Toward Individual Fairness Without Centralized Data: Selective Counterfactual Consistency for Vertical Federated Learning

DAWOOD WASIF, Virginia Tech, USA
CHANDAN K. REDDY, Virginia Tech, USA
TERRENCE J. MOORE, U.S. Army Research Laboratory, USA
JIN-HEE CHO, Virginia Tech, USA

When algorithmic decisions depend on data distributed across institutions, how can we ensure that an individual's outcome does not change arbitrarily based on a protected attribute? We study this question in vertical federated learning (VFL), where features are split across parties, sensitive attributes may be private, and proxies for protected characteristics can be scattered across institutional boundaries under strict privacy constraints. Our focus is on individual-level counterfactual stability, i.e., per-instance prediction consistency under protected-attribute interventions as formalized in the causal fairness literature, rather than group parity guarantees such as demographic parity or equalized odds. We propose SCC-VFL, a server-centric framework for enforcing selective counterfactual consistency (SCC) at the individual level in VFL. SCC-VFL operationalizes a given policy specification by combining three components: (i) differentially private, graph-free discovery of feature roles into non-descendants, policy-permitted mediators, and impermissible proxies using only a formally private sketch of the sensitive attribute, with a formal per-release privacy that does not extend to the full training pipeline; (ii) masked counterfactual generation that edits only mediators while fixing non-descendants and suppressing proxy leakage; and (iii) server-side enforcement via an SCC consistency loss that penalizes impermissible prediction changes under protected-attribute interventions. Across three real-world datasets spanning credit, healthcare, and criminal justice, SCC-VFL maintains or improves predictive accuracy while sharply reducing decision flip rates by up to 98% relative to strong baselines. It also lowers attribute-inference attack success and improves robustness, demonstrating favorable utility-fairness-privacy trade-offs in realistic VFL deployments.

CCS Concepts: • **Computing methodologies** → **Machine learning**; *Learning paradigms*; • **Security and privacy** → *Privacy protections*; Differential privacy; Privacy-preserving protocols; • **Applied computing** → Law, social and behavioral sciences.

Additional Key Words and Phrases: vertical federated learning, individual fairness, counterfactual consistency, differential privacy, algorithmic accountability

ACM Reference Format:
Dawood Wasif, Chandan K. Reddy, Terrence J. Moore, and Jin-Hee Cho. 2026. Toward Individual Fairness Without Centralized Data: Selective Counterfactual Consistency for Vertical Federated Learning. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 35 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' Contact Information: Dawood Wasif, dawoodwasif@vt.edu, Virginia Tech, Alexandria, Virginia, USA; Chandan K. Reddy, reddy@cs.vt.edu, Virginia Tech, Alexandria, Virginia, USA; Terrence J. Moore, terrence.j.moore.civ@army.mil, U.S. Army Research Laboratory, Adelphi, Maryland, USA; Jin-Hee Cho, jicho@vt.edu, Virginia Tech, Alexandria, Virginia, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

FAccT '26, Montreal, QC, Canada
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Do sensitive attributes change the decision? Algorithmic decision systems increasingly shape high-stakes outcomes in credit, hiring, healthcare, and criminal justice, directly determining who receives resources and opportunities. This raises expectations that decisions must be accurate, defensible, and socially legitimate [1]. A core requirement is fairness: An individual's outcome should not change simply because a protected attribute such as race, gender, or age changes, or because the model relies on an impermissible proxy. Yet many deployed systems satisfy population-level fairness constraints while still permitting decision changes for specific individuals under hypothetical changes to protected attributes [30]. Although the community has developed a substantial fairness toolkit [1], much of it emphasizes group-level criteria such as demographic parity and equalized odds [16, 21], alongside individual notions of treating similar individuals similarly [13]. These criteria can be mutually incompatible in practice, forcing explicit trade-offs that are often opaque to affected individuals [4, 29]. As accountability demands intensify, there is growing need for individual-level protections that stakeholders can understand, audit, and contest [1]. Following the causal fairness literature [7, 30], we use *individual fairness* throughout this paper to denote per-instance counterfactual stability under a declared policy specification, distinct from group-level parity constraints such as demographic parity or equalized odds.

Practical motivation. Consider a credit consortium where Bank A holds account histories, Employer B holds income records, and Bureau C holds third-party risk indicators. These parties jointly train a lending model via VFL without centralizing sensitive data. But how can they ensure that a 25-year-old applicant receives a decision stable under a counterfactual change in age, holding policy-relevant qualifications constant? This is a legal requirement, not merely a technical preference: age discrimination in lending is prohibited [51], yet no single party observes all features, and the sensitive attribute may be privately held. Similar challenges arise wherever high-stakes decisions rely on data distributed across organizational boundaries, including healthcare, employment, and criminal justice.

Data are split, but decisions are shared. As predictive data become increasingly distributed across organizations, data islands are becoming common [53]. Federated learning enables collaborative training without centralizing raw data [26, 35]. While most work assumes horizontal partitioning, many real settings are vertically partitioned, with parties holding disjoint feature subsets for overlapping individuals [52, 53]. In VFL, party-local representations are aggregated into a joint predictor, but fairness becomes harder because proxies may be distributed across parties, counterfactual interventions may be implausible under partial visibility, and sensitive attributes may still leak through shared representations or updates [17, 25, 33, 36, 52]. Existing approaches such as statistical constraints, representation invariance, and adversarial debiasing often fail to cleanly separate permissible mediators from impermissible proxies, leading to under-protection or utility loss [57, 58].

Fairness is about causal what-if questions. Fairness in high-stakes decision-making hinges on causal influence rather than correlation [32, 40]. A natural standard is counterfactual consistency: an individual's outcome should remain unchanged under a hypothetical intervention on the protected attribute, holding fixed what should remain the same for that person [30]. Crucially, this framing recognizes that not all causal pathways are alike. Some pathways are impermissible, others may be acceptable under policy, and observed features may act as proxies that should not transmit sensitive influence [28, 60]. Operationalizing this reasoning typically presumes a structural causal model or a vetted causal graph [3, 27]. In practice, such graphs are rarely available, are often contested, and are costly to maintain in multi-party settings subject to distribution shift.

Our proposal: SCC-VFL. We introduce SCC-VFL, or Selective Counterfactual Consistency for Vertical Federated Learning, a training-time framework that targets individual-level counterfactual stability while respecting privacy and partial visibility. SCC-VFL integrates three components: (i) a graph-free, differentially private procedure to identify candidate non-descendants, permissible mediators, and impermissible proxies; (ii) party-local masked counterfactual generation that edits only policy-permitted mediators while suppressing proxy leakage; and (iii)

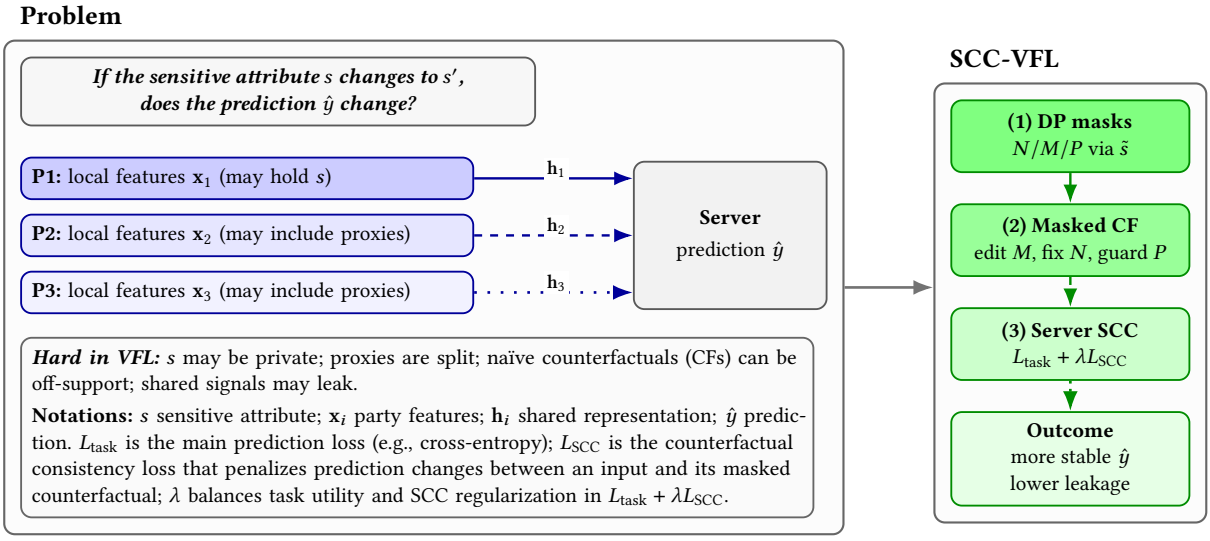


Fig. 1. Overview of SCC-VFL. *Left:* In VFL, features are distributed across parties and sensitive attributes may be private, making naive counterfactual edits unstable or leaky. *Right:* SCC-VFL discovers feature roles via differentially private masks, edits only policy-permitted mediators, and enforces prediction stability at server, improving stability while reducing leakage.

server-side enforcement via a selective counterfactual consistency loss that penalizes impermissible prediction changes while preserving legitimate pathways. Figure 1 summarizes the VFL fairness challenge and the design of SCC-VFL.

Key contributions. This paper makes four contributions toward individual-level fairness and auditability in VFL:

(1) **Problem formulation: selective counterfactual consistency for VFL.** We define SCC for VFL: predictions should be stable under counterfactual changes to a sensitive attribute, with non-descendants fixed and only policy-permitted mediators allowed to vary.
(2) **Graph-free, policy-aware mask discovery under differential privacy.** We propose a DP, server-assisted mask discovery that partitions each party's features into non-descendants (N), mediators (M), and proxies (P), thereby separating policy specification from enforcement.
(3) **Masked counterfactual generation with leakage control.** We design party-local masked generators that edit only mediators, preserve non-descendant identity, and reduce proxy leakage while sharing only representations.
(4) **Server-centric enforcement and empirical validation.** We enforce SCC at aggregation and evaluate across multiple domains, reducing individual flip rates (up to 98%) while maintaining accuracy and lowering empirical sensitive leakage under inference attacks.

2 Related Work

2.1 Foundations of Algorithmic Fairness

Foundational work formalizes algorithmic fairness through complementary group- and individual-level criteria, while exposing tensions among statistical parity goals, calibration, and utility [2, 10, 21, 29]. Group notions such as demographic parity and equalized odds target parity in outcomes or error rates across protected groups

[21, 29], while individual notions seek consistent treatment of similar individuals but typically depend on a task-specific similarity metric [14, 55]. Surveys and critiques emphasize that observational metrics can obscure causal pathways from sensitive attributes to predictions, and that satisfying one fairness criterion can preclude another [8, 56]. Although auditing and mitigation toolkits have matured, real deployments still report per-person inconsistencies even when aggregate metrics appear acceptable [42]. These limitations motivate approaches that move beyond correlational parity toward individual-level guarantees and pathway-aware reasoning about how protected attributes and proxies influence decisions.

2.2 Counterfactual and Causal Fairness

Causal perspectives define fairness via interventions on protected attributes and by restricting which causal pathways may influence decisions. Building on counterfactual explanations as a contestability tool [47], counterfactual fairness deems a decision fair if it remains unchanged when the protected attribute is counterfactually altered, typically using structural causal models and a separation of descendant from non-descendant features [30]. Path-specific fairness blocks impermissible routes while allowing policy-accepted mediators [7, 38], and causal reinterpretations of equalized odds connect group criteria to interpretable mechanisms [59]. Since full causal graphs are rarely available, prior work also studies partial-graph guarantees, safe feature sets, and graph-free surrogates that reduce sensitive influence without hand-crafted causal structure [61, 62], along with practical approximations such as adversarial removal of sensitive signal and counterfactual augmentation [18, 34]. Generative approaches further improve counterfactual validity by producing edits under identity and support constraints [7, 43, 46]. Overall, these methods provide principled individual-level guarantees in centralized settings, but many assume full feature access or trusted causal structure, limiting applicability under privacy and partial observability constraints.

2.3 Fairness in Federated and Vertical Federated Learning

Federated learning (FL) raises fairness challenges beyond centralized training due to client heterogeneity, non-IID data, and privacy constraints [31, 49]; existing work largely targets group fairness via reweighting, constrained or min–max objectives, and aggregation modifications to improve worst-group outcomes [9, 15, 39, 50], and studies joint group/individual formulations that surface fairness–utility trade-offs under non-IID regimes [54]. Vertical federated learning (VFL) extends FL to vertically partitioned features for overlapping individuals, training via representation sharing and secure aggregation [5, 52], but fairness in VFL remains comparatively underexplored, with early constrained approaches imposing group-level objectives despite limited access to sensitive attributes [31]. Bringing causal and counterfactual fairness into VFL is particularly challenging because mediators and impermissible proxies may be distributed across parties and the causal structure is rarely known or agreed upon [7, 30, 32, 38]; although graph-free proxy screening and generative counterfactual modeling provide useful ingredients, most prior work is centralized or lacks per-instance guarantees under realistic VFL threat models [43, 46, 62], motivating server-centric mechanisms that separate permissible from impermissible pathways, and promote stable decisions under protected-attribute interventions.

3 Proposed Approach: SCC-VFL

3.1 VFL System Setup

We consider VFL with m parties over n shared entities. Each party p holds disjoint features $x_i^{(p)} \in \mathbb{R}^{d_p}$ and shares no raw data; the server aligns IDs, aggregates party representations, and hosts the prediction head. A protected attribute $s_i \in \mathcal{S}$, binary or multi-category, is held by one trusted party or processed in an enclave and is never revealed to the server, which accesses only privacy-preserving summaries when needed. The supervised label is y_i, and the full feature vector is $x_i = (x_i^{(1)}, \ldots, x_i^{(m)})$. Table 1 summarizes the notation.

Table 1. Consolidated notation. Symbols are grouped by scope: system-level, per-party, and training-related.

Symbol	Owner	Description
n, m	System	Number of shared entities; number of parties
$x_i^{(p)} \in \mathbb{R}^{d_p}$	Party p	Local feature vector for entity i
$s_i \in \mathcal{S}$	Trusted holder	Protected (sensitive) attribute for entity i
y_i	Server	Supervised label (classification or regression)
$\phi^{(p)}$, $h_\phi^{(p)}$	Party p	Local encoder parameters and mapping
θ, f_θ	Server	Prediction head parameters and mapping
$z_i \in \mathbb{R}^{d_z}$	Server	Fused representation; $z_s = \psi(s)$: DP sketch of s
$N^{(p)}$, $M^{(p)}$, $P^{(p)}$	Party p	Non-descendant, mediator, and proxy index sets
$x_i^{cf,(p)}$, z_i^{cf}	Party p / Server	Counterfactual input and fused representation
$\varphi^{(p)}$, $\omega^{(p)}$	Party p	Generator $g_\varphi^{(p)}$ and adversary $a_\omega^{(p)}$ parameters
λ, α', β', η'	Server	Loss weights (SCC, identity, support, leakage)

Each party p operates a local encoder $h_\phi^{(p)}$ parameterized by $\phi^{(p)}$; the server hosts a prediction head f_θ parameterized by θ. The model follows split learning: each party sends an activation $h_\phi^{(p)}(x_i^{(p)})$ to the server, which fuses them into

$$z_i = \mathrm{Fuse}\big(h_\phi^{(1)}(x_i^{(1)}), \ldots, h_\phi^{(m)}(x_i^{(m)})\big) \in \mathbb{R}^{d_z}, \tag{3.1}$$

and predicts $f_\theta(z_i) \in \mathbb{R}^k$ (k=1 for regression; k classes for classification). To define counterfactual interventions on s, each party's coordinates are partitioned into non-descendants $N^{(p)}$ (fixed), permitted mediators $M^{(p)}$ (may change), and impermissible proxies $P^{(p)}$ (should not transmit sensitive influence). Under $s \leftarrow s'$, we construct masked counterfactual inputs $x_i^{cf,(p)}$ with $N^{(p)}$ unchanged, $M^{(p)}$ edited on-support for s', and proxy leakage from $P^{(p)}$ suppressed; the full counterfactual is $x_i^{cf} = (x_i^{cf,(1)}, \ldots, x_i^{cf,(m)})$ with fused representation z_i^{cf}. The server compares $f_\theta(z_i)$ and $f_\theta(z_i^{cf})$ and penalizes differences via a consistency term (Section 3.4), while the supervised objective is

$$\mathcal{L}_{\text{task}}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(f_\theta(z_i),\, y_i\big), \qquad z_i = \mathrm{Fuse}\big(h_\phi^{(1)}(x_i^{(1)}), \ldots, h_\phi^{(m)}(x_i^{(m)})\big). \tag{3.2}$$

Figure 2 illustrates the VFL system model, in which parties keep raw features local and exchange only intermediate representations with a coordinating server, while the sensitive attribute is held by a trusted party and accessed solely through a differentially private summary.

The remainder of this section introduces three components that together enable privacy-preserving, server-centric enforcement of individual-level counterfactual stability: selective mask discovery, masked counterfactual generation, and server-side consistency enforcement. We then describe the end-to-end privacy and security mechanisms that support deployment under realistic VFL threat models.

Specifically, SCC–VFL assumes a domain policy that divides the characteristics of each party into non-descendants

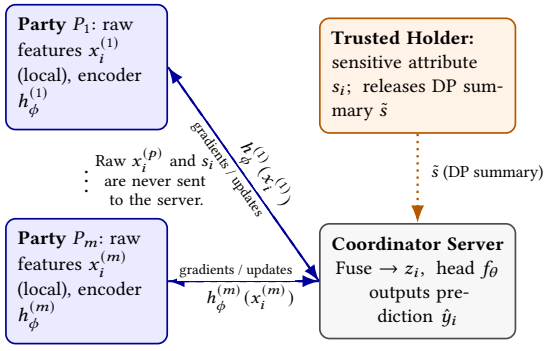


Fig. 2. System model for VFL.

$N^{(p)}$ (held fixed under $s \leftarrow s'$), permissible mediators $M^{(p)}$ (able to change on-support) and impermissible proxies $P^{(p)}$ (descendants of s treated as disallowed pathways). Our contribution is enforcement given a declared policy, not deciding the policy: the DP-sketch screening in Section 3.2 surfaces candidate descendants, while the final $N/M/P$ assignment and its governance are specified and auditable via the protocol in Appendix D. Consistency is interpreted as promoting local decision stability under policy-permitted mediator edits (recourse-style changes), rather than claiming removal of all mediator pathway effects.

3.2 Selective Descendant Discovery

Selective descendant discovery identifies, within each party, which local feature coordinates plausibly respond to interventions on the protected attribute after conditioning on other available information. For each party $p \in \{1, \dots, m\}$, it outputs a three-way mask $(N^{(p)}, M^{(p)}, P^{(p)})$ over local feature indices: non-descendants $N^{(p)}$, candidate mediators $M^{(p)}$, and stress-test proxies $P^{(p)} \subseteq M^{(p)}$. The goal is not causal identification, but a policy-aware, graph-free screening of features whose behavior is consistent with sensitivity to interventions on s. The procedure is graph-free and relies on conditional predictability evidence with a lightweight interventional validation. The server first produces a DP sketch $z_s = \psi(s)$, a low-dimensional encoding of s with clipping and Gaussian noise calibrated to a privacy budget (Proposition 3.1). Parties never receive raw s and use only z_s as an auxiliary covariate in local tests.

Proposition 3.1 (DP guarantee for sketch release). *For a single feature coordinate j, let $c_j \in \mathbb{R}^{2\times K}$ be the contingency table of s versus the discretised feature, and let neighboring datasets D, D' differ by the addition or removal of one record. The mechanism that (i) clips c_j to ℓ_2-sensitivity S via $\bar{c}_j = c_j \cdot \min(1, S/\|c_j\|_2)$ and (ii) releases $\tilde{c}_j = \bar{c}_j + \mathcal{N}(0, \sigma^2_{\text{sketch}} S^2 I)$ satisfies (ε, δ)-differential privacy with*

$$\varepsilon \;=\; \frac{1}{\sigma_{\text{sketch}}}\sqrt{2\ln(1.25/\delta)},$$

for any $\delta \in (0,1)$, under the standard Gaussian mechanism analysis [14].

Scope of the DP claim. Differential privacy applies *only* to the released contingency-table sketch used for mediator/proxy screening (Eqs. (C.7)–(C.8) in Appendix C.3). It does *not* cover representations, gradients, or model parameters. We do not claim end-to-end DP for the full SCC-VFL training pipeline. Formal composition accounting across mask refreshes is deferred to future work.

For coordinate j at party p, let $x_j^{(p)}$ be the target and $\tilde{x}_{\setminus j}^{(p)}$ the remaining local coordinates, optionally compressed by a small shared encoder. Party p fits two predictors of $x_j^{(p)}$ from $\tilde{x}_{\setminus j}^{(p)}$: $\hat{g}^{(p)}_{\text{w/ } z_s}$, which also takes z_s as input, and $\hat{g}^{(p)}_{\text{w/o } z_s}$, which does not. The predictors share the same architecture. Using a held-out risk $\mathcal{R}$ (e.g., squared error or negative log-likelihood), the party computes a risk difference and a Hilbert-Schmidt Independence Criterion (HSIC) statistic [20]:

$$\underbrace{\Delta_j^{(p)}}_{\substack{\text{Risk gain}\\ \text{from adding } z_s}} = \underbrace{\mathcal{R}\big(\hat{g}^{(p)}_{\text{w/ } z_s};\, x_j^{(p)} \mid \tilde{x}_{\setminus j}^{(p)}, z_s\big)}_{\text{Held-out risk with } z_s} - \underbrace{\mathcal{R}\big(\hat{g}^{(p)}_{\text{w/o } z_s};\, x_j^{(p)} \mid \tilde{x}_{\setminus j}^{(p)}\big)}_{\text{Held-out risk without } z_s}, \qquad \underbrace{\widehat{\text{HSIC}}_j^{(p)}}_{\substack{\text{Residual}\\ \text{dependence on } z_s}} = \underbrace{\text{HSIC}\big(r_j^{(p)}, z_s \mid \tilde{x}_{\setminus j}^{(p)}\big)}_{\text{Conditional contrast score}}, \tag{3.3}$$

where $r_j^{(p)}$ are residuals from one predictor, for example $\hat{g}^{(p)}_{\text{w/o } z_s}$. Intuitively, $\widehat{\text{HSIC}}_j^{(p)}$ measures residual dependence on z_s after conditioning, while $\Delta_j^{(p)}$ captures how much predictability improves when z_s is available.

Intuition: How Mask Discovery Works

Goal: Identify which features at each party are statistically responsive to interventions on sensitive attribute s.
Challenge: Most parties cannot directly observe s due to privacy constraints.
Approach: The server constructs a differentially private sketch z_s, a blurred representation of s that preserves aggregate patterns while protecting individual values. Each party then tests whether access to z_s improves prediction of its local features, conditional on other available information.
Outcome: Features whose predictability changes when z_s is available are flagged as candidates for policy review. This ranking supports, but does not determine, the $N/M/P$ partition used for counterfactual enforcement.

The server collects $(\Delta_j^{(p)}, \widehat{\mathrm{HSIC}}_j^{(p)})$ across parties and forms the tri-partition by assigning small $\widehat{\mathrm{HSIC}}_j^{(p)}$ to $N^{(p)}$, large $\widehat{\mathrm{HSIC}}_j^{(p)}$ to $M^{(p)}$, and within $M^{(p)}$ designating proxies $P^{(p)}$ as those with large $|\Delta_j^{(p)}|$, indicating strong sensitivity to s but limited task relevance. These scores serve as lightweight surrogates for full graph-based causal discovery and are intended to surface candidates for policy review rather than definitive causal roles. Finally, a lightweight interventional validation closes the loop. The server toggles z_s to a target $z_{s'}$ and instructs parties to apply the masked generator. Mediators in $M^{(p)}$ are edited toward s', while non-descendants in $N^{(p)}$ are copied exactly. If identity on $N^{(p)}$ is violated or mediators fail to respond, assignments are revised, yielding masks that are both statistically supported and intervention-consistent.

3.3 Masked Counterfactual Generation

Given per-party masks $(N^{(p)}, M^{(p)}, P^{(p)})$, masked counterfactual generation edits only mediator coordinates in a manner consistent with the conditional distribution under a target sensitive embedding. The objective is to preserve identity on non-descendants and suppress proxy leakage while producing plausible, on-support edits. Each party p trains a compact conditional generator $g_\varphi^{(p)}$ that takes local non-descendants $x_N^{(p)}$, current mediators $x_M^{(p)}$, a server-provided context vector c summarizing cross-party information, and the target embedding $z_{s'}$. The generator outputs edited mediators $x_M^{cf,(p)}$ intended to lie on the support of $p(x_M^{(p)} \mid x_N^{(p)}, c, z_{s'})$, while enforcing $x_N^{cf,(p)} = x_N^{(p)}$. Proxy features $x_P^{(p)}$ pass through unchanged but are guarded by an adversary that attempts to recover z_s from $(x_P^{cf,(p)}, x_N^{cf,(p)})$. The generators and encoders minimize this recovery success, reducing residual sensitive signal without discarding features. Parties share only intermediate activations for $(x^{(p)}, x^{cf,(p)})$ and scalar validity metrics with the server, where the SCC loss further discourages proxy-dependent prediction variation.

The per-party generator loss combines identity, support, and leakage-control terms. Writing q_φ for a conditional variational autoencoder and $a_\omega^{(p)}$ for the s-adversary, we define

$$\begin{aligned}\mathcal{L}_{\mathrm{gen}}^{(p)}(\varphi,\omega) \;=\; & \underbrace{\alpha\,\|x_N^{cf,(p)} - x_N^{(p)}\|_2^2}_{\text{Identity on non-descendants } N^{(p)}} \;+\; \underbrace{\beta\,\mathbb{E}\Big[-\log q_\varphi\Big(x_M^{cf,(p)} \mid x_N^{(p)}, c, z_{s'}\Big)\Big]}_{\text{On-support mediator likelihood under } z_{s'}} \\ & +\; \underbrace{\gamma\,\mathrm{MMD}\Big(x_M^{cf,(p)},\, \mathcal{D}_{M|N,c,z_{s'}}\Big)}_{\text{Kernel two-sample support matching for mediators}} \;-\; \underbrace{\eta\,\mathbb{E}\Big[\log\big(1 - a_\omega^{(p)}(x_P^{cf,(p)}, x_N^{cf,(p)})\big)\Big]}_{\text{Proxy leakage suppression via } s\text{-adversary}},\end{aligned} \tag{3.4}$$

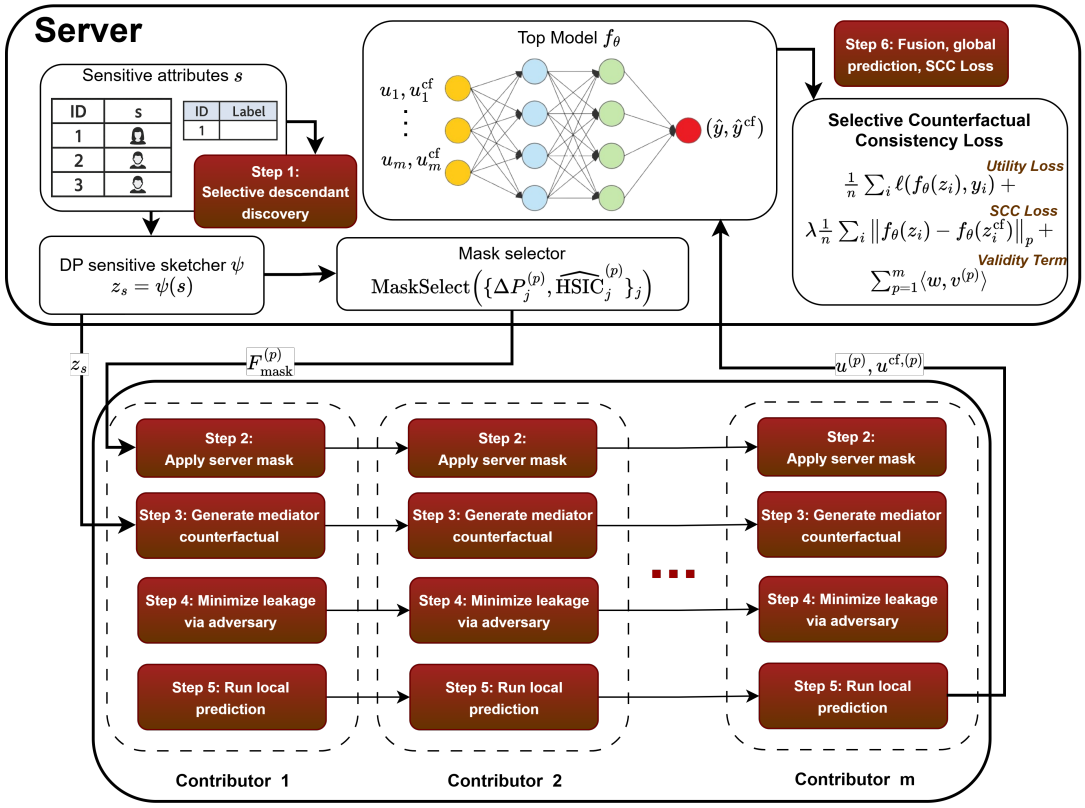


Fig. 3. Overview of SCC–VFL. The server discovers descendants (Step 1), sketches sensitive attributes z_s, selects a per-client mask via $\Delta P_j^{(p)}, \mathrm{HSIC}_j^{(p)}$, sends $F_{\mathrm{mask}}^{(p)}$ to contributors for Steps 2–5, receives $\{u^{(p)}, u^{cf,(p)}\}$, and performs fusion with the SCC loss (Step 6).

with nonnegative weights $(\alpha, \beta, \gamma, \eta)$. The first term enforces identity on non-descendants, the second and third promote on-support mediator edits, and the last suppresses residual sensitive predictability through proxies. An optional cycle-consistency term regenerates mediators from $z_{s'}$ back to z_s to further stabilize edits. The generator operates only on mediator coordinates, limiting communication and computational overhead while helping preserve task accuracy.

Appendix I provides a worked credit example illustrating how mediator edits and SCC enforcement interact under a counterfactual intervention.

3.4 Server-side Aggregation and Consistency

Fairness is enforced at the server because it is the only point where the fused decision is formed. For each minibatch, the server instructs parties to compute local activations on both the original inputs and masked counterfactual inputs targeted to one or more sensitive alternatives. It then fuses the received activations to obtain z and z^{cf}, evaluates the task loss on z, and applies a Selective Counterfactual Consistency (SCC) penalty that constrains prediction changes between z and z^{cf}, measured on logits for classification and scalar outputs for regression. Importantly, SCC is applied to *bounded, on-support* mediator edits produced by the generator, and is intended to prevent brittle decision flips under such permissible variability rather than to eliminate all mediator-driven effects. Because counterfactuals differ only along policy-permitted mediators, with non-descendants fixed

and proxies guarded, this penalty targets impermissible sensitive influence rather than suppressing legitimate causal pathways.

The server minimizes a joint objective combining supervised utility, per-example consistency, and per-party validity summaries: $\mathcal{V}_{\text{id}}^{(p)}$ (identity error on $N^{(p)}$), $\mathcal{V}_{\text{supp}}^{(p)}$ (mediator support adherence), and $\mathcal{V}_{\text{leak}}^{(p)}$ (proxy leakage measured by adversarial recovery). Formally,

$$\min_{\theta,\phi,\{\varphi^{(p)}\},\{\omega^{(p)}\}} \underbrace{\mathcal{L}_{\text{task}}(\theta,\phi)}_{\substack{\text{Supervised utility}\\ \text{on original inputs}}} + \underbrace{\lambda \frac{1}{n}\sum_{i=1}^{n} \left\| f_\theta(z_i) - f_\theta(z_i^{cf}) \right\|_p}_{\substack{\text{SCC penalizing prediction}\\ \text{changes under masked counterfactuals}}} + \underbrace{\sum_{p=1}^{m} \left[\alpha' \mathcal{V}_{\text{id}}^{(p)} + \beta' \mathcal{V}_{\text{supp}}^{(p)} - \eta' \mathcal{V}_{\text{leak}}^{(p)} \right]}_{\substack{\text{Validity control: identity on } N^{(p)},\\ \text{on-support edits for } M^{(p)}, \text{ reduced leakage via } P^{(p)}}} , \tag{3.5}$$

with weights $(\lambda, \alpha', \beta', \eta')$ chosen via validation on utility and stability metrics, where $p = 1$ for classification (logits) and $p = 2$ for regression.

The server backpropagates supervised and consistency gradients through the fusion operator and broadcasts split gradients to parties. Parties update local encoders and generators independently, while the server updates the prediction head, preserving separation of responsibilities and privacy constraints. The server also governs mask refresh. At fixed intervals, it recomputes descendant statistics under secure aggregation, applies false discovery rate control, and updates masks with hysteresis to prevent oscillations.

3.5 Privacy and Training Overview

SCC-VFL preserves sensitive information under the vertical federated threat model by never transmitting the protected attribute s in the clear. The server computes a clipped and noisy sketch $z_s = \psi(s)$ that satisfies (ε, δ)-DP for the released contingency tables (Proposition 3.1; optionally computed inside a trusted execution environment when stronger isolation is required), and parties receive only z_s and target variants $z_{s'}$, never raw sensitive values. We emphasize that this DP guarantee covers the sketch release only and does not extend to representations, gradients, or model parameters exchanged during training. Descendant statistics and validity metrics are shared via secure aggregation so the server observes only minibatch-level aggregates; only fixed-width representations (optionally randomized) cross authenticated channels. Audit logs record mask versions, privacy budgets, leakage metrics, and refresh events for governance review.

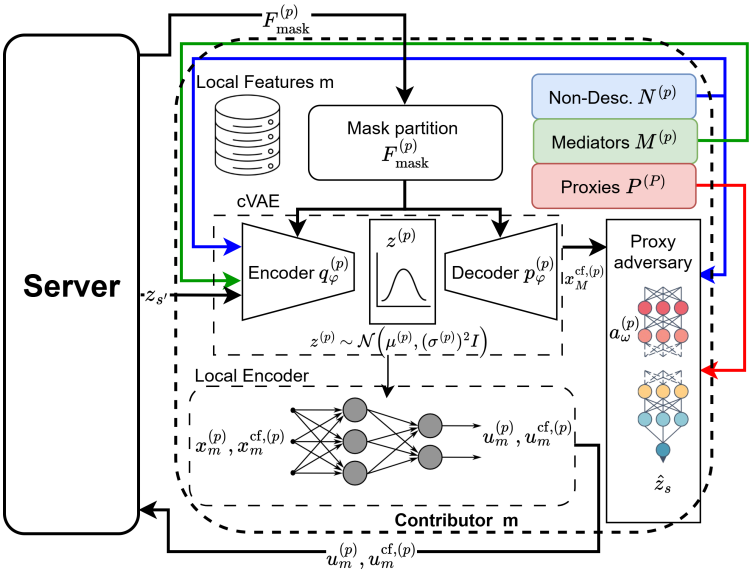


Fig. 4. Contributor m in SCC-VFL.

Training proceeds in three phases that may partially overlap: (1) selective descendant discovery for a few epochs with frozen encoders (securely aggregated statistics, FDR control, mask initialization), (2) party-local masked counterfactual generator training with identity, on-support, and leakage-control objectives while warm-starting the supervised head, and (3) joint optimization with the server-side

Selective Counterfactual Consistency penalty, interleaving short fine-tuning rounds with periodic mask refresh events to stabilize the utility-stability trade-off. Hyperparameters are chosen from a narrow grid using validation curves of accuracy versus flip rate and consistency gap; for classification we compute consistency on pre-softmax logits (to avoid saturation), and for regression we use squared differences. Defining $c_i = \|f_\theta(z_i) - f_\theta(z_i^{cf})\|_p$ and $v^{(p)} = (\mathcal{V}_{\text{id}}^{(p)}, \mathcal{V}_{\text{supp}}^{(p)}, \mathcal{V}_{\text{leak}}^{(p)})$, the overall objective is

$$\min_{\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell\big(f_\theta(z_i), y_i\big) + \lambda \frac{1}{n}\sum_{i=1}^{n} c_i + \sum_{p=1}^{m} \Big(\alpha' \mathcal{V}_{\text{id}}^{(p)} + \beta' \mathcal{V}_{\text{supp}}^{(p)} - \eta' \mathcal{V}_{\text{leak}}^{(p)}\Big), \qquad \Theta = \big(\theta, \phi, \{\varphi^{(p)}\}, \{\omega^{(p)}\}\big). \quad (3.6)$$

This decomposition keeps enforcement server-centric (stability on fused predictions) while parties ensure identity preservation on non-descendants, on-support mediator edits, and reduced proxy leakage. Figure 3 summarizes the end-to-end SCC-VFL pipeline, and Figure 4 illustrates Contributor m (masking, mediator counterfactual generation, adversarial leakage suppression, and representation return for fusion).

4 Experimental Setup

The evaluation of SCC-VFL tests whether it improves individual-level counterfactual fairness while preserving utility and privacy in realistic VFL settings. We run experiments across three domains (banking, healthcare, and criminal justice), using consistent vertical partitions, standard metrics, and baselines to enable cross-domain comparison.

4.1 Datasets

(1) Banking (German Credit, UCI): We use German Credit [22] with a binary default label. The sensitive attribute is age, binarized at 25 and used only for fairness evaluation (not as input). We form realistic vertical silos: Bank A (account/credit history), Employer/Payroll (employment/income), and an optional Bureau/Platform view (savings/collateral/third-party debtor), and evaluate under IID and non-IID splits. **(2) Healthcare (UCI Heart Disease):** We use UCI Heart Disease [24] with a binary disease label. The sensitive attribute is sex, treated as protected and omitted from inputs. We create three vertical views: Hospital A (clinical measurements/tests), Clinic B (demographic and structural risk factors), and Payer (remaining lab/administrative indicators), yielding clinically plausible mediators and proxies. **(3) Criminal Justice (COMPAS Cox violent subset):** We use a COMPAS-style subset [12] with a binary violent recidivism outcome. The sensitive attribute is race (African American vs. other), used only for fairness evaluation and excluded from inputs. Features are partitioned into Court system (charges/history), Community services (socioeconomic/supervision), and Agency/Platform (demographic screening), supporting both approximately IID and race-shifted non-IID splits.

We additionally evaluate SCC-VFL on a larger multi-group sensitive-attribute benchmark using **ACS Folktables** [11], with race (RAC1P) as a multi-category sensitive attribute and Income/Mobility as binary targets. Results under both IID and protected-attribute shift (Non-IID) settings are reported in Appendix B.

4.2 Evaluation Metrics

We evaluate each model using a fixed set of complementary metrics: **(1) Accuracy** (%) ↑, standard predictive performance; **(2) LogLoss** ↓, cross-entropy over predicted probabilities; **(3) Selective Consistency Gap (SCG)** ↓, the mean $\|f(z) - f(z^{cf})\|$ over logits, capturing per-instance counterfactual stability; **(4) Flip Rate (FR, %)** ↓, the fraction of samples whose predicted label changes under counterfactual edits (e.g., FR= 0.3% corresponds to ≈3 flips per 1000 individuals); **(5) Attribute Inference Attack Success Rate (AIA SR, %)** ↓, the post-hoc attacker success rate for predicting s from latent representations under increasing attack strength; and **(6) Subspace-PGD Attack Success Rate (PGD SR, %)** ↓, the success rate of an ℓ_∞-bounded adversary that perturbs mediator

coordinates to flip predictions as ϵ grows. Together, these quantify utility (Accuracy, LogLoss), fairness (SCG, FR), and privacy leakage and robustness (attack success rates).

4.3 Existing Baselines for Performance Comparison with SCC-VFL

We compare SCC-VFL against four baselines: **(1) No-mask adversarial debiasing (Adv-NoMask)** [23], which applies a server-level adversary to remove s-signal from fused activations without counterfactual reasoning; **(2) Uniform counterfactuals (Uniform-CF)** [18], which uses a naïve generator that edits all coordinates indiscriminately; **(3) Policy-blind mask (Policy-blind Mask)** [48], which performs descendant discovery without an explicit proxy partition; and **(4) Server-only consistency (Server-Consistency)** [6], which directly penalizes prediction changes between original and "flipped-s" DP-sketch embeddings, without generator modules.

4.4 Considered Adversaries

We stress-test all models with two adversarial probes: **(1) Attribute Inference Attack (AIA)** [36], a post-hoc attacker that trains a separate classifier on frozen latent representations to recover s under increasing attacker training budgets; and **(2) Subspace-constrained Projected Gradient Descent (Subspace-PGD)** [41], a gradient-based attack that perturbs only coordinates in a designated sensitive subspace within an ℓ_∞ ball of radius ϵ to flip the predicted label. Complete configurations and additional results for AIA and the Subspace-PGD variant are reported in Appendix E.

4.5 Implementation Details

All models are implemented in PyTorch. Each party encoder is a two-layer MLP (see Appendix C) with ReLU and dropout $d \in [0.05, 0.10]$, and the server applies a linear prediction head (output dimension equals the number of classes). SCC-VFL uses a conditional VAE-style masked generator that perturbs only mediator coordinates, with counterfactual scale $\gamma \in [0.20, 0.25]$; baselines use deterministic masked generators applied either to mediators or to all features. Mediators are selected via a DP-motivated discovery score that combines $|\Delta_j^{(p)}|$ and $\widehat{\mathrm{HSIC}}_j^{(p)}$ (Section 3.2); thresholds are set by percentile ranking, retaining the top $\rho_M = 0.60$ coordinates as mediators $M^{(p)}$ and defining proxies $P^{(p)}$ as the top $\rho_P = 0.50$ fraction within $M^{(p)}$. Training uses AdamW with learning rate 0.005–0.015 and optional weight decay 5×10^{-4} for up to 80 epochs for baselines and 150–300 epochs for SCC-VFL, with early stopping on a composite validation objective combining log loss, SCG, and FR. SCC-VFL logits are calibrated via temperature scaling on a held-out validation set (see Appendix G). All metrics are averaged over 30 random seeds and reported under both IID and non-IID client splits.

4.6 Experimental Design

We conduct three experiments: **(E1)** utility–fairness comparisons across all datasets using the core metrics; **(E2)** attack evaluations that measure sensitive leakage under party-side and server-side probes; and **(E3)** ablations that isolate the effects of mask discovery, masked generation, and the server-side consistency penalty on utility and stability. Together, these experiments show that SCC-VFL delivers counterfactual stability, improved privacy protection, and strong predictive performance across diverse VFL settings; (E2) and (E3) are conducted on the German Credit dataset under the IID split.

5 Experimental Results & Analyses

5.1 Utility–Fairness Trade-offs Across Datasets

We first compare SCC-VFL and all baselines on the joint utility–fairness metrics across the three datasets under both IID and non-IID partitions (Table 2). Across domains, SCC-VFL either matches or is very close to the best

Table 2. Performance comparison on UCI German Credit, UCI Heart Disease, and COMPAS Cox datasets under IID and Non-IID splits. Best values per block are highlighted in bold.

Dataset	Split	Method	Acc ↑	LogLoss ↓	SCG ↓	FR (%) ↓
German Credit	IID	Adv-NoMask	$0.7176_{\pm 0.0217}$	$0.7131_{\pm 0.0666}$	$0.0561_{\pm 0.0076}$	$0.97_{\pm 0.35}$
		Uniform-CF	$0.7311_{\pm 0.0197}$	$0.6305_{\pm 0.0423}$	$0.0891_{\pm 0.0094}$	$1.56_{\pm 0.77}$
		Policy-blind Mask	$0.7322_{\pm 0.0214}$	$0.6114_{\pm 0.0433}$	$0.0532_{\pm 0.0053}$	$1.30_{\pm 0.75}$
		Server-Consistency	$\mathbf{0.7351_{\pm 0.0221}}$	$0.5747_{\pm 0.0409}$	$0.0218_{\pm 0.0029}$	$0.68_{\pm 0.48}$
		SCC-VFL (ours)	$0.7224_{\pm 0.0244}$	$\mathbf{0.5676_{\pm 0.0277}}$	$\mathbf{0.0031_{\pm 0.0036}}$	$\mathbf{0.09_{\pm 0.15}}$
	Non-IID	Adv-NoMask	$0.7200_{\pm 0.0291}$	$0.7143_{\pm 0.0676}$	$0.5688_{\pm 0.0306}$	$8.22_{\pm 1.75}$
		Uniform-CF	$0.7293_{\pm 0.0229}$	$0.6376_{\pm 0.0510}$	$0.3652_{\pm 0.0297}$	$7.00_{\pm 1.54}$
		Policy-blind Mask	$\mathbf{0.7324_{\pm 0.0242}}$	$0.6155_{\pm 0.0454}$	$0.3185_{\pm 0.0270}$	$6.24_{\pm 1.35}$
		Server-Consistency	$0.7300_{\pm 0.0230}$	$0.5904_{\pm 0.0460}$	$0.3335_{\pm 0.0273}$	$6.90_{\pm 1.39}$
		SCC-VFL (ours)	$0.7243_{\pm 0.0246}$	$\mathbf{0.5716_{\pm 0.0259}}$	$\mathbf{0.0036_{\pm 0.0041}}$	$\mathbf{0.12_{\pm 0.20}}$
UCI Heart	IID	Adv-NoMask	$0.9712_{\pm 0.0053}$	$\mathbf{0.0907_{\pm 0.0155}}$	$1.9882_{\pm 0.4007}$	$12.60_{\pm 3.01}$
		Uniform-CF	$0.9530_{\pm 0.0080}$	$0.1806_{\pm 0.0105}$	$0.6694_{\pm 0.1049}$	$10.52_{\pm 2.61}$
		Policy-blind Mask	$0.9488_{\pm 0.0067}$	$0.1941_{\pm 0.0083}$	$0.6067_{\pm 0.1171}$	$10.45_{\pm 2.88}$
		Server-Consistency	$0.8844_{\pm 0.0211}$	$0.3442_{\pm 0.0165}$	$0.2869_{\pm 0.0319}$	$9.13_{\pm 1.48}$
		SCC-VFL (ours)	$\mathbf{0.9820_{\pm 0.0077}}$	$0.0910_{\pm 0.0136}$	$\mathbf{0.2503_{\pm 0.0394}}$	$\mathbf{2.11_{\pm 0.86}}$
	Non-IID	Adv-NoMask	$0.8538_{\pm 0.0386}$	$0.5798_{\pm 0.0605}$	$0.4646_{\pm 0.1562}$	$15.21_{\pm 5.57}$
		Uniform-CF	$0.8557_{\pm 0.0340}$	$0.5804_{\pm 0.0418}$	$0.0935_{\pm 0.0182}$	$2.89_{\pm 1.46}$
		Policy-blind Mask	$0.8500_{\pm 0.0390}$	$0.5883_{\pm 0.0489}$	$0.0248_{\pm 0.0064}$	$0.83_{\pm 0.67}$
		Server-Consistency	$0.8145_{\pm 0.0432}$	$0.6282_{\pm 0.0555}$	$0.1840_{\pm 0.0522}$	$7.86_{\pm 2.95}$
		SCC-VFL (ours)	$\mathbf{0.9492_{\pm 0.0153}}$	$\mathbf{0.4594_{\pm 0.0497}}$	$\mathbf{0.0092_{\pm 0.0058}}$	$\mathbf{0.05_{\pm 0.19}}$
COMPAS Cox	IID	Adv-NoMask	$\mathbf{0.9710_{\pm 0.0021}}$	$\mathbf{0.0499_{\pm 0.0021}}$	$1.0357_{\pm 0.2266}$	$0.66_{\pm 0.46}$
		Uniform-CF	$0.9705_{\pm 0.0034}$	$0.0548_{\pm 0.0016}$	$0.3839_{\pm 0.0747}$	$0.56_{\pm 0.57}$
		Policy-blind Mask	$0.9705_{\pm 0.0035}$	$0.0548_{\pm 0.0021}$	$0.3901_{\pm 0.0812}$	$0.55_{\pm 0.55}$
		Server-Consistency	$0.9706_{\pm 0.0034}$	$0.0553_{\pm 0.0021}$	$0.3191_{\pm 0.0360}$	$0.63_{\pm 0.64}$
		SCC-VFL (ours)	$0.9697_{\pm 0.0042}$	$0.0514_{\pm 0.0024}$	$\mathbf{0.0266_{\pm 0.0049}}$	$\mathbf{0.02_{\pm 0.03}}$
	Non-IID	Adv-NoMask	$0.9758_{\pm 0.0015}$	$\mathbf{0.0428_{\pm 0.0017}}$	$1.0712_{\pm 0.3014}$	$0.57_{\pm 0.34}$
		Uniform-CF	$0.9750_{\pm 0.0047}$	$0.0485_{\pm 0.0027}$	$0.3812_{\pm 0.0694}$	$0.41_{\pm 0.34}$
		Policy-blind Mask	$0.9755_{\pm 0.0032}$	$0.0485_{\pm 0.0018}$	$0.3751_{\pm 0.0793}$	$0.34_{\pm 0.28}$
		Server-Consistency	$\mathbf{0.9759_{\pm 0.0014}}$	$0.0489_{\pm 0.0019}$	$0.3238_{\pm 0.0441}$	$0.34_{\pm 0.28}$
		SCC-VFL (ours)	$0.9721_{\pm 0.0065}$	$0.0453_{\pm 0.0035}$	$\mathbf{0.0255_{\pm 0.0044}}$	$\mathbf{0.02_{\pm 0.03}}$

baseline in Accuracy and LogLoss while sharply improving Selective Consistency Gap (SCG) and Flip Rate (FR). On German Credit, SCC-VFL slightly trails Server-Consistency in Accuracy by about one percentage point but attains the lowest LogLoss and reduces SCG and FR by roughly an order of magnitude in both IID and non-IID settings. On UCI Heart, SCC-VFL achieves the highest Accuracy in both splits and essentially ties the best LogLoss, while cutting SCG and FR from double digits to around 0.25 / 2.1 in IID and almost zero in non-IID. On COMPAS Cox, SCC-VFL maintains Accuracy within 0.3 percentage points of the strongest baselines, with

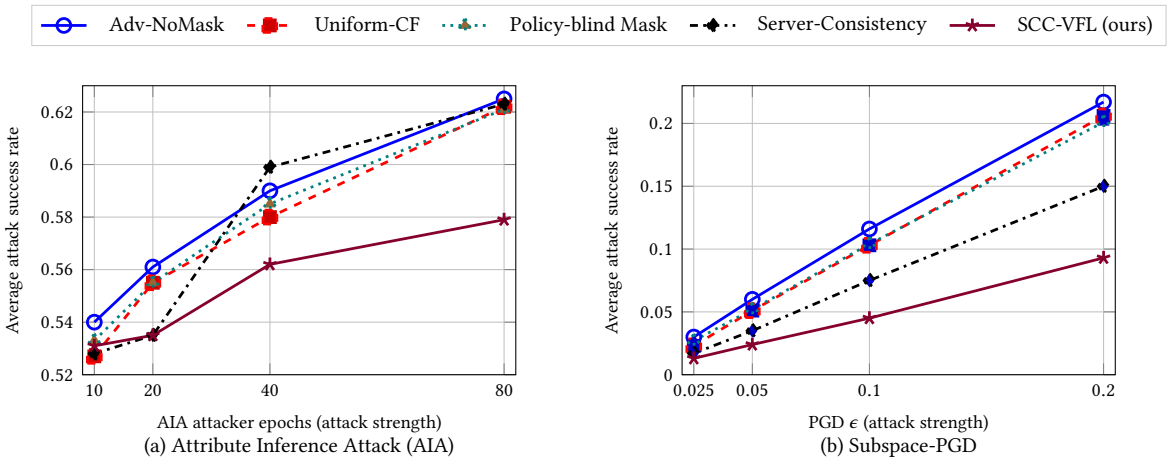


Fig. 5. Average attack success rate for AIA and Subspace-PGD across attack strengths (30 seeds).

only a minor LogLoss gap, yet collapses SCG to about 0.026 and FR to 0.02 compared to 0.3–1.0 SCG and 0.3–0.7 FR for all baselines. non-IID partitions intensify these gaps: baselines often suffer large SCG and FR spikes when distributions shift, whereas SCC-VFL keeps SCG in the 0.003–0.026 range and FR near zero while staying competitive in utility.

These patterns follow from how SCC-VFL enforces counterfactual stability. Adv-NoMask operates only at the fused representation, so it can suppress some global s-signal but lacks instance-level control over which features move, leaving many per-example flips and high SCG. Uniform-CF perturbs all coordinates, often preserving or slightly improving utility but introducing off-support edits that inflate SCG and FR, especially under non-IID splits. Policy-blind Mask performs better by focusing edits on descendants, yet treating all descendants as acceptable mediators lets proxies transmit residual sensitive signal and limits stability gains. Server-Consistency constrains predictions under "flipped-s" embeddings but has no explicit generator or identity constraint, so it cannot guarantee that only permissible mediators move or that non-descendants stay fixed. In contrast, SCC-VFL combines selective mask discovery, mediator-only cVAE edits, and a per-example SCC loss at the server, which jointly preserves on-support edits, freezes non-descendants, and collapses proxy influence. This architecture yields models that keep task performance high while producing much lower SCG and FR across datasets and under both IID and non-IID VFL regimes.

5.2 Privacy Leakage and Robustness Analysis

We next evaluate privacy leakage and robustness using the Attribute Inference Attack (AIA) and Subspace-constrained Projected Gradient Descent (Subspace-PGD). As Fig. 5 shows, as attacker strength increases, baselines cluster around 62% AIA success at 80 epochs, whereas SCC-VFL stays at 57.9%, a 4–5 point reduction in balanced attack accuracy. For Subspace-PGD, which perturbs only mediator coordinates, SCC-VFL consistently attains the lowest success rate, from 1.0% at $\epsilon = 0.02$ to 9.3% at $\epsilon = 0.20$, about half of Adv-NoMask (21.7%) and below all other baselines. These gains stem from combining selective mask discovery (to isolate mediators and proxies), masked generation (editing only mediators while enforcing identity on non-descendants and suppressing proxy leakage), and an SCC loss that flattens the decision boundary along mediator directions, reducing usable s-signal and strengthening resistance to mediator-space attacks relative to Adv-NoMask, Uniform-CF, Policy-blind Mask, and Server-Consistency.

Table 3. Ablation of SCC-VFL components on German Credit (IID split).

Method	Acc ↑	LogLoss ↓	SCG ↓	FR (%) ↓
SCC-VFL (full)	$0.7186_{\pm 0.022}$	$0.5737_{\pm 0.027}$	$\mathbf{0.0045_{\pm 0.004}}$	$\mathbf{0.11_{\pm 0.180}}$
w/o mask discovery	$\mathbf{0.7220_{\pm 0.024}}$	$\mathbf{0.5713_{\pm 0.026}}$	$0.0094_{\pm 0.016}$	$0.36_{\pm 0.60}$
w/o generator	$0.7089_{\pm 0.017}$	$0.5761_{\pm 0.027}$	$0.0233_{\pm 0.005}$	$0.94_{\pm 0.54}$
w/o consistency	$0.7174_{\pm 0.022}$	$0.5747_{\pm 0.027}$	$0.0076_{\pm 0.000}$	$0.63_{\pm 0.100}$

5.3 Ablation Analysis

Table 3 studies the contribution of each component of SCC-VFL. The full model attains the best overall trade-off, with the lowest SCG (0.0045) and FR (0.11%) while keeping Accuracy and LogLoss on par with or slightly better than the ablations. Removing mask discovery and treating all features as mediators yields marginally lower LogLoss but roughly doubles SCG and more than triples FR, showing that selective masks help target edits and avoid unnecessary flips. Removing the generator (leaving only server-side consistency) substantially degrades stability, with SCG increasing to 0.0233 and FR to 0.94%, indicating that explicit on-support mediator counterfactuals are important for effective consistency enforcement. Finally, dropping the consistency term while keeping masks and the generator retains good utility but leaves SCG and FR significantly higher than the full model, confirming that the per-example SCC loss at the server is necessary to translate selective edits into stable predictions. We also report runtime and communication overhead measurements in Appendix F.

5.4 Discussion & Implications

What SCC-VFL does and does not guarantee. Our results show that SCC-VFL substantially reduces prediction instability under counterfactual interventions on the sensitive attribute. This is a *consistency* property: for a given individual, predictions remain stable when the protected attribute is counterfactually changed, non-descendant features are held fixed, and only policy-permitted mediators are allowed to vary. Viewed this way, mediator edits function as recourse-style interventions rather than arbitrary perturbations [45, 47], and our guarantees inherit standard caveats about counterfactual-based fairness [3, 27].

Consistency, however, does not imply justice [44]. A system can be consistently discriminatory if the underlying policy specification encodes biased or unjustified judgments about which causal pathways are deemed permissible [44, 51]. SCC-VFL enforces the provided policy model rather than correcting it. Moreover, counterfactual stability captures only one dimension of fairness and does not ensure group-level parity properties such as equalized odds or demographic parity, nor does it resolve intersectional concerns. Deployments should therefore complement SCC-VFL with group-level audits and substantive review of whether the policy model itself reflects defensible values.

Individual stability as an auditable fairness target. The results position individual-level counterfactual stability as a practically auditable fairness target for VFL. Metrics such as Selective Consistency Gap (SCG) and Flip Rate (FR) directly quantify whether a person's prediction changes under an intervention on the protected attribute when non-descendants are held fixed. Across datasets, SCC-VFL achieves large reductions in SCG and FR while maintaining competitive Accuracy and LogLoss, indicating that it removes arbitrary per-example sensitivity rather than trading utility for coarse group-level constraints [4]. This aligns with FAccT goals of accountability and contestability, since stability can be assessed at the individual level and aggregated for monitoring.

Robustness under heterogeneity and shift. The non-IID results highlight how institutional heterogeneity and distribution shift can amplify proxy pathways and off-support edits, leading baseline methods to exhibit sharp stability degradation. SCC-VFL remains more stable by restricting edits to mediator coordinates, enforcing

identity on non-descendants, and applying a per-example consistency penalty that flattens the decision boundary along the intended intervention. The observed reductions in attribute inference and Subspace-PGD attack success suggest reduced usable sensitive signal in shared representations under the evaluated threats, though these results should be interpreted as empirical robustness rather than formal privacy guarantees. A key implication is that policy choices embedded in the mask should be documented and audited, including shift-aware reporting of SCG and FR, to ensure that enforced intervention semantics align with domain requirements.

5.5 Limitations and Broader Considerations

Technical scope. We evaluate SCC-VFL on tabular VFL benchmarks with discrete protected attributes and a fixed policy specification. Extending the approach to high-dimensional modalities (text, images), continuous protected attributes, and settings with frequent policy updates requires additional design for (i) scalable mask discovery and (ii) valid, on-support counterfactual generation in richer feature spaces. The proposed mask discovery is graph-free and relies on statistical signals in order to support the enforcement of a declared policy rather than to recover ground-truth causal structure.

Policy dependence and governance. SCC-VFL enforces counterfactual stability relative to the chosen $N/M/P$ specification; it does not determine which pathways *are normatively permissible*. If the specification is incomplete or outdated, the model can be stable while still misaligned with institutional or legal intent. For deployment, we recommend documenting the policy rationale, tracking mask versions, and pairing SCC metrics (e.g., SCG/FR) with routine outcome monitoring, including intersectional slices when feasible, following model-reporting practices [37].

Privacy and accountability boundaries. The formal DP guarantee applies only to the sketch used for mask discovery (Proposition 3.1), not to representations, gradients, or model parameters; composition across mask refreshes is capped but not tracked via a formal composition accountant (e.g., Rényi DP). Beyond the sketch, privacy evidence is empirical and does not rule out inference under stronger threat models. We position SCC-VFL as an accountability mechanism to be used alongside established privacy protections, audits, and external review.

6 Conclusion and Future Work

This paper introduced SCC-VFL, a server-centric framework for enforcing individual-level counterfactual stability in vertical federated learning while preserving predictive performance. The approach integrates three components: selective mask discovery that partitions each party's features into non-descendants, mediators, and proxies; masked counterfactual generation that edits only policy-permitted mediators while preserving identity on non-descendants and suppressing proxy leakage; and a server-side selective counterfactual consistency loss that penalizes prediction changes under on-support mediator interventions. Across banking, healthcare, and criminal justice datasets, SCC-VFL achieves strong predictive utility while substantially reducing the selective consistency gap, flip rate, and adversarial attack success under both attribute inference and subspace-constrained PGD, in IID and non-IID vertical settings.

Future work includes extending SCC-VFL along methodological, privacy, and deployment dimensions. An immediate direction is supporting higher-dimensional modalities such as text or imaging, and more complex vertical consortia with many parties, partial entity overlap, and asynchronous participation. On the privacy side, integrating formal differential privacy accounting and secure computation primitives into the mask discovery stage would provide end-to-end privacy guarantees beyond the current functional protections. Another avenue is adaptive and multi-attribute mask discovery that can track distribution shift and handle multiple protected attributes simultaneously. Finally, the masked generators developed for counterfactual stability could also support actionable recourse and auditing [45], where policy-compliant mediator edits are exposed to stakeholders as interpretable explanations or intervention recommendations.

Code and data availability

The implementation is available at https://github.com/dawoodwasif/SCC-VFL. The datasets used in this work are publicly available from their official sources; links and preprocessing scripts are provided in the repository.

Generative AI Usage

We used generative models in two limited and transparent ways. First, a generative model is used as a *method component*: the masked conditional VAE that produces mediator-only counterfactual edits within the proposed SCC-VFL framework. Second, standard generative tools were used during manuscript preparation for formatting assistance and language polishing. Generative models were not used to generate scientific hypotheses, select evaluation metrics, design experiments, interpret results, or draw conclusions. No external generative system was used to create, label, or augment the experimental datasets, and no model outputs were used as ground-truth labels for evaluation. All reported quantitative results, tables, and figures are produced by our implemented training and evaluation code on the stated datasets and splits.

Ethical Considerations

This work studies fairness and privacy risks in vertical federated learning, where disjoint feature holders collaborate without sharing raw features. The protected attribute s is treated as sensitive: it is not broadcast to parties and is used only to define and evaluate the counterfactual intervention underlying Selective Consistency Gap (SCG) and Flip Rate (FR). Our approach encodes an explicit normative choice about which descendants of s are permissible mediators versus impermissible proxies. These choices can shape downstream incentives and institutional behavior and should therefore be defined using domain policy, legal guidance, and stakeholder input. We emphasize that deployments should document these choices and audit resulting models using both utility metrics and stability metrics such as SCG and FR across relevant subgroups and distribution shifts.

Adverse Impacts

A stability objective can be misused to justify suppressing legitimate signal, to enforce overly rigid decisions, or to mask discrimination by optimizing a narrow metric. Errors in mediator or proxy specification can unintentionally amplify harm by freezing features that should change or allowing pathways that should have been blocked. While our privacy results show reduced post hoc leakage under the evaluated attribute inference attacks and increased robustness under Subspace-PGD, the method does not provide a formal privacy guarantee for training data, representations, gradients, or model parameters. Adversaries outside our threat model may still infer sensitive information. As with many fairness interventions, the risks of mis-specification or misuse are most likely to be borne by protected or marginalized groups, underscoring the need for external accountability. To mitigate these risks, we recommend documenting intervention semantics and policy specifications, reporting SCG and FR alongside standard group fairness and utility metrics under non-IID and temporal shifts, and monitoring deployed systems with clear review and rollback processes.

Acknowledgements

This research was partially supported by the U.S. National Science Foundation (NSF) under Grant No. 2416728. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities.* MIT Press. https://fairmlbook.org/

[2] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[3] Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 80–89.

[4] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 514–524.

[5] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 1175–1191.

[6] Yuhang Chen, Wenke Huang, and Mang Ye. 2024. Fair federated learning under domain skew with local consistency and domain diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12077–12086.

[7] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 7801–7808.

[8] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

[9] Lingyang Chu, Lanjun Wang, Yanjie Dong, Jian Pei, Zirui Zhou, and Yong Zhang. 2021. Fedfair: Training fair models in cross-silo federated learning. *arXiv preprint arXiv:2109.05662* (2021).

[10] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 797–806.

[11] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems* 34 (2021), 6478–6490.

[12] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018), eaao5580.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through Awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*. ACM, 214–226. doi:10.1145/2090236.2090255

[14] Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science* 9, 3-4 (2014), 211–487.

[15] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. 2023. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 7494–7502.

[16] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*. ACM, 259–268. doi:10.1145/2783258.2783311

[17] Chong Fu, Xuhong Zhang, Shouling Ji, Jinyin Chen, Jingzheng Wu, Shanqing Guo, Jun Zhou, Alex X Liu, and Ting Wang. 2022. Label inference attacks against vertical federated learning. In *31st USENIX security symposium (USENIX Security 22)*. 1397–1414.

[18] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 219–226.

[19] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[20] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*. Springer, 63–77.

[21] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29 (2016).

[22] Hans Hofmann. 1994. Statlog (German credit data). *UCI Machine Learning Repository* 10 (1994), C5NC77.

[23] Junyuan Hong, Zhuangdi Zhu, Shuyang Yu, Zhangyang Wang, Hiroko H Dodge, and Jiayu Zhou. 2021. Federated adversarial debiasing for fair and transferable representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 617–627.

[24] Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. 1988. UCI machine learning repository-heart disease data set. *School Inf. Comput. Sci., Univ. California, Irvine, CA, USA* (1988).

[25] Xiao Jin, Pin-Yu Chen, Chia-Yi Hsu, Chia-Mu Yu, and Tianyi Chen. 2021. Cafe: Catastrophic data leakage in vertical federated learning. *Advances in neural information processing systems* 34 (2021), 994–1006.

[26] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning* 14, 1–2 (2021), 1–210. doi:10.1561/2200000083

[27] Atoosa Kasirzadeh and Andrew Smart. 2021. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 228–236.

[28] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems* 30 (2017).

[29] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, Berkeley, CA, USA, January 9-11, 2017 (LIPIcs, Vol. 67)*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 43:1–43:23. doi:10.4230/LIPICS.ITCS.2017.43

[30] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in Neural Information Processing Systems* 30 (2017).

[31] Yang Liu, Yan Kang, Tianyuan Zou, Yanhong Pu, Yuanqin He, Xiaozhou Ye, Ye Ouyang, Ya-Qin Zhang, and Qiang Yang. 2024. Vertical federated learning: Concepts, advances, and challenges. *IEEE Transactions on Knowledge and Data Engineering* 36, 7 (2024), 3615–3634.

[32] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859* (2018).

[33] Xinjian Luo, Yuncheng Wu, Xiaokui Xiao, and Beng Chin Ooi. 2021. Feature Inference Attack on Model Predictions in Vertical Federated Learning. In *37th IEEE International Conference on Data Engineering (ICDE)*. IEEE, 181–192. doi:10.1109/ICDE51399.2021.00023

[34] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*. PMLR, 3384–3393.

[35] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) (Proceedings of Machine Learning Research, Vol. 54)*. PMLR, 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html

[36] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 691–706.

[37] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[38] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[39] Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. 2022. Minimax demographic group fairness in federated learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 142–159.

[40] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference* (2 ed.). Cambridge University Press.

[41] Samuel Pfrommer, Brendon G Anderson, and Somayeh Sojoudi. 2023. Projected randomized smoothing for certified adversarial robustness. *arXiv preprint arXiv:2309.13794* (2023).

[42] Teresa Salazar, Helder AraÃ̃ējo, Alberto Cano, and Pedro Henriques Abreu. 2024. A survey on group fairness in federated learning: challenges, taxonomy of solutions and directions for future research. *arXiv preprint arXiv:2410.03855* (2024).

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

[44] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[45] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*. 10–19.

[46] Boris Van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela Van der Schaar. 2021. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems* 34 (2021), 22221–22233.

[47] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2018), 841–887.

[48] Yu Wang, Yuying Zhao, Yushun Dong, Huiyuan Chen, Jundong Li, and Tyler Derr. 2022. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1938–1948.

[49] Dawood Wasif, Dian Chen, Sindhuja Madabushi, Nithin Alluru, Terrence J Moore, and Jin-Hee Cho. 2025. Empirical analysis of privacy-fairness-accuracy trade-offs in federated learning: a step towards responsible AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 8. 2666–2677.

[50] Dawood Wasif, Terrence J Moore, and Jin-Hee Cho. 2025. RESFL: An Uncertainty-Aware Framework for Responsible Federated Learning by Balancing Privacy, Fairness and Utility in Autonomous Vehicles. *arXiv preprint arXiv:2503.16251* (2025).

[51] Hilde Weerts, Raphaële Xenidis, Fabien Tarissan, Henrik Palmer Olsen, and Mykola Pechenizkiy. 2023. Algorithmic unfairness through the lens of EU non-discrimination law: Or why the law is not a decision tree. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 805–816.

[52] Liu Yang, Di Chai, Junxue Zhang, Yilun Jin, Leye Wang, Hao Liu, Han Tian, Qian Xu, and Kai Chen. 2023. A survey on vertical federated learning: From a layered perspective. *arXiv preprint arXiv:2304.01829* (2023).

[53] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology* 10, 2 (2019), 12:1–12:19. doi:10.1145/3298981

[54] Xubo Yue, Maher Nouiehed, and Raed Al Kontar. 2023. Gifair-fl: A framework for group and individual fairness in federated learning. *INFORMS Journal on Data Science* 2, 1 (2023), 10–23.

[55] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. 2019. Training individually fair ML models with sensitive subspace robustness. *arXiv preprint arXiv:1907.00020* (2019).

[56] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000* (2021).

[57] Richard Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, Vol. 28. PMLR, 325–333. https://proceedings.mlr.press/v28/zemel13.html

[58] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM, 335–340. doi:10.1145/3278721.3278779

[59] Junzhe Zhang and Elias Bareinboim. 2018. Equality of opportunity in classification: A causal approach. *Advances in Neural Information Processing Systems* 31 (2018).

[60] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in Decision-Making: The Causal Explanation Formula. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. 2037–2045. https://aaai.org/papers/11564-fairness-in-decision-making-the-causal-explanation-formula/

[61] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. 2022. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1433–1442.

[62] Aoqi Zuo, Susan Wei, Tongliang Liu, Bo Han, Kun Zhang, and Mingming Gong. 2022. Counterfactual fairness with partially known causal graph. *Advances in Neural Information Processing Systems* 35 (2022), 1238–1252.

APPENDICES

A Datasets and Client Partitions

We use three tabular benchmarks with a held-out protected attribute (used only for fairness/privacy evaluation). Dataset-level details are in Table A.1; vertical partitions are in Table A.2.

A.1 Overview of datasets

Preprocessing and protected-attribute operationalization choices below are documented in the spirit of dataset transparency practices [19].

German Credit (Banking). Binary credit risk (bad vs good); protected attribute: age binarised at 25; age excluded from inputs.

UCI Heart Disease (Healthcare). Binary heart disease presence; protected attribute: sex; sex excluded from inputs; remaining features standardised.

COMPAS Cox violent subset (Criminal justice). Binary violent recidivism within horizon; protected attribute: race (African American vs other); race excluded from inputs; categorical encoded, numerics imputed then scaled.

A.2 Client partitions and VFL views

Each dataset is split into three feature-holding clients plus a coordinating server; clients send only encodings. Concrete per-dataset client feature blocks are in Table A.2.

Table A.1. Dataset summary. Here n is the number of samples and d is the number of numerical features after preprocessing. Protected attributes are used only for fairness and privacy evaluation, and are never provided as inputs to the prediction models.

Dataset	Domain	n	d	Protected	Label	Splits
German Credit	Banking	≈ 1000	≈ 20	Age (< 25 vs ≥ 25)	Default (bad vs good)	IID and non-IID
Heart Disease	Healthcare	≈ 300	13	Sex (female vs male)	Presence of heart disease	IID and non-IID
COMPAS Cox	Criminal justice	≈ 6000	≈ 20	Race (African American vs other)	Violent recidivism event	IID and non-IID

A.3 Splitting strategies and non-IID settings

We use stratified train/val/test splits (70/30 train/test; within train, 20% validation). IID: stratified random by individuals. non-IID: shift protected-attribute distribution between train/test (race shift for COMPAS; sex shift for Heart Disease; analogous shift for German Credit).

A.4 Protected-attribute operationalization and shift rationale

Protected attribute thresholds. We operationalize the protected attribute by a fixed, dataset-standard binarization to enable auditable group metrics while keeping the attribute out of model inputs (Table A.1). For German Credit we use age < 25 vs. ≥ 25 as an interpretable "younger vs. older" split; the SCC-VFL pipeline is unchanged under alternative fixed cutoffs because it only requires re-defining s for evaluation and mask discovery.

Non-IID shift realism. Our Non-IID setting is a targeted demographic shift: we alter the train/test distribution of the protected attribute (race for COMPAS, sex for Heart Disease, analogous for German) while keeping the feature space, labels, and client partitions unchanged. This models a common deployment mismatch where

Table A.2. Vertical client partitions for each dataset. Each client holds a disjoint subset of feature groups and participates in the VFL protocol by sending local encodings to the server. The server coordinates training and aggregation but never observes raw features.

Dataset	Client	Feature groups (examples)
German Credit	Bank A	Account status, credit history, credit amount, credit duration, instalment rate
	Employer / Payroll	Employment duration, job category, housing, number of dependents
	Bureau / Platform	Savings accounts, collateral, other debtors or guarantors, foreign worker flag
Heart Disease	Hospital A	Core clinical measures such as resting blood pressure, serum cholesterol, ST depression, maximum heart rate
	Clinic B	Demographics and structural risk factors including age, chest pain type, vessel count, slope category
	Payer	Fasting blood sugar, resting ECG findings, thallium scan result, simple administrative codes
COMPAS Cox	Court system	Charge degree, offense type, prior counts, custody history, incarceration days
	Community services	Employment status, education level, supervision and program participation indicators
	Agency / Platform	Screening age, marital status, residence stability and other demographic records

training data is collected under institution- or region-specific coverage, but evaluation occurs on a different demographic mix, stressing counterfactual stability under group shift.

Partition and feature-family intuition. Client partitions (Table A.2) are chosen to mimic realistic institutional boundaries and to keep the mediator/proxy discussion interpretable: German Credit concentrates socio-economic and employment-related variables in parties that plausibly mediate age effects; Heart partitions separate clinical measures from demographics and payer-coded variables; COMPAS partitions separate court records from community services and platform demographics. These choices make it transparent which feature families can enter the editable set M and which can concentrate proxy signal P under the same discovery score.

B Additional Benchmark

We extend the experiments with a larger, multi-group sensitive-attribute benchmark using Folktables. We use the ACS 2018 1-Year Person survey (CA-only), subsample to 60k records, and follow the same evaluation protocol as the main paper: identical training pipeline and aggregation over 30 random seeds. We evaluate both IID and Non-IID settings.

Tasks and labels. We consider two Folktables prediction tasks: Income (binary label indicating whether income exceeds the task-defined threshold) and Mobility (binary mobility outcome as defined by Folktables).

Sensitive attribute. The sensitive attribute is race RAC1P, treated as multi-group and remapped to $s \in \{0, \ldots, K-1\}$ (typically $K = 9$ groups in this CA-only subset). As in the main paper, s is used only for fairness/privacy evaluation and counterfactual construction, and is not provided as an input feature to the predictor.

Table B.1. Folktables ACS 2018 (CA-only, 60k) Income under IID splits, mean±std over 30 seeds. Best per column in bold.

Method	Acc.↑	LogLoss↓	SCG↓	FR↓	DP↓	EO↓
Baseline	0.7954 ± 0.0031	0.4432 ± 0.0035	0.0261 ± 0.0092	0.7911 ± 0.3142	0.5822 ± 0.1663	0.6055 ± 0.1888
Uniform CF	0.7951 ± 0.0028	0.4435 ± 0.0031	0.0214 ± 0.0038	0.5409 ± 0.1263	0.5786 ± 0.1728	0.5815 ± 0.1773
Policy-blind	0.7951 ± 0.0032	0.4441 ± 0.0037	**0.0128 ± 0.0016**	0.3224 ± 0.0538	0.5826 ± 0.1712	0.5809 ± 0.1756
Server-only	0.7953 ± 0.0024	0.4477 ± 0.0031	0.0213 ± 0.0053	0.7328 ± 0.1637	**0.5759 ± 0.1717**	**0.5679 ± 0.1827**
SCC-VFL	**0.7987 ± 0.0033**	**0.4327 ± 0.0044**	0.0164 ± 0.0118	**0.2046 ± 0.1038**	0.5771 ± 0.1714	0.5746 ± 0.1883

Table B.2. Folktables ACS 2018 (CA-only, 60k) Mobility under IID splits, mean±std over 30 seeds. Best per column in bold.

Method	Acc.↑	LogLoss↓	SCG↓	FR↓	DP↓	EO↓
Baseline	0.7074 ± 0.0072	0.5738 ± 0.0041	0.0137 ± 0.0040	0.4631 ± 0.1849	0.1955 ± 0.0704	0.1336 ± 0.0503
Uniform CF	0.7055 ± 0.0092	0.5751 ± 0.0035	0.0105 ± 0.0026	0.3539 ± 0.1797	0.1788 ± 0.0827	0.1235 ± 0.0612
Policy-blind	0.7034 ± 0.0098	0.5752 ± 0.0045	0.0084 ± 0.0022	0.2528 ± 0.1472	**0.1505 ± 0.0817**	**0.1020 ± 0.0591**
Server-only	0.7022 ± 0.0102	0.5768 ± 0.0040	0.0114 ± 0.0035	0.3541 ± 0.2614	0.1624 ± 0.1672	0.1198 ± 0.1702
SCC-VFL	**0.7237 ± 0.0044**	**0.5451 ± 0.0030**	**0.0072 ± 0.0199**	**0.0220 ± 0.8754**	0.1689 ± 0.0433	0.1104 ± 0.0350

Table B.3. Folktables ACS 2018 (CA-only, 60k) Income under Non-IID splits, mean±std over 30 seeds. Best per column in bold.

Method	Acc.↑	LogLoss↓	SCG↓	FR↓	DP↓	EO↓
Baseline	0.7337 ± 0.0072	0.5325 ± 0.0085	0.0325 ± 0.0109	1.1957 ± 0.3764	0.6041 ± 0.1363	0.5355 ± 0.2910
Uniform CF	0.7327 ± 0.0077	0.5343 ± 0.0093	0.0206 ± 0.0031	0.6731 ± 0.1176	**0.5968 ± 0.1301**	0.5338 ± 0.2919
Policy-blind	0.7353 ± 0.0061	0.5316 ± 0.0069	0.0150 ± 0.0026	0.4704 ± 0.0990	0.6036 ± 0.1347	0.5452 ± 0.2804
Server-only	**0.7378 ± 0.0052**	**0.5267 ± 0.0057**	0.0248 ± 0.0064	1.0670 ± 0.2953	0.5994 ± 0.1374	0.5372 ± 0.2842
SCC-VFL	0.7314 ± 0.0112	0.5273 ± 0.0136	**0.0141 ± 0.0066**	**0.3665 ± 0.1614**	0.6016 ± 0.1294	**0.5020 ± 0.2611**

Splits and shifts. IID uses stratified random splits by individuals. Non-IID uses a protected-attribute distribution shift between train and test induced by reweighting/sampling across race groups, mirroring the main paper's shift-based protocol. For counterfactual evaluation with multi-group s, we sample an alternative group $s' \neq s$ per example and compute stability metrics under the corresponding intervention.

Metrics. We report utility (Accuracy, LogLoss), individual stability (SCG, FR), and group-fairness context metrics. Specifically, *Demographic Parity (DP) difference* is the absolute difference (or, for multi-group, the max pairwise gap) in positive prediction rates across sensitive groups, and *Equalized Odds (EO) gap* is the maximum absolute gap in TPR and FPR across groups.

Across both tasks, SCC-VFL attains the lowest flip rate in all four settings and remains competitive on utility. In IID, SCC-VFL achieves the best LogLoss in both tasks and the best Accuracy on Mobility (Table B.2), while substantially reducing decision instability (FR) relative to all baselines (Tables B.1–B.2). Under Non-IID shifts, SCC-VFL preserves low FR and achieves the best EO in both tasks (Tables B.3–B.4), indicating that the counterfactual stability objective remains well-controlled even under demographic distribution shift with multi-group s.

Table B.4. Folktables ACS 2018 (CA-only, 60k) Mobility under Non-IID splits, mean±std over 30 seeds. Best per column in bold.

Method	Acc.↑	LogLoss↓	SCG↓	FR↓	DP↓	EO↓
Baseline	0.6368 ± 0.0083	0.6613 ± 0.0050	0.0238 ± 0.0064	1.5080 ± 0.4802	0.4092 ± 0.0285	0.3132 ± 0.0184
Uniform CF	0.6362 ± 0.0087	0.6602 ± 0.0068	0.0157 ± 0.0037	0.7915 ± 0.1598	0.4060 ± 0.0214	0.3147 ± 0.0178
Policy-blind	0.6345 ± 0.0095	0.6601 ± 0.0073	**0.0121 ± 0.0023**	0.6357 ± 0.1536	0.4096 ± 0.0227	0.3175 ± 0.0185
Server-only	**0.6483 ± 0.0079**	**0.6462 ± 0.0063**	0.0225 ± 0.0087	1.5256 ± 0.6073	**0.3877 ± 0.0323**	0.2946 ± 0.0395
SCC-VFL	0.6320 ± 0.0184	0.6542 ± 0.0093	0.0133 ± 0.0280	**0.3567 ± 2.1679**	0.4024 ± 0.0463	**0.2941 ± 0.0350**

C Models and Hyperparameters

All methods share the same backbone, optimiser schedule, and partitions (Table A.2); differences are only in enabled components (Table C.2). Shared hyperparameters are in Table C.1.

C.1 Common architecture and client–server protocol

Client $c \in \{1, 2, 3\}$ holds $x_i^{(c)} \in \mathbb{R}^{d_c}$ and sends an encoding to the server:

$$h_i^{(c)} \;=\; h_\phi^{(c)}\Big(x_i^{(c)}\Big), \qquad z_i \;=\; \mathrm{Fuse}\Big(h_i^{(1)},\, h_i^{(2)},\, h_i^{(3)}\Big)\,. \tag{C.1}$$

We implement this as a 2-layer MLP over concatenated $x_i = [x_i^{(1)}, x_i^{(2)}, x_i^{(3)}]$ with fixed index ranges:

$$h_i \;=\; \sigma(W_2\,\sigma(W_1 x_i + b_1) + b_2)\,, \tag{C.2}$$

with ReLU σ. The server uses a linear classifier

$$f_\theta(z_i) \;=\; W_{\mathrm{cls}} h_i + b_{\mathrm{cls}}, \tag{C.3}$$

for binary $y_i \in \{0, 1\}$.

C.2 Mediator and proxy discovery

Server forms a DP sketch of the protected attribute:

$$z_s \;=\; \psi(s). \tag{C.4}$$

For each feature coordinate j, compute (i) signed probability gap ΔP_j and (ii) HSIC_j, then rank by

$$S_j \;=\; f\big(|\Delta P_j|, \mathrm{HSIC}_j\big), \tag{C.5}$$

and derive per-party sets $(M^{(p)}, P^{(p)}, N^{(p)})$ via fractions (ρ_M, ρ_P):

$$P^{(p)} \;=\; \Big\{ j \in M^{(p)} : S_j \text{ lies in the top } \rho_P \text{ of } \{S_k : k \in M^{(p)}\} \Big\}. \tag{C.6}$$

Fractions are selected on validation, then fixed across seeds/methods for that dataset.

Governance note. The partition into permissible mediators M and impermissible proxies $P \subset M$ is a normative policy choice; we make this choice auditable via a concrete protocol in Appendix D.

C.3 Differential privacy and secure aggregation

Scope and threat model. We use differential privacy *only* to privatize the released sketch used for mediator/proxy screening in mask discovery. This provides record-level protection for the contingency tables (and any downstream statistics computed from them) against an observer who sees the released sketch. It does *not* provide end-to-end DP for learned representations, model parameters, or gradients, and it does not rule out inference attacks outside the evaluated threat model.

DP sketch for mediator discovery. A trusted holder of the sensitive attribute (or a trusted enclave) computes and releases the DP sketch; the coordinating server and other parties never observe raw s. For feature j, discretise into K bins and build contingency table $c_j \in \mathbb{R}^{2\times K}$, clip:

$$\bar{c}_j \;=\; c_j \cdot \min\!\left(1, \frac{S}{\|c_j\|_2}\right), \tag{C.7}$$

then add Gaussian noise:

$$\tilde{c}_j \;=\; \bar{c}_j + \mathcal{N}\!\left(0, \sigma^2_{\text{sketch}} S^2 I\right), \tag{C.8}$$

and compute $\Delta P_j, \mathrm{HSIC}_j$ from $\tilde{c}_j$.

Proof of Proposition 3.1. Under addition or removal of one record, the contingency table $c_j \in \mathbb{R}^{2\times K}$ changes in exactly one cell by ± 1, so the ℓ_2-sensitivity of the unclipped table is 1. After clipping (Eq. (C.7)), the ℓ_2-norm of $\bar{c}_j$ is bounded by S, and the ℓ_2-sensitivity of the clipped table under neighboring datasets is at most S (since each record contributes at most S to the clipped output). The Gaussian mechanism (Eq. (C.8)) adds noise with standard deviation $\sigma_{\text{sketch}} \cdot S$ per coordinate. By the standard Gaussian mechanism guarantee (Theorem A.1 of Dwork and Roth [14]), releasing $\tilde{c}_j$ satisfies (ε, δ)-DP with $\varepsilon = (1/\sigma_{\text{sketch}})\sqrt{2\ln(1.25/\delta)}$ for any $\delta \in (0, 1)$. □

When masks are refreshed, the same mechanism is re-run; we cap the number of refreshes, and each refresh consumes an additional (ε, δ) budget. A full composition accountant (e.g., via Rényi DP or the moments accountant) is outside the scope of this work but is a natural extension.

Secure aggregation. Secure aggregation protects per-client updates from being inspected individually by the coordinating server. Client gradient $g_t^{(c)}$ is masked as

$$u_t^{(c)} \;=\; g_t^{(c)} + r_t^{(c)}, \tag{C.9}$$

with masks satisfying $\sum_{c=1}^{m} r_t^{(c)} = 0$, so the server recovers only

$$\sum_{c=1}^{m} u_t^{(c)} \;=\; \sum_{c=1}^{m} g_t^{(c)}. \tag{C.10}$$

We add no DP noise to training gradients or model updates in the main experiments; DP is applied only to the mask-discovery sketch above.

C.4 SCC–VFL model

Masked conditional generators. Client generator edits only mediators:

$$x_M^{\text{cf},(c)} \;=\; g_\varphi^{(c)}\!\left(x_N^{(c)}, x_M^{(c)}, c_i, s'\right), \qquad x_N^{\text{cf},(c)} \;=\; x_N^{(c)}. \tag{C.11}$$

We instantiate $g_\varphi^{(c)}$ as a cVAE:

$$z^{(c)} \sim q_\varphi^{(c)}\!\left(z^{(c)} \mid x_N^{(c)}, x_M^{(c)}, s\right), \qquad x_M^{\text{cf},(c)} = d_\varphi^{(c)}\!\left(z^{(c)}, x_N^{(c)}, c_i, s'\right). \tag{C.12}$$

Table C.1. Shared architectural and optimisation hyperparameters per dataset. These values are held fixed across SCC–VFL and all baselines.

Setting	German Credit	Heart Disease	COMPAS Cox
# feature clients m	3	3	3
Encoder hidden dim d_z	64	32	32
Encoder layers	2	2	2
Dropout (encoder)	0.05	0.10	0.10
Classifier output units	2	2	2
Generator type	cVAE	cVAE	cVAE
Generator latent dim	8	16	8
Optimiser (backbone)	AdamW	Adam	Adam
Learning rate (backbone)	0.015	0.010	0.005
Learning rate (generators)	0.010	0.010	0.005
Learning rate (adversary)	0.005	0.005	0.003
Batch size	full batch	full batch	full batch
Max epochs (baselines)	300	200	80
Max epochs (SCC–VFL)	300	200	150
Warmup epochs (no L_{cons})	40	30	25
Early stopping patience	35	25	25
Number of seeds	30	30	30

Proxy adversaries and gradient reversal. Per-party adversary with GRL:

$$\tilde{u}_i^{(p)} = \mathrm{GRL}_{\lambda_{\text{grl}}}\big(u_i^{(p)}\big), \qquad \hat{s}_i^{(p)} = a_\omega^{(p)}\big(\tilde{u}_i^{(p)}\big), \tag{C.13}$$

and total adversarial loss

$$L_{\text{adv}} = \sum_{p=1}^{m} L_{\text{adv}}^{(p)}. \tag{C.14}$$

Objective.

$$L = L_{\text{cls}} + \lambda_{\text{cons}} L_{\text{cons}} + \lambda_{\text{gen}} L_{\text{gen}} + \lambda_{\text{adv}} L_{\text{adv}}. \tag{C.15}$$

C.5 Baseline implementations

Baselines are ablations: same backbone Eqs. (C.2)–(C.3), same partitions (Table A.2), same hyperparameters (Table C.1), and activation pattern in Table C.2.

Adv-NoMask. Task loss plus adversarial debiasing (no mask discovery; adversary active).

Uniform-CF. Generators edit all coordinates; uses $L_{\text{cls}} + \lambda_{\text{cons}} L_{\text{cons}}$; no adversary.

Policy-blind mask. Uses mediator mask M but no proxy handling; generators and consistency active; adversary off.

Server only consistency. No generators; server perturbs/shuffles mediator coordinates and applies consistency; adversary off.

C.6 Hyperparameter settings

Shared settings are in Table C.1. Active components per method are in Table C.2.

Table C.2. Loss terms and generator masks per method. All methods share the same backbone and optimiser hyperparameters; only the active components differ.

Method	Gen. mask	L_{cls}	L_{cons}	L_{gen}	L_{adv}	Adv. on h
Adv-NoMask	none	✓	×	×	✓	✓
Uniform CF	all coords	✓	✓	✓	×	×
Policy blind mask	M	✓	✓	✓	×	×
Server only consistency	random M shuffle	✓	✓	×	×	×
SCC–VFL (ours)	M per client	✓	✓	✓	✓	✓

D Policy Specification and Audit Protocol for $N/M/P$

SCC-VFL distinguishes *editable mediators* M from *impermissible proxies* $P \subset M$. This split is a normative policy decision: M encodes which descendants of s are considered acceptable pathways for counterfactual edits, while P marks descendants judged unacceptable to rely on (even if predictive). We therefore make the $N/M/P$ specification explicit, versioned, and contestable.

Domain examples (illustrative). **German Credit (protected: age).** A lending policy board may treat financially grounded, action-relevant factors as permissible mediators (e.g., repayment history signals) while flagging demographic-adjacent or structurally discriminatory signals as proxies if they encode age-related opportunity differences. The Policy Card records which feature groups are placed in M vs. P, and why, and documents the appeal path for disputed assignments.

Heart Disease (protected: sex). A clinical governance group may allow physiological variables that plausibly sit on causal pathways for disease risk as mediators, while treating administrative or measurement artifacts that encode sex in a non-clinical way as proxies. Reviews are triggered when clinical guidelines change or when the patient population shifts.

COMPAS Cox (protected: race). A justice-policy panel may classify variables tied to enforcement intensity or structural inequities as proxies (even if predictive), while allowing narrowly defined, policy-justified pathways as mediators if explicitly permitted by the deployment context. The appeal process supports challenges to specific role assignments and logs any updates under a new policy version.

Policy Specification and Audit Protocol (auditable $N/M/P$)

A1. Policy owner (who decides). A named governance group defines and approves the mediator/proxy policy: regulator or compliance lead (if applicable, consistent with non-discrimination law [51]), an institutional policy board, domain experts, and at least one affected-stakeholder representative.
A2. Inputs (what is being decided). A dataset-specific *feature dictionary* (definitions, measurement process, manipulability, known confounders, and known correlations with s) and the SCC-VFL discovery outputs (DP-ranked coordinates).
A3. Evidence (what counts). Accepted evidence includes written statutes or institutional policy, domain guidelines, documented stakeholder input, and empirical checks (feature meaning, stability under shift, and whether a feature is a plausible descendant of s in the application context).
A4. Decision rule (how $N/M/P$ is produced). Start from the DP-ranked candidates as *candidate descendants*. The policy owner assigns: (i) N: coordinates treated as non-descendants (held fixed), (ii) M: permissible descendants (editable under CF), (iii) $P \subset M$: impermissible descendants (treated as leakage-risk and not relied on). Any override to the ranking must include a written rationale.
A5. Artifacts (what is logged). Create a Policy Card per dataset, modeled after Model Cards [37] and Datasheets for Datasets [19], with: version ID, date, owners, rationale per role decision, evidence links, and a machine-readable $N/M/P$ mask file used in training and evaluation.
A6. Contestability (appeals). Provide a documented process for an individual or auditor to dispute a role assignment. The appeal triggers: (i) review of the feature dictionary entry, (ii) re-evaluation of evidence, and (iii) a logged decision with an updated version if changed.
A7. Review cadence (when it is revisited). Re-approve the policy on a fixed schedule (e.g., quarterly) and additionally upon major distribution shift, feature set changes, or policy changes. Each review re-runs discovery on the new data slice and records deltas in $N/M/P$.

E Privacy and Robustness Evaluation

This appendix evaluates (i) *representation leakage* via an attribute inference attacker trained on latent vectors, and (ii) *input robustness* via a constrained adversary that perturbs only a designated feature subspace. Threat model and notation are in Section E; AIA in Section E.2; subspace-PGD in Section E.3; results are reported in Tables E.2 and E.3.

E.1 Threat Model and Notation

We consider a trained model with encoder h_ϕ and prediction head f_θ under the VFL threat surface that spans feature inference [25, 33] and label inference [17] attacks. For entity i with input x_i, the encoder produces a latent representation $h_i \in \mathbb{R}^{d_h}$ and the head outputs a predictive distribution over labels $\mathcal{Y}$:

$$h_i = h_\phi(x_i) \in \mathbb{R}^{d_h}, \quad \hat{y}_i = f_\theta(h_i) \in \Delta(\mathcal{Y}). \tag{E.1}$$

The protected attribute is $s_i \in \{0, 1\}$. We use $\hat{y}_i$ (overloading notation) for the predicted class obtained by the argmax over class probabilities:

$$\hat{y}_i = \arg\max_y f_\theta(h_i)_y. \tag{E.2}$$

The adversary is post-hoc (after training) and is evaluated by success rate (SR, in %) where higher SR indicates stronger attack and thus weaker privacy/robustness.

E.2 Attribute Inference Attack (AIA)

AIA measures how much information about s remains in the learned representation h_i. The attacker trains a binary classifier

$$g_\psi : \mathbb{R}^{d_h} \to [0, 1], \tag{E.3}$$

where $g_\psi(h_i)$ estimates $\Pr(s_i = 1 \mid h_i)$. Given attacker training data $\{(h_i, s_i)\}_{i=1}^n$, parameters ψ are fit by minimizing binary cross-entropy:

$$\mathcal{L}_{\text{AIA}}(\psi) = \frac{1}{n} \sum_{i=1}^{n} \ell_{\text{BCE}}\big(g_\psi(h_i), s_i\big), \quad \ell_{\text{BCE}}(p, s) = -s \log p - (1 - s) \log(1 - p). \tag{E.4}$$

The attacker predicts $\hat{s}_i$ by thresholding at $1/2$:

$$\hat{s}_i = \mathbb{I}\{g_\psi(h_i) \geq 1/2\}. \tag{E.5}$$

For an attacker training budget T (epochs), we report the attack success rate on a balanced test split $\mathcal{I}_{\text{test}}$:

$$\text{AIA-SR}(T) = \frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{i \in \mathcal{I}_{\text{test}}} \mathbb{I}\{\hat{s}_i^{(T)} = s_i\} \times 100\,\%. \tag{E.6}$$

As a reference, AIA-SR* denotes the Bayes-optimal success rate achievable by the optimal classifier $g^\star$ under the induced representation distribution:

$$\text{AIA-SR}^\star = \mathbb{E}\big[\mathbb{I}\{g^\star(h) = s\}\big]. \tag{E.7}$$

Lower AIA-SR indicates less s-signal in h.

E.3 Subspace-Constrained PGD Attack

Subspace-PGD measures robustness to targeted feature manipulation. The attacker chooses a perturbation δ_i that is (i) supported only on a designated coordinate subset S (the "attackable" subspace) and (ii) bounded in ℓ_∞ norm by ϵ:

$$\text{supp}(\delta_i) \subseteq S, \qquad \|\delta_i\|_\infty \leq \epsilon. \tag{E.8}$$

Starting from $\delta_i^{(0)} = 0$, PGD iteratively updates δ_i using signed gradients of the task loss (with respect to the input) and then projects back to the feasible set:

$$\delta_i^{(t+1)} = \Pi_{C_\epsilon}\Big(\delta_i^{(t)} + \alpha\, \text{sign}\big(\nabla_x \ell_{\text{task}}(f_\theta(h_\phi(x_i + \delta_i^{(t)})), \hat{y}_i)\big)\Big), \tag{E.9}$$

where α is the step size and Π_{C_ϵ} projects onto

$$C_\epsilon = \big\{\delta \in \mathbb{R}^{d_x} : \text{supp}(\delta) \subseteq S,\ \|\delta\|_\infty \leq \epsilon\big\}. \tag{E.10}$$

After T_{PGD} steps, the adversarial example is

$$x_i^{\text{adv}} = x_i + \delta_i^{(T_{\text{PGD}})}. \tag{E.11}$$

The PGD success rate reports the fraction of inputs whose predicted class changes under attack:

$$\text{PGD-SR}(\epsilon) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\hat{y}_i^{\text{adv}} \neq \hat{y}_i\} \times 100\,\%. \tag{E.12}$$

Lower PGD-SR indicates stronger robustness to targeted subspace perturbations.

E.4 Implementation Details

Table E.1 summarises the attack implementations and evaluation settings used for AIA and Subspace-PGD.

Table E.1. Attack implementation details and evaluation protocol.

Component	Setting
AIA data	Balanced split over s; compute and freeze representations $h_\phi(x)$
AIA attacker	Train g_ψ for $T \in \{10, 20, 40, 80\}$ epochs
AIA metric	Report AIA-SR via Eq. (E.6), aggregated over 30 seeds
PGD subspace	Set sensitive subspace $S = M$
PGD steps	$T_{\text{PGD}} = 20$ with step size $\alpha = \epsilon/5$
PGD radii	$\epsilon \in \{0.02, 0.05, 0.10, 0.20\}$
PGD metric	Report PGD-SR via Eq. (E.12), aggregated over 30 seeds

Table E.2. Attribute inference attack success rate (AIA-SR, %) as a function of attacker training budget (epochs), averaged over 30 seeds. Lower is better.

Method	**10 epochs**	**20 epochs**	**40 epochs**	**80 epochs**
Adv-NoMask	54.03 ± 4.57	56.15 ± 5.35	59.06 ± 6.00	62.52 ± 5.55
Uniform-CF	52.64 ± 4.31	55.41 ± 4.61	57.93 ± 4.63	62.51 ± 4.26
Policy-blind Mask	53.11 ± 5.03	55.47 ± 5.55	58.35 ± 6.50	62.04 ± 5.71
Server-Consistency	52.84 ± 4.03	53.49 ± 3.98	59.79 ± 5.42	62.27 ± 4.91
SCC-VFL (ours)	53.08 ± 4.38	53.45 ± 4.67	56.17 ± 6.39	57.89 ± 5.72

Table E.3. Subspace-constrained PGD attack success rate (PGD-SR, %) as a function of perturbation radius ϵ, averaged over 30 seeds. Lower is better.

Method	$\epsilon = 0.02$	$\epsilon = 0.05$	$\epsilon = 0.10$	$\epsilon = 0.20$
Adv-NoMask	2.24 ± 0.90	5.97 ± 1.43	11.41 ± 1.80	21.67 ± 2.85
Uniform-CF	1.94 ± 0.81	5.14 ± 1.08	10.36 ± 1.64	20.64 ± 3.05
Policy-blind Mask	2.09 ± 0.67	5.36 ± 1.41	10.37 ± 1.96	20.11 ± 2.75
Server-Consistency	1.28 ± 0.57	3.50 ± 1.15	7.54 ± 1.85	15.02 ± 2.89
SCC-VFL (ours)	1.03 ± 0.61	2.44 ± 0.93	4.53 ± 1.40	9.30 ± 2.29

E.5 Empirical Results

AIA-SR is reported in Table E.2. PGD-SR is reported in Table E.3.

F Runtime and Communication Overhead

We report wall-clock runtime and communication for SCC-VFL on German Credit (IID), aggregated as mean±std over 30 random seeds (CPU runs). Over training, SCC-VFL requires 1.0251 ± 0.2233 seconds total, corresponding to 0.0218 ± 0.0032 seconds per epoch and 7.1015 ± 1.0378 ms per optimization step, with 40.1333 ± 4.7380 epochs executed. Communication totals (sum over training) are 2.87 MB for model-side messages and 124.69–146.51 MB for feature-side messages across the shown seeds (batch size 256); the per-seed table indicates a stable model communication footprint and feature communication that scales with the number of steps/epochs. These measurements quantify the additional passes and per-party generator components in SCC-VFL in a cross-silo

Table G.1. Sensitivity to mediator mask threshold τ_M (top_frac). Other hyperparameters fixed.

Setting	Acc.↑	LogLoss↓	SCG↓	FR(%)↓
$\tau_M = 0.15$	0.7353 ± 0.0138	0.5738 ± 0.0117	0.0051 ± 0.0041	0.13 ± 0.16
$\tau_M = 0.25$	0.7020 ± 0.0045	0.5792 ± 0.0195	0.0068 ± 0.0074	0.13 ± 0.16
$\tau_M = 0.33$	0.7187 ± 0.0233	0.5780 ± 0.0284	**0.0027 ± 0.0011**	**0.00 ± 0.00**
$\tau_M = 0.45$	0.7180 ± 0.0088	0.5742 ± 0.0148	0.0083 ± 0.0067	0.33 ± 0.30
$\tau_M = 0.60$	0.7140 ± 0.0083	**0.5714 ± 0.0130**	0.0053 ± 0.0056	0.27 ± 0.39

Table G.2. Sensitivity to counterfactual edit magnitude γ_{cf} (cf_scale). Other hyperparameters fixed.

Setting	Acc.↑	LogLoss↓	SCG↓	FR(%)↓
$\gamma_{\text{cf}} = 0.05$	0.7187 ± 0.0233	0.5780 ± 0.0284	**0.0015 ± 0.0008**	**0.00 ± 0.00**
$\gamma_{\text{cf}} = 0.10$	0.7187 ± 0.0233	0.5780 ± 0.0284	0.0019 ± 0.0010	**0.00 ± 0.00**
$\gamma_{\text{cf}} = 0.20$	0.7187 ± 0.0233	0.5780 ± 0.0284	0.0027 ± 0.0011	**0.00 ± 0.00**
$\gamma_{\text{cf}} = 0.35$	0.7187 ± 0.0233	0.5780 ± 0.0284	0.0039 ± 0.0011	0.07 ± 0.13
$\gamma_{\text{cf}} = 0.50$	0.7187 ± 0.0233	0.5780 ± 0.0284	0.0050 ± 0.0011	0.07 ± 0.13

style setting: while model communication remains small, feature-side traffic dominates and should be considered when bandwidth is constrained.

G Sensitivity Analysis

German Credit sensitivity is studied under the same protocol as the main results (fixed split sizes, identical pipeline, 30 seeds). Each sweep changes one hyperparameter; report mean±std of Accuracy, LogLoss, SCG, and FR(%). Tables G.1–G.5 contain all results.

G.1 Mediator Mask Threshold Sensitivity (τ_M)

Mediator set (top-τ_M by DP ranking score S_j):

$$M \;=\; \{j \in [d] \;:\; S_j \text{ is in the top } \tau_M \text{ fraction of } \{S_\ell\}_{\ell=1}^{d}\}, \tag{G.1}$$

with S_j defined from $|\Delta P_j|$ and HSIC_j (Section 3.2). Table G.1 shows moderate utility variation and consistently low SCG/FR; $\tau_M = 0.33$ minimizes SCG and yields FR= 0 (within resolution).

G.2 Counterfactual Edit Magnitude Sensitivity (γ_{cf})

Edit scale for mediator-only perturbations:

$$\tilde{x} \;=\; x \;+\; \gamma_{\text{cf}} \cdot \Pi_M\big(\Delta(x)\big), \tag{G.2}$$

where Π_M projects to mediator coordinates and $\Delta(x)$ is generator output. Table G.2 shows utility invariant across the tested range; SCG increases smoothly with γ_{cf}; FR remains near zero until the largest scales.

G.3 Consistency Weight Sensitivity (λ_{cons})

Objective with scheduled max weight λ_{cons}:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{task}} \;+\; \lambda_{\text{cons}} \cdot \mathcal{L}_{\text{cons}} \;+\; \lambda_{\text{adv}} \cdot \mathcal{L}_{\text{adv}} \;+\; \lambda_{\text{gen}} \cdot \mathcal{L}_{\text{gen}}, \tag{G.3}$$

Table G.3. Sensitivity to consistency weight λ_{cons} (lam_cons_max). Other hyperparameters fixed.

Setting	Acc.↑	LogLoss↓	SCG↓	FR(%)↓
$\lambda_{\text{cons}} = 0.0$	0.7187 ± 0.0233	0.5781 ± 0.0285	**0.0017 ± 0.0002**	**0.00 ± 0.00**
$\lambda_{\text{cons}} = 0.4$	0.7187 ± 0.0233	0.5780 ± 0.0284	0.0020 ± 0.0003	**0.00 ± 0.00**
$\lambda_{\text{cons}} = 0.8$	0.7187 ± 0.0233	0.5780 ± 0.0284	0.0024 ± 0.0007	**0.00 ± 0.00**
$\lambda_{\text{cons}} = 1.2$	0.7187 ± 0.0233	0.5780 ± 0.0284	0.0027 ± 0.0011	**0.00 ± 0.00**
$\lambda_{\text{cons}} = 1.6$	0.7187 ± 0.0233	0.5780 ± 0.0284	0.0031 ± 0.0014	**0.00 ± 0.00**

Table G.4. Sensitivity to adversary weight λ_{adv}. Other hyperparameters fixed.

Setting	Acc.↑	LogLoss↓	SCG↓	FR(%)↓
$\lambda_{\text{adv}} = 0.00$	0.7120 ± 0.0150	0.5777 ± 0.0185	**0.0026 ± 0.0012**	**0.00 ± 0.00**
$\lambda_{\text{adv}} = 0.01$	**0.7220 ± 0.0173**	**0.5771 ± 0.0176**	0.0027 ± 0.0013	**0.00 ± 0.00**
$\lambda_{\text{adv}} = 0.03$	0.7187 ± 0.0233	0.5780 ± 0.0284	0.0027 ± 0.0011	**0.00 ± 0.00**
$\lambda_{\text{adv}} = 0.06$	0.7120 ± 0.0133	0.5741 ± 0.0225	0.0028 ± 0.0013	0.07 ± 0.13
$\lambda_{\text{adv}} = 0.10$	0.7080 ± 0.0096	0.5869 ± 0.0115	0.0031 ± 0.0020	0.13 ± 0.16

Table G.5. Sensitivity to generator regularization λ_{gen} (lam_gen). Other hyperparameters fixed.

Setting	Acc.↑	LogLoss↓	SCG↓	FR(%)↓
$\lambda_{\text{gen}} = 0.000$	0.7207 ± 0.0241	0.5786 ± 0.0281	0.0294 ± 0.0084	0.93 ± 0.44
$\lambda_{\text{gen}} = 0.005$	0.7187 ± 0.0233	0.5780 ± 0.0284	0.0037 ± 0.0019	**0.00 ± 0.00**
$\lambda_{\text{gen}} = 0.010$	0.7187 ± 0.0233	0.5780 ± 0.0284	0.0027 ± 0.0011	**0.00 ± 0.00**
$\lambda_{\text{gen}} = 0.020$	0.7187 ± 0.0233	0.5780 ± 0.0284	0.0021 ± 0.0006	**0.00 ± 0.00**
$\lambda_{\text{gen}} = 0.050$	0.7187 ± 0.0233	0.5780 ± 0.0284	**0.0018 ± 0.0002**	**0.00 ± 0.00**

where $\mathcal{L}_{\text{cons}}$ enforces agreement under masked edits. Table G.3 shows utility stable; SCG rises mildly with λ_{cons}; FR stays at 0.

G.4 Adversary Weight Sensitivity (λ_{adv})

We vary λ_{adv} holding others fixed. Table G.4 shows low-to-moderate λ_{adv} maintains FR= 0 and good utility; larger λ_{adv} slightly increases FR and can reduce utility.

G.5 Generator Regularization Sensitivity (λ_{gen})

We vary λ_{gen} holding others fixed. Table G.5 shows $\lambda_{\text{gen}} = 0$ yields large SCG and non-zero FR; small regularization restores FR= 0 and reduces SCG without hurting utility.

Overall (Tables G.1–G.5), utility is stable across sweeps; SCG changes smoothly with edit strength/consistency; λ_{gen} is the main safeguard against high SCG and non-zero FR.

H Qualitative Analysis

German Credit IID qualitative views (representative seed 0; patterns consistent with 30-seed aggregates). Figures H.1–H.3 and Table H.1 summarize mask roles, latent-space structure, and training dynamics.

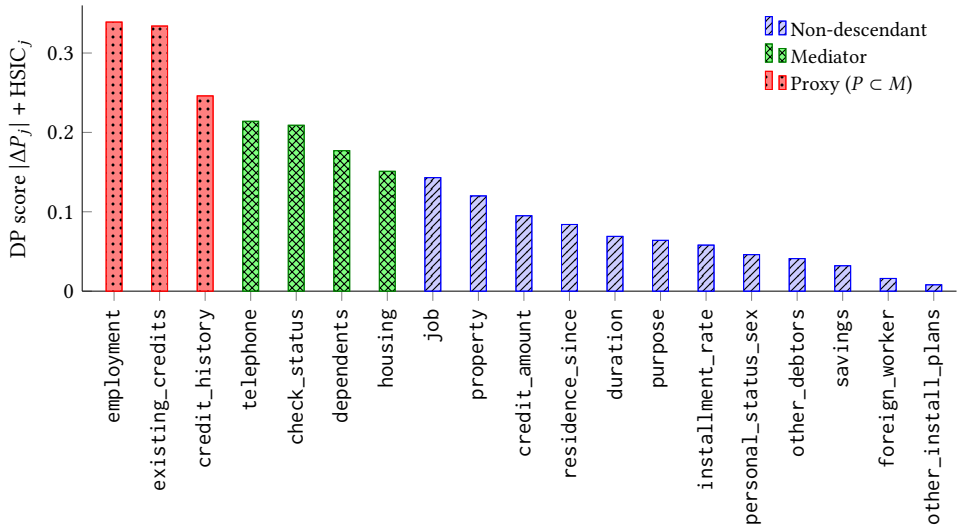


Fig. H.1. Differentially private mask scores $|\Delta P_j| + \mathrm{HSIC}_j$ for German Credit features, colored by SCC-VFL mask roles: non descendants N (blue), mediators M (green), and proxies $P \subset M$ (red).

H.1 Feature roles and DP sketch scores

We rank features by the DP score $|\Delta P_j| + \mathrm{HSIC}_j$ (Section 3.2) and assign roles N, M, and $P \subset M$. Figure H.1 shows the highest scores as proxies (red), mid as mediators (green), and low as non-descendants (blue).

H.2 Mask interpretability summary

Table H.1 reports a representative ranking of features by $\mathrm{MI}(x_j, s)$ together with their assigned roles. High-MI coordinates are consistently categorized as proxies (P), while mid-range coordinates fall into mediators (M). Features with low or near-zero MI are assigned to non-descendants (N). This monotone alignment between dependence strength and role supports the intended semantics of $P/M/N$ and indicates the mask is not arbitrary.

H.3 Latent representation PCA views

Figure H.2 compares PCA projections across methods; SCC-VFL shows the strongest mixing between s groups, aligning with reduced sensitive signal in the fused representation.

H.4 Training dynamics of SCC-VFL

Across German Credit IID experiments aggregated over 30 seeds, we inspect SCC-VFL training curves to verify that the gains reflect stable optimization rather than a single favorable run. Figure H.3 shows smooth loss convergence and consistently competitive validation accuracy, while SCG increases and then saturates once the consistency term activates and FR remains low, indicating SCC-VFL primarily refines logits and confidence under counterfactual edits instead of causing frequent label flips.

H.5 Counterfactual case studies across methods

Instance-level summaries (seed 0) illustrating that SCC-VFL typically preserves labels with smaller, structured mediator edits, while some baselines require larger shifts and can flip labels on borderline cases.

Table H.1. Example mask interpretation on German Credit: high MI demographic variables fall into proxies P, mid MI structural variables into mediators M, and low MI attributes into non-descendants N.

Feature name	MI(x_j, s)	Role
personal_status_sex	0.0539	Proxy (P)
employment	0.0476	Proxy (P)
housing	0.0471	Proxy (P)
other_install_plans	0.0335	Mediator (M)
property	0.0267	Mediator (M)
purpose	0.0226	Mediator (M)
existing_credits	0.0215	Mediator (M)
residence_since	0.0181	Non-desc (N)
dependents	0.0170	Non-desc (N)
credit_amount	0.0000	Non-desc (N)

Case study 1: correctly accepted applicant, $y = 0, s = 1$

Top standardized features: credit_amount, duration, purpose (M), employment (P), other_install_plans (M). Baseline and all fairness baselines keep the label $y = 0$ but Uniform-CF requires a large mediator shift $\|\Delta_M\| \approx 0.52$. SCC-VFL keeps the label stable with a smaller and more structured change $\|\Delta_M\| \approx 0.25$, mainly nudging employment and installment related mediators.

Case study 2: ambiguous low-risk applicant, $y = 0, s = 1$

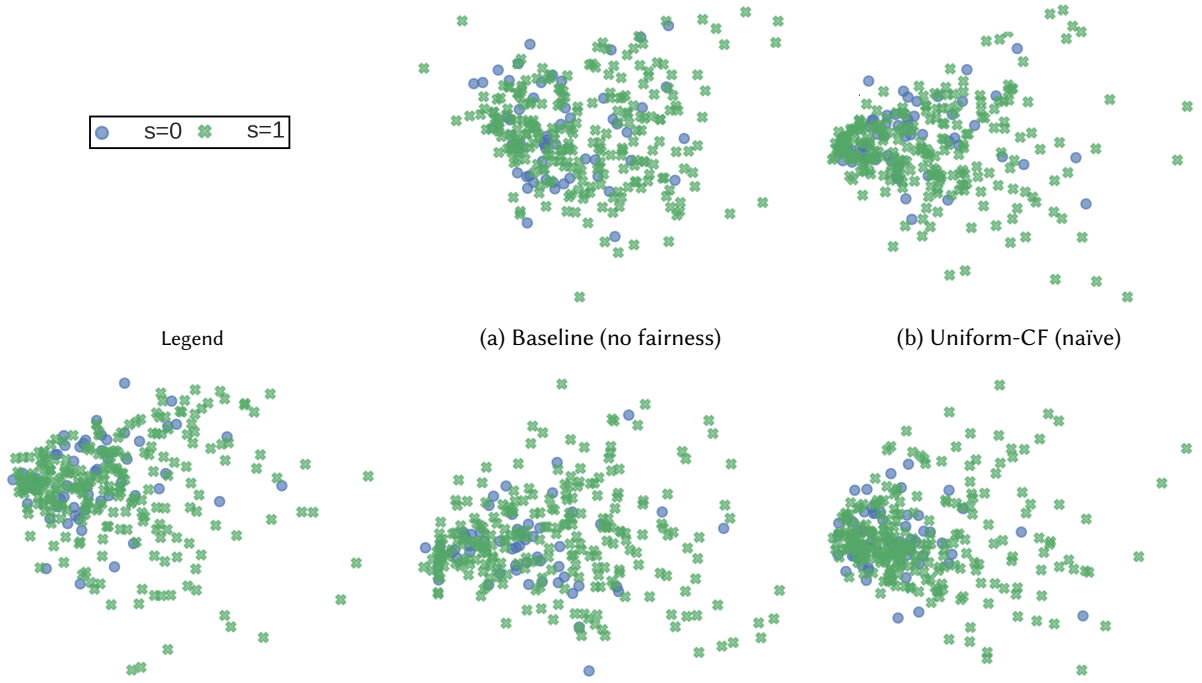
The baseline predicts $y = 0$ with low confidence, while Server-Consistency flips to $y = 1$ with high confidence, revealing residual dependence on s. Uniform-CF and Policy-blind Mask leave $y = 0$ but change probabilities with large $\|\Delta_M\|$. SCC-VFL preserves the decision and adjusts confidence in a moderate way, indicating that its generator learns a minimal recourse-style edit rather than over-correcting. This pattern is typical across similar borderline cases sampled over multiple seeds.

Case study 3: high-risk applicant, $y = 1, s = 1$

Uniform-CF flips the label from 1 to 0 under strong mediator perturbations, which corresponds to a brittle counterfactual. Policy-blind Mask and Server-Consistency keep $y = 1$ but still rely on large or opaque changes. SCC-VFL maintains the positive decision with moderate mediator edits, confirming that it does not use counterfactuals to hide risk but instead enforces stability around the original decision.

H.6 Recourse-style mediator edits in SCC-VFL

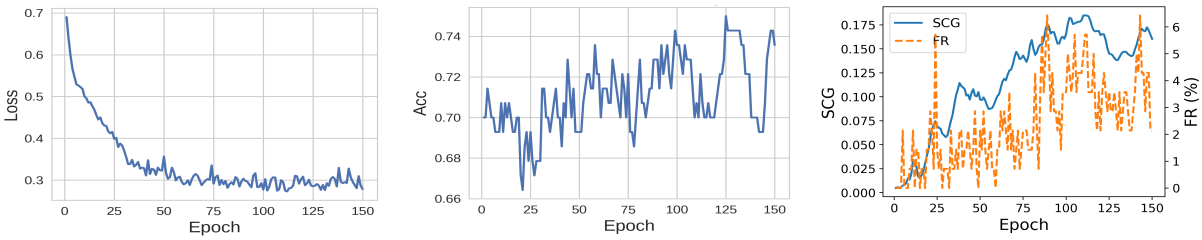
Seed-0 examples highlighting that SCC-VFL edits concentrate on mediator coordinates while leaving non-descendants fixed, producing small, policy-aligned directions of change reminiscent of actionable recourse [45] and counterfactual explanations [47].



Legend (a) Baseline (no fairness) (b) Uniform-CF (naïve)

(c) Policy-blind Mask (d) Server-Consistency (e) SCC–VFL (ours)

Fig. H.2. PCA of fused representations colored by the sensitive-group indicator s for all methods on German Credit (IID, seed 0), where $s = 0$ and $s = 1$ denote the two groups defined by the sensitive attribute. SCC–VFL shows the strongest mixing between the two groups while preserving a coherent manifold, suggesting reduced sensitive-attribute signal relative to the baselines.



(a) Training loss (b) Validation accuracy (c) SCG and FR over epochs

Fig. H.3. SCC–VFL training dynamics on German Credit IID: train loss, validation accuracy, and the evolution of SCG and FR over epochs. The model converges to a regime with low loss, strong accuracy, and low flip rate while maintaining non-trivial counterfactual consistency.

Recourse example A: low-risk, $\hat{y} = 0$

SCC-VFL slightly increases `employment` tenure and `housing` quality, and reduces `existing_credits`, while keeping non-descendants such as `credit_amount` and `duration` fixed. These changes resemble realistic recourse actions that a bank might recommend to improve future credit decisions without directly editing demographic attributes, and we observe similar patterns for other low-risk clients.

Recourse example B: high-risk, $\hat{y} = 1$

For a predicted defaulter, SCC-VFL proposes small shifts in `housing`, `property`, and `other_install_plans`, again leaving non-descendants unchanged. The model preserves the current label but provides a concrete direction for improvement in mediator space, illustrating that SCC-VFL implements policy-aligned, on-support counterfactuals rather than arbitrary feature perturbations. Across random draws, mediator edits stay within similar magnitude ranges, which matches the low flip rates reported in the main tables.

I Worked Example: Counterfactual Enforcement in Credit Decisions

This appendix provides an illustrative example of how SCC-VFL enforces selective counterfactual consistency in a credit decision setting. The example is intended to clarify the mechanics of mediator editing and server-side enforcement and does not constitute a normative judgment about which features should be considered permissible or impermissible in practice.

Setup. Consider a loan applicant whose features are vertically partitioned across institutions. The protected attribute is age, which is not provided as a model input and is held by a trusted party. A policy review specifies the following feature roles:

- Non-descendants N: loan amount and loan duration
- Permissible mediators M: employment tenure and credit history
- Impermissible proxies P: housing status and number of dependents

The applicant is initially observed with age corresponding to a younger group.

Counterfactual generation. To evaluate the counterfactual scenario in which the applicant is older, the server provides a target sensitive embedding corresponding to the older age group. Each party applies the masked generator as follows. Non-descendants in N are copied exactly, ensuring identity on features that should not change under the intervention. Mediators in M are edited to remain on-support under the conditional distribution for an older applicant, for example by increasing employment tenure or modestly improving credit history. Proxy features in P pass through unchanged but are guarded by the proxy adversary to suppress residual sensitive signal.

Server-side enforcement. The server fuses representations from the original and counterfactual inputs and applies the Selective Counterfactual Consistency loss to penalize prediction differences. Because only policy-permitted mediators are edited and proxy leakage is controlled, this penalty targets impermissible influence of age on the lending decision rather than suppressing legitimate pathways.

Interpretation. Framed as a counterfactual explanation [47], if the predicted decision changes substantially between the original and counterfactual representations, the SCC penalty increases, discouraging age-driven instability. If the decision remains stable, the model satisfies selective counterfactual consistency for this individual under the given policy specification. This example illustrates how SCC-VFL operationalizes individual-level stability without requiring access to raw sensitive attributes or a fully specified causal graph.