ONLINE MISSING-UPDATE CORRECTION UNDER INFORMATIVE CLIENT PARTICIPATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-device federated learning relies on partial participation, yet client availability can depend on prior activity, staleness, and optimization state. The returned updates can therefore be an informative rather than representative subset, causing standard averaging to favor repeatedly available clients. Existing methods typically correct participation probabilities or reconstruct missing updates; either correction can become misspecified as availability and client trajectories drift. We introduce FEDDORA, a server-side framework that treats participation as sequential missingness and combines leakage-free online propensity estimation with vector-valued update prediction. A population-normalized orthogonal aggregate couples these two nuisance models: under history-conditional ignorability and overlap, either a correct propensity or update model removes conditional aggregation bias, while joint misspecification enters through a product remainder. We establish an exact bias identity, an explicit variance formula under conditional client independence, and a general non-convex convergence bound. Across CIFAR-10, CIFAR-100, and OfficeHome, controlled observable and latent participation mechanisms, residual adaptation, and oracle interventions, FEDDORA achieves competitive endpoint performance while providing strong tail and mechanism-level evidence for informative-participation correction. Ablations and hidden-driver tests further delineate its identification boundary.

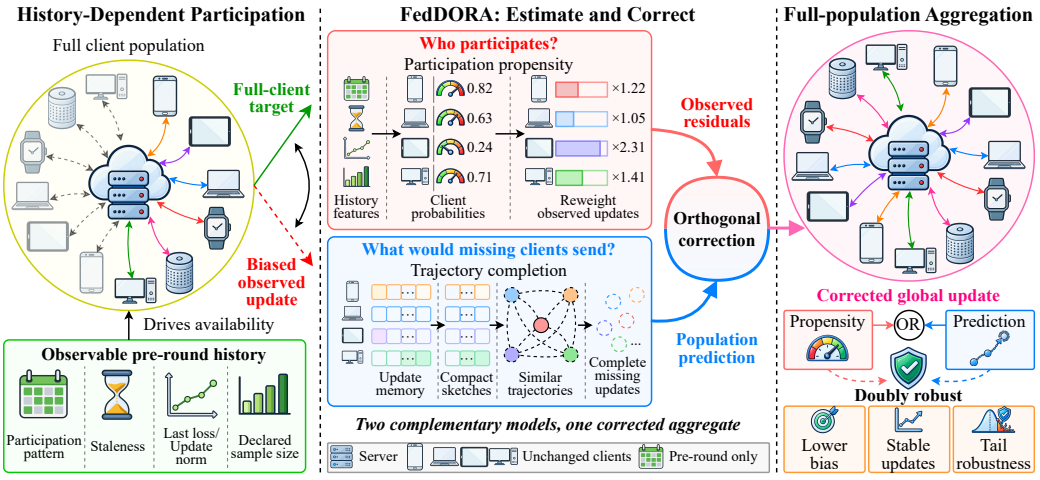

Figure 1: FEDDORA **at a glance.** History-dependent participation makes the returned updates an informative subset of the full client population. Using only the observable pre-round history shown on the left, FEDDORA estimates (i) participation propensities and (ii) missing client trajectories, then applies a propensity-aware residual correction to the population prediction. Under Assumption 1, either correct nuisance removes conditional first-order bias. The displayed gauges and weights are schematic rather than measured results; all operations are server-side and clients remain unchanged.

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.

1 INTRODUCTION

Federated learning (FL) trains a shared model without collecting raw client data (McMahan et al., 2017), but cross-device systems can observe only a changing subset of clients at each round. Standard aggregation implicitly treats the returned updates as representative of the intended population. That premise fails when availability depends on recurring schedules, past participation, staleness, or the most recently reported optimization state. As Figure 1 illustrates, the server can then optimize for repeatedly available clients rather than the full client population, even when every local update is computed correctly.

Existing approaches attack one side of this selection problem. Participation-aware methods estimate weights from participation intervals or stationary frequencies (Wang & Ji, 2024; Sun et al., 2025), while unavailability- and memory-aware methods reuse stored innovations, client updates, or historical models (Jhunjhunwala et al., 2022; Xiang et al., 2024; Wang et al., 2024). The first route remains sensitive to misspecified propensities and high inverse-weight variance; the second substitutes stale or predicted updates whose error can be systematic as client trajectories drift. A single-model correction can therefore exchange selection bias for model bias. What is missing is an online aggregate that learns both *who is observed* and *what absent clients would have contributed*, without allowing the current observation event to leak into its own correction.

Our key insight is that the two imperfect views can protect one another through orthogonal residual correction. We formulate informative participation as sequential optimization with missing vector updates: the server targets the full-client update, but each potential update is hidden whenever its client is unavailable. Let $A_{i,t}$ indicate participation, $\Delta_{i,t}$ the potential update, $\pi_{i,t}$ its pre-round propensity, and $m_{i,t}$ its conditional mean. The augmented score

$$\psi_{i,t}(\hat{\pi}, \hat{m}) = \hat{m}_{i,t} + \frac{A_{i,t}}{\hat{\pi}_{i,t}}(\Delta_{i,t} - \hat{m}_{i,t}) \tag{1}$$

is Neyman-orthogonal: its conditional bias is a product of propensity and prediction errors, so either correct nuisance removes first-order selection bias. Classical AIPW supplies this statistical primitive (Chernozhukov et al., 2018; Kurz, 2022); making it valid for evolving FL is not a plug-in exercise. The missing outcome is a high-dimensional vector whose law changes after every global update, both nuisances must be learned online from censored histories, same-round state would reveal the observation event being corrected, and vector aggregation error must ultimately be propagated through optimization.

FEDDORA resolves these requirements with three server-side modules (Figure 2). A pre-round online learner estimates history-conditional participation; a temporal and cross-client router predicts missing updates in the original parameter space while using random-projection sketches only for neighbor search; and a canonical N^{-1}-normalized augmented aggregate combines population predictions with propensity-weighted observed residuals. Five in-budget FedAvg rounds initialize shared histories, while an explicit pre-round prototype handles clients still unseen afterward. Clients run unchanged local SGD and return the same single update as in FedAvg.

We evaluate the correction mechanism rather than relying only on endpoint accuracy. Eight cited methods and four nuisance/oracle controls are compared across three representation benchmarks, residual adaptation, observable and hidden participation mechanisms, and a small raw-pixel scope check. Controlled selection is necessary here because oracle propensities and full-client counterfactuals needed to test identification cannot be recovered from ordinary deployment logs. Propensity error, update-residual error, their product, gradient-bias proxies, tail accuracy, and calibration connect the empirical analysis directly to the theory. The primary evidence remains controlled and representation-level; the raw-pixel experiment is only an execution scope check.

Our contributions are:

- We formalize informative participation as sequential missing-update optimization over a drifting vector estimand, specify the pre-round information boundary required for identification, and give a constructive non-identifiability result for unlogged selection drivers.
- We develop the server-side FEDDORA estimator by combining leakage-free online propensity learning, cold-start-safe full-dimensional trajectory completion, and population-normalized orthogonal aggregation. For the implemented estimator, we derive an exact bias identity, a conditional-

independence variance formula, a routing-error condition, and a general non-convex convergence bound with a nuisance-product penalty.
- We design a mechanism-oriented evaluation across three benchmarks, eight cited methods, and four nuisance/oracle controls. Tail and calibration metrics, observable-versus-hidden mechanisms, routing ablations, and overhead and raw-pixel checks jointly test both the predicted correction behavior and its scope boundaries.

2 Related Work

Optimization and partial-participation anchors. FedAvg is the standard averaging baseline (McMahan et al., 2017); SCAFFOLD corrects client drift with control variates (Karimireddy et al., 2020); and FedVARP reduces sampling variance with client memories (Jhunjhunwala et al., 2022). Complementary work exploits otherwise unused non-aligned samples through cross-institutional representation augmentation in vertical healthcare FL (Yin et al., 2024); our horizontal setting instead corrects the selection bias of missing client updates. The three optimization methods are retained as established anchors, not presented as recent state of the art.

Recent availability-aware baselines. FedAU (ICLR 2024) estimates inverse participation weights from cutoff participation intervals (Wang & Ji, 2024). FedAWE (NeurIPS 2024) uses staleness-scaled innovation echoing and persistent client states (Xiang et al., 2024). Debiasing FedAvg (ICLR 2025) rescales local updates by empirical stationary frequencies under correlated participation (Sun et al., 2025). Our harness follows these defining recursions; Section A states the correspondence. Unlike these one-model corrections, FedDORA remains unbiased if either of two nuisance models is correct.

Concurrent FedSWE studies non-stationary unavailability through compensated local computation and implicit gossiping (Xiang et al., 2026). FedSWE, FedAWE, and Debias-FL seek robust or unbiased optimization through availability compensation, persistent state, or frequency correction. FedDORA instead casts each round's full-client vector update as a sequentially missing outcome and couples an observation nuisance with an outcome nuisance through a pre-round-measurable augmented score; this yields an exact product-form misspecification remainder rather than requiring either component alone to be correct.

Cross-round completion and doubly robust estimation. FedCDA selects historical client models by cross-round divergence (Wang et al., 2024); FedDORA uses sketches only for neighbor lookup and predicts in the original update space. Most closely, AAggFF-D uses a doubly robust estimator for partially observed client responses in fairness-aware online convex aggregation under client sampling with a known inclusion probability (Hahn et al., 2024). FedDORA instead targets the full high-dimensional update under history-dependent informative availability, learns both nuisances online from censored pre-round histories, and carries their product-form misspecification into a non-convex FL bound. Offline AIPW/DML assumes a fixed sample with static outcomes and sample splitting (Chernozhukov et al., 2018); here the outcome law changes with w_t, and same-round selection cannot enter its own correction. The contribution is therefore the prequential missing-vector formulation and optimization consequence, not the DR identity alone.

3 Problem Formulation

There are N clients with objective $F(w) = N^{-1}\sum_i f_i(w)$. At round t, $\Delta_{i,t} \in \mathbb{R}^d$ is the potential update client i would produce from w_t, but it is observed only when $A_{i,t} = 1$. The realized full-client update is $\bar{\Delta}_t = N^{-1}\sum_i \Delta_{i,t}$; its history-conditional target is $\Delta_t^\star = \mathbb{E}[\bar{\Delta}_t \mid \mathcal{H}_t] = N^{-1}\sum_i \mu_{i,t}$. This distinction separates irreducible local-update randomness from selection bias. The pre-round history $\mathcal{H}_t$ contains prior participation, staleness, last reported loss and update norm, round phase, and declared sample count. Define

$$\pi_{i,t} = \Pr(A_{i,t} = 1 \mid \mathcal{H}_t), \qquad \mu_{i,t} = \mathbb{E}[\Delta_{i,t} \mid \mathcal{H}_t]. \tag{2}$$

Assumption 1 (History-conditional observation). *$A_{i,t} \perp\!\!\!\perp \Delta_{i,t} \mid \mathcal{H}_t$ and $\pi_{i,t} \geq \pi_0 > 0$. Nuisance estimates are measurable with respect to $\mathcal{H}_t$, and $\hat{\pi}_{i,t} \geq \underline{p} > 0$.*

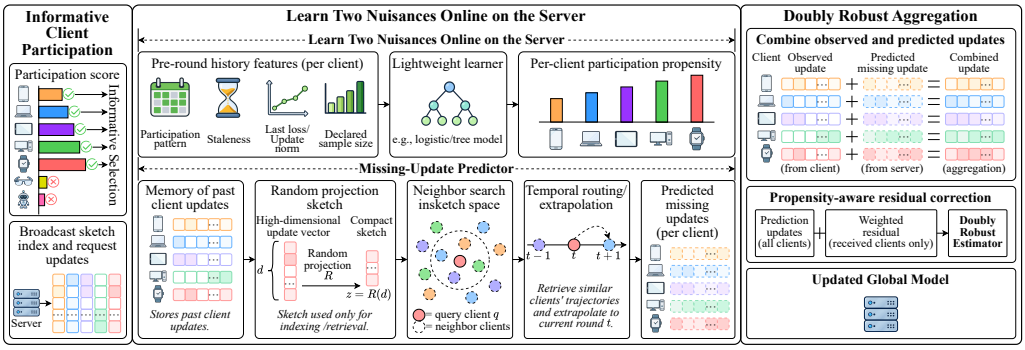

Figure 2: FEDDORA **architecture and executable data flow.** Informative participation exposes only the active updates. Before observing round-t availability, the server maps observable histories to client propensities and maps update memories through a fixed random-projection index, temporal extrapolation, and full-dimensional neighbor routing to missing-update predictions. The canonical augmented score combines the population prediction with propensity-weighted observed residuals and updates the global model. Sketches support lookup only; local training and the client payload are unchanged.

Marginal dependence between participation and updates is allowed. If an unlogged state still affects both, the full-population update is not identified; Proposition 1 formalizes this boundary, and our mean-matched latent-shift stress test intentionally violates it.

4 METHOD

4.1 OVERVIEW AND LEAKAGE-FREE ROUND TIMELINE

FEDDORA is a server-side realization of the augmented score in Eq. (1). Its three modules answer the three questions encoded by Figure 2: the propensity learner models *which* potential updates are observed; trajectory routing predicts *what* absent clients would have returned; and canonical aggregation determines *how* predictions and observations should be combined without first-order sensitivity to either nuisance alone. The design changes neither local SGD nor the client message: each active client still receives w_t and returns one d-dimensional update.

The execution order is essential for identification. At the start of round t, the server state contains only history through $t-1$: the latest two reported updates, participation counts, staleness, last reported loss and update norm, and the logistic parameter β_t. From this frozen state it constructs every $\hat{\pi}_{i,t}$ and $\hat{m}_{i,t}$ *before* observing $A_{i,t}$. The server then receives updates from $S_t = \{i : A_{i,t} = 1\}$, forms the aggregate, updates w_t, and only afterward writes round-t observations into memory and fits β_{t+1}. Consequently, a round-t update can improve future nuisance estimates but cannot predict or weight itself. This chronological separation is the implementation-level counterpart of the pre-round measurability condition in Assumption 1.

4.2 LEAKAGE-FREE ONLINE PROPENSITY LEARNING

Participation can depend on observable client history, so a uniform sampling probability systematically misweights the returned updates. We model the conditional inclusion probability with a lightweight online logistic learner shared across clients. For client i, the pre-round feature vector is

$$x_{i,t} = \left[1, \operatorname{logit}(\bar{a}_{i,t}), \frac{\log(1+s_{i,t})}{3}, \tilde{\ell}_{i,t}, \widetilde{\|\Delta\|}_{i,t}, \sin(t/5), \widetilde{\log(1+n_i)}\right]. \tag{3}$$

Here $\bar{a}_{i,t}$ is smoothed activity, $s_{i,t}$ is staleness, $\ell_{i,t}$ and $\|\Delta\|_{i,t}$ are the last reported loss and update norm, and n_i is sample count; tildes denote cross-client standardization. Cold-start raw loss/norm memories are $1/0$ until the first report; with $\beta_0 = 0$, every first-round propensity is 0.5 before clipping. These features remain available when client i is absent. The pre-round prediction is

$$\hat{\pi}_{i,t} = \operatorname{clip}\big(\sigma(x_{i,t}^{\top}\beta_t), \underline{p}, 0.98\big), \tag{4}$$

with the main-suite floor $\underline{p} = 0.04$. After completing round t, the server refreshes its stored history and takes three gradient steps of size 0.12 on the observed binary availability vector, with $10^{-3}\|\beta\|_2^2/2$ regularization. This produces β_{t+1} for the next pre-round prediction. Thus $A_{\cdot,t}$ may inform future propensities through the updated history, but it never changes $\hat{\pi}_{\cdot,t}$ or the round-t aggregate retroactively.

The module has two practical advantages. Pooling the binary outcomes of all N clients makes each online update data efficient, while client-specific histories still induce different probabilities. Clipping controls the variance of inverse weights under weak overlap; it is not an extra residual heuristic, and any resulting propensity approximation remains explicitly represented by $\epsilon_{\pi,t}$ in Appendix B. Figure 4(d) measures its error against the simulator's oracle probabilities rather than inferring calibration only from downstream accuracy.

4.3 Sketch-based trajectory routing

The outcome nuisance is an interface of the framework, not a commitment to a particular router. Any pre-round-measurable vector-update predictor is admissible, with its error entering the analysis through $\epsilon_{m,t}$. Our lightweight realization combines temporal and cross-client evidence because propensity correction alone can have high variance under sparse participation. The server maintains two zero-initialized buffers: $u_{i,t}^{(1)}$ stores the latest reported update, while $u_{i,t}^{(2)}$ stores the preceding report when one exists. Its client-specific temporal proposal is

$$b_{i,t} = u_{i,t}^{(1)} + \kappa\big(u_{i,t}^{(1)} - u_{i,t}^{(2)}\big), \qquad \kappa = 0.35. \tag{5}$$

After exactly one report, $u_{i,t}^{(2)} = 0$ and the implemented proposal is $b_{i,t} = (1+\kappa)u_{i,t}^{(1)}$; from the second report onward, Eq. (5) extrapolates observed velocity. For unseen clients, let $\mathcal{O}_t$ contain those with a pre-round report and set $c_t = |\mathcal{O}_t|^{-1}\sum_{j\in\mathcal{O}_t} u_{j,t}^{(1)}$ if nonempty, else $c_t = 0$. For each $i \in \mathcal{O}_t$, a fixed seeded Gaussian $R \in \mathbb{R}^{d\times s}$, $s = 48$, produces $z_{i,t} = u_{i,t}^{(1)}R$. The server retrieves up to $r = 6$ largest off-diagonal similarities in $\mathcal{O}_t \setminus \{i\}$ and computes

$$\alpha_{ij,t} = \frac{\exp(\cos(z_{i,t}, z_{j,t})/\tau)}{\sum_{k\in\mathcal{N}_{i,t}} \exp(\cos(z_{i,t}, z_{k,t})/\tau)}, \quad q_{i,t} = \sum_{j\in\mathcal{N}_{i,t}} \alpha_{ij,t} u_{j,t}^{(1)}, \qquad \tau = 0.25. \tag{6}$$

For a previously observed client with at least one eligible neighbor, the final outcome nuisance is

$$\hat{m}_{i,t} = \rho b_{i,t} + (1-\rho)q_{i,t}, \qquad \rho = 0.55. \tag{7}$$

If $i \notin \mathcal{O}_t$, it has no valid query sketch and we set $\hat{m}_{i,t} = c_t$; if $|\mathcal{O}_t| = 1$, the sole observed client uses $b_{i,t}$. Before the first report all predictions are zero. This fallback remains active after burn-in, which does not assume complete client coverage. After aggregation, only participating rows shift from $u^{(1)}$ to $u^{(2)}$ and receive a new $u^{(1)}$, so missing clients are never assigned fabricated observations.

The projection is solely a routing index: Eq. (6) retrieves and combines full d-dimensional updates, with no low-rank reconstruction. This preserves coordinate-level residual correction while moving similarity search to 48 dimensions. Appendix B gives one sufficient local error condition, and Figure 4(c) reports observed residual RMSE as a proxy. A stronger temporal model, learned sequence predictor, or alternative retrieval rule can replace Eqs. (5)–(7) without changing the score or theory, provided it uses only pre-round history. If prediction is inaccurate but propensity is correct, conditional bias is still removed, although variance can increase.

4.4 Canonical orthogonal aggregation

The two nuisance modules are joined only through the canonical augmented score. After five in-budget burn-in rounds, the server computes

$$\hat{\Delta}_t = \frac{1}{N}\sum_{i=1}^{N} \hat{m}_{i,t} + \frac{1}{N}\sum_{i:A_{i,t}=1} \frac{\Delta_{i,t} - \hat{m}_{i,t}}{\hat{\pi}_{i,t}}, \qquad w_{t+1} = w_t + \eta\hat{\Delta}_t. \tag{8}$$

The first term is a full-population plug-in estimate, including absent clients. The second term replaces the predictions of observed clients by propensity-corrected residual information. Together they

estimate the conditional target $\Delta_t^\star$; the realized deviation $\bar{\Delta}_t - \Delta_t^\star$ remains ordinary local-update noise. Both sums use the exact population normalization N^{-1}; in particular, the residual sum is not divided by the realized participant count or weight sum. On a post-burn-in empty round the residual sum is zero and the model advances using the mean prediction. During burn-in, a nonempty round uses the ordinary participant mean; an empty round follows the same memory-prediction fallback. All five initialization rounds count toward the common 80-round communication budget.

Eq. (8) uses neither self-normalization, residual clipping, reliability shrinkage, nor server momentum. Its conditional error is therefore exactly the nuisance-product remainder analyzed in Appendix B. Trajectory prediction reduces residual dispersion, propensity weighting repairs history-dependent observation, and their errors interact only at second order around the true nuisances. Algorithm 1 gives the exact round sequence, while Table 2 maps every module to its theorem quantity and empirical diagnostic.

4.5 Complexity and practical trade-offs

Client computation and communication are unchanged from FedAvg: each active client performs the same local steps and uploads one d-dimensional update. The exact server stores two $N \times d$ memories and a fixed $d \times s$ projection, for $O(Nd + ds)$ storage. Forming sketches costs $O(Nds)$, exact all-pairs similarity costs $O(N^2 s)$, routing full updates costs $O(Nrd)$, and canonical aggregation costs $O(Nd)$. These costs buy a correction for informative missingness without an additional client pass or payload.

We report only the exact-memory, exhaustive-routing implementation. Appendix C measures its overhead defensively rather than claiming scalability as a contribution; compressed memories and approximate retrieval remain unevaluated extensions.

5 Experiments

5.1 Protocol

We use CIFAR-10, CIFAR-100, and OfficeHome 512-dimensional ResNet-18 representations, each with six held-out corruption/domain episodes. One hundred clients receive Dirichlet label-skewed partitions ($\alpha = 0.2$). The simulator is a controlled identification testbed, not a deployment-trace surrogate: it (i) isolates bias induced by observable history, (ii) exposes oracle propensities for nuisance diagnosis, and (iii) creates a mean-matched hidden-driver violation. The primary logit uses the seven pre-round features above with no forced participant count; the violation condition adds an unobserved label-mixture term only outside the primary evidence.

All methods use identical client batches, local steps, client learning rate, and server learning rate. We report five seeds (2027–2031), accuracy, macro-F1, worst-decile client accuracy, ECE, paired differences with 95% t intervals, and two-sided paired Wilcoxon tests. The baselines are FedAvg (McMahan et al., 2017), SCAFFOLD (Karimireddy et al., 2020), FedVARP (Jhunjhunwala et al., 2022), FedAU (Wang & Ji, 2024), FedAWE (Xiang et al., 2024), Debiasing FedAvg (Debias-FL) (Sun et al., 2025), and FedCDA (Wang et al., 2024); four controls isolate trajectory-only, oracle IPW, oracle canonical DR, and learned canonical DR. The comparison is scoped to informative participation correction, with FedAvg/SCAFFOLD/FedVARP as optimization anchors rather than an exhaustive optimizer benchmark.

Every primary run uses exactly 80 communication rounds for every method. FedDORA's five burn-in rounds are included in these 80 rather than added to them: during rounds 1–5 it uses the same selected clients, two local steps, one model broadcast, and one update upload as FedAvg, then switches to Eq. (8) at round 6. Baselines receive the same 80-round client-computation and communication budget and no extra rounds are granted to FedDORA.

Table 1: **Primary mean and tail accuracy with 100 clients.** Mean ± standard deviation over five seeds; best is bold and second is underlined per benchmark and metric. The complete macro-F1 and ECE results are in Table 3 of Appendix C.

	CIFAR-10		CIFAR-100		OfficeHome	
Method	Acc. ↑	Worst-10 ↑	Acc. ↑	Worst-10 ↑	Acc. ↑	Worst-10 ↑
FedAvg	57.01±0.53	44.03±3.45	36.21±0.32	**32.32±1.53**	69.52±0.71	60.28±1.81
SCAFFOLD	56.51±0.28	43.66±2.65	36.00±0.22	31.80±1.06	70.92±0.40	62.33±2.60
FedVARP	56.92±0.66	42.85±2.47	35.36±0.10	32.30±0.81	70.67±0.46	62.57±0.58
FedAU	**57.09±0.30**	43.84±2.45	35.46±0.24	31.65±1.27	70.20±0.40	61.72±0.61
FedAWE	55.87±1.25	42.51±3.49	33.01±0.62	29.45±2.29	67.48±1.35	58.76±1.87
Debias-FL	53.75±1.06	41.39±4.42	35.69±0.42	31.27±1.59	69.99±0.89	61.10±2.45
FedCDA	56.50±0.41	41.95±2.03	36.13±0.49	31.90±1.35	69.61±0.57	61.03±1.75
FedDORA	57.07±0.52	**44.70±4.62**	**36.29±0.20**	32.14±0.86	**71.51±0.25**	**63.67±1.52**

5.2 Endpoint context and distributional robustness

Endpoint accuracy is context for the bias-correction claim, not its primary evidence. FedDORA obtains 71.51% mean and 63.67% worst-decile accuracy on OfficeHome, gains of 0.59 and 1.10 points over the strongest baselines. On CIFAR-100 its 0.08-point mean gain over FedAvg accompanies a 0.18-point tail deficit; on CIFAR-10 FedAU leads mean accuracy while FedDORA leads macro-F1 and the tail. We therefore make no practical-significance claim from these small gaps. Tail distributions, ECE, paired uncertainty, nuisance-oracle error, and intervention controls jointly assess whether correction behaves as predicted; Figure 4 and the latent stress test are the stronger evidence.

Figure 3 complements endpoint means: the convergence curve rules out a single late fluctuation; the violin, quantile, and ECDF panels expose client and class tails; and the accuracy–tail plane shows the relevant trade-off. Average rank is descriptive only. We interpret these moderate gaps jointly with the mechanism diagnostics below, not as universal state of the art.

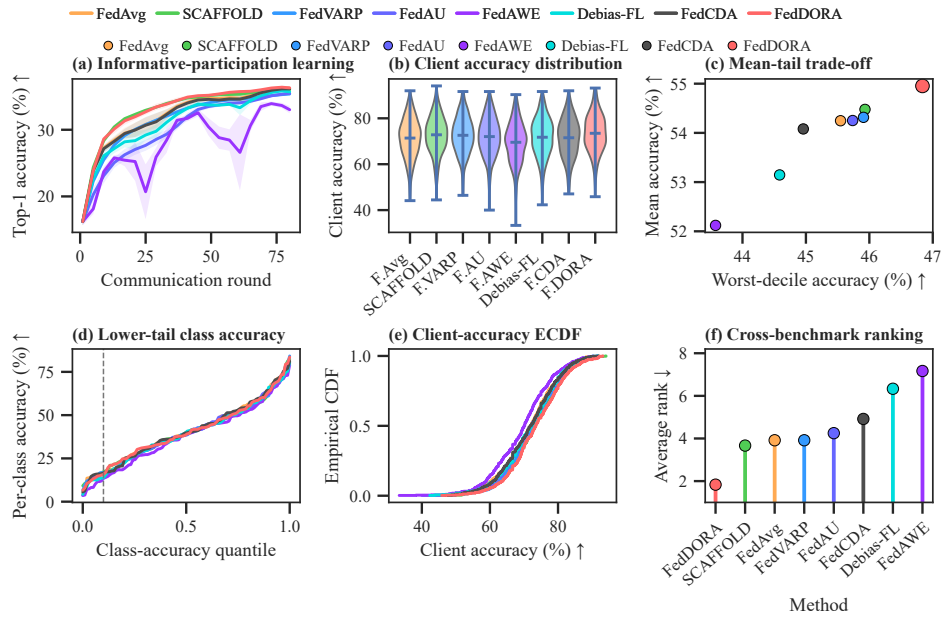

Figure 3: **Endpoint and distributional context.** (a) CIFAR-100 convergence with 95% intervals. (b) OfficeHome client distributions. (c) accuracy–tail tradeoff. (d) class quantiles. (e) client ECDF. (f) descriptive average rank. All subplots use five seeds; arrows state metric direction.

5.3 NUISANCE DIAGNOSTICS AND ASSUMPTION BOUNDARY

The primary mechanistic claim is correction behavior under informative observation, not a universal accuracy gain. The learned propensity reaches MSE around 3–5 $\times$ 10^{-3} against the simulator oracle. Figure 4 tests the theoretical chain: residual RMSE proxies outcome error, propensity-oracle MSE measures selection error, their product is compared with gradient bias, and four controls isolate completion, reweighting, and augmentation. Tail accuracy and ECE expose distributional and calibration effects. The mean-matched latent grid deliberately violates Assumption 1; its degradation is a boundary diagnostic, not robustness beyond identification.

The controls are distinct interventions: trajectory-only removes residual correction; oracle IPW removes completion; DR-oracle combines completion with true propensities; and DR-learned uses the online learner. Their ordering need not be monotone because weighting trades bias for variance. Together they separate *who is observed* from *what was missing* and test augmentation against either nuisance alone.

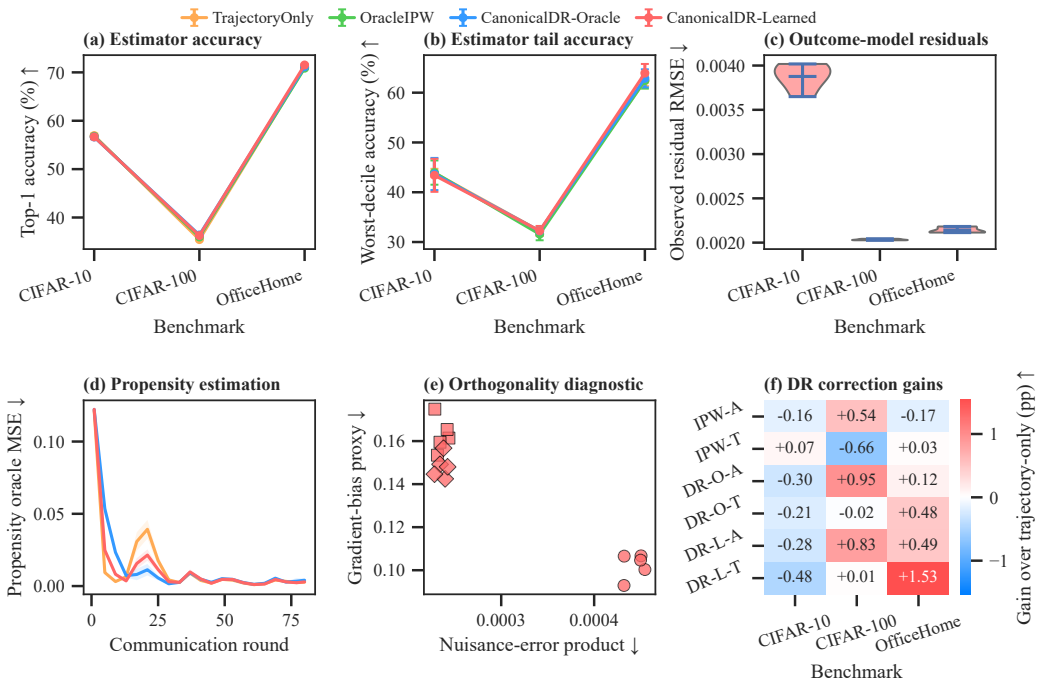

Figure 4: **Nuisance and oracle diagnostics.** (a–b) accuracy and tail accuracy of four diagnostic estimators across benchmarks with 95% intervals. (c) observed update-residual RMSE, an empirical outcome-nuisance proxy. (d) round-wise propensity-oracle error with 95% bands. (e) empirical nuisance-product proxy versus the gradient-bias proxy. (f) estimator gains over trajectory-only completion; A/T denotes mean/tail accuracy.

These diagnostics distinguish the optimization framework from post-hoc AIPW: both proxies evolve with the global model, use only pre-round state, and concern vector updates rather than fixed scalar outcomes. Panels 4(c–e) test the online nuisances and product remainder connecting correction to optimization. Oracle probabilities are available precisely because this is a controlled simulator; they enable causal debugging but do not constitute, or replace, privacy-preserving deployment traces.

5.4 NON-CONVEX ADAPTATION, PIXEL SCOPE CHECK, AND STRESS EVALUATION

Beyond a linear head, a CIFAR-10 residual MLP adapter gives FEDDORA the best eight-method mean (48.44%). On an 8,000-image held-out test split, a 20-client raw-pixel MNIST ResNet-18 check gives FEDDORA $79.07 \pm 1.37\%$, second to FedAvg's $81.19 \pm 3.04\%$ and above Debias-FL's $78.25 \pm 2.80\%$ (Appendix C). This small check establishes end-to-end execution but is not evidence

of comparative superiority. A 480-run grid varies heterogeneity, participation, and observable versus latent coupling.

Panels 5(c,d) vary participation and heterogeneity inside identification without retuning. Panels 5(e,f) mean-match participation volume before adding a hidden label-mixture driver, separating identified robustness from failure outside the assumptions.

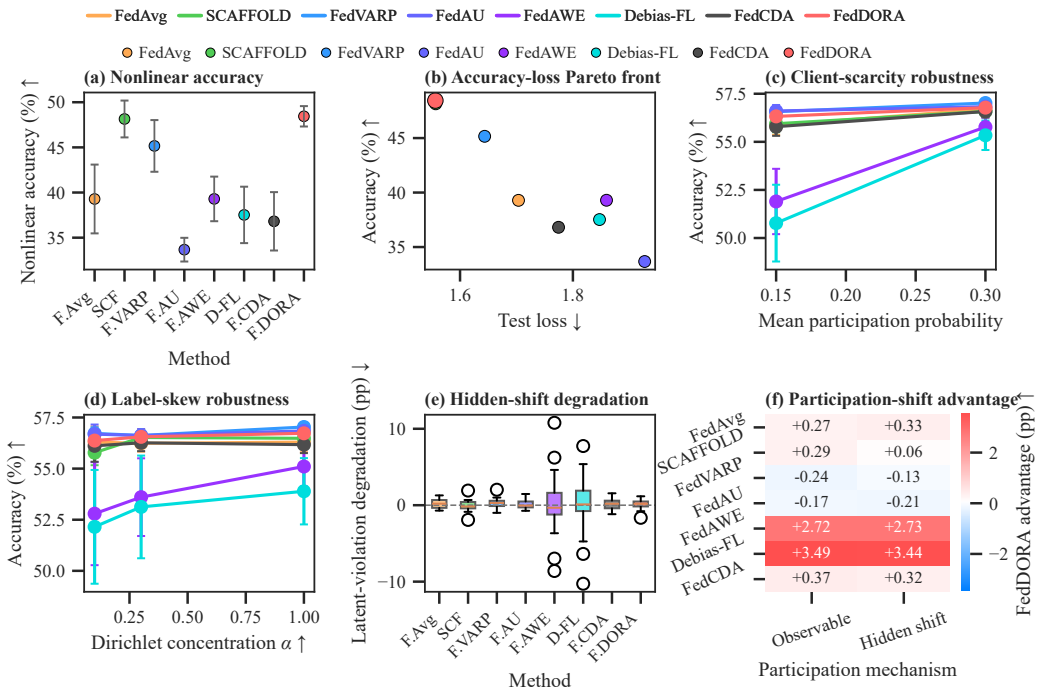

Figure 5: **Representation-level adaptation and stress evidence.** (a–b) eight-method fixed-representation residual-adapter accuracy and accuracy–loss Pareto behavior. (c–d) all-eight-method participation-rate and heterogeneity sensitivity with 95% intervals. (e) degradation under a mean-matched unobserved label-mixture driver. (f) FEDDORA's mean accuracy contrast against all seven baselines under observable and hidden-shift mechanisms.

6 Conclusion and Limitations

FEDDORA casts informative participation as sequential missing-update optimization and couples online participation and trajectory models through a leakage-free orthogonal aggregate. The exact bias identity, conditional-independence variance formula, and general non-convex analysis expose how nuisance errors enter aggregation and optimization, while oracle controls, tail metrics, and assumption-violation tests support the predicted mechanism across three benchmarks. The resulting evidence is strongest for correcting identifiable selection effects rather than claiming universal endpoint dominance.

Limitations are threefold. First, identification requires overlap and sufficiently informative pre-round histories; never-observed clients and unlogged drivers remain unrecoverable. Second, most evidence is controlled and representation-level: the small raw-pixel check establishes feasibility but not superiority over FedAvg, so large-model and deployment-trace validation remain necessary. Third, exact routing stores two full updates per client and performs quadratic sketch search; privacy-preserving compression and approximate retrieval are important future directions.

AI USE STATEMENT

Generative AI tools were used to assist with generating and refining portions of the experimental code and with checking and revising intermediate steps in the mathematical proofs. All AI-assisted code was manually reviewed and tested, and all mathematical derivations and proofs were independently verified by the authors. Generative AI was also used for language polishing and improving the clarity of the manuscript; the research problem, proposed method, experimental design, and interpretation of results were developed by the authors. The authors take full responsibility for the final content of the paper, including its code, theoretical claims, experimental results, and text.

ETHICS STATEMENT

We use public representations and simulations, without human subjects or private logs. Deployment histories require protection; erroneous weights can amplify noisy clients.

REPRODUCIBILITY STATEMENT

Fixed configurations, a run manifest, seed-level CSVs, plot/table sources, and JSON/PT outputs reproduce all results; Appendix C documents protocol and compute.

REFERENCES

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.

Seok-Ju Hahn, Gi-Soo Kim, and Junghye Lee. Pursuing overall welfare in federated learning through sequential decision making. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17246–17278. PMLR, 2024. URL `https://proceedings.mlr.press/v235/hahn24a.html`.

Divyansh Jhunjhunwala, Pranay Sharma, Aushim Nagarkatti, and Gauri Joshi. FedVARP: Tackling the variance due to partial client participation in federated learning. In *UAI*, pp. 906–916, 2022.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *ICML*, pp. 5132–5143, 2020.

Christoph F. Kurz. Augmented inverse probability weighting and the double robustness property. *Medical Decision Making*, 42(2):156–167, 2022.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pp. 1273–1282, 2017.

Zhenyu Sun, Ziyang Zhang, Zheng Xu, Gauri Joshi, Pranay Sharma, and Ermin Wei. Debiasing federated learning with correlated client participation. In *ICLR*, 2025.

Haozhao Wang, Haoran Xu, Yichen Li, Yuan Xu, Ruixuan Li, and Tianwei Zhang. FedCDA: Federated learning with cross-rounds divergence-aware aggregation. In *ICLR*, 2024.

Shiqiang Wang and Mingyue Ji. A lightweight method for tackling unknown participation statistics in federated averaging. In *ICLR*, 2024.

Ming Xiang, Stratis Ioannidis, Edmund Yeh, Carlee Joe-Wong, and Lili Su. Efficient federated learning against heterogeneous and non-stationary client unavailability. In *NeurIPS*, volume 37, 2024.

Ming Xiang, Stratis Ioannidis, Edmund Yeh, Carlee Joe-Wong, and Lili Su. Resilience beyond stationary client unavailability: Unlocking efficient and unbiased federated learning. *Journal of Machine Learning Research*, 27(173):1–73, 2026.

Zilong Yin, Haoyu Wang, Bin Chen, Xin Zhang, Xiaogang Lin, Hangling Sun, Anji Li, and Chenyu Zhou. Federated semi-supervised representation augmentation with cross-institutional knowledge transfer for healthcare collaboration. *Knowledge-Based Systems*, 300:112208, 2024. doi: 10.1016/j.knosys.2024.112208.

A Algorithm and Implementation Correspondence

Algorithm 1 FedDORA at the server

Initialize w_0, $\beta_0 = 0$, both update buffers to zero, loss/norm memories to $1/0$, and fixed R.
for $t = 0, \dots, T-1$ **do**
  Build $x_{i,t}$ from pre-round history and predict $\hat{\pi}_{i,t}$.
  Let $\mathcal{O}_t$ contain clients with a pre-round report and set c_t to their mean latest update (zero if empty).
  For $i \in \mathcal{O}_t$, route within $\mathcal{O}_t \setminus \{i\}$ and mix with $b_{i,t}$ (if only one report, $b_{i,t} = (1+\kappa)u_{i,t}^{(1)}$); for $i \notin \mathcal{O}_t$, set $\hat{m}_{i,t} = c_t$.
  Broadcast w_t and receive $\Delta_{i,t}$ from active set S_t.
  if $t < 5$ and $S_t \neq \emptyset$ **then**
    $\hat{\Delta}_t \leftarrow |S_t|^{-1} \sum_{i \in S_t} \Delta_{i,t}$ (memory burn-in).
  else if $S_t = \emptyset$ **then**
    $\hat{\Delta}_t \leftarrow N^{-1} \sum_i \hat{m}_{i,t}$ (prediction fallback).
  else
    $\hat{\Delta}_t \leftarrow N^{-1} \sum_i \hat{m}_{i,t} + N^{-1} \sum_{i \in S_t} \hat{\pi}_{i,t}^{-1} (\Delta_{i,t} - \hat{m}_{i,t})$.
  end if
  $w_{t+1} \leftarrow w_t + \eta \hat{\Delta}_t$; update memories and fit β_{t+1} using $A_{\cdot,t}$.
end for

Baseline fidelity. FedAU maintains the cutoff interval length, number of completed intervals, and online mean interval weight from Algorithm 2 of Wang & Ji (2024). FedAWE maintains a persistent client model and last-active round, scales the local innovation by elapsed rounds, averages uploaded models, and returns the new global model to active clients as in Algorithm 1 of Xiang et al. (2024). Debias-FL uses the current empirical frequency $t_i/((t+1)B)$ to rescale the local update as in Algorithm 1 of Sun et al. (2025), generalized to random B_t through cumulative participation slots. FedCDA stores a three-round candidate window and selects the candidate closest to the current participating-model center, matching its defining cross-round divergence rule (Wang et al., 2024). FedAvg, SCAFFOLD, FedVARP follow their defining server/client recursions (McMahan et al., 2017; Karimireddy et al., 2020; Jhunjhunwala et al., 2022).

Complexity. Client computation and communication match FedAvg. The direct server implementation stores two $N \times d$ memories and a $d \times 48$ fixed projection. Exact all-pairs sketch routing costs $O(N^2 s + Nrd)$ after sketch construction. FP16 storage halves memory bytes; low-bit quantization or K prototypes can reduce storage toward $O(Kd)$, and approximate nearest-neighbor indices can avoid exhaustive pairwise retrieval. These variants remain pre-round measurable and therefore preserve the score identity, but may enlarge trajectory error; none is evaluated here.

B Theoretical Analysis and Proofs

All formal guarantees are collected in this appendix so that the main paper can describe the complete executable method without interrupting its round-level flow. The analysis concerns the exact canonical aggregate in Eq. (8) from round $t_0 = 5$ onward; the finite burn-in initializes histories but is not claimed to be doubly robust. We do not analyze self-normalization, residual clipping, reliability shrinkage, or momentum because none appears in the implementation.

Identification boundary. An unlogged driver of both availability and updates leaves the missing part of the full-client target unrestricted.

Table 2: **Algorithm–theory–evidence correspondence.** Every empirical diagnostic is attached to the implemented quantity that appears in the analysis; no theorem is justified only by end-task accuracy.

Implemented component	Mathematical role	Assumption or guarantee	Direct evidence
Frozen pre-round state and online logistic learner	$\hat{\pi}_{i,t}$ and $\epsilon_{\pi,t}$	$\mathcal{H}_t$-measurability, positivity, Theorem 1	Figure 4(d); oracle/learned DR controls
Two-update memory, temporal proposal, and sketch routing	$\hat{m}_{i,t}$ and $\epsilon_{m,t}$	Eq. 15; prediction term in Eq. 13	Observed-residual proxy in Figure 4(c); trajectory-only control
Population-normalized augmented score	$N^{-1}\sum_i \psi_{i,t}$	Product bias, orthogonality, and general conditional variance	Figure 4(a,b,e,f)
Five in-budget initialization rounds followed by the exact score	$t_0 = 5$ and $w_{t+1} = w_t + \eta\hat{\Delta}_t$	Theorem 2 applies post-burn-in	Figures 3(a) and 5(a)
Mean-matched hidden label-mixture driver	Deliberate violation of $A \perp \Delta \mid \mathcal{H}_t$	Non-identification outside Assumption 1	Figure 5(e,f)

Proposition 1 (Non-identifiability under unlogged selection)**.** *Without conditional independence in Assumption 1, the observed law of $(\mathcal{H}_t, A_{i,t}, A_{i,t}\Delta_{i,t})$ does not generally identify $\mu_{i,t}$, even with participation bounded away from zero.*

Proof. Fix a history and nonzero vector v. In two processes let $\Pr(A = 1) = 1/2$ and $\Delta = 0$ when $A = 1$; when $A = 0$, set $\Delta = 0$ in the first and $\Delta = v$ in the second. Both induce the same observed law, $(A, A\Delta) \in \{(1, 0), (0, 0)\}$ equiprobably, but their conditional means are 0 and $v/2$. Hence observed trajectories do not identify the target without Assumption 1. □

B.1 Bias, orthogonality, and variance

The potential full-client update $\bar{\Delta}_t = N^{-1}\sum_i \Delta_{i,t}$ is random because local minibatches are random even if every client participates. Selection-bias analysis therefore targets its history-conditional mean. Define the target and the implemented post-burn-in aggregate as

$$\Delta_t^\star = \frac{1}{N}\sum_{i=1}^{N} \mu_{i,t}, \qquad \hat{\Delta}_t = \frac{1}{N}\sum_{i=1}^{N} \psi_{i,t}(\hat{\pi}, \hat{m}), \tag{9}$$

where every nuisance value is fixed before participation at round t.

This target choice does not replace the full-client objective with a participating-client objective: $\Delta_t^\star = \mathbb{E}[\bar{\Delta}_t \mid \mathcal{H}_t]$ still averages all N potential clients. It only integrates out minibatch/update noise already present under full participation. The optimization analysis below places the corresponding zero-mean deviation in ζ_t.

Theorem 1 (Exact double-robust bias identity)**.** *Under Assumption 1, conditionally on $\mathcal{H}_t$,*

$$\mathbb{E}[\psi_{i,t} - \mu_{i,t} \mid \mathcal{H}_t] = \frac{\pi_{i,t} - \hat{\pi}_{i,t}}{\hat{\pi}_{i,t}}(\mu_{i,t} - \hat{m}_{i,t}). \tag{10}$$

Thus the aggregate is conditionally unbiased if either nuisance is correct. With $\epsilon_{\pi,t}^2 = N^{-1}\sum_i(\hat{\pi}_{i,t} - \pi_{i,t})^2$ and $\epsilon_{m,t}^2 = N^{-1}\sum_i \|\hat{m}_{i,t} - \mu_{i,t}\|^2$,

$$\|\mathbb{E}[\hat{\Delta}_t - \Delta_t^\star \mid \mathcal{H}_t]\| \le \underline{p}^{-1}\epsilon_{\pi,t}\epsilon_{m,t}. \tag{11}$$

Proposition 2 (Conditional Neyman orthogonality). *Let* $\mathcal{M}_t(p, m) = N^{-1} \sum_i \mathbb{E}[m_i + A_{i,t} p_i^{-1}(\Delta_{i,t} - m_i) - \mu_{i,t} \mid \mathcal{H}_t]$. *For bounded directions* h_p, h_m, *the Gateaux derivative at the true nuisances vanishes:*

$$\frac{\partial}{\partial \varepsilon} \mathcal{M}_t(\pi + \varepsilon h_p, \mu + \varepsilon h_m)\Big|_{\varepsilon=0} = 0. \tag{12}$$

Hence first-order perturbations do not affect the target moment; Eq. (10) is the exact joint-misspecification remainder.

Proposition 3 (Explicit variance and the value of prediction). *If clients are conditionally independent given* $\mathcal{H}_t$ *and* $v_{i,t} = \mathbb{E}[\|\Delta_{i,t} - \mu_{i,t}\|^2 \mid \mathcal{H}_t]$, *then*

$$\sigma^2_{\mathrm{DR},t} = \frac{1}{N^2} \sum_i \frac{\pi_{i,t}}{\hat{\pi}^2_{i,t}} \left[v_{i,t} + (1 - \pi_{i,t}) \|\mu_{i,t} - \hat{m}_{i,t}\|^2 \right]. \tag{13}$$

equals $V_t := \operatorname{tr} \operatorname{Cov}(\hat{\Delta}_t \mid \mathcal{H}_t)$. *With correct propensities,* $\hat{m} = \mu$ *minimizes this expression. Relative to oracle IPW (*$\hat{m} = 0$*), the resulting variance reduction is* $N^{-2} \sum_i (1 - \pi_{i,t}) \pi_{i,t}^{-1} \|\mu_{i,t}\|^2$.

Corollary 1 (Conditional mean-squared aggregation error). *Under the conditions of Proposition 3,*

$$\mathbb{E}\Big[\|\hat{\Delta}_t - \Delta_t^\star\|^2 \mid \mathcal{H}_t \Big] = \sigma^2_{\mathrm{DR},t} + \Big\| \mathbb{E}[\hat{\Delta}_t - \Delta_t^\star \mid \mathcal{H}_t] \Big\|^2 \le \sigma^2_{\mathrm{DR},t} + \underline{p}^{-2} \epsilon^2_{\pi,t} \epsilon^2_{m,t}. \tag{14}$$

Thus trajectory completion and propensity learning affect different, measurable parts of error: the former can reduce residual variance, while their joint misspecification controls squared conditional bias.

These statements expose the estimator's trade-off. Better update prediction decreases both the second-order bias and residual variance; better propensity calibration decreases bias, but small propensities amplify variance; and the propensity floor regularizes variance while its approximation remains in $\epsilon_{\pi,t}$. Figure 4(c–e) measures both nuisances, their product, and a gradient-bias proxy, whereas panels (a,b,f) compare the completion, weighting, and augmentation controls predicted by Eqs. (13)–(14).

B.2 A SUFFICIENT CONDITION FOR TRAJECTORY ROUTING

The routing module is a practical nuisance estimator, not a universally consistent predictor. For an observed client, let d_c be cosine distance between normalized stored updates. Suppose (i) conditional means are locally smooth, $\|\mu_{i,t} - \mu_{j,t}\| \le L_c d_c(u_i, u_j) + \delta_t$; (ii) stored neighbor errors obey $\|u_j - \mu_{j,t}\| \le \epsilon_{\mathrm{mem}}$; and (iii) the sketch distorts relevant cosine distances by at most ϵ_R. If the retrieved neighbors lie within sketch radius ϵ_{nn} and the temporal proposal error is at most ϵ_{temp}, convexity of the weights in Eq. (6) yields

$$\|\hat{m}_{i,t} - \mu_{i,t}\| \le \rho \epsilon_{\mathrm{temp}} + (1 - \rho)[L_c(\epsilon_{\mathrm{nn}} + \epsilon_R) + \delta_t + \epsilon_{\mathrm{mem}}]. \tag{15}$$

Thus routing is accurate when update direction identifies locally similar conditional means and model drift is slow relative to memory refresh. For unseen clients, the prototype error $\|c_t - \mu_{i,t}\|$ enters the same aggregate quantity $\epsilon_{m,t}$; no separate consistency claim is made. When these conditions fail, Eq. (11) still removes conditional bias if the propensity nuisance is correct, while Eq. (13) shows why a poor predictor can increase variance.

B.3 FEDERATED OPTIMIZATION CONSEQUENCE

Assumption 2 (Optimization regularity). *After burn-in,* F *is* L*-smooth and lower bounded. The history-conditional full-client update satisfies* $\Delta_t^\star = -\nabla F(w_t) + b_t^{\mathrm{loc}}$, *where the local-SGD drift obeys* $\mathbb{E}\|b_t^{\mathrm{loc}}\|^2 \le B_K^2$. *Random deviation of* $\bar{\Delta}_t$ *around* $\Delta_t^\star$ *is included in the conditional aggregation variance.*

Theorem 2 (Non-convex convergence). *Under Assumptions 1 and 2, define the general conditional variance* $V_t = \operatorname{tr} \operatorname{Cov}(\hat{\Delta}_t \mid \mathcal{H}_t)$ *and let* $\eta \le 1/(4L)$. *After burn-in* t_0, *canonical* FEDDORA *satisfies*

$$\begin{aligned} \frac{1}{T - t_0} \sum_{t=t_0}^{T-1} \mathbb{E}\|\nabla F(w_t)\|^2 \le & \frac{4(F(w_{t_0}) - F_\star)}{\eta(T - t_0)} + 4L\eta \frac{\sum_t \mathbb{E} V_t}{T - t_0} + 8B_K^2 \\ & + \frac{8}{\underline{p}^2 (T - t_0)} \sum_t \mathbb{E}[\epsilon^2_{\pi,t} \epsilon^2_{m,t}]. \end{aligned} \tag{16}$$

Both sums run from t_0 to $T-1$. Under the conditional client-independence condition of Proposition 3, $V_t = \sigma^2_{\mathrm{DR},t}$; otherwise V_t also contains cross-client covariance.

Corollary 2 (Nuisance-rate regimes)**.** *If almost surely $\epsilon_{\pi,t} \le C_\pi(t+1)^{-a}$ and $\epsilon_{m,t} \le C_m(t+1)^{-b}$, the last term in Eq. (16) is $O(\underline{p}^{-2}C_\pi^2 C_m^2 \varphi_T)$, where*

$$\varphi_T = \begin{cases} T^{-2(a+b)}, & 2(a+b) < 1, \\ (\log T)/T, & 2(a+b) = 1, \\ T^{-1}, & 2(a+b) > 1. \end{cases} \tag{17}$$

If either nuisance is exact, this optimization penalty vanishes regardless of the other nuisance.

The bound separates optimization progress, conditional DR variance, local-update drift, and the product of sequential nuisance errors. It supports the diagnostics in Figure 4 but does not imply that correction must outperform averaging for every non-convex parameterization; the pixel-level check is empirical evidence of this distinction.

B.4 Proofs

Proof of Theorem 1. Measurability and conditional ignorability give

$$\mathbb{E}[\psi_{i,t} \mid \mathcal{H}_t] = \hat{m}_{i,t} + \frac{\pi_{i,t}}{\hat{\pi}_{i,t}}(\mu_{i,t} - \hat{m}_{i,t}).$$

Subtracting $\mu_{i,t}$ gives Eq. (10). Let $\epsilon^2_{\pi,t} = N^{-1}\sum_i(\hat{\pi}_{i,t} - \pi_{i,t})^2$ and $\epsilon^2_{m,t} = N^{-1}\sum_i \|\hat{m}_{i,t} - \mu_{i,t}\|^2$. Triangle inequality, $\hat{\pi} \ge \underline{p}$, and Cauchy–Schwarz yield aggregate bias at most $\underline{p}^{-1}\epsilon_{\pi,t}\epsilon_{m,t}$. □

Proof of Proposition 2. Conditional ignorability gives the exact moment map

$$\mathcal{M}_t(p, m) = \frac{1}{N}\sum_i \left[m_i - \mu_i + \frac{\pi_i}{p_i}(\mu_i - m_i) \right].$$

Along $p_i(\varepsilon) = \pi_i + \varepsilon h_{p,i}$ and $m_i(\varepsilon) = \mu_i + \varepsilon h_{m,i}$, differentiation at zero gives $h_{m,i} - h_{m,i} - h_{p,i}(\mu_i - \mu_i)/\pi_i = 0$ for every client. Averaging proves Eq. (12). Expanding the same display at arbitrary $(\hat{\pi}, \hat{m})$ leaves exactly the product remainder in Eq. (10). □

Proof of Proposition 3. Write $\Delta_i = \mu_i + \xi_i$ with $\mathbb{E}[\xi_i \mid \mathcal{H}_t] = 0$. After subtracting the score mean, the random component is $A_i\hat{\pi}_i^{-1}\xi_i + (A_i - \pi_i)\hat{\pi}_i^{-1}(\mu_i - \hat{m}_i)$. The cross term is zero by conditional ignorability. Its squared second moment is $\pi_i\hat{\pi}_i^{-2}v_i + \pi_i(1-\pi_i)\hat{\pi}_i^{-2}\|\mu_i - \hat{m}_i\|^2$. Conditional client independence removes cross-client covariances, and division by N^2 proves Eq. (13). With $\hat{\pi}_i = \pi_i$, subtracting the expression at $\hat{m}_i = \mu_i$ from that at $\hat{m}_i = 0$ yields $N^{-2}\sum_i(1-\pi_i)\pi_i^{-1}\|\mu_i\|^2$. □

Proof of Corollary 1. Condition on $\mathcal{H}_t$ and write $X = \hat{\Delta}_t - \Delta_t^\star$. The vector bias–variance identity gives $\mathbb{E}\|X\|^2 = \mathrm{tr}\,\mathrm{Cov}(X) + \|\mathbb{E}X\|^2$. Proposition 3 identifies the trace term as $\sigma^2_{\mathrm{DR},t}$, and Theorem 1 bounds the squared bias by $\underline{p}^{-2}\epsilon^2_{\pi,t}\epsilon^2_{m,t}$. □

Lemma 1 (Biased smooth descent)**.** *If $w_{t+1} = w_t - \eta(\nabla F(w_t) + \zeta_t + b_t)$, $\mathbb{E}[\zeta_t \mid w_t] = 0$, $\mathbb{E}\|\zeta_t\|^2 \le V_t$, and $\eta \le 1/(4L)$, then*

$$\frac{\eta}{4}\mathbb{E}\|\nabla F(w_t)\|^2 \le \mathbb{E}[F(w_t) - F(w_{t+1})] + L\eta^2 V_t + \eta\mathbb{E}\|b_t\|^2.$$

Proof. Apply L-smoothness, insert the update, take conditional expectation, use $2\langle a, b\rangle \le \frac{1}{2}\|a\|^2 + 2\|b\|^2$, and absorb the $L\eta^2\|\nabla F\|^2$ term. □

Proof of Theorem 2. Let $g_t = -\hat{\Delta}_t$ and decompose

$$g_t = \nabla F(w_t) - b_t^{\mathrm{loc}} + b_t^{\mathrm{nuis}} + \xi_t, \qquad \mathbb{E}[\xi_t \mid \mathcal{H}_t] = 0,$$

where $\mathbb{E}[\|\xi_t\|^2 \mid \mathcal{H}_t] = V_t$ and $b_t^{\mathrm{nuis}} = -\mathbb{E}[\hat{\Delta}_t - \Delta_t^\star \mid \mathcal{H}_t]$. Theorem 1 implies

$$\mathbb{E}\|b_t^{\mathrm{nuis}}\|^2 \le \underline{p}^{-2}\mathbb{E}[\epsilon_{\pi,t}^2 \epsilon_{m,t}^2].$$

Apply the biased smooth-descent lemma with $b_t = -b_t^{\mathrm{loc}} + b_t^{\mathrm{nuis}}$ and $\|b_t\|^2 \le 2\|b_t^{\mathrm{loc}}\|^2 + 2\|b_t^{\mathrm{nuis}}\|^2$:

$$\begin{aligned}\frac{\eta}{4}\mathbb{E}\|\nabla F(w_t)\|^2 \le \mathbb{E}[F(w_t) - F(w_{t+1})] + L\eta^2 \mathbb{E}V_t \\ + 2\eta B_K^2 + 2\eta \underline{p}^{-2}\mathbb{E}[\epsilon_{\pi,t}^2 \epsilon_{m,t}^2].\end{aligned}$$

Summing from t_0 to $T-1$, using $F(w_T) \ge F_\star$, and dividing by $\eta(T - t_0)/4$ gives Eq. (16). □

Proof of Corollary 2. The assumed bounds imply $\epsilon_{\pi,t}^2 \epsilon_{m,t}^2 \le C_\pi^2 C_m^2 (t+1)^{-2(a+b)}$. The integral test gives an average of order $T^{-2(a+b)}$ below exponent one, $(\log T)/T$ at exponent one, and T^{-1} above exponent one. Multiplication by the constant in Eq. (16) proves the stated regimes. Exactness of either nuisance sets every product to zero. □

C Experimental Details and Additional Results

Primary runs use 80 total communication rounds, two local steps, batch size at most 96, client learning rate 0.12, weight decay 2×10^{-4}, and server learning rate 1.0 for every method. FEDDORA uses $r = 6$, $\rho = 0.55$, sketch dimension 48, and propensity floor 0.04 across all three benchmarks. Its five FedAvg memory-initialization rounds occupy rounds 1–5 of the same 80-round budget; canonical aggregation starts at round 6 without additional client computation, uploads, or rounds. Empty rounds are retained: zero-update baselines do not update, whereas model-memory estimators use their defining predictions. Evaluation occurs every four rounds.

Data and metric construction. For each seed, the source and calibration representations are pooled and partitioned once into 100 clients by label-wise Dirichlet draws; tiny clients are repaired by transferring examples without duplicating or discarding the global training set. The test population is partitioned to approximately match each client's training label mixture. Reported top-1 accuracy and ECE are means over the six held-out corruption/domain episodes. Macro-F1, per-class quantiles, and client-level distributions use the canonical first episode, and worst-decile accuracy is the 0.1 quantile of the resulting 100 client accuracies. This distinction is fixed across all methods and seeds.

Representation provenance and leakage controls. All 512-dimensional features are frozen before federated partitioning, tuning, and training. We use the L2-normalized penultimate global-average-pool output of torchvision ResNet-18 with ImageNet-1K V1 weights, in evaluation mode, after resizing to 256, center-cropping to 224, and applying the associated ImageNet normalization; no target sample or label updates the backbone. CIFAR-10 uses fixed class-balanced subsets of 5,000 training and 2,000 test images, while CIFAR-100 uses all 50,000/10,000 images; a seeded 15% source split forms calibration data. OfficeHome uses Art, Clipart, and Product as source domains and holds out Real World as target, again splitting 15% of source data for calibration. Six deterministic target episodes apply Gaussian noise, motion blur, contrast, brightness, pixelation, and JPEG corruption with severities $(1, 2, 3, 4, 5, 3)$. Target labels are serialized only for evaluation and are never read by adaptation, propensity learning, routing, or model selection. The anonymous artifact includes `prepare_representations.py` with the exact subset, split, corruption, and episode seeds.

Table 3: **Complete primary evaluation.** Mean ± standard deviation over five seeds; best is bold and second is underlined per dataset and metric.

Dataset	Method	Accuracy ↑	Macro-F1 ↑	Worst-10 ↑	ECE ↓
CIFAR-10	FedAvg	57.01±0.53	61.53±1.66	44.03±3.45	15.42±0.58
	SCAFFOLD	56.51±0.28	61.20±0.64	43.66±2.65	17.08±1.68
	FedVARP	56.92±0.66	61.27±1.18	42.85±2.47	13.99±0.82
	FedAU	**57.09±0.30**	61.39±1.57	43.84±2.45	**12.97±1.08**
	FedAWE	55.87±1.25	60.65±1.84	42.51±3.49	13.34±2.00
	Debias-FL	53.75±1.06	58.85±1.55	41.39±4.42	20.95±3.06
	FedCDA	56.50±0.41	60.88±1.20	41.95±2.03	16.91±1.00
	FedDORA	57.07±0.52	**62.09±2.21**	**44.70±4.62**	14.75±1.19
CIFAR-100	FedAvg	36.21±0.32	**41.58±0.20**	**32.32±1.53**	10.21±0.13
	SCAFFOLD	36.00±0.22	41.39±0.36	31.80±1.06	10.03±1.04
	FedVARP	35.36±0.10	40.73±0.33	32.30±0.81	12.19±0.98
	FedAU	35.46±0.24	40.92±0.14	31.65±1.27	13.77±0.51
	FedAWE	33.01±0.62	38.19±0.46	29.45±2.29	14.24±1.21
	Debias-FL	35.69±0.42	40.59±0.63	31.27±1.59	11.03±0.82
	FedCDA	36.13±0.49	41.44±0.58	31.90±1.35	10.09±0.80
	FedDORA	**36.29±0.20**	41.51±0.32	32.14±0.86	**10.02±0.96**
OfficeHome	FedAvg	69.52±0.71	68.70±0.80	60.28±1.81	8.90±0.29
	SCAFFOLD	70.92±0.40	70.57±0.48	62.33±2.60	**2.10±0.69**
	FedVARP	70.67±0.46	70.10±0.64	62.57±0.58	2.71±1.06
	FedAU	70.20±0.40	69.39±0.61	61.72±0.61	14.28±0.66
	FedAWE	67.48±1.35	66.13±1.43	58.76±1.87	18.31±0.19
	Debias-FL	69.99±0.89	69.39±1.18	61.10±2.45	8.63±0.73
	FedCDA	69.61±0.57	68.82±0.58	61.03±1.75	8.06±0.34
	FedDORA	**71.51±0.25**	**71.18±0.39**	**63.67±1.52**	8.14±0.76

Statistical summaries. Table 1 reports the sample mean and standard deviation across the five matched seeds. Curve bands and diagnostic error bars are mean ± 1.96 standard errors. For every dataset–metric–baseline pair, the saved paired-difference file contains the seed-matched mean difference, a 95% Student-t interval with four degrees of freedom, and a two-sided paired Wilcoxon statistic. These small-sample summaries quantify uncertainty but are not interpreted as asymptotic proof of superiority.

The primary inclusion probability is generated by

$$\mathrm{logit}(\pi_{i,t}) = \mathrm{logit}(q) + 0.32h_{i,t}^{\mathrm{rate}} - 0.55h_{i,t}^{\mathrm{stale}} - 0.58\gamma h_{i,t}^{\mathrm{loss}} + 0.18h_{i,t}^{\mathrm{norm}} + 0.95h_{t}^{\mathrm{phase}} + 0.28h_{i}^{\mathrm{size}}. \tag{18}$$

Every term belongs to $\mathcal{H}_t$. Bernoulli draws are used directly; no post-selection repair is applied. The latent stress condition subtracts $1.5r_i$, where r_i is the standardized KL divergence between client i's label mixture and the global label mixture, intentionally withheld from $\mathcal{H}_t$. A bisection-solved intercept mean-matches participation probabilities between latent and observable conditions, isolating hidden dependence from participation volume.

Diagnostic estimands. The simulator exposes the oracle $\pi_{i,t}$, so propensity error is the client-mean $(\hat{\pi}_{i,t} - \pi_{i,t})^2$. Because the conditional mean update $\mu_{i,t}$ is not analytically available, Figure 4(c) uses the RMSE of $\Delta_{i,t} - \hat{m}_{i,t}$ over observed clients and rounds as a labeled empirical proxy rather than claiming direct access to $\epsilon_{m,t}$. Panel (e) multiplies the square root of mean propensity MSE by this residual proxy and compares it with the observable participant-versus-memory gradient-bias proxy. The oracle/learned estimator controls provide the complementary intervention evidence: they replace one nuisance or remove augmentation while leaving the benchmark protocol fixed.

Stress and scope mapping. The representation-level adapter changes the optimized parameterization while retaining the aggregation interface. The raw-pixel experiment trains all 701,178 ResNet-18

parameters and tests implementation portability, but its non-win is not a theorem violation: Theorem 2 bounds optimization error and never asserts dominance over FedAvg. The observable stress cells preserve Assumption 1; mean-matched hidden-driver cells break it by construction.

Table 4: **Representation-level non-convex adaptation.** Residual-adapter results over five seeds; best and second-best formatting follows the main table.

Method	Accuracy ↑	Test loss ↓
FedAvg	39.29±4.35	1.70±0.05
SCAFFOLD	<u>48.15±2.33</u>	<u>1.56±0.02</u>
FedVARP	45.16±3.27	1.64±0.03
FedAU	33.67±1.50	1.93±0.01
FedAWE	39.29±2.82	1.86±0.01
Debias-FL	37.52±3.56	1.85±0.03
FedCDA	36.81±3.69	1.77±0.09
FedDORA	**48.44±1.29**	**1.56±0.01**

Trajectory-router ablations. We vary one router component at a time on the primary CIFAR-10 setting using five matched seeds. Accuracy spans 56.45–57.07%, indicating low sensitivity in the tested range. The full router is best on accuracy and second on both worst-decile accuracy and ECE. Relative to neighbor-only routing it improves mean accuracy by 0.32 points with a 0.11-point tail trade-off; relative to temporal-only it improves accuracy, tail accuracy, and ECE by 0.62, 0.95, and 0.60 points, respectively. Thus temporal and neighbor signals are complementary on average, without implying monotone gains for every metric and seed.

Table 5: **Trajectory-router ablation.** Mean ± standard deviation over five seeds. The full setting uses sketch dimension $s = 48$, $r = 6$ neighbors, and temporal interpolation $\rho = 0.55$.

Variant	Accuracy ↑	Worst-10 ↑	ECE ↓
Full (s=48, r=6, ρ=0.55)	**57.07±0.52**	<u>44.70±4.62</u>	<u>14.75±1.19</u>
Temporal only	56.45±0.72	43.75±2.95	15.35±0.99
Neighbor only	56.75±0.57	**44.81±5.22**	15.16±1.07
Sketch s=16	56.54±0.56	43.39±3.65	15.05±1.57
Sketch s=32	56.50±0.45	42.60±3.45	14.96±1.06
Sketch s=64	56.76±0.61	43.08±5.18	14.91±0.88
Neighbors r=3	56.63±0.88	43.07±4.04	**14.60±0.43**
Neighbors r=9	56.94±0.67	44.56±4.52	14.98±1.15
Mix ρ=0.35	<u>56.97±0.33</u>	44.63±3.60	14.88±1.25
Mix ρ=0.75	56.60±0.73	44.05±5.76	15.37±1.22

Raw-pixel end-to-end scope check. We train a width-16 GroupNorm ResNet-18 (701,178 parameters) from scratch on MNIST pixels with 20 Dirichlet clients ($\alpha = 0.3$), history-dependent mean participation 0.45, 80 rounds, four local steps, batch size 64, client/server learning rates 0.03/0.5, and three seeds. A fixed 2,000-image development split selects among six predeclared server-side FEDDORA candidates; the remaining 8,000 images are evaluated once. Validation selects propensity floor $p = 0.12$, trajectory mix $\rho = 0.55$, and server rate 0.5. Baseline configurations are fixed a priori and their checkpoints are trained by our harness with identical data, rounds, local computation, and evaluation split; FedDORA alone receives this six-candidate server-side search. Table 6 ranks FEDDORA second in both accuracy and loss, behind FedAvg. Because tuning budgets are unequal and the dataset is small, this table verifies end-to-end execution only and is not a comparative superiority claim.

Table 6: **End-to-end raw-pixel ResNet-18.** Held-out 8,000-image test results over three seeds; best is bold and second is underlined.

Method	Accuracy ↑	Test loss ↓
FedAvg	**81.19±3.04**	**0.77±0.08**
FedVARP	75.04±1.75	1.01±0.14
Debias-FL	78.25±2.80	0.85±0.07
FedDORA	<u>79.07±1.37</u>	<u>0.82±0.06</u>

Server scalability. On the RTX 4080, we time only the CUDA aggregation kernel after warm-up for a 5,130-dimensional update, 15% participation, $s = 48$, and $r = 6$; these are not end-to-end training times. Exact FEDDORA remains below 1 ms at 500 clients but is substantially slower than averaging, while measured persistent state follows the predicted linear-in-N memory term. The result supports feasibility at the tested sizes, not Internet-scale deployment.

Table 7: **GPU aggregation microbenchmark.** Lower is better; state includes two client memories and the fixed projection.

Clients	FedAvg (ms) ↓	FEDDORA (ms) ↓	State (MB) ↓
50	0.005	0.137	2.9
100	0.005	0.177	4.9
200	0.006	0.294	8.8
500	0.024	0.957	20.5