WHEN MISSINGNESS CARRIES MEANING: CAUSAL FEDERATED MULTI-VIEW CLUSTERING WITH INFORMATIVE VIEW MISSINGNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated multi-view clustering enables multiple institutions to cluster data collaboratively without sharing raw features, yet in practice many samples lack one or more views. Existing methods treat the missing pattern as a mere index: they use what is observed, impute what is not, and implicitly assume that views are missing at random. In practice, missingness often carries meaning, as which views a sample lacks is correlated with its latent semantics: clinicians typically order an imaging examination only when initial symptoms suggest severe disease, so a missing imaging view indicates a likely milder case. Ignoring this discards useful information and introduces selection bias, since each client observes only samples with an available view. We propose `CFIMC`, which places the missing mechanism inside a causal graph and builds three components on it: (i) a mechanism model that turns the missing pattern into additional evidence for cluster assignment and into inverse-probability weights that correct client-side selection bias; (ii) sample-level counterfactual view imputation, replacing cluster-level imputation that fills a missing view with the prototype of the currently assigned cluster and thus reinforces assignment errors; and (iii) a label-free reliability score and doubly robust prototypes that determine how much imputed views are trusted. Extensive experiments demonstrate the superiority of `CFIMC` against state-of-the-art baselines.

1 INTRODUCTION

Multi-view clustering (MVC) discovers latent groups from heterogeneous views of the same objects by exploiting their consistency and complementarity (Yang & Wang, 2018; Chao et al., 2021; Moujahid & Dornaika, 2025), and deep MVC learns view-specific encoders and a shared cluster-friendly space end-to-end (Xie et al., 2016; Cui et al., 2025). When views are held by institutions that cannot pool raw features, federated learning (McMahan et al., 2017; Kairouz et al., 2021), notably its vertical variant (Yang et al., 2019), yields federated MVC (FedMVC) (Huang et al., 2022; Chen et al., 2023); as views are also rarely complete, a problem studied by centralized incomplete MVC (IMVC) (Wen et al., 2023; Liu et al., 2018; Lin et al., 2021), this leads to *federated incomplete multi-view clustering* (FedIMVC): each client observes its view for only some samples, the server never sees raw features, and the partition must be inferred from incomplete, distributed and unlabeled evidence.

Related work. FedIMVC has advanced rapidly. FedDMVC (Chen et al., 2023) pioneered global self-supervision, where the server fuses uploaded embeddings into global prototypes and pseudo-labels; FCUIF (Ren et al., 2024) elegantly removed the alignment and completeness assumptions at once; later methods bridged the client and view gaps (Chen et al., 2024), injected globally fused graphs into client encoders (Chao et al., 2025), achieved remarkable communication efficiency (Liu et al., 2025; Feng et al., 2026), reconciled privacy and performance (Chen et al., 2025) and imputed missing views with Transformers (Wang et al., 2026), while FedMVC for complete data matured in parallel (Feng et al., 2024b; Jiang et al., 2024; Feng et al., 2025; Liu et al., 2026; Zhang et al., 2026; Wei et al., 2026). Appendix B reviews these works and the causal and missing-data literature.

Missingness carries meaning. Existing FedIMVC methods treat the missing mask as an *index set*: observed entries are used, missing ones are imputed or skipped, and views are implicitly assumed missing completely at random (MCAR). Yet which views a sample lacks often reflects its semantics:

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

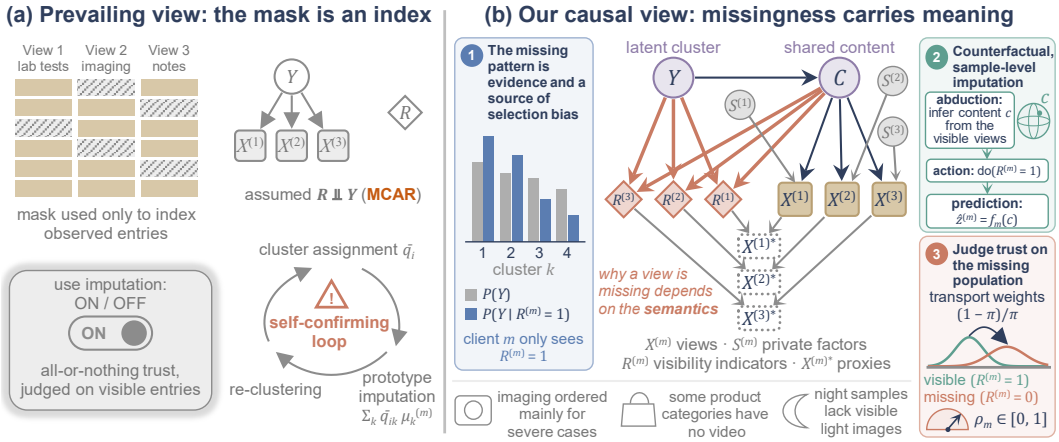

Figure 1: **Why missingness carries meaning.** (a) Existing FedIMVC: the mask is an index set, imputation is prototype-based, and trust is binary. (b) Our missingness graph: Y and C cause both the views $X^{(m)}$ and the indicators $R^{(m)}$, exposing problems ① to ③.

clinicians order imaging mainly when early symptoms suggest severe disease, so a missing imaging view hints at a milder case. Views are then missing not at random (MNAR) (Little & Rubin, 2019), and a causal graph of the mechanism (Figure 1) exposes three problems.

Problem 1: ignoring the pattern discards evidence and induces selection bias. The pattern $\boldsymbol{r}_i$ is a noisy observation of the cluster of sample i, yet it is ignored, and client m only sees samples with $R^{(m)} = 1$, whose cluster law differs from $P(Y)$, so clusters that rarely carry view m get almost no local gradient. **Problem 2: cluster-level imputation is circular.** Filling a missing view with (a learned map of) $\sum_k q_{ik}\boldsymbol{\mu}_k^{(m)}$ votes for the current assignment in the next round, so clustering errors confirm themselves (Lemma D.8), and Section 4.6 checks this against the hidden views. **Problem 3: trust in imputation is all-or-nothing and judged on the wrong population.** Imputations are adopted or discarded by a binary switch, and their quality is measured on visible entries but matters on missing ones, which differ under MNAR, so a naive evaluation is optimistic (Theorem 3.3). We address all three problems by making the missingness mechanism a first-class citizen of FedIMVC; our contributions are as follows.

- To our knowledge, this is the first FedIMVC work that models *why* views are missing. We prove that the pattern is both a source of information, whose expected log-likelihood ratio is a Kullback-Leibler divergence (Theorem 3.1), and a source of bias, as every client optimizes a selected rather than an interventional objective (Theorem D.2).
- We propose `CFIMC`**, a mechanism-aware, counterfactual and doubly robust FedIMVC method** with three modules derived from one missingness graph: **(I)** a mechanism model that yields Bayesian evidence and inverse-probability and transport weights; **(II)** sample-level counterfactual imputation by a server-side cross-view structural causal model; and **(III)** a label-free transported reliability with doubly robust prototypes that weighs each imputation.
- A comprehensive theory: counterfactual identifiability, the transport gap of associational imputers, consistency of the transported reliability, bias and asymptotic normality of the doubly robust prototypes, convergence of the server clustering, and communication and privacy guarantees.
- Experiments on one synthetic and eight real multi-view datasets under a class-dependent MNAR and a label-free latent MNAR compare `CFIMC` with eleven FedMVC and FedIMVC methods.

2 Preliminaries

Notation is summarized in Table 2 (Appendix A); we write $\mathrm{nrm}(\boldsymbol{a}) = \boldsymbol{a}/\|\boldsymbol{a}\|_2$, $\mathrm{clip}_{[a,b]}(x) = \min\{\max\{x, a\}, b\}$ and $P_k, \mathbb{E}_k, \mathrm{Var}_k, \mathrm{Cov}_k$ for probability and moments given $Y = k$. The theorems of Section 3 are abridged; Appendix D states them in full and Appendix E proves them.

Definition 2.1 (FedIMVC)**.** A dataset of N samples and M views, $\boldsymbol{x}_i^{(m)} \in \mathbb{R}^{D_m}$, comes with a mask $\boldsymbol{R} = [r_{im}] \in \{0,1\}^{N\times M}$, where $r_{im} = 1$ if and only if view m of sample i is observed; its rows, the *missing patterns* $\boldsymbol{r}_i$, lie in $\mathfrak{R} = \{0,1\}^M \setminus \{\mathbf{0}\}$, and $\mathcal{V}_i = \{m : r_{im} = 1\}$. Client $m \in [M]$ privately holds $\{\boldsymbol{x}_i^{(m)} : i \in \mathcal{O}_m\}$ with $\mathcal{O}_m = \{i : r_{im} = 1\}$ and $N_m = |\mathcal{O}_m|$. The server knows $\boldsymbol{R}$ through the sample identifiers, as in any vertical protocol after entity resolution, but never receives raw features. FedIMVC seeks a partition of $[N]$ into K clusters through T rounds of low-dimensional message exchange.

Missingness mechanism. Each sample is a draw of $(Y, C, S^{(1:M)}, X^{(1:M)}, \mathbf{R})$ with latent cluster $Y \in [K]$, shared content $C \in \mathbb{S}^{d_c-1}$ (semantics common to all views), private factors $S^{(m)}$, full views $X^{(m)}$ and pattern $\mathbf{R} \in \mathfrak{R}$; view m is observed as $X^{(m)*} = X^{(m)}$ if $R^{(m)} = 1$ and $\star$ otherwise (Mohan & Pearl, 2021). The per-cluster pattern law and visibility and the fine-grained propensity are

$$\varpi_k(\boldsymbol{r}) = P_k(\mathbf{R} = \boldsymbol{r}), \qquad \pi_m(k) = P_k\big(R^{(m)} = 1\big), \qquad \pi_m^\star(Y, C) = P\big(R^{(m)} = 1 \mid Y, C\big), \quad (1)$$

and $\bar{\pi}_m = P(R^{(m)} = 1)$. The missingness is *informative about the clusters* (IVM) if $\varpi_k \neq \varpi_l$ for some $k \neq l$, i.e., $I(Y; \mathbf{R}) > 0$; as Y is never observed, IVM is MNAR with respect to the clustering target (Definition C.3).

Assumption 2.2 (Causal model of view missingness)**.** *(i) Structure:* the data are generated by the SCM

$$Y := F_Y(U_Y), \quad C := F_C(Y, U_C), \quad X^{(m)} := g_m\big(C, S^{(m)}\big), \quad \mathbf{R} := F_R(Y, C, U_R), \quad (2)$$

with private factors $S^{(m)} := U_{S^{(m)}}$ and mutually independent exogenous variables $(U_Y, U_C, U_{S^{(1)}}, \ldots, U_{S^{(M)}}, U_R)$; its m-graph $\mathcal{G}_{\text{IVM}}$ is shown in Figure 1(b), and $\mathbf{R}$ has no parent in $\{X^{(m)}, S^{(m)}\}_{m\in[M]}$. *(ii) View-wise conditional independence:* $R^{(m)} \perp\!\!\!\perp \mathbf{R}^{(-m)} \mid (Y, C)$ for every $m \in [M]$. *(iii) Positivity:* there exists $\eta \in (0, 1)$ such that $\pi_m^\star(Y, C) \geq \eta$ almost surely for all m, and $\bar{\pi}_m < 1$ for every view with missing entries.

Assumption 2.3 (Regularity of the embeddings)**.** For fixed client encoders Enc_m, the embeddings $Z^{(m)} = \text{Enc}_m(X^{(m)})$ satisfy: *(i) additive noise:* $Z^{(m)} = f_m(C) + \boldsymbol{\xi}^{(m)}$, where f_m is measurable, $\boldsymbol{\xi}^{(m)} = \zeta_m(S^{(m)})$, $\mathbb{E}\boldsymbol{\xi}^{(m)} = \mathbf{0}$ and $\mathbb{E}\|Z^{(m)}\|_2^2 < \infty$; *(ii) content recoverability:* there exist measurable maps $h_m^\star : \mathbb{R}^d \to \mathbb{S}^{d_c-1}$ such that $h_m^\star(Z^{(m)}) = C$ almost surely.

Condition (P) and scope. Condition (P) (cluster-level propensity) states that $\mathbf{R} \perp\!\!\!\perp C \mid Y$, i.e., $\pi_m^\star(Y, C) = \pi_m(Y)$ for every m. It holds for MCAR and cluster-dependent MNAR but is *not* a standing assumption: the algorithm never requires it, and every result that uses it says so. Assumption 2.2 lets propensities depend arbitrarily on the semantics and makes the missingness *latently ignorable* (Frangakis & Rubin, 1999), $\mathbf{R} \perp\!\!\!\perp X^{(1:M)} \mid C$, so view mechanisms are invariant between visible and missing entries (Lemma C.4); it excludes self-masking, which is not identifiable without further assumptions (Mohan & Pearl, 2021) (Appendix C.2).

3 The Proposed CFIMC Method

CFIMC keeps the global self-supervision protocol of Chen et al. (2023) and Ren et al. (2024) but replaces every component that silently assumes MCAR (Figure 2; Algorithms 1 to 3). In each round, client m uploads embeddings $\boldsymbol{Z}^{(m)}$, soft assignments $\boldsymbol{Q}^{(m)}$ and local centers $\boldsymbol{U}^{(m)}$; the server runs Modules I to III and returns prototypes $\{\boldsymbol{\mu}_k^{(m)}\}_k$, pseudo-labels $\boldsymbol{P}^{(m)}$ and inverse-probability weights $\boldsymbol{w}^{(m)}$. The only extra message, $\boldsymbol{w}^{(m)}$, adds less than a fraction $1/(d+K)$ to the communication and leaks at most $K \log_2(N+1)$ bits beyond the client's own cluster labels (Appendix G).

3.1 Client Side: Selection-Debiased Self-Training

Client m owns an autoencoder with encoder $\text{Enc}_m : \mathbb{R}^{D_m} \to \mathbb{R}^d$ and decoder Dec_m, pretrained on its visible samples by the reconstruction loss $\mathcal{L}_{\text{rec}}^{(m)} = N_m^{-1} \sum_{i\in\mathcal{O}_m} \|\boldsymbol{x}_i^{(m)} - \text{Dec}_m(\text{Enc}_m(\boldsymbol{x}_i^{(m)}))\|_2^2$.

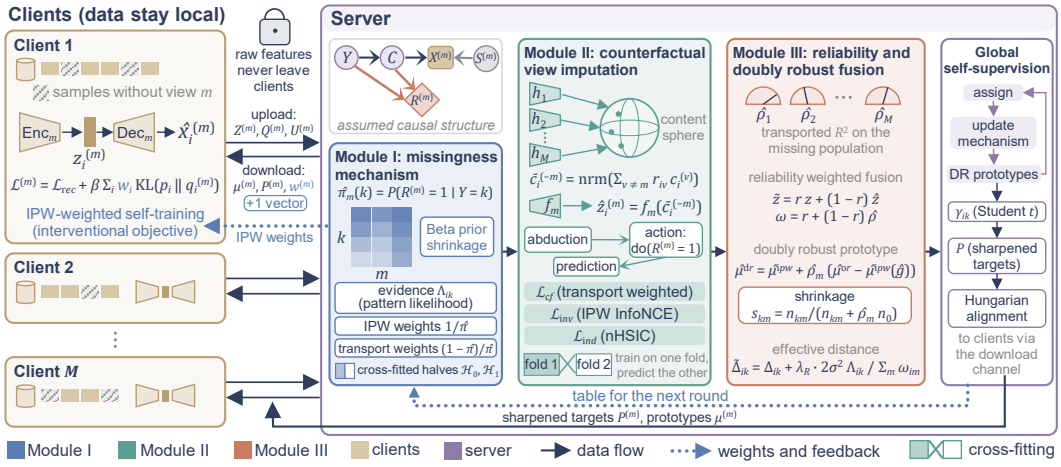

Figure 2: **Framework of** CFIMC**.** Clients run inverse-probability-weighted self-training; the server estimates the mechanism (Module I), imputes missing blocks counterfactually (Module II) and weighs the imputations by a transported reliability in a doubly robust clustering (Module III).

In round $t \geq 1$ it computes the Student's t assignments $q_{ik}^{(m)}$ (Definition C.9) of $\boldsymbol{z}_i^{(m)} = \mathrm{Enc}_m(\boldsymbol{x}_i^{(m)})$ to local centers $\boldsymbol{U}^{(m)} = \{\boldsymbol{u}_k^{(m)}\}_{k\in[K]}$, initialized by the downloaded prototypes, and minimizes

$$\mathcal{L}^{(m)} = \mathcal{L}_{\mathrm{rec}}^{(m)} + \frac{\beta}{N_m} \sum_{i\in\mathcal{O}_m} w_{im}^{\mathrm{ipw}}\, D_{\mathrm{KL}}\big(\boldsymbol{p}_i \,\big\|\, \boldsymbol{q}_i^{(m)}\big), \qquad w_{im}^{\mathrm{ipw}} = \frac{1/\hat{\pi}_{im}}{\frac{1}{N_m}\sum_{j\in\mathcal{O}_m} 1/\hat{\pi}_{jm}}, \tag{3}$$

where $\boldsymbol{p}_i$ is the server target, $\hat{\pi}_{im}$ the estimated propensity of Eq. (4) and $\beta > 0$. Without the weights, Eq. (3) is the local loss of FedDMVC and FCUIF; under IVM, however, client m sees the shifted cluster law $P(Y = k \mid R^{(m)} = 1) = \pi_m(k)P(Y = k)/\bar{\pi}_m$ (Proposition D.1). The weights undo this shift: under Assumption 2.2 and Condition (P) with $\hat{\pi}_{im} = \pi_m(y_i)$, the weighted objective is consistent for the interventional risk $\mathbb{E}_{\mathrm{do}(R^{(m)}=1)}[D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}^{(m)})]$, as if view m were observed for every sample, whereas the unweighted one converges to the selected risk $\mathbb{E}[D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}^{(m)}) \mid R^{(m)} = 1]$ (Theorem D.2). Under MCAR or complete data, the weights are identically one and Eq. (3) reduces to the backbone loss.

3.2 Module I: Modeling the Missingness Mechanism

Mechanism estimation. Given hard assignments $\{a_i\}$ (Section 3.5), let $a_{ik} = \mathbb{I}[a_i = k]$, $n_k = \sum_i a_{ik}$, $n_{km} = \sum_i a_{ik} r_{im}$ and $\hat{\bar{\pi}}_m = N_m/N$, and set

$$\hat{\pi}_m(k) = \frac{n_{km} + \alpha\,\hat{\bar{\pi}}_m}{n_k + \alpha}, \qquad \hat{\pi}_{im} = \mathrm{clip}_{[\underline{\pi},1]}\big(\hat{\pi}_m(a_i)\big), \qquad \check{\pi}_m(k) = \mathrm{clip}_{[\underline{\pi},1-\underline{\pi}]}\big(\hat{\pi}_m(k)\big), \tag{4}$$

with a pseudo-count $\alpha > 0$ and a floor $\underline{\pi} \in (0, \frac{1}{2})$, a Beta-posterior estimate that shrinks small clusters toward the marginal visibility (Lemma D.4). The inverse-probability weights $w_{im}^{\mathrm{ipw}} \propto 1/\hat{\pi}_{im}$ and transport weights $w_{im}^{\mathrm{tr}} \propto (1 - \hat{\pi}_{im})/\hat{\pi}_{im}$, normalized to mean one over $\mathcal{O}_m$, turn averages over visible entries into averages over the whole population and over the *missing* entries, respectively (Lemma C.7).

Missingness as evidence. The table induces the negative log-likelihood of the observed pattern, which existing methods discard,

$$\Lambda_{ik} = -\sum_{m=1}^{M} \Big[r_{im} \log \check{\pi}_m(k) + (1 - r_{im}) \log\big(1 - \check{\pi}_m(k)\big) \Big]. \tag{5}$$

Theorem 3.1 (Missingness is evidence; full version: Theorem D.3)**.** *Let Assumption 2.2(i) and Condition (P) hold, and let* $\check{P}(Y = k \mid \boldsymbol{x}_{\boldsymbol{r}}) \propto P(Y = k)\, p_k(\boldsymbol{x}_{\boldsymbol{r}})$ *be the* pattern-agnostic *posterior given the visible features* $\boldsymbol{x}_{\boldsymbol{r}}$ *of a pattern* $\boldsymbol{r}$*. (i)* $P(Y = k \mid \boldsymbol{x}_{\boldsymbol{r}}, \mathbf{R} = \boldsymbol{r}) \propto \check{P}(Y = k \mid \boldsymbol{x}_{\boldsymbol{r}})\, \varpi_k(\boldsymbol{r})$*, and under Assumption 2.2(ii) the pattern log-odds* $\log \frac{\varpi_k(\boldsymbol{r})}{\varpi_l(\boldsymbol{r})}$ *is the population version of* $\Lambda_{il} - \Lambda_{ik}$*; (ii) its expectation given* $Y = k$ *is* $D_{\mathrm{KL}}(\varpi_k \| \varpi_l) \geq 0$*, with equality if and only if* $\varpi_k = \varpi_l$*; (iii) the rule* $\arg\max_k \check{P}(k \mid \boldsymbol{x}_{\boldsymbol{r}})$ *is Bayes optimal for every feature law compatible with Assumption 2.2(i) if and only if the missingness is not informative.*

Evidence-aware assignment. With completed blocks $\tilde{\boldsymbol{z}}_i^{(m)}$, fusion weights ω_{im} (Eq. (11)) and prototypes $\boldsymbol{\mu}_k^{(m)}$, samples are assigned by an evidence-aware distance, in-sample or cross-fitted:

$$\Delta_{ik} = \frac{\sum_m \omega_{im} \left\| \tilde{\boldsymbol{z}}_i^{(m)} - \boldsymbol{\mu}_k^{(m)} \right\|_2^2}{\sum_m \omega_{im}}, \qquad \tilde{\Delta}_{ik} = \Delta_{ik} + \lambda_R \frac{2\sigma^2\, \Lambda_{ik}}{\sum_m \omega_{im}}, \tag{6}$$

$$\tilde{\Delta}_{ik}^{\mathrm{cf}} = \Delta_{ik} + \lambda_R \frac{2\sigma^2\, \Lambda_{ik}^{\mathrm{cf}}}{\sum_m \omega_{im}}, \tag{7}$$

where $\lambda_R \geq 0$ weighs the evidence and σ^2 is a per-dimension within-cluster variance fixed after pretraining; $\arg\min_k \tilde{\Delta}_{ik}$ is then a MAP rule (Proposition D.5), and the evidence is inert once the table is flat, as under MCAR (Corollary D.6). As in-sample evidence is self-confirming (Lemma D.7), $\Lambda_{ik}^{\mathrm{cf}}$ uses the table fitted on the other half of a random split $\mathcal{H}_0 \cup \mathcal{H}_1 = [N]$, redrawn each round.

3.3 Module II: Counterfactual View Imputation

Existing methods fill block m of sample i with $\hat{\boldsymbol{z}}_i^{\mathrm{proto},(m)} = \sum_k \bar{q}_{ik} \boldsymbol{\mu}_k^{(m)}$ for a consensus soft assignment $\bar{\boldsymbol{q}}_i$, which never argues against the current assignment (Lemma D.8); `CFIMC` instead answers, sample by sample, the query "what would $\boldsymbol{z}_i^{(m)}$ be had view m been observed" with an abduction network h_m and a mechanism network f_m per view, following Pearl's three steps:

$$\text{abduction:} \quad \boldsymbol{c}_i^{(v)} = \mathrm{nrm}\left(h_v(\boldsymbol{z}_i^{(v)})\right), \qquad \bar{\boldsymbol{c}}_i^{(-m)} = \mathrm{nrm}\Big(\sum_{v \neq m} r_{iv}\, \boldsymbol{c}_i^{(v)}\Big), \tag{8}$$

$$\text{action and prediction:} \quad \mathrm{do}\big(R^{(m)} = 1\big), \qquad \hat{\boldsymbol{z}}_i^{(m)} = f_m\big(\bar{\boldsymbol{c}}_i^{(-m)}\big). \tag{9}$$

Eq. (9) never reads the assignment; it relies on the invariance of mechanisms under selection:

Theorem 3.2 (Counterfactual identifiability; full version: Theorem D.9)**.** *Under Assumptions 2.2 and 2.3: (i)* $P(Z^{(m)} \mid C, \mathbf{R}) = P(Z^{(m)} \mid C)$ *and* $\mathbb{E}[Z^{(m)} \mid C, \mathbf{R}] = f_m(C)$*; (ii)* f_m *is identified from co-observed pairs of views; (iii) for a sample with* $R^{(m)} = 0$ *and visible embeddings* $(\boldsymbol{z}^{(v)})_{v \in \mathcal{V}}$*, the counterfactual embedding* $Z^{(m)*}_{\mathrm{do}(R^{(m)}=1)}$ *has law* $f_m(\boldsymbol{c}^\star) + \boldsymbol{\xi}^{(m)}$*, with* $\boldsymbol{c}^\star = \mathrm{nrm}\left(\sum_{v \in \mathcal{V}} h_v^\star(\boldsymbol{z}^{(v)})\right)$ *and* $\boldsymbol{\xi}^{(m)}$ *independent of the evidence, so Eq. (9) with oracle networks is the MSE-optimal counterfactual imputation.*

Associational imputers, which regress one view on others, lack this invariance; their transport gap is bounded by the abduction uncertainty, which the pooling in Eq. (8) reduces (Proposition D.10).

Training and cross-fitting. The server fits $\{h_m, f_m\}_{m \in [M]}$ on all samples with at least two visible views, under random view dropout, by minimizing a structural causal objective (Eq. (27)): a counterfactual squared-error loss, transport weighted as the imputer serves the missing population, plus an inverse-probability-weighted InfoNCE term (van den Oord et al., 2018) and an nHSIC noise-cause independence term; it is consistent for the missing-population risk (Proposition D.11). With L-fold cross-fitting, every entry is predicted by networks fitted on other folds (Eq. (32)), so visible-entry errors are unbiased for out-of-sample ones (Lemma D.12).

3.4 Module III: Transported Reliability and Doubly Robust Fusion

Transported reliability. Let $\hat{g}_m$ be the cross-fitted imputer, a function of the other views defined on $\{\mathbf{R}^{(-m)} \neq \mathbf{0}\}$ (Remark E.4 treats the other samples), and $\hat{g}_{im}$ its value at sample i ($\hat{g}_{im} = \boldsymbol{z}_i^{(m)}$

if i has no other visible view). The label-free reliability of view m is the transport-weighted R^2 of the cross-fitted imputations on visible entries,

$$\hat{\rho}_m = \mathrm{clip}_{[0,1]}\left(1 - \frac{\sum_{i\in\mathcal{O}_m^+} \hat{w}_{im}^{\mathrm{tr}} \left\|\boldsymbol{z}_i^{(m)} - \hat{\boldsymbol{z}}_i^{(m)}\right\|_2^2}{\sum_{i\in\mathcal{O}_m^+} \hat{w}_{im}^{\mathrm{tr}} \left\|\boldsymbol{z}_i^{(m)} - \bar{\boldsymbol{z}}_m^{\mathrm{tr}}\right\|_2^2}\right), \qquad \bar{\boldsymbol{z}}_m^{\mathrm{tr}} = \sum_{i\in\mathcal{O}_m^+} \hat{w}_{im}^{\mathrm{tr}} \boldsymbol{z}_i^{(m)}, \tag{10}$$

where $\mathcal{O}_m^+ = \{i \in \mathcal{O}_m : \sum_{v\neq m} r_{iv} \geq 1\}$, $\hat{w}_{im}^{\mathrm{tr}} = w_{im}^{\mathrm{tr}} / \sum_{j\in\mathcal{O}_m^+} w_{jm}^{\mathrm{tr}}$, and $\hat{\rho}_m = 1$ for a view without missing entries.

Theorem 3.3 (Transported reliability; full version: Theorem D.13)**.** *Under Assumption 2.2 and Condition (P), with a consistent plug-in propensity and cross-fitted imputers converging to a fixed $\hat{g}_m$: (i) $\hat{\rho}_m \to \mathrm{clip}_{[0,1]}(\rho_m)$ in probability, where ρ_m is the R^2 of $\hat{g}_m$ on the* missing *population; (ii) the unweighted visible-entry and missing-population errors satisfy* $\mathrm{MSE}_{\mathrm{vis}} - \mathrm{MSE}_{\mathrm{mis}} = \mathrm{Cov}\big(e_m(Y), \pi_m(Y)\big)/\big(\bar{\pi}_m(1-\bar{\pi}_m)\big)$ *with $e_m(k) = \mathbb{E}\big[\|Z^{(m)} - \hat{g}_m\|^2 \mid Y = k\big]$, so unweighted evaluation is optimistic if rarely visible clusters are the hardest to impute.*

Prior work restricts $\hat{\rho}_m$ to $\{0, 1\}$; `CF IMC` interpolates, using completed blocks and fusion weights

$$\tilde{\boldsymbol{z}}_i^{(m)} = r_{im}\, \boldsymbol{z}_i^{(m)} + (1 - r_{im})\, \hat{\boldsymbol{z}}_i^{(m)}, \qquad \omega_{im} = r_{im} + (1 - r_{im})\, \hat{\rho}_m. \tag{11}$$

Doubly robust prototypes. With the cross-fitted imputations $\hat{g}_{im}$, the prototypes are built as

$$\hat{\boldsymbol{\mu}}_{km}^{\mathrm{ipw}} = \frac{\sum_i a_{ik}\,(r_{im}/\hat{\pi}_{im})\,\boldsymbol{z}_i^{(m)}}{\sum_i a_{ik}\,(r_{im}/\hat{\pi}_{im})}, \qquad \hat{\boldsymbol{\mu}}_{km}^{\mathrm{or}} = \frac{\sum_i a_{ik}\,\hat{g}_{im}}{\sum_i a_{ik}}, \tag{12}$$

$$\hat{\boldsymbol{\mu}}_{km}^{\mathrm{dr}} = \hat{\boldsymbol{\mu}}_{km}^{\mathrm{ipw}} + \hat{\rho}_m \left(\hat{\boldsymbol{\mu}}_{km}^{\mathrm{or}} - \frac{\sum_i a_{ik}\,(r_{im}/\hat{\pi}_{im})\,\hat{g}_{im}}{\sum_i a_{ik}\,(r_{im}/\hat{\pi}_{im})}\right), \tag{13}$$

$$\boldsymbol{\mu}_k^{(m)} = (1 - s_{km})\,\hat{\boldsymbol{\mu}}_{km}^{\mathrm{or}} + s_{km}\,\hat{\boldsymbol{\mu}}_{km}^{\mathrm{dr}}, \qquad s_{km} = \frac{n_{km}}{n_{km} + \hat{\rho}_m n_0}. \tag{14}$$

For $\hat{\rho}_m = 1$, Eq. (13) is the augmented inverse-probability-weighted estimator (Bang & Robins, 2005), for $\hat{\rho}_m = 0$ the Hájek estimator; Eq. (14) protects sparsely visible clusters, and the correction targets *within-cluster* selection, missed by cluster-level weights (Remark D.15).

Theorem 3.4 (Doubly robust prototypes; full version: Theorem D.14)**.** *Fix k and m, assume oracle memberships ($a_i = y_i$) and cross-fitted imputers converging in $L^2(P_k)$ to $\hat{g}_m$, and let $\pi^\star = \pi_m^\star(Y, C)$, $\bar{\pi}_k^\star = \mathbb{E}_k[\pi^\star]$, $g^\star = f_m(C)$ and $\boldsymbol{\mu}_{km} = \mathbb{E}_k[Z^{(m)}]$. Under Assumptions 2.2 and 2.3(i), for every fixed $\rho \in [0, 1]$ in place of $\hat{\rho}_m$,* plim $\hat{\boldsymbol{\mu}}_{km}^{\mathrm{dr}}(\rho) = \boldsymbol{\mu}_{km} + \big[(1-\rho)\mathrm{Cov}_k(\pi^\star, g^\star) + \rho\,\mathrm{Cov}_k(\pi^\star, g^\star - \hat{g}_m)\big]/\bar{\pi}_k^\star$. *This limit equals $\boldsymbol{\mu}_{km}$ under Condition (P) for every ρ and every imputer, or if $\hat{g}_m = g^\star$ and $\rho = 1$ even when (P) fails, and its bias is at most $\mathrm{CV}_{km}\big[(1-\rho)\,\mathrm{sd}_k(g^\star) + \rho\,\|\hat{g}_m - g^\star\|_{L^2(P_k)}\big]$ with $\mathrm{CV}_{km} = \mathrm{sd}_k(\pi^\star)/\bar{\pi}_k^\star$ and sd_k the standard deviation given $Y = k$.*

Accurate imputers thus remove selection bias, and $\hat{\rho}_m$ down-weights poor ones; the estimator is asymptotically normal and minimizes a doubly robust risk (Theorem D.16, Proposition D.17).

3.5 Evidence-Aware Clustering and Global Self-Supervision

The server seeds the prototypes by masked k-means++ (Arthur & Vassilvitskii, 2007), since k-means on zero-filled blocks tends to recover missing patterns (Appendix D.5), then alternates the assignment $a_i = \arg\min_k \tilde{\Delta}_{ik}^{\mathrm{cf}}$, the table update Eq. (4) and the prototype update Eq. (12) to Eq. (14) until convergence (Algorithm 3); with in-sample evidence and plain means, this is a classification EM with a Beta prior that terminates in finitely many sweeps (Proposition D.18). The soft assignments and targets

$$\gamma_{ik} \propto \big(1 + \tilde{\Delta}_{ik}\big)^{-1}, \qquad p_{ik} \propto \gamma_{ik}^2 \Big/ \textstyle\sum_j \gamma_{jk}, \qquad \text{both normalized over } k, \tag{15}$$

thus inherit the evidence and reliability-weighted imputations; indices are aligned across rounds by the Hungarian algorithm (Kuhn, 1955), and $\hat{y}_i = \arg\max_k \gamma_{ik}$. On complete data `CF IMC` reduces to its backbone; under MCAR, only the counterfactual imputation stays active (Corollary D.19).

4 Experimental Evaluation

4.1 Experiment Settings

Datasets We use eight public multi-view datasets, MSRC-v1, Caltech-5V, Scene-15, LandUse-21, 3V-Fashion, 100leaves, DryBean and Hdigit (Chen et al., 2022), and one synthetic dataset (Appendix H.1). Missing views are simulated on complete data, so the hidden views are available for evaluation but never given to any method. Under MNAR, the visibility of a view depends on the cluster only, and Condition (P) holds. Under MNAR_latent, it depends on a continuous latent factor of the content generated without the labels, so it also varies within clusters and (P) fails. The missing rate $\eta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, and every sample keeps at least one view (Appendix H.2). We do not run MCAR, under which only the counterfactual imputation of `CFIMC` remains active (Corollary D.19).

Comparison Baselines Five baselines assume complete views: CeFMC (Liu et al., 2026), FedMVL (Huang et al., 2022), FedCMv and FedFCMv (Bárcena et al., 2024), and HFMVC (Jiang et al., 2024)[1]. Six are FedIMVC methods: EFIMVC (Liu et al., 2025), ESFMC (Chen et al., 2025), FCUIF (Ren et al., 2024), FedDMVC (Chen et al., 2023), FIMCI (Wang et al., 2026) and FIMVC-TLRC (Feng et al., 2026). Details are in Appendix H.

Evaluation Metrics We report clustering accuracy (ACC), normalized mutual information (NMI) and adjusted Rand index (ARI) in percent.

Implementation Details `CFIMC` uses the defaults of Table 4 on every dataset, and every method uses K equal to the number of classes and the same five seeds, each fixing one mask per setting that all methods share. Tables pool seeds and missing rates, so their standard deviations include the spread across rates, and figures show each rate separately. All experiments were conducted on a workstation equipped with an Intel Core Ultra 7 265K processor, NVIDIA RTX 5070 Ti GPU (16GB VRAM), and 48GB system memory.

4.2 Results on Synthetic Data

FCUIF and EFIMVC are the strongest baselines, with a mean ACC of 88.19 and 88.01 under MNAR and of 87.29 and 88.09 under MNAR_latent (Table 6 in Appendix H.3). From MNAR to MNAR_latent, FIMCI, FedDMVC, HFMVC, FedCMv and FedFCMv lose 2.6 to 3.8 points of ACC, whereas EFIMVC and FIMVC-TLRC change by less than one point.

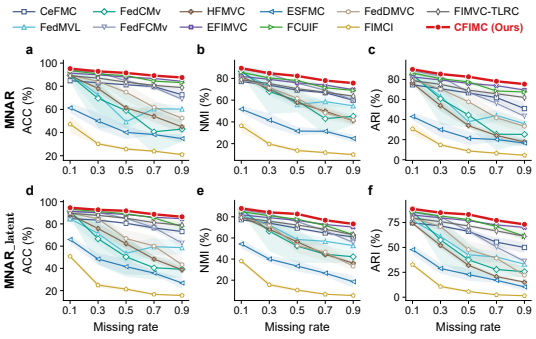

Figure 3: Synthetic data: ACC, NMI and ARI (%) against η under MNAR (top) and MNAR_latent (bottom). Lines are means over five seeds, and bands are $\pm$ one standard deviation.

4.3 Results on Real-World Data

FCUIF is the strongest baseline in 41 of the 48 cells of Table 1, and EFIMVC in the other seven (all of 100leaves and the ACC of 3V-Fashion under MNAR_latent). Over all cells, the six FedIMVC baselines average 45.7 and the five complete-view baselines 39.4, although FIMCI has the lowest average of all (31.0). From MNAR to MNAR_latent, the mean ACC of each baseline over the eight datasets changes by -2.5 to $+0.5$ points.

4.4 Ablation Study

Each variant removes one component and shares the mask, the seed and the pretrained autoencoders with the full model (MNAR, $\eta = 0.5$). *w/o mechanism* replaces $\hat{\pi}_m(k)$ by the marginal visibility

[1] We have pre-filled missing data using publicly available federated data imputation techniques.

Table 1: Clustering performance (%) on real-world datasets under two missing mechanisms, averaged over 5 seeds × 5 missing rates. The best results are in **bold** and the second best are underlined.

Dataset	Mechanism	Metric	CeFMC	FedMVL	FedCMv	FedFCMv	HFMVC	EFIMVC	ESFMC	FCUIF	FedDMVC	FIMCI	FIMVC-TLRC	CFIMC (Ours)
MSRC-v1	MNAR	ACC	63.85±8.93	58.21±9.62	54.95±10.85	58.32±10.30	52.29±9.71	66.84±6.24	58.69±12.65	70.04±7.14	54.61±7.79	39.87±10.54	64.46±7.08	**78.09**±4.07
		NMI	53.13±8.94	48.28±9.19	45.64±10.08	47.31±9.84	44.07±8.89	58.92±6.76	48.12±13.95	63.41±7.11	44.31±8.18	26.50±13.50	51.34±7.64	**68.16**±5.29
		ARI	42.11±11.94	32.19±13.21	29.76±13.17	32.64±13.24	28.46±11.53	48.87±8.47	37.92±15.15	53.23±9.28	30.50±9.10	16.61±12.46	42.33±9.06	**61.28**±6.20
	MNAR_latent	ACC	65.94±11.07	59.03±9.41	55.09±9.05	59.56±9.29	50.21±8.99	68.91±4.02	57.31±11.32	71.35±4.88	53.22±7.86	37.16±11.15	64.02±6.45	**78.05**±5.31
		NMI	55.38±11.62	48.36±9.00	46.01±9.39	49.01±8.86	42.05±8.58	59.23±4.83	47.00±12.95	64.46±5.35	43.69±8.51	24.85±13.77	51.62±6.38	**68.88**±6.09
		ARI	45.22±13.74	33.43±13.12	31.01±11.93	35.33±12.45	26.76±11.13	50.24±6.08	36.37±13.92	54.06±6.65	29.59±9.37	14.61±12.57	42.19±7.84	**62.25**±7.45
Caltech-5V	MNAR	ACC	73.04±5.53	61.63±11.98	58.11±14.70	61.08±10.38	61.13±11.93	79.31±1.86	58.71±12.75	81.56±3.85	65.63±9.03	31.04±10.02	68.70±4.34	**82.96**±1.83
		NMI	63.66±5.40	52.23±11.56	50.06±12.81	50.73±10.55	49.98±11.83	67.37±3.65	49.39±11.74	70.88±3.99	52.61±9.70	17.30±13.41	54.87±4.59	**71.64**±2.01
		ARI	56.51±8.30	34.67±17.94	32.93±18.65	34.14±16.33	32.63±16.40	64.60±3.61	41.57±12.85	66.84±4.98	45.35±11.54	12.80±11.74	49.81±5.26	**68.58**±2.26
	MNAR_latent	ACC	75.00±5.30	61.85±12.03	57.41±12.53	57.91±10.93	57.04±12.15	79.62±2.15	61.86±13.55	**83.01**±3.00	64.66±10.10	28.22±11.67	67.73±3.43	82.64±2.35
		NMI	63.99±6.24	51.98±11.14	49.00±11.92	48.23±11.29	46.32±12.53	67.68±3.39	53.16±12.20	**72.30**±3.10	51.43±11.36	14.32±14.31	54.20±4.45	71.34±3.45
		ARI	56.72±10.19	33.20±18.17	30.58±17.38	30.26±17.31	28.03±14.16	65.00±3.53	45.64±13.41	**68.63**±3.87	43.76±13.65	10.53±12.55	48.97±4.31	67.77±3.48
Scene-15	MNAR	ACC	23.49±6.35	31.88±3.69	31.50±4.03	30.46±3.25	30.04±3.22	34.34±2.96	25.18±3.96	39.90±1.92	25.02±1.25	23.63±5.40	31.04±1.66	**41.56**±0.77
		NMI	20.21±7.04	32.91±3.89	32.91±4.21	30.68±3.55	29.99±5.36	32.79±3.04	23.65±4.92	39.10±2.19	24.37±2.60	21.76±5.40	28.14±2.48	**40.83**±1.53
		ARI	7.01±5.53	15.42±3.62	15.11±3.97	14.37±3.59	12.33±3.83	18.67±2.28	11.10±3.11	23.59±1.86	12.54±1.89	10.51±4.02	15.29±1.51	**24.38**±0.79
	MNAR_latent	ACC	18.42±3.06	30.80±4.92	30.23±4.61	29.93±4.79	28.40±5.00	33.32±3.76	23.88±4.21	39.13±2.11	24.94±1.95	21.09±6.98	30.85±2.57	**41.34**±1.03
		NMI	13.92±3.83	32.07±4.87	31.88±4.84	29.52±4.47	28.32±6.60	31.44±3.95	22.65±5.00	38.83±2.98	24.39±2.18	20.07±6.51	27.51±2.98	**40.51**±1.80
		ARI	2.90±1.73	14.78±4.32	14.49±4.47	13.66±4.27	11.04±4.84	17.69±3.02	10.61±3.21	23.62±2.33	13.22±2.40	8.91±4.85	14.89±2.03	**24.17**±1.19
LandUse-21	MNAR	ACC	14.51±1.42	19.37±2.44	19.22±2.60	14.29±1.86	17.23±1.77	24.72±2.27	16.32±2.01	25.23±1.33	22.36±2.15	15.51±2.57	23.82±1.82	**26.82**±0.92
		NMI	12.75±1.55	25.27±3.44	25.18±3.43	17.38±2.91	22.66±2.90	26.19±3.23	15.59±2.41	27.83±1.68	23.33±2.68	17.16±3.65	23.70±2.30	**31.62**±1.73
		ARI	2.72±0.73	6.17±2.37	6.04±2.34	3.07±1.39	5.09±1.73	10.61±2.08	3.72±1.28	10.82±1.17	8.92±1.77	4.28±1.80	9.81±1.35	**12.84**±0.98
	MNAR_latent	ACC	14.10±1.57	19.38±2.73	19.05±2.44	14.34±1.77	16.83±2.10	24.19±2.15	15.42±2.38	24.54±1.72	22.06±2.38	14.26±3.09	23.23±1.58	**26.54**±0.87
		NMI	12.07±2.02	25.04±3.70	25.06±3.54	17.14±3.10	22.20±2.96	25.85±3.18	14.38±3.52	27.04±2.26	23.10±3.14	15.61±4.65	23.29±2.59	**30.80**±1.72
		ARI	2.45±1.04	6.14±2.43	5.96±2.52	3.12±1.28	4.83±1.92	10.29±2.16	3.48±1.50	10.33±1.47	8.73±2.07	3.60±2.18	9.46±1.38	**12.47**±0.75
3V-Fashion	MNAR	ACC	60.74±3.65	56.75±11.88	51.48±8.72	51.10±8.07	49.92±22.92	70.89±6.91	59.31±10.07	73.72±4.78	66.04±5.85	48.58±12.73	69.13±3.30	**74.08**±2.73
		NMI	57.68±3.89	57.39±10.72	55.00±9.14	54.28±8.50	44.26±24.64	66.28±6.45	64.83±5.90	75.49±4.31	66.71±5.87	52.84±12.04	65.86±4.26	**77.32**±3.95
		ARI	43.72±6.41	44.05±13.84	40.38±11.06	39.97±11.07	27.46±22.13	58.41±8.04	48.99±8.55	63.61±5.51	55.05±6.56	36.13±13.87	55.79±4.61	**65.63**±3.83
	MNAR_latent	ACC	59.99±2.52	55.59±11.24	52.09±9.72	51.30±7.67	44.09±25.72	72.23±5.60	60.65±7.46	70.91±6.11	63.64±4.74	46.33±13.75	67.18±5.34	**74.77**±5.88
		NMI	56.90±3.60	56.59±9.77	55.45±9.34	54.93±8.02	39.21±27.00	66.54±6.43	66.69±4.28	74.64±4.51	65.52±5.36	51.25±12.78	64.58±5.09	**77.34**±4.78
		ARI	41.49±6.60	42.75±13.10	40.68±11.48	40.23±10.29	23.09±24.04	59.26±7.68	51.05±6.40	61.17±6.20	53.00±5.91	34.33±14.76	53.65±6.11	**67.09**±6.30
100leaves	MNAR	ACC	25.84±5.18	56.70±15.20	52.87±12.41	14.00±1.76	35.40±12.53	67.03±13.02	36.69±15.07	65.04±15.89	50.96±12.91	43.55±13.76	54.31±11.95	**72.18**±11.72
		NMI	50.73±5.09	76.43±9.10	75.38±8.97	51.48±7.22	62.32±10.72	81.64±7.58	61.87±10.74	80.11±9.80	74.53±9.11	71.35±8.98	74.24±6.89	**85.74**±7.50
		ARI	7.65±3.91	38.58±19.26	35.70±17.66	9.34±3.67	19.54±14.82	54.11±16.72	19.89±15.86	50.92±20.77	38.78±16.47	31.39±16.86	38.95±13.51	**62.20**±15.90
	MNAR_latent	ACC	25.27±4.91	55.93±16.02	52.43±12.50	13.72±2.08	33.63±13.21	66.64±13.88	34.53±14.60	64.87±16.60	50.47±14.23	41.78±14.36	53.63±12.53	**72.56**±12.25
		NMI	50.07±5.08	76.15±9.80	75.17±9.24	51.06±7.75	60.74±11.74	81.43±8.10	60.23±10.44	80.13±10.07	74.28±9.68	70.22±9.55	73.83±7.29	**86.53**±7.49
		ARI	7.27±3.73	38.34±20.79	34.73±17.27	9.13±3.72	18.17±15.14	53.54±17.93	17.67±14.49	50.71±21.46	38.14±17.68	29.51±17.36	37.91±14.19	**62.96**±16.75
DryBean	MNAR	ACC	60.47±7.16	66.70±7.00	65.43±5.44	63.77±8.02	34.48±3.95	59.37±2.23	54.32±12.62	67.46±1.06	57.87±7.92	46.15±9.43	50.59±2.89	**85.78**±0.16
		NMI	43.08±6.42	58.06±5.33	57.42±5.14	58.01±4.94	27.70±5.08	51.45±3.41	45.91±11.12	60.88±1.02	55.67±6.29	40.77±9.05	43.50±2.56	**72.71**±0.24
		ARI	35.05±8.66	46.50±8.14	45.80±7.58	46.44±7.30	7.07±2.76	40.12±2.88	35.24±12.51	51.51±1.41	42.72±8.68	28.99±11.01	31.98±3.26	**69.08**±0.30
	MNAR_latent	ACC	59.25±4.88	64.31±8.51	64.65±9.29	60.24±8.37	33.64±3.14	60.52±1.09	51.94±10.56	67.65±0.94	60.58±5.18	42.89±11.22	53.06±3.84	**85.72**±0.20
		NMI	41.18±4.66	56.10±5.97	56.16±6.01	55.39±5.79	25.99±5.75	52.58±1.87	43.98±10.04	60.56±0.91	55.15±6.24	38.87±9.83	43.61±2.52	**72.69**±0.28
		ARI	31.73±5.91	44.24±8.92	44.15±8.90	42.99±8.06	5.98±2.58	41.47±1.23	32.82±11.81	51.17±1.13	43.32±7.04	27.36±12.07	34.68±4.63	**68.95**±0.40
Hdigit	MNAR	ACC	56.87±7.52	58.70±11.36	54.72±10.62	59.40±16.52	41.91±13.04	68.07±8.41	65.51±9.48	82.94±10.72	72.08±13.56	51.38±11.25	55.59±5.39	**92.72**±2.94
		NMI	44.43±7.12	54.20±11.56	53.36±11.07	53.95±13.85	40.30±18.92	54.93±7.88	64.58±9.30	73.18±10.32	60.99±14.39	44.81±10.27	46.38±4.80	**83.96**±5.35
		ARI	36.00±8.42	39.55±15.08	36.87±14.81	40.44±19.45	19.42±11.95	49.32±9.50	54.21±10.32	70.59±13.73	55.43±18.31	33.72±12.39	36.49±5.44	**84.76**±5.84
	MNAR_latent	ACC	56.28±9.44	61.23±14.76	56.32±11.75	60.53±16.66	38.31±17.83	68.58±8.61	67.89±11.35	88.04±4.13	72.57±13.08	50.05±14.35	56.15±4.45	**92.75**±2.79
		NMI	44.06±8.06	54.22±13.44	51.79±11.93	53.56±13.93	36.10±22.68	54.80±8.31	65.47±9.37	76.13±6.74	60.71±14.28	43.70±12.24	45.82±5.01	**83.90**±5.26
		ARI	35.74±9.75	41.00±18.29	36.50±14.94	40.30±19.47	18.28±16.38	48.91±10.26	55.67±12.37	75.65±7.80	55.46±18.01	32.59±15.00	36.36±5.08	**84.77**±5.59

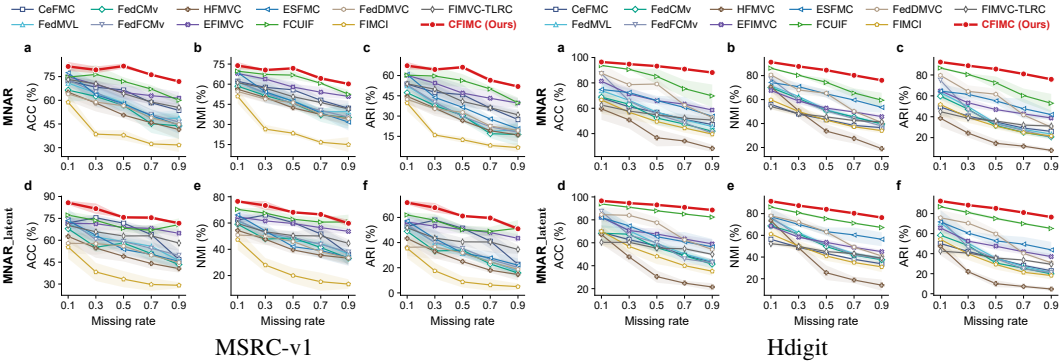

Figure 4: ACC, NMI and ARI (%) against η on MSRC-v1 (left) and Hdigit (right), with MNAR in the top and MNAR_latent in the bottom row of each panel (means over five seeds). Appendix H.3 shows the other six datasets.

$\hat{\tilde{\pi}}_m$, which also makes the evidence and both weights inert (Problem 1). *w/o evidence* drops only the evidence term of Eq. (6) (Problem 1). *w/o CF imputation* uses the cluster-level imputation of FCUIF instead of Module II (Problem 2). *w/o reliability and DR* sets $\hat{\rho}_m \equiv 1$ and uses plain means of the completed blocks as prototypes (Problem 3). Appendix H.4 shows ACC over the rounds.

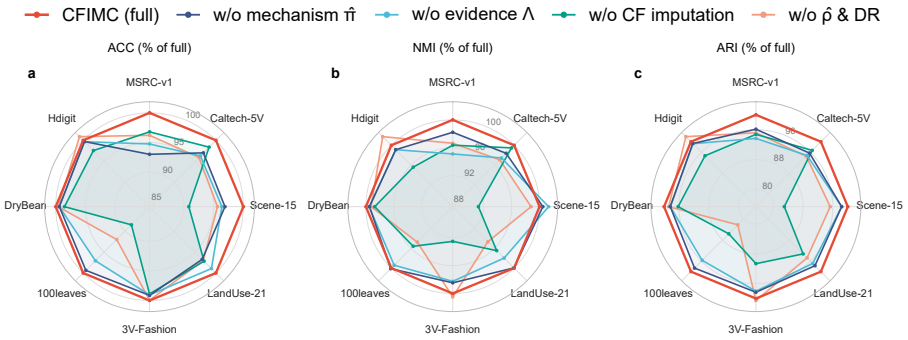

Figure 5: Ablation (MNAR, $\eta = 0.5$). Each axis is a dataset, and each variant is shown as a percentage of the full model (the circle at 100), averaged over the same five seeds for every variant.

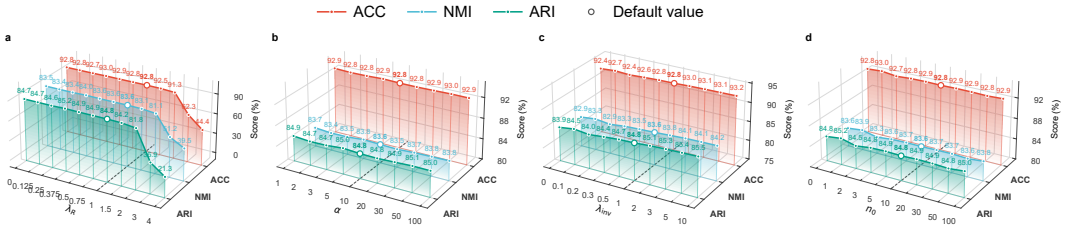

Figure 6: Sensitivity on Hdigit (MNAR, $\eta = 0.5$). Each panel varies one hyperparameter while the others are at their defaults (white markers).

4.5 Parameter Sensitivity

On MSRC-v1 and Hdigit (MNAR, $\eta = 0.5$) we vary one hyperparameter at a time: the evidence weight λ_R and table pseudo-count α of Module I, the InfoNCE weight λ_{inv} of Module II and the shrinkage strength n_0 of Module III (grids in Table 7). The grids include the switched-off values $\lambda_R = 0$ (equal to *w/o evidence*), $\lambda_{\text{inv}} = 0$ and $n_0 = 0$, and large values at which the theory predicts a limit: the evidence becomes inert as α grows (Lemma D.4, Corollary D.6), and for $\hat{\rho}_m > 0$ the prototypes tend to $\hat{\boldsymbol{\mu}}_{km}^{\text{or}}$ as n_0 grows (Eq. (14)).

4.6 Oracle Audit of Imputed Views

Since missingness is simulated, we audit the imputations of a trained `CFIMC` against the hidden views (MSRC-v1 and Hdigit, both mechanisms, $\eta = 0.5$). A block $\boldsymbol{b}$ in view m enters Δ_{ik} through $\|\boldsymbol{b} - \boldsymbol{\mu}_k^{(m)}\|_2^2$ (Eq. (6)), so it votes for its nearest prototype. For each missing entry, we compare the prototype of the current cluster, the cluster-level imputation of FCUIF, the counterfactual imputation $\hat{\boldsymbol{z}}_i^{(m)}$, the visible mean, and an oracle block that encodes the hidden view with the final client encoder. *Due to space limitations, the results are provided in Appendix H.6.*

5 Conclusions and Future Work

We showed that in FedIMVC the missing pattern is both evidence about the clusters and a source of selection bias, that cluster-level imputation confirms the current assignment, and that judging imputations on visible entries is optimistic when rarely visible clusters are the hardest to impute. `CFIMC` addresses these problems with a mechanism model, sample-level counterfactual imputation, and a label-free transported reliability that scales doubly robust prototypes.

Two limitations point to future work. (1) Self-masking is excluded, as it is not identifiable without further assumptions. (2) The propensity is cluster-level, so within-cluster selection is corrected only by the doubly robust term, which needs an accurate imputer when Condition (P) fails (Theorem 3.4), and finer propensity models fit the same estimator (Remark D.15).

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance and language polishing, including editing and refining parts of the manuscript text. We have not used generative AI tools for research ideation, experimental design, data collection, or analysis, and the remaining tasks requiring disclosure do not apply to this work. We have reviewed all AI-assisted work and verified that it accurately reflects our research and claims. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops clustering methods for data distributed across institutions, a setting motivated by privacy regulations. Raw features never leave the clients; the only additional message has a provable information bound (Proposition G.2), and the method is compatible with differential privacy (Proposition G.3). Because `CFIMC` exploits the fact that missingness correlates with latent semantics, practitioners should be aware that missing patterns may themselves encode sensitive attributes; the server already holds the mask in all vertical protocols, and our analysis makes this channel explicit rather than creating it. All experiments use public benchmarks or synthetic data.

REPRODUCIBILITY STATEMENT

All assumptions are stated in Section 2 and discussed in Appendix C.2, complete proofs of all theoretical results are given in Appendix E, and the federated training protocol and the full server procedures are listed in Algorithms 1 to 3 of Appendix F. A consolidated notation table is provided in Appendix A. The experimental protocol is described in Section 4.1 and Appendix H. Source code, configuration files, and the scripts that generate both missingness mechanisms will be released as supplementary material.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, volume 7, pp. 1027–1035, 2007.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Shifeng Bao, Zhe Xue, Qi Chen, Shilong Ou, Amin Beheshti, Quan Z Sheng, Anton van den Hengel, and Yuankai Qi. CausalMVC: Causal content-style representation learning for deep multi-view clustering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1598–1606, 2025.

José Luis Corcuera Bárcena, Francesco Marcelloni, Alessandro Renda, Alessio Bechini, and Pietro Ducange. Federated c-means and fuzzy c-means clustering algorithms for horizontally and vertically partitioned data. *IEEE Transactions on Artificial Intelligence*, 5(12):6426–6441, 2024.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Gilles Celeux and Gérard Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3):315–332, 1992.

Guoqing Chao, Shiliang Sun, and Jinbo Bi. A survey on multiview clustering. *IEEE Transactions on Artificial Intelligence*, 2(2):146–168, 2021.

Guoqing Chao, Yi Jiang, and Dianhui Chu. Incomplete contrastive multi-view clustering with high-confidence guiding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11221–11229, 2024.

Guoqing Chao, Zhenghao Zhang, Lei Meng, Jie Wen, and Dianhui Chu. Federated incomplete multi-view clustering with globally fused graph guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 7417–7429, 2025.

Man-Sheng Chen, Jia-Qi Lin, Xiang-Long Li, Bao-Yu Liu, Chang-Dong Wang, Dong Huang, and Jian-Huang Lai. Representation learning in multi-view clustering: A literature review. *Data Science and Engineering*, 7(3):225–241, 2022.

Xinyue Chen, Jie Xu, Yazhou Ren, Xiaorong Pu, Ce Zhu, Xiaofeng Zhu, Zhifeng Hao, and Lifang He. Federated deep multi-view clustering with global self-supervision. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3498–3506, 2023.

Xinyue Chen, Yazhou Ren, Jie Xu, Fangfei Lin, Xiaorong Pu, and Yang Yang. Bridging gaps: Federated multi-view clustering in heterogeneous hybrid views. *Advances in Neural Information Processing Systems*, 37:37020–37049, 2024.

Xinyue Chen, Jinfeng Peng, Yuhao Li, Xiaorong Pu, Yang Yang, and Yazhou Ren. An effective and secure federated multi-view clustering method with information-theoretic perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 8871–8889, 2025.

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

Jinrong Cui, Bang Liufu, Chongjie Dong, Jingcheng Ke, and Jie Wen. Deep multi-view clustering with meta information compression. *IEEE Transactions on Image Processing*, 34:7514–7527, 2025.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E. Vogt. Identifiability results for multimodal contrastive learning. In *International Conference on Learning Representations*, 2023.

Shijie Deng, Jie Wen, Chengliang Liu, Ke Yan, Gehui Xu, and Yong Xu. Projective incomplete multi-view clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8): 10539–10551, 2023.

Xiaojian Ding, Lin Zhao, Xian Li, and Xiaoying Zhu. Incomplete multi-view clustering via hierarchical semantic alignment and cooperative completion. *Advances in Neural Information Processing Systems*, 38:52175–52195, 2026.

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

Bradley Efron. The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

Wei Feng, Zhenwei Wu, Qianqian Wang, Bo Dong, and Quanxue Gao. Federated fuzzy c-means with Schatten-p norm minimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9407–9416, 2024a.

Wei Feng, Zhenwei Wu, Qianqian Wang, Bo Dong, Zhiqiang Tao, and Quanxue Gao. Efficient federated multi-view clustering with integrated matrix factorization and k-means. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 3971–3979, 2024b.

Wei Feng, Zhenwei Wu, Qianqian Wang, Bo Dong, Zhiqiang Tao, and Quanxue Gao. Federated multi-view clustering via tensor factorization. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 3962–3970, 2024c.

Wei Feng, Danting Liu, Qianqian Wang, Wenqi Liang, and Zheng Yan. Scalable federated one-step multi-view clustering with tensorized regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16586–16594, 2025.

Wei Feng, Danting Liu, Qianqian Wang, Mengping Jiang, and Bin Liu. Federated incomplete multi-view clustering with tensorized low-rank constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21083–21091, 2026.

Constantine E. Frangakis and Donald B. Rubin. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, 86(2):365–379, 1999.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alex J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20, pp. 585–592, 2008.

Renxiang Guan, Xiang Yang, Hao Yu, Siwei Wang, Suyuan Liu, Wenjing Yang, Jun-Jie Huang, Ao Li, Xinwang Liu, and Yuhua Tang. Federated multi-view clustering for remote sensing data. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=kHLdixnMdi.

Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 1753–1759, 2017.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21, pp. 689–696, 2008.

Xingchen Hu, Jindong Qin, Yinghua Shen, Witold Pedrycz, Xinwang Liu, and Jiyuan Liu. An efficient federated multiview fuzzy c-means clustering method. *IEEE Transactions on Fuzzy Systems*, 32(4):1886–1899, 2024.

Shudong Huang, Wei Shi, Zenglin Xu, Ivor W. Tsang, and Jiancheng Lv. Efficient federated multi-view learning. *Pattern Recognition*, 131:108817, 2022.

Xiaorui Jiang, Zhongyi Ma, Yulin Fu, Yong Liao, and Pengyuan Zhou. Heterogeneity-aware federated deep multi-view clustering towards diverse feature representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9184–9193, 2024.

Peter Kairouz, H. Brendan McMahan, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1):1–210, 2021.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3519–3529, 2019.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1):83–97, 1955.

Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11169–11178, 2021.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, third edition, 2019.

Chengliang Liu, Zhihao Wu, Jie Wen, Yong Xu, and Chao Huang. Localized sparse incomplete multi-view clustering. *IEEE Transactions on Multimedia*, 25:5539–5551, 2022a.

Jiyuan Liu, Xinwang Liu, Siqi Wang, Xinhang Wan, Dongsheng Li, Kai Lu, and Kunlun He. Communication-efficient federated multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):17–32, 2026.

Suyuan Liu, Xinwang Liu, Siwei Wang, Xin Niu, and En Zhu. Fast incomplete multi-view clustering with view-independent anchors. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7740–7751, 2022b.

Suyuan Liu, Hao Yu, Hao Tan, Ke Liang, Siwei Wang, Shengju Yu, En Zhu, and Xinwang Liu. Efficient federated incomplete multi-view clustering. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 40100–40114, 2025.

Xinwang Liu, Xinzhong Zhu, Miaomiao Li, Lei Wang, Chang Tang, Jianping Yin, Dinggang Shen, Huaimin Wang, and Wen Gao. Late fusion incomplete multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2410–2423, 2018.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021.

Abdelmalik Moujahid and Fadi Dornaika. Advanced unsupervised learning: a comprehensive overview of multi-view clustering techniques. *Artificial Intelligence Review*, 58(8):234, 2025.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Yazhou Ren, Xinyue Chen, Jie Xu, Jingyu Pu, Yonghao Huang, Xiaorong Pu, Ce Zhu, Xiaofeng Zhu, Zhifeng Hao, and Lifang He. A novel federated multi-view clustering method for unaligned and incomplete data fusion. *Information Fusion*, 108:102357, 2024.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1670–1679, 2016.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Shuaijun Wang, Hui Huang, Nan Zhang, and Shiliang Sun. Federated incomplete multi-view clustering with cross-view relationship imputation. *IEEE Transactions on Knowledge and Data Engineering*, 38(7):4344–4356, 2026.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6638–6647, 2019.

Bingjun Wei, Xuemei Cao, Jiafen Liu, Haoyang Liang, and Xin Yang. Enhanced federated deep multi-view clustering under uncertainty scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26806–26813, 2026.

Jie Wen, Zheng Zhang, Zhao Zhang, Zhihao Wu, Lunke Fei, Yong Xu, and Bob Zhang. DIMC-net: Deep incomplete multi-view clustering network. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3753–3761, 2020.

Jie Wen, Zheng Zhang, Lunke Fei, Bob Zhang, Yong Xu, Zhao Zhang, and Jinxing Li. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2):1136–1149, 2023.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 478–487, 2016.

Jie Xu, Wenyuan Yang, Yazhou Ren, Lifang He, Philip S. Yu, and Xiaofeng Zhu. Missing pattern tree based decision grouping and ensemble for enhancing pair utilization in deep incomplete multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/TPAMI.2026.3729698.

Miin-Shen Yang and Kristina P. Sinaga. Federated multi-view k-means clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2446–2459, 2025.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):1–19, 2019.

Xihong Yang, Siwei Wang, Jiaqi Jin, Fangdi Wang, Tianrui Liu, Yueming Jin, Xinwang Liu, En Zhu, and Kunlun He. Generalized deep multi-view clustering via causal learning with partially aligned cross-view correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1990–1999, 2025.

Yan Yang and Hao Wang. Multi-view clustering: A survey. *Big Data Mining and Analytics*, 1(2): 83–107, 2018.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations*, volume 2024, pp. 34817–34848, 2024.

Jun Yin, Pei Wang, Shiliang Sun, and Zhonglong Zheng. Incomplete multi-view clustering via multi-level contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*, 37(8): 4716–4727, 2025.

Honglin Yuan, Yuan Sun, Fei Zhou, Jing Wen, Shihua Yuan, Xiaojian You, and Zhenwen Ren. Prototype matching learning for incomplete multi-view clustering. *IEEE Transactions on Image Processing*, 34:828–841, 2025.

Li Zhang, Pinhan Fu, Li Lv, Qian Guo, Liang Du, and Xinyan Liang. EvoFMVC: Trusted federated multi-view clustering with evolutionary fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28292–28300, 2026.

CONTENTS OF THE APPENDIX

A Notation **17**

B Detailed Related Work **19**

B.1 Multi-View Clustering 19

B.2 Incomplete Multi-View Clustering 19

B.3 Federated Multi-View Clustering 19

B.4 Federated Incomplete Multi-View Clustering 20

B.5 Missing Data, Causal Inference and Debiasing 20

B.6 Causal Representation Learning and Identifiability 20

B.7 Positioning of CFIMC 21

C Supplementary Preliminaries **21**

C.1 Causal Background 21

C.2 Discussion of the Assumptions 21

C.3 Selection, Reweighting and Transport 22

C.4 Kernel Independence and Deep Embedded Clustering 23

D Additional Method Details and Theoretical Results **24**

D.1 Client Side 24

D.2 Module I: Missingness Mechanism 25

D.3 Module II: Counterfactual Imputation 26

D.4 Module III: Reliability and Doubly Robust Fusion 28

D.5 Server Clustering 30

E Proofs **30**

E.1 Conventions and Auxiliary Results 30

E.2 Proofs for Section 2 and Appendix C 32

E.3 Proofs for the Client Side (Section 3.1 and Appendix D.1) 34

E.4 Proofs for Module I (Section 3.2 and Appendix D.2) 34

E.5 Proofs for Module II (Section 3.3 and Appendix D.3) 37

E.6 Proofs for Module III (Section 3.4 and Appendix D.4) 39

E.7 Proofs for the Server Clustering (Section 3.5 and Appendix D.5) 43

E.8 Proofs for Communication and Privacy (Appendix G) 44

F Detailed Algorithms and Default Hyperparameters **46**

G Communication and Privacy Analysis **48**

G.1 Communication Cost Analysis 48

G.2 Privacy Analysis 48

H Additional Experimental Details and Results **49**

H.1 Datasets 49

H.2 Missingness Mechanisms 49

H.3 Additional Results 49

H.4 Ablation over Communication Rounds 49

H.5 Sensitivity Grids and Results on MSRC-v1 49

H.6 Oracle Audit Protocol 49

A NOTATION

Table 2 collects all notations used in the paper, grouped by the component in which they first appear. Random variables are written in capitals, vectors in bold lowercase and matrices in bold uppercase, and $[N] = \{1, \dots, N\}$. The colored group headers follow the color code of Figures 1 and 2.

Table 2: Summary of notation.

Symbol	Description	Where
Data and federation		
$N,\ M,\ K$	numbers of samples, views (clients) and clusters	Def. 2.1
$i, j \in [N]$; $m, v \in [M]$; $k, l \in [K]$	indices of samples, views and clusters	Sec. 2
$\boldsymbol{x}_i^{(m)} \in \mathbb{R}^{D_m}$	raw feature of view m of sample i; D_m is its dimension	Def. 2.1
$\boldsymbol{R} = [r_{im}], \boldsymbol{r}_i$	mask matrix; missing pattern of sample i	Def. 2.1
$\mathcal{O}_m, N_m$	samples whose view m is visible, and their number	Def. 2.1
$\mathcal{O}_m^+$	samples in $\mathcal{O}_m$ with at least one other visible view	Eq. (10)
$\mathcal{V}_i, \mathfrak{R}$	visible views of sample i; admissible patterns $\{0,1\}^M \setminus \{\mathbf{0}\}$	Sec. 2
T, t	number of communication rounds; round index	Def. 2.1
Clients and self-supervision		
$\mathrm{Enc}_m,\ \mathrm{Dec}_m$	encoder and decoder of client m	Sec. 3.1
$\boldsymbol{z}_i^{(m)} \in \mathbb{R}^d, \boldsymbol{Z}^{(m)}$	embedding of view m of sample i and embedding matrix; d is the dimension	Sec. 3.1
$\bar{\boldsymbol{z}}^{(m)}, \varsigma_m$	visible mean and scale used for server-side standardization	App. F
$\boldsymbol{u}_k^{(m)}, \boldsymbol{U}^{(m)}$	local cluster centers of client m	Sec. 3.1
$q_{ik}^{(m)}, \boldsymbol{Q}^{(m)}$	client soft assignment (Student's t)	Eq. (21)
$p_{ik}, \boldsymbol{P}, \boldsymbol{P}^{(m)}$	global sharpened targets; rows of client m	Eq. (15)
$\mathcal{L}_{\mathrm{rec}}^{(m)}, \mathcal{L}^{(m)}, \beta$	reconstruction loss; local objective; clustering weight	Eq. (3)
$\boldsymbol{w}^{(m)}$	downloaded inverse-probability weights $(w_{im}^{\mathrm{ipw}})_{i \in \mathcal{O}_m}$	Alg. 1
$T_{\mathrm{pre}}, T_{\mathrm{loc}}$	pretraining epochs; local epochs per round	Alg. 2
Causal model		
Y	latent cluster	Sec. 2
$C \in \mathbb{S}^{d_c - 1}$	shared content (semantics) on the unit sphere of $\mathbb{R}^{d_c}$	Sec. 2
$S^{(m)}$	private factor of view m	Eq. (2)
$X^{(m)}, X^{(m)*}$	full view; observed proxy ($\star$ if missing)	Sec. 2
$Z^{(m)}$	embedding of the full view, $Z^{(m)} = \mathrm{Enc}_m(X^{(m)})$	Ass. 2.3
$R^{(m)}, \mathbf{R}, \mathbf{R}^{(-m)}$	visibility indicator; pattern; pattern without view m	Sec. 2
$\mathcal{U}, U_{(\cdot)}$	exogenous variables	Def. C.1
g_m	generating mechanism of view m	Eq. (2)
$f_m, \boldsymbol{\xi}^{(m)}$	mechanism in embedding space (also its network estimate); additive noise	Ass. 2.3
$h_m^\star$	content-recovery map	Ass. 2.3
$\mathcal{G}_{\mathrm{IVM}}, \mathrm{do}(\cdot)$	m-graph of informative view missingness; intervention operator	Ass. 2.2
η	positivity constant	Ass. 2.2
(P), (O)	cluster-level propensity condition $\mathbf{R} \perp\!\!\!\perp C \mid Y$; accurate-imputer condition	Sec. 2, Thm. D.14
Module I: missingness mechanism		
$\varpi_k(\boldsymbol{r})$	law of the pattern given $Y = k$	Eq. (1)
$\pi_m(k), \bar{\pi}_m$	cluster-conditional and marginal visibility of view m	Eq. (1)
$\pi_m^\star(Y, C)$	fine-grained propensity	Eq. (1)
$\hat{\pi}_m(k), \hat{\bar{\pi}}_m$	estimated visibility table; estimated marginal	Eq. (4)

continued on next page

Table 2 continued

Symbol	Description	Where
$\hat{\pi}_{im}$, $\check{\pi}_m(k)$	sample-level propensity; clipped table	Eq. (4)
α, $\underline{\pi}$	pseudo-count of the Beta prior; propensity floor	Eq. (4)
a_i, a_{ik}	hard assignment of sample i; its indicator $\mathbb{I}[a_i = k]$	Sec. 3.2
n_k, n_{km}	size of cluster k; its members with view m visible	Sec. 3.2
Λ_{ik}, $\Lambda_{ik}^{\mathrm{cf}}$	missingness evidence; cross-fitted evidence	Eq. (5), Eq. (7)
$\mathcal{H}_0$, $\mathcal{H}_1$	random halves used for cross-fitted evidence	Eq. (7)
λ_R	weight of the missingness evidence	Eq. (6)
σ^2, σ_0^2	per-dimension within-cluster variance; value frozen in round 0	Eq. (6)
w_{im}^{ipw}, w_{im}^{tr}	inverse-probability and transport weights (mean one over $\mathcal{O}_m$)	Def. C.6
Module II: counterfactual imputation		
h_m, d_h	abduction network of view m; hidden width	Eq. (8)
$\boldsymbol{c}_i^{(v)}$, $\bar{\boldsymbol{c}}_i^{(-m)}$	per-view content; pooled content of the other visible views	Eq. (8)
$\hat{\boldsymbol{z}}_i^{(m)}$	counterfactual imputation of view m of sample i	Eq. (9)
$\mathcal{L}_{\mathrm{cf}}^{(m)}$, $\mathcal{L}_{\mathrm{inv}}^{(m)}$, $\mathcal{L}_{\mathrm{ind}}^{(m)}$	counterfactual, invariance and independence losses	Eq. (28) to Eq. (30)
$\mathcal{L}_{\mathrm{SCM}}$, λ_{inv}, λ_{ind}, τ	server objective; loss weights; InfoNCE temperature	Eq. (27)
p_{drop}, $\tilde{r}_{iv}$, $\tilde{\boldsymbol{c}}_i^{(-m)}$	view-dropout rate; dropped mask; pooled content under dropout	App. D.3
B, $\mathcal{B}_m$, T_{s}	batch size; usable batch for view m; server epochs	App. D.3
L, $\mathcal{I}_\ell$	number of cross-fitting folds; fold ℓ	Eq. (32)
g_r^{as}, W_1	associational imputer of population r; Wasserstein-1 distance	Prop. D.10
ν, L_f	content-estimation error scale; Lipschitz constant of f_m	Prop. D.10
Module III: reliability and doubly robust fusion		
$\hat{g}_m$, $\hat{g}_{im}$	cross-fitted imputer as a function; its value at sample i	Sec. 3.4
$\hat{\rho}_m$, ρ_m	transported reliability; its population target	Eq. (10), Eq. (34)
$\tilde{\boldsymbol{z}}_i^{(m)}$, ω_{im}	completed block; fusion weight	Eq. (11)
$\boldsymbol{\mu}_k^{(m)}$, $\boldsymbol{\mu}_{km}$	prototype of cluster k in view m; population cluster mean	Eq. (14)
$\hat{\boldsymbol{\mu}}_{km}^{\mathrm{ipw}}$, $\hat{\boldsymbol{\mu}}_{km}^{\mathrm{or}}$, $\hat{\boldsymbol{\mu}}_{km}^{\mathrm{dr}}$	IPW, outcome-regression and doubly robust estimates	Eq. (12) to Eq. (13)
s_{km}, n_0	shrinkage weight; prior strength	Eq. (14)
B_{km}^{sel}	within-cluster selection bias of the visible mean	Thm. D.14
CV_{km}, sd_k	coefficient of variation of $\pi^\star$ within cluster k; standard deviation	Thm. D.14
$\hat{\mathcal{L}}_{km}^{\mathrm{dr}}$	doubly robust risk	Eq. (40)
$\mathrm{MSE}_{\mathrm{vis}}$, $\mathrm{MSE}_{\mathrm{mis}}$, $e_m(k)$	imputation errors on visible and missing populations; per cluster	Eq. (35)
Server clustering		
Δ_{ik}, $\tilde{\Delta}_{ik}$, $\tilde{\Delta}_{ik}^{\mathrm{cf}}$	feature distance; effective distance; cross-fitted version	Eq. (6) to Eq. (7)
$\delta(i, j)$	partial distance for masked k-means++ seeding	Eq. (41)
γ_{ik}, $\hat{y}_i$	global soft assignment; final label	Eq. (15)
J	classification EM objective	Eq. (42)
n_{init}, I_{cl}	seeding restarts; alternation sweeps	App. D.5
Communication and privacy		
b, Comm	bits per scalar; total communication in bits	Eq. (54)
$\mathcal{D}_m$, $\mathcal{D}_{-m}$	information held and not held by client m	Eq. (56)
Γ, Π_Γ	clipping radius; projection onto the ball of radius Γ	Prop. G.3
σ_{dp}, $(\varepsilon_{\mathrm{dp}}, \delta_{\mathrm{dp}})$	Gaussian noise level; privacy parameters	Prop. G.3
Operators		

continued on next page

Table 2 continued

Symbol	Description	Where
$\perp\!\!\!\perp$, $\mathbb{I}[\cdot]$	(conditional) independence; indicator	Sec. 2
$\mathrm{nrm}(\boldsymbol{a})$, $\mathrm{clip}_{[a,b]}$	projection $\boldsymbol{a}/\|\boldsymbol{a}\|_2$ onto the sphere; clipping to $[a, b]$	Sec. 2
D_{KL}, kl	Kullback-Leibler divergence; binary version	Thm. D.3
HSIC, nHSIC	Hilbert-Schmidt independence criterion; normalized estimator	Def. C.8
P_k, $\mathbb{E}_k$, Var_k, Cov_k	probability, expectation, variance, covariance given $Y = k$	Sec. 2
plim, $\Rightarrow$	probability limit; convergence in distribution	Thm. D.14

B Detailed Related Work

B.1 Multi-View Clustering

Multi-view clustering partitions samples described by several feature sets by exploiting the consistency of the cluster structure across views and the complementary information that each view contributes; comprehensive surveys are given by Yang & Wang (2018) and Chao et al. (2021). Classical approaches include co-training and co-regularized spectral clustering, multiple kernel k-means, subspace and matrix-factorization models, and graph-based fusion. Deep embedded clustering (Xie et al., 2016) and its locality-preserving extension (Guo et al., 2017) introduced the Student's t soft assignment and the sharpened self-training target that remain the backbone of most deep multi-view methods, including the global self-supervision protocol used in this paper. Recent causal views of MVC separate shared content from view-specific style: CausalMVC (Bao et al., 2025) learns content and style representations with a causal model to purify cluster semantics, and CauMVC (Yang et al., 2025) uses causal learning to generalize under partially aligned cross-view correspondence. These works are centralized and address alignment or style confounding, not the mechanism by which views go missing.

B.2 Incomplete Multi-View Clustering

IMVC handles samples that lack some views (Wen et al., 2023). Shallow methods integrate incomplete kernels or graphs: LF-IMVC (Liu et al., 2018) performs late fusion of incomplete base partitions and imputation in linear complexity, FIMVC-VIA (Liu et al., 2022b) learns view-independent anchors for large-scale data, PIMVC (Deng et al., 2023) learns a projective consensus representation with graph regularization, and LSIMVC (Liu et al., 2022a) learns a sparse, locally structured consensus that balances information across views. Deep methods learn view-specific encoders and recover or bypass missing views: DIMC-net (Wen et al., 2020) uses a weighted fusion layer and a fusion-graph constraint, COMPLETER (Lin et al., 2021) unifies consistent representation learning and data recovery through contrastive learning and dual prediction from an information-theoretic viewpoint, ICMVC (Chao et al., 2024) guides completion with high-confidence cross-view neighbors, PMIMC (Yuan et al., 2025) matches misaligned cross-view prototypes, IMC-MCL (Yin et al., 2025) recovers missing views by minimizing conditional entropy with multi-level contrastive learning, and HSACC (Ding et al., 2026) couples hierarchical semantic alignment with cooperative completion. TreeEIC (Xu et al., 2026) is the closest in spirit to our use of the pattern: it groups samples into decision sets according to their missing patterns and ensembles the resulting clusterings to improve the utilization of available pairs. In TreeEIC, the pattern is a structural device for grouping, whereas in `CFIMC` it is a random variable caused by the latent semantics and therefore evidence about the cluster. More broadly, all IMVC methods above treat the mask as an index set and implicitly assume that it carries no information about the clusters.

B.3 Federated Multi-View Clustering

Federated learning trains models on decentralized data (McMahan et al., 2017; Kairouz et al., 2021); when parties hold different feature sets of overlapping samples, the setting is vertical (Yang et al., 2019), which is the natural home of multi-view data. Early work established federated multi-view learning (Huang et al., 2022) and lossless federated c-means for horizontally and vertically partitioned data (Bárcena et al., 2024). Subsequent FedMVC methods improved efficiency and structure

mining: FMVC-IMK (Feng et al., 2024b) integrates matrix factorization with k-means in one step, TensorFMVC (Feng et al., 2024c) imposes tensor factorization across clients, FMVFCMSP (Feng et al., 2024a) combines fuzzy c-means with Schatten p-norm minimization, SFOMVC-TR (Feng et al., 2025) obtains labels directly with tensorized regularization, and CeFMC (Liu et al., 2026) approximates representations by pseudo-labels and centroids to reduce communication with linear complexity. Heterogeneity and trust have also been addressed: HFMVC (Jiang et al., 2024) perceives the degree of heterogeneity and adapts contrastive learning accordingly, EFDMVC (Wei et al., 2026) tackles view and aggregation uncertainty, EvoFMVC (Zhang et al., 2026) communicates lightweight trusted evidence and searches fusion operators evolutionarily, and FedRSMVC (Guan et al., 2026) shares privatized prototypes for remote sensing data. Horizontal variants include federated multi-view fuzzy clustering (Hu et al., 2024) and federated multi-view k-means (Yang & Sinaga, 2025). These methods assume complete views.

B.4 Federated Incomplete Multi-View Clustering

FedDMVC (Chen et al., 2023) is the first deep FedIMVC method: the server explores complementary cluster structures through sample alignment and data extension and distributes global prototypes and pseudo-labels as global self-supervision, and clients train autoencoders with this signal. FCUIF (Ren et al., 2024) removes both the alignment and the completeness assumptions; it generates alignment matrices from sample commonality and view versatility, imputes missing embeddings from global prototypes weighted by consensus assignments, and evaluates and refines imputation quality to decide whether imputations are adopted. FMCSC (Chen et al., 2024) addresses heterogeneous hybrid views with local-synergistic contrastive learning and global-specific weighting aggregation. FIMCFG (Chao et al., 2025) injects a globally fused graph into client training. EFIMVC (Liu et al., 2025) uses view-specific and shared anchor graphs as communication variables with a localized optimization strategy and convergence guarantees. ESFMC (Chen et al., 2025) splits client features from an information-theoretic perspective, sharing only task-relevant parts. FIMCI (Wang et al., 2026) captures global cross-view dependencies with Transformer encoders and dynamic view fusion to impute missing views under high missing rates. FIMVC-TLRC (Feng et al., 2026) aligns local anchor graphs under a tensorized low-rank constraint for efficiency. These contributions established strong baselines in accuracy, communication, and privacy. Common to all of them is the treatment of the mask as an index set: none models the mechanism that generates it, uses the pattern as evidence, corrects the selection it induces on clients and imputers, or evaluates imputations on the missing population.

B.5 Missing Data, Causal Inference and Debiasing

Rubin's taxonomy of MCAR, MAR and MNAR (Little & Rubin, 2019) underlies all principled treatments of missing data. Latent ignorability (Frangakis & Rubin, 1999) allows missingness to depend on unobserved latent quantities, such as compliance types, but not on the missing value itself given them; our Lemma C.4 derives it from the SCM of Assumption 2.2(i). Missingness graphs (Mohan & Pearl, 2021) encode missingness mechanisms in causal DAGs and characterize recoverability; transportability theory (Bareinboim & Pearl, 2016) formalizes when mechanisms learned in one population transfer to another, which is the logic behind our invariance result (Theorem D.9). Inverse probability weighting (Horvitz & Thompson, 1952), augmented inverse probability weighting and double robustness (Bang & Robins, 2005), efficient semiparametric estimation (van der Vaart, 1998) and cross-fitting (Chernozhukov et al., 2018) provide the estimation toolbox that Module III adapts to prototype estimation. Importance weighting under covariate shift (Sugiyama et al., 2007) motivates our transport weights, and the optimism of in-sample error (Efron, 2004) motivates cross-fitted reliability. In recommender systems, inverse propensity scoring (Schnabel et al., 2016) and doubly robust learning (Wang et al., 2019) correct ratings that are missing not at random; to our knowledge, such mechanism-aware estimators have not been used in IMVC or FedIMVC.

B.6 Causal Representation Learning and Identifiability

Causal representation learning seeks latent variables with causal semantics (Schölkopf et al., 2021). For multi-view data, content shared across views is identifiable up to an invertible map under contrastive or correlation-based objectives (Daunhawer et al., 2023), including under partial observability (Yao et al., 2024); this literature justifies Assumption 2.3(ii) and the InfoNCE term (van den

Oord et al., 2018) in Eq. (29). The additive noise model (Hoyer et al., 2008) and kernel independence measures (Gretton et al., 2008; Kornblith et al., 2019) motivate the independence term Eq. (30).

B.7 Positioning of CFIMC

CFIMC is complementary to the architectural advances above: it keeps a standard global self-supervision backbone and replaces the components that silently assume MCAR by components derived from an explicit m-graph. To our knowledge it is the first FedIMVC method that (i) treats the missing pattern as evidence about the cluster, (ii) corrects client-side and imputer-side selection with inverse-probability and transport weights, (iii) imputes counterfactually at the sample level with cross-fitting, and (iv) replaces binary adoption switches by a label-free reliability that scales a doubly robust correction. Several of these components, such as the transported reliability and the doubly robust prototypes, could be combined with other backbones.

C Supplementary Preliminaries

This appendix supports Section 2. Appendix C.1 recalls the causal vocabulary and the taxonomy of missingness mechanisms, Appendix C.2 states the latent-ignorability lemma, discusses the assumptions and lists those used by every result, Appendix C.3 states the identities behind the inverse-probability and transport weights, and Appendix C.4 defines the kernel independence measure and the self-supervision targets used by the method.

C.1 Causal Background

Definition C.1 (Structural causal model, intervention, counterfactual; Pearl, 2009)**.** A structural causal model (SCM) $\mathfrak{M} = \langle \mathcal{U}, \mathcal{W}, \mathcal{F}, P(\mathcal{U}) \rangle$ consists of exogenous variables $\mathcal{U}$ with law $P(\mathcal{U})$, endogenous variables $\mathcal{W}$ and assignments $W_j := F_j(\mathrm{Pa}(W_j), U_j)$ that induce a directed acyclic graph. The intervention $\mathrm{do}(W_j = w)$ replaces F_j by the constant w and yields the submodel $\mathfrak{M}_w$. Given evidence $\boldsymbol{e}$, the counterfactual law of $W_{j'}$ under $\mathrm{do}(W_j = w)$ is obtained in three steps: *abduction* (update $P(\mathcal{U})$ to $P(\mathcal{U} \mid \boldsymbol{e})$), *action* (form $\mathfrak{M}_w$) and *prediction* (evaluate $W_{j'}$ in $\mathfrak{M}_w$ under $P(\mathcal{U} \mid \boldsymbol{e})$).

Definition C.2 (Missingness graph and recoverability; Mohan & Pearl, 2021)**.** A missingness graph (m-graph) is a causal DAG over $\mathcal{W} \cup \mathbf{R} \cup \mathcal{W}^* \cup \mathcal{U}$ in which each proxy $X^{(m)*}$ has parents $\{X^{(m)}, R^{(m)}\}$. A functional Q of the full-data law is *recoverable* if it is uniquely determined by the observed-data law $P(X^{(1:M)*}, \mathbf{R})$ for every model compatible with the m-graph that satisfies positivity.

Definition C.3 (Missingness mechanisms for views)**.** The views are *missing completely at random* (MCAR) if $\mathbf{R} \perp\!\!\!\perp (Y, C, X^{(1:M)})$, *missing at random* (MAR) if $P(\mathbf{R} \mid X^{(1:M)})$ depends only on the observed coordinates $X^{(\mathcal{V})}$, and *missing not at random* (MNAR) otherwise (Little & Rubin, 2019). The missingness is *informative about the clusters*, abbreviated IVM, if $\varpi_k \neq \varpi_l$ for some $k \neq l$, equivalently $I(Y; \mathbf{R}) > 0$. Since Y is never observed, IVM is MNAR with respect to the clustering target.

In this vocabulary, Assumption 2.2(i) specifies an SCM whose m-graph $\mathcal{G}_{\mathrm{IVM}}$ is shown in Figure 1(b), Theorem D.9(ii) expresses the view mechanism f_m as a functional of the observed-data law, and Module II answers the counterfactual query $\mathrm{do}(R^{(m)} = 1)$ by the three steps of Definition C.1.

C.2 Discussion of the Assumptions

The structural part of Assumption 2.2 has the following consequences, which underlie all components of CFIMC.

Lemma C.4 (Latent ignorability and mechanism invariance)**.** *Under Assumption 2.2(i): (i)* $\mathbf{R} \perp\!\!\!\perp \left(X^{(1:M)}, S^{(1:M)}\right) \mid (Y, C)$ *and* $\mathbf{R} \perp\!\!\!\perp X^{(1:M)} \mid C$*, that is, missingness is* latently ignorable *(Frangakis & Rubin, 1999); (ii)* $X^{(m)} \perp\!\!\!\perp \left(Y, \mathbf{R}, X^{(-m)}\right) \mid C$ *for every* m*; (iii) consequently* $P(X^{(m)} \mid C, \mathbf{R}) = P(X^{(m)} \mid C)$*, i.e., the generating mechanism of every view is invariant between the visible and the missing subpopulations.*

Latent ignorability. Lemma C.4 shows that Assumption 2.2(i) makes the missingness *latently ignorable*: it may depend on the semantics (Y, C), but not on the value of a view given the semantics. This covers MCAR as well as mechanisms driven by the unobserved semantics, which are MNAR with respect to the observed data. It excludes only self-masking, in which the value of a view causes its own absence given the semantics; no method that uses only $(\mathbf{R}, X^{(\mathcal{V})})$ can identify such mechanisms without further assumptions (Mohan & Pearl, 2021). Assumption 2.2(i) does not by itself tie the mechanism to the clusters; Condition (P) does, and every result states explicitly whether it needs it.

Remark C.5 (Mechanisms in our experiments). The two protocols used in our experiments instantiate IVM with $\pi_m^\star$ depending on Y only (`mnar`) and IVM with $\pi_m^\star$ depending on a continuous latent factor of the content, hence heterogeneous within clusters (`mnar_latent`). Condition (P) holds for `mnar` but fails for `mnar_latent`. We do not run MCAR, under which only the counterfactual imputation of `CFIMC` remains active (Corollary D.19).

View-wise conditional independence. Assumption 2.2(ii) states that, given the semantics, each view is revealed independently of the others. It is violated only through the admissibility constraint $\mathbf{R} \neq \mathbf{0}$, whose effect is confined to samples that observe a single view (Lemma E.3 and Remark E.4 in Appendix E.1).

Positivity. Assumption 2.2(iii) is the usual overlap condition of causal inference. It bounds the population inverse-probability weights by $1/\eta$ and guarantees co-observed pairs of views for almost every value of the content, which Theorem D.9(ii) requires; the floor $\underline{\pi}$ in Eq. (4) is its empirical counterpart.

Regularity of the embeddings. Assumption 2.3(i) is the additive noise model of Hoyer et al. (2008) written in embedding space; it implies $f_m(C) = \mathbb{E}[Z^{(m)} \mid C]$. Assumption 2.3(ii) holds whenever g_m and Enc_m are injective, which is the regime in which multi-view content is identifiable up to an invertible map that we absorb into C (Daunhawer et al., 2023; Yao et al., 2024); Proposition D.10 quantifies what happens when it holds only approximately. Both parts concern the learned embeddings rather than the data-generating process, and they are used only by the results on counterfactual imputation and on the doubly robust prototypes.

Condition (P). Condition (P) makes the propensity a function of the cluster alone. It holds for MCAR and for cluster-dependent MNAR, but fails when visibility also varies within clusters (Remark C.5). The algorithm never requires it. It is used by the results that analyze cluster-level propensities, namely the identities of Lemma C.7, the debiasing of the clients (Theorem D.2), the factorization of the posterior (Theorem D.3) and the transported reliability (Theorem D.13). Without it, the doubly robust prototypes remain consistent for an accurate imputer (condition (O) of Theorem D.14), and their bias obeys the product-form bound Eq. (37) in general. Table 3 lists the assumptions used by every result.

C.3 Selection, Reweighting and Transport

Module I turns the visibility table into the following weights (Section 3.2).

Definition C.6 (Inverse-probability and transport weights)**.** For a working propensity $\tilde{\pi}_m$ with values in $[\eta, 1]$, the inverse-probability weight is $w_m^{\mathrm{ipw}} = R^{(m)}/\tilde{\pi}_m$ and the transport weight is $w_m^{\mathrm{tr}} = R^{(m)}(1 - \tilde{\pi}_m)/\tilde{\pi}_m$. Their empirical counterparts w_{im}^{ipw} and w_{im}^{tr} are normalized to have mean one over $\mathcal{O}_m$.

Table 3: Assumptions used by each result. A1 and A2 abbreviate Assumptions 2.2 and 2.3; roman numerals refer to their parts, ✓ means all parts and · means not used. Notes: [a]A1(ii) only for the product forms in parts (ii) and (iii); [b]A2(ii) only for parts (ii) and (iii); [c]only for part (ii), where (O) is the accurate-imputer condition of Theorem D.14; [d]only for the population statement in part (i). The abridged Theorems 3.1 to 3.4 of Section 3 use the same assumptions as their full versions.

Result	Where	A1	A2	(P)
Lemma C.4 (latent ignorability)	App. C.2	(i)	·	·
Lemma C.7 (selection, reweighting, transport)	App. C.3	✓	·	✓
Lemma E.5 (propensity given the full data)	App. E.2	(i), (ii)	·	✓
Proposition D.1 (selection-induced cluster shift)	App. D.1	✓	·	✓
Theorem D.2 (IPW local training)	App. D.1	✓	·	✓
Theorem D.3 (missingness as evidence)	App. D.2	(i), (ii)[a]	·	✓
Theorem D.9 (invariance, identifiability)	App. D.3	✓	(i), (ii)[b]	·
Proposition D.10 (transport gap)	App. D.3	(i)	(i)	·
Proposition D.11 (structural causal objective)	App. D.3	✓	✓	✓
Lemma D.12 (cross-fitting)	App. D.3	(i)	(i)	·
Theorem D.13 (transported reliability)	App. D.4	✓	·	✓
Theorem D.14 (doubly robust prototypes)	App. D.4	✓	(i)	(P) or (O)[c]
Theorem D.16 (asymptotic normality)	App. D.4	✓	(i)	✓
Proposition D.17 (doubly robust risk)	App. D.4	✓[d]	(i)[d]	(P) or (O)[d]
Lemmas D.4, D.7; Proposition D.5; Corollary D.6	App. D.2	·	·	·
Lemma D.8 (circularity of cluster-level imputation)	App. D.3	·	·	·
Proposition D.18; Corollary D.19	App. D.5	·	·	·
Propositions G.1, G.2 and G.3	App. G	·	·	·

Lemma C.7 (Selection, reweighting and transport)**.** *Let Assumption 2.2 and Condition (P) hold and let $\phi = \phi(Y, C, X^{(1:M)}, \mathbf{R}^{(-m)})$ be integrable. Then*

$$\textit{(a)}\quad \mathbb{E}\big[\phi \mid R^{(m)} = 1\big] = \frac{\mathbb{E}[\pi_m(Y)\,\phi]}{\bar{\pi}_m}; \tag{16}$$

$$\textit{(b)}\quad \mathbb{E}\Big[\frac{R^{(m)}}{\pi_m(Y)}\,\phi\Big] = \mathbb{E}[\phi]; \tag{17}$$

$$\textit{(c)}\quad \mathbb{E}\Big[\frac{1-\pi_m(Y)}{\pi_m(Y)}\,\phi \;\Big|\; R^{(m)} = 1\Big] = \frac{1-\bar{\pi}_m}{\bar{\pi}_m}\,\mathbb{E}\big[\phi \mid R^{(m)} = 0\big]; \tag{18}$$

$$\textit{(d)}\quad \mathbb{E}\big[\phi \mid R^{(m)} = 1\big] - \mathbb{E}\big[\phi \mid R^{(m)} = 0\big] = \frac{\mathrm{Cov}\big(\mathbb{E}[\phi \mid Y],\, \pi_m(Y)\big)}{\bar{\pi}_m(1-\bar{\pi}_m)}. \tag{19}$$

In particular, (a) with $\phi = \mathbb{I}[Y = k]$ gives $P\big(Y = k \mid R^{(m)} = 1\big) = \pi_m(k)P(Y = k)/\bar{\pi}_m$.

Identity (a) describes the selected population seen by client m, (b) reweights it back to the full population, (c) transports it to the missing population, and (d) quantifies the gap between the visible and the missing populations. Identity (c) also explains the normalization in Definition C.6: after rescaling to mean one on the visible entries, transport-weighted averages over visible entries estimate averages over the *missing* entries.

C.4 Kernel Independence and Deep Embedded Clustering

The independence term of Module II and the self-supervision of the clients and of the server rely on two standard tools.

Definition C.8 (Hilbert-Schmidt independence criterion)**.** For random vectors U, V with characteristic kernels, $\mathrm{HSIC}(U, V) = \|\mathcal{C}_{UV}\|^2_{\mathrm{HS}}$ is the squared Hilbert-Schmidt norm of their cross-covariance operator, and $\mathrm{HSIC}(U, V) = 0$ if and only if $U \perp\!\!\!\perp V$ (Gretton et al., 2008). On

a batch of size B with Gaussian Gram matrices $\boldsymbol{K}_U, \boldsymbol{K}_V$ (median-heuristic bandwidth) and $\boldsymbol{H}_B = \boldsymbol{I}_B - \frac{1}{B}\mathbf{1}\mathbf{1}^\top$, we use the normalized estimator

$$\mathrm{nHSIC}(U,V) = \frac{\mathrm{Tr}\left(\tilde{\boldsymbol{K}}_U \tilde{\boldsymbol{K}}_V\right)}{\sqrt{\mathrm{Tr}\left(\tilde{\boldsymbol{K}}_U \tilde{\boldsymbol{K}}_U\right)\ \mathrm{Tr}\left(\tilde{\boldsymbol{K}}_V \tilde{\boldsymbol{K}}_V\right)}} \in [0,1], \qquad \tilde{\boldsymbol{K}} = \boldsymbol{H}_B \boldsymbol{K} \boldsymbol{H}_B, \tag{20}$$

which coincides with centered kernel alignment (Kornblith et al., 2019).

Definition C.9 (Deep embedded clustering)**.** For embeddings $\{\boldsymbol{z}_i\}$ and centers $\{\boldsymbol{u}_k\}$, the Student's t soft assignment and its sharpened target (Xie et al., 2016) are

$$q_{ik} = \frac{(1 + \|\boldsymbol{z}_i - \boldsymbol{u}_k\|_2^2)^{-1}}{\sum_{l=1}^{K}(1 + \|\boldsymbol{z}_i - \boldsymbol{u}_l\|_2^2)^{-1}}, \qquad p_{ik} = \frac{q_{ik}^2 / \sum_j q_{jk}}{\sum_{l=1}^{K} q_{il}^2 / \sum_j q_{jl}}. \tag{21}$$

D Additional Method Details and Theoretical Results

Section 3 states four main theorems of CFIMC in abridged form and summarizes a fifth in words. This appendix gives their full statements (Theorems D.2, D.3, D.9, D.13 and D.14; each main-text version restates parts of the corresponding full statement) and collects the remaining results and implementation details, organized by the component of the method to which they belong; proofs are given in Appendix E.

D.1 Client Side

The following result makes precise why the unweighted local loss is biased under IVM; Theorem D.2 shows that the weights in Eq. (3) remove exactly this shift.

Proposition D.1 (Selection-induced cluster shift)**.** *Let Assumption 2.2 and Condition (P) hold and let $\ell = \ell(Y, C, X^{(1:M)}, \mathbf{R}^{(-m)}; \theta)$ be a loss whose gradient admits an integrable envelope. The population seen by client m has cluster law $P(Y = k \mid R^{(m)} = 1) = \pi_m(k)P(Y = k)/\bar{\pi}_m$, and the expected unweighted local gradient equals*

$$\mathbb{E}\left[\nabla_\theta \ell \mid R^{(m)} = 1\right] = \sum_{k=1}^{K} P(Y = k)\, \frac{\pi_m(k)}{\bar{\pi}_m}\, \mathbb{E}\left[\nabla_\theta \ell \mid Y = k\right]. \tag{22}$$

Hence cluster k is under-weighted by the factor $\pi_m(k)/\bar{\pi}_m$, and its contribution vanishes as $\pi_m(k) \to 0$ although its population mass is positive.

Inverse-probability-weighted local training (summarized in Section 3.1). The interventional identity holds at the population level and, under an integrable envelope, for the gradients:

Theorem D.2 (Inverse-probability-weighted local training is interventional)**.** *Let Assumption 2.2 and Condition (P) hold, let $\ell^{(m)} = D_{\mathrm{KL}}\big(\boldsymbol{p}(X^{(1:M)}, \mathbf{R}) \,\|\, \boldsymbol{q}^{(m)}(Z^{(m)}; \theta_m)\big)$ for any target $\boldsymbol{p}$ that is a measurable function of $(X^{(1:M)}, \mathbf{R})$, and let $\hat{\pi}_{im} = \pi_m(y_i)$. Then:*

- *(i) $\mathbb{E}\big[\frac{R^{(m)}}{\pi_m(Y)}\,\ell^{(m)}\big] = \mathbb{E}_{\mathrm{do}(R^{(m)}=1)}\big[\ell^{(m)}\big] = \mathbb{E}\big[\ell^{(m)}\big(X^{(1:M)}, (1, \mathbf{R}^{(-m)})\big)\big]$, where the middle term is the risk in the submodel of Eq. (2) in which view m is observed for every sample; under an integrable envelope the identity also holds for ∇_{θ_m};*
- *(ii) the weighted empirical objective in Eq. (3) is consistent for this interventional risk, whereas the unweighted objective converges to the selected risk $\mathbb{E}\big[\ell^{(m)} \mid R^{(m)} = 1\big]$;*
- *(iii) for any $\hat{\pi}_{im} \in [\underline{\pi}, 1]$, the Kish effective sample size of Eq. (3) satisfies*

$$\frac{\left(\sum_{i \in \mathcal{O}_m} w_{im}^{\mathrm{ipw}}\right)^2}{\sum_{i \in \mathcal{O}_m} \left(w_{im}^{\mathrm{ipw}}\right)^2} \;\geq\; \frac{4\underline{\pi}}{(1+\underline{\pi})^2}\, N_m. \tag{23}$$

D.2 Module I: Missingness Mechanism

Full statement of Theorem 3.1. The full version below also writes the evidence in log-odds form, relates it to the mutual information $I(Y;\mathbf{R})$ and quantifies the excess risk of ignoring the pattern.

Theorem D.3 (Missingness is evidence about the cluster)**.** *Let Assumption 2.2(i) and Condition (P) hold. For an admissible pattern $\boldsymbol{r}$ with visible features $\boldsymbol{x}_{\boldsymbol{r}} = (\boldsymbol{x}^{(v)})_{v:r_v=1}$, let $p_k(\boldsymbol{x}_{\boldsymbol{r}})$ be the density of the visible views given $Y=k$ and let $\check{P}(Y=k \mid \boldsymbol{x}_{\boldsymbol{r}}) \propto P(Y=k)\, p_k(\boldsymbol{x}_{\boldsymbol{r}})$ be the* pattern-agnostic *posterior used by existing methods. Then:*

(i) Factorization: $P\big(Y=k \mid X_{\boldsymbol{r}}=\boldsymbol{x}_{\boldsymbol{r}}, \mathbf{R}=\boldsymbol{r}\big) = \dfrac{\check{P}(Y=k \mid \boldsymbol{x}_{\boldsymbol{r}})\, \varpi_k(\boldsymbol{r})}{\sum_{l=1}^{K} \check{P}(Y=l \mid \boldsymbol{x}_{\boldsymbol{r}})\, \varpi_l(\boldsymbol{r})}$.

(ii) Evidence as log-odds: $\log \frac{P(k\mid \boldsymbol{x}_{\boldsymbol{r}},\boldsymbol{r})}{P(l\mid \boldsymbol{x}_{\boldsymbol{r}},\boldsymbol{r})} = \log \frac{\check{P}(k\mid \boldsymbol{x}_{\boldsymbol{r}})}{\check{P}(l\mid \boldsymbol{x}_{\boldsymbol{r}})} + \log \frac{\varpi_k(\boldsymbol{r})}{\varpi_l(\boldsymbol{r})}$, *and under Assumption 2.2(ii) the pattern term equals $\sum_m \big[r_m \log \frac{\pi_m(k)}{\pi_m(l)} + (1-r_m)\log \frac{1-\pi_m(k)}{1-\pi_m(l)}\big]$, the population counterpart of $\Lambda_{il} - \Lambda_{ik}$.*

(iii) Expected evidence: $\mathbb{E}\big[\log \frac{\varpi_Y(\mathbf{R})}{\varpi_l(\mathbf{R})} \,\big|\, Y=k\big] = D_{\mathrm{KL}}(\varpi_k \| \varpi_l) \geq 0$ *with equality if and only if $\varpi_k = \varpi_l$; under Assumption 2.2(ii) this equals $\sum_m \mathrm{kl}\big(\pi_m(k)\|\pi_m(l)\big)$, where* kl *is the binary divergence. Moreover $I(Y;\mathbf{R}) = \sum_k P(Y=k)\, D_{\mathrm{KL}}(\varpi_k \| \bar{\varpi})$ with $\bar{\varpi} = \sum_l P(Y=l)\varpi_l$.*

(iv) Bayes suboptimality of ignoring the pattern: *the rule $\check{h}(\boldsymbol{x}_{\boldsymbol{r}}) = \arg\max_k \check{P}(k \mid \boldsymbol{x}_{\boldsymbol{r}})$ is Bayes optimal for every feature law compatible with Assumption 2.2(i) if and only if the missingness is not informative; otherwise there exist such feature laws under which its excess 0-1 risk is strictly positive.*

Hard assignments and shrinkage. The table Eq. (4) is computed from hard rather than soft assignments on purpose: Student's t assignments are not calibrated posteriors, and their heavy tails would spread every sample over all clusters and flatten the table toward the marginal, underestimating the selection. The resulting estimate is a shrunken Beta-posterior mode:

Lemma D.4 (Beta shrinkage)**.** *(i) $\check{\pi}_m(k)$ is the unique maximizer over $[\underline{\pi}, 1-\underline{\pi}]$ of the log-posterior $\sum_{i:a_i=k}\big[r_{im}\log\pi + (1-r_{im})\log(1-\pi)\big] + \log \mathrm{Beta}\big(\pi; 1+\alpha\hat{\bar{\pi}}_m, 1+\alpha(1-\hat{\bar{\pi}}_m)\big)$. (ii) Equivalently, $\hat{\pi}_m(k) = \lambda_k \frac{n_{km}}{n_k} + (1-\lambda_k)\hat{\bar{\pi}}_m$ with $\lambda_k = \frac{n_k}{n_k+\alpha}$, a shrinkage toward the marginal visibility that stabilizes small clusters. (iii) If the within-cluster frequencies satisfy $\max_k |n_{km}/n_k - \hat{\bar{\pi}}_m| \to 0$, as under MCAR with a feature-driven partition, then $\max_k |\hat{\pi}_m(k) - \hat{\bar{\pi}}_m| \to 0$.*

The evidence-aware assignment is a MAP rule. The scaling in Eq. (6) is not heuristic:

Proposition D.5 (Evidence-aware assignment is a MAP rule)**.** *Consider the working model in which, given $Y=k$, (W1) the completed blocks $\tilde{\boldsymbol{z}}^{(m)}$ are independent $\mathcal{N}(\boldsymbol{\mu}_k^{(m)}, \sigma^2 \boldsymbol{I}_d)$ and the likelihood of block m is tempered by ω_{im}; (W2) $\mathbf{R} \sim \bigotimes_m \mathrm{Bernoulli}\big(\check{\pi}_m(k)\big)$; and (W3) the clusters are a priori uniform. Then $\arg\max_k P_\omega(Y=k \mid \tilde{\boldsymbol{z}}_i, \boldsymbol{r}_i) = \arg\min_k \tilde{\Delta}_{ik}$ with $\lambda_R = 1$; for general $\lambda_R \geq 0$ the same holds with the Bernoulli likelihood in (W2) tempered by λ_R.*

Corollary D.6 (Safety and boundedness of the evidence)**.** *(i) If $\check{\pi}_m(k)$ does not depend on k for every m, then $\arg\min_k \tilde{\Delta}_{ik} = \arg\min_k \Delta_{ik}$ for every i; in particular the evidence cannot alter any assignment under MCAR once the table has reached the marginal visibility (Lemma D.4(iii)). (ii) For all i, k, l,*

$$\Big|\big(\tilde{\Delta}_{ik} - \tilde{\Delta}_{il}\big) - \big(\Delta_{ik} - \Delta_{il}\big)\Big| \;\leq\; \frac{2\lambda_R \sigma^2 M}{\sum_m \omega_{im}} \log \frac{1-\underline{\pi}}{\underline{\pi}} \;\leq\; 2\lambda_R \sigma^2 M \log \frac{1-\underline{\pi}}{\underline{\pi}}, \tag{24}$$

so the evidence can overturn a feature-based preference only when the feature margin is below this bound.

Corollary D.6 is the formal reason why the evidence cannot collapse clusters onto missing patterns: its influence is capped by the floor $\underline{\pi}$ and scaled by σ^2. We fix $\sigma^2 = \sigma_0^2$, the per-dimension within-cluster variance measured on the pretrained embeddings in round 0: self-training on pseudo-labels tightens clusters artificially, and a recomputed σ^2 would switch the evidence off after a few rounds.

Why the evidence is cross-fitted. The cross-fitted evidence Eq. (7) is motivated by the following lemma.

Lemma D.7 (Self-confirmation of in-sample evidence)**.** *Fix a sample i with $a_i = k$ and let $\Lambda_{i\cdot}^{(-i)}$ be the evidence computed from the table Eq. (4) without sample i (with $\hat{\pi}_m$ held fixed). Then $\Lambda_{il} = \Lambda_{il}^{(-i)}$ for every $l \neq k$, whereas*

$$0 \;\le\; \Lambda_{ik}^{(-i)} - \Lambda_{ik} \;\le\; \frac{M}{(n_k + \alpha)\,\underline{\pi}}. \tag{25}$$

Hence in-sample evidence systematically favors the current assignment, a self-reinforcing loop that can drift clusters toward missing patterns; Eq. (7) removes this first-order term because sample i never enters the table used for its own evidence.

D.3 Module II: Counterfactual Imputation

Circularity of cluster-level imputation. For the prototype-based filling $\hat{\boldsymbol{z}}_i^{\mathrm{proto},(m)} = \sum_k \bar{q}_{ik} \boldsymbol{\mu}_k^{(m)}$ of Section 3.3:

Lemma D.8 (Cluster-level imputation is self-confirming)**.** *(i) $\hat{\boldsymbol{z}}_i^{\mathrm{proto},(m)}$ is a deterministic function of $(\bar{\boldsymbol{q}}_i, \{\boldsymbol{\mu}_k^{(m)}\}_k)$, hence $I\big(\hat{\boldsymbol{z}}_i^{\mathrm{proto},(m)}; Y \,\big|\, \bar{\boldsymbol{q}}_i, \{\boldsymbol{\mu}_k^{(m)}\}_k\big) = 0$: it carries no information about the cluster beyond the current assignment. (ii) If $\bar{\boldsymbol{q}}_i = \boldsymbol{e}_{a_i}$, then for every $k \neq a_i$ the imputed block adds $\omega_{im}\|\boldsymbol{\mu}_{a_i}^{(m)} - \boldsymbol{\mu}_k^{(m)}\|_2^2 / \sum_v \omega_{iv} \ge 0$ to $\Delta_{ik} - \Delta_{ia_i}$, with strict inequality whenever $\boldsymbol{\mu}_k^{(m)} \neq \boldsymbol{\mu}_{a_i}^{(m)}$: the filled view never argues against the current assignment and makes errors harder to correct.*

Full statement of Theorem 3.2. Part (i) needs only Assumption 2.3(i):

Theorem D.9 (Mechanism invariance and counterfactual identifiability)**.** *Let Assumptions 2.2 and 2.3(i) hold.*

- *(i)* Invariance: $P(Z^{(m)} \mid C, \mathbf{R}) = P(Z^{(m)} \mid C)$ *and* $\mathbb{E}[Z^{(m)} \mid C, \mathbf{R}] = f_m(C)$.
- *(ii)* Recoverability: *under Assumption 2.3(ii), for every $v \neq m$, $f_m(\boldsymbol{c}) = \mathbb{E}\big[Z^{(m)} \,\big|\, h_v^\star(Z^{(v)}) = \boldsymbol{c},\, R^{(m)} = R^{(v)} = 1\big]$ for P_C-almost every $\boldsymbol{c}$; thus f_m is a functional of the observed-data law, identified from co-observed pairs.*
- *(iii)* Counterfactual identifiability: *under Assumption 2.3(ii), for evidence $\boldsymbol{e} = \big(\mathbf{R} = \boldsymbol{r},\; Z^{(v)} = \boldsymbol{z}^{(v)}\;\forall v \in \mathcal{V}\big)$ with $r_m = 0$ and $\mathcal{V} = \{v : r_v = 1\}$, the counterfactual observed embedding $Z^{(m)*}_{\mathrm{do}(R^{(m)}=1)}$ has conditional law $f_m(\boldsymbol{c}^\star) + \boldsymbol{\xi}^{(m)}$, where $\boldsymbol{c}^\star = \mathrm{nrm}\big(\sum_{v\in\mathcal{V}} h_v^\star(\boldsymbol{z}^{(v)})\big)$ and $\boldsymbol{\xi}^{(m)} \sim P_{\boldsymbol{\xi}^{(m)}}$ is independent of $\boldsymbol{e}$. In particular $\mathbb{E}\big[Z^{(m)*}_{\mathrm{do}(R^{(m)}=1)} \mid \boldsymbol{e}\big] = f_m(\boldsymbol{c}^\star)$, and Eq. (9) with $(h_v, f_m) = (h_v^\star, f_m)$ is the minimum mean-squared-error counterfactual imputation.*

Transport gap of associational imputers. Associational imputers, which regress a target view directly on source views, do not share the invariance of Theorem D.9 whenever the sources do not pin down the content:

Proposition D.10 (Transport gap of associational imputers)**.** *Let Assumptions 2.2(i) and 2.3(i) hold, let f_m be L_f-Lipschitz, and let $\mathcal{A} \subseteq [M] \setminus \{m\}$ be a nonempty set of source views. For $r \in \{0,1\}$ let $g_r^{\mathrm{as}}(\boldsymbol{z}_\mathcal{A}) = \mathbb{E}\big[Z^{(m)} \mid Z^{(\mathcal{A})} = \boldsymbol{z}_\mathcal{A}, R^{(m)} = r\big]$ be the associational imputer of population r, and $P_{C|\boldsymbol{z}_\mathcal{A},r}$ the corresponding law of C.*

(i) For every abduction map $\hat{\boldsymbol{c}}(\cdot)$,

$$\begin{aligned}\big\|g_1^{\mathrm{as}}(\boldsymbol{z}_{\mathcal{A}}) - g_0^{\mathrm{as}}(\boldsymbol{z}_{\mathcal{A}})\big\| &\le L_f\, W_1\big(P_{C|\boldsymbol{z}_{\mathcal{A}},1}, P_{C|\boldsymbol{z}_{\mathcal{A}},0}\big)\\ &\le L_f \sum_{r\in\{0,1\}} \mathbb{E}\big[\|C - \hat{\boldsymbol{c}}(\boldsymbol{z}_{\mathcal{A}})\| \,\big|\, Z^{(\mathcal{A})} = \boldsymbol{z}_{\mathcal{A}}, R^{(m)} = r\big].\end{aligned} \tag{26}$$

(ii) The gap vanishes when C is $Z^{(\mathcal{A})}$-measurable, but not in general: in the two-cluster Gaussian model of Example E.6, $g_1^{\mathrm{as}}(z) - g_0^{\mathrm{as}}(z) = \mu\big[\tanh\big(\frac{\mu z}{\vartheta^2} + \frac{1}{2}\log\frac{\pi_m(1)}{\pi_m(2)}\big) - \tanh\big(\frac{\mu z}{\vartheta^2} + \frac{1}{2}\log\frac{1-\pi_m(1)}{1-\pi_m(2)}\big)\big]$, which is nonzero for every z whenever $\pi_m(1) \neq \pi_m(2)$, although f_m is the same in both populations.

(iii) If the per-view content estimates satisfy $\boldsymbol{c}_i^{(v)} = C_i + \boldsymbol{\delta}_i^{(v)}$ with independent zero-mean errors and $\mathbb{E}\|\boldsymbol{\delta}_i^{(v)}\|^2 \le \nu^2$, then $\mathbb{E}\|\bar{\boldsymbol{c}}_i^{(-m)} - C_i\| \le 2\nu/\sqrt{|\mathcal{V}_i \setminus \{m\}|}$.

Proposition D.10 is the formal version of the intuition behind Eq. (8): the transport gap of any imputer is controlled by the uncertainty of abduction, and pooling all visible views on the content sphere shrinks this uncertainty at the rate $|\mathcal{V}_i \setminus \{m\}|^{-1/2}$.

View dropout. When the server trains the objective Eq. (27) on a mini-batch for a target view m, it uses the dropped mask $\tilde{r}_{iv} = r_{iv}b_{iv}$ ($v \neq m$) with $b_{iv} \sim \mathrm{Bernoulli}(1 - p_{\mathrm{drop}})$ applied to a random half of the batch, reverting to r_{iv} if all sources are dropped, and $\tilde{\boldsymbol{c}}_i^{(-m)}$ denotes Eq. (8) computed with $\tilde{r}_{iv}$. View dropout makes the training-time visibility resemble that of the samples that actually miss view m, which typically see fewer views.

The structural causal objective in full. For a mini-batch and a target view m, let $\mathcal{B}_m = \{i : r_{im} = 1, \sum_{v\neq m}\tilde{r}_{iv} \ge 1\}$, and let $\hat{w}_{im}^{\mathrm{tr}}$ and $\hat{w}_{im}^{\mathrm{ipw}}$ be the weights of Module I renormalized to sum to one over $\mathcal{B}_m$. The server minimizes

$$\mathcal{L}_{\mathrm{SCM}} = \sum_{m=1}^{M}\Big[\mathcal{L}_{\mathrm{cf}}^{(m)} + \lambda_{\mathrm{inv}}\mathcal{L}_{\mathrm{inv}}^{(m)} + \lambda_{\mathrm{ind}}\mathcal{L}_{\mathrm{ind}}^{(m)}\Big], \tag{27}$$

whose three terms are

$$\mathcal{L}_{\mathrm{cf}}^{(m)} = \sum_{i\in\mathcal{B}_m} \hat{w}_{im}^{\mathrm{tr}}\,\big\|f_m(\tilde{\boldsymbol{c}}_i^{(-m)}) - \boldsymbol{z}_i^{(m)}\big\|_2^2, \tag{28}$$

$$\mathcal{L}_{\mathrm{inv}}^{(m)} = -\sum_{i\in\mathcal{B}_m} \hat{w}_{im}^{\mathrm{ipw}} \log\frac{\exp\big(\langle \boldsymbol{c}_i^{(m)}, \tilde{\boldsymbol{c}}_i^{(-m)}\rangle/\tau\big)}{\sum_{j\in\mathcal{B}_m}\exp\big(\langle \boldsymbol{c}_i^{(m)}, \tilde{\boldsymbol{c}}_j^{(-m)}\rangle/\tau\big)}, \tag{29}$$

$$\mathcal{L}_{\mathrm{ind}}^{(m)} = \mathrm{nHSIC}\Big(\big\{\tilde{\boldsymbol{c}}_i^{(-m)}\big\}_{i\in\mathcal{B}_m}, \big\{\boldsymbol{z}_i^{(m)} - f_m(\tilde{\boldsymbol{c}}_i^{(-m)})\big\}_{i\in\mathcal{B}_m}\Big). \tag{30}$$

The counterfactual loss Eq. (28) is transport weighted because the imputer serves the missing population; the InfoNCE term Eq. (29) (van den Oord et al., 2018) aligns per-view content with the pooled content of the other views and is inverse-probability weighted because content should be shared by the whole population; and the nHSIC term Eq. (30) (Definition C.8) enforces the independence between exogenous noise and cause that defines the additive noise model. All samples with at least two visible views are used, not only complete ones.

Consistency of the structural causal objective. The weights and regularizers in Eq. (28) to Eq. (30) target the right population and introduce no bias at the truth:

Proposition D.11 (The structural causal objective is consistent)**.** *Let Assumptions 2.2 and 2.3 and Condition (P) hold. (i) For any fixed (h, f), the population transport-weighted risk on visible entries equals the missing-population risk,*

$$\begin{aligned}&\frac{\bar{\pi}_m}{1-\bar{\pi}_m}\,\mathbb{E}\Big[\frac{1-\pi_m(Y)}{\pi_m(Y)}\big\|Z^{(m)} - f_m(\bar{\boldsymbol{c}}^{(-m)})\big\|^2 \,\Big|\, R^{(m)} = 1\Big]\\ &\qquad = \mathbb{E}\Big[\big\|Z^{(m)} - f_m(\bar{\boldsymbol{c}}^{(-m)})\big\|^2 \,\Big|\, R^{(m)} = 0\Big].\end{aligned} \tag{31}$$

(ii) The oracle pair $(h^\star, f)$ minimizes the population versions of Eq. (28) and Eq. (30) simultaneously, attains the Bayes risk $\mathbb{E}\|\boldsymbol{\xi}^{(m)}\|^2$, and achieves perfect alignment $\langle \boldsymbol{c}^{(m)}, \bar{\boldsymbol{c}}^{(-m)} \rangle = 1$ in Eq. (29); the regularizers therefore introduce no bias at the truth.

Cross-fitting. We partition $[N]$ at random into folds $\mathcal{I}_1, \dots, \mathcal{I}_L$, fit $(h^{[-\ell]}, f^{[-\ell]})$ by minimizing $\mathcal{L}_{\mathrm{SCM}}$ on $[N] \setminus \mathcal{I}_\ell$, and predict every entry, visible or missing, of the held-out fold:

$$\hat{\boldsymbol{z}}_i^{(m)} = f_m^{[-\ell]}\Big(\operatorname{nrm}\Big(\sum_{v \neq m} r_{iv} \operatorname{nrm}\big(h_v^{[-\ell]}(\boldsymbol{z}_i^{(v)})\big)\Big)\Big), \qquad i \in \mathcal{I}_\ell, \;\; \textstyle\sum_{v \neq m} r_{iv} \geq 1. \tag{32}$$

Cross-fitting removes optimism. The cross-fitted predictions Eq. (32) make visible-entry errors honest estimates of out-of-sample errors:

Lemma D.12 (Cross-fitting removes optimism)**.** *Condition on the client encoders, let (i, m) be a visible entry of a sample with another visible view, and let $\boldsymbol{z}_i'^{(m)} = f_m(C_i) + \boldsymbol{\xi}_i'^{(m)}$ with $\boldsymbol{\xi}_i'^{(m)}$ an independent copy of $\boldsymbol{\xi}_i^{(m)}$. Then*

$$\mathbb{E}\big\|\boldsymbol{z}_i'^{(m)} - \hat{\boldsymbol{z}}_i^{(m)}\big\|^2 - \mathbb{E}\big\|\boldsymbol{z}_i^{(m)} - \hat{\boldsymbol{z}}_i^{(m)}\big\|^2 = 2\,\mathbb{E}\big\langle \boldsymbol{\xi}_i^{(m)}, \hat{\boldsymbol{z}}_i^{(m)} \big\rangle. \tag{33}$$

For the cross-fitted predictions Eq. (32) the right-hand side vanishes, so visible-entry errors are unbiased for out-of-sample errors; for an in-sample interpolating fit it equals $2\mathbb{E}\|\boldsymbol{\xi}^{(m)}\|^2$.

D.4 Module III: Reliability and Doubly Robust Fusion

Full statements of Theorems 3.3 and 3.4. The population target of the transported reliability Eq. (10) is the R^2 of the imputer on the *missing* population,

$$\rho_m = 1 - \frac{\mathbb{E}\big[\|Z^{(m)} - \hat{g}_m\|^2 \mid R^{(m)} = 0\big]}{\mathbb{E}\big[\|Z^{(m)} - \mathbb{E}[Z^{(m)} \mid R^{(m)} = 0]\|^2 \mid R^{(m)} = 0\big]}. \tag{34}$$

Theorem D.13 (Transported reliability)**.** *Let Assumption 2.2 and Condition (P) hold, let the plug-in propensity be consistent, $\max_i |\hat{\pi}_{im} - \pi_m(y_i)| \to 0$ in probability, and let the cross-fitted imputers converge to a fixed $\hat{g}_m$. Then:*

(i) *$\hat{\rho}_m \to \mathrm{clip}_{[0,1]}(\rho_m)$ in probability;*

(ii) *the unweighted visible-entry error $\mathrm{MSE}_{\mathrm{vis}} = \mathbb{E}\big[\|Z^{(m)} - \hat{g}_m\|^2 \mid R^{(m)} = 1\big]$ and the missing-population error $\mathrm{MSE}_{\mathrm{mis}} = \mathbb{E}\big[\|Z^{(m)} - \hat{g}_m\|^2 \mid R^{(m)} = 0\big]$ satisfy*

$$\mathrm{MSE}_{\mathrm{vis}} - \mathrm{MSE}_{\mathrm{mis}} = \frac{\mathrm{Cov}\big(e_m(Y), \pi_m(Y)\big)}{\bar{\pi}_m(1 - \bar{\pi}_m)} \tag{35}$$

with $e_m(k) = \mathbb{E}\big[\|Z^{(m)} - \hat{g}_m\|^2 \mid Y = k\big]$, so that an unweighted evaluation is optimistic whenever the clusters that rarely show view m are also the hardest to impute;

(iii) *under MCAR the transport weights are constant and both evaluations coincide; the all-or-nothing switch of prior work corresponds to restricting $\hat{\rho}_m$ to $\{0, 1\}$.*

Theorem D.14 (Bias of reliability-scaled doubly robust prototypes)**.** *Fix k and m, assume oracle memberships ($a_i = y_i$), and let the cross-fitted imputers converge in $L^2(P_k)$ to $\hat{g}_m$. Write $\pi^\star = \pi_m^\star(Y, C)$, $\bar{\pi}_k^\star = \mathbb{E}_k[\pi^\star]$, $g^\star = f_m(C)$ and $\boldsymbol{\mu}_{km} = \mathbb{E}_k[Z^{(m)}]$. Under Assumptions 2.2 and 2.3(i), and for every fixed $\rho \in [0, 1]$ in place of $\hat{\rho}_m$:*

(i) Exact decomposition:

$$\operatorname{plim} \hat{\boldsymbol{\mu}}_{km}^{\mathrm{dr}}(\rho) = \boldsymbol{\mu}_{km} + (1 - \rho)\, B_{km}^{\mathrm{sel}} + \rho\, \frac{\mathrm{Cov}_k\big(\pi^\star,\, g^\star - \hat{g}_m\big)}{\bar{\pi}_k^\star}, \tag{36}$$

where $B_{km}^{\mathrm{sel}} = \mathrm{Cov}_k(\pi^\star, g^\star)/\bar{\pi}_k^\star = \mathbb{E}[Z^{(m)} \mid Y = k, R^{(m)} = 1] - \boldsymbol{\mu}_{km}$ is the within-cluster selection bias of the visible mean.

(ii) Double robustness: *the limit equals $\boldsymbol{\mu}_{km}$ if Condition (P) holds, for every ρ, every imputer and every cluster-level $\hat{\pi}$; or if* ***(O)*** *$\hat{g}_m = g^\star$ and $\rho = 1$, even when (P) fails.*

(iii) Product-form bound: *with* $\mathrm{CV}_{km} = \mathrm{sd}_k(\pi^\star)/\bar{\pi}_k^\star$ *and* $\mathrm{sd}_k(g^\star) = (\mathbb{E}_k\|g^\star - \boldsymbol{\mu}_{km}\|^2)^{1/2}$,

$$\left\| \operatorname{plim} \hat{\boldsymbol{\mu}}_{km}^{\mathrm{dr}}(\rho) - \boldsymbol{\mu}_{km} \right\| \;\le\; \mathrm{CV}_{km}\Big[(1-\rho)\,\mathrm{sd}_k(g^\star) + \rho\,\big\|\hat{g}_m - g^\star\big\|_{L^2(P_k)}\Big]. \tag{37}$$

Remark D.15 (Division of labor between weights and the doubly robust correction). Because the working propensity Eq. (4) is cluster level, $r_{im}/\hat{\pi}_{im}$ is constant within each cluster and cancels in Eq. (12) and Eq. (13): $\hat{\boldsymbol{\mu}}_{km}^{\mathrm{ipw}}$ is the mean of the visible members of cluster k, and the correction in Eq. (13) is the gap between the average imputation over the whole cluster and over its visible members. The doubly robust correction therefore acts precisely on the *within-cluster* selection that a cluster-level propensity cannot see, whereas *across-cluster* selection is handled by the weights in Eq. (3), Eq. (28), Eq. (29) and Eq. (10). The general form Eq. (12) to Eq. (13) remains valid for finer propensity models.

Theorem D.16 (Asymptotic normality and variance reduction)**.** *Under the conditions of Theorem D.14 with Condition (P), let $\pi_k = \pi_m(k)$ and $n_k \to \infty$. For every fixed $\rho \in [0,1]$ and unit vector $\boldsymbol{u} \in \mathbb{R}^d$,*

$$\sqrt{n_k}\,\boldsymbol{u}^\top\big(\hat{\boldsymbol{\mu}}_{km}^{\mathrm{dr}}(\rho) - \boldsymbol{\mu}_{km}\big) \Rightarrow \mathcal{N}\big(0, V_k(\rho;\boldsymbol{u})\big), \tag{38}$$

$$V_k(\rho;\boldsymbol{u}) = \frac{\mathrm{Var}_k(\boldsymbol{u}^\top Z^{(m)})}{\pi_k} - \frac{1-\pi_k}{\pi_k}\Big[2\rho\,\mathrm{Cov}_k\big(\boldsymbol{u}^\top\hat{g}_m, \boldsymbol{u}^\top Z^{(m)}\big) - \rho^2\,\mathrm{Var}_k\big(\boldsymbol{u}^\top\hat{g}_m\big)\Big]. \tag{39}$$

Consequently:

(i) the variance-optimal coefficient and its gain over the visible mean ($\rho = 0$) are

$$\rho^{\mathrm{opt}} = \frac{\mathrm{Cov}_k(\boldsymbol{u}^\top\hat{g}_m, \boldsymbol{u}^\top Z^{(m)})}{\mathrm{Var}_k(\boldsymbol{u}^\top\hat{g}_m)},$$

$$V_k(0;\boldsymbol{u}) - V_k(\rho^{\mathrm{opt}};\boldsymbol{u}) = \frac{1-\pi_k}{\pi_k}\,\mathrm{Var}_k(\boldsymbol{u}^\top Z^{(m)})\,\mathrm{Corr}_k^2\big(\boldsymbol{u}^\top\hat{g}_m, \boldsymbol{u}^\top Z^{(m)}\big);$$

(ii) if $\hat{g}_m$ is calibrated, $\mathrm{Cov}_k(\hat{g}_m, Z^{(m)} - \hat{g}_m) = \mathbf{0}$ (e.g., $\hat{g}_m = g^\star$), then $\rho^{\mathrm{opt}} = 1$ and the gain equals $\frac{1-\pi_k}{\pi_k} R_k^2\,\mathrm{Var}_k(\boldsymbol{u}^\top Z^{(m)})$ with $R_k^2 = \mathrm{Var}_k(\boldsymbol{u}^\top\hat{g}_m)/\mathrm{Var}_k(\boldsymbol{u}^\top Z^{(m)})$;

(iii) the shrinkage Eq. (14) leaves the limit law unchanged because $1 - s_{km} = O_P(1/n_k)$.

Theorems D.13 and D.16 connect: the transported reliability is a label-free estimate of the R^2 that governs how much variance the imputations can remove, and its clipped value lies in the range in which the doubly robust estimator improves on the visible mean.

Proposition D.17 (Doubly robust risk and oracle shrinkage)**.** *The prototypes of Module III solve explicit risk minimization problems:*

(i) $\hat{\boldsymbol{\mu}}_{km}^{\mathrm{dr}}$ is the unique minimizer of the doubly robust risk

$$\hat{\mathcal{L}}_{km}^{\mathrm{dr}}(\boldsymbol{\mu}) = \frac{\hat{\rho}_m}{n_k}\sum_i a_{ik}\|\hat{g}_{im} - \boldsymbol{\mu}\|^2 + \frac{\sum_i a_{ik}(r_{im}/\hat{\pi}_{im})\big(\|\boldsymbol{z}_i^{(m)} - \boldsymbol{\mu}\|^2 - \hat{\rho}_m\|\hat{g}_{im} - \boldsymbol{\mu}\|^2\big)}{\sum_i a_{ik}(r_{im}/\hat{\pi}_{im})}, \tag{40}$$

whose population counterpart equals $\mathbb{E}_k\|Z^{(m)} - \boldsymbol{\mu}\|^2$ under Condition (P) for every $\hat{\rho}_m$ and under condition (O) of Theorem D.14 with $\hat{\rho}_m = 1$;

(ii) $\boldsymbol{\mu}_k^{(m)}$ in Eq. (14) is the unique minimizer of $\hat{\mathcal{L}}_{km}^{\mathrm{dr}}(\boldsymbol{\mu}) + \frac{\hat{\rho}_m n_0}{n_{km}}\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{km}^{\mathrm{or}}\|^2$, a ridge prior centered at the outcome-regression estimate;

(iii) if $\hat{\boldsymbol{\mu}}_{km}^{\mathrm{dr}}$ is unbiased with $\mathbb{E}\|\hat{\boldsymbol{\mu}}_{km}^{\mathrm{dr}} - \boldsymbol{\mu}_{km}\|^2 = v_{km}/n_{km}$ and $\hat{\boldsymbol{\mu}}_{km}^{\mathrm{or}}$ has bias $\mathbf{b}_{km}$ and negligible variance, the mean-squared-error optimal convex weight is $s_{km}^{\mathrm{opt}} = n_{km}/(n_{km} + v_{km}/\|\mathbf{b}_{km}\|^2)$, which Eq. (14) matches when $\hat{\rho}_m n_0 = v_{km}/\|\mathbf{b}_{km}\|^2$; in all cases $s_{km} \to 1$ as $n_{km} \to \infty$.

D.5 Server Clustering

Initialization. Zero-filling missing blocks is harmful: two samples with the same pattern share identical zero blocks, and the squared distance between samples with different patterns contains a pattern term of order d times the Hamming distance of their patterns, so k-means on zero-filled data tends to recover missing patterns rather than clusters. We instead seed with k-means++ (Arthur & Vassilvitskii, 2007) on the partial distance

$$\delta(i,j) = \frac{\sum_m \omega_{im}\omega_{jm} \left\|\tilde{\boldsymbol{z}}_i^{(m)} - \tilde{\boldsymbol{z}}_j^{(m)}\right\|_2^2}{\sum_m \omega_{im}\omega_{jm}}, \tag{41}$$

which only compares blocks available for both samples (pairs without a common block receive the mean of the defined distances), with n_{init} restarts and selection by the smallest average $\min_k \Delta_{ik}$. In round 0 the server first clusters the visible blocks alone ($\omega_{im} = r_{im}$) to obtain the first table and σ_0^2, then imputes and re-seeds on the completed blocks; later rounds warm-start from the uploaded client centers.

Classification EM interpretation. The alternation of Section 3.5 decreases an explicit objective:

Proposition D.18 (Classification-EM interpretation and finite convergence)**.** *Fix $\hat{\rho}$, ω and σ^2, set $\lambda_R = 1$, use in-sample evidence, and replace Eq. (12) to Eq. (14) by ω-weighted means. Then each sweep of the alternation is an exact block-coordinate descent step on*

$$\begin{aligned} J\big(\{a_i\}, \{\boldsymbol{\mu}_k^{(m)}\}, \check{\pi}\big) = \sum_{i=1}^{N} \Big[\frac{1}{2\sigma^2} \sum_m \omega_{im} \big\|\tilde{\boldsymbol{z}}_i^{(m)} - \boldsymbol{\mu}_{a_i}^{(m)}\big\|_2^2 + \Lambda_{i a_i}(\check{\pi})\Big] \\ - \sum_{k,m} \log \mathrm{Beta}\big(\check{\pi}_m(k); 1 + \alpha\hat{\bar{\pi}}_m, 1 + \alpha(1 - \hat{\bar{\pi}}_m)\big) \end{aligned} \tag{42}$$

over $\check{\pi} \in [\underline{\pi}, 1 - \underline{\pi}]^{K \times M}$, i.e., a classification EM (Celeux & Govaert, 1992) for the working model of Proposition D.5 with a Beta prior. Hence J is non-increasing and, with ties broken in favor of the current label, the alternation terminates after finitely many sweeps at a partial optimum. With the doubly robust step, the prototype block instead minimizes the penalized risk of Proposition D.17, which targets the same population quantity.

Corollary D.19 (Graceful degeneration)**.** *(i) On complete data ($\boldsymbol{R} \equiv 1$) all weights equal one, the evidence is constant in k and is therefore dropped, no block is imputed ($\omega \equiv 1$, $\tilde{\boldsymbol{z}} = \boldsymbol{z}$), the prototypes are plain means, and* `CFIMC` *coincides with its self-supervised backbone. (ii) Under MCAR, as $\max_k |\hat{\pi}_m(k) - \hat{\bar{\pi}}_m| \to 0$, the evidence becomes uninformative (Corollary D.6), w^{ipw} and w^{tr} become uniform after normalization, and $B_{km}^{\text{sel}} = \mathbf{0}$, so only the counterfactual imputation remains active.*

E Proofs

This appendix proves all results of Sections 2 and 3 and of Appendices C to D. Appendix E.1 fixes conventions and collects auxiliary lemmas; the remaining subsections follow the order of the main text, and each of them also proves the supplementary results that Appendices C to D attach to the same part of the paper.

E.1 Conventions and Auxiliary Results

All random elements live on a common probability space, and equalities between conditional expectations hold almost surely. For a fixed view m we abbreviate

$$\mathbf{W}_m = \big(Y, C, X^{(1:M)}, \mathbf{R}^{(-m)}\big), \qquad \mathcal{E}_m = \big\{\mathbf{R}^{(-m)} \neq \mathbf{0}\big\}. \tag{43}$$

We use the semi-graphoid properties of conditional independence (Pearl, 2009): for random elements A, B, D, E,

(G1) decomposition: $A \perp\!\!\!\perp (B, E) \mid D \implies A \perp\!\!\!\perp B \mid D$;
(G2) weak union: $A \perp\!\!\!\perp (B, E) \mid D \implies A \perp\!\!\!\perp B \mid (D, E)$;
(G3) contraction: $A \perp\!\!\!\perp B \mid D$ and $A \perp\!\!\!\perp E \mid (B, D) \implies A \perp\!\!\!\perp (B, E) \mid D$;
(G4) functions: $A \perp\!\!\!\perp B \mid D \implies \phi(A, D) \perp\!\!\!\perp \psi(B, D) \mid D$ for measurable ϕ, ψ.

Lemma E.1 (Kantorovich-type bound on the effective sample size)**.** *Let $w_1, \dots, w_n \in [a, b]$ with $0 < a \le b$. Then $\big(\sum_i w_i\big)^2 \ge \frac{4ab}{(a+b)^2}\, n \sum_i w_i^2$.*

Proof. For each i, $(w_i - a)(b - w_i) \ge 0$, i.e., $w_i^2 \le (a + b)w_i - ab$. Summing over i and writing $S = \sum_i w_i$ gives $\sum_i w_i^2 \le (a + b)S - nab$. Therefore

$$\begin{aligned}(a+b)^2 S^2 - 4ab\, n \sum_i w_i^2 &\ge (a+b)^2 S^2 - 4ab\, n\big[(a+b)S - nab\big] \\ &= (a+b)^2 S^2 - 4abn(a+b)S + 4a^2 b^2 n^2 \\ &= \big[(a+b)S - 2abn\big]^2 \;\ge\; 0,\end{aligned}$$

which is the claim after division by $(a + b)^2$. ∎

Lemma E.2 (Normalization onto the sphere is 2-Lipschitz toward the sphere)**.** *For every $\boldsymbol{a} \in \mathbb{R}^{d_c}$ and every unit vector $\boldsymbol{c}$, $\|\operatorname{nrm}(\boldsymbol{a}) - \boldsymbol{c}\| \le 2\|\boldsymbol{a} - \boldsymbol{c}\|$, where $\operatorname{nrm}(\mathbf{0})$ is any unit vector.*

Proof. If $\boldsymbol{a} = \mathbf{0}$, then $\|\operatorname{nrm}(\boldsymbol{a}) - \boldsymbol{c}\| \le \|\operatorname{nrm}(\boldsymbol{a})\| + \|\boldsymbol{c}\| = 2 = 2\|\boldsymbol{a} - \boldsymbol{c}\|$. Otherwise, by the triangle inequality and the reverse triangle inequality,

$$\begin{aligned}\Big\|\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \boldsymbol{c}\Big\| &\le \Big\|\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \boldsymbol{a}\Big\| + \|\boldsymbol{a} - \boldsymbol{c}\| = \big|1 - \|\boldsymbol{a}\|\big| + \|\boldsymbol{a} - \boldsymbol{c}\| \\ &= \big|\|\boldsymbol{c}\| - \|\boldsymbol{a}\|\big| + \|\boldsymbol{a} - \boldsymbol{c}\| \le 2\|\boldsymbol{a} - \boldsymbol{c}\|.\end{aligned}$$

∎

Lemma E.3 (Admissibility truncation)**.** *Let $\tilde{\mathbf{R}}$ have components that are conditionally independent given (Y, C) with $P(\tilde{R}^{(v)} = 1 \mid Y, C) = \tilde{\pi}_v(Y, C) \in (0, 1)$, and let $\mathbf{R}$ be distributed as $\tilde{\mathbf{R}}$ conditioned on $\{\tilde{\mathbf{R}} \neq \mathbf{0}\}$. Then for every m and every $\boldsymbol{r}_{-m} \neq \mathbf{0}$,*

$$P\big(R^{(m)} = 1 \mid Y, C, \mathbf{R}^{(-m)} = \boldsymbol{r}_{-m}\big) = \tilde{\pi}_m(Y, C), \quad P\big(R^{(m)} = 1 \mid Y, C, \mathbf{R}^{(-m)} = \mathbf{0}\big) = 1. \tag{44}$$

Hence Assumption 2.2(ii) holds on $\mathcal{E}_m$, and it fails only on $\mathcal{E}_m^c = \{\mathbf{R} = \boldsymbol{e}_m\}$, whose conditional probability is $\tilde{\pi}_m \prod_{v \neq m}(1 - \tilde{\pi}_v)/\big(1 - \prod_v (1 - \tilde{\pi}_v)\big)$.

Proof. For $\boldsymbol{r} \neq \mathbf{0}$, with $Z_0(Y, C) = 1 - \prod_v (1 - \tilde{\pi}_v(Y, C))$,

$$P(\mathbf{R} = \boldsymbol{r} \mid Y, C) = \frac{1}{Z_0(Y, C)} \prod_{v=1}^{M} \tilde{\pi}_v^{r_v} (1 - \tilde{\pi}_v)^{1 - r_v}.$$

If $\boldsymbol{r}_{-m} \neq \mathbf{0}$, both $(1, \boldsymbol{r}_{-m})$ and $(0, \boldsymbol{r}_{-m})$ are admissible, and the common factors $Z_0^{-1} \prod_{v \neq m} \tilde{\pi}_v^{r_v} (1 - \tilde{\pi}_v)^{1 - r_v}$ cancel in

$$\begin{aligned}P\big(R^{(m)} = 1 \mid Y, C, \mathbf{R}^{(-m)} = \boldsymbol{r}_{-m}\big) &= \frac{P\big((1, \boldsymbol{r}_{-m}) \mid Y, C\big)}{P\big((1, \boldsymbol{r}_{-m}) \mid Y, C\big) + P\big((0, \boldsymbol{r}_{-m}) \mid Y, C\big)} \\ &= \frac{\tilde{\pi}_m}{\tilde{\pi}_m + (1 - \tilde{\pi}_m)} = \tilde{\pi}_m.\end{aligned}$$

If $\boldsymbol{r}_{-m} = \mathbf{0}$, the only admissible completion is $\boldsymbol{e}_m$, so the probability equals one. The first identity does not depend on $\boldsymbol{r}_{-m}$, which is the conditional independence $R^{(m)} \perp\!\!\!\perp \mathbf{R}^{(-m)} \mid (Y, C)$ on $\mathcal{E}_m$. The last statement is the displayed formula at $\boldsymbol{r} = \boldsymbol{e}_m$. ■

Remark E.4 (Samples that observe a single view). In Theorems D.13, D.14 and D.16 we work on $\mathcal{E}_m$, where Assumption 2.2(ii) holds by Lemma E.3, and all propensities are understood conditionally on $\mathcal{E}_m$, e.g., $\pi^\star_m(Y, C) = P(R^{(m)} = 1 \mid Y, C, \mathcal{E}_m)$. Every missing entry of view m lies in $\mathcal{E}_m$, so the reliability target is unaffected. For the prototypes, samples in $\mathcal{E}_m^c$ enter $\hat{\boldsymbol{\mu}}^{\mathrm{or}}_{km}$ in Eq. (12) with $\hat{g}_{im} = \boldsymbol{z}_i^{(m)}$. Let $q_k = P(\mathcal{E}_m^c \mid Y = k)$, $\pi_{k,\mathcal{E}} = P(R^{(m)} = 1 \mid Y = k, \mathcal{E}_m)$ and $\pi_{k,\mathrm{tot}} = q_k + (1 - q_k)\pi_{k,\mathcal{E}}$. Repeating the computation of Appendix E.6 on the whole cluster gives: under (O) with $\rho = 1$, the whole-cluster estimator is still consistent for $\boldsymbol{\mu}_{km}$, because $\mathbb{E}[R^{(m)}(Z^{(m)} - g^\star) \mid Y = k, \mathcal{E}_m] = \mathbf{0}$ and $(1-q_k)\mathbb{E}[g^\star \mid Y = k, \mathcal{E}_m] + q_k\mathbb{E}[Z^{(m)} \mid Y = k, \mathcal{E}_m^c] = \boldsymbol{\mu}_{km}$; under (P), it acquires the additional term

$$\rho\, \frac{q_k(1 - q_k)(1 - \pi_{k,\mathcal{E}})}{\pi_{k,\mathrm{tot}}} \Big(\mathbb{E}\big[\hat{g}_m \mid Y = k, \mathcal{E}_m\big] - \boldsymbol{\mu}_{km} \Big), \tag{45}$$

which is of order q_k and vanishes for imputers that are unbiased on average within the cluster.

E.2 Proofs for Section 2 and Appendix C

Proof of Lemma C.4. Write $\mathbf{S} = (S^{(1)}, \dots, S^{(M)}) = (U_{S^{(1)}}, \dots, U_{S^{(M)}})$ and $\mathbf{U}_0 = (U_Y, U_C, U_R)$. By Eq. (2), $Y = F_Y(U_Y)$, $C = F_C(F_Y(U_Y), U_C)$ and $\mathbf{R} = F_R(Y, C, U_R)$ are measurable functions of $\mathbf{U}_0$, while $X^{(m)} = g_m(C, S^{(m)})$.

Step 1 (independence of the private factors). Since the exogenous variables are mutually independent, $\mathbf{S} \perp\!\!\!\perp \mathbf{U}_0$, hence by (G4)

$$\mathbf{S} \perp\!\!\!\perp (Y, C, \mathbf{R}), \qquad \text{and more precisely} \qquad S^{(m)} \perp\!\!\!\perp \big(Y, C, \mathbf{R}, \mathbf{S}^{(-m)}\big) \ \forall m. \tag{46}$$

Step 2 (part (i), first statement). By (G2) applied to Eq. (46), $\mathbf{S} \perp\!\!\!\perp \mathbf{R} \mid (Y, C)$. Since $X^{(1:M)}$ is a measurable function of $(C, \mathbf{S})$, (G4) yields $(X^{(1:M)}, \mathbf{S}) \perp\!\!\!\perp \mathbf{R} \mid (Y, C)$. Explicitly, for bounded measurable ϕ, ψ and $P_{\mathbf{S}}$ the law of $\mathbf{S}$,

$$\begin{aligned}
\mathbb{E}\big[\phi(X^{(1:M)}, \mathbf{S})\, \psi(\mathbf{R}) \,\big|\, Y, C\big] &= \mathbb{E}\Big[\int \phi\big((g_m(C, \boldsymbol{s}^{(m)}))_m, \boldsymbol{s}\big)\, dP_{\mathbf{S}}(\boldsymbol{s})\ \psi(\mathbf{R}) \,\Big|\, Y, C\Big] \\
&= \underbrace{\int \phi\big((g_m(C, \boldsymbol{s}^{(m)}))_m, \boldsymbol{s}\big)\, dP_{\mathbf{S}}(\boldsymbol{s})}_{= \mathbb{E}[\phi(X^{(1:M)}, \mathbf{S}) \mid Y, C]}\ \mathbb{E}\big[\psi(\mathbf{R}) \mid Y, C\big].
\end{aligned}$$

Step 3 (part (i), second statement). Let $\Phi(C) = \int \phi\big((g_m(C, \boldsymbol{s}^{(m)}))_m\big) dP_{\mathbf{S}}(\boldsymbol{s})$, which equals both $\mathbb{E}[\phi(X^{(1:M)}) \mid Y, C]$ and $\mathbb{E}[\phi(X^{(1:M)}) \mid C]$ by Step 2. By the tower property,

$$\begin{aligned}
\mathbb{E}\big[\phi(X^{(1:M)})\psi(\mathbf{R}) \mid C\big] &= \mathbb{E}\Big[\mathbb{E}\big[\phi(X^{(1:M)}) \mid Y, C\big]\, \mathbb{E}\big[\psi(\mathbf{R}) \mid Y, C\big] \,\Big|\, C\Big] \\
&= \Phi(C)\, \mathbb{E}\Big[\mathbb{E}\big[\psi(\mathbf{R}) \mid Y, C\big] \,\Big|\, C\Big] = \mathbb{E}\big[\phi(X^{(1:M)}) \mid C\big]\, \mathbb{E}\big[\psi(\mathbf{R}) \mid C\big],
\end{aligned}$$

i.e., $\mathbf{R} \perp\!\!\!\perp X^{(1:M)} \mid C$.

Step 4 (part (ii)). The tuple $(Y, \mathbf{R}, X^{(-m)})$ is a function of $(\mathbf{U}_0, \mathbf{S}^{(-m)})$, and C is a function of $\mathbf{U}_0$. By Eq. (46), $S^{(m)}$ is independent of $(Y, C, \mathbf{R}, X^{(-m)})$. Hence, for bounded measurable ϕ, ψ and $\Phi_m(\boldsymbol{c}) = \int \phi(g_m(\boldsymbol{c}, \boldsymbol{s}))\, dP_{S^{(m)}}(\boldsymbol{s})$,

$$\begin{aligned}
&\mathbb{E}\big[\phi(X^{(m)})\, \psi(Y, \mathbf{R}, X^{(-m)}) \mid C\big] \\
&\quad = \mathbb{E}\Big[\mathbb{E}\big[\phi(g_m(C, S^{(m)})) \mid C, Y, \mathbf{R}, X^{(-m)}\big]\, \psi(Y, \mathbf{R}, X^{(-m)}) \,\Big|\, C\Big] \\
&\quad = \Phi_m(C)\, \mathbb{E}\big[\psi(Y, \mathbf{R}, X^{(-m)}) \mid C\big] = \mathbb{E}\big[\phi(X^{(m)}) \mid C\big]\, \mathbb{E}\big[\psi(Y, \mathbf{R}, X^{(-m)}) \mid C\big].
\end{aligned}$$

Step 5 (part (iii)). By (G1) applied to part (ii), $X^{(m)} \perp\!\!\!\perp \mathbf{R} \mid C$, which is $P(X^{(m)} \mid C, \mathbf{R}) = P(X^{(m)} \mid C)$. ■

We next isolate the key property behind Lemma C.7.

Lemma E.5 (Propensity given the full data)**.** *Under Assumption 2.2(i) and (ii) and Condition (P),* $\mathbb{E}\big[R^{(m)} \mid \mathbf{W}_m\big] = \pi_m(Y)$ *and* $R^{(m)} \perp\!\!\!\perp \big(C, X^{(1:M)}, \mathbf{R}^{(-m)}\big) \mid Y$.

Proof. Step 1. Lemma C.4(i) gives $\mathbf{R} \perp\!\!\!\perp X^{(1:M)} \mid (Y, C)$ and Condition (P) gives $\mathbf{R} \perp\!\!\!\perp C \mid Y$; by (G3),

$$\mathbf{R} \perp\!\!\!\perp \big(C, X^{(1:M)}\big) \mid Y. \tag{47}$$

Step 2. For patterns $r \in \{0,1\}$ and $\boldsymbol{r}' \in \{0,1\}^{M-1}$, Assumption 2.2(ii) and then Condition (P) (applied to the marginals of $\mathbf{R}$) give

$$\begin{aligned} P\big(R^{(m)} = r, \mathbf{R}^{(-m)} = \boldsymbol{r}' \mid Y\big) &= \mathbb{E}\Big[P\big(R^{(m)} = r \mid Y, C\big)\, P\big(\mathbf{R}^{(-m)} = \boldsymbol{r}' \mid Y, C\big) \,\Big|\, Y\Big] \\ &= \mathbb{E}\Big[P\big(R^{(m)} = r \mid Y\big)\, P\big(\mathbf{R}^{(-m)} = \boldsymbol{r}' \mid Y\big) \,\Big|\, Y\Big] \\ &= P\big(R^{(m)} = r \mid Y\big)\, P\big(\mathbf{R}^{(-m)} = \boldsymbol{r}' \mid Y\big), \end{aligned}$$

so $R^{(m)} \perp\!\!\!\perp \mathbf{R}^{(-m)} \mid Y$. *Step 3.* By Eq. (47) and (G2), $R^{(m)} \perp\!\!\!\perp (C, X^{(1:M)}) \mid (Y, \mathbf{R}^{(-m)})$. Together with Step 2 and (G3), $R^{(m)} \perp\!\!\!\perp (\mathbf{R}^{(-m)}, C, X^{(1:M)}) \mid Y$. Hence $\mathbb{E}[R^{(m)} \mid \mathbf{W}_m] = \mathbb{E}[R^{(m)} \mid Y] = \pi_m(Y)$. ∎

Proof of Lemma C.7. Let $\phi = \phi(\mathbf{W}_m)$ be integrable; Assumption 2.2(iii) guarantees $\pi_m(Y) \geq \eta > 0$, so all ratios below are integrable.

(b) Reweighting. By the tower property and Lemma E.5,

$$\mathbb{E}\Big[\frac{R^{(m)}}{\pi_m(Y)}\,\phi\Big] = \mathbb{E}\Big[\frac{\mathbb{E}[R^{(m)} \mid \mathbf{W}_m]}{\pi_m(Y)}\,\phi\Big] = \mathbb{E}\Big[\frac{\pi_m(Y)}{\pi_m(Y)}\,\phi\Big] = \mathbb{E}[\phi].$$

(a) Selection. Since $P(R^{(m)} = 1) = \mathbb{E}[\mathbb{E}[R^{(m)} \mid \mathbf{W}_m]] = \mathbb{E}[\pi_m(Y)] = \bar{\pi}_m$,

$$\mathbb{E}\big[\phi \mid R^{(m)} = 1\big] = \frac{\mathbb{E}[R^{(m)}\phi]}{\bar{\pi}_m} = \frac{\mathbb{E}\big[\mathbb{E}[R^{(m)} \mid \mathbf{W}_m]\,\phi\big]}{\bar{\pi}_m} = \frac{\mathbb{E}[\pi_m(Y)\,\phi]}{\bar{\pi}_m}.$$

Taking $\phi = \mathbb{I}[Y = k]$ gives $P(Y = k \mid R^{(m)} = 1) = \pi_m(k)P(Y = k)/\bar{\pi}_m$.

(c) Transport. Using (a) with $\phi(1 - \pi_m(Y))/\pi_m(Y)$ in place of ϕ, and the analogue of (a) for $R^{(m)} = 0$,

$$\mathbb{E}\Big[\frac{1 - \pi_m(Y)}{\pi_m(Y)}\,\phi \;\Big|\; R^{(m)} = 1\Big] = \frac{1}{\bar{\pi}_m}\mathbb{E}\Big[\pi_m(Y)\,\frac{1 - \pi_m(Y)}{\pi_m(Y)}\,\phi\Big] = \frac{\mathbb{E}[(1 - \pi_m(Y))\phi]}{\bar{\pi}_m},$$

$$\mathbb{E}\big[\phi \mid R^{(m)} = 0\big] = \frac{\mathbb{E}[(1 - R^{(m)})\phi]}{1 - \bar{\pi}_m} = \frac{\mathbb{E}[(1 - \pi_m(Y))\phi]}{1 - \bar{\pi}_m}.$$

Dividing the two lines yields Eq. (18). With $\phi \equiv 1$ it also shows $\mathbb{E}\big[\frac{1-\pi_m(Y)}{\pi_m(Y)} \mid R^{(m)} = 1\big] = \frac{1-\bar{\pi}_m}{\bar{\pi}_m}$, which is the normalizing constant absorbed by the mean-one normalization of Definition C.6.

(d) Selection gap. Combining the two expressions above,

$$\begin{aligned} \mathbb{E}\big[\phi \mid R^{(m)} = 1\big] - \mathbb{E}\big[\phi \mid R^{(m)} = 0\big] &= \frac{\mathbb{E}[\pi_m(Y)\phi]}{\bar{\pi}_m} - \frac{\mathbb{E}[(1 - \pi_m(Y))\phi]}{1 - \bar{\pi}_m} \\ &= \frac{\mathbb{E}\big[\phi\,\{\pi_m(Y)(1 - \bar{\pi}_m) - (1 - \pi_m(Y))\bar{\pi}_m\}\big]}{\bar{\pi}_m(1 - \bar{\pi}_m)} \\ &= \frac{\mathbb{E}\big[\phi\,(\pi_m(Y) - \bar{\pi}_m)\big]}{\bar{\pi}_m(1 - \bar{\pi}_m)} = \frac{\mathrm{Cov}\big(\phi, \pi_m(Y)\big)}{\bar{\pi}_m(1 - \bar{\pi}_m)} \\ &= \frac{\mathrm{Cov}\big(\mathbb{E}[\phi \mid Y], \pi_m(Y)\big)}{\bar{\pi}_m(1 - \bar{\pi}_m)}, \end{aligned}$$

where we used $\mathbb{E}[\pi_m(Y)] = \bar{\pi}_m$ and, in the last step, the tower property with the Y-measurability of $\pi_m(Y)$. ∎

E.3 Proofs for the Client Side (Section 3.1 and Appendix D.1)

Proof of Proposition D.1. The cluster law is Lemma C.7(a) with $\phi = \mathbb{I}[Y = k]$. For the gradient, the integrable envelope allows us to exchange differentiation and expectation. Decomposing over clusters,

$$\mathbb{E}\big[\nabla_\theta \ell \mid R^{(m)} = 1\big] = \sum_{k=1}^{K} P\big(Y = k \mid R^{(m)} = 1\big)\, \mathbb{E}\big[\nabla_\theta \ell \mid Y = k, R^{(m)} = 1\big]$$
$$= \sum_{k=1}^{K} \frac{\pi_m(k) P(Y = k)}{\bar{\pi}_m}\, \mathbb{E}\big[\nabla_\theta \ell \mid Y = k\big],$$

where the second equality uses $R^{(m)} \perp\!\!\!\perp \mathbf{W}_m \mid Y$ (Lemma E.5), so that conditioning on $R^{(m)} = 1$ does not change the conditional law of $\nabla_\theta \ell$ given $Y = k$. Comparing with the population gradient $\sum_k P(Y = k)\mathbb{E}[\nabla_\theta \ell \mid Y = k]$ shows that cluster k is reweighted by $\pi_m(k)/\bar{\pi}_m$, and its term vanishes as $\pi_m(k) \to 0$. ∎

Proof of Theorem D.2. (i) Interventional risk. By Assumption 2.2(ii) and the transfer principle for conditional distributions, we may represent the components of the pattern by separate assignments $R^{(v)} := F_{R,v}(Y, C, U_{R,v})$ with mutually independent $U_{R,v}$, without changing the joint law of $(\mathbf{W}_m, R^{(m)})$. No endogenous variable other than the proxy $X^{(m)*}$ has $R^{(m)}$ as a parent, so in the submodel of $\mathrm{do}(R^{(m)} = 1)$ the law of $\mathbf{W}_m$ is unchanged and $R^{(m)} \equiv 1$. Hence

$$\mathbb{E}_{\mathrm{do}(R^{(m)}=1)}\big[\ell^{(m)}\big] = \mathbb{E}\Big[\ell^{(m)}\big(X^{(1:M)}, (1, \mathbf{R}^{(-m)})\big)\Big]. \tag{48}$$

On $\{R^{(m)} = 1\}$ we have $\ell^{(m)}(X^{(1:M)}, \mathbf{R}) = \ell^{(m)}(X^{(1:M)}, (1, \mathbf{R}^{(-m)}))$, and on $\{R^{(m)} = 0\}$ both sides of the next display vanish. Therefore, with $\phi = \ell^{(m)}(X^{(1:M)}, (1, \mathbf{R}^{(-m)}))$, which is a function of $\mathbf{W}_m$,

$$\mathbb{E}\Big[\frac{R^{(m)}}{\pi_m(Y)}\, \ell^{(m)}\big(X^{(1:M)}, \mathbf{R}\big)\Big] = \mathbb{E}\Big[\frac{R^{(m)}}{\pi_m(Y)}\, \phi\Big] \overset{Eq.\ (17)}{=} \mathbb{E}[\phi] \overset{Eq.\ (48)}{=} \mathbb{E}_{\mathrm{do}(R^{(m)}=1)}\big[\ell^{(m)}\big].$$

If $\nabla_{\theta_m} \ell^{(m)}$ admits an integrable envelope, dominated convergence permits differentiation under both expectations, and the same chain of equalities holds for the gradients.

(ii) Consistency. With $\hat{\pi}_{im} = \pi_m(y_i)$, the weighted objective in Eq. (3) (without the reconstruction term) is the ratio

$$\frac{\frac{1}{N}\sum_{i=1}^{N} \frac{r_{im}}{\pi_m(y_i)} \ell_i^{(m)}}{\frac{1}{N}\sum_{i=1}^{N} \frac{r_{im}}{\pi_m(y_i)}} \xrightarrow{P} \frac{\mathbb{E}\big[\frac{R^{(m)}}{\pi_m(Y)} \ell^{(m)}\big]}{\mathbb{E}\big[\frac{R^{(m)}}{\pi_m(Y)}\big]} = \frac{\mathbb{E}_{\mathrm{do}(R^{(m)}=1)}[\ell^{(m)}]}{1},$$

by the law of large numbers for numerator and denominator, the continuous mapping theorem, and (i) applied to $\ell^{(m)}$ and to the constant 1. The unweighted objective $\frac{1}{N_m}\sum_{i \in \mathcal{O}_m} \ell_i^{(m)}$ is a sample mean over the selected population and converges to $\mathbb{E}[\ell^{(m)} \mid R^{(m)} = 1]$.

(iii) Effective sample size. The normalization in Eq. (3) multiplies all weights by the same constant, which leaves the ratio $(\sum w)^2 / \sum w^2$ unchanged. Before normalization the weights are $1/\hat{\pi}_{im} \in [1, 1/\underline{\pi}]$. Lemma E.1 with $a = 1$, $b = 1/\underline{\pi}$ and $n = N_m$ gives

$$\frac{\big(\sum_{i \in \mathcal{O}_m} 1/\hat{\pi}_{im}\big)^2}{\sum_{i \in \mathcal{O}_m} 1/\hat{\pi}_{im}^2} \geq \frac{4 \cdot 1 \cdot \frac{1}{\underline{\pi}}}{\big(1 + \frac{1}{\underline{\pi}}\big)^2} N_m = \frac{4\underline{\pi}}{(1 + \underline{\pi})^2} N_m.$$

∎

E.4 Proofs for Module I (Section 3.2 and Appendix D.2)

Proof of Lemma D.4. Fix (k, m) and write $s = n_{km}$, $n = n_k$ and $\bar{\pi} = \hat{\bar{\pi}}_m$. Since $\mathrm{Beta}(\pi; 1 + \alpha\bar{\pi}, 1 + \alpha(1 - \bar{\pi})) \propto \pi^{\alpha\bar{\pi}}(1 - \pi)^{\alpha(1-\bar{\pi})}$, the log-posterior is, up to a constant,

$$\psi_{km}(\pi) = \big(s + \alpha\bar{\pi}\big) \log \pi + \big(n - s + \alpha(1 - \bar{\pi})\big) \log(1 - \pi), \qquad \pi \in (0, 1). \tag{49}$$

(i) Differentiating twice,

$$\psi'_{km}(\pi) = \frac{s+\alpha\bar{\pi}}{\pi} - \frac{n-s+\alpha(1-\bar{\pi})}{1-\pi},$$
$$\psi''_{km}(\pi) = -\frac{s+\alpha\bar{\pi}}{\pi^2} - \frac{n-s+\alpha(1-\bar{\pi})}{(1-\pi)^2} < 0,$$

where strictness holds because the two nonnegative coefficients sum to $n+\alpha > 0$. Hence ψ_{km} is strictly concave, and its stationary point solves $(s+\alpha\bar{\pi})(1-\pi) = (n-s+\alpha(1-\bar{\pi}))\pi$, i.e.,

$$s+\alpha\bar{\pi} = \pi\,(n+\alpha) \qquad \Longleftrightarrow \qquad \pi^\circ = \frac{s+\alpha\bar{\pi}}{n+\alpha} = \hat{\pi}_m(k).$$

Because ψ'_{km} is strictly decreasing, it is positive on $(0,\pi^\circ)$ and negative on $(\pi^\circ,1)$. On the interval $[\underline{\pi}, 1-\underline{\pi}]$ the maximum is therefore attained at π° if π° lies in the interval, at $\underline{\pi}$ if $\pi^\circ < \underline{\pi}$, and at $1-\underline{\pi}$ if $\pi^\circ > 1-\underline{\pi}$; in all cases at $\mathrm{clip}_{[\underline{\pi},1-\underline{\pi}]}(\pi^\circ) = \check{\pi}_m(k)$, and uniquely by strict concavity.

(ii) For $n \geq 1$, $\frac{s+\alpha\bar{\pi}}{n+\alpha} = \frac{n}{n+\alpha}\cdot\frac{s}{n} + \frac{\alpha}{n+\alpha}\cdot\bar{\pi}$.

(iii) By (ii), $|\hat{\pi}_m(k) - \bar{\pi}| = \lambda_k\,|n_{km}/n_k - \bar{\pi}| \leq |n_{km}/n_k - \bar{\pi}|$; taking the maximum over k gives the claim. ■

Proof of Theorem D.3. (i) Factorization. By Assumption 2.2(i), Lemma C.4(i) and Condition (P), Eq. (47) holds: $\mathbf{R} \perp\!\!\!\perp (C, X^{(1:M)}) \mid Y$. For the fixed pattern $\boldsymbol{r}$, the visible block $X_{\boldsymbol{r}} = (X^{(v)})_{v:r_v=1}$ is a fixed sub-vector of $X^{(1:M)}$, so by (G1) $\mathbf{R} \perp\!\!\!\perp X_{\boldsymbol{r}} \mid Y$. Consequently, for every measurable set A,

$$P\big(Y=k,\ X_{\boldsymbol{r}} \in A,\ \mathbf{R}=\boldsymbol{r}\big) = P(Y=k)\ P\big(\mathbf{R}=\boldsymbol{r} \mid Y=k\big)\ P\big(X_{\boldsymbol{r}} \in A \mid Y=k, \mathbf{R}=\boldsymbol{r}\big)$$
$$= P(Y=k)\ \varpi_k(\boldsymbol{r}) \int_A p_k(\boldsymbol{x}_{\boldsymbol{r}})\, d\boldsymbol{x}_{\boldsymbol{r}}.$$

The conditional law of Y given $(X_{\boldsymbol{r}}, \mathbf{R}) = (\boldsymbol{x}_{\boldsymbol{r}}, \boldsymbol{r})$ is therefore proportional to $P(Y = k)\, p_k(\boldsymbol{x}_{\boldsymbol{r}})\, \varpi_k(\boldsymbol{r})$. Dividing numerator and denominator by $\sum_l P(Y=l) p_l(\boldsymbol{x}_{\boldsymbol{r}}) > 0$ gives

$$P\big(Y=k \mid \boldsymbol{x}_{\boldsymbol{r}}, \boldsymbol{r}\big) = \frac{P(Y=k)p_k(\boldsymbol{x}_{\boldsymbol{r}})\varpi_k(\boldsymbol{r})}{\sum_l P(Y=l)p_l(\boldsymbol{x}_{\boldsymbol{r}})\varpi_l(\boldsymbol{r})} = \frac{\check{P}(Y=k \mid \boldsymbol{x}_{\boldsymbol{r}})\,\varpi_k(\boldsymbol{r})}{\sum_l \check{P}(Y=l \mid \boldsymbol{x}_{\boldsymbol{r}})\,\varpi_l(\boldsymbol{r})}.$$

(ii) Evidence as log-odds. Taking the logarithm of the ratio of (i) for k and l cancels the common denominator. Under Assumption 2.2(ii) and Condition (P), Step 2 in the proof of Lemma E.5 gives $R^{(m)} \perp\!\!\!\perp \mathbf{R}^{(-m)} \mid Y$ for every m; applying it successively to $m = 1, 2, \ldots$ together with (G1),

$$\varpi_k(\boldsymbol{r}) = \prod_{m=1}^{M} \pi_m(k)^{r_m}\big(1-\pi_m(k)\big)^{1-r_m}, \tag{50}$$

so that $\log\frac{\varpi_k(\boldsymbol{r})}{\varpi_l(\boldsymbol{r})} = \sum_m \big[r_m \log\frac{\pi_m(k)}{\pi_m(l)} + (1-r_m)\log\frac{1-\pi_m(k)}{1-\pi_m(l)}\big]$. This is $\Lambda_{il} - \Lambda_{ik}$ with $\check{\pi}$ replaced by π. If the product model is truncated to admissible patterns as in Lemma E.3, then $\varpi_k(\boldsymbol{r})$ acquires the factor $1/Z_0(k)$ with $Z_0(k) = 1 - \prod_m(1-\check{\pi}_m(k))$, which does not depend on $\boldsymbol{r}$ and merely shifts the prior log-odds by $\log(Z_0(l)/Z_0(k))$.

(iii) Expected evidence. By definition,

$$\mathbb{E}\Big[\log\frac{\varpi_Y(\mathbf{R})}{\varpi_l(\mathbf{R})}\,\Big|\,Y=k\Big] = \sum_{\boldsymbol{r}\in\mathfrak{R}} \varpi_k(\boldsymbol{r})\log\frac{\varpi_k(\boldsymbol{r})}{\varpi_l(\boldsymbol{r})} = D_{\mathrm{KL}}(\varpi_k\|\varpi_l),$$

which is nonnegative with equality if and only if $\varpi_k = \varpi_l$ by Gibbs' inequality. Under Eq. (50) both laws are product measures, and the chain rule of the Kullback-Leibler divergence gives $D_{\mathrm{KL}}(\varpi_k\|\varpi_l) = \sum_m \mathrm{kl}(\pi_m(k)\|\pi_m(l))$. Finally, by the definition of mutual information,

$$I(Y;\mathbf{R}) = \mathbb{E}\Big[\log\frac{P(\mathbf{R}\mid Y)}{P(\mathbf{R})}\Big] = \sum_k P(Y=k)\sum_{\boldsymbol{r}} \varpi_k(\boldsymbol{r})\log\frac{\varpi_k(\boldsymbol{r})}{\bar{\varpi}(\boldsymbol{r})}$$
$$= \sum_k P(Y=k) D_{\mathrm{KL}}(\varpi_k\|\bar{\varpi}).$$

(iv) Bayes suboptimality. If $\varpi_k \equiv \varpi$ for all k, then (i) gives $P(Y = k \mid \boldsymbol{x}_{\boldsymbol{r}}, \boldsymbol{r}) = \check{P}(Y = k \mid \boldsymbol{x}_{\boldsymbol{r}})$, so $\check{h}$ is the Bayes rule for $(X_{\mathbf{R}}, \mathbf{R})$ whatever the feature law. Conversely, suppose $\varpi_k \neq \varpi_l$ for some $k \neq l$. Since both are probability mass functions on $\mathfrak{R}$, there is $\boldsymbol{r}^\circ$ with $\varpi_l(\boldsymbol{r}^\circ) > \varpi_k(\boldsymbol{r}^\circ)$; otherwise $\varpi_l \leq \varpi_k$ pointwise with a strict inequality somewhere, contradicting $\sum_{\boldsymbol{r}} \varpi_l(\boldsymbol{r}) = \sum_{\boldsymbol{r}} \varpi_k(\boldsymbol{r}) = 1$. Construct a feature law compatible with Assumption 2.2(i): $P(Y = k) = P(Y = l) = \frac{1}{2}$, $C = \boldsymbol{c}_Y$ with distinct unit vectors $\boldsymbol{c}_k \neq \boldsymbol{c}_l$, and $X^{(v)} = C + S^{(v)}$ with $S^{(v)} \sim \mathcal{N}(\mathbf{0}, s^2 \boldsymbol{I})$. The likelihood ratio on the visible block of $\boldsymbol{r}^\circ$,

$$\frac{p_k(\boldsymbol{x})}{p_l(\boldsymbol{x})} = \exp\Big(\frac{1}{s^2} \sum_{v: r_v^\circ = 1} \big\langle \boldsymbol{x}^{(v)}, \boldsymbol{c}_k - \boldsymbol{c}_l \big\rangle + \text{const}\Big),$$

is the exponential of a non-constant affine function, hence the set $\mathcal{A} = \{\boldsymbol{x} : 1 < p_k(\boldsymbol{x})/p_l(\boldsymbol{x}) < \varpi_l(\boldsymbol{r}^\circ)/\varpi_k(\boldsymbol{r}^\circ)\}$ (with upper bound $+\infty$ if $\varpi_k(\boldsymbol{r}^\circ) = 0$) is a nonempty open slab of positive Gaussian measure. For $\boldsymbol{x} \in \mathcal{A}$ we have $\check{P}(k \mid \boldsymbol{x}) > \check{P}(l \mid \boldsymbol{x})$, so $\check{h}(\boldsymbol{x}) = k$, whereas by (i)

$$\frac{P(k \mid \boldsymbol{x}, \boldsymbol{r}^\circ)}{P(l \mid \boldsymbol{x}, \boldsymbol{r}^\circ)} = \frac{p_k(\boldsymbol{x})\, \varpi_k(\boldsymbol{r}^\circ)}{p_l(\boldsymbol{x})\, \varpi_l(\boldsymbol{r}^\circ)} < 1,$$

so the Bayes rule chooses l. The excess 0-1 risk of $\check{h}$ is at least

$$\mathbb{E}\Big[\big(P(l \mid X_{\boldsymbol{r}^\circ}, \boldsymbol{r}^\circ) - P(k \mid X_{\boldsymbol{r}^\circ}, \boldsymbol{r}^\circ)\big)\, \mathbb{I}[X_{\boldsymbol{r}^\circ} \in \mathcal{A}, \mathbf{R} = \boldsymbol{r}^\circ]\Big] > 0,$$

because the integrand is strictly positive on $\mathcal{A}$ and $P(X_{\boldsymbol{r}^\circ} \in \mathcal{A}, \mathbf{R} = \boldsymbol{r}^\circ) \geq \frac{1}{2}\varpi_l(\boldsymbol{r}^\circ) P(X_{\boldsymbol{r}^\circ} \in \mathcal{A} \mid Y = l) > 0$. ∎

Proof of Proposition D.5. Under (W1) to (W3), and writing $-\log \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d) = \frac{1}{2\sigma^2}\|\boldsymbol{z} - \boldsymbol{\mu}\|^2 + c_d(\sigma^2)$ with c_d independent of $\boldsymbol{\mu}$,

$$\begin{aligned} -\log P_\omega\big(Y = k \mid \tilde{\boldsymbol{z}}_i, \boldsymbol{r}_i\big) &= \sum_{m=1}^{M} \omega_{im} \Big[\frac{\|\tilde{\boldsymbol{z}}_i^{(m)} - \boldsymbol{\mu}_k^{(m)}\|^2}{2\sigma^2} + c_d(\sigma^2)\Big] + \log K + \log \mathcal{Z}_i \\ &\quad - \sum_{m=1}^{M} \Big[r_{im} \log \check{\pi}_m(k) + (1 - r_{im}) \log(1 - \check{\pi}_m(k))\Big] \\ &= \frac{\sum_m \omega_{im}}{2\sigma^2} \Big[\Delta_{ik} + \frac{2\sigma^2\, \Lambda_{ik}}{\sum_m \omega_{im}}\Big] + \underbrace{c_d(\sigma^2) \sum_m \omega_{im}}_{\text{independent of } k} + \log K + \log \mathcal{Z}_i, \end{aligned}$$

where $\mathcal{Z}_i$ is the normalizer of the posterior. Since $\sum_m \omega_{im} \geq \sum_m r_{im} \geq 1 > 0$, maximizing the posterior over k is equivalent to minimizing the bracket, which is $\tilde{\Delta}_{ik}$ with $\lambda_R = 1$. Tempering the Bernoulli likelihood in (W2) by λ_R multiplies Λ_{ik} by λ_R, which gives Eq. (6) for general λ_R. ∎

Proof of Corollary D.6. (i) If $\check{\pi}_m(k) = \check{\pi}_m$ for all k, then $\Lambda_{ik} = \Lambda_i$ does not depend on k, so $\tilde{\Delta}_{ik} = \Delta_{ik} + 2\lambda_R \sigma^2 \Lambda_i / \sum_m \omega_{im}$ differs from Δ_{ik} by a constant in k, and the minimizers coincide. Under MCAR, Lemma D.4(iii) drives the table to the marginal, and clipping preserves this.

(ii) For each m and both values of r_{im}, the difference of the m-th terms of Λ_{ik} and Λ_{il} is either $\log(\check{\pi}_m(l)/\check{\pi}_m(k))$ or $\log\big((1 - \check{\pi}_m(l))/(1 - \check{\pi}_m(k))\big)$. Both $\check{\pi}_m(\cdot)$ and $1 - \check{\pi}_m(\cdot)$ lie in $[\underline{\pi}, 1 - \underline{\pi}]$, so each difference is bounded in absolute value by $\log \frac{1-\underline{\pi}}{\underline{\pi}}$, and $|\Lambda_{ik} - \Lambda_{il}| \leq M \log \frac{1-\underline{\pi}}{\underline{\pi}}$. Since

$$\big(\tilde{\Delta}_{ik} - \tilde{\Delta}_{il}\big) - \big(\Delta_{ik} - \Delta_{il}\big) = \frac{2\lambda_R \sigma^2}{\sum_m \omega_{im}} \big(\Lambda_{ik} - \Lambda_{il}\big)$$

and $\sum_m \omega_{im} \geq 1$, Eq. (24) follows. If $\Delta_{il} - \Delta_{ik}$ exceeds the bound, then $\tilde{\Delta}_{il} - \tilde{\Delta}_{ik} > 0$, so the evidence cannot reverse the preference. ∎

Proof of Lemma D.7. For $l \neq k$ the table of cluster l does not involve sample i, so $\Lambda_{il} = \Lambda_{il}^{(-i)}$. Fix m, let $S = \sum_{j \neq i : a_j = k} r_{jm}$, $n = n_k$ (including i) and $\bar{\pi} = \hat{\bar{\pi}}_m$. Then

$$\pi^{\mathrm{in}} = \frac{S + r_{im} + \alpha\bar{\pi}}{n + \alpha}, \qquad \pi^{\mathrm{loo}} = \frac{S + \alpha\bar{\pi}}{n - 1 + \alpha},$$

and a direct computation gives

$$\begin{aligned}\pi^{\mathrm{in}} - \pi^{\mathrm{loo}} &= \frac{(S + r_{im} + \alpha\bar{\pi})(n - 1 + \alpha) - (S + \alpha\bar{\pi})(n + \alpha)}{(n + \alpha)(n - 1 + \alpha)} \\ &= \frac{r_{im}(n - 1 + \alpha) - (S + \alpha\bar{\pi})}{(n + \alpha)(n - 1 + \alpha)} = \frac{r_{im} - \pi^{\mathrm{loo}}}{n + \alpha}.\end{aligned}$$

Thus $\pi^{\mathrm{in}} - \pi^{\mathrm{loo}} \in [0, \frac{1}{n+\alpha}]$ if $r_{im} = 1$ and $\in [-\frac{1}{n+\alpha}, 0]$ if $r_{im} = 0$. Clipping to $[\underline{\pi}, 1 - \underline{\pi}]$ is monotone and 1-Lipschitz, so the clipped values $\check{\pi}^{\mathrm{in}}, \check{\pi}^{\mathrm{loo}}$ satisfy the same sign pattern and $|\check{\pi}^{\mathrm{in}} - \check{\pi}^{\mathrm{loo}}| \leq \frac{1}{n+\alpha}$. The m-th term of $\Lambda_{ik}^{(-i)} - \Lambda_{ik}$ equals $\log \check{\pi}^{\mathrm{in}} - \log \check{\pi}^{\mathrm{loo}}$ if $r_{im} = 1$ and $\log(1 - \check{\pi}^{\mathrm{in}}) - \log(1 - \check{\pi}^{\mathrm{loo}})$ if $r_{im} = 0$. In both cases it is nonnegative by the sign pattern, and by the mean value theorem, since the logarithm is $1/\underline{\pi}$-Lipschitz on $[\underline{\pi}, 1 - \underline{\pi}]$,

$$0 \leq (m\text{-th term}) \leq \frac{|\check{\pi}^{\mathrm{in}} - \check{\pi}^{\mathrm{loo}}|}{\underline{\pi}} \leq \frac{1}{(n + \alpha)\underline{\pi}}.$$

Summing over m yields Eq. (25). For the cross-fitted evidence, the table used for $i \in \mathcal{H}_b$ is computed from $\mathcal{H}_{1-b}$ and does not contain $(a_i, \boldsymbol{r}_i)$, so the above self-influence term is absent. ■

E.5 Proofs for Module II (Section 3.3 and Appendix D.3)

Proof of Lemma D.8. (i) $\hat{\boldsymbol{z}}_i^{\mathrm{proto},(m)} = \sum_k \bar{q}_{ik} \boldsymbol{\mu}_k^{(m)}$ is a measurable function of $\mathbf{V}_i := (\bar{\boldsymbol{q}}_i, \{\boldsymbol{\mu}_k^{(m)}\}_k)$, so $\sigma(\mathbf{V}_i, \hat{\boldsymbol{z}}_i^{\mathrm{proto},(m)}) = \sigma(\mathbf{V}_i)$. By the chain rule of mutual information,

$$I\big(\hat{\boldsymbol{z}}_i^{\mathrm{proto},(m)}; Y \,\big|\, \mathbf{V}_i\big) = I\big(Y; (\mathbf{V}_i, \hat{\boldsymbol{z}}_i^{\mathrm{proto},(m)})\big) - I\big(Y; \mathbf{V}_i\big) = 0.$$

(ii) With $\bar{\boldsymbol{q}}_i = \boldsymbol{e}_{a_i}$, the imputed block equals $\boldsymbol{\mu}_{a_i}^{(m)}$. Its contribution to Δ_{ik} in Eq. (6) is $\omega_{im} \|\boldsymbol{\mu}_{a_i}^{(m)} - \boldsymbol{\mu}_k^{(m)}\|^2 / \sum_v \omega_{iv}$ and its contribution to Δ_{ia_i} is 0. Hence the block adds $\omega_{im} \|\boldsymbol{\mu}_{a_i}^{(m)} - \boldsymbol{\mu}_k^{(m)}\|^2 / \sum_v \omega_{iv} \geq 0$ to $\Delta_{ik} - \Delta_{ia_i}$, strictly if $\omega_{im} > 0$ and $\boldsymbol{\mu}_k^{(m)} \neq \boldsymbol{\mu}_{a_i}^{(m)}$. ■

Proof of Theorem D.9. (i) Invariance. Lemma C.4(ii) gives $X^{(m)} \perp\!\!\!\perp (Y, \mathbf{R}, X^{(-m)}) \mid C$. Since $Z^{(m)} = \mathrm{Enc}_m(X^{(m)})$ with a fixed encoder, (G4) yields $Z^{(m)} \perp\!\!\!\perp (Y, \mathbf{R}, X^{(-m)}) \mid C$ and in particular $P(Z^{(m)} \mid C, \mathbf{R}) = P(Z^{(m)} \mid C)$. By Assumption 2.3(i), $\boldsymbol{\xi}^{(m)} = \zeta_m(S^{(m)})$ with $S^{(m)} \perp\!\!\!\perp (C, \mathbf{R})$ by Eq. (46); hence

$$\mathbb{E}\big[Z^{(m)} \mid C, \mathbf{R}\big] = f_m(C) + \mathbb{E}\big[\boldsymbol{\xi}^{(m)} \mid C, \mathbf{R}\big] = f_m(C) + \mathbb{E}\big[\boldsymbol{\xi}^{(m)}\big] = f_m(C).$$

(ii) Recoverability. Fix $v \neq m$ and the event $\mathcal{C} = \{R^{(m)} = R^{(v)} = 1\}$. By Assumption 2.2(ii) and (iii), $P(\mathcal{C} \mid Y, C) = \pi_m^\star \pi_v^\star \geq \eta^2 > 0$, so conditioning on $\mathcal{C}$ is well defined for P_C-almost every value of C. By Assumption 2.3(ii), $h_v^\star(Z^{(v)}) = C$ almost surely, so the two random variables generate the same conditional expectations. Using (i) with the $\mathbf{R}$-measurable event $\mathcal{C}$,

$$\begin{aligned}\mathbb{E}\big[Z^{(m)} \,\big|\, h_v^\star(Z^{(v)}), \mathcal{C}\big] = \mathbb{E}\big[Z^{(m)} \,\big|\, C, \mathcal{C}\big] &= \frac{\mathbb{E}\big[Z^{(m)} \mathbb{I}_{\mathcal{C}} \mid C\big]}{P(\mathcal{C} \mid C)} \\ &= \frac{\mathbb{E}\big[\mathbb{E}[Z^{(m)} \mid C, \mathbf{R}]\, \mathbb{I}_{\mathcal{C}} \mid C\big]}{P(\mathcal{C} \mid C)} = f_m(C).\end{aligned}$$

On $\mathcal{C}$ both $Z^{(m)}$ and $Z^{(v)}$ are observed, so the left-hand side is a functional of the observed-data law, and so is f_m on the support of P_C.

(iii) Counterfactual identifiability. Augment the SCM of Assumption 2.2(i) with the proxies $Z^{(v)*} = Z^{(v)}$ if $R^{(v)} = 1$ and $Z^{(v)*} = \star$ otherwise, and let $\boldsymbol{e} = (\mathbf{R} = \boldsymbol{r},\ Z^{(v)*} = \boldsymbol{z}^{(v)}\ \forall v \in \mathcal{V})$ with $r_m = 0$. We follow Definition C.1.

- *Abduction.* Under $P(\mathcal{U} \mid \boldsymbol{e})$, Assumption 2.3(ii) gives $h_v^\star(\boldsymbol{z}^{(v)}) = C$ for all $v \in \mathcal{V}$, hence $\sum_{v\in\mathcal{V}} h_v^\star(\boldsymbol{z}^{(v)}) = |\mathcal{V}|\, C$ and $C = \mathrm{nrm}\big(\sum_{v\in\mathcal{V}} h_v^\star(\boldsymbol{z}^{(v)})\big) = \boldsymbol{c}^\star$ almost surely. The evidence $\boldsymbol{e}$ is a function of $(U_Y, U_C, U_R, (U_{S^{(v)}})_{v\in\mathcal{V}})$, which is independent of $U_{S^{(m)}}$; therefore the posterior of $U_{S^{(m)}}$ equals its prior.
- *Action.* Replace the assignment of $R^{(m)}$ by the constant 1. Since no variable among $(Y, C, \mathbf{S}, X^{(1:M)})$ is a descendant of $R^{(m)}$, their assignments are unchanged, and the proxy becomes $Z^{(m)*} = Z^{(m)}$.
- *Prediction.* In the modified model, $Z^{(m)*}_{\mathrm{do}(R^{(m)}=1)} = Z^{(m)} = f_m(C) + \zeta_m(U_{S^{(m)}})$. Under the abducted posterior this equals $f_m(\boldsymbol{c}^\star) + \zeta_m(U_{S^{(m)}})$ with $U_{S^{(m)}}$ distributed according to its prior, i.e., $f_m(\boldsymbol{c}^\star) + \boldsymbol{\xi}^{(m)}$ with $\boldsymbol{\xi}^{(m)} \sim P_{\boldsymbol{\xi}^{(m)}}$ independent of $\boldsymbol{e}$.

Taking expectations and using $\mathbb{E}\boldsymbol{\xi}^{(m)} = \mathbf{0}$ gives $\mathbb{E}[Z^{(m)*}_{\mathrm{do}(R^{(m)}=1)} \mid \boldsymbol{e}] = f_m(\boldsymbol{c}^\star)$. Since the conditional mean minimizes the conditional mean-squared error among all functions of $\boldsymbol{e}$, Eq. (9) with the oracle maps is the minimum mean-squared-error counterfactual imputation. ■

Example E.6 (Two-cluster Gaussian model)**.** Let $K = 2$, $P(Y = 1) = P(Y = 2) = \frac{1}{2}$, and let the content be the scalar $C = \mu s_Y$ with $s_1 = 1$, $s_2 = -1$ and $\mu > 0$ (the sphere constraint is dropped for illustration). Let a single source view be $Z^{(a)} = C + \vartheta\epsilon_a$ and the target view $Z^{(m)} = C + \vartheta\epsilon_m$ with $\epsilon_a, \epsilon_m \sim \mathcal{N}(0, 1)$ independent of everything else, $R^{(a)} \equiv 1$ and $R^{(m)} \mid Y \sim \mathrm{Bernoulli}(\pi_m(Y))$. Then $f_m(c) = c$ with $L_f = 1$ in both populations, and Assumption 2.2 (whenever $\pi_m(1), \pi_m(2) \in (0, 1)$), Condition (P) and Assumption 2.3(i) hold, while Assumption 2.3(ii) fails because $Z^{(a)}$ is noisy.

Proof of Proposition D.10. (i) By Assumption 2.3(i), $Z^{(m)} = f_m(C) + \boldsymbol{\xi}^{(m)}$, where $\boldsymbol{\xi}^{(m)}$ is a function of $S^{(m)}$ and is therefore independent of $(C, Z^{(\mathcal{A})}, R^{(m)})$, which are functions of $(\mathbf{U}_0, (S^{(v)})_{v\in\mathcal{A}})$ by Eq. (46). Hence

$$g_r^{\mathrm{as}}(\boldsymbol{z}_\mathcal{A}) = \mathbb{E}\big[f_m(C) \mid Z^{(\mathcal{A})} = \boldsymbol{z}_\mathcal{A}, R^{(m)} = r\big] + \mathbb{E}\big[\boldsymbol{\xi}^{(m)}\big] = \int f_m(\boldsymbol{c})\, P_{C|\boldsymbol{z}_\mathcal{A},r}(d\boldsymbol{c}).$$

For any unit vector $\boldsymbol{u}$, the scalar function $\boldsymbol{c} \mapsto \boldsymbol{u}^\top f_m(\boldsymbol{c})$ is L_f-Lipschitz, so by the Kantorovich-Rubinstein duality (Villani, 2009),

$$\boldsymbol{u}^\top\big(g_1^{\mathrm{as}}(\boldsymbol{z}_\mathcal{A}) - g_0^{\mathrm{as}}(\boldsymbol{z}_\mathcal{A})\big) = \int \boldsymbol{u}^\top f_m\, d\big(P_{C|\boldsymbol{z}_\mathcal{A},1} - P_{C|\boldsymbol{z}_\mathcal{A},0}\big) \le L_f\, W_1\big(P_{C|\boldsymbol{z}_\mathcal{A},1}, P_{C|\boldsymbol{z}_\mathcal{A},0}\big).$$

Taking the supremum over $\boldsymbol{u}$ gives the first inequality in Eq. (26). For the second, let $\hat{\boldsymbol{c}} = \hat{\boldsymbol{c}}(\boldsymbol{z}_\mathcal{A})$ and let $\delta_{\hat{\boldsymbol{c}}}$ be the point mass at $\hat{\boldsymbol{c}}$. Since $W_1(P, \delta_{\hat{\boldsymbol{c}}}) = \mathbb{E}_P\|C - \hat{\boldsymbol{c}}\|$ (the only coupling with a point mass is the product coupling), the triangle inequality for W_1 gives

$$\begin{aligned} W_1\big(P_{C|\boldsymbol{z}_\mathcal{A},1}, P_{C|\boldsymbol{z}_\mathcal{A},0}\big) &\le W_1\big(P_{C|\boldsymbol{z}_\mathcal{A},1}, \delta_{\hat{\boldsymbol{c}}}\big) + W_1\big(\delta_{\hat{\boldsymbol{c}}}, P_{C|\boldsymbol{z}_\mathcal{A},0}\big) \\ &= \sum_{r\in\{0,1\}} \mathbb{E}\big[\|C - \hat{\boldsymbol{c}}\| \,\big|\, Z^{(\mathcal{A})} = \boldsymbol{z}_\mathcal{A}, R^{(m)} = r\big]. \end{aligned}$$

(ii) If $C = h(Z^{(\mathcal{A})})$ almost surely, both conditional laws equal $\delta_{h(\boldsymbol{z}_\mathcal{A})}$ and the bound is zero. In Example E.6, with φ the standard normal density and $\mathrm{sig}(u) = 1/(1 + e^{-u})$,

$$\begin{aligned} P\big(Y = 1 \mid Z^{(a)} = z, R^{(m)} = 1\big) &= \frac{\frac{1}{2}\varphi\big(\frac{z-\mu}{\vartheta}\big)\pi_m(1)}{\frac{1}{2}\varphi\big(\frac{z-\mu}{\vartheta}\big)\pi_m(1) + \frac{1}{2}\varphi\big(\frac{z+\mu}{\vartheta}\big)\pi_m(2)} \\ &= \mathrm{sig}\Big(\frac{(z+\mu)^2 - (z-\mu)^2}{2\vartheta^2} + \log\frac{\pi_m(1)}{\pi_m(2)}\Big) \\ &= \mathrm{sig}\Big(\frac{2\mu z}{\vartheta^2} + \log\frac{\pi_m(1)}{\pi_m(2)}\Big). \end{aligned}$$

Since $\mathbb{E}[C \mid \cdot] = \mu\,(2P(Y = 1 \mid \cdot) - 1)$ and $2\,\mathrm{sig}(2u) - 1 = \tanh(u)$,

$$g_1^{\mathrm{as}}(z) = \mu\tanh\Big(\frac{\mu z}{\vartheta^2} + \frac{1}{2}\log\frac{\pi_m(1)}{\pi_m(2)}\Big), \qquad g_0^{\mathrm{as}}(z) = \mu\tanh\Big(\frac{\mu z}{\vartheta^2} + \frac{1}{2}\log\frac{1-\pi_m(1)}{1-\pi_m(2)}\Big),$$

the second line following in the same way with $1-\pi_m(\cdot)$. The two arguments differ by the constant $\frac{1}{2}\log\frac{\pi_m(1)(1-\pi_m(2))}{\pi_m(2)(1-\pi_m(1))}$, which is nonzero if and only if $\pi_m(1)\neq\pi_m(2)$; since $\tanh$ is strictly increasing, the gap is then nonzero for every z. In contrast, $\mathbb{E}[Z^{(m)}\mid C,R^{(m)}]=C=f_m(C)$ in both populations. Note that, by Lemma C.7(c), fitting any regression of $Z^{(m)}$ on $Z^{(a)}$ with the transport weights on the visible population reproduces the joint law of the missing population and thus targets g_0^{as}, which is the role of the weights in Eq. (28).

(iii) Let $n_v=|\mathcal{V}_i\setminus\{m\}|$ and $\boldsymbol{a}=\frac{1}{n_v}\sum_{v\in\mathcal{V}_i\setminus\{m\}}\boldsymbol{c}_i^{(v)}=C_i+\bar{\boldsymbol{\delta}}_i$. Since nrm is invariant to positive scaling, $\bar{\boldsymbol{c}}_i^{(-m)}=\mathrm{nrm}(\boldsymbol{a})$, and Lemma E.2 with Jensen's inequality gives

$$\mathbb{E}\|\bar{\boldsymbol{c}}_i^{(-m)}-C_i\|\le 2\,\mathbb{E}\|\bar{\boldsymbol{\delta}}_i\|\le 2\big(\mathbb{E}\|\bar{\boldsymbol{\delta}}_i\|^2\big)^{1/2}=2\Big(\frac{1}{n_v^2}\sum_{v}\mathbb{E}\|\boldsymbol{\delta}_i^{(v)}\|^2\Big)^{1/2}\le\frac{2\nu}{\sqrt{n_v}},$$

where the cross terms vanish because the errors are independent with zero mean. ∎

Proof of Proposition D.11. *(i)* For fixed (h,f) and encoders, $\phi=\|Z^{(m)}-f_m(\bar{\boldsymbol{c}}^{(-m)})\|^2$ is a measurable function of $(X^{(1:M)},\mathbf{R}^{(-m)})$, since $\bar{\boldsymbol{c}}^{(-m)}$ depends only on the other visible embeddings. Lemma C.7(c) gives the claim; the factor $\bar{\pi}_m/(1-\bar{\pi}_m)$ is the population version of the mean-one normalization.

(ii) At $(h^\star,f)$, Assumption 2.3(ii) gives $\boldsymbol{c}^{(v)}=h_v^\star(Z^{(v)})=C$ for all visible v, so $\bar{\boldsymbol{c}}^{(-m)}=\mathrm{nrm}(|\mathcal{V}\setminus\{m\}|\,C)=C$ on $\mathcal{E}_m$ for any subset of sources kept by view dropout. The residual is $Z^{(m)}-f_m(C)=\boldsymbol{\xi}^{(m)}$.

- *Counterfactual loss.* For any (h,f), since $\boldsymbol{\xi}^{(m)}$ is independent of $(C,\bar{\boldsymbol{c}}^{(-m)},R^{(m)})$ and centered,

$$\begin{aligned}&\mathbb{E}\big[\|Z^{(m)}-f(\bar{\boldsymbol{c}}^{(-m)})\|^2\mid R^{(m)}=0\big]\\&\quad=\mathbb{E}\big[\|f_m(C)-f(\bar{\boldsymbol{c}}^{(-m)})\|^2\mid R^{(m)}=0\big]+\mathbb{E}\|\boldsymbol{\xi}^{(m)}\|^2\\&\qquad+2\,\mathbb{E}\big[\langle\boldsymbol{\xi}^{(m)},f_m(C)-f(\bar{\boldsymbol{c}}^{(-m)})\rangle\mid R^{(m)}=0\big]\\&\quad=\mathbb{E}\big[\|f_m(C)-f(\bar{\boldsymbol{c}}^{(-m)})\|^2\mid R^{(m)}=0\big]+\mathbb{E}\|\boldsymbol{\xi}^{(m)}\|^2\;\ge\;\mathbb{E}\|\boldsymbol{\xi}^{(m)}\|^2,\end{aligned}$$

  with equality at the oracle; by (i) this is also the minimum of the transport-weighted population loss. It is the Bayes risk because $\mathbb{E}[Z^{(m)}\mid\text{other views}]=f_m(C)$ under Assumption 2.3(ii).

- *Independence loss.* HSIC ≥ 0 always, and $\mathrm{HSIC}(C,\boldsymbol{\xi}^{(m)})=0$ because $\boldsymbol{\xi}^{(m)}\perp\!\!\!\perp C$; hence the population nHSIC vanishes at the oracle and attains its minimum.

- *Invariance loss.* $\langle\boldsymbol{c}^{(m)},\bar{\boldsymbol{c}}^{(-m)}\rangle=\langle C,C\rangle=1$, the maximal inner product on the sphere.

∎

Proof of Lemma D.12. Condition on $\mathcal{F}_i$, the sigma-field generated by the encoders, C_i, the other views of sample i and the training folds. Write $\boldsymbol{z}=\boldsymbol{z}_i^{(m)}=f_m(C_i)+\boldsymbol{\xi}$ and $\boldsymbol{z}'=f_m(C_i)+\boldsymbol{\xi}'$ with $\boldsymbol{\xi}'$ an independent copy of $\boldsymbol{\xi}$. Expanding the squares,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{z}'-\hat{\boldsymbol{z}}\|^2-\mathbb{E}\|\boldsymbol{z}-\hat{\boldsymbol{z}}\|^2&=\mathbb{E}\|\boldsymbol{z}'\|^2-\mathbb{E}\|\boldsymbol{z}\|^2-2\,\mathbb{E}\langle\boldsymbol{z}'-\boldsymbol{z},\hat{\boldsymbol{z}}\rangle=-2\,\mathbb{E}\langle\boldsymbol{\xi}'-\boldsymbol{\xi},\hat{\boldsymbol{z}}\rangle\\&=-2\,\langle\mathbb{E}\boldsymbol{\xi}',\mathbb{E}\hat{\boldsymbol{z}}\rangle+2\,\mathbb{E}\langle\boldsymbol{\xi},\hat{\boldsymbol{z}}\rangle=2\,\mathbb{E}\langle\boldsymbol{\xi},\hat{\boldsymbol{z}}\rangle,\end{aligned}$$

using that $\boldsymbol{z}$ and $\boldsymbol{z}'$ have the same law, that $\boldsymbol{\xi}'$ is independent of $\hat{\boldsymbol{z}}$, and $\mathbb{E}\boldsymbol{\xi}'=\mathbf{0}$. For the cross-fitted prediction Eq. (32), $\hat{\boldsymbol{z}}$ is $\mathcal{F}_i$-measurable: the model is trained on other folds, and its input consists of the other views of sample i, which are functions of $(C_i,(S_i^{(v)})_{v\neq m})$. Since $\boldsymbol{\xi}=\zeta_m(S_i^{(m)})$ is independent of $\mathcal{F}_i$ and centered, $\mathbb{E}\langle\boldsymbol{\xi},\hat{\boldsymbol{z}}\rangle=\mathbb{E}\langle\mathbb{E}[\boldsymbol{\xi}\mid\mathcal{F}_i],\hat{\boldsymbol{z}}\rangle=0$. For an interpolating in-sample fit, $\hat{\boldsymbol{z}}=\boldsymbol{z}$ and $\mathbb{E}\langle\boldsymbol{\xi},\boldsymbol{z}\rangle=\mathbb{E}\langle\boldsymbol{\xi},f_m(C_i)\rangle+\mathbb{E}\|\boldsymbol{\xi}\|^2=\mathbb{E}\|\boldsymbol{\xi}\|^2$. ∎

E.6 Proofs for Module III (Section 3.4 and Appendix D.4)

Throughout this subsection we work on $\mathcal{E}_m$ (Remark E.4) and use two consequences of the assumptions. First, by Lemma C.4(i), (G2) and Assumption 2.2(ii) (which holds on $\mathcal{E}_m$ by Lemma E.3),

and (G3),

$$R^{(m)} \perp\!\!\!\perp \big(X^{(1:M)}, \mathbf{R}^{(-m)}\big) \mid (Y, C), \quad \text{hence} \quad \mathbb{E}\big[R^{(m)} \mid Y, C, X^{(1:M)}, \mathbf{R}^{(-m)}\big] = \pi_m^\star(Y, C). \tag{51}$$

Second, for a scalar $\pi^\star$ and an integrable random vector V, under P_k,

$$\mathbb{E}_k[\pi^\star V] = \bar{\pi}_k^\star\, \mathbb{E}_k[V] + \mathrm{Cov}_k(\pi^\star, V), \qquad \mathrm{Cov}_k(\pi^\star, V) = \mathbb{E}_k\big[(\pi^\star - \bar{\pi}_k^\star)(V - \mathbb{E}_k V)\big]. \tag{52}$$

Cross-fitting makes the imputation of every sample a function of its other views and of data from other folds only. Conditionally on the training folds, the summands within a fold are therefore independent and identically distributed, and the laws of large numbers and central limit theorems below are applied fold by fold and then pooled over the L folds.

Proof of Theorem D.13. Let $\psi = \|Z^{(m)} - \hat{g}_m\|^2$, where $\hat{g}_m$ is evaluated at the other views of the sample, and let $\chi(Y) = (1 - \pi_m(Y))/\pi_m(Y)$ denote the oracle transport weight.

Step 1 (oracle weights). By the law of large numbers and the tower property with Lemma E.5,

$$\frac{1}{N}\sum_i r_{im}\,\chi(y_i)\,\psi_i \xrightarrow{P} \mathbb{E}\big[R^{(m)}\chi(Y)\psi\big] = \mathbb{E}\big[\pi_m(Y)\chi(Y)\psi\big] = \mathbb{E}\big[(1-\pi_m(Y))\psi\big],$$
$$\frac{1}{N}\sum_i r_{im}\,\chi(y_i) \xrightarrow{P} \mathbb{E}\big[(1-\pi_m(Y))\big] = 1 - \bar{\pi}_m.$$

Their ratio converges to $\mathbb{E}[(1-\pi_m(Y))\psi]/(1-\bar{\pi}_m) = \mathbb{E}[\psi \mid R^{(m)} = 0] = \mathrm{MSE}_{\mathrm{mis}}$, as in the proof of Lemma C.7(c). Replacing ψ by $Z^{(m)}$ and by $\|Z^{(m)}\|^2$ shows that the weighted mean $\bar{\boldsymbol{z}}_m^{\mathrm{tr}}$ converges to $\boldsymbol{m}_0 = \mathbb{E}[Z^{(m)} \mid R^{(m)} = 0]$ and the weighted second moment to $\mathbb{E}[\|Z^{(m)}\|^2 \mid R^{(m)} = 0]$; hence the denominator of Eq. (10) converges to $V_0 = \mathbb{E}[\|Z^{(m)} - \boldsymbol{m}_0\|^2 \mid R^{(m)} = 0]$, which we assume positive.

Step 2 (cross-fitted imputations). If the fold imputers converge to $\hat{g}_m$ in L^2, then by the Cauchy-Schwarz inequality the averages computed with $\hat{\boldsymbol{z}}_i^{(m)}$ differ from those computed with $\hat{g}_m$ by $o_P(1)$.

Step 3 (plug-in weights). Since $\hat{\pi}_{im} \geq \underline{\pi}$ and $\pi_m(y_i) \geq \eta$,

$$\big|\hat{\chi}_i - \chi(y_i)\big| = \Big|\frac{1}{\hat{\pi}_{im}} - \frac{1}{\pi_m(y_i)}\Big| \leq \frac{\max_j |\hat{\pi}_{jm} - \pi_m(y_j)|}{\underline{\pi}\,\eta} = o_P(1)$$

uniformly in i, so $\frac{1}{N}\sum_i r_{im}|\hat{\chi}_i - \chi(y_i)|\psi_i \leq o_P(1)\cdot\frac{1}{N}\sum_i \psi_i = o_P(1)$, and the same holds for the other weighted averages.

Step 4 (continuous mapping). Steps 1 to 3 give numerator and denominator of the ratio in Eq. (10) converging to $\mathrm{MSE}_{\mathrm{mis}}$ and $V_0 > 0$. The map $(a, b) \mapsto \mathrm{clip}_{[0,1]}(1 - a/b)$ is continuous at $(\mathrm{MSE}_{\mathrm{mis}}, V_0)$, so $\hat{\rho}_m \to \mathrm{clip}_{[0,1]}(\rho_m)$ in probability.

(ii) Apply Lemma C.7(d) to $\phi = \psi$, which is a function of $(X^{(1:M)}, \mathbf{R}^{(-m)})$ and of independent training data; $\mathbb{E}[\psi \mid Y] = e_m(Y)$ gives Eq. (35).

(iii) Under MCAR, $\pi_m(k) \equiv \bar{\pi}_m$, so the weights are constant, the normalized weights are uniform on $\mathcal{O}_m^+$, $\hat{\rho}_m$ becomes the unweighted coefficient of determination, and $\mathrm{Cov}(e_m(Y), \pi_m(Y)) = 0$ in Eq. (35). A binary adoption switch uses the fusion weights $\omega_{im} \in \{0, 1\}$ for missing entries, which is Eq. (11) with $\hat{\rho}_m \in \{0, 1\}$. ■

Proof of Theorem D.14. Fix k, m and ρ. With oracle memberships the cluster-level propensity $\hat{\pi}_{im} = \hat{\pi}_m(k)$ is constant over the members of cluster k and cancels in Eq. (12) and Eq. (13), so that

$$\hat{\boldsymbol{\mu}}_{km}^{\mathrm{dr}}(\rho) = \frac{\sum_i a_{ik} r_{im} \boldsymbol{z}_i^{(m)}}{\sum_i a_{ik} r_{im}} + \rho\left(\frac{\sum_i a_{ik}\hat{g}_{im}}{\sum_i a_{ik}} - \frac{\sum_i a_{ik} r_{im}\hat{g}_{im}}{\sum_i a_{ik} r_{im}}\right). \tag{53}$$

This shows in particular that the result holds for every cluster-level $\hat{\pi}$.

Step 1 (limits of the four averages). Dividing each sum by n_k and applying the law of large numbers under P_k,

$$\tfrac{1}{n_k}\textstyle\sum_i a_{ik} r_{im} \boldsymbol{z}_i^{(m)} \to \mathbb{E}_k\big[R^{(m)} Z^{(m)}\big], \qquad \tfrac{1}{n_k}\textstyle\sum_i a_{ik} r_{im} \to \mathbb{E}_k\big[R^{(m)}\big] = \bar\pi_k^\star,$$
$$\tfrac{1}{n_k}\textstyle\sum_i a_{ik} \hat g_{im} \to \mathbb{E}_k\big[\hat g_m\big], \qquad \tfrac{1}{n_k}\textstyle\sum_i a_{ik} r_{im} \hat g_{im} \to \mathbb{E}_k\big[R^{(m)} \hat g_m\big].$$

Step 2 (evaluating the limits). By Eq. (51) and Theorem D.9(i) (which with Lemma C.4(ii) gives $\mathbb{E}[Z^{(m)} \mid Y, C] = f_m(C) = g^\star$),

$$\begin{aligned}\mathbb{E}_k\big[R^{(m)} Z^{(m)}\big] &= \mathbb{E}_k\Big[\mathbb{E}\big[R^{(m)} \mid Y, C, X^{(1:M)}, \mathbf{R}^{(-m)}\big]\, Z^{(m)}\Big] = \mathbb{E}_k\big[\pi^\star Z^{(m)}\big] \\ &= \mathbb{E}_k\big[\pi^\star\, \mathbb{E}[Z^{(m)} \mid Y, C]\big] = \mathbb{E}_k\big[\pi^\star g^\star\big], \\ \mathbb{E}_k\big[R^{(m)} \hat g_m\big] &= \mathbb{E}_k\big[\pi^\star \hat g_m\big],\end{aligned}$$

the second line because $\hat g_m$ is a function of $(X^{(-m)}, \mathbf{R}^{(-m)})$ and of independent training data.

Step 3 (assembling). By the continuous mapping theorem and Eq. (52),

$$\begin{aligned}\operatorname{plim} \hat{\boldsymbol{\mu}}_{km}^{\mathrm{dr}}(\rho) &= \frac{\mathbb{E}_k[\pi^\star g^\star]}{\bar\pi_k^\star} + \rho\Big(\mathbb{E}_k[\hat g_m] - \frac{\mathbb{E}_k[\pi^\star \hat g_m]}{\bar\pi_k^\star}\Big) \\ &= \mathbb{E}_k[g^\star] + \frac{\mathrm{Cov}_k(\pi^\star, g^\star)}{\bar\pi_k^\star} - \rho\,\frac{\mathrm{Cov}_k(\pi^\star, \hat g_m)}{\bar\pi_k^\star} \\ &= \boldsymbol{\mu}_{km} + (1-\rho)\,\frac{\mathrm{Cov}_k(\pi^\star, g^\star)}{\bar\pi_k^\star} + \rho\,\frac{\mathrm{Cov}_k(\pi^\star, g^\star) - \mathrm{Cov}_k(\pi^\star, \hat g_m)}{\bar\pi_k^\star} \\ &= \boldsymbol{\mu}_{km} + (1-\rho)\,B_{km}^{\mathrm{sel}} + \rho\,\frac{\mathrm{Cov}_k(\pi^\star, g^\star - \hat g_m)}{\bar\pi_k^\star},\end{aligned}$$

where $\mathbb{E}_k[g^\star] = \mathbb{E}_k[Z^{(m)}] - \mathbb{E}_k[\boldsymbol{\xi}^{(m)}] = \boldsymbol{\mu}_{km}$. The same computation with $\rho = 0$ shows $\mathbb{E}[Z^{(m)} \mid Y = k, R^{(m)} = 1] = \mathbb{E}_k[R^{(m)} Z^{(m)}]/\bar\pi_k^\star = \boldsymbol{\mu}_{km} + B_{km}^{\mathrm{sel}}$, which identifies B_{km}^{sel} as the selection bias of the visible mean. This proves (i).

(ii) Under (P), $\pi^\star = \pi_m(Y)$ is constant on $\{Y = k\}$, so both covariances vanish. Under (O), $(1-\rho) = 0$ and $g^\star - \hat g_m = \mathbf{0}$.

(iii) For any integrable V and any constant vector $\boldsymbol{c}$, by Jensen's and the Cauchy-Schwarz inequalities,

$$\begin{aligned}\big\|\mathrm{Cov}_k(\pi^\star, V)\big\| &= \big\|\mathbb{E}_k\big[(\pi^\star - \bar\pi_k^\star)(V - \mathbb{E}_k V)\big]\big\| \le \mathbb{E}_k\big[|\pi^\star - \bar\pi_k^\star|\, \|V - \mathbb{E}_k V\|\big] \\ &\le \mathrm{sd}_k(\pi^\star)\,\big(\mathbb{E}_k\|V - \mathbb{E}_k V\|^2\big)^{1/2} \le \mathrm{sd}_k(\pi^\star)\,\big(\mathbb{E}_k\|V - \boldsymbol{c}\|^2\big)^{1/2},\end{aligned}$$

since the mean minimizes the mean squared deviation. With $V = g^\star$, $\boldsymbol{c} = \boldsymbol{\mu}_{km}$ the last factor is $\mathrm{sd}_k(g^\star)$; with $V = g^\star - \hat g_m$, $\boldsymbol{c} = \mathbf{0}$ it is $\|\hat g_m - g^\star\|_{L^2(P_k)}$. Dividing by $\bar\pi_k^\star$ and applying the triangle inequality to (i) gives Eq. (37). ∎

Proof of Theorem D.16. Under (P), $\pi^\star \equiv \pi_k$ on $\{Y = k\}$, and by Eq. (51) and Lemma E.5 the indicators $\{r_{im}\}_{i:a_i=k}$ are independent $\mathrm{Bernoulli}(\pi_k)$ variables, independent of the embeddings and of the imputer inputs. Write $n = n_k$ and, for a fixed unit vector $\boldsymbol{u}$,

$$A_i = \boldsymbol{u}^\top\big(\boldsymbol{z}_i^{(m)} - \boldsymbol{\mu}_{km}\big), \qquad G_i = \boldsymbol{u}^\top\big(\hat g_m(\cdot_i) - \mathbb{E}_k \hat g_m\big), \qquad \Delta_i = \boldsymbol{u}^\top\big(\hat g_{im}^{[-\ell(i)]} - \hat g_m(\cdot_i)\big),$$

where $\hat g_m(\cdot_i)$ is the limit imputer evaluated at sample i and $\hat g_{im}^{[-\ell(i)]}$ the cross-fitted value actually used. The correction term in Eq. (53) is invariant to adding a constant to all imputations, so we may center them at $\mathbb{E}_k \hat g_m$.

Step 1 (linearization). Let $\bar r = \frac{1}{n}\sum_i r_{im} = \pi_k + O_P(n^{-1/2})$. Since $\mathbb{E}_k[R^{(m)} A] = \pi_k \mathbb{E}_k[A] = 0$, the sum $\frac{1}{n}\sum_i r_{im} A_i$ is $O_P(n^{-1/2})$ and

$$\frac{\sum_i r_{im} A_i}{\sum_i r_{im}} = \frac{1}{n\pi_k}\sum_i r_{im} A_i - \frac{\bar r - \pi_k}{\pi_k \bar r}\cdot\frac{1}{n}\sum_i r_{im} A_i = \frac{1}{n\pi_k}\sum_i r_{im} A_i + O_P(n^{-1}).$$

The same expansion applied to the imputations gives

$$\frac{1}{n}\sum_i (G_i+\Delta_i) - \frac{\sum_i r_{im}(G_i+\Delta_i)}{\sum_i r_{im}} = \frac{1}{n}\sum_i \Big(1-\frac{r_{im}}{\pi_k}\Big)G_i + \frac{1}{n}\sum_i \Big(1-\frac{r_{im}}{\pi_k}\Big)\Delta_i + O_P(n^{-1}) + o_P(n^{-1/2}),$$

where the $o_P(n^{-1/2})$ term is $(\bar r-\pi_k)\cdot \frac{1}{n}\sum_i r_{im}\Delta_i/(\pi_k\bar r) = O_P(n^{-1/2})\,o_P(1)$.

Step 2 (cross-fitting remainder). Within fold ℓ and conditionally on the data used to train $\hat g^{[-\ell]}$, the terms $(1-r_{im}/\pi_k)\Delta_i$ are independent, have mean $\mathbb{E}[1-R^{(m)}/\pi_k]\,\mathbb{E}[\Delta_i \mid \cdot] = 0$ because $R^{(m)}$ is independent of Δ_i, and have variance at most

$$\mathbb{E}\Big[\Big(1-\frac{R^{(m)}}{\pi_k}\Big)^2\Big]\,\mathbb{E}[\Delta_i^2 \mid \cdot] = \frac{1-\pi_k}{\pi_k}\,\big\|\boldsymbol{u}^\top(\hat g_m^{[-\ell]}-\hat g_m)\big\|^2_{L^2(P_k)} \to 0.$$

By Chebyshev's inequality, $\frac{1}{n}\sum_{i\in\mathcal{I}_\ell}(1-r_{im}/\pi_k)\Delta_i = o_P(n^{-1/2})$ for each of the finitely many folds.

Step 3 (central limit theorem). Combining Steps 1 and 2,

$$\sqrt{n}\,\boldsymbol{u}^\top\big(\hat{\boldsymbol{\mu}}^{\mathrm{dr}}_{km}(\rho)-\boldsymbol{\mu}_{km}\big) = \frac{1}{\sqrt{n}}\sum_i \varphi_i + o_P(1), \qquad \varphi_i = \frac{r_{im}}{\pi_k}\big(A_i-\rho G_i\big)+\rho\,G_i.$$

The φ_i are independent and identically distributed with $\mathbb{E}\varphi_i = \mathbb{E}[A_i-\rho G_i]+\rho\mathbb{E}G_i = 0$ and finite variance because $\mathbb{E}\|Z^{(m)}\|^2<\infty$ and $\hat g_m \in L^2(P_k)$. Using $\mathbb{E}[r_{im}] = \mathbb{E}[r_{im}^2] = \pi_k$ and the independence of r_{im} from (A_i,G_i),

$$\begin{aligned}
\mathbb{E}\varphi_i^2 &= \frac{1}{\pi_k}\mathbb{E}\big(A-\rho G\big)^2 + 2\rho\,\mathbb{E}\big[(A-\rho G)G\big] + \rho^2\,\mathbb{E}G^2\\
&= \frac{1}{\pi_k}\Big[\mathbb{E}A^2 - 2\rho\,\mathbb{E}AG + \rho^2\,\mathbb{E}G^2\Big] + 2\rho\,\mathbb{E}AG - 2\rho^2\,\mathbb{E}G^2 + \rho^2\,\mathbb{E}G^2\\
&= \frac{\mathbb{E}A^2}{\pi_k} - 2\rho\,\mathbb{E}AG\Big(\frac{1}{\pi_k}-1\Big) + \rho^2\,\mathbb{E}G^2\Big(\frac{1}{\pi_k}-1\Big)\\
&= \frac{\mathrm{Var}_k(\boldsymbol{u}^\top Z^{(m)})}{\pi_k} - \frac{1-\pi_k}{\pi_k}\Big[2\rho\,\mathrm{Cov}_k\big(\boldsymbol{u}^\top\hat g_m, \boldsymbol{u}^\top Z^{(m)}\big) - \rho^2\,\mathrm{Var}_k\big(\boldsymbol{u}^\top\hat g_m\big)\Big] = V_k(\rho;\boldsymbol{u}).
\end{aligned}$$

The Lindeberg-Lévy central limit theorem and Slutsky's lemma give Eq. (38).

Step 4 (optimal coefficient). Write $c = \mathrm{Cov}_k(\boldsymbol{u}^\top\hat g_m, \boldsymbol{u}^\top Z^{(m)})$ and $v = \mathrm{Var}_k(\boldsymbol{u}^\top\hat g_m) > 0$. Then $V_k(\rho;\boldsymbol{u})$ is a convex quadratic in ρ with $\partial_\rho V_k = -\frac{1-\pi_k}{\pi_k}(2c-2\rho v)$, minimized at $\rho^{\mathrm{opt}} = c/v$, and

$$\begin{aligned}
V_k(0;\boldsymbol{u}) - V_k(\rho^{\mathrm{opt}};\boldsymbol{u}) &= \frac{1-\pi_k}{\pi_k}\Big(\frac{2c^2}{v}-\frac{c^2}{v}\Big) = \frac{1-\pi_k}{\pi_k}\,\frac{c^2}{v}\\
&= \frac{1-\pi_k}{\pi_k}\,\mathrm{Var}_k(\boldsymbol{u}^\top Z^{(m)})\,\mathrm{Corr}_k^2\big(\boldsymbol{u}^\top\hat g_m, \boldsymbol{u}^\top Z^{(m)}\big).
\end{aligned}$$

If $\hat g_m$ is calibrated, $c = \mathrm{Cov}_k(\boldsymbol{u}^\top\hat g_m, \boldsymbol{u}^\top\hat g_m) + \mathrm{Cov}_k(\boldsymbol{u}^\top\hat g_m, \boldsymbol{u}^\top(Z^{(m)}-\hat g_m)) = v$, so $\rho^{\mathrm{opt}} = 1$ and the gain is $\frac{1-\pi_k}{\pi_k}v = \frac{1-\pi_k}{\pi_k}R_k^2\mathrm{Var}_k(\boldsymbol{u}^\top Z^{(m)})$. The oracle $g^\star$ is calibrated because $Z^{(m)}-g^\star = \boldsymbol{\xi}^{(m)}$ is independent of C and centered.

Step 5 (shrinkage). By Eq. (14), $\boldsymbol{\mu}_k^{(m)} - \hat{\boldsymbol{\mu}}^{\mathrm{dr}}_{km} = (1-s_{km})(\hat{\boldsymbol{\mu}}^{\mathrm{or}}_{km} - \hat{\boldsymbol{\mu}}^{\mathrm{dr}}_{km})$ with $0 \le 1-s_{km} = \frac{\hat\rho_m n_0}{n_{km}+\hat\rho_m n_0} \le \frac{n_0}{n_{km}} = O_P(n^{-1})$, since $n_{km}/n \to \pi_k > 0$, while $\hat{\boldsymbol{\mu}}^{\mathrm{or}}_{km} - \hat{\boldsymbol{\mu}}^{\mathrm{dr}}_{km} = O_P(1)$. Hence $\sqrt{n}\,(\boldsymbol{\mu}_k^{(m)} - \hat{\boldsymbol{\mu}}^{\mathrm{dr}}_{km}) = o_P(1)$, and Slutsky's lemma transfers Eq. (38) to the shrunk prototype. ■

Proof of Proposition D.17. *(i)* Let $\boldsymbol{m}_1 = \hat{\boldsymbol{\mu}}^{\mathrm{or}}_{km}$, $\boldsymbol{m}_2 = \hat{\boldsymbol{\mu}}^{\mathrm{ipw}}_{km}$ and

$$\boldsymbol{m}_3 = \frac{\sum_i a_{ik}(r_{im}/\hat\pi_{im})\,\hat g_{im}}{\sum_i a_{ik}(r_{im}/\hat\pi_{im})}.$$

Each of the three weighted sums in Eq. (40) has convex weights, and for convex weights c_i one has $\sum_i c_i\|\boldsymbol{v}_i-\boldsymbol{\mu}\|^2=\|\boldsymbol{\mu}-\bar{\boldsymbol{v}}\|^2+\sum_i c_i\|\boldsymbol{v}_i-\bar{\boldsymbol{v}}\|^2$ with $\bar{\boldsymbol{v}}=\sum_i c_i\boldsymbol{v}_i$. Hence

$$\begin{aligned}\hat{\mathcal{L}}^{\mathrm{dr}}_{km}(\boldsymbol{\mu}) &= \hat{\rho}_m\|\boldsymbol{\mu}-\boldsymbol{m}_1\|^2+\|\boldsymbol{\mu}-\boldsymbol{m}_2\|^2-\hat{\rho}_m\|\boldsymbol{\mu}-\boldsymbol{m}_3\|^2+\text{const}\\ &= \|\boldsymbol{\mu}\|^2-2\big\langle\boldsymbol{\mu},\ \boldsymbol{m}_2+\hat{\rho}_m(\boldsymbol{m}_1-\boldsymbol{m}_3)\big\rangle+\text{const}'=\big\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}^{\mathrm{dr}}_{km}\big\|^2+\text{const}'',\end{aligned}$$

because the quadratic coefficients add up to $\hat{\rho}_m+1-\hat{\rho}_m=1$. The unique minimizer is $\hat{\boldsymbol{\mu}}^{\mathrm{dr}}_{km}$. For the population counterpart, replace the averages by their limits. Under (P), $R^{(m)}$ is independent of $(Z^{(m)},\hat{g}_m)$ given $Y=k$, so the weighted averages converge to unweighted expectations and

$$\mathcal{L}^{\mathrm{dr}}_{km}(\boldsymbol{\mu})=\rho\,\mathbb{E}_k\|\hat{g}_m-\boldsymbol{\mu}\|^2+\mathbb{E}_k\|Z^{(m)}-\boldsymbol{\mu}\|^2-\rho\,\mathbb{E}_k\|\hat{g}_m-\boldsymbol{\mu}\|^2=\mathbb{E}_k\|Z^{(m)}-\boldsymbol{\mu}\|^2.$$

Under (O) with $\rho=1$ and $\hat{g}_m=g^\star$, using $\|Z^{(m)}-\boldsymbol{\mu}\|^2-\|g^\star-\boldsymbol{\mu}\|^2=\|Z^{(m)}-g^\star\|^2+2\langle Z^{(m)}-g^\star,g^\star-\boldsymbol{\mu}\rangle$ and Eq. (51),

$$\begin{aligned}\mathcal{L}^{\mathrm{dr}}_{km}(\boldsymbol{\mu}) &= \mathbb{E}_k\|g^\star-\boldsymbol{\mu}\|^2+\frac{\mathbb{E}_k\big[\pi^\star\,\mathbb{E}[\|\boldsymbol{\xi}^{(m)}\|^2+2\langle\boldsymbol{\xi}^{(m)},g^\star-\boldsymbol{\mu}\rangle\mid Y,C]\big]}{\bar{\pi}^\star_k}\\ &= \mathbb{E}_k\|g^\star-\boldsymbol{\mu}\|^2+\mathbb{E}\|\boldsymbol{\xi}^{(m)}\|^2=\mathbb{E}_k\|Z^{(m)}-\boldsymbol{\mu}\|^2.\end{aligned}$$

(ii) With $\lambda=\hat{\rho}_m n_0/n_{km}$ and (i), $\hat{\mathcal{L}}^{\mathrm{dr}}_{km}(\boldsymbol{\mu})+\lambda\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}^{\mathrm{or}}_{km}\|^2=\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}^{\mathrm{dr}}_{km}\|^2+\lambda\|\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}^{\mathrm{or}}_{km}\|^2+\text{const}$, whose unique minimizer is

$$\frac{\hat{\boldsymbol{\mu}}^{\mathrm{dr}}_{km}+\lambda\,\hat{\boldsymbol{\mu}}^{\mathrm{or}}_{km}}{1+\lambda}=s_{km}\,\hat{\boldsymbol{\mu}}^{\mathrm{dr}}_{km}+(1-s_{km})\,\hat{\boldsymbol{\mu}}^{\mathrm{or}}_{km},\qquad s_{km}=\frac{1}{1+\lambda}=\frac{n_{km}}{n_{km}+\hat{\rho}_m n_0}.$$

(iii) Under the stated conditions,

$$\mathbb{E}\big\|s\,(\hat{\boldsymbol{\mu}}^{\mathrm{dr}}_{km}-\boldsymbol{\mu}_{km})+(1-s)(\hat{\boldsymbol{\mu}}^{\mathrm{or}}_{km}-\boldsymbol{\mu}_{km})\big\|^2=s^2\,\frac{v_{km}}{n_{km}}+(1-s)^2\|\boldsymbol{b}_{km}\|^2,$$

whose derivative in s vanishes at $s^{\mathrm{opt}}_{km}=\|\boldsymbol{b}_{km}\|^2/(\|\boldsymbol{b}_{km}\|^2+v_{km}/n_{km})=n_{km}/(n_{km}+v_{km}/\|\boldsymbol{b}_{km}\|^2)$. Comparing with s_{km} gives the matching condition, and $s_{km}\to 1$ as $n_{km}\to\infty$ for fixed $\hat{\rho}_m n_0$. ■

E.7 Proofs for the Server Clustering (Section 3.5 and Appendix D.5)

Proof of Proposition D.18. Write $J=J_{\mathrm{feat}}(\{a_i\},\{\boldsymbol{\mu}\})+J_{\mathrm{mech}}(\{a_i\},\check{\pi})$ with

$$\begin{aligned}J_{\mathrm{feat}} &= \sum_{i=1}^{N}\frac{1}{2\sigma^2}\sum_m\omega_{im}\big\|\tilde{\boldsymbol{z}}_i^{(m)}-\boldsymbol{\mu}_{a_i}^{(m)}\big\|^2,\\ J_{\mathrm{mech}} &= \sum_{i=1}^{N}\Lambda_{ia_i}(\check{\pi})-\sum_{k,m}\log\mathrm{Beta}\big(\check{\pi}_m(k);1+\alpha\hat{\bar{\pi}}_m,1+\alpha(1-\hat{\bar{\pi}}_m)\big).\end{aligned}$$

Assignment block. For fixed $(\{\boldsymbol{\mu}\},\check{\pi})$, J is a sum over samples of terms that depend on a_i only (the Beta term does not depend on the assignments). Minimizing each term,

$$\begin{aligned}&\arg\min_k\Big[\frac{1}{2\sigma^2}\sum_m\omega_{im}\|\tilde{\boldsymbol{z}}_i^{(m)}-\boldsymbol{\mu}_k^{(m)}\|^2+\Lambda_{ik}\Big]\\ &\quad=\arg\min_k\frac{\sum_m\omega_{im}}{2\sigma^2}\Big[\Delta_{ik}+\frac{2\sigma^2\Lambda_{ik}}{\sum_m\omega_{im}}\Big]=\arg\min_k\tilde{\Delta}_{ik}\quad(\lambda_R=1).\end{aligned}$$

Mechanism block. For fixed assignments, $J_{\mathrm{mech}}=-\sum_{k,m}\psi_{km}(\check{\pi}_m(k))+\text{const}$ with ψ_{km} the log-posterior Eq. (49); it separates over (k,m), and by Lemma D.4(i) each term is minimized over $[\underline{\pi},1-\underline{\pi}]$ exactly by the clipped table Eq. (4).

Prototype block. For fixed assignments, $J_{\mathrm{feat}}=\frac{1}{2\sigma^2}\sum_{k,m}\sum_{i:a_i=k}\omega_{im}\|\tilde{\boldsymbol{z}}_i^{(m)}-\boldsymbol{\mu}_k^{(m)}\|^2$ separates over (k,m), and each term is a strictly convex quadratic minimized by the weighted mean

$\sum_{i:a_i=k} \omega_{im} \tilde{\boldsymbol{z}}_i^{(m)} / \sum_{i:a_i=k} \omega_{im}$ whenever the denominator is positive; otherwise J does not depend on $\boldsymbol{\mu}_k^{(m)}$ and any value, e.g., the previous one, is a minimizer.

Monotonicity and finite termination. Each block update is an exact minimization, so J is non-increasing along the iterates. Define $\bar{J}(\{a_i\}) = \min_{\{\boldsymbol{\mu}\},\check{\pi}} J(\{a_i\}, \{\boldsymbol{\mu}\}, \check{\pi})$; after the mechanism and prototype blocks, J equals $\bar{J}$ of the current partition because the two blocks decouple. If the assignment block changes at least one label, it does so only by a strict improvement (ties are broken in favor of the current label), hence for consecutive partitions $\mathcal{P}^t \neq \mathcal{P}^{t+1}$,

$$\bar{J}(\mathcal{P}^{t+1}) \leq J\big(\mathcal{P}^{t+1}, \{\boldsymbol{\mu}\}^t, \check{\pi}^t\big) < J\big(\mathcal{P}^t, \{\boldsymbol{\mu}\}^t, \check{\pi}^t\big) = \bar{J}(\mathcal{P}^t).$$

Thus no partition can occur twice, and since there are at most K^N partitions, the alternation stops after finitely many sweeps. At termination the assignments are a best response to $(\{\boldsymbol{\mu}\}, \check{\pi})$ and vice versa, which is a partial optimum of J. This is the classification EM algorithm of Celeux & Govaert (1992) for the working model of Proposition D.5 with a Beta prior on the table. With the doubly robust step, the prototype block minimizes the penalized doubly robust risk of Proposition D.17(ii) rather than J_{feat}; J need not be monotone then, which is why Algorithm 3 also caps the number of sweeps, while by Proposition D.17(i) the prototype block still targets the within-cluster risk $\mathbb{E}_k \|Z^{(m)} - \boldsymbol{\mu}\|^2$. ■

Proof of Corollary D.19. (i) If $\boldsymbol{R} \equiv 1$, then $n_{km} = n_k$ and $\hat{\bar{\pi}}_m = 1$, so $\hat{\pi}_m(k) = (n_k + \alpha)/(n_k + \alpha) = 1$ and $\hat{\pi}_{im} = 1$; after normalization $w_{im}^{\text{ipw}} = 1$, and Eq. (3) is the backbone loss. The clipped table is $\check{\pi}_m(k) = 1 - \underline{\pi}$ for all k, so $\Lambda_{ik} = -M \log(1 - \underline{\pi})$ does not depend on k and is dropped. No block is missing, so $\omega \equiv 1$ and $\tilde{\boldsymbol{z}} = \boldsymbol{z}$; by the convention after Eq. (14), $\hat{g}_{im} = \boldsymbol{z}_i^{(m)}$ and all prototype estimates equal the cluster mean. The server therefore runs multi-view k-means with Student's t targets, which is the backbone.

(ii) Under MCAR and $\max_k |\hat{\pi}_m(k) - \hat{\bar{\pi}}_m| \to 0$, the clipped table becomes constant in k, so the evidence does not affect assignments by Corollary D.6(i). The weights $1/\hat{\pi}_{im}$ and $(1 - \hat{\pi}_{im})/\hat{\pi}_{im}$ become constant in i and equal one after normalization. Finally $\pi_m^\star(Y, C) \equiv \bar{\pi}_m$ under MCAR, so $B_{km}^{\text{sel}} = \text{Cov}_k(\pi^\star, g^\star)/\bar{\pi}_k^\star = \mathbf{0}$. ■

E.8 Proofs for Communication and Privacy (Appendix G)

Proof of Proposition G.1. In round $t \geq 1$, client m uploads $\boldsymbol{Z}^{(m)} \in \mathbb{R}^{N_m \times d}$, $\boldsymbol{Q}^{(m)} \in \mathbb{R}^{N_m \times K}$ and $\boldsymbol{U}^{(m)} \in \mathbb{R}^{K \times d}$, i.e., $N_m(d + K) + Kd$ scalars, and downloads $\boldsymbol{P}^{(m)} \in \mathbb{R}^{N_m \times K}$, $\boldsymbol{w}^{(m)} \in \mathbb{R}^{N_m}$ and $\boldsymbol{\mu}^{(m)} \in \mathbb{R}^{K \times d}$, i.e., $N_m(K + 1) + Kd$ scalars. In round 0 it uploads only $\boldsymbol{Z}^{(m)}$ and downloads the same three objects. Summing,

$$\begin{aligned}\frac{\text{Comm}_{\texttt{CFIMC}}}{b} &= \sum_{m=1}^{M} \Big[T\big(N_m(d + K) + Kd + N_m(K + 1) + Kd\big) + N_m d + N_m(K + 1) + Kd\Big] \\ &= \sum_{m=1}^{M} \Big[T\big(N_m(d + 2K + 1) + 2Kd\big) + N_m(d + K + 1) + Kd\Big].\end{aligned}$$

The backbone sends the same messages without $\boldsymbol{w}^{(m)}$, i.e., N_m fewer scalars per client in each of the $T + 1$ rounds. Hence the relative overhead is

$$\begin{aligned}\frac{(T + 1)\sum_m N_m}{\sum_m \big[T\big(N_m(d + 2K) + 2Kd\big) + N_m(d + K) + Kd\big]} &< \frac{(T + 1)\sum_m N_m}{\sum_m N_m\big[T(d + 2K) + d + K\big]} \\ &= \frac{T + 1}{(T + 1)(d + K) + TK} < \frac{1}{d + K},\end{aligned}$$

where the first inequality drops the positive terms $(2T + 1)Kd$. For $T = 3$ and $d = K = 10$ the middle expression equals $4/110 < 3.7\%$. Since w_{im}^{ipw} depends on i only through a_i (see the proof of Proposition G.2), $\boldsymbol{w}^{(m)}$ is determined by $(\hat{\pi}_m(k))_{k \in [K]}$ and the labels $\{a_i\}_{i \in \mathcal{O}_m}$, which can be encoded with $\lceil \log_2 K \rceil$ bits each. None of the messages depends on D_m. ■

Proof of Proposition G.2. Client m knows $\mathcal{O}_m$, hence N_m, and N is public. Given its own labels, it can compute $n_{km} = |\{i \in \mathcal{O}_m : a_i = k\}|$ and $\hat{\bar{\pi}}_m = N_m/N$. By Eq. (4) and Eq. (3),

$$w_{im}^{\mathrm{ipw}} = \frac{1/\operatorname{clip}_{[\underline{\pi},1]}\big(\hat{\pi}_m(a_i)\big)}{\frac{1}{N_m}\sum_{j\in\mathcal{O}_m} 1/\operatorname{clip}_{[\underline{\pi},1]}\big(\hat{\pi}_m(a_j)\big)}, \qquad \hat{\pi}_m(k) = \frac{n_{km} + \alpha N_m/N}{n_k + \alpha},$$

so $\boldsymbol{w}^{(m)} = F\big(\{a_i\}_{i\in\mathcal{O}_m}, (n_k)_{k\in[K]}\big)$ for a deterministic map F that depends only on quantities known to client m and on public constants. Therefore

$$\begin{aligned} I\big(\boldsymbol{w}^{(m)}; \mathcal{D}_{-m} \,\big|\, \{a_i\}_{i\in\mathcal{O}_m}, \mathcal{D}_m\big) &\le H\big(\boldsymbol{w}^{(m)} \,\big|\, \{a_i\}_{i\in\mathcal{O}_m}, \mathcal{D}_m\big) \le H\big((n_k)_k \,\big|\, \{a_i\}_{i\in\mathcal{O}_m}, \mathcal{D}_m\big) \\ &\le H\big((n_k)_k\big) \le \log_2\big|\{0,1,\dots,N\}^K\big| = K\log_2(N+1), \end{aligned}$$

using, in order, $I(A; B \mid D) \le H(A \mid D)$ for discrete A, the fact that a deterministic function of $(n_k)_k$ given the conditioning cannot have larger entropy, the fact that conditioning does not increase entropy, and the uniform bound on the entropy of a variable with at most $(N+1)^K$ values. Finally, w_{im}^{ipw} depends on i only through a_i, so $w_{im} = w_{jm}$ whenever $a_i = a_j$, and the pattern $\boldsymbol{r}_i$ of an individual sample does not enter $\boldsymbol{w}^{(m)}$ except through the aggregate counts $(n_k)_k$. ■

Proof of Proposition G.3. Fix the encoder. The released object of one round is $\breve{\boldsymbol{Z}}^{(m)} = \mathbf{F}(\mathcal{D}_m) + \mathbf{E}$ with $\mathbf{F}(\mathcal{D}_m) = \big(\Pi_\Gamma(\mathrm{Enc}_m(\boldsymbol{x}_i^{(m)}))\big)_{i\in\mathcal{O}_m}$ and $\mathbf{E}$ a matrix of independent $\mathcal{N}(0, \sigma_{\mathrm{dp}}^2)$ entries. If $\mathcal{D}_m$ and $\mathcal{D}'_m$ differ in the record of one sample j, only the j-th row of $\mathbf{F}$ changes, and both rows lie in the ball of radius Γ, so the ℓ_2-sensitivity is

$$\big\|\mathbf{F}(\mathcal{D}_m) - \mathbf{F}(\mathcal{D}'_m)\big\|_F = \big\|\Pi_\Gamma(\boldsymbol{z}_j) - \Pi_\Gamma(\boldsymbol{z}'_j)\big\| \le 2\Gamma.$$

By the Gaussian mechanism (Dwork & Roth, 2014, Theorem A.1), noise with standard deviation $\sigma_{\mathrm{dp}} \ge 2\Gamma\sqrt{2\ln(1.25/\delta_{\mathrm{dp}})}/\varepsilon_{\mathrm{dp}}$ yields $(\varepsilon_{\mathrm{dp}}, \delta_{\mathrm{dp}})$-differential privacy for $\varepsilon_{\mathrm{dp}} \in (0,1)$. The soft assignments $\boldsymbol{Q}^{(m)}$ are computed from $\breve{\boldsymbol{Z}}^{(m)}$ and from server outputs, and every server computation (Modules I to III, clustering and the downloads) is a function of the released embeddings of all clients, of the mask and of public constants; by the post-processing property, none of them incurs additional privacy cost with respect to the records of client m. Over $T+1$ rounds, each with fresh noise, the adaptive basic composition theorem gives $\big((T+1)\varepsilon_{\mathrm{dp}}, (T+1)\delta_{\mathrm{dp}}\big)$. If the noisy embeddings are released once and reused, there is a single release and the cost is $(\varepsilon_{\mathrm{dp}}, \delta_{\mathrm{dp}})$. If the encoder itself is the output of an $(\varepsilon_0, \delta_0)$-differentially private training procedure such as DP-SGD (Abadi et al., 2016), composing this release with the above gives the stated sum of the privacy parameters. ■

F Detailed Algorithms and Default Hyperparameters

Algorithm 1 summarizes the federated training protocol of CFIMC. Algorithm 2 details one server aggregation (lines 11 to 19 of Algorithm 1), and Algorithm 3 details the evidence-aware doubly robust clustering that it calls. Table 4 lists the default hyperparameters. Each client uses the Input-500-500-2000-d autoencoder of Ren et al. (2024), and the server networks h_m and f_m of Module II are two-hidden-layer perceptrons of width d_h. Before any computation, the server standardizes each view block, $\boldsymbol{z}_i^{(m)} \leftarrow (\boldsymbol{z}_i^{(m)} - \bar{\boldsymbol{z}}^{(m)})/\varsigma_m$, with the visible mean $\bar{\boldsymbol{z}}^{(m)}$ and a scalar scale ς_m; this preserves the geometry of each view while preventing any client from dominating fused distances, and all quantities are mapped back to client coordinates before download.

Algorithm 1: CFIMC (federated training protocol)

Input: private views $\{\boldsymbol{x}_i^{(m)}\}_{i\in\mathcal{O}_m}$ at client $m \in [M]$; mask $\boldsymbol{R}$ at the server; K; rounds T; hyperparameters $\beta, \alpha, \underline{\pi}, \lambda_R, \lambda_{\text{inv}}, \lambda_{\text{ind}}, \tau, p_{\text{drop}}, L, n_0$
Output: cluster labels $\hat{y}_i = \arg\max_k \gamma_{ik}$ for all $i \in [N]$

Round 0: local pretraining and server initialization

1 **for** each client $m \in [M]$ **in parallel do**
2 pretrain $(\text{Enc}_m, \text{Dec}_m)$ by minimizing $\mathcal{L}_{\text{rec}}^{(m)}$ (Section 3.1); upload $\boldsymbol{Z}^{(m)}$ ▷ *raw data stay local*
3 **end for**
4 server: masked k-means++ on visible blocks; initial table $\hat{\pi}$; freeze $\sigma^2 \leftarrow \sigma_0^2$ ▷ M-I
5 server: execute lines 11 to 19; each client sets $\boldsymbol{U}^{(m)} \leftarrow \{\boldsymbol{\mu}_k^{(m)}\}_k$

Rounds $t = 1, \ldots, T$: selection-debiased self-training and causal aggregation

6 **for** $t = 1, \ldots, T$ **do**
7 **for** each client $m \in [M]$ **in parallel do**
8 update $(\text{Enc}_m, \text{Dec}_m, \boldsymbol{U}^{(m)})$ on $\mathcal{L}^{(m)}$ in Eq. (3) with weights $\boldsymbol{w}^{(m)}$ ▷ *IPW* M-I
9 upload $\boldsymbol{Z}^{(m)}, \boldsymbol{Q}^{(m)}, \boldsymbol{U}^{(m)}$
10 **end for**
11 standardize blocks; table $\hat{\pi}_m(k)$ by Eq. (4); weights $w_{im}^{\text{ipw}}, w_{im}^{\text{tr}}$ ▷ M-I
12 counterfactual imputation $\hat{\boldsymbol{z}}_i^{(m)}$ by Eq. (8), Eq. (9) and Eq. (32), L-fold cross-fitting ▷ M-II
13 reliability $\hat{\rho}_m$ by Eq. (10); completed blocks $\tilde{\boldsymbol{z}}_i^{(m)}$ and weights ω_{im} by Eq. (11) ▷ M-III
14 **repeat** ▷ *evidence-aware DR clustering*
15 $a_i \leftarrow \arg\min_k \tilde{\Delta}_{ik}^{\text{cf}}$ by Eq. (6) and Eq. (7) ▷ M-I
16 update $\hat{\pi}_m(k)$ by Eq. (4); update $\boldsymbol{\mu}_k^{(m)}$ by Eq. (12) to Eq. (14) ▷ M-III
17 **until** the assignments $\{a_i\}$ no longer change
18 γ_{ik} by Eq. (15); targets $\boldsymbol{P}$ by Eq. (21); align cluster indices with round $t-1$
19 download $\boldsymbol{\mu}^{(m)}$ (client coordinates), $\boldsymbol{P}^{(m)} = [\boldsymbol{p}_i]_{i\in\mathcal{O}_m}$, $\boldsymbol{w}^{(m)} = (w_{im}^{\text{ipw}})_{i\in\mathcal{O}_m}$ ▷ $+N_m$ *scalars*
20 **end for**
21 **return** $\hat{y}_i = \arg\max_k \gamma_{ik}$

Table 4: Default hyperparameters of CFIMC.

Clients and protocol			Server		
Symbol	Meaning	Value	Symbol	Meaning	Value
T	communication rounds	3	d_c	content dimension	16
T_{pre}	pretraining epochs	300	d_h	hidden width of h_m, f_m	64
T_{loc}	local epochs per round	150	T_{s}	server epochs	150
d	embedding dimension	20	$\lambda_{\text{inv}},\ \lambda_{\text{ind}}$	loss weights	1.0, 0.1
β	clustering weight	0.1	τ	InfoNCE temperature	0.5
Enc_m	encoder widths	$500, 500, 2000$	p_{drop}	view-dropout rate	0.5
$B_{\text{c}},\ \epsilon_{\text{c}}$	batch size, learning rate	$256,\ 10^{-3}$	L	cross-fitting folds	2
α	table pseudo-count	10	n_0	shrinkage strength	10
$\underline{\pi}$	propensity floor	0.05	λ_R	evidence weight	1.0
			n_{init}	seeding restarts	20

Algorithm 2: Server-side causal aggregation in round t

Input: uploads $\{\boldsymbol{Z}^{(m)}, \boldsymbol{Q}^{(m)}, \boldsymbol{U}^{(m)}\}_{m\in[M]}$ ($\boldsymbol{Z}^{(m)}$ only if $t = 0$); mask $\boldsymbol{R}$; previous soft assignment $\boldsymbol{\gamma}^{(t-1)}$ (if $t \geq 1$); hyperparameters of Table 4

Output: prototypes $\{\boldsymbol{\mu}_k^{(m)}\}$, targets $\boldsymbol{P}^{(m)}$ and weights $\boldsymbol{w}^{(m)}$ for every client m

standardize visible blocks, $\boldsymbol{z}_i^{(m)} \leftarrow (\boldsymbol{z}_i^{(m)} - \bar{\boldsymbol{z}}^{(m)})/\varsigma_m$

Module I: missingness mechanism

if $t = 0$ **then**
  $\{a_i\} \leftarrow$ masked k-means++ on visible blocks ($\omega_{im} = r_{im}$), n_{init} restarts ▷ *Eq. (41)*
  $\sigma_0^2 \leftarrow$ per-dimension within-cluster variance of $\{a_i\}$; keep it fixed for all rounds
else
  $a_i \leftarrow \arg\max_k \gamma_{ik}^{(t-1)}$
end if
table $\hat{\pi}_m(k)$, propensities $\hat{\pi}_{im}$ and clipped table $\check{\pi}_m(k)$ by Eq. (4) ▷ *M-I*
$w_{im}^{\text{ipw}} \propto 1/\hat{\pi}_{im}$, $w_{im}^{\text{tr}} \propto (1 - \hat{\pi}_{im})/\hat{\pi}_{im}$, each normalized to mean one over $\mathcal{O}_m$

Module II: counterfactual view imputation with cross-fitting

partition $[N]$ at random into folds $\mathcal{I}_1, \ldots, \mathcal{I}_L$
for $\ell = 1, \ldots, L$ **do**
  train $(h^{[-\ell]}, f^{[-\ell]})$ on $\{i \notin \mathcal{I}_\ell : |\mathcal{V}_i| \geq 2\}$ for T_{s} epochs on $\mathcal{L}_{\text{SCM}}$ ▷ *Eq. (27)*
    with view dropout p_{drop}, transport weights in $\mathcal{L}_{\text{cf}}$ and IPW weights in $\mathcal{L}_{\text{inv}}$ ▷ *M-II*
  predict $\hat{\boldsymbol{z}}_i^{(m)}$ by Eq. (32) for all $i \in \mathcal{I}_\ell$, $m \in [M]$ with $\sum_{v\neq m} r_{iv} \geq 1$
end for

Module III: transported reliability, fusion and clustering

$\hat{\rho}_m$ by Eq. (10) ($\hat{\rho}_m = 1$ if view m is complete); $\tilde{\boldsymbol{z}}_i^{(m)}$ and ω_{im} by Eq. (11) ▷ *M-III*
if $t = 0$ **then**
  initial prototypes by masked k-means++ on $(\tilde{\boldsymbol{z}}, \omega)$ with n_{init} restarts
else
  initial prototypes from the standardized client centers $\boldsymbol{U}^{(m)}$
end if
$(\{a_i\}, \{\boldsymbol{\mu}_k^{(m)}\}, \hat{\pi}) \leftarrow$ Algorithm 3

Global self-supervision and download

γ_{ik} and p_{ik} by Eq. (15); permute cluster indices to match $\boldsymbol{\gamma}^{(t-1)}$ (Hungarian)
de-standardize, $\boldsymbol{\mu}_k^{(m)} \leftarrow \varsigma_m \boldsymbol{\mu}_k^{(m)} + \bar{\boldsymbol{z}}^{(m)}$
return $\boldsymbol{\mu}^{(m)}$, $\boldsymbol{P}^{(m)} = [\boldsymbol{p}_i]_{i\in\mathcal{O}_m}$, $\boldsymbol{w}^{(m)} = (w_{im}^{\text{ipw}})_{i\in\mathcal{O}_m}$ for every m

Algorithm 3: Evidence-aware doubly robust clustering

Input: completed blocks $\tilde{\boldsymbol{z}}$, weights ω, mask $\boldsymbol{R}$, imputations $\hat{g}$, reliabilities $\hat{\rho}$, initial prototypes $\{\boldsymbol{\mu}_k^{(m)}\}$, frozen σ_0^2, α, $\underline{\pi}$, λ_R, n_0, maximal sweeps $I_{\max}$

Output: assignments $\{a_i\}$, prototypes $\{\boldsymbol{\mu}_k^{(m)}\}$, table $\hat{\pi}$

split $[N]$ uniformly at random into halves $\mathcal{H}_0, \mathcal{H}_1$
for $s = 1, \ldots, I_{\max}$ **do**
  for $b \in \{0, 1\}$ **do** ▷ *cross-fitted evidence* *M-I*
    estimate $\hat{\pi}_m^{[1-b]}(k)$ by Eq. (4) from the current labels of $\mathcal{H}_{1-b}$
    Λ_{ik}^{cf} by Eq. (5) with $\hat{\pi}^{[1-b]}$ for all $i \in \mathcal{H}_b$
  end for
  $a_i \leftarrow \arg\min_k \tilde{\Delta}_{ik}^{\text{cf}}$ by Eq. (6) and Eq. (7), ties broken in favor of the current label
  if no label changed and $s > 1$ **then break**
  end if
  re-seed empty clusters with the samples farthest from their prototypes
  update $\hat{\pi}_m(k)$ and $\hat{\pi}_{im}$ by Eq. (4) on all samples ▷ *M-I*
  for $k \in [K]$, $m \in [M]$ **do** ▷ *doubly robust prototypes* *M-III*
    $\hat{\boldsymbol{\mu}}_{km}^{\text{ipw}}, \hat{\boldsymbol{\mu}}_{km}^{\text{or}}, \hat{\boldsymbol{\mu}}_{km}^{\text{dr}}$ by Eq. (12) to Eq. (13); $\boldsymbol{\mu}_k^{(m)} \leftarrow (1 - s_{km})\hat{\boldsymbol{\mu}}_{km}^{\text{or}} + s_{km}\hat{\boldsymbol{\mu}}_{km}^{\text{dr}}$ ▷ *Eq. (14)*
  end for
end for
return $\{a_i\}$, $\{\boldsymbol{\mu}_k^{(m)}\}$, $\hat{\pi}$

G Communication and Privacy Analysis

This appendix analyzes the communication and computation costs of CFIMC (Appendix G.1) and what its messages reveal (Appendix G.2). Proofs are given in Appendix E.8.

G.1 Communication Cost Analysis

Proposition G.1 (Communication cost). *In every round $t \geq 1$, client m uploads $N_m(d+K)+Kd$ scalars $(\boldsymbol{Z}^{(m)}, \boldsymbol{Q}^{(m)}, \boldsymbol{U}^{(m)})$ and downloads $N_m(K+1)+Kd$ scalars $(\boldsymbol{P}^{(m)}, \boldsymbol{w}^{(m)}, \boldsymbol{\mu}^{(m)})$; in round 0 it uploads $N_m d$ and downloads $N_m(K+1)+Kd$. With b bits per scalar, the total cost over T rounds is*

$$\mathrm{Comm}_{\texttt{CFIMC}} = b\sum_{m=1}^{M}\Big[T\big(N_m(d+2K+1)+2Kd\big)+N_m(d+K+1)+Kd\Big], \tag{54}$$

and the overhead relative to the same backbone without $\boldsymbol{w}^{(m)}$ satisfies

$$\frac{\mathrm{Comm}_{\texttt{CFIMC}}-\mathrm{Comm}_{\mathrm{backbone}}}{\mathrm{Comm}_{\mathrm{backbone}}} < \frac{T+1}{(T+1)(d+K)+TK} < \frac{1}{d+K}, \tag{55}$$

i.e., under 3.7% for $T=3$ and $d=K=10$. Since $\boldsymbol{w}^{(m)}$ takes at most K distinct values, it can also be sent as the vector $(\hat{\pi}_m(k))_{k\in[K]}$ plus $\lceil\log_2 K\rceil$-bit indices. The cost does not depend on the raw dimensions D_m.

Computation. Clients run the same computation as the backbone. On the server, the structural causal model costs $O\big(L\,T_{\mathrm{s}}\,MN_2(d_h^2+d_h(d+d_c))+L\,T_{\mathrm{s}}\,MN_2B(d+d_c)\big)$ per round, where N_2 is the number of samples with at least two views, T_{s} the number of server epochs and B the batch size (the last term comes from the $B\times B$ InfoNCE and Gram matrices); the alternation costs $O(I_{\mathrm{cl}}NKMd)$ for I_{cl} sweeps, and seeding costs $O(n_{\mathrm{init}}(2+\ln K)KNMd)$. All terms are linear in N and independent of D_m.

G.2 Privacy Analysis

As in FedDMVC and FCUIF, raw features never leave the clients, and the server knows the mask $\boldsymbol{R}$ because it aligns sample identifiers; CFIMC uses $\boldsymbol{R}$ only through this existing channel. The only new message is the weight vector $\boldsymbol{w}^{(m)}$. The next result bounds what it reveals.

Proposition G.2 (Leakage of the additional download). *Fix a round. Conditionally on the global hard labels $\{a_i\}_{i\in\mathcal{O}_m}$ of its own samples and on the public constants $(N,\alpha,\underline{\pi})$, the vector $\boldsymbol{w}^{(m)}$ is a deterministic function of the cluster sizes $(n_k)_{k\in[K]}$. Consequently, with $\mathcal{D}_{-m}$ the information not held by client m (other views and other patterns),*

$$I\big(\boldsymbol{w}^{(m)};\mathcal{D}_{-m}\,\big|\,\{a_i\}_{i\in\mathcal{O}_m},\mathcal{D}_m\big)\leq H\big((n_k)_{k\in[K]}\big)\leq K\log_2(N+1)\ \textit{bits}, \tag{56}$$

and $w_{im}=w_{jm}$ whenever $a_i=a_j$: the download reveals nothing about which other views an individual sample possesses beyond its cluster.

Proposition G.3 (Compatibility with differential privacy). *Consider the variant in which client m freezes its encoder after pretraining, uploads $\breve{\boldsymbol{z}}_i^{(m)}=\Pi_\Gamma(\boldsymbol{z}_i^{(m)})+\mathcal{N}(\boldsymbol{0},\sigma_{\mathrm{dp}}^2\boldsymbol{I}_d)$ with Π_Γ the projection onto the ball of radius Γ and $\sigma_{\mathrm{dp}}=2\Gamma\sqrt{2\ln(1.25/\delta_{\mathrm{dp}})}/\varepsilon_{\mathrm{dp}}$, $\varepsilon_{\mathrm{dp}}\in(0,1)$, and computes $\boldsymbol{Q}^{(m)}$ from $\breve{\boldsymbol{z}}^{(m)}$ and the downloaded prototypes. Then, conditionally on the frozen encoder, each round is $(\varepsilon_{\mathrm{dp}},\delta_{\mathrm{dp}})$-differentially private with respect to replacing one record of client m; every server computation of CFIMC (Modules I to III, clustering and all downloads) is post-processing; and $T+1$ rounds are $\big((T+1)\varepsilon_{\mathrm{dp}},(T+1)\delta_{\mathrm{dp}}\big)$-differentially private by basic composition (Dwork & Roth, 2014). If the noisy embeddings are released once and reused in all rounds, the total cost is $(\varepsilon_{\mathrm{dp}},\delta_{\mathrm{dp}})$; if the encoder is pretrained with DP-SGD (Abadi et al., 2016) at level (ε_0,δ_0), the guarantees add.*

Remark G.4 (Scope of the guarantees). Latent ignorability excludes self-masking, in which the value of a view causes its own absence given the semantics; no method that uses only $(\mathbf{R}, X^{(\mathcal{V})})$ can identify such mechanisms without further assumptions (Mohan & Pearl, 2021). Condition (P) is needed only by the propensity-based parts, and Theorem D.14 shows how the doubly robust prototypes degrade gracefully when it fails.

H Additional Experimental Details and Results

H.1 Datasets

Table 5 lists the datasets.

Table 5: Statistics of the datasets.

Dataset	Content	Samples	Views	Clusters	View dimensions
MSRC-v1	object images	210	5	7	24, 576, 512, 256, 254
Caltech-5V	object images	1,400	5	7	40, 254, 928, 512, 1,984
Scene-15	scene images	4,485	3	15	20, 59, 40
LandUse-21	aerial images	2,100	3	21	20, 59, 40
3V-Fashion	product images	10,000	3	10	784, 784, 784
100leaves	plant leaves	1,600	3	100	64, 64, 64
DryBean	bean measurements	13,611	2	7	12, 4
Hdigit	handwritten digits	10,000	2	10	784, 256
Synthetic	simulated	2,000	3	10	100, 150, 125

H.2 Missingness Mechanisms

Under MNAR, $\pi^{\star}_m(Y, C) = \pi_m(Y)$, so Condition (P) holds. Under MNAR_latent, $\pi^{\star}_m$ varies with C inside a cluster, so (P) fails (Remark C.5).

H.3 Additional Results

Table 6 averages the synthetic results over the missing rates, and Figures 7 to 9 show the curves on the six real datasets that Section 4.3 does not plot.

H.4 Ablation over Communication Rounds

H.5 Sensitivity Grids and Results on MSRC-v1

Table 7 lists the grids. We run every setting with three seeds, and the default-value run is shared across the four sweeps. Figure 11 shows the results on MSRC-v1.

H.6 Oracle Audit Protocol

Since missingness is simulated, we audit the imputations of a trained `CFIMC` against the hidden views (MSRC-v1 and Hdigit, both mechanisms, $\eta = 0.5$). A block $\boldsymbol{b}$ in view m enters Δ_{ik} through $\|\boldsymbol{b} - \boldsymbol{\mu}_k^{(m)}\|_2^2$ (Eq. (6)), so it votes for its nearest prototype. For each missing entry we compare the prototype of the current cluster, the cluster-level imputation of FCUIF, the counterfactual imputation $\hat{\boldsymbol{z}}_i^{(m)}$, the visible mean, and an oracle block that encodes the hidden view with the final client encoder (Appendix H.6). By Lemma D.8(ii), the current prototype always votes for the current cluster, so it is right on every correctly assigned sample and wrong on every misassigned one. The informative comparison is thus on misassigned samples, where a block helps only by voting for the true cluster.

The audit trains the full `CFIMC` once per dataset, mechanism, and seed (three seeds) without changing its code, and a probe records the final-round quantities while the models are in memory. All comparisons use the standardized server coordinates of Algorithm 2: each view block is centered by

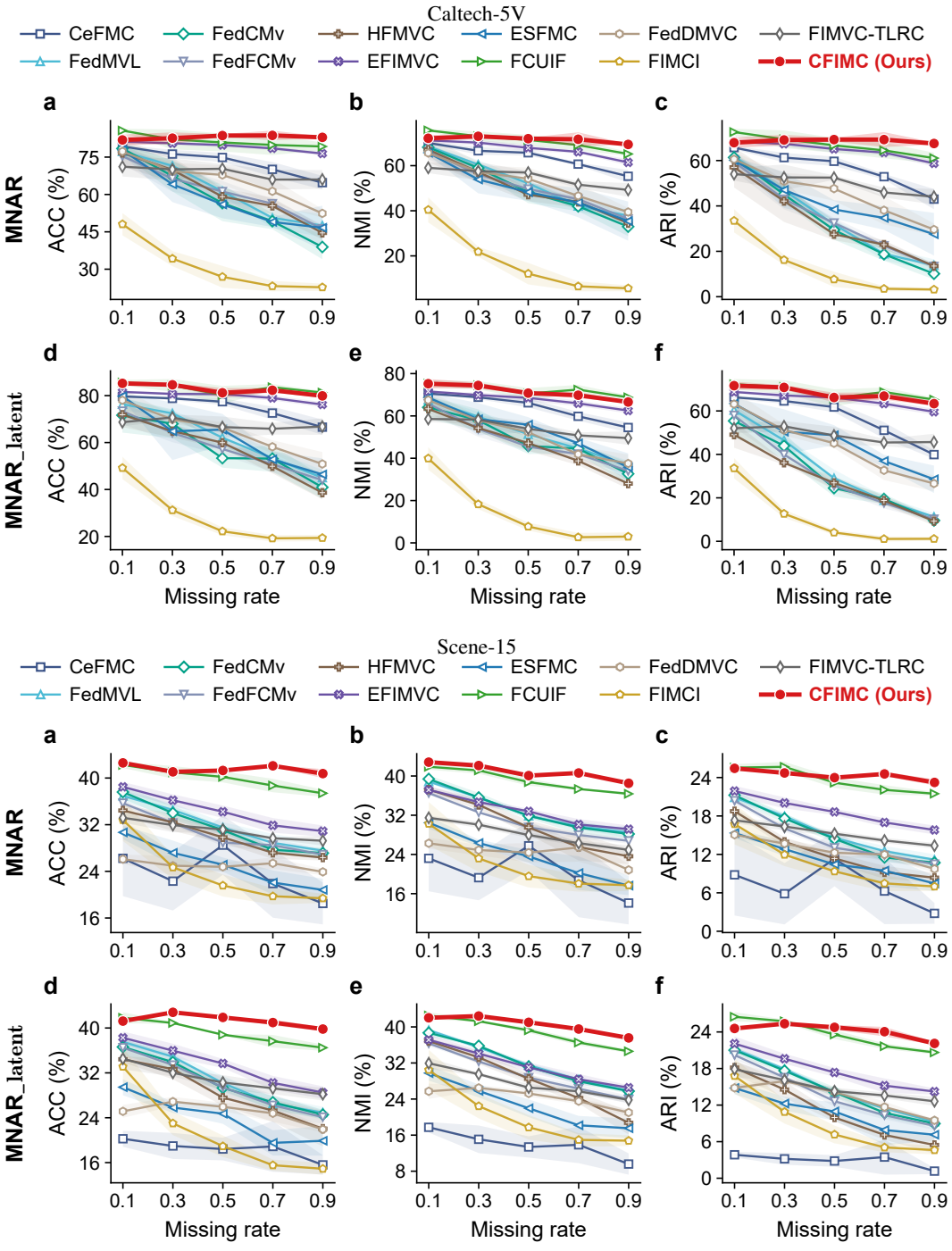

Figure 7: Clustering performance (%) against the missing rate on Caltech-5V (top) and Scene-15 (bottom). Within each panel, the rows are MNAR and MNAR_latent. Lines are means over five seeds and bands show ± one standard deviation.

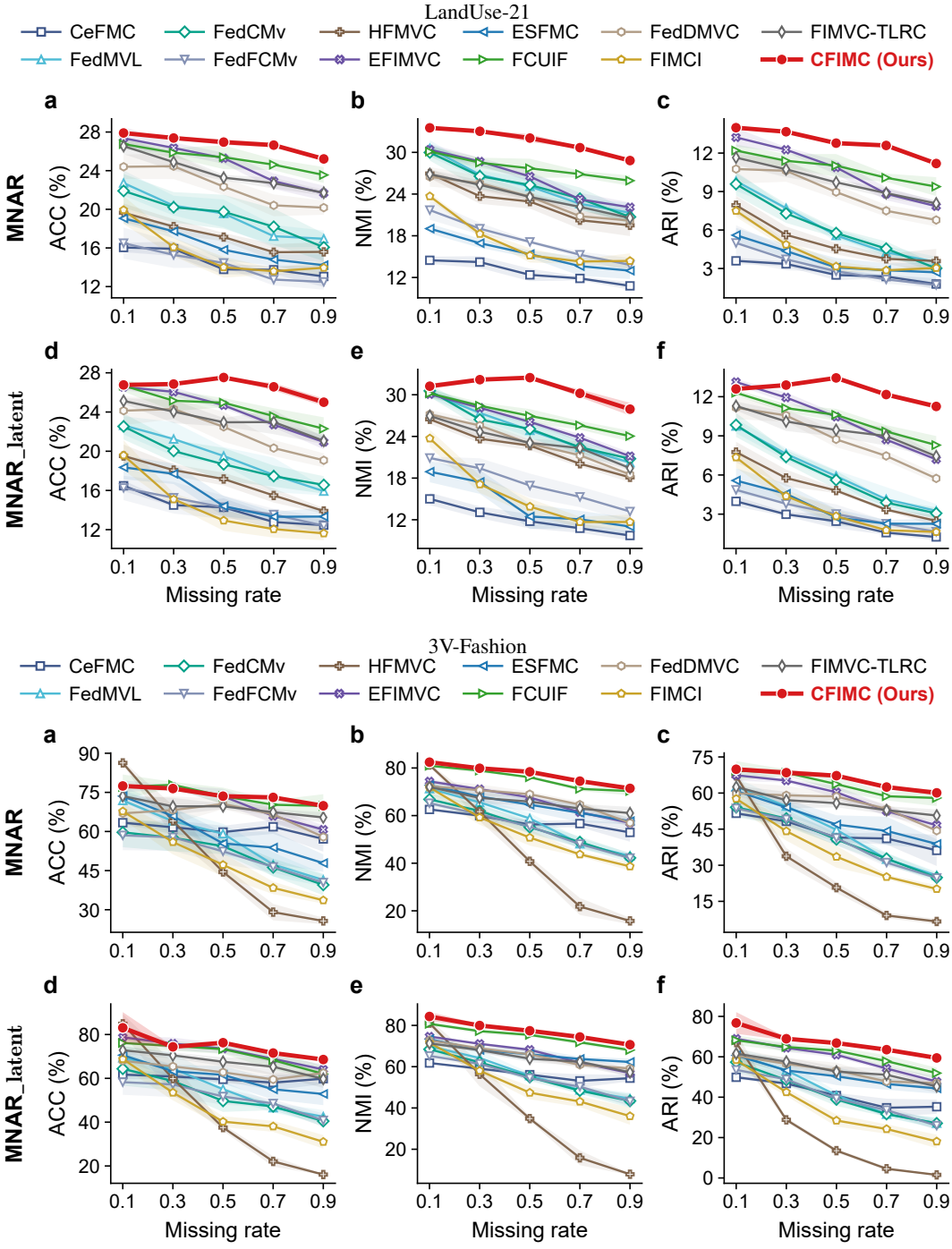

Figure 8: Clustering performance (%) against the missing rate on LandUse-21 (top) and 3V-Fashion (bottom), in the layout of Figure 7.

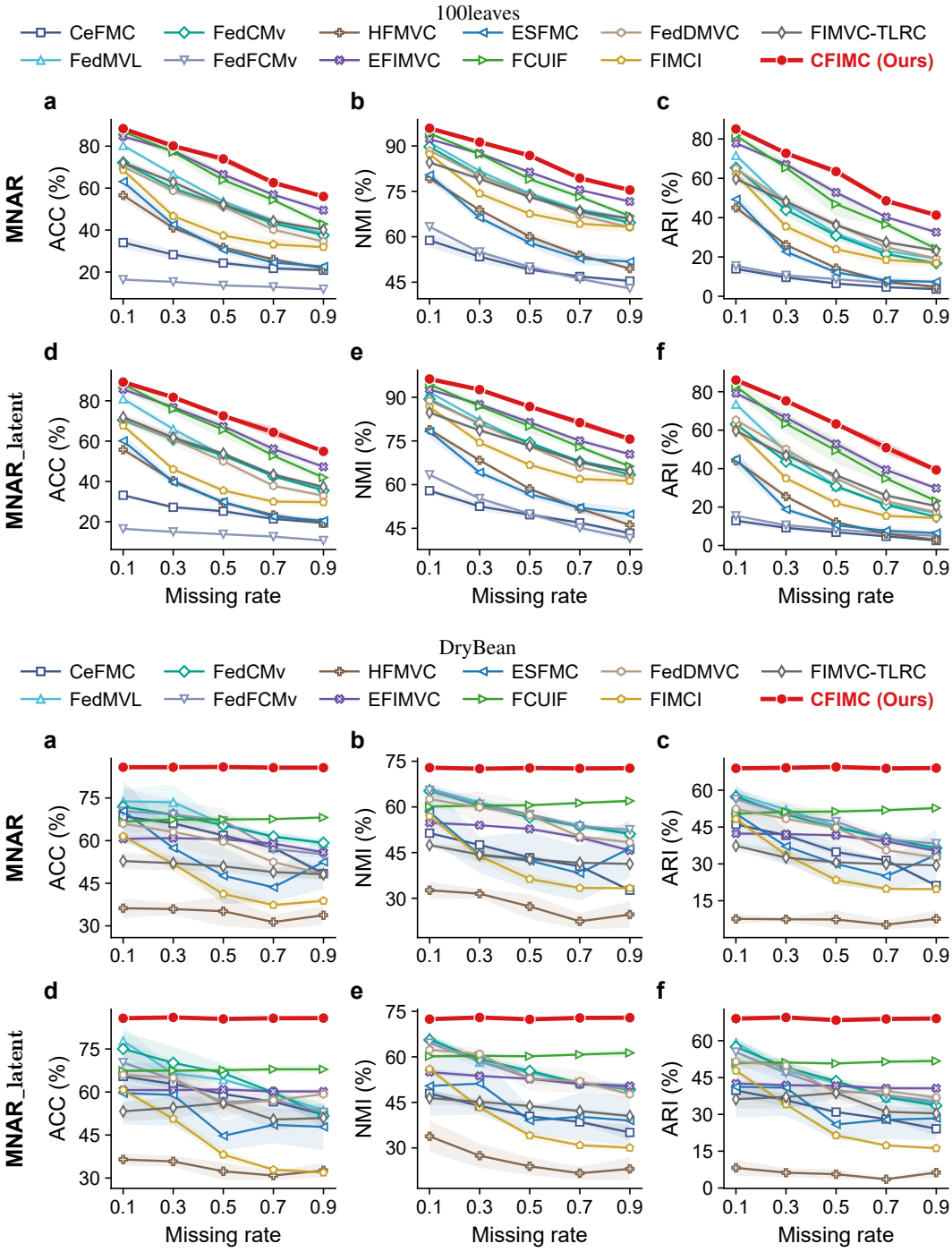

Figure 9: Clustering performance (%) against the missing rate on 100leaves (top) and DryBean (bottom), in the layout of Figure 7.

Table 6: Clustering performance (%) on the synthetic dataset under three missing mechanisms, averaged over selected seeds × 5 missing rates (mean±std). The best results are in **bold** and the second best are underlined.

Method	MNAR			MNAR_latent		
	ACC	NMI	ARI	ACC	NMI	ARI
CeFMC	80.28±4.69	69.36±6.26	64.96±8.83	79.65±4.70	69.19±6.08	63.54±10.11
FedMVL	66.64±16.02	65.49±12.86	53.00±19.84	65.53±13.58	64.44±11.85	50.88±17.53
FedCMv	60.31±24.53	60.44±20.63	48.09±27.32	57.64±24.39	58.55±20.56	46.92±26.47
FedFCMv	83.67±9.44	75.58±10.42	67.72±17.13	81.11±11.21	73.42±11.42	63.84±18.97
HFMVC	65.52±16.56	60.08±14.64	41.20±21.70	62.68±18.26	57.07±15.97	38.85±21.89
EFIMVC	88.01±2.55	76.06±4.36	76.24±4.65	88.09±2.59	76.15±4.33	76.39±4.70
ESFMC	44.88±10.77	36.18±10.31	26.23±10.85	43.69±14.04	34.56±12.97	25.34±13.64
FCUIF	88.19±4.05	76.85±6.30	76.13±7.55	87.29±5.83	75.72±7.94	75.50±8.99
FedDMVC	72.63±14.65	60.63±14.91	56.64±18.20	68.93±17.93	56.90±17.82	51.96±22.14
FIMCI	29.64±9.96	18.31±10.10	13.03±10.00	25.86±12.98	15.34±12.03	10.62±11.60
FIMVC-TLRC	83.85±4.10	71.09±5.60	70.19±6.42	84.46±4.31	71.71±5.82	70.96±6.69
CFIMC (Ours)	**91.27**±2.71	**81.99**±4.87	**82.30**±5.21	**90.81**±2.91	**81.00**±5.28	**81.40**±5.55

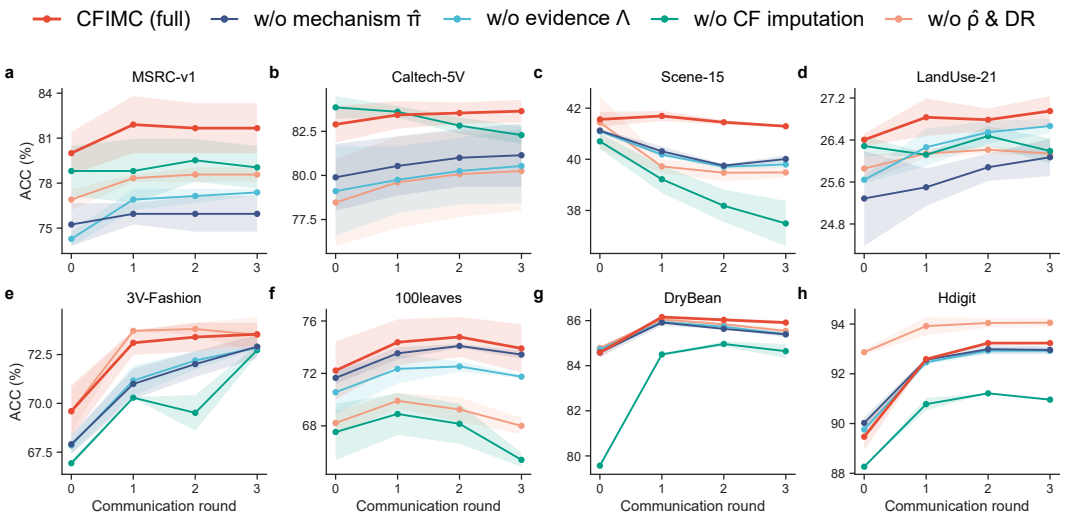

Figure 10: ACC (%) of the ablation variants after each communication round (MNAR, $\eta = 0.5$). Lines are means over five seeds, the same for every variant, and bands show ± one standard deviation.

Table 7: Grids of the sensitivity analysis. Defaults (Table 4) are underlined.

Parameter	Role	Grid
λ_R	evidence weight, Module I	0, 0.125, 0.25, 0.375, 0.5, 0.75, 1, 1.5, 2, 3, 4
α	table pseudo-count, Module I	1, 2, 3, 5, 10, 20, 30, 50, 100
λ_{inv}	InfoNCE weight, Module II	0, 0.1, 0.2, 0.3, 0.5, 1, 2, 3, 5, 10
n_0	shrinkage strength, Module III	0, 1, 2, 3, 5, 10, 20, 30, 50, 100

its visible mean and divided by the square root of its mean per-dimension variance, both recomputed from the final visible embeddings. The five blocks for a missing entry (i, m) are defined as follows.

1. Current prototype: $\boldsymbol{\mu}_{\hat{y}_i}^{(m)}$, the case of Lemma D.8(ii).

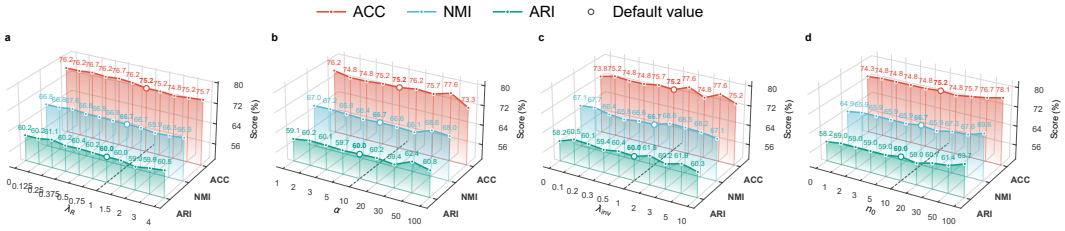

Figure 11: Sensitivity on MSRC-v1 under MNAR with $\eta = 0.5$, in the layout of Figure 6.

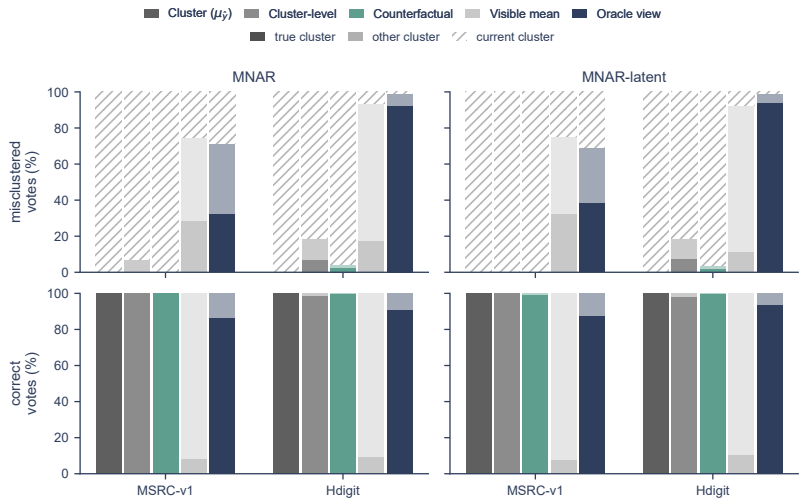

Figure 12: Votes of the imputed blocks on MSRC-v1 and Hdigit ($\eta = 0.5$, three seeds): shares of missing entries whose block is nearest to the prototype of the true cluster, of the current wrong cluster, or of another cluster, for misassigned (top) and correctly assigned (bottom) samples.

2. Cluster-level imputation: $\sum_k \bar{q}_{ik} \boldsymbol{\mu}_k^{(m)}$, the form used by FCUIF, where $\bar{\boldsymbol{q}}_i$ averages the soft assignments of the visible views of sample i, each recomputed with the final centers of its client.
3. Counterfactual imputation: the last-round cross-fitted prediction $\hat{\boldsymbol{z}}_i^{(m)}$ used by `CFIMC`.
4. Visible mean: the zero vector in standardized coordinates.
5. Oracle: the hidden view $\boldsymbol{x}_i^{(m)}$ encoded by the final encoder of client m.

Three limits apply. The oracle block contains the out-of-sample error of the encoder, which was trained on visible samples only. All quantities come from the last communication round. The audit needs the hidden views, so it applies only to simulated missingness.