42, 9.33)	3.78 (2.94, 5.96)	6.4
CNN (LeCun et al., 1998)	83.56 (81.66, 85.27)	**46.65 (42.02, 51.09)**	44.57 (40.93, 48.03)	6.18 (5.30, 8.37)	3.03 (2.50, 4.02)	5.0	73.74 (71.89, 75.66)	42.46 (38.93, 45.94)	44.07 (41.29, 46.46)	7.26 (6.45, 10.41)	3.81 (2.91, 5.75)	5.4
FairEHR-CLP (Wang et al., 2024)	83.45 (81.65, 85.14)	42.07 (37.63, 46.57)	40.48 (36.38, 44.17)	5.84 (5.20, 9.54)	2.35 (2.12, 3.53)	6.6	74.36 (72.39, 76.32)	41.62 (37.93, 45.08)	42.60 (39.73, 45.40)	7.27 (6.20, 9.90)	3.11 (2.46, 4.85)	6.2
FLMD (Liu et al., 2023b)	83.69 (81.78, 85.50)	44.80 (40.16, 49.45)	43.44 (39.47, 47.33)	3.63 (3.46, 7.39)	2.36 (1.92, 3.37)	3.4	74.24 (72.23, 76.14)	42.43 (38.68, 46.25)	42.87 (39.83, 46.03)	4.03 (3.75, 7.01)	2.69 (2.25, 4.22)	3.8
FFVAE (Creager et al., 2019)	82.46 (80.54, 84.28)	41.44 (36.72, 46.22)	42.74 (38.59, 46.48)	4.55 (4.05, 8.02)	2.15 (1.70, 3.18)	5.8	73.20 (71.16, 75.14)	41.99 (38.45, 45.85)	42.37 (39.61, 45.26)	4.47 (3.96, 7.23)	2.70 (2.36, 4.45)	5.8
FarconVAE (Oh et al., 2022)	83.14 (81.30, 84.88)	40.30 (35.66, 45.10)	42.54 (38.74, 45.86)	4.78 (2.92, 7.76)	1.74 (1.24, 2.69)	6.0	73.42 (71.46, 75.37)	42.47 (38.99, 46.01)	44.01 (41.21, 47.08)	4.87 (3.15, 7.51)	2.40 (1.54, 3.47)	3.6
Diffusion-style (Gandikota et al., 2023)	81.24 (79.22, 83.17)	38.95 (34.70, 43.59)	43.21 (39.73, 46.69)	4.29 (4.07, 8.00)	2.48 (2.11, 3.67)	7.0	73.48 (71.45, 75.32)	40.43 (36.68, 44.25)	40.87 (37.83, 44.03)	6.04 (5.20, 10.90)	5.11 (3.46, 5.85)	8.0
CI-CFM (Ours)	**83.88 (82.15, 85.50)**	44.13 (39.46, 48.90)	**45.33 (40.14, 47.52)**	**3.49 (3.17, 5.01)**	**1.66 (1.49, 2.88)**	**1.6**	**74.86 (73.03, 76.81)**	**43.72 (40.24, 47.48)**	**44.72 (41.78, 47.65)**	**3.98 (3.66, 7.13)**	**2.33 (2.13, 4.07)**	**1.0**

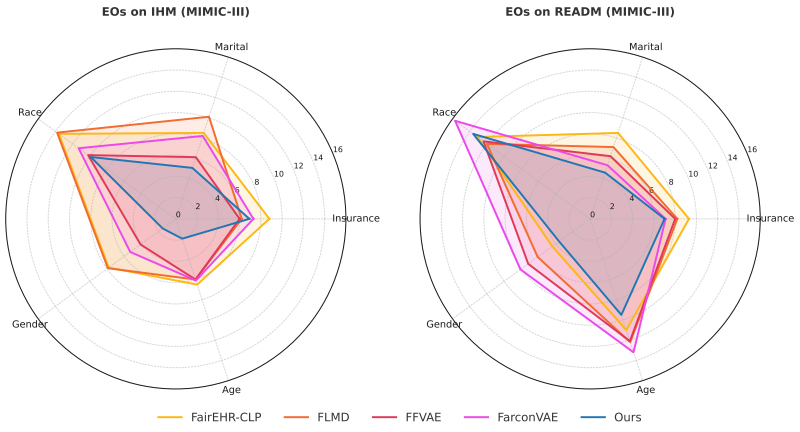

Figure 2: **Attribute-wise equalized-odds comparison on MIMIC-III** for IHM and READM.

their paired difference intervals include zero. Appendix C.2 reports all eight metrics and the score-alignment ablation.

5.2 Clinical prediction: MIMIC-III and MIMIC-IV

We predict in-hospital mortality (IHM) and 30-day readmission (READM) from the first 48 hours of clinical time series. The sensitive attributes are insurance, marital status, race, gender, and age. CI-CFM averages five attribute-specific log-ratio penalties, targeting conditional independence between the prediction score and each attribute given the outcome. We compare CI-CFM with backbone-only predictors, fairness-aware systems, and an adapted diffusion-style baseline under the same EHR protocol.

Predictive performance and fairness. Table 4 reports absolute AUROC, AUPR, F1, EO, and EDDI values for every method, together with 95% patient-cluster bootstrap intervals. CI-CFM has the highest AUROC and F1 and the lowest EO and EDDI point estimates in all four dataset–task settings. CNN has the highest AUPR for mortality on both datasets. Across the four settings, CI-CFM improves AUROC by 0.05–0.50 percentage points and reduces EO and EDDI by 0.05–1.44 and 0.07–0.32 percentage points, respectively, relative to the best competing point estimate for each metric.

Table 5: **Image-generation quality.** All values are means over three runs, using 50,000 generated images per run and EMA weights. Bold marks the best mean for each metric within each dataset.

Dataset	Method	FID ↓	IS ↑	SWD ↓
CIFAR-10	CFM	31.08	5.96	0.21497
	CI-CFM	**30.64**	**6.47**	0.21603
	DDPM	54.77	3.94	0.113
	PGGAN	220.98	1.77	**0.103**
CelebA	CFM	18.601	2.533	0.3733
	CI-CFM	**17.873**	2.614	0.3673
	DDPM	169.400	**2.769**	0.1607
	DCGAN	244.182	1.092	0.3439
	PGGAN	342.682	1.350	**0.1572**

Attribute-wise fairness. Figure 2 shows how the aggregate EO results vary across sensitive attributes on MIMIC-III. Smaller radii indicate smaller group disparities. CI-CFM has shorter radii on most attributes in both tasks, with different improvements across attributes.

5.3 Image generation: CIFAR-10 and CelebA

CFM and CI-CFM share a class-conditional generator in standardized 2048-dimensional Inception-v3 feature space and a decoder to images. An independent Gaussian nuisance, withheld from the decoder, provides a null control for evaluating optimization and sample quality under CI regularization. DDPM and GAN baselines generate images directly in pixel space.

Table 5 reports three-run means for all methods. CI-CFM has the lowest mean FID on both datasets and the highest mean IS on CIFAR-10. DDPM has the highest mean IS on CelebA, while PGGAN has the lowest mean SWD on both datasets, showing that the metrics rank these models differently. Within the shared flow architecture, CI-CFM has lower mean FID and higher mean IS than CFM on CIFAR-10, but slightly higher mean SWD. On CelebA, all three mean metrics favor CI-CFM over CFM. These results show that CI regularization maintains comparable sample quality within the shared flow architecture. Appendix C.3 gives training, sampling, and selection details.

6 Conclusion

CI-CFM couples conditional flow matching with a compatible density-ratio regularizer evaluated on the deployed prediction score. Constructing the reference by mixing the same conditional components connects the training surrogate to population CMI. Controlled experiments show that score alignment improves measured conditional dependence, while Adult and MIMIC evaluations characterize the resulting fairness–utility trade-offs. The vision experiments extend the construction to feature-space generation with comparable sample quality under an independent nuisance.

Limitations & future work. The method relies on estimating conditional score densities, which motivates further study under group imbalance and distribution shift. The vision experiments use an independent nuisance; future work will evaluate real sensitive attributes and extend the approach to structured outputs.

AI USE STATEMENT

We used generative AI tools solely to improve the grammar, clarity, and readability of the manuscript. The authors reviewed and approved all AI-assisted edits and take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

Section 3 specifies the model and training objectives, and Sections 4 and 5 describe the experimental comparisons and evaluation. Appendices A.1 and A.2 provide the theoretical derivations, assumptions, and approximation error analysis; Appendix A.3 gives the training algorithm. Synthetic data generation, training, evaluation metrics, and paired statistical comparisons are detailed in Appendices B.2, B.3, and B.5. Dataset-specific protocols for clinical prediction, Adult, and image generation appear in Appendices C.1, C.2, and C.3, respectively.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018.

Kartik Ahuja, Prasanna Sattigeri, Karthikeyan Shanmugam, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Murat Kocaoglu. Conditionally independent data generation. In *Uncertainty in Artificial Intelligence*, pp. 2050–2060. PMLR, 2021.

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201, 2004.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, pp. 1397–1405. PMLR, 2019.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pp. 1436–1445. PMLR, 2019.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. *Advances in neural information processing systems*, 20, 2007.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2426–2436, 2023.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, and David Duvenaud. Scalable reversible generative models with free-form continuous dynamics. In *International Conference on Learning Representations*, 2019.

Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(96), 2019.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

Jensen Hwa, Qingyu Zhao, Aditya Lahiri, Adnan Masood, Babak Salimi, and Ehsan Adeli. Enforcing conditional independence for fair representation learning and causal image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 103–112, 2024.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

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

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023a.

Zheng Liu, Xiaohan Li, and S Yu Philip. A counterfactual fair model for longitudinal electronic health records via deconfounder. In *2023 IEEE International Conference on Data Mining*, pp. 1175–1180. IEEE, 2023b.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884. PMLR, 2020.

Changdae Oh, Heeji Won, Junhyuk So, Taero Kim, Yewon Kim, Hosik Choi, and Kyungwoo Song. Learning fair representation via distributional contrastive disentanglement. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1295–1305, 2022.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

Roman Pogodin, Namrata Deka, Yazhe Li, Danica J. Sutherland, Victor Veitch, and Arthur Gretton. Efficient conditionally invariant representation learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International conference on machine learning*, pp. 5171–5180. PMLR, 2019.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019.

Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Expert Certification.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yuqing Wang, Malvika Pillai, Yun Zhao, Catherine M Curtin, and Tina Hernandez-Boussard. Fairehr-clp: Towards fairness-aware clinical predictions with contrastive learning in multimodal electronic health records. In *Machine Learning for Healthcare Conference*. PMLR, 2024.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on learning theory*, pp. 1920–1953. PMLR, 2017.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575. IEEE, 2018.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 804–813, 2011.

A Theory and Algorithmic Details

This appendix gives the population identity, approximation results, and training algorithm used in Section 3.1. We distinguish the population CMI from the fitted density-ratio objective and its minibatch estimate.

A.1 Proof of Theorem 3.1

The proof uses the expanded notation $Y_{\text{pred}} = S$ and $Y_{\text{true}} = Y$. Its reference kernel $q(z \mid y) = \nu(z \mid y)$ is distinct from the learned density q_ϕ. The predictor is fixed throughout.

A.1.1 Notation and assumptions

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be an underlying probability space. We work on standard measurable spaces

$$(\mathcal{Y}_{\text{pred}}, \mathscr{Y}_p), \qquad (\mathcal{Z}, \mathscr{Z}), \qquad (\mathcal{Y}_{\text{true}}, \mathscr{Y}_t),$$

with product σ-algebras and reference (dominating) σ-finite measures μ_p, μ_z, μ_t. Write $\mu := \mu_p \otimes \mu_z \otimes \mu_t$ and $\mu_{zt} := \mu_z \otimes \mu_t$. Random variables $Y_{\text{pred}} : \Omega \to \mathcal{Y}_{\text{pred}}$, $Z : \Omega \to \mathcal{Z}$, $Y_{\text{true}} : \Omega \to \mathcal{Y}_{\text{true}}$ are measurable.

Denote by $P_{Y_{\text{pred}}, Z, Y_{\text{true}}}$ the joint law of $(Y_{\text{pred}}, Z, Y_{\text{true}})$ and by $P_{Z, Y_{\text{true}}}$ the (Z, Y_{true}) marginal. Assume absolute continuity and densities:

$$p(y_p, z, y_t) = \frac{dP_{Y_{\text{pred}}, Z, Y_{\text{true}}}}{d\mu}(y_p, z, y_t), \qquad p(z, y_t) = \frac{dP_{Z, Y_{\text{true}}}}{d\mu_{zt}}(z, y_t),$$

exist, and $p(\cdot) > 0$ on its support. By disintegration, there are conditional densities $p(y_p, z \mid y_t)$, $p(y_p \mid y_t)$, $p(z \mid y_t)$ with $p(y_p, z, y_t) = p(y_t)\, p(y_p, z \mid y_t)$, etc., for μ_t-a.e. y_t.

Let $q(\cdot \mid y_t)$ be a Markov kernel on $(\mathcal{Z}, \mathscr{Z})$ such that $q(z \mid y_t)$ is a density (w.r.t. μ_z), strictly positive for μ_t-a.e. y_t. The auxiliary factorized law is defined as

$$Q_{Y_{\text{pred}}, Z', Y_{\text{true}}} \;\text{ with density }\; q(y_p, z', y_t) = p(y_p, y_t)\, q(z' \mid y_t) \quad \text{w.r.t. } \mu,$$

and its (Z', Y_{true})-marginal $Q_{Z', Y_{\text{true}}}$ with density $q(z', y_t) = p(y_t)\, q(z' \mid y_t)$ w.r.t. μ_{zt}. As in Theorem 3.1, assume that each pair of laws in the two KL terms is mutually absolutely continuous and that both KL divergences are finite. Finiteness permits their subtraction; absolute continuity alone does not guarantee it.

For any probability measures $P \ll Q$ on the same measurable space, we use the convention

$$D_{\text{KL}}(P \| Q) = \int \log\left(\frac{dP}{dQ}\right) dP \in [0, \infty],$$

and the information inequality: $D_{\text{KL}}(P \| Q) = 0$ iff $P = Q$ a.s. Conditional mutual information is defined as

$$I(Y_{\text{pred}}; Z \mid Y_{\text{true}}) = \mathbb{E}_{Y_{\text{true}}}\Big[D_{\text{KL}}\big(p(Y_{\text{pred}}, Z \mid Y_{\text{true}}) \,\big\|\, p(Y_{\text{pred}} \mid Y_{\text{true}})\, p(Z \mid Y_{\text{true}})\big)\Big].$$

A.1.2 Theorem and proof

Theorem 3.1 (Equivalence of ξ and Conditional Mutual Information). Under the assumptions above, the population divergence difference

$$\xi = D_{\text{KL}}\big(p(Y_{\text{pred}}, Z, Y_{\text{true}}) \,\big\|\, q(Y_{\text{pred}}, Z', Y_{\text{true}})\big) - D_{\text{KL}}\big(p(Z, Y_{\text{true}}) \,\big\|\, q(Z', Y_{\text{true}})\big)$$

coincides with the conditional mutual information:

$$\xi = I\big(Y_{\text{pred}}; Z \mid Y_{\text{true}}\big).$$

In particular, $\xi \geq 0$, and $\xi = 0$ if and only if $Y_{\text{pred}} \perp Z \mid Y_{\text{true}}$.

Proof. We write y_p for Y_{pred} and y_t for Y_{true} to lighten notation.

By the Radon-Nikodym definitions and $q(y_p, z', y_t) = p(y_p, y_t)q(z' \mid y_t)$,

$$D_{\mathrm{KL}}\big(P_{y_p,z,y_t} \,\|\, Q_{y_p,z',y_t}\big) = \int_{\mathcal{Y}_p\times\mathcal{Z}\times\mathcal{Y}_t} p(y_p, z, y_t)\, \log \frac{p(y_p, z, y_t)}{p(y_p, y_t)\, q(z \mid y_t)}\, d\mu,$$

$$D_{\mathrm{KL}}\big(P_{z,y_t} \,\|\, Q_{z',y_t}\big) = \int_{\mathcal{Z}\times\mathcal{Y}_t} p(z, y_t)\, \log \frac{p(z, y_t)}{p(y_t)\, q(z \mid y_t)}\, d\mu_{zt}. \tag{8}$$

Using Fubini's theorem under the finite-KL assumptions and the identity $p(z, y_t) = \int p(y_p, z, y_t)\, d\mu_p(y_p)$, the marginal KL in (8) can be expressed over the product space (as in the main text):

$$D_{\mathrm{KL}}\big(P_{z,y_t} \,\|\, Q_{z',y_t}\big) = \int_{\mathcal{Y}_p\times\mathcal{Z}\times\mathcal{Y}_t} p(y_p, z, y_t)\, \log \frac{p(z, y_t)}{p(y_t)\, q(z \mid y_t)}\, d\mu.$$

Subtracting the two KL expressions yields

$$\begin{aligned}\xi &= \int p(y_p, z, y_t) \left[\log \frac{p(y_p, z, y_t)}{p(y_p, y_t)\, q(z \mid y_t)} - \log \frac{p(z, y_t)}{p(y_t)\, q(z \mid y_t)}\right] d\mu \\ &= \int p(y_p, z, y_t)\, \log \frac{p(y_p, z, y_t)\, p(y_t)}{p(y_p, y_t)\, p(z, y_t)}\, d\mu,\end{aligned} \tag{9}$$

where the factor $q(z \mid y_t)$ cancels algebraically.

Using the chain rule $p(y_p, z, y_t) = p(y_t)\, p(y_p, z \mid y_t)$, $p(y_p, y_t) = p(y_t)\, p(y_p \mid y_t)$, $p(z, y_t) = p(y_t)\, p(z \mid y_t)$, we rewrite the logarithm as

$$\log \frac{p(y_p, z \mid y_t)}{p(y_p \mid y_t)\, p(z \mid y_t)}.$$

Therefore,

$$\xi = \int_{\mathcal{Y}_t} p(y_t) \left[\int_{\mathcal{Y}_p\times\mathcal{Z}} p(y_p, z \mid y_t)\, \log \frac{p(y_p, z \mid y_t)}{p(y_p \mid y_t)\, p(z \mid y_t)}\, d\mu_p\, d\mu_z\right] d\mu_t(y_t).$$

By the disintegration theorem this is precisely

$$\xi = \mathbb{E}_{Y_{\mathrm{true}}}\Big[D_{\mathrm{KL}}\big(p(Y_{\mathrm{pred}}, Z \mid Y_{\mathrm{true}}) \,\big\|\, p(Y_{\mathrm{pred}} \mid Y_{\mathrm{true}})\, p(Z \mid Y_{\mathrm{true}})\big)\Big] = I\big(Y_{\mathrm{pred}}; Z \mid Y_{\mathrm{true}}\big).$$

Each inner KL divergence is nonnegative, hence $\xi \geq 0$. Moreover, $\xi = 0$ iff $D_{\mathrm{KL}}\big(p(Y_{\mathrm{pred}}, Z \mid Y_{\mathrm{true}}) \| p(Y_{\mathrm{pred}} \mid Y_{\mathrm{true}}) p(Z \mid Y_{\mathrm{true}})\big) = 0$ almost surely in Y_{true}, which (by strict convexity of D_{KL}) is equivalent to $p(Y_{\mathrm{pred}}, Z \mid Y_{\mathrm{true}}) = p(Y_{\mathrm{pred}} \mid Y_{\mathrm{true}}) p(Z \mid Y_{\mathrm{true}})$ a.s., i.e., the conditional independence $Y_{\mathrm{pred}} \perp Z \mid Y_{\mathrm{true}}$. □

A.1.3 Remarks

Population identity and fitted density ratios. For a fixed predictor, the compatible mixture

$$m_\phi(s \mid y) = \sum_z \widehat{\pi}(z \mid y) q_\phi(s \mid z, y)$$

places the component and marginal in a single fitted probability model. The data-induced law need not coincide with that model. The resulting log-ratio estimate can therefore differ from CMI and can be negative in a finite evaluation sample. Proposition A.3 quantifies population value error, while the following results address held-out estimation and numerical mixture approximation.

Evaluation at the same observation. Both terms in a density ratio must be evaluated at the same observation. In contrast, evaluating the numerator under P and the denominator under Q gives $\mathbb{E}_P[\log p] - \mathbb{E}_Q[\log q]$, not $\mathbb{E}_P[\log p - \log q]$. With constant score and label, $Z \sim \mathrm{Bernoulli}(1/2)$ and $Z' \sim \mathrm{Bernoulli}(3/4)$, these quantities are approximately -0.1308 and 0.1438, respectively. Theorem 3.1 instead uses genuine KL divergences evaluated under their first argument.

For the next proposition, write $U = S$, $R = Y$, and let P be the fixed joint law of (U, Z, R). Let $a(u \mid z, r) = p(u \mid z, r)$ and $b(u \mid r) = p(u \mid r)$ be normalized conditional densities. Then $I_P(U; Z \mid R) = \mathbb{E}_P[\log a(U \mid Z, R) - \log b(U \mid R)]$.

Proposition A.1 (Held-out same-point plug-in consistency)**.** *Assume $I = I_P(U; Z \mid R)$ is finite. Let positive normalized conditional density estimates a_n, b_n be fitted using data $\mathcal{D}_n$ independent of an evaluation batch $W_1, \ldots, W_B \overset{\text{iid}}{\sim} P$. Define $f_n(u, z, r) = \log a_n(u \mid z, r) - \log b_n(u \mid r)$ and $\widehat{I}_{n,B} = B^{-1} \sum_i f_n(W_i)$. Suppose, in probability over $\mathcal{D}_n$,*

$$\epsilon_n = \mathbb{E}_P |\log a_n - \log a| + \mathbb{E}_P |\log b_n - \log b| \longrightarrow 0, \qquad \mathbb{E}_P[f_n^2 \mid \mathcal{D}_n] = O_P(1).$$

Then $\widehat{I}_{n,B} - I = O_P(B^{-1/2}) + O_P(\epsilon_n)$ and $\widehat{I}_{n,B} \to I$ in probability as $n, B \to \infty$.

Proof. Conditioning on $\mathcal{D}_n$, the evaluation observations are independent and identically distributed. Conditional Chebyshev's inequality and the tight conditional second moment imply $\widehat{I}_{n,B} - \mathbb{E}_P f_n = O_P(B^{-1/2})$. Also, by the triangle inequality, $|\mathbb{E}_P f_n - I| \le \epsilon_n$. Adding these two errors proves the assertion. □

Proposition A.2 (Finite-mixture value and endpoint-gradient error)**.** *Fix an evaluation pair (s, r) and write $m(s, r) = \mathbb{E}_{Z \sim p_\eta(\cdot \mid r)} q_\phi(s \mid Z, r)$ and $m_K(s, r) = K^{-1} \sum_{k=1}^K q_\phi(s \mid Z_k, r)$. Assume differentiation can be interchanged with integration under the fixed mixing law. Suppose that on the evaluation region the component density satisfies $q_\phi(s \mid z, r) \ge a > 0$, its endpoint gradient has norm at most G, and the conditional second moments*

$$\mathrm{Var}[q_\phi(s \mid Z, r)] \le \sigma_0^2, \qquad \mathbb{E}\|\nabla_s q_\phi(s \mid Z, r) - \nabla_s m(s, r)\|^2 \le \sigma_1^2$$

are finite. For iid mixture draws,

$$\mathbb{E}|\log m_K - \log m| \le \frac{\sigma_0}{a\sqrt{K}},$$

and

$$\mathbb{E}\|\nabla_s \log m_K - \nabla_s \log m\| \le \frac{\sigma_1}{a\sqrt{K}} + \frac{G\sigma_0}{a^2\sqrt{K}}.$$

The same $K^{-1/2}$ order holds for independent antithetic pairs when their pair-level second moments are finite, with pair-level constants.

Proof. The sample means for the density value and its endpoint gradient have expected absolute errors at most $\sigma_0/\sqrt{K}$ and $\sigma_1/\sqrt{K}$ by Cauchy–Schwarz. Since both m and m_K are at least a, the logarithm is $1/a$-Lipschitz. Writing $g = \nabla_s m$ and $g_K = \nabla_s m_K$ gives

$$\left\| \frac{g_K}{m_K} - \frac{g}{m} \right\| \le \frac{\|g_K - g\|}{a} + \frac{\|g\|\,|m_K - m|}{a^2}.$$

The gradient bound implies $\|g\| \le G$; taking expectations proves the claims. Grouping antithetic samples into independent pairs yields the same argument at the pair level. □

Proposition A.2 controls only numerical integration under the fitted mixing model on a region where the density is bounded away from zero. It does not control misspecification of $p_\eta(z \mid r)$, density-estimation error, generator drift, or alternating-optimization convergence.

Proposition A.1 requires log-density consistency and an evaluation sample independent of density fitting. Normalization alone provides neither condition. These fixed-predictor results do not establish convergence or CMI-gradient accuracy during alternating training.

Independence of the choice of q. The equality $\xi = I(Y_{\text{pred}}; Z \mid Y_{\text{true}})$ does not depend on the specific conditional kernel $q(\cdot \mid y_t)$ (as long as the KLs are finite). The auxiliary law q thus serves only as a factorized reference to make the two KL terms well defined.

Integrability and supports. If either absolute-continuity condition fails, the corresponding KL divergence is infinite and the preceding identity should be interpreted only on the common finite-KL regime.

On strict convexity and separability. These assumptions guarantee the information inequality and equality conditions for a broad class of divergences. For D_{KL} they are standard and ensure that $\xi = 0$ iff the conditional factorization holds almost surely.

A.2 Approximation error of the density-ratio surrogate

We make the fixed mixing-law dependence explicit by writing

$$m_{\phi,\eta}(s \mid y) = \int q_\phi(s \mid z, y)\widehat{p}_\eta(z \mid y)\, dz,$$

with sums for discrete sensitive attributes. This is the exact mixture corresponding to m_ϕ in the main text, before numerical integration. Define its population log-ratio value by

$$J(\theta; \phi, \eta) = \mathbb{E}_{p_\theta}\left[\log \frac{q_\phi(S \mid Z, Y)}{m_{\phi,\eta}(S \mid Y)}\right]. \tag{10}$$

The expectation uses the data-induced distribution of (S, Z, Y), not samples from the fitted density.

Proposition A.3 (Error of the compatible density-ratio surrogate)**.** *Fix θ, ϕ, η. Assume conditional densities exist with respect to common dominating measures, the true conditional laws are absolutely continuous with respect to their fitted counterparts, and CMI and the expected KL divergences below are finite. Define*

$$\begin{aligned}
\mathcal{E}_{\mathrm{cond}} &= \mathbb{E}_{Y,Z} D_{\mathrm{KL}}(p_\theta(S \mid Z, Y) \,\|\, q_\phi(S \mid Z, Y))\,, \\
\mathcal{E}_{\mathrm{marg}} &= \mathbb{E}_{Y} D_{\mathrm{KL}}(p_\theta(S \mid Y) \,\|\, m_{\phi,\eta}(S \mid Y))\,, \\
\mathcal{E}_{\mathrm{mix}} &= \mathbb{E}_{Y} D_{\mathrm{KL}}(p(Z \mid Y) \,\|\, \widehat{p}_\eta(Z \mid Y))\,.
\end{aligned}$$

For the exact mixture defined above,

$$J - I_\theta = -\mathcal{E}_{\mathrm{cond}} + \mathcal{E}_{\mathrm{marg}}, \qquad -\mathcal{E}_{\mathrm{cond}} \le J - I_\theta \le \mathcal{E}_{\mathrm{mix}}. \tag{11}$$

In particular, if $\widehat{p}_\eta(Z \mid Y) = p(Z \mid Y)$ almost surely, then

$$0 \le I_\theta - J \le \mathcal{E}_{\mathrm{cond}}.$$

Proof.

Fix the predictor and fitted models under the assumptions of Proposition A.3. All expectations below use the true data-induced law. For compactness, write $I_\theta = I_\theta(S; Z \mid Y)$. Expanding the two log ratios gives

$$\begin{aligned}
J - I_\theta &= \mathbb{E}_{p_\theta} \log \frac{q_\phi(S \mid Z, Y)}{p_\theta(S \mid Z, Y)} + \mathbb{E}_{p_\theta} \log \frac{p_\theta(S \mid Y)}{m_{\phi,\eta}(S \mid Y)} \\
&= -\mathcal{E}_{\mathrm{cond}} + \mathcal{E}_{\mathrm{marg}}.
\end{aligned}$$

The finiteness and absolute-continuity assumptions justify these operations.

For each label y, compare the true and fitted conditional joint densities

$$p_\theta(s, z \mid y) = p(z \mid y)p_\theta(s \mid z, y), \qquad \widetilde{q}_{\phi,\eta}(s, z \mid y) = \widehat{p}_\eta(z \mid y)q_\phi(s \mid z, y).$$

The conditional KL chain rule, averaged over Y, gives

$$\mathbb{E}_Y D_{\mathrm{KL}}(p_\theta(S, Z \mid Y) \,\|\, \widetilde{q}_{\phi,\eta}(S, Z \mid Y)) = \mathcal{E}_{\mathrm{mix}} + \mathcal{E}_{\mathrm{cond}}.$$

Their score marginals are $p_\theta(s \mid y)$ and the exact compatible mixture $m_{\phi,\eta}(s \mid y)$. Marginalization cannot increase KL, hence

$$0 \le \mathcal{E}_{\mathrm{marg}} \le \mathcal{E}_{\mathrm{mix}} + \mathcal{E}_{\mathrm{cond}}.$$

Combining this inequality with the error identity yields

$$-\mathcal{E}_{\mathrm{cond}} \le J - I_\theta \le \mathcal{E}_{\mathrm{mix}}.$$

When the mixing law is exact, $\mathcal{E}_{\mathrm{mix}} = 0$, proving the final claim. These steps specialize the standard KL chain rule and marginalization inequality to the compatible-mixture construction.

The result concerns the exact mixture under a fixed fitted mixing law. For the continuous finite-mixture approximation, the additional difference in the population log-ratio value is

$$\mathbb{E}_{p_\theta}\left[\log m_{\phi,\eta}(S \mid Y) - \log \widehat{m}_{\phi,\eta}(S \mid Y)\right],$$

whenever this expectation exists. Its absolute value is bounded by the expected absolute log-mixture error. The local integration bounds in Proposition A.2 apply on regions satisfying that proposition's assumptions; an integrated bound requires corresponding integrable control. Neither this decomposition nor pointwise integration control implies convergence of alternating training.

Algorithm 1 CI-CFM with a compatible conditional mixture

Require: Training data $\mathcal{D} = \{(X_i, Y_i, Z_i)\}_{i=1}^{N}$; weight λ; integration steps L; fixed four-draw antithetic bank $\mathcal{E}$

Ensure: Predictor G_θ and auxiliary density $q_\phi(s \mid z, y)$

Fit $\widehat{\pi}(z \mid y)$ for categorical Z, or $\widehat{p}_\eta(z \mid y)$ for continuous Z, on training data; freeze the mixing law
Initialize predictor and density parameters θ, ϕ; fix the continuous-mixture base draws when needed
for each training epoch **do**
  for each minibatch $\{(X_i, Y_i, Z_i)\}_{i=1}^{B}$ **do**
    $h_i \leftarrow \mathrm{Enc}_\theta(X_i)$
    Draw $t_i \sim \mathrm{Unif}[0,1]$, $a_{0i} \sim \mathcal{N}(0,1)$, independently of $\mathcal{E}$
    $a_{ti} \leftarrow (1-t_i)a_{0i} + t_i Y_i, \quad u_{ti} \leftarrow Y_i - a_{0i}$
    $\mathcal{L}_{\mathrm{CFM}} \leftarrow B^{-1} \sum_i \|v_\theta(a_{ti}, h_i, t_i) - u_{ti}\|_2^2$
    Integrate each $\epsilon \in \mathcal{E}$ for L steps; average terminal states to obtain $S_i = G_\theta(X_i)$
    $\mathcal{L}_{\mathrm{den}} \leftarrow -B^{-1} \sum_i \log q_\phi(\mathrm{sg}[S_i] \mid Z_i, Y_i)$
    Update ϕ using $\nabla_\phi \mathcal{L}_{\mathrm{den}}$
    Freeze ϕ; retain gradients with respect to S_i
    Compute $m_\phi(S_i \mid Y_i)$ from the same conditional components using an exact sum or fixed Monte Carlo mixture
    $\widehat{I}_\phi \leftarrow B^{-1} \sum_i [\log q_\phi(S_i \mid Z_i, Y_i) - \log m_\phi(S_i \mid Y_i)]$
    Update θ using $\nabla_\theta[\mathcal{L}_{\mathrm{CFM}} + \lambda \widehat{I}_\phi]$
  end for
end for

A.3 Training algorithm

Algorithm 1 summarizes the prediction setting using the notation of Section 3.3. The mixing law is fitted once on training data. Each iteration then fits the conditional density on detached deployed scores and updates the predictor with the density parameters fixed. Dataset-specific warm-up schedules and update frequencies are specified in the experimental protocols.

Prediction and density evaluation. Inference averages four ODE endpoints from the fixed antithetic bank and does not evaluate the auxiliary density. During training, a categorical mixture evaluates one conditional component per group. For continuous attributes, the finite mixture uses fixed draws from the fitted conditional mixing law. Both numerator and denominator are evaluated at the deployed score, with gradients passing through the score into the predictor.

A.4 Why the historical two-head estimator is superseded

The historical implementation evaluated independent free-score networks on different tuples, pairing the generated output with observed Z in one term and a permuted or perturbed Z' in the other. Three failures prevent those experiments from validating the final method. First, a scalar network trained only by increasing its mean output is not a normalized log density: adding a constant to its final bias improves that objective without changing any probability law. Second, even two individually normalized heads need not be compatible. Independently fitted $a(s \mid z, r)$ and $b(s \mid r)$ need not satisfy $b(s \mid r) = \sum_z \pi(z \mid r) a(s \mid z, r)$, so their log difference does not describe a single fitted joint model. Third, an unconditional permutation samples the product $p(z)p(r)$, not the required conditional reference $p(r)p(z \mid r)$ when Z and R are associated. Label-wise permutation is closer for discrete labels but still introduces finite-stratum sampling error, while Gaussian perturbation changes the reference law.

These observations motivate the final same-endpoint compatible mixture; they do not establish that every historical empirical trend was spurious. We retain the old source and outputs for provenance, but exclude their tables, Pareto curves, calibration claims, and head-count ablations from the evidence for the final CI-CFM method.

Table 6: **System-level comparison of fairness enforcement paradigms.** The rows describe representative formulations rather than every variant in each family. A normalized auxiliary density permits log-density evaluation but does not by itself establish calibration or optimization stability.

Paradigm	Primary criterion / objective	Min–max?	Explicit normalized density required?	Typical trade-off
EO constraints / post-processing	Rate parity (e.g., TPR/FPR) via constraints or thresholding	No	No	Efficient inference; may sacrifice utility; depends on constraint handling
MI regularizers	MI penalties (often via variational/contrastive bounds)	No	No	Bound tightness and estimator variance depend on the estimator
Adversarial CI enforcement	Reduce Z-predictability from $(Y_{\mathrm{pred}}, Y_{\mathrm{true}})$ or match joint vs. factorized	Yes	No	Flexible; can be unstable; requires capacity and update balancing
CI-CFM (ours)	Compatible density-ratio regularizer (population identity: CMI)	**No**	**Yes**	Auxiliary density-fitting cost; fixed few-step ODE prediction

A.5 System-Level Comparison of CI/EO Enforcement Paradigms

Table 6 provides a system-level comparison of four families of approaches for enforcing group fairness. We organize the comparison along three axes: (i) the *target criterion* (rate parity vs. conditional independence), (ii) the *optimization form* (constrained/regularized vs. min–max), and (iii) whether the method specifies *an explicit normalized conditional density*; this permits log-density evaluation but does not by itself establish calibration.

EO-style constrained learning and post-processing. Equalized-odds-style methods directly target group-conditional error-rate parity (e.g., TPR/FPR) via constrained ERM or post-hoc thresholding. These procedures can be operationally appealing for classification and keep test-time inference lightweight. However, they typically do not specify a normalized probabilistic model over $(Y_{\mathrm{pred}}, Z, Y_{\mathrm{true}})$, so assessment typically relies on rate gaps, calibration curves, and related predictive metrics.

MI-based regularizers. Information-theoretic methods use mutual-information objectives, often estimated through variational or contrastive formulations. Their interpretation depends on the chosen estimator and its assumptions; a critic need not define a normalized conditional density.

Adversarial CI enforcement. Adversarial methods reduce sensitive-attribute predictability from the score and label, or distinguish joint from conditionally factorized samples. They alternate predictor and discriminator updates. The discriminator defines a prediction or density-ratio objective, without necessarily supplying separate normalized conditional densities.

CI-CFM (ours). At the population level, the divergence characterization of Ahuja et al. (2021) equals the target conditional mutual information $I(Y_{\mathrm{pred}}; Z \mid Y_{\mathrm{true}})$ (Theorem 3.1). The implemented objective uses one normalized group-conditional density and its algebraically compatible finite mixture as a regularizer in a non-adversarial alternating procedure. These experiments do not establish the fitted objective as a calibrated estimator of conditional mutual information. Held-out signed log ratios are retained only as model diagnostics; empirical fairness is assessed with independent held-out metrics. Inference requires four short forward ODE solves from a fixed antithetic bank and their mean, without density evaluation, test-time optimization, or group-specific post-processing.

Dimensionality and computation. In prediction tasks, the auxiliary model estimates the scalar score density conditional on the label and sensitive attributes, rather than the full covariate distribution. This density model is trained separately from the CFM predictor. Its cost depends on conditional-density fitting and the number of mixture components or numerical integration samples. The feature-generation experiments instead model 2048-dimensional features and use the separate protocol in Appendix C.3.

A.6 Finite-sample surrogate diagnostics

We assess the fitted regularizer alongside independent dependence diagnostics rather than interpret its value as exact CMI. For continuous-variable prediction, Appendix B.2 defines residual HSIC, partial correlation, binned mean gap, and within-bin MMD. For binary prediction, Appendix C.2 reports threshold-based group gaps and macro-EDDI together with predictive utility. Held-out signed log ratios provide an additional model diagnostic; a negative value is retained rather than clipped.

The historical sample-size, update-frequency, permutation, and calibration tables use the superseded free-score implementation. They are retained in the source for provenance but excluded from the

current evidence. Calibration and mixture-size studies for the compatible estimator remain further work.

B Experiment details: Simulation (for Section 4)

This appendix specifies the synthetic data mechanism, four dependence diagnostics, and training protocol used in Section 4. It also reports paired comparisons, deterministic-predictor controls, and development ablations.

B.1 Simulation setup: data generation

This subsection specifies the complete synthetic data-generation process. All instantiated numerical values (dimensions, splits, and hyperparameters) used in Section 4 are reported separately in Appendix B.3.

Executed Gaussian mechanism. For each observation, independently draw $H, \epsilon_Z, \epsilon_Y, \epsilon_X, \epsilon_3, \epsilon_4 \sim \mathcal{N}(0,1)$ and set

$$Z = 0.8H + 0.6\epsilon_Z, \qquad Y = H + 0.35\epsilon_Y, \qquad X = (H + 0.2\epsilon_X,\ Z,\ \epsilon_3,\ \epsilon_4). \tag{12}$$

The unregularized predictor can exploit the observed Z coordinate in X, which creates residual prediction–Z dependence after conditioning on the noisy label Y. This is the complete data-generating process used by the locked compatible-mixture experiments; no time-series, Fourier, or discrete group construction is used.

Independent partitions and seeds. Training, validation, and test partitions are generated independently by adding the fixed offsets recorded in the locked manifest to each experiment seed. The confirmatory seeds and partition sizes are stated in Appendix B.3. Randomness from data generation, optimization, and endpoint sampling is controlled by separate deterministic generators in the runner.

B.2 Diagnostics of Conditional Independence

We define the four continuous-variable diagnostics reported in Section 4. The confirmatory and stress-test results use the v2 metric implementation. For these definitions, let $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ denote i.i.d. samples, with $Y_i \in \mathbb{R}$ and $Z_i \in \mathbb{R}^{d_Z}$ a sensitive covariate vector. Let $\widehat{Y}_i$ be the model's prediction for Y_i. We write $\|\cdot\|$ for the Euclidean norm, $\mathrm{med}\{\cdot\}$ for the median, and $\mathbb{1}\{\cdot\}$ for the indicator.

B.2.1 Binned Prediction Mean Gap

The historical runner labels this quantity `EDDIcont_macro`, but its implementation is a binned absolute mean gap, not a kernel MMD and not a distribution-level EDDI. We therefore denote it by $\mathrm{BMG}_{\mathrm{cont}}$ in the manuscript.

Binning. Choose five empirical quantile bins of $\{Y_i\}_{i=1}^n$, removing duplicate boundaries. Intervals are left-closed and right-open, except that the last interval includes the maximum observed label. If all labels coincide, use one bin containing every observation. Write $\mathcal{I}_k$ for the resulting bins.

Within-bin groups. For coordinate j, define $G_{ij} = \mathbb{1}\{Z_{ij} > \mathrm{median}(Z_{\cdot j})\}$ and split each outcome bin into $\mathcal{I}_{kj}^{(g)}$. For a valid bin with at least five observations in each group, compute

$$d_{kj} = \left| \frac{1}{|\mathcal{I}_{kj}^{(1)}|} \sum_{i \in \mathcal{I}_{kj}^{(1)}} \widehat{Y}_i - \frac{1}{|\mathcal{I}_{kj}^{(0)}|} \sum_{i \in \mathcal{I}_{kj}^{(0)}} \widehat{Y}_i \right|.$$

Aggregation. Let $\mathcal{K}_j$ contain the valid bins for coordinate j. We report

$$\mathrm{BMG}_{\mathrm{cont}} = \frac{1}{d_Z} \sum_{j=1}^{d_Z} \frac{1}{|\mathcal{K}_j|} \sum_{k \in \mathcal{K}_j} d_{kj}.$$

The active v2 suite reports this mean gap separately from distributional MMD. Historical values are not backfilled or relabelled.

B.2.2 Within-bin Gaussian MMD^2

For each valid outcome bin and median-split sensitive group, let A and B be the two prediction samples. With a Gaussian kernel whose bandwidth is the median pairwise distance in $A \cup B$, v2 reports the biased statistic

$$\widehat{\mathrm{MMD}}^2(A, B) = \frac{1}{|A|^2} \sum_{a,a' \in A} k(a, a') + \frac{1}{|B|^2} \sum_{b,b' \in B} k(b, b') - \frac{2}{|A||B|} \sum_{a \in A, b \in B} k(a, b).$$

We average over valid bins and then over sensitive coordinates. Unlike BMG, this statistic is sensitive to distributional differences beyond the mean.

B.2.3 Partial Correlation after Residualization

This diagnostic estimates the linear association between $\widehat{Y}$ and each component of Z after removing linear dependence on Y.

Residualization. Fit OLS of $\widehat{Y}$ on Y with intercept: $\widehat{Y}_i = a_0 + a_1 Y_i + r_i^{(\widehat{Y})}$. For each $j \in \{1, \ldots, d_Z\}$, fit OLS of Z_{ij} on Y_i with intercept: $Z_{ij} = b_{0j} + b_{1j} Y_i + r_i^{(Z_j)}$. Collect residuals $\mathbf{r}^{(\widehat{Y})}$ and $\mathbf{r}^{(Z_j)}$.

Partial correlation. The sample Pearson partial correlation between $\widehat{Y}$ and Z_j given Y equals the correlation between residuals:

$$\widehat{\rho}_j = \frac{\sum_{i=1}^n \left(r_i^{(\widehat{Y})} - \overline{r}^{(\widehat{Y})}\right)\left(r_i^{(Z_j)} - \overline{r}^{(Z_j)}\right)}{\sqrt{\sum_{i=1}^n \left(r_i^{(\widehat{Y})} - \overline{r}^{(\widehat{Y})}\right)^2} \sqrt{\sum_{i=1}^n \left(r_i^{(Z_j)} - \overline{r}^{(Z_j)}\right)^2}}.$$

We summarize by the mean absolute partial correlation

$$|\rho|_{\text{partial}} = \frac{1}{d_Z} \sum_{j=1}^{d_Z} |\widehat{\rho}_j|.$$

B.2.4 Residual HSIC

Residual HSIC measures nonlinear association between prediction and sensitive-attribute residuals after adjusting for Y. It is a diagnostic of residual dependence, not an exact measure of conditional mutual information.

Residuals. The historical suite uses the linear residuals from Appendix B.2.3. In v2, the primary HSIC diagnostic instead uses five folds. On each complementary training fold, we fit separate Gaussian-kernel ridge regressions of $\widehat{Y}$ and Z_j on Y, using 128 Nyström features, ridge coefficient 10^{-3}, and a training-fold median bandwidth. Only held-out predictions form the residual vectors. Thus neither the residualizer nor its bandwidth is fitted on the residual being evaluated.

Kernels. For each j, define Gaussian kernels $k_{\sigma_{\widehat{Y}}}(a, b) = \exp(-|a - b|^2/(2\sigma_{\widehat{Y}}^2))$ on $\mathbb{R}$ and $\ell_{\sigma_j}(a, b) = \exp(-|a - b|^2/(2\sigma_j^2))$ on $\mathbb{R}$. Choose bandwidths by the median heuristic: $\sigma_{\widehat{Y}} = \mathrm{med}\{|r_i^{(\widehat{Y})} - r_{i'}^{(\widehat{Y})}| : i \neq i'\}$ and $\sigma_j = \mathrm{med}\{|r_i^{(Z_j)} - r_{i'}^{(Z_j)}| : i \neq i'\}$.

Empirical HSIC (biased). Let $K \in \mathbb{R}^{n\times n}$ with $K_{ii'} = k_{\sigma_{\widehat{Y}}}(r_i^{(\widehat{Y})}, r_{i'}^{(\widehat{Y})})$, $L^{(j)} \in \mathbb{R}^{n\times n}$ with $L^{(j)}_{ii'} = \ell_{\sigma_j}(r_i^{(Z_j)}, r_{i'}^{(Z_j)})$, and $H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. The v2 diagnostic is

$$\mathrm{HSIC}_j = \frac{1}{n^2}\,\mathrm{tr}\big(KHL^{(j)}H\big).$$

It uses the cross-fitted nonlinear residuals described above. The historical implementation used linear residuals and the normalization $(n-1)^{-2}$; those values are not pooled with the v2 results. We report the mean residual HSIC

$$\mathrm{HSIC}_{\text{partial}} \;=\; \frac{1}{d_Z}\sum_{j=1}^{d_Z}\mathrm{HSIC}_j.$$

B.3 Experimental protocol: architectures and training

This subsection specifies the continuous-Z benchmark reported in the main text.

Synthetic data instantiation and splits. For each split and seed we independently draw a latent signal $H \sim \mathcal{N}(0,1)$, set $Z = 0.8H + 0.6\epsilon_Z$ and $Y = H + 0.35\epsilon_Y$, and construct

$$X = (H + 0.2\epsilon_X,\ Z,\ \epsilon_3,\ \epsilon_4),$$

with independent standard-Gaussian noises. Ten seeds are generated before the confirmatory runner is implemented; each has independent training, validation, and test samples of sizes 1,024, 1,024, and 2,048. The endpoint generator uses a two-hidden-layer tanh vector field of width 48. Training uses eight epochs, batch size 256, Adam with learning rate 2×10^{-3}, and eight forward- Euler endpoint steps. The score in every partition averages endpoints from the same fixed four-draw antithetic Gaussian bank.

Conditional-density estimator. The component and mixture densities are evaluated at the same deployed score. For continuous Z, we approximate

$$m_\phi(s \mid y) = \int q_\phi(s \mid z, y)\widehat{p}_\eta(z \mid y)\,dz$$

using a conditional Gaussian mixing model fitted on training data and fixed antithetic Monte Carlo draws. The approximation remains a normalized mixture, with numerical integration error treated separately from density-fitting error in Proposition A.2.

Training and evaluation protocol. The implementation first fits $\widehat{p}_\eta(z \mid y)$ for 120 training-only steps, shares that fit across the prespecified methods, and freezes it. Alternating updates fit the branchless conditional spline $q_\phi(s \mid z, y)$ on detached fixed scores and update the generator through the fixed-density component-minus-mixture ratio. The mixture size is $K = 32$ and $\lambda = 0.2$. Validation cannot select a method or checkpoint. Test evaluation occurs once for all eight fixed methods after 80 training/validation tasks, finite-metric checks, and checkpoint/source hashes pass. The task order is cyclically rotated by seed, and the primary decision uses paired-seed bootstrap inference rather than selecting favorable seeds.

B.4 Computational cost

Density fitting and differentiation through the prediction ODE add training work to CI-CFM. The synthetic comparisons match generator-update budgets, not wall-clock time. On the main ten-seed benchmark, mean standalone training time is 11.04 seconds for CI-CFM, including the mixing-law fit, versus 1.12 seconds for CFM (9.86 times). The adversarial, CIRCE, and InfoNCE alternatives take 4.72, 4.77, and 5.65 seconds, respectively. On the five-seed non-Gaussian stress test, CI-CFM takes 20.15 seconds versus 1.44 seconds for CFM (13.97 times). These are training-time comparisons, not inference-latency measurements. Adult timing is reported in Appendix C.2; vision training budgets are specified in Appendix C.3.

Table 7: Effect of regularizing the deployed score (ten seeds). Negative differences favor the aligned method for every row. Intervals are paired-seed bootstrap 95% intervals from 10,000 resamples; the primary HSIC interval uses the preregistered gate seed.

CI-CFM − single-endpoint variant	Mean paired difference	Paired-seed bootstrap 95% interval
RMSE	-0.00219	$[-0.00420, -0.00036]$
Cross-fitted residual HSIC	-0.00689	$[-0.00931, -0.00441]$
Linear partial correlation	-0.11754	$[-0.16002, -0.07456]$
Binned mean gap	-0.07934	$[-0.10704, -0.05232]$
Within-bin MMD^2	-0.09182	$[-0.12351, -0.05940]$

Table 8: Alternative regularizers on the same CFM predictor (ten seeds). Negative differences favor the first method for every metric. All comparisons use the same fixed deployed score; absolute values appear in Table 1.

Paired comparison	ΔRMSE	ΔHSIC	$\Delta\lvert\rho\rvert$	ΔBMG
CI-CFM − adversary	-0.00101	-0.00704	-0.12033	-0.08353
CI-CFM − CIRCE	-0.00319	-0.00904	-0.14661	-0.10378
CI-CFM − InfoNCE	-0.00326	-0.00904	-0.14664	-0.10357

B.5 Alignment, stress-test comparisons, and development controls

The ten-seed confirmation fixed eight methods, $\lambda = 0.2$, 32 antithetic mixture samples, and the four-draw deployed score after a three-seed validation-only pilot. Validation could check integrity but could not select methods, coefficients, checkpoints, metrics, or thresholds. Test data were created only after all 80 training/validation tasks and checkpoint hashes passed. The primary comparison was aligned compatible versus single-endpoint compatible v1; the other regularizers and deterministic predictors were fixed secondary comparisons. Earlier alignment and cross-fitting candidates failed their prespecified gates and were not promoted. Their results were not reused as independent confirmation of the final method.

The five-seed non-Gaussian migration fixed all eight methods and unsealed test data after 40 training/validation tasks. Relative to the adversary and CIRCE, the final method's mean HSIC differences were -0.000363 and -0.001207, with four and five favorable seeds, respectively.

A separate five-seed 2-by-2 attribution study varied the mixing family (affine/GMM) and regularization schedule (fixed/adaptive). GMM improved mean control NLL but won on only three seeds, below the four-of-five stability criterion. Affine-adaptive minus affine-fixed changed RMSE by $+0.000085$, partial correlation by $+0.00395$, HSIC by $+0.000108$, and BMG by $+0.00405$; all paired intervals crossed zero. Cost increased from 15.9 to 18.5 CFM candidate times. These negative ablations do not motivate replacing the fixed affine mixing model. Historical free-score, head-count, and auxiliary-CNF ablations remain excluded because they do not evaluate the final estimator.

B.6 Deterministic-predictor controls

The main comparison holds the CFM backbone fixed. Here we examine the compatible regularizer on a deterministic predictor, alongside CFM and CI-CFM. The deterministic baseline predicts directly from the covariates and minimizes squared error; Deterministic + CI adds the same component-to-mixture penalty to that predictor. Both controls use the benchmark's fixed splits and seeds, with $\lambda = 0.2$ for the regularized model. Table 10 reports all five metrics for both benchmarks.

Regularizer effect and accuracy trade-off. On the main benchmark, adding the compatible penalty to the deterministic predictor reduces mean HSIC from 0.00666 to 0.00195, while RMSE increases from 0.427 to 0.442. The paired changes are $+0.01528$ in RMSE, -0.00471 in HSIC, -0.20065 in absolute partial correlation, and -0.13207 in BMG. On the stress test, HSIC decreases from 0.00215 to 0.00024 and RMSE increases from 0.442 to 0.447. Thus the regularizer also reduces

Paired comparison	ΔRMSE	ΔHSIC	$\Delta\lvert\rho\rvert$	ΔBMG
Aligned compatible − compatible v1	$+0.00141$	-0.000725	-0.04218	-0.02379
Aligned compatible − CFM	$+0.00175$	-0.001223	-0.06819	-0.04049
Aligned compatible − adversary	$+0.00143$	-0.000363	-0.02659	-0.01366
Aligned compatible − CIRCE	$+0.00174$	-0.001207	-0.06749	-0.04003

Table 9: Paired differences on the non-Gaussian stress test (five seeds). Negative differences favor CI-CFM and positive differences favor the comparator. Absolute values appear in Table 2.

Table 10: Cross-backbone controls on the main benchmark (ten seeds) and non-Gaussian stress test (five seeds). Values are mean ± sample standard deviation; arrows indicate lower is better. HSIC is multiplied by 10^3. Bold values mark the lowest mean in each dependence column within each benchmark panel, not statistical significance. CFM and CI-CFM are repeated from the main tables for direct comparison.

Method	RMSE ↓	HSIC ↓ ($\times 10^3$)	$\lvert\rho\rvert$ ↓	BMG ↓	MMD^2 ↓
Main benchmark					
CFM	0.423 ± 0.014	20.09 ± 6.13	0.625 ± 0.073	0.537 ± 0.048	0.375 ± 0.068
CI-CFM (ours)	0.420 ± 0.013	10.92 ± 5.54	0.477 ± 0.099	0.432 ± 0.062	0.256 ± 0.069
Deterministic	0.427 ± 0.010	6.66 ± 3.17	0.383 ± 0.081	0.401 ± 0.071	0.215 ± 0.062
Deterministic + CI	0.442 ± 0.012	$\mathbf{1.95 \pm 2.16}$	$\mathbf{0.182 \pm 0.122}$	$\mathbf{0.269 \pm 0.096}$	$\mathbf{0.110 \pm 0.066}$
Non-Gaussian stress test					
CFM	0.443 ± 0.005	2.56 ± 1.28	0.248 ± 0.040	0.187 ± 0.031	0.054 ± 0.013
CI-CFM (ours)	0.445 ± 0.004	1.34 ± 1.05	0.180 ± 0.051	0.147 ± 0.031	0.035 ± 0.013
Deterministic	0.442 ± 0.007	2.15 ± 1.28	0.235 ± 0.044	0.185 ± 0.024	0.056 ± 0.010
Deterministic + CI	0.447 ± 0.008	$\mathbf{0.24 \pm 0.18}$	$\mathbf{0.089 \pm 0.027}$	$\mathbf{0.094 \pm 0.009}$	$\mathbf{0.019 \pm 0.006}$

measured dependence on a non-flow predictor. Compared with CI-CFM, Deterministic + CI has lower values for all four dependence diagnostics on both benchmarks, with higher mean RMSE.

Scope of the backbone comparison. These runs compare fixed operating points. The deterministic model uses a squared-error objective, whereas CFM uses a flow-matching objective; setting $\lambda = 0.2$ for both does not match their effective regularization strengths. The results therefore characterize accuracy–dependence trade-offs at the reported settings, rather than establish a ranking at matched accuracy or dependence. Comparing those trade-offs across backbones would require validation-selected regularization sweeps. The evidence in Section 4.2 concerns regularization and deployed-score alignment within CFM, not the necessity of a flow predictor for point regression.

B.7 Additional implementation ablations

The completed development comparisons are summarized in Appendix B.5. Mixture-sample count, held-out calibration, density-update frequency, and predictor–density timescale separation remain additional ablation questions for the compatible estimator. Historical warm-up, auxiliary-CNF budget, and head-count experiments concern the earlier estimator and are retained only in the source.

B.8 Additional Results: Pareto Fronts with EDDI and HSIC

The fixed operating-point comparison appears in Appendix B.6. A regularization sweep for the compatible estimator is needed to compare full accuracy–dependence curves. The historical curves use the superseded estimator and are excluded from the current evidence.

Table 11: Summary statistics of the datasets used.

Dataset	MIMIC-III			MIMIC-IV		
Split	Training	Validation	Test	Training	Validation	Test
Total	12,672	1,833	3,638	15,112	2,188	4,473
Missing rate	72.81%	72.71%	72.73%	71.16%	71.17%	71.23%
Positive (**IHM**)	1,481	237	446	1,702	243	511
Positive (**READM**)	2,268	364	647	2,648	390	814

C Experiment Details: Real-data (for Section 5)

C.1 EHR: datasets, preprocessing, tasks, and metrics

Compatible-mixture implementation. For each of the five sensitive attributes Z_j, the clinical implementation fits a normalized component $q_{\phi_j}(s \mid z_j, y)$ to the deployed risk score s. It does not train a separate marginal head. Instead, it constructs the empirical training-law mixture

$$m_{\phi_j}(s \mid y) = \sum_g \widehat{p}_{\text{train}}(Z_j = g \mid y) q_{\phi_j}(s \mid Z_j = g, y).$$

The five component-minus-mixture log ratios are averaged before multiplication by the CI coefficient. Table 4 gives the full numerical results and uncertainty intervals; Figure 2 shows attribute-wise EO in the main text.

Summary of datasets. We evaluate on two large-scale, real-world EHR datasets: MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023). Both datasets contain de-identified health records of patients admitted to intensive care units (ICUs) or emergency departments at Beth Israel Deaconess Medical Center (BIDMC), comprising multivariate time-series data (e.g., vital signs, lab tests) and demographic information.

Following established preprocessing protocols (Harutyunyan et al., 2019; Wang et al., 2024), we extract 26 continuous-valued clinical variables from MIMIC-III and 25 from MIMIC-IV, sampled hourly during the first 48 hours of ICU admission. Table 11 summarizes dataset statistics: we obtain 18,143 ICU stays from MIMIC-III and 21,773 from MIMIC-IV, and randomly split each dataset into training, validation, and test sets using a 7:1:2 ratio.

We extract five categories of continuous-valued clinical predictors from the MIMIC-III and MIMIC-IV datasets: vital signs, blood gases, renal function, metabolic panel, and hematology. We focus exclusively on the first 48 hours of patient data recorded after ICU admission, sampling observations at hourly intervals. Admissions with fewer than 48 hours of recorded data are excluded. Detailed predictor information is summarized in Table 12. We preprocess the raw data using the pipeline proposed by (Harutyunyan et al., 2019). Notably, due to the extreme sparsity of arterial oxygen pressure data in MIMIC-IV, we include this predictor only for the MIMIC-III dataset. Consequently, the total number of predictors is 26 for MIMIC-III and 25 for MIMIC-IV.
We also extract five sensitive attributes from the MIMIC-III and MIMIC-IV datasets: insurance type, marital status, race, gender, and age. Each attribute contains several subgroups, whose compositions vary between the two datasets. For instance, the insurance attribute comprises five subgroups (*Medicare, Medicaid, Government, Self Pay, and Private*) in MIMIC-III, whereas it includes only three subgroups (*Medicare, Medicaid, and Other*) in MIMIC-IV due to differences in recording standards.

Implementation Details and Hyperparameters. The clinical experiments use Python 3.11, PyTorch 2.0, and a single NVIDIA RTX-4090 GPU. Models are trained with Adam, batch size 128, and a maximum of 100 epochs. Training stops when validation AUROC does not improve for 10 consecutive epochs. The checkpoint with the highest validation AUROC is evaluated on the test set.

For CI-CFM, a six-layer Transformer encoder with two attention heads and hidden dimension 128 encodes the temporal input; a five-layer MLP with hidden dimension 128 parameterizes the vector field. The single deployed scalar risk score is the mean of four ODE endpoints obtained from two

Table 12: Summary of clinical predictors in longitudinal data for MIMIC-III/IV datasets, * indicates the predictor is only available in MIMIC-III.

Category	Predictors
Vital Signs	Heart Rate, Systolic Blood Pressure, Diastolic Blood Pressure, Mean Blood Pressure, Respiratory Rate, Body Temperature, Oxygen Saturation
Blood Gases	Arterial Base Excess, Arterial Carbon Dioxide Pressure, Arterial Oxygen Pressure*, Arterial pH
Renal Function	Blood Urea Nitrogen, Creatinine
Metabolic Panel	Ionized Calcium, Serum Chloride, Serum Glucose, Fingerstick Glucose, Anion Gap, Serum Bicarbonate, Magnesium, Serum Potassium, Serum Sodium
Hematology	Serum Hematocrit, Hemoglobin, Platelet Count, White Blood Cell Count

prespecified Gaussian base draws and their antithetic pairs. We use eight fixed-step Euler updates. The draw bank, integrator, and averaging rule are identical in training, validation, and test; test-time sampling introduces no new random base draw.

Each predictor minibatch triggers one density update followed by one predictor update. The density step maximizes conditional likelihood on a detached copy of the four-endpoint mean score and cannot update the predictor. The predictor step freezes all density parameters, evaluates every numerator and compatible-mixture denominator at that same score, and retains gradients only through the score. The update ratio is therefore 1:1, with `density_update_interval=1`. The CI coefficient is $\lambda = 0.2$ on the arithmetic mean of the five attribute-wise log ratios. During epochs 1–5 the CI coefficient is zero while predictor and density fitting continue; it increases linearly from epochs 6–10 to 0.2 and remains fixed thereafter.

Each sensitive attribute is represented as an unordered categorical variable using a one-hot or categorical embedding, never an ordinal integer code. Missing or training-unseen categories map to an explicit Other/Unknown level. Age uses four training-fixed groups: 18–44, 45–64, 65–79, and at least 80 years; test data never redefine these bins. For every method–dataset–task combination, a single global threshold maximizing validation F1 is selected and then frozen. The numerical threshold may differ across methods, but the selection rule is shared; within a fitted model it is shared across all attributes and groups for test F1, EO, and EDDI.

Fairness metrics. Let $\mathcal{A}$ denote the five sensitive attributes and $\mathcal{G}_a$ the predefined groups of attribute a. All rates below are computed on the test set using the validation-selected global threshold. For each group, let

$$\mathrm{TPR}_{a,g} = \widehat{\Pr}(\widehat{Y} = 1 \mid Y = 1, Z_a = g), \qquad \mathrm{FPR}_{a,g} = \widehat{\Pr}(\widehat{Y} = 1 \mid Y = 0, Z_a = g),$$

where $\widehat{\Pr}$ denotes the empirical proportion.

For the clinical experiments, we use an averaged equalized-odds gap:

$$\mathrm{EO}_a = \frac{1}{2\binom{|\mathcal{G}_a|}{2}} \sum_{\substack{g,h \in \mathcal{G}_a \\ g<h}} \left(|\mathrm{TPR}_{a,g} - \mathrm{TPR}_{a,h}| + |\mathrm{FPR}_{a,g} - \mathrm{FPR}_{a,h}| \right).$$

Each unordered group pair is counted once.

For EDDI, define the overall and group-specific error rates as

$$\mathrm{OER} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\widehat{Y}_i \neq Y_i\}, \qquad \mathrm{ER}_{a,g} = \frac{1}{N_{a,g}} \sum_{i: Z_{i,a} = g} \mathbf{1}\{\widehat{Y}_i \neq Y_i\},$$

where N is the test-set size and $N_{a,g}$ is the number of test observations in group g. The attribute-specific EDDI is

$$\mathrm{EDDI}_a = \frac{1}{|\mathcal{G}_a|} \sum_{g \in \mathcal{G}_a} \frac{|\mathrm{ER}_{a,g} - \mathrm{OER}|}{\max(\mathrm{OER}, 1 - \mathrm{OER})}.$$

The reported clinical fairness metrics are the unweighted macro-averages across attributes:

$$\mathrm{EO}_{\mathrm{MIMIC}} = \frac{1}{5} \sum_{a \in \mathcal{A}} \mathrm{EO}_a, \qquad \mathrm{EDDI}_{\mathrm{MIMIC}} = \frac{1}{5} \sum_{a \in \mathcal{A}} \mathrm{EDDI}_a.$$

Each attribute therefore receives equal weight. Table 4 reports both metrics multiplied by 100.

We form 95% percentile intervals from 10,000 patient-cluster bootstrap samples of the test set, retaining all ICU stays of each resampled patient. The frozen validation threshold is not reselected inside a bootstrap draw. Within each replicate, we recompute all subgroup TPRs, FPRs, and error rates, together with the overall error rate, and then recompute the attribute-specific metrics and their five-attribute macro-averages. Because aligned patient-level predictions across methods are unavailable, Table 4 reports method-specific intervals and point-estimate rankings rather than paired superiority tests.

For all baselines, we concatenate time series data and its missing mask (0 for observed, 1 for missing) as input. For methods involving adversarial fine-tuning, we initiate low-rank fine-tuning after pre-training the MINE module for 30 epochs. Hyperparameter tuning is performed via grid search on the validation set with the following search spaces:

- Dropout ratio: $\{0, 0.1, 0.2, 0.3\}$
- Learning rate: $\{1 \times 10^{-4}, 5 \times 10^{-5}, 1 \times 10^{-5}\}$
- Mutual information regularization coefficient λ_{MI}: $\{0.1, 0.2, 0.5\}$

Detailed Descriptions of Baseline Methods.

- **CNN** (LeCun et al., 1998): Uses convolutional layers to extract predictive features.
- **RNN** (Elman, 1990): Processes the clinical sequence through recurrent hidden-state updates.
- **LSTM** (Graves, 2012): Uses gated recurrent states to model temporal dependencies.
- **Transformer** (Vaswani et al., 2017): Uses self-attention to model dependencies across the clinical sequence.
- **FFVAE** (Creager et al., 2019): Uses a variational representation-learning objective to disentangle sensitive attributes for downstream prediction.
- **FarconVAE** (Oh et al., 2022): Combines variational representation learning with a contrastive objective.
- **FairEHR-CLP** (Wang et al., 2024): Uses generated counterfactual patient data and contrastive learning to reduce demographic prediction disparities.
- **FLMD** (Liu et al., 2023b): Models latent confounders in counterfactual prediction.

The EHR model below is a task-specific adaptation.

C.1.1 Additional comparison to a diffusion-style fairness baseline

To broaden the EHR baseline family beyond representation-learning and clinical fairness models, we additionally include a diffusion-style baseline adapted from the negative-guidance concept-erasure idea of Gandikota et al. (2023). Since Gandikota et al. (2023) studies text-to-image diffusion rather than fairness-aware EHR prediction, this baseline should be understood as a task-specific adaptation to our prediction-space setting, not as a literal reproduction of the original method.

Adapted protocol. We instantiate a conditional diffusion model over the scalar prediction variable (risk logit), conditioned on the EHR history X. Following the teacher–student negative-guidance idea of Gandikota et al. (2023), we keep a frozen teacher model and fine-tune a student model using two conditioning branches for each sensitive attribute Z: a subgroup-conditioned branch that includes the observed subgroup information, and a subgroup-neutral branch that removes subgroup information from the conditioning input. The student is trained to follow the subgroup-neutral teacher signal while moving away from the subgroup-conditioned one, thereby suppressing subgroup-specific influence on the generated prediction. We use a single adapted configuration throughout, rather than introducing separate ESD-style variants. For a controlled comparison, we match this adapted baseline to CI-CFM in preprocessing, train/validation/test splits, prediction targets, fairness metrics, early stopping, and model-selection protocol.

Table 4 reports the quantitative comparison: CI-CFM has higher AUROC point estimates and lower EO and EDDI point estimates than the adapted diffusion-style comparator in all four MIMIC task–dataset combinations. These rankings do not establish statistical superiority.

C.2 Adult: protocol and additional results

We evaluate the label-conditional criterion $S \perp Z \mid Y$ on Adult. The three fixed methods are CFM, the single-endpoint compatible variant, and CI-CFM. They share preprocessing, data splits, the threshold-selection rule, and evaluation metrics. Earlier adversarial results are excluded because the retained artifacts do not establish a matched five-seed protocol.

Dataset and protocol. The prediction task is income classification and the sensitive attribute is sex. We use a frozen public-source snapshot (OpenML data ID 1590, version 2) and retain its hash and cleaning metadata. Each of five seeds uses a deterministic 60/20/20 train/validation/test split. Preprocessing is fitted on training data only. CFM, the single-endpoint variant, and CI-CFM are fixed before training, both regularized methods use $\lambda = 0.2$, and validation selects one global classification threshold by minimum error. The test split is evaluated after training and threshold selection. We report AUROC, accuracy, and F1 for utility, and EO, TPR, FPR, demographic-parity, and macro-EDDI gaps for fairness.

Implementation details. Adult has a binary sensitive attribute. An MLP parameterizes the CFM vector field, and a conditional rational-quadratic spline models $q_\phi(s \mid z, y)$. Training-only estimates $\widehat{\pi}(z \mid y)$ define the compatible mixture

$$m_\phi(s \mid y) = \sum_z \widehat{\pi}(z \mid y) q_\phi(s \mid z, y).$$

The predictor minimizes $\mathcal{L}_{\mathrm{CFM}} + \lambda \widehat{I}_\phi$, with density parameters fixed during its update. CI-CFM evaluates both density terms at the fixed four-endpoint mean score used in training, validation, and test. The single-endpoint variant changes the regularization target, not the deployed prediction rule. All methods share the optimizer family and stopping budget. The held-out signed log ratio is reported but does not determine run eligibility.

Results. Absolute Adult results are reported in Table 13. Table 14 supplements them with paired differences and bootstrap intervals. Relative to CFM, CI-CFM lowers Adult EO, TPR, FPR, and DP gaps on all five seeds. The paired Adult AUROC, accuracy, and F1 intervals include zero. CI-CFM has the lowest mean Adult EDDI among the three methods. Its paired EDDI difference from CFM is -0.00588 (95% bootstrap CI: $[-0.00681, -0.00498]$). Relative to the single-endpoint variant, CI-CFM lowers Adult EO by 0.0256 and F1 by 0.0086. These results identify a fairness–utility trade-off on Adult.

Effect of score alignment. Relative to the single-endpoint variant, CI-CFM lowers mean EO gap from 0.0884 to 0.0628, while mean F1 decreases from 0.6656 to 0.6569. The corresponding paired differences are -0.02555 (95% CI: $[-0.03832, -0.01323]$) for EO and -0.00863 (95% CI: $[-0.01374, -0.00429]$) for F1. CI-CFM also lowers mean EDDI from 0.0658 to 0.0605. Thus score alignment improves the threshold-based group gaps on Adult, with a small F1 cost relative to that variant.

Table 13: Absolute test metrics on Adult (five paired seeds). Values are mean ± sample standard deviation on the 0–1 scale. Arrows indicate the preferred direction. Bold marks the lowest mean for each fairness metric among the three methods, not statistical significance. Paired differences and intervals are reported below.

Method	AUROC ↑	Accuracy ↑	F1 ↑	EDDI ↓
CFM	0.9024 ± 0.0019	0.8469 ± 0.0013	0.6610 ± 0.0109	0.0663 ± 0.0012
CFM + single-endpoint CI	0.9025 ± 0.0023	0.8468 ± 0.0010	0.6656 ± 0.0110	0.0658 ± 0.0017
CI-CFM (ours)	0.9004 ± 0.0050	0.8464 ± 0.0008	0.6569 ± 0.0068	$\mathbf{0.0605 \pm 0.0021}$
Method	**EO gap ↓**	**TPR gap ↓**	**FPR gap ↓**	**DP gap ↓**
CFM	0.1206 ± 0.0354	0.1183 ± 0.0396	0.0807 ± 0.0107	0.1871 ± 0.0178
CFM + single-endpoint CI	0.0884 ± 0.0200	0.0832 ± 0.0288	0.0774 ± 0.0104	0.1811 ± 0.0168
CI-CFM (ours)	$\mathbf{0.0628 \pm 0.0086}$	$\mathbf{0.0331 \pm 0.0268}$	$\mathbf{0.0628 \pm 0.0086}$	$\mathbf{0.1590 \pm 0.0128}$

Adult metric	Paired difference	95% bootstrap interval
AUROC ↑	-0.00195	$[-0.00473, 0.00060]$
Accuracy ↑	-0.00049	$[-0.00159, 0.00057]$
EO gap ↓	-0.05776	$[-0.07898, -0.03417]$
$\text{EDDI}_{\text{macro}}$ ↓	-0.00588	$[-0.00681, -0.00498]$

Table 14: Paired CI-CFM minus CFM test differences on Adult over five seeds. Brackets are paired-seed bootstrap 95% intervals. Negative fairness differences and positive utility differences favor CI-CFM. Absolute metrics appear in Table 13.

Computational cost. The earlier five-seed Adult experiment recorded mean training times of 109.89 seconds for CI-CFM and 11.78 seconds for CFM (9.32 times). These timings describe that run, not a new timing measurement for the updated Adult results.

C.3 Vision: feature-space feasibility protocol

The vision experiments apply compatible-mixture regularization to generated Inception features. The generator maps initial noise ϵ and class label c to a feature $h = G_\theta(\epsilon, c)$, which a class-conditional decoder maps to an image. The auxiliary nuisance Z is sampled independently of ϵ and c and is used only to condition the density model $q_\phi(h \mid z, c)$. Neither the generator nor the decoder receives Z. This construction provides a conditional-independence null for evaluating the effect of CI regularization on sample quality.

CIFAR-10 representation and decoder. A fixed ImageNet-pretrained Inception-v3 maps each image to its 2,048-dimensional global-average-pooling feature. Each feature dimension is standardized using training-set statistics, and CFM generates this standardized feature h, not pixels. A deterministic class-conditional convolutional decoder $x = g(h, c)$ receives the generated feature and a class embedding, but never the nuisance variable. Its loss weights are 1.0 for L_1, 2.0 for Inception feature matching, 0.2 for VGG perceptual loss, and 0.25 for SWD; its learning rate is three times the generator learning rate.

Continuous-nuisance compatible mixture. The null nuisance is $Z \in \mathbb{R}^{32}$ with $Z \sim \mathcal{N}(0, I_{32})$, independent of image and class; c is the ten-way CIFAR-10 label. A single normalized component $q_\phi(h \mid z, c)$ is fitted to detached generated features. Because $p(z \mid c) = \mathcal{N}(0, I)$ is known, a fixed bank of 32 Gaussian draws arranged in antithetic pairs constructs

$$m_\phi(h \mid c) \approx \frac{1}{32} \sum_{k=1}^{32} q_\phi(h \mid z_k, c).$$

The signed component-minus-mixture log ratio uses the same generated feature h in numerator and denominator. Density parameters are frozen during its generator update, and the coefficient is $\lambda_{\text{vision}} = 0.01$.

Updates and compute budget. A CI event occurs every 8,192 generator updates. Each event performs 64 density minibatch steps before one generator step carrying the CI penalty. During the first 300 GPU-seconds, the generator, decoder, and density are fitted without applying that penalty. Every CIFAR-10 run then uses a fixed total budget of 1,200 GPU-seconds, batch size 32, generator learning rate 10^{-4}, model capacity 96, and the same decoder configuration for CFM and CI-CFM. EMA uses a 60-second time constant with per-update weight capped at 0.001; validation and test report EMA weights only.

CIFAR-10 selection, sampling, and reporting. Classifier-free conditioning dropout is 10%. Guidance 2.75 is selected using only the seed-2026 validation screen and then frozen for seeds 2027 and 2028. Within each run, the checkpoint minimizes the validation average rank of FID, negative IS, and SWD, with FID as the tie-breaker; no metric-specific or test-selected checkpoint is allowed. Each selected EMA checkpoint generates exactly 50,000 images using a separate fixed sampling seed and 100-step Heun ODE integration, with 5,000 images per CIFAR-10 class. FID uses fixed statistics of the official 10,000-image test split; IS uses ten splits, and SWD is multi-scale. The three seeds 2026, 2027, and 2028 are summarized by mean and sample standard deviation.

CelebA protocol. The CelebA study uses aligned images with the fixed crop coordinates $(25, 57, 153, 185)$, resized to 128×128. The conditioning label is the binary *Smiling* attribute. The CFM generator operates in the fixed 2,048-dimensional global-average-pooling representation of ImageNet-pretrained Inception-v3, standardized coordinate-wise with CelebA training-split statistics. A deterministic class-conditional convolutional decoder maps (h, c) to 128×128 pixels and never receives the nuisance variable. The nuisance is $Z \in \mathbb{R}^{32}$ with $Z \sim \mathcal{N}(0, I_{32})$, independent of the image and *Smiling* label. A single normalized component is combined with 32 fixed antithetic Gaussian draws (16 pairs):

$$m_\phi(h \mid c) \approx \frac{1}{32} \sum_{k=1}^{32} q_\phi(h \mid z_k, c),$$

with numerator and denominator evaluated at the same generated feature h. The signed compatible-mixture coefficient is $\lambda_{\text{CelebA}} = 0.01$. Every 1,024 generator steps, an update event first fits the density on detached generated features for 64 minibatch steps, then freezes the density and takes one generator step with the CI penalty. The first 600 GPU-seconds fit generator, decoder, and density without applying that penalty, while density fitting continues normally. Each model uses 2,400 GPU-seconds, batch size 32, generator learning rate 10^{-4}, and decoder learning-rate multiplier 3. EMA has a 60-second time constant with per-update weight capped at 0.001; selection and evaluation use EMA weights. Classifier-free conditioning dropout is 10%, guidance scale is 1.5, and sampling uses 100-step Heun integration.

For CelebA, the checkpoint is selected on the official validation split by the average rank of FID, negative IS, and SWD, with FID as tie-breaker; the test split is never used for selection. The main results are means over three runs. From each run's validation-selected EMA checkpoint we generate exactly 50,000 images, with 25,000 for each *Smiling* value. FID, IS, and multi-scale SWD use the same fixed implementation as the other vision rows; FID uses the official 19,962-image CelebA test partition. Each run's selected checkpoint is evaluated once on the test set.

The complete test comparison is presented in Table 5 in the main text. All entries in that table are three-run means. The recorded sample standard deviations for CIFAR-10 CFM are 2.69 (FID), 0.75 (IS), and 0.00082 (SWD); those for CI-CFM are 1.78, 0.28, and 0.00409, respectively. Across-run standard deviations were not recorded for the other entries.