REWARD-FREE CLASS DIRECTIONS FOR COLLABORATIVE FAIRNESS AND GLOBAL MODEL RECOVERY IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Collaborative fairness in federated learning concerns how participants' model benefits relate to their contributions. Under label skew, a single client score can obscure substantial differences in class participation. We study participation-aware class utility allocation together with its effect on shared training. Our framework, SICAR, constructs class directions from relative client distributions on the probability simplex, without separately valuing each class and mapping those values to rewards. This reward-free direction construction is followed by a response sensitive classifier fit and an explicit rank-based amplitude rule. We then subtract each saved additive perturbation from the trained client model before aggregation. The cancellation removes the injected offset, while our analysis characterizes the remaining effect of local optimization. This allows clients to receive differentiated class-wise utility while substantially decoupling the fairness intervention from global model optimization. Experiments show that the proposed framework maintains stable contribution–utility alignment while recovering global model performance.

1 INTRODUCTION

Federated learning (FL) enables multiple clients to collaboratively train a global model without directly sharing their raw data (McMahan et al., 2017; Kairouz et al., 2021), and has become a fundamental paradigm for distributed learning under privacy constraints. However, classic FL primarily focuses on maximizing global model performance and typically allows all clients to access the same or nearly the same model utility (Lyu et al., 2020; Sim et al., 2020; Xu et al., 2021). This uniform benefit is not always reasonable for clients with different data quantity, class coverage, and data quality. In collaborative settings, if the utility received by a client is persistently decoupled from what it contributes, high-contribution clients may have little incentive to continue providing valuable data or computing resources. Therefore, collaborative fairness requires not only learning a high-performance global model, but also ensuring that the utility of the model ultimately obtained by the client matches its contribution to the federated system (Wu et al., 2024).

Class imbalance makes this objective more specific. A client may provide substantial data for some classes and almost none for others, yet benefit from knowledge of every class through the global model. Such sharing is a normal benefit of FL. Under a contribution-based allocation objective, however, it raises the question of whether the utility received for a class should reflect participation in that class. We use class wise free-riding to describe an allocation mismatch under this participation-based policy. We assume truthful metadata and evaluate dispatched models; preventing strategic misreporting or subsequent client recalibration is outside this setting. A single client score does not specify which classes should receive more favorable responses. We therefore study how relative class participation can guide differentiated class utility while limiting its interference with shared training.

Existing approaches use reputation and gradient rewards (Lyu et al., 2020; Xu et al., 2021), Shapley-style valuation (Sim et al., 2020; Tastan et al., 2024), or model-capacity allocation (Wang et al., 2024b; 2026b; Tastan et al., 2025b). Their designs motivate three distinct questions. The first concerns contribution representation: reducing a class profile to a client scalar can discard differences

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.

between classes. Class-specific valuation already exists; ShapFed, for example, averages class-specific classifier similarities into a client weight for aggregation and personalized mixing (Tastan et al., 2024). The second concerns reward construction: even a vector of contribution estimates does not determine the parameter changes needed to produce corresponding class responses. The third concerns training interference: a modified client model changes the starting point of local optimization, so its effects can return through aggregation (Wu et al., 2024). The first two questions motivate geometric reward construction, while the third motivates cancellation before aggregation.

We construct class target directions from the relative geometry of client distributions on the probability simplex. We compare each smoothed client distribution with the federation distribution in centered log-ratio coordinates (Aitchison, 1982), preserving relative class participation in a zero-sum vector. We call this class direction construction reward-free because it bypasses separate estimation of marginal class values and the assignment of rewards from those estimates. A labeled server probe measures classifier responses in class margins; a regularized inverse fit realizes the scaled target in a response-sensitive subspace outside retained clean-update directions. This fit balances response error against perturbation norm. Standalone validation rank explicitly controls the perturbation amplitude, while class distributions determine the target direction. Thus, the reward-free designation applies to class direction construction, and the realization and amplitude rules remain explicit. Class counts specify relative participation, which we use as an allocation policy rather than a causal valuation of data. We evaluate how well this policy is realized and how it interacts with shared optimization.

We then apply known offset cancellation to the returned client models. The server saves each injected classifier perturbation and subtracts that same tensor after local training, before aggregation. This differs operationally from IAFL's stochastic reference recovery, which probabilistically supplies clients with a shared reference model (Wu et al., 2024): we correct the known additive offset in the uploaded models. The subtraction exactly removes that offset, while local gradients may still reflect the perturbed starting point. Our analysis characterizes this remaining discrepancy and derives a condition under which cancellation reduces it. The mechanism therefore prevents direct accumulation of the saved reward without asserting recovery of the exact clean training trajectory.

We assess whether class utility changes track relative class participation and whether cancellation recovers shared-model utility within the same procedure. On CIFAR-100 with ResNet18, we obtain within-client class Pearson alignment of 0.855 ± 0.009, alongside paired global accuracy gains of 2.40 ± 0.42 percentage points over reward retention and 1.53 ± 0.35 points over clean training. These observations jointly support differentiated class allocation and shared-model recovery. All completed CIFAR-100 and Tiny-ImageNet settings show positive mean recovery over reward retention. Architecture-dependent dispersion, a measurable cost in dispatched utility, and weaker allocation signals on Street View House Numbers define the scope of these findings.

Our contributions are twofold. First, we formulate a geometric allocation mechanism that retains relative class participation and realizes class response targets through regularized classifier perturbations. We use spectral analysis to characterize which target components the selected subspace can express. Second, we couple these rewards with known offset cancellation, derive its remaining optimization discrepancy, and evaluate class alignment, shared-model recovery, and dispatched utility through paired experiments and a carrier ablation. Related work, broader comparisons, and the scope of the available experimental evidence are detailed in Appendix A.

2 Problem Setup

There are $N \geq 2$ clients and $K \geq 2$ classes. Client i has $n_i > 0$ labeled training samples, with n_{ic} in class c. Define aggregation weights $w_i = n_i / \sum_j n_j$, local proportions $p_{ic} = n_{ic}/n_i$, and federation proportions $q_c = \sum_i w_i p_{ic}$. The shared model has parameters $\theta \in \mathbb{R}^D$; its packed classifier weights and biases have dimension M. The coordinate embedding $E \in \mathbb{R}^{D \times M}$ satisfies $E^\top E = I_M$. A reward $\delta_i \in \mathbb{R}^M$ changes the received model to $\theta + E\delta_i$. Local training $F_{it} : \mathbb{R}^D \to \mathbb{R}^D$ updates the whole model, including the backbone. Table 27 provides a consolidated notation reference.

Allocation and utility. On the common endpoint evaluation distribution, global accuracy is $u_g = \mathrm{Acc}(\theta_T)$ and client dispatch utility is $u_i = \mathrm{Acc}(\theta_T + E\delta_i)$, before further local adaptation.

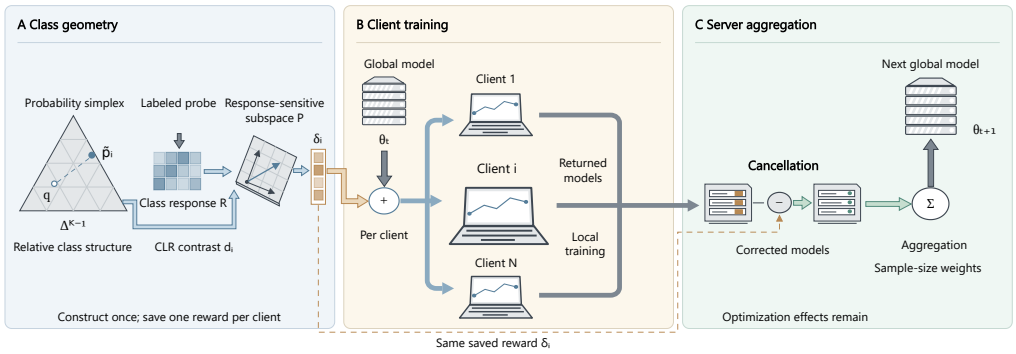

Figure 1: SICAR separates class-structured allocation from aggregation. Relative simplex geometry supplies a class direction, mapped through probe responses to a classifier perturbation. Distinct saved rewards produce distinct received models. Clients train normally; the server applies Known-δ Cancellation before aggregation. Computer and server icons denote local training and shared model states. The triangle schematically represents the K-class simplex. Rewards are constructed once and reused.

Class-conditional dispatch accuracy is u_{ic}. We measure across-client alignment of u_i with standalone quality c_i, and within-client alignment of class-accuracy changes with the target d_i. These capture complementary allocation dimensions. Collaborative fairness here concerns contribution–utility alignment, distinct from demographic parity (Salazar et al., 2024a), uniform client accuracy (Li et al., 2020), and personalization on client-specific distributions (Li et al., 2021b; Chen & Chao, 2022). We report absolute utility alongside both alignment measures. Because c_i and d_i also enter the allocation rule, these correlations measure policy alignment, not independent contribution valuation.

Information and assumptions. The server uses training class counts, a labeled validation probe V_c for each class, standalone validation accuracies c_i, a reference model $\theta_\star$, clean classifier-update history, and a positive calibration norm m_{ref}. We require global class support $q_c > 0$, nonempty probe classes, a positive calibration denominator, and compliant clients. We use client-provided counts and server-validation data for construction and calibration, excluding final-test data from those operations. Appendix E describes our training configuration, data access, and calibration choices.

3 Method

3.1 Overview

We construct each client's classifier reward once and reuse it throughout training. Relative class proportions determine its direction, and standalone validation rank controls its amplitude. As Figure 1 shows, each client trains from its rewarded model; the server subtracts the saved injection before aggregation. The procedure returns a shared model and the reward-bearing models dispatched to clients.

3.2 From the probability simplex to a class reward

Relative class direction. We represent class proportions on the probability simplex manifold. Its interior is $\Delta^{K-1}_{++} = \{p \in \mathbb{R}^K : p_c > 0,\ \mathbf{1}^\top p = 1\}$. To handle missing local classes, smooth toward q with total pseudocount mass α:

$$\widetilde{p}_{ic} = \frac{n_{ic} + \alpha q_c}{n_i + \alpha}, \qquad \widetilde{p}_i = \frac{n_i}{n_i + \alpha} p_i + \frac{\alpha}{n_i + \alpha} q \in \Delta^{K-1}_{++}. \tag{1}$$

Thus the logarithms are defined even when $n_{ic} = 0$. Let $C_K = I_K - K^{-1}\mathbf{1}\mathbf{1}^\top$. CLR coordinates and the normalized displacement are

$$\mathrm{clr}(p) = C_K \log p, \qquad s_i = \mathrm{clr}(\widetilde{p}_i) - \mathrm{clr}(q), \qquad d_i = \frac{s_i}{\max(\|s_i\|_2, \varepsilon)}. \tag{2}$$

Logs and vector ratios are componentwise. The simplex geometry enters through relative log-ratios, rather than an unconstrained difference of counts. In particular,

$$\mathbf{1}^\top s_i = 0, \qquad s_{ic} - s_{ik} = \log \frac{\widetilde{p}_{ic}/q_c}{\widetilde{p}_{ik}/q_k}. \tag{3}$$

A positive coordinate compares that class's relative representation with the geometric mean across classes. The zero-sum contrast preserves relative structure; normalization removes its overall magnitude above the ε floor. This construction uses simplex coordinates with ordinary local optimization. If $\widetilde{p}_i = q$, then $d_i = 0$. We allocate relative class emphasis, not rewards proportional to absolute class counts: a large client can have a negative coordinate for an underrepresented class. The ε floor makes the map continuous at zero, although directions above that floor remain sensitive near the federation distribution.

Class responses in parameter space. To turn d_i into a model reward, define a class margin from logits f_k on the labeled probe:

$$m_c(\theta) = \frac{1}{|V_c|} \sum_{x \in V_c} \left[f_c(x;\theta) - \frac{1}{K-1} \sum_{k \neq c} f_k(x;\theta) \right], \qquad R_{c,:} = \nabla_{E^\top \theta} m_c(\theta_\star)^\top. \tag{4}$$

For our linear classifier and fixed reference features, the margin response is exact: $m(\theta_\star + E\delta) - m(\theta_\star) = R\delta$. We use this operator to connect class targets to classifier parameters. It remains a reference-state model of response because the backbone subsequently changes during local training.

Let Q contain retained orthonormal directions of clean classifier-update history. To avoid placing the class reward directly in those historical directions, form

$$W = \mathrm{diag}((q{+}10^{-8})^{-1}), \qquad H = (I_M - QQ^\top)R^\top W^{1/2}, \qquad P = \mathrm{orth}(H_r), \quad G = RP. \tag{5}$$

H_r denotes the leading left singular directions, with $r \leq K-1$ capped by numerical rank. Thus $P \in \mathbb{R}^{M \times r}$, $P^\top P = I_r$, and $Q^\top P = 0$ in exact arithmetic. The constraint applies to the reward at construction.

The two operations in Equation (5) serve different roles. Multiplication by $W^{1/2}$ emphasizes responses according to federation class frequencies. The projection removes the retained historical component:

$$H = R^\top W^{1/2} - Q(Q^\top R^\top W^{1/2}), \qquad Q^\top H = (Q^\top - Q^\top QQ^\top)R^\top W^{1/2} = 0. \tag{6}$$

Selecting singular directions from this range preserves the constraint without constructing a dense projector. The requested $K-1$ carrier dimension matches the class contrast's dimension, but numerical rank can make the achievable response space smaller.

Regularized response fit. Writing the class reward as Pz, fit a desired margin vector $v \in \mathbb{R}^K$ while penalizing its parameter norm:

$$J(z;v) = (Gz - v)^\top W(Gz - v) + \beta^{-1} z^\top z. \tag{7}$$

Because P is orthonormal, $\|Pz\|_2 = \|z\|_2$. Differentiation yields the actual solve:

$$\nabla_z J = 2G^\top W(Gz - v) + 2\beta^{-1} z = 0,$$
$$(G^\top WG + \beta^{-1} I_r)z = G^\top W v, \qquad z(v) = (G^\top WG + \beta^{-1} I_r)^{-1} G^\top W v. \tag{8}$$

For $\beta > 0$, the coefficient matrix is positive definite even if G is rank deficient. One factorization serves all clients. The penalty is $1/\beta$; Section 4 characterizes the resulting response.

Contribution-modulated class amplitude. Let c_i be the fixed standalone clean-model score under the server-validation protocol, excluding final test. Define its ascending, zero-based average rank and the rank-based amplitude factor by

$$\rho_i = \frac{\mathrm{rank}_0(c_i)}{N-1}, \qquad f_i = 1 + \lambda(1 - 2\rho_i), \qquad \lambda = 0.25. \tag{9}$$

Ties receive their average rank. The score affects only the amplitude; d_i remains determined by class counts. For a fixed response fit and calibration, larger f_i means a larger parameter intervention, not a guaranteed utility gain for that client. Let $m_{\mathrm{ref}} > 0$ be the median reward norm in the frozen reference artifact. Using the unit-direction fit from Equation (8), set

$$h_i = \|Pz(d_i)\|_2, \quad a_0 = \frac{m_{\mathrm{ref}}}{\mathrm{median}_i(f_i h_i)}, \quad a_i = s a_0 f_i, \quad t_i = a_i d_i, \qquad s = 0.5. \tag{10}$$

The denominator must be positive. The single parameter reward is

$$z_i = z(t_i) = (G^\top W G + \beta^{-1} I_r)^{-1} G^\top W t_i, \qquad \boxed{\delta_i = P z_i}. \tag{11}$$

Since z is linear, $\mathrm{median}_i \|\delta_i\|_2 = s m_{\mathrm{ref}}$. The rank-based amplitude factor f_i and response norm h_i jointly determine the amplitude. The server saves one resulting tensor per client and fixes the construction quantities for subsequent training.

3.3 Cancellation and the complete training flow

A client starts from $x_{it,0} = \theta_t + E\delta_i$ and trains the entire model. For the stated momentum-SGD convention and local minibatch losses ℓ_{itj},

$$v_{it,0} = 0, \qquad v_{it,j+1} = \mu v_{it,j} + \nabla \ell_{itj}(x_{it,j}), \qquad x_{it,j+1} = x_{it,j} - \eta_t v_{it,j+1}. \tag{12}$$

The optimizer is fresh each communication round; momentum persists within that round. Let $F_{it}(\theta_t + E\delta_i) = x_{it,J_i}$. Direct cancellation and aggregation give

$$\widetilde{\phi}_{it} = F_{it}(\theta_t + E\delta_i) - E\delta_i, \qquad \theta_{t+1} = \sum_i w_i \widetilde{\phi}_{it}. \tag{13}$$

The client uploads its full trained model. The server then subtracts the exact tensor saved at injection, in the classifier scope only, before aggregation. Cancellation leaves all backbone parameters as trained. Expanding the local recursion shows exactly which term is removed:

$$\begin{aligned} x_{it,J_i} &= \theta_t + E\delta_i - \sum_{j=0}^{J_i-1} \eta_t v_{it,j+1}, \\ \widetilde{\phi}_{it} &= \theta_t - \sum_{j=0}^{J_i-1} \eta_t v_{it,j+1}, \qquad \theta_{t+1} = \theta_t - \sum_i w_i \sum_{j=0}^{J_i-1} \eta_t v_{it,j+1}. \end{aligned} \tag{14}$$

Cancellation exactly removes the known additive reward; optimization-induced effects remain in the velocities. The saved tensor is therefore sufficient for deployment, with the remaining discrepancy characterized below.

Algorithm 1 gives the deployment flow; Appendix C expands its construction and edge cases.

4 Mechanism analysis

A well-defined response map. Equations (2)–(3) retain relative class representation, and the ridge fit provides a unique response map even under rank deficiency. For $B = W^{1/2}G = U_B \Sigma V_B^\top$ and $\kappa = \beta^{-1}$,

$$W^{1/2} G z(v) = U_B \,\mathrm{diag}\left(\frac{\sigma_j^2}{\sigma_j^2 + \kappa}\right) U_B^\top W^{1/2} v. \tag{15}$$

Equation (15) identifies the part of the target supported by the selected carrier and the attenuation imposed by regularization. Stronger singular directions retain more of the weighted target. This makes class response sensitivity a concrete design criterion and motivates comparing the selected carrier with a random history-orthogonal alternative.

Algorithm 1 SICAR: simplex class reward with known offset cancellation

Input: Reference $\theta_\star$, clean history, counts n_{ic}, labeled V_c, scores c_i, calibration m_{ref}, start round $t_\star$, final round T

1: Compute d_i by (1)–(2); construct R, P, G by (4)–(5).
2: Solve (8), calibrate (10), and save the complete rewards (11).
3: Set $\theta_{t_\star} = \theta_\star$; freeze the reward construction.
4: **for** $t = t_\star, \dots, T-1$ **do**
5: Each client trains and uploads $\widehat{\theta}_{t+1}^{(i)} = F_{it}(\theta_t + E\delta_i)$.
6: Server subtracts saved $E\delta_i$ and aggregates $\theta_{t+1} = \sum_i w_i[\widehat{\theta}_{t+1}^{(i)} - E\delta_i]$.
7: **end for**
8: Return shared θ_T and dispatch models $\theta_T + E\delta_i$.

Relation to logit calibration. For features $h(x)$ and classifier perturbation $(\Delta W_i, \Delta b_i)$, the logit change is $\Delta W_i h(x) + \Delta b_i$. In the unrestricted classifier space, the bias response satisfies

$$R_b = \frac{K}{K-1} C_K, \qquad \Delta b = \frac{K-1}{K} v \implies R_b \Delta b = v \quad (\mathbf{1}^\top v = 0). \tag{16}$$

Thus bias shifts alone can realize a zero-sum margin target before our subspace and norm constraints are imposed. This connects our intervention to logit adjustment and FedLC (Menon et al., 2021; Zhang et al., 2022). Our fit also permits feature-dependent weight changes and enforces $\delta_i \in \text{range}(P)$; an unrestricted bias solution need not satisfy this constraint. We use $C - B$ to isolate cancellation for a fixed intervention. Improvement over clean A may also reflect label-skew calibration, which our carrier comparison does not separate from other effects of the chosen directions.

A criterion for useful cancellation. Fix a client, round, and realization of training randomness. With $a = E\delta_i$, write the counterfactual influence as $\tau = F(\theta+a) - F(\theta)$. The corrected discrepancy is exactly

$$e = F(\theta + a) - a - F(\theta) = \tau - a. \tag{17}$$

Expanding its squared norm gives the criterion

$$\|e\|_2^2 - \|\tau\|_2^2 = \|a\|_2^2 - 2\langle \tau, a \rangle, \qquad \|e\|_2 < \|\tau\|_2 \iff 2\langle \tau, a \rangle > \|a\|_2^2 . \tag{18}$$

We use this identity to specify a working regime: cancellation reduces the parameter discrepancy when the retained influence has sufficient alignment with the saved reward. For one SGD step, $e = -\eta[\nabla \ell(\theta + a) - \nabla \ell(\theta)]$; an L-Lipschitz minibatch gradient gives $\|e\| \le \eta L \|a\|$. This relates the residual to local step size and sensitivity. Appendix D derives the multi-step bounds and a scalar example delimiting the condition.

For our momentum setting, let x_j and x'_j be clean and rewarded local trajectories, $\xi_j = x'_j - x_j$, and b_j the difference of their velocities. Starting from $\xi_0 = a$ and $b_0 = 0$, the stepwise discrepancy obeys

$$b_{j+1} = \mu b_j + \nabla \ell_j(x'_j) - \nabla \ell_j(x_j), \qquad \xi_{j+1} = \xi_j - \eta_j b_{j+1}, \qquad e = \xi_J - a. \tag{19}$$

Equation (19) separates the known initial offset from accumulated changes in local gradients. At a common server state, the corrected aggregate discrepancy is $\sum_i w_i e_i$. The recurrence provides a direct connection between local response sensitivity and the behavior of the aggregation step.

A clean twin at the *same* server state measures τ for diagnosis, while the paired A/B/C endpoints measure the effect across training trajectories. These two comparisons answer complementary questions. The response diagnostics and their scope are given in Appendix F.13.

Training-time and terminal allocation. We differentiate models throughout training. A dispatch-only comparator would follow clean A and add the perturbation only at final delivery, preserving the shared endpoint by construction. Our $C - B$ comparison instead isolates subtraction during recurrent intervention. The distinction matters when clients receive intermediate models; endpoint accuracy alone cannot establish an advantage over terminal-only allocation. Measuring intermediate utility and retained information is a separate evaluation question.

5 EXPERIMENTS

We evaluate class-target realization together with its effect on shared-model performance. Experiments cover CIFAR-100, Tiny-ImageNet, and Street View House Numbers (SVHN) (Krizhevsky, 2009; Le & Yang, 2015; Netzer et al., 2011), with ResNet18, VGG16, and DenseNet121 (He et al., 2016; Simonyan & Zisserman, 2015; Huang et al., 2017) using two-group GroupNorm (Wu & He, 2018). Ten clients participate fully under label-Dirichlet concentration 0.5; the default strength is $s = 0.5$. We use one local epoch of SGD with momentum 0.9, and construct the fixed intervention after 12 clean reference rounds. Most settings use 200 total rounds; CIFAR-100 / DenseNet121 and SVHN / DenseNet121 use 100, and SVHN / VGG16 uses 120. Results give means and sample standard deviations over three seeds. Appendix E specifies learning rates, batch sizes, calibration, and data access. Client utility is evaluated at dispatch, before further local training.

Paired controls use clean training (A), reward retention (B), and SICAR with saved-offset cancellation (C). Thus, $C - B$ isolates cancellation's effect, while $C - A$ compares the full procedure with clean training. Fine Pearson measures alignment with the within-client class target d_i; Spearman measures utility ranking against standalone validation quality c_i. Both inputs enter our allocation policy, so these are realization metrics rather than independent tests of contribution validity. Definitions and implementation details appear in Appendices B and E.

5.1 MAIN RESULTS

Table 1 places SICAR alongside nine comparison methods on two illustrative settings; Appendix F retains the complete nine-setting inventory. We obtain mean global accuracy of 58.14% on CIFAR-100 / ResNet18 and 46.57% on Tiny-ImageNet / VGG16, with Spearman 0.841 and 0.904. These are strong ranking results in the displayed settings, although the full inventory includes settings in which other methods align better. We interpret cross-method outcomes as descriptive context because the implementations differ in allocation rules and data access. The matched A/B/C controls provide our primary mechanism evidence.

Table 1: Descriptive comparison on two illustrative settings. Global and client utility are accuracies (%); Spearman measures utility rank correlation with standalone validation quality. Values are means ± sample SD over three seeds. Bold: largest mean per column; dash: undefined correlation for identical client utilities.

	CIFAR-100 / ResNet18			Tiny-ImageNet / VGG16		
Method	Global	Client utility	Spearman	Global	Client utility	Spearman
FedAvg	54.43 ± 0.13	54.43 ± 0.13	–	$\mathbf{46.92} \pm 0.35$	$\mathbf{46.92} \pm 0.35$	–
CFFL	30.61 ± 3.07	27.25 ± 1.41	0.801 ± 0.102	28.12 ± 0.19	24.10 ± 0.20	0.652 ± 0.052
CGSV	48.60 ± 9.16	50.94 ± 5.00	0.471 ± 0.097	42.77 ± 1.68	43.27 ± 0.94	-0.419 ± 0.217
IAFL	41.46 ± 3.41	36.16 ± 5.29	0.606 ± 0.301	38.80 ± 1.90	33.30 ± 3.27	0.773 ± 0.091
ShapFed	36.08 ± 1.63	33.00 ± 0.81	0.582 ± 0.303	31.81 ± 0.80	23.90 ± 0.88	0.409 ± 0.385
Quota surrogate	55.45 ± 0.60	$\mathbf{55.45} \pm 0.60$	–	44.48 ± 0.13	44.48 ± 0.13	–
Aequa	43.83 ± 1.15	43.32 ± 1.16	0.665 ± 0.387	38.03 ± 0.31	37.03 ± 0.15	0.858 ± 0.089
FedAKD	27.59 ± 0.42	24.83 ± 0.33	0.746 ± 0.074	40.08 ± 0.51	27.00 ± 0.04	0.522 ± 0.117
FedRAC	45.02 ± 1.22	39.51 ± 2.76	0.674 ± 0.372	42.85 ± 0.68	37.29 ± 1.28	0.881 ± 0.082
SICAR	$\mathbf{58.14} \pm 0.72$	52.75 ± 1.07	$\mathbf{0.841} \pm 0.079$	46.57 ± 0.53	43.41 ± 0.51	$\mathbf{0.904} \pm 0.020$

The independent FedAvg runs are distinct from paired A: on CIFAR-100 / ResNet18, their global accuracies are 54.43% and 56.61%. The displayed settings use the same nominal training budget across methods, so budget alone does not explain this difference between runs. Some of our baseline adaptations also use historical test-informed scores, and Quota is a shared-model surrogate. We therefore avoid attributing SICAR–baseline gaps exclusively to the proposed mechanism. Extended-setting budgets and protocol differences are documented in Appendix E.

5.2 CLASS ALIGNMENT AND GLOBAL RECOVERY

Figure 2 jointly reports class alignment, global recovery, and dispatched utility across all nine settings. On all six CIFAR-100 and Tiny-ImageNet settings, SICAR improves mean global accuracy

over reward retention by 0.61–2.40 pp and over clean training by 0.12–1.53 pp. Fine Pearson remains between 0.765 and 0.855, improving over B in five of the six settings. On CIFAR-100 / ResNet18, the 2.40 ± 0.42-pp recovery accompanies alignment of 0.855 ± 0.009. These paired results support the joint objective of class-target realization and shared-model recovery in the tested regime. The completed Tiny-ImageNet / DenseNet121 setting gives $C - B = 0.92 \pm 0.33$ pp, $C - A = 0.60 \pm 0.29$ pp, and Fine 0.836 ± 0.015, extending this pattern to all three architectures on both object-recognition datasets.

	(a) Class alignment		(b) Global recovery		(c) Client utility	
Dataset / model	Fine B	Fine C	$C-A$	$C-B$	$C-A$	$C-B$
CIFAR-100 / ResNet18	0.447 ± 0.055	0.855 ± 0.009	+1.53 ± 0.35	+2.40 ± 0.42	−3.86 ± 0.89	−1.94 ± 0.92
CIFAR-100 / VGG16	0.516 ± 0.044	0.844 ± 0.015	+0.12 ± 0.84	+1.93 ± 0.30	−4.48 ± 0.45	−1.10 ± 0.65
CIFAR-100 / DenseNet121	0.835 ± 0.026	0.830 ± 0.029	+0.90 ± 0.44	+1.18 ± 0.76	−3.25 ± 0.68	+0.48 ± 0.20
Tiny-ImageNet / ResNet18	0.665 ± 0.029	0.833 ± 0.001	+0.49 ± 0.05	+1.20 ± 0.31	−3.52 ± 0.18	−0.85 ± 0.06
Tiny-ImageNet / VGG16	0.607 ± 0.014	0.765 ± 0.007	+0.41 ± 0.81	+0.61 ± 0.84	−2.75 ± 0.81	−1.09 ± 0.73
Tiny-ImageNet / DenseNet121	0.766 ± 0.015	0.836 ± 0.015	+0.60 ± 0.29	+0.92 ± 0.33	−3.15 ± 0.28	−0.43 ± 0.25
SVHN / ResNet18	0.010 ± 0.038	−0.040 ± 0.028	+0.11 ± 0.20	+0.02 ± 0.26	+0.11 ± 0.20	+0.02 ± 0.27
SVHN / VGG16	0.490 ± 0.128	0.602 ± 0.112	−0.04 ± 0.02	−0.03 ± 0.07	−0.09 ± 0.02	−0.02 ± 0.05
SVHN / DenseNet121	0.121 ± 0.071	0.067 ± 0.104	+0.04 ± 0.06	+0.01 ± 0.05	+0.05 ± 0.07	+0.02 ± 0.07

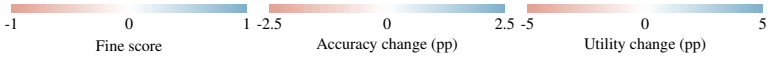

Figure 2: Class alignment and utility across all nine settings. A: clean training; B: reward retention; C: SICAR with saved-offset cancellation. Fine denotes within-client class Pearson. Cells report mean ± sample SD; difference SDs use paired seeds. Color scales are separate for alignment, global recovery, and dispatch-utility changes. Client utility is evaluated before further local training.

Global recovery and differentiated dispatch measure distinct outcomes. Mean dispatched utility is 2.75–4.48 pp below clean A on these six settings. The client-level records sharpen this result: all 180 client–seed comparisons on CIFAR-100 and Tiny-ImageNet have negative dispatch $C - A$. On CIFAR-100 / ResNet18, the within-seed minimum and maximum changes average -5.67 ± 1.00 and -1.89 ± 0.76 pp. Thus relative allocation alignment coexists with an absolute cost even for the least affected client; it does not establish a participation advantage over clean shared-model access. Table 5 gives extrema and positive-change counts for every setting. SVHN has near-zero global $C - B$ contrasts and architecture-dependent alignment, identifying a less responsive regime.

Local response evidence. Endpoint contrasts compare different training trajectories. To examine cancellation at a common state, we also compare a rewarded local update with its clean twin. With classifier influence $\tau_h = E^\top[F(\theta + E\delta) - F(\theta)]$, we measure

$$\eta_{\text{resp}} = 1 - \frac{\|R(\tau_h - \delta)\|}{\|R\tau_h\| + \varepsilon}. \tag{20}$$

A positive value indicates reduced discrepancy in the reference class-response space after subtraction. Across the six CIFAR-100 and Tiny-ImageNet settings, the mean seed-level medians range from 0.059 to 0.444; the seed-level medians are positive in all three seeds for each setting (Table 26). This connects global recovery with measured local response reduction. The diagnostic uses a fixed response projection, so it does not by itself establish the full-parameter condition in Equation (18). Together, the local and endpoint comparisons provide complementary evidence for when cancellation helps.

5.3 Carrier Ablation and Strength Sensitivity

Response-aware direction selection. The response-aware carrier improves global C accuracy over the random history-orthogonal carrier by 1.57 ± 0.28 pp and the $C - B$ contrast by 3.24 ± 0.22 pp (Figure 3a,b). Both gains occur in all three paired seeds. Macro-F1 also increases by 1.74 ± 0.39 pp. This supports selecting directions through class responses beyond imposing clean-history orthogonality alone. The complete carrier endpoints in Table 25 show an important boundary: random-carrier B is approximately 57.41%, above clean A (56.61%), while random-carrier C is 56.57%.

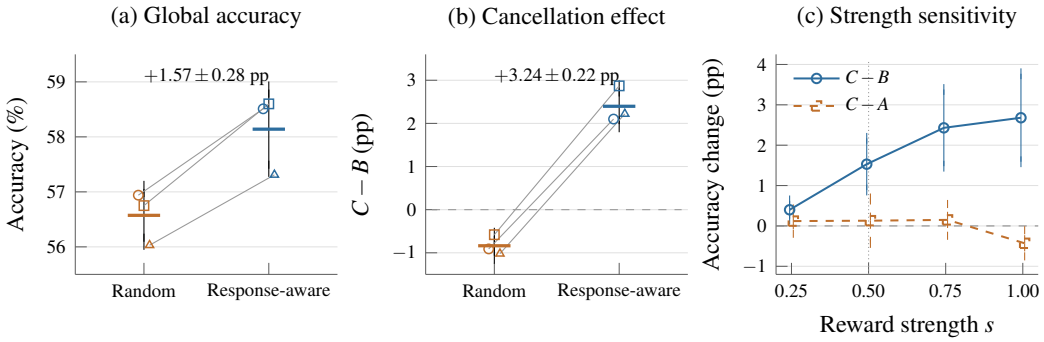

Figure 3: Carrier ablation and strength sensitivity on CIFAR-100 / ResNet18. (a,b) Test-endpoint comparisons of random and response-aware carriers. Symbols and connecting lines show paired seeds; horizontal marks and error bars show arm means and sample SDs computed from our two-decimal seed records. Annotations retain the directly computed paired gains. (c) Short-run server-validation scan, with mean ± sample SD; the dotted line marks the fixed default $s = 0.5$.

Retaining a perturbation therefore need not harm training; the benefit of cancellation depends on the induced local response. This ablation does not distinguish feature-dependent effects from bias calibration in Equation (16).

Strength and performance preservation. In the validation scan (Figure 3c), stronger interventions increase recovery over B, but $s = 1.0$ reduces accuracy relative to clean A (-0.43 ± 0.29 pp). The default $s = 0.5$ gives $C - B = 1.53 \pm 0.64$ pp and $C - A = 0.13 \pm 0.54$ pp, showing a positive mean contrast against both controls. The scan ends after 22 rounds on server validation, whereas the formal result uses the 200-round test endpoint. Both the training length and evaluation split differ; the 0.13-pp and 1.53-pp $C - A$ values therefore cannot isolate either source of variation. This short-run experiment describes sensitivity; it does not establish an optimal strength at the formal test endpoint.

Contribution validity and scope. We additionally examine a restricted coalition-value proxy with a frozen backbone and a single classifier update (Appendix F.13). Its correlations vary in sign across seeds and utility definitions, so this check leaves contribution validity unresolved. The current evidence supports realization of a declared relative-participation policy. It does not yet isolate the role of rank-based amplitude, establish resistance to client recalibration, or determine whether a bias-only intervention can reproduce the gains. These questions require different controls from the A/B/C comparison.

6 Conclusion

We connect relative class participation on the simplex to client-specific classifier rewards and known offset cancellation at aggregation. Paired recovery and carrier ablation provide the strongest evidence for the mechanism, with class alignment retained in representative CIFAR-100 and Tiny-ImageNet settings. The observed dispatch cost and weaker separation on SVHN define the current operating boundary. Our results establish a concrete mechanism for participation-aware allocation and characterize a regime of shared-model recovery. Independent class valuation, client adaptation, and comparisons with bias-only calibration remain directions for further evaluation.

AI USE STATEMENT

We used an LLM to assist with drafting and revising text, organizing experimental summaries, and generating LaTeX and visualization code. The numerical results come from our experimental records.

ETHICS STATEMENT

We study participation-aware model allocation in federated learning. Its application requires a justification for unequal client benefits and is distinct from fairness across protected demographic groups. We assume compliant clients and server access to client class counts and a labeled probe; these inputs can reveal information about participants, and our mechanism provides no additional privacy or truthful-reporting guarantee. Our experiments use image classification benchmarks without new human-subject data collection or deployment in sensitive decision-making settings.

REFERENCES

Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating Bias in Federated Learning, 2020. URL https://arxiv.org/abs/2012.02447. arXiv:2012.02447, preprint.

Nimesh Agrawal, Anuj Kumar Sirohi, Sandeep Kumar, and Jayadeva. No Prejudice! Fair Federated Graph Neural Networks for Personalized Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):10775–10783, 2024. doi: 10.1609/aaai.v38i10.28950. URL https://doi.org/10.1609/aaai.v38i10.28950.

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–177, 1982. doi: 10.1111/j.2517-6161.1982.tb01195.x.

Saba Amiri, Adam Belloum, Eric Nalisnick, Sander Klous, and Leon Gommans. On the impact of non-IID data on the performance and fairness of differentially private federated learning. In *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, pp. 52–58, 2022. doi: 10.1109/dsn-w54100.2022.00018. URL https://doi.org/10.1109/dsn-w54100.2022.00018.

Maryam Badar, Sandipan Sikdar, Wolfgang Nejdl, and Marco Fisichella. FairTrade: Achieving Pareto-Optimal Trade-Offs between Balanced Accuracy and Fairness in Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):10962–10970, 2024. doi: 10.1609/aaai.v38i10.28971. URL https://ojs.aaai.org/index.php/AAAI/article/view/28971.

Alycia N. Carey, Wei Du, and Xintao Wu. Robust Personalized Federated Learning under Demographic Fairness Heterogeneity. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 1425–1434, 2022. doi: 10.1109/bigdata55660.2022.10020554. URL https://doi.org/10.1109/bigdata55660.2022.10020554.

Hongyan Chang and Reza Shokri. Bias Propagation in Federated Learning, 2023. URL https://arxiv.org/abs/2309.02160. arXiv:2309.02160, preprint.

Xin Che, Jingdi Hu, Zirui Zhou, Yong Zhang, and Lingyang Chu. Training Fair Models in Federated Learning without Data Privacy Infringement, 2021. URL https://arxiv.org/abs/2109.05662. arXiv:2109.05662, preprint.

Chaomeng Chen, Zhenhong Zhou, Peng Tang, Longzhu He, and Sen Su. Enforcing group fairness in privacy-preserving Federated Learning. *Future Generation Computer Systems*, 160:890–900, 2024. doi: 10.1016/j.future.2024.06.040. URL https://doi.org/10.1016/j.future.2024.06.040.

Chunling Chen, Haiwei Pan, Kejia Zhang, Zhe Li, and Fengming Yu. A Federated Fairness-Aware Incentive Mechanism for medical image classification. *Engineering Applications of Artificial Intelligence*, 160:111855, 2025a. doi: 10.1016/j.engappai.2025.111855. URL https://doi.org/10.1016/j.engappai.2025.111855.

Guorong Chen, Chao Li, Wei Wang, Li Duan, Bin Wang, Zhen Han, and Xiangliang Zhang. FairReward: Towards Fair Reward Distribution Using Equity Theory in Blockchain-Based Federated Learning. *IEEE Transactions on Dependable and Secure Computing*, 22(2):1612–1626, 2025b. doi: 10.1109/tdsc.2024.3446864. URL https://doi.org/10.1109/tdsc.2024.3446864.

Hong-You Chen and Wei-Lun Chao. On Bridging Generic and Personalized Federated Learning for Image Classification. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2107.00778.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client Selection in Federated Learning: Convergence Analysis and Power-of-Choice Selection Strategies, 2020. URL https://arxiv.org/abs/2010.01243. arXiv:2010.01243, preprint.

Sujit Chowdhury, Aritra Bhaduri, and Raju Halder. FLAIR: Federated Learning with Augmented and Improved Feature Representations, 2024. URL https://openreview.net/forum?id=C7XoUdJ5ZC. OpenReview submission, posted September 2024; acceptance not verified.

Mingshu Cong, Han Yu, Xi Weng, Jiabao Qu, Yang Liu, and Siu Ming Yiu. A VCG-based Fair Incentive Mechanism for Federated Learning, 2020. URL https://arxiv.org/abs/2008.06680. arXiv:2008.06680, preprint.

Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. Addressing Algorithmic Disparity and Performance Inconsistency in Federated Learning. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2108.08435.

Yutong Dai, Zeyuan Chen, Junnan Li, Shelby Heinecke, Lichao Sun, and Ran Xu. Tackling Data Heterogeneity in Federated Learning with Class Prototypes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. URL https://arxiv.org/abs/2212.02758.

Yiqun Diao, Qinbin Li, and Bingsheng He. Exploiting Label Skews in Federated Learning with Model Concatenation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10): 11784–11792, 2024. doi: 10.1609/aaai.v38i10.29063. URL https://ojs.aaai.org/index.php/AAAI/article/view/29063.

Yasmine Djebrouni, Nawel Benarba, Ousmane Touat, Pasquale De Rosa, Sara Bouchenak, Angela Bonifati, Pascal Felber, Vania Marangozova, and Valerio Schiavoni. Bias Mitigation in Federated Learning for Edge Computing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(4):1–35, 2023. doi: 10.1145/3631455. URL https://doi.org/10.1145/3631455.

Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware Agnostic Federated Learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 181–189, 2021. doi: 10.1137/1.9781611976700.21. URL https://doi.org/10.1137/1.9781611976700.21.

Yuying Duan, Gelei Xu, Yiyu Shi, and Michael Lemmon. The Cost of Local and Global Fairness in Federated Learning, 2025. URL https://arxiv.org/abs/2503.22762. arXiv:2503.22762, preprint.

Yahya H. Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A. Salman Avestimehr. FairFed: Enabling Group Fairness in Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7494–7502, 2023. doi: 10.1609/aaai.v37i6.25911. URL https://ojs.aaai.org/index.php/AAAI/article/view/25911.

Jiashi Gao, Xin Yao, and Xuetao Wei. Anti-Matthew FL: Bridging the Performance Gap in Federated Learning to Counteract the Matthew Effect, 2023. URL https://arxiv.org/abs/2309.16338. arXiv:2309.16338, preprint.

Xiuting Gu, Zhu Tianqing, Jie Li, Tao Zhang, Wei Ren, and Kim-Kwang Raymond Choo. Privacy, accuracy, and model fairness trade-offs in federated learning. *Computers & Security*, 122: 102907, 2022. doi: 10.1016/j.cose.2022.102907. URL https://doi.org/10.1016/j.cose.2022.102907.

Kaiqi Guan, Wenke Huang, Xianda Guo, Yueyang Yuan, Bin Yang, and Mang Ye. Rethinking Fair Federated Learning from Parameter and Client View. In *Advances in Neural Information Processing Systems*, volume 38, pp. 18714–18735, 2025. doi: 10.52202/085713-0630. URL https://papers.nips.cc/paper_files/paper/2025/hash/1b1205172b8a57ad504002a081b37859-Abstract-Conference.html.

Kuangpu Guo, Yuhe Ding, Jian Liang, Zilei Wang, Ran He, and Tieniu Tan. Exploring Vacant Classes in Label-Skewed Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):16960–16968, 2025. doi: 10.1609/aaai.v39i16.33864. URL https://doi.org/10.1609/aaai.v39i16.33864.

Faisal Hamman and Sanghamitra Dutta. Demystifying Local and Global Fairness Trade-offs in Federated Learning Using Partial Information Decomposition, 2023. URL https://arxiv.org/abs/2307.11333. arXiv:2307.11333, preprint.

Mengde Han, Tianqing Zhu, and Wanlei Zhou. Fair Federated Learning with Opposite GAN. *Knowledge-Based Systems*, 287:111420, 2024. doi: 10.1016/j.knosys.2024.111420. URL https://doi.org/10.1016/j.knosys.2024.111420.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://arxiv.org/abs/1512.03385.

Shihao Hou, Xinyi Shang, Shreyank N Gowda, Yang Lu, Chao Wu, Yan Yan, and Hanzi Wang. CAPT: Class-Aware Prompt Tuning for Federated Long-Tailed Learning with Vision-Language Model, 2025. URL https://arxiv.org/abs/2503.06993. arXiv:2503.06993, preprint.

Guangzheng Hu, Patricia Menéndez, Feng Liu, Mingming Gong, Guanghui Wang, and Liuhua Peng. FedReLa: Imbalanced Federated Learning via Re-Labeling, 2026. URL https://arxiv.org/abs/2606.26037. arXiv:2606.26037, preprint.

Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Fair Federated Learning via Bounded Group Loss. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 140–160, 2024. doi: 10.1109/satml59370.2024.00015. URL https://doi.org/10.1109/satml59370.2024.00015.

Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. Federated Learning Meets Multi-Objective Optimization. *IEEE Transactions on Network Science and Engineering*, 9 (4):2039–2051, 2022. URL https://cs.uwaterloo.ca/~y328yu/publication/hu-szy-22/hu-szy-22.pdf.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017. URL https://arxiv.org/abs/1608.06993.

Meirui Jiang, Holger R Roth, Wenqi Li, Dong Yang, Can Zhao, Vishwesh Nath, Daguang Xu, Qi Dou, and Ziyue Xu. Fair Federated Medical Image Segmentation via Client Contribution Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2303.16520.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021. doi: 10.1561/2200000083. URL https://arxiv.org/abs/1912.04977.

Samhita Kanaparthy, Manisha Padala, Sankarshan Damle, and Sujit Gujar. Fair Federated Learning for Heterogeneous Data. In *Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, pp. 298–299, 2022. doi: 10.1145/3493700.3493750. URL https://doi.org/10.1145/3493700.3493750.

Zahra Kharaghani, Ali Dadras, and Tommy Löfstedt. FAIRVAR: Fair Federated Learning via Variance Regularization, 2025. URL https://arxiv.org/abs/2508.12042. arXiv:2508.12042, preprint.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. Technical report, Stanford University, 2015. URL https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf.

Bizhi Lei, Fan Chen, YuXin Xie, Shaobo Zhang, Entao Luo, and Qing Yang. FedDAR: Multidimensional Contribution Assessment and Adaptive Gradient Rewards for Collaborative Fairness in Heterogeneous Federated Learning. In *International Conference on Intelligent Computing*, pp. 100–113, 2025. doi: 10.1007/978-981-96-9884-4_9. URL https://link.springer.com/10.1007/978-981-96-9884-4_9.

Afei Li, JunHui Song, Zhangqi Zheng, Zhixin Xia, Jiadong Ren, and Yongshan Liu. From local bias to global consensus: Group-wise prototype federated learning under heterogeneous and cross-domain settings. *Information Sciences*, 741:123284, 2026a. doi: 10.1016/j.ins.2026.123284. URL https://doi.org/10.1016/j.ins.2026.123284.

Qi Li, YuCan Zhou, Hong Zhao, Jiang Zhou, and XiaoYan Gu. Decoupling and alleviating knowledge bias for federated learning under data heterogeneity. *Knowledge-Based Systems*, 337: 115421, 2026b. doi: 10.1016/j.knosys.2026.115421. URL https://doi.org/10.1016/j.knosys.2026.115421.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair Resource Allocation in Federated Learning. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1905.10497.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted Empirical Risk Minimization. In *International Conference on Learning Representations*, 2021a. URL https://arxiv.org/abs/2007.01162.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 6357–6368. PMLR, 2021b. URL https://proceedings.mlr.press/v139/li21h.html.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B. Allen, Randy P. Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think Locally, Act Globally: Federated Learning with Local and Global Representations, 2020. URL https://arxiv.org/abs/2001.01523. arXiv:2001.01523, preprint.

Changxin Liu, Zhenan Fan, Zirui Zhou, Yang Shi, Jian Pei, Lingyang Chu, and Yong Zhang. Achieving Model Fairness in Vertical Federated Learning, 2021. URL https://arxiv.org/abs/2109.08344. arXiv:2109.08344, preprint.

Yiqi Liu, Shan Chang, Ye Liu, Bo Li, and Cong Wang. FairFed: Improving Fairness and Efficiency of Contribution Evaluation in Federated Learning via Cooperative Shapley Value. In *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications*, pp. 621–630, 2024. doi: 10.1109/infocom52122.2024.10621438. URL https://doi.org/10.1109/infocom52122.2024.10621438.

Zelei Liu, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui. GTG-Shapley: Efficient and Accurate Participant Contribution Evaluation in Federated Learning. *ACM Transactions on Intelligent Systems and Technology*, 13(4):1–21, 2022a. doi: 10.1145/3501811. URL https://doi.org/10.1145/3501811.

Zelei Liu, Yuanyuan Chen, Yansong Zhao, Han Yu, Yang Liu, Renyi Bao, Jinpeng Jiang, Zaiqing Nie, Qian Xu, and Qiang Yang. Contribution-Aware Federated Learning for Smart Healthcare. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12396–12404, 2022b. doi: 10.1609/aaai.v36i11.21505. URL https://ojs.aaai.org/index.php/AAAI/article/view/21505.

Lingjuan Lyu, Xinyi Xu, and Qian Wang. Collaborative fairness in federated learning. In *International Workshop on Federated Learning for User Privacy and Data Confidentiality, in conjunction with IJCAI*, 2020. URL https://arxiv.org/abs/2008.12161.

Zhiyong Ma, Yuanjie Shi, Yan Yan, and Jian Chen. Federated Rényi Fair Inference in Federated Heterogeneous System. In *Conference on Uncertainty in Artificial Intelligence*, pp. 2804–2843, 2025. URL https://proceedings.mlr.press/v286/ma25a.html.

Samuel Maddock, Graham Cormode, and Carsten Maple. Private Federated Multiclass Post-hoc Calibration, 2025. URL https://arxiv.org/abs/2510.01987. arXiv:2510.01987, preprint.

Erich Malan, Valentino Peluso, Andrea Calimera, and Enrico Macii. Gradient-Aware Participation for Energy Reduction in Federated Learning with Extreme Label Skew. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2025. doi: 10.1109/ijcnn64981.2025.11227886. URL https://doi.org/10.1109/ijcnn64981.2025.11227886.

Sarim Malik, Muhammad Ahmad Ashraf, and Murtaza Taj. FedChill: Adaptive Temperature Scaling for Federated Learning in Heterogeneous Client Environments, 2025. URL https://openreview.net/forum?id=mRXv4cYncW. OpenReview submission, posted September 2025; acceptance not verified.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Ag"uera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Syed Irfan Ali Meerza, Luyang Liu, Jiaxin Zhang, and Jian Liu. GLOCALFAIR: Jointly Improving Global and Local Group Fairness in Federated Learning, 2024. URL https://arxiv.org/abs/2401.03562. arXiv:2401.03562, preprint.

Ninareh Mehrabi, Cyprien de Lichy, John McKay, Cynthia He, and William Campbell. Towards Multi-Objective Statistically Fair Federated Learning, 2022. URL https://arxiv.org/abs/2201.09917. arXiv:2201.09917, preprint.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2007.07314.

Gerry Windiarto Mohamad Dunda and Shenghui Song. Fairness-aware Federated Minimax Optimization with Convergence Guarantee. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pp. 563–568, 2024. doi: 10.1109/cai59869.2024.00111. URL https://doi.org/10.1109/cai59869.2024.00111.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic Federated Learning. In *Proceedings of the International Conference on Machine Learning*, 2019. URL https://arxiv.org/abs/1902.00146.

Noorain Mukhtiar, Adnan Mahmood, and Quan Z. Sheng. CoRe-Fed: Bridging collaborative and representation fairness via federated embedding distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24458–24466, 2026. doi: 10.1609/aaai.v40i29.39628. URL https://ojs.aaai.org/index.php/AAAI/article/view/39628.

Aniket Murhekar, Jiaxin Song, Parnian Shahkar, Bhaskar Ray Chaudhury, and Ruta Mehta. You get what you give: Reciprocally fair federated learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 45289–45310. PMLR, 2025. URL https://proceedings.mlr.press/v267/murhekar25a.html.

Mohamed Nafea, Eugine Shin, and Aylin Yener. Proportional Fair Clustered Federated Learning. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 2022–2027, 2022. doi: 10.1109/isit50566.2022.9834545. URL https://doi.org/10.1109/isit50566.2022.9834545.

Lokesh Nagalapatti and Ramasuri Narayanam. Game of Gradients: Mitigating Irrelevant Clients in Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. URL https://arxiv.org/abs/2110.12257.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. URL https://ai.stanford.edu/~twangcat/papers/nips2011_housenumbers.pdf.

Manisha Padala, Sankarshan Damle, and Sujit Gujar. Federated Learning Meets Fairness and Differential Privacy. In *International Conference on Neural Information Processing*, pp. 692–699, 2021. doi: 10.1007/978-3-030-92310-5_80. URL https://link.springer.com/chapter/10.1007/978-3-030-92310-5_80.

Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. Minimax Demographic Group Fairness in Federated Learning. In *2022 ACM Conference on Fairness Accountability and Transparency*, pp. 142–159, 2022. doi: 10.1145/3531146.3533081. URL https://doi.org/10.1145/3531146.3533081.

Haemin Park, Diego Klabjan, Martin W. Braun, Xiuqi Li, and Balakrishnan Ananthanarayanan. Class-Grouped Normalized Momentum and Faster Hyperparameter Exploration to Tackle Class Imbalance in Federated Learning, 2026. URL https://arxiv.org/abs/2607.01474. arXiv:2607.01474, preprint.

Sikha Pentyala, Nicola Neophytou, Anderson Nascimento, Martine De Cock, and Golnoosh Farnadi. PrivFairFL: Privacy-Preserving Group Fairness in Federated Learning, 2022. URL https://arxiv.org/abs/2205.11584. arXiv:2205.11584, preprint.

Raphael Poulain, Mirza Farhan Bin Tarek, and Rahmatollah Beheshti. Improving Fairness in AI Models on Electronic Health Records: The Case for Federated Learning Methods. In *2023 ACM Conference on Fairness Accountability and Transparency*, pp. 1599–1608, 2023. doi: 10.1145/3593013.3594102. URL https://doi.org/10.1145/3593013.3594102.

Adam Richardson, Aris Filos-Ratsikas, and Boi Faltings. Budget-Bounded Incentives for Federated Learning. In *Federated Learning*, volume 12500 of *Lecture Notes in Computer Science*, pp. 176–188. Springer, 2020. doi: 10.1007/978-3-030-63076-8_13. URL https://link.springer.com/chapter/10.1007/978-3-030-63076-8_13.

Borja Rodríguez-Gálvez, Filip Granqvist, Rogier van Dalen, and Matt Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers, 2021. URL https://arxiv.org/abs/2109.08604. arXiv:2109.08604, preprint.

Pramit Saha, Divyanshu Mishra, Felix Wagner, Konstantinos Kamnitsas, and J. Alison Noble. FedExIT - missing class-agnostic semi-supervised federated learning with extreme imbalance tackling scheme. *Information Fusion*, 130:104080, 2026. doi: 10.1016/j.inffus.2025.104080. URL https://doi.org/10.1016/j.inffus.2025.104080.

Teresa Salazar, Miguel Fernandes, Helder Araújo, and Pedro Henriques Abreu. FAIR-FATE: Fair Federated Learning with Momentum. In *International Conference on Computational Science*, pp. 524–538, 2023. doi: 10.1007/978-3-031-35995-8_37. URL https://link.springer.com/chapter/10.1007/978-3-031-35995-8_37.

Teresa Salazar, Helder Araújo, Alberto Cano, and Pedro Henriques Abreu. A Survey on Group Fairness in Federated Learning: Challenges, Taxonomy of Solutions and Directions for Future Research, 2024a. URL https://arxiv.org/abs/2410.03855. arXiv:2410.03855, preprint.

Teresa Salazar, João Gama, Helder Araújo, and Pedro Henriques Abreu. Unveiling Group-Specific Distributed Concept Drift: A Fairness Imperative in Federated Learning, 2024b. URL https://arxiv.org/abs/2402.07586. arXiv:2402.07586, preprint.

Khotso Selialia, Yasra Chandio, and Fatima M. Anwar. Mitigating Group Bias in Federated Learning for Heterogeneous Devices. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1043–1054, 2024. doi: 10.1145/3630106.3658954. URL https://doi.org/10.1145/3630106.3658954.

Xinyi Shang, Yang Lu, Gang Huang, and Hanzi Wang. Federated Learning on Heterogeneous and Long-Tailed Data via Classifier Re-Training with Federated Features. In *Thirty-First International Joint Conference on Artificial Intelligence*, volume 3, pp. 2218–2224, 2022. doi: 10.24963/ijcai.2022/308. URL https://www.ijcai.org/proceedings/2022/308.

Fanfan Shen, Wenzhang Su, Chao Xu, Zhiquan Liu, Jun Feng, and Yanxiang He. FedDPKD: Federated learning with dual-phase knowledge distillation for label distribution skew. *Information Processing & Management*, 63(5):104657, 2026. doi: 10.1016/j.ipm.2026.104657. URL https://doi.org/10.1016/j.ipm.2026.104657.

Tao Sheng, Chengchao Shen, Yuan Liu, Yeyu Ou, Zhe Qu, and Jianxin Wang. Modeling Global Distribution for Federated Learning with Label Distribution Skew, 2022. URL https://arxiv.org/abs/2212.08883. arXiv:2212.08883, preprint.

Jiangming Shi, Shanshan Zheng, Xiangbo Yin, Yang Lu, Yuan Xie, and Yanyun Qu. CLIP-Guided Federated Learning on Heterogeneity and Long-Tailed Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):14955–14963, 2024a. doi: 10.1609/aaai.v38i13.29416. URL https://ojs.aaai.org/index.php/AAAI/article/view/29416.

Yuxin Shi, Zelei Liu, Zhuan Shi, and Han Yu. Fairness-Aware Client Selection for Federated Learning. In *IEEE International Conference on Multimedia and Expo*, 2023. URL https://arxiv.org/abs/2307.10738.

Zhuan Shi, Lan Zhang, Zhenyu Yao, Lingjuan Lyu, Cen Chen, Li Wang, Junhao Wang, and Xiang-Yang Li. FedFAIM: A Model Performance-Based Fair Incentive Mechanism for Federated Learning. *IEEE Transactions on Big Data*, 10(6):1038–1050, 2024b. doi: 10.1109/tbdata.2022.3183614. URL https://doi.org/10.1109/tbdata.2022.3183614.

Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low. Collaborative machine learning with incentive-aware model rewards. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 8927–8936. PMLR, 2020. URL https://proceedings.mlr.press/v119/sim20a.html.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1409.1556.

Tianshu Song, Yongxin Tong, and Shuyue Wei. Profit Allocation for Federated Learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 2577–2586, 2019. doi: 10.1109/bigdata47090.2019.9006327. URL https://doi.org/10.1109/bigdata47090.2019.9006327.

Cong Su, Guoxian Yu, Jun Wang, Hui Li, Qingzhong Li, and Han Yu. Multi-Dimensional Fair Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):15083–15090, 2024. doi: 10.1609/aaai.v38i13.29430. URL https://ojs.aaai.org/index.php/AAAI/article/view/29430.

Qiheng Sun, Xiang Li, Jiayao Zhang, Li Xiong, Weiran Liu, Jinfei Liu, Zhan Qin, and Kui Ren. ShapleyFL: Robust Federated Learning Based on Shapley Value. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2096–2108, 2023. doi: 10.1145/3580305.3599500. URL https://doi.org/10.1145/3580305.3599500.

Nurbek Tastan, Samar Fares, Toluwani Aremu, Samuel Horváth, and Karthik Nandakumar. Redefining contributions: Shapley-driven federated learning. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pp. 5009–5017, 2024. doi: 10.24963/ijcai.2024/554. URL https://www.ijcai.org/proceedings/2024/554.

Nurbek Tastan, Samuel Horvath, and Karthik Nandakumar. CYCle: Choosing Your Collaborators Wisely to Enhance Collaborative Fairness in Decentralized Learning, 2025a. URL https://arxiv.org/abs/2501.12344. arXiv:2501.12344, preprint.

Nurbek Tastan, Samuel Horváth, and Karthik Nandakumar. Aequa: Fair model rewards in collaborative learning via slimmable networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 59210–59236. PMLR, 2025b. URL https://proceedings.mlr.press/v267/tastan25a.html.

Sebastian Shenghong Tay, Xinyi Xu, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Incentivizing Collaboration in Machine Learning via Synthetic Data Rewards. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. URL https://arxiv.org/abs/2112.09327.

Ousmane Touat and Sara Bouchenak. Towards Robust and Bias-free Federated Learning. In *Proceedings of the 3rd Workshop on Machine Learning and Systems*, pp. 49–55, 2023. doi: 10.1145/3578356.3592576. URL https://doi.org/10.1145/3578356.3592576.

Asim Ukaye, Nurbek Tastan, Mubarak Abdu-Aguye, and Karthik Nandakumar. Class-wise contribution estimation via logit maximization for federated learning, 2026. URL https://arxiv.org/pdf/2605.18892. arXiv:2605.18892v2, preprint.

Ganghua Wang, Ali Payani, Myungjin Lee, and Ramana Kompella. Mitigating Group Bias in Federated Learning: Beyond Local Fairness, 2023. URL https://arxiv.org/abs/2305.09931. arXiv:2305.09931, preprint.

Guan Wang, Charlie Xiaoqian Dang, and Ziye Zhou. Measure Contribution of Participants in Federated Learning, 2019. URL https://arxiv.org/abs/1909.08525. arXiv:1909.08525, preprint.

Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Addressing Class Imbalance in Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):10165–10173, 2021a. doi: 10.1609/aaai.v35i11.17219. URL https://ojs.aaai.org/index.php/AAAI/article/view/17219.

Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A Principled Approach to Data Valuation for Federated Learning. In *Federated Learning*, volume 12500 of *Lecture Notes in Computer Science*, pp. 153–167. Springer, 2020. doi: 10.1007/978-3-030-63076-8_11. URL https://link.springer.com/chapter/10.1007/978-3-030-63076-8_11.

Yingchao Wang, Wenqi Niu, and Hanpo Hou. FedAKD: Federated Adaptive Knowledge Distillation via Global Knowledge Calibration and Decoupling. In *Proceedings of the ACM Web Conference 2026*, pp. 5415–5426, 2026a. doi: 10.1145/3774904.3792386. URL https://doi.org/10.1145/3774904.3792386.

Zheng Wang, Xiaoliang Fan, Jianzhong Qi, Chenglu Wen, Cheng Wang, and Rongshan Yu. Federated Learning with Fair Averaging. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2021b. URL https://arxiv.org/abs/2104.14937.

Zihui Wang, Zhaopeng Peng, Xiaoliang Fan, Zheng Wang, Shangbin Wu, Rongshan Yu, Peizhen Yang, Chuanpan Zheng, and Cheng Wang. FedAVE: Adaptive data value evaluation framework for collaborative fairness in federated learning. *Neurocomputing*, 574:127227, 2024a. doi: 10.1016/j.neucom.2023.127227. URL https://doi.org/10.1016/j.neucom.2023.127227.

Zihui Wang, Zheng Wang, Lingjuan Lyu, Zhaopeng Peng, Zhicheng Yang, Chenglu Wen, Rongshan Yu, Cheng Wang, and Xiaoliang Fan. FedSAC: Dynamic submodel allocation for collaborative fairness in federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024b. doi: 10.1145/3637528.3671748. URL https://arxiv.org/abs/2405.18291.

Zihui Wang, Yuhang Fu, Mengmeng Du, Zhimin Yuan, Yachen Liu, Weisheng Liao, Kaiyu Wang, and Zheng Wang. FedRAC: Rolling submodel allocation for collaborative fairness in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_FedRAC_Rolling_Submodel_Allocation_for_Collaborative_Fairness_in_Federated_Learning_CVPR_2026_paper.html.

Nannan Wu, Li Yu, Xuefeng Jiang, Kwang-Ting Cheng, and Zengqiang Yan. FedNoRo: Towards Noise-Robust Federated Learning by Addressing Class Imbalance and Label Noise Heterogeneity, 2023. URL https://arxiv.org/abs/2305.05230. arXiv:2305.05230, preprint.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018. URL https://arxiv.org/abs/1803.08494.

Zhaoxuan Wu, Mohammad Mohammadi Amiri, Ramesh Raskar, and Bryan Kian Hsiang Low. Incentive-aware federated learning with training-time model rewards. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/a6250ffb9403aa7fa5304f525b5ae0e4-Abstract-Conference.html.

Zikai Xiao, Zihan Chen, Songshang Liu, Hualiang Wang, Yang Feng, Jin Hao, Joey Tianyi Zhou, Jian Wu, Howard Hao Yang, and Zuozhu Liu. Fed-GraB: Federated Long-tailed Learning with Self-Adjusting Gradient Balancer. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f4b8ddb9b1aa3cb11462d64a70b84db2-Paper-Conference.pdf.

Xinyi Xu and Lingjuan Lyu. A Reputation Mechanism Is All You Need: Collaborative Fairness and Adversarial Robustness in Federated Learning, 2020. URL https://arxiv.org/abs/2011.10464. arXiv:2011.10464, preprint.

Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16104–16117, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8682cc30db9c025ecd3fee433f8ab54c-Abstract.html.

Yurui Yang and Bo Jiang. Towards Group Fairness via Semi-Centralized Adversarial Training in Federated Learning. In *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*, pp. 482–487, 2022. doi: 10.1109/mdm55031.2022.00103. URL https://doi.org/10.1109/mdm55031.2022.00103.

Mang Ye, Wenke Huang, Zekun Shi, He Li, and Bo Du. Revisiting federated learning with label skew: an over-confidence perspective. *Science China Information Sciences*, 68(9), 2025. doi: 10.1007/s11432-023-4202-1. URL https://doi.org/10.1007/s11432-023-4202-1.

Rongguang Ye, Wei-Bin Kou, and Ming Tang. PraFFL: A Preference-Aware Scheme in Fair Federated Learning, 2024. URL https://arxiv.org/abs/2404.08973. arXiv:2404.08973, preprint.

Qichuan Yin, Zexian Wang, Junzhou Huang, Huaxiu Yao, and Linjun Zhang. Distribution-Free Fair Federated Learning with Small Samples, 2024. URL https://arxiv.org/abs/2402.16158. arXiv:2402.16158, preprint.

Di Yu, Xin Du, Linshan Jiang, Huijing Zhang, and Shuiguang Deng. Exploiting Label Skewness for Spiking Neural Networks in Federated Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 6895–6903, 2025a. URL https://www.ijcai.org/proceedings/2025/0767.pdf.

Han Yu, Zelei Liu, Yang Liu, Tianjian Chen, Mingshu Cong, Xi Weng, Dusit Niyato, and Qiang Yang. A Fairness-aware Incentive Scheme for Federated Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 393–399, 2020. doi: 10.1145/3375627.3375840. URL https://doi.org/10.1145/3375627.3375840.

Tianrun Yu, Jiaqi Wang, Haoyu Wang, Mingquan Lin, Han Liu, Nelson S. Yee, and Fenglong Ma. Towards collaborative fairness in federated learning under imbalanced covariate shift. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3645–3656, 2025b. doi: 10.1145/3711896.3737161. URL https://arxiv.org/abs/2507.08617.

Rongfei Zeng, Shixun Zhang, Jiaqi Wang, and Xiaowen Chu. FMore: An Incentive Scheme of Multi-dimensional Auction for Federated Learning in MEC. In *IEEE International Conference on Distributed Computing Systems*, 2020. doi: 10.1109/ICDCS47774.2020.00094. URL https://arxiv.org/abs/2002.09699.

Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving Fairness via Federated Learning, 2021. URL https://arxiv.org/abs/2110.15545. arXiv:2110.15545, preprint.

Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Federated Learning with Local Fairness Constraints. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 1937–1942, 2023. doi: 10.1109/isit54713.2023.10206590. URL https://doi.org/10.1109/isit54713.2023.10206590.

Daniel Yue Zhang, Ziyi Kou, and Dong Wang. FairFL: A Fair Federated Learning Approach to Reducing Demographic Bias in Privacy-Sensitive Classification Models. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 1051–1060, 2020. doi: 10.1109/bigdata50022.2020.9378043. URL https://doi.org/10.1109/bigdata50022.2020.9378043.

Fengda Zhang, Kun Kuang, Yuxuan Liu, Long Chen, Chao Wu, Fei Wu, Jiaxun Lu, Yunfeng Shao, and Jun Xiao. Unified Group Fairness on Federated Learning, 2021. URL https://arxiv.org/abs/2111.04986. arXiv:2111.04986, preprint.

Fengda Zhang, Zitao Shuai, Kun Kuang, Fei Wu, Yueting Zhuang, and Jun Xiao. Unified fair federated learning for digital healthcare. *Patterns*, 5(1):100907, 2024. doi: 10.1016/j.patter.2023.100907. URL https://doi.org/10.1016/j.patter.2023.100907.

Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated Learning with Label Distribution Skew via Logits Calibration. In *International Conference on Machine Learning*, pp. 26311–26329, 2022. URL https://proceedings.mlr.press/v162/zhang22p.html.

Jing Zhang, Chuanwen Li, Jianzgong Qi, and Jiayuan He. A Survey on Class Imbalance in Federated Learning, 2023. URL https://arxiv.org/abs/2303.11673. arXiv:2303.11673, preprint.

Kejun Zhang, Jun Wang, Wenbin Wang, Taiheng Zeng, Pengcheng Li, Xunxi Wang, and Tingrui Zhang. Federated learning with heterogeneous data and models based on global decision boundary distillation. *Journal of King Saud University Computer and Information Sciences*, 37(5), 2025a. doi: 10.1007/s44443-025-00097-0. URL https://doi.org/10.1007/s44443-025-00097-0.

Li Zhang, Zhongxuan Han, XiaoHua Feng, Jiaming Zhang, Yuyuan Li, and Chaochao Chen. FedFACT: A Provable Framework for Controllable Group-Fairness Calibration in Federated Learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 16177–16228, 2025b. doi: 10.52202/085713-0546. URL https://papers.nips.cc/paper_files/paper/2025/hash/17ca2b635fdb94dbd3292da790e3fd46-Abstract-Conference.html.

A Related work

A.1 Collaborative fairness and contribution valuation

Federated learning exchanges model information while keeping raw training data at its source (McMahan et al., 2017; Kairouz et al., 2021). Collaborative fairness relates the benefit received by each participant to its contribution (Lyu et al., 2020; Sim et al., 2020). This objective differs from equalizing client performance (Mohri et al., 2019; Li et al., 2020), demographic group fairness (Salazar et al., 2024a), and predictive accuracy under class imbalance (Zhang et al., 2023). We study participation-aware model allocation: relative class composition specifies a class-response direction, and standalone validation rank modulates its amplitude. We evaluate how well this policy controls dispatched utility and how its training-time intervention affects the shared model.

Contribution valuation provides several possible allocation principles. Grouped influence measures evaluate instances or features (Wang et al., 2019), while profit allocation explicitly distributes proceeds according to estimated contribution (Song et al., 2019). Federated Shapley values adapt valuation to sequential training (Wang et al., 2020); GTG-Shapley uses update reconstruction, guided sampling, and truncation to reduce its cost (Liu et al., 2022a). Game of Gradients identifies relevant participants, including those relevant to a target label (Nagalapatti & Narayanam, 2021). ShapleyFL (Sun et al., 2023), the contribution-evaluation FairFed (Liu et al., 2024), contribution-aware healthcare learning (Liu et al., 2022b), and FedAVE (Wang et al., 2024a) use valuation to guide participation or aggregation. We use relative class participation as an allocation input rather than estimating a marginal value for each client–class pair. Accordingly, our class alignment metric evaluates the realization of that input, while an independent valuation would assess the input itself.

Incentive mechanisms also differ in their utility and resource definitions. Fairness-aware incentives consider participation timing and costs (Yu et al., 2020); budget-bounded incentives consider expenditure and redundancy (Richardson et al., 2020). FMore uses a multidimensional procurement auction (Zeng et al., 2020), and VCG-based incentives use a mechanism-design formulation (Cong et al., 2020). FedFAIM allocates model-performance rewards (Shi et al., 2024b), FairReward studies equity-based reward distribution (Chen et al., 2025b), and synthetic-data rewards provide an alternative to monetary payments or predictors (Tay et al., 2022). Reciprocally fair FL relates transfers, costs, and participant payoffs (Murhekar et al., 2025). We evaluate the predictive utility of a dispatched model under compliant participation; payments, truthful reporting, and equilibrium behavior require additional specifications.

A.2 Model rewards and class-specific structure

CFFL differentiates received models using reputation (Lyu et al., 2020). Incentive-aware collaborative learning studies contribution-dependent model rewards (Sim et al., 2020), and RFFL combines reputation with collaborative fairness and adversarial robustness (Xu & Lyu, 2020). CGSV/GDR allocates gradient rewards using gradient similarity (Xu et al., 2021); FedDAR combines multidimensional contribution assessment with adaptive gradient rewards (Lei et al., 2025). IAFL addresses the effect of training-time rewards on subsequent shared learning through stochastic reference recovery (Wu et al., 2024). We instead save each additive classifier perturbation and subtract that tensor before aggregation. Equation (17) describes the optimization-induced discrepancy that remains after this operation.

Capacity allocation provides a different control. FedSAC dynamically allocates submodels (Wang et al., 2024b), FedRAC uses rolling submodel allocation (Wang et al., 2026b), and Aequa uses slimmable networks (Tastan et al., 2025b). We retain a common architecture and vary classifier parameters. FairFedCS changes client-selection probabilities using participation and reputation (Shi et al., 2023). CYCle studies collaborator choice and reputation and shows why accuracy correlations need to be considered alongside collaboration gains (Tastan et al., 2025a); we therefore report absolute dispatch utility as well as alignment.

The closest class-based approaches use class information at different stages. ShapFed computes class-specific classifier similarities and averages them into a client weight for aggregation and personalized mixing (Tastan et al., 2024). CELM uses logit maximization to estimate class wise contribution evidence without server data or client count metadata (Ukaye et al., 2026). FedFIM uses prototype-based contribution estimates and prototype-level aggregation (Chen et al., 2025a). We re-

tain a vector of relative class participation as the target of a parameter-response fit. This construction preserves class structure through the allocation step, while its marginal-contribution interpretation remains an empirical question.

FedCE evaluates contribution in gradient and data spaces for medical segmentation (Jiang et al., 2023); collaborative-fairness FedAKD addresses imbalanced covariate shift through distillation (Yu et al., 2025b); CoRe-Fed combines representation alignment and contribution-aware aggregation (Mukhtiar et al., 2026). We evaluate label heterogeneity and classifier perturbations. Our baseline implementations and data-access choices are described separately from these methods' original formulations.

A.3 Client performance fairness and personalization

Agnostic FL optimizes against mixtures of client distributions (Mohri et al., 2019); q-FFL encourages more uniform client performance through a resource-allocation objective (Li et al., 2020). TERM changes the influence of losses through exponential tilting (Li et al., 2021a). FedMGDA+ uses multi-objective optimization (Hu et al., 2022), FedFV modifies conflicting gradients (Wang et al., 2021b), and variance regularization further addresses client disparities (Kharaghani et al., 2025). FedPW separates parameter selection from client weighting (Guan et al., 2025). We use a rank-based amplitude rule for model allocation, rather than optimize these performance-fairness objectives. Power-of-Choice studies the effect of biased client selection (Cho et al., 2020); our experiments use full participation.

Personalization changes another aspect of utility. LG-FedAvg combines local representations with global learning (Liang et al., 2020), Ditto addresses fairness and robustness with personalized models (Li et al., 2021b), and Fed-RoD separates generic and personalized prediction (Chen & Chao, 2022). Other work considers heterogeneous fairness requirements (Carey et al., 2022) or clients' fairness–performance preferences (Ye et al., 2024). We evaluate dispatched models on a common endpoint distribution before additional local adaptation. This measures a different utility from evaluating a personalized predictor on its own client's data distribution.

A.4 Statistical group fairness

Group fairness in FL has been addressed through preprocessing and training interventions (Abay et al., 2020), demographic-fairness objectives in FairFL (Zhang et al., 2020), unknown test distributions in AgnosticFair (Du et al., 2021), decentralized estimation in FedFair (Che et al., 2021), and heterogeneous client data (Kanaparthy et al., 2022). FCFL combines algorithmic disparity and performance inconsistency (Cui et al., 2021); FedVal considers multiple statistical fairness objectives (Mehrabi et al., 2022); FMDA-M considers client or demographic groups (Zhang et al., 2021); FairSCAT uses semi-centralized adversarial training (Yang & Jiang, 2022). Minimax demographic fairness (Papadaki et al., 2022), proportional fairness in clustered learning (Nafea et al., 2022), and bounded group loss (Hu et al., 2024) define further explicit criteria.

Local and global fairness can differ. FedFB compares federated and local fairness (Zeng et al., 2021); analyses of local constraints (Zeng et al., 2023) and FedGFT (Wang et al., 2023) examine what survives aggregation. Information-theoretic trade-offs (Hamman & Dutta, 2023), GLOCAL-FAIR (Meerza et al., 2024), and Anti-Matthew FL (Gao et al., 2023) address different combinations of local, global, and client disparities. Aggregation approaches include fairness-aware momentum (Salazar et al., 2023), the group-fairness FairFed (Ezzeldin et al., 2023), heterogeneous-device bias mitigation (Selialia et al., 2024), and Astral (Djebrouni et al., 2023). Other formulations use opposite generative adversarial networks (Han et al., 2024), augmented-Lagrangian minimax optimization (Mohamad Dunda & Song, 2024), or gradient-conflict handling for joint group and client fairness (Su et al., 2024).

Privacy further changes the information available for fairness control. Related work studies fairness with differential privacy (Padala et al., 2021), differential multipliers (Rodríguez-Gálvez et al., 2021), secure computation in PrivFairFL (Pentyala et al., 2022), private non-IID learning (Amiri et al., 2022), and privacy–accuracy–fairness trade-offs (Gu et al., 2022). FairVFL considers vertically partitioned features (Liu et al., 2021), and privacy-preserving enforcement provides another

approach (Chen et al., 2024). We assume server access to client class counts and a labeled probe, and we provide no additional privacy protection for those quantities.

Further studies connect fairness to robustness (Touat & Bouchenak, 2023), bias propagation (Chang & Shokri, 2023), concept drift (Salazar et al., 2024b), electronic health records (Poulain et al., 2023), digital healthcare (Zhang et al., 2024), and personalized recommendation (Agrawal et al., 2024). FairTrade studies accuracy–fairness optimization (Badar et al., 2024); FedFaiREE develops post-processing guarantees (Yin et al., 2024); FedFACT controls local and global calibration (Zhang et al., 2025b). The accuracy cost of jointly imposing fairness constraints (Duan et al., 2025) and dependence-based fairness in FedRényi (Ma et al., 2025) further illustrate the variety of objectives. We use these works to distinguish demographic fairness from our participation-aware allocation objective.

A.5 Label heterogeneity, missing classes, and calibration

Class imbalance can occur locally and globally (Zhang et al., 2023). Ratio Loss combines imbalance monitoring and a modified loss (Wang et al., 2021a); FedLC calibrates logits using class occurrence (Zhang et al., 2022); CReFF retrains a classifier using federated features (Shang et al., 2022); FedNH combines uniform and semantic class prototypes (Dai et al., 2023). FedMGD models a global distribution (Sheng et al., 2022), and Fed-GraB adjusts gradient balance (Xiao et al., 2023). These methods motivate an alternative explanation for part of our global accuracy gains: class-informed perturbations can also influence optimization under label skew. Our paired C–B contrast identifies the effect of removing the saved offset within our procedure; C–A alone does not separate this effect from label-skew correction.

Missing classes can amplify local forgetting. FedVLS uses distillation and logit suppression (Guo et al., 2025); over-confidence analysis connects label skew with logits and classifier bias (Ye et al., 2025); FedLEC combines label calibration and distillation for spiking networks (Yu et al., 2025a); FedExIT studies extreme imbalance with missing classes (Saha et al., 2026). We treat access to cross-client class knowledge as an allocation choice under the stated participation policy, rather than evidence of malicious behavior.

Different resources support further approaches: CLIP2FL transfers vision-language knowledge and retrains the classifier (Shi et al., 2024a); CAPT uses class-aware prompts (Hou et al., 2025); FedConcat combines models for label-distribution clusters (Diao et al., 2024); FedNoRo handles imbalance and label noise (Wu et al., 2023). FedReLa reallocates labels (Hu et al., 2026), while groupwise momentum normalization and resampling address other aspects of skew (Park et al., 2026). Global knowledge calibration in the WWW FedAKD paper (Wang et al., 2026a), dual-phase distillation in FedDPKD (Shen et al., 2026), knowledge-bias decoupling (Li et al., 2026b), groupwise prototypes (Li et al., 2026a), and decision-boundary distillation (Zhang et al., 2025a) also modify knowledge sharing. The WWW FedAKD formulation is distinct from the collaborative-fairness FedAKD formulation (Yu et al., 2025b).

FLAIR studies augmented feature representations (Chowdhury et al., 2024), and FedChill studies adaptive temperature scaling (Malik et al., 2025). Post-hoc federated calibration relates predicted probabilities to outcomes (Maddock et al., 2025); gradient-aware participation under extreme skew considers energy and cohort size (Malan et al., 2025). We fit a desired class-margin response under a fixed participation protocol. Centered log-ratios provide the compositional representation (Aitchison, 1982); our contribution is to connect that class vector to a constrained response fit and saved-offset cancellation. The response analysis and absolute-utility measurements describe both the control this construction provides and the effects that persist through local training.

B Additional problem analysis and metric definitions

We evaluate shared-model quality, received utility, and allocation alignment separately. We define each quantity below and describe our experimental protocol in Appendix E.

B.1 EVALUATION STAGES

For a common labeled evaluation set $\mathcal{T}$, define

$$\mathrm{Acc}(\theta;\mathcal{T}) = \frac{1}{|\mathcal{T}|}\sum_{(x,y)\in\mathcal{T}} \mathbf{1}\{\arg\max_k f_k(x;\theta) = y\}. \tag{21}$$

The global statistic uses θ_T^C. Client i's dispatch statistic uses $\theta_T^C + E\delta_i$. We compute mean dispatch utility as the unweighted average $N^{-1}\sum_i u_i$. A sample-weighted client average, if desired, is a different quantity and should be named separately. Reported accuracy percentages multiply Equation (21) by 100. A difference of reported percentages is a percentage-point difference, not a relative percent gain.

Class accuracy uses $\mathcal{T}_c = \{(x,y) \in \mathcal{T} : y = c\}$. An empty $\mathcal{T}_c$ makes this accuracy undefined; it must not be assigned zero. Local training absence does not make evaluation accuracy undefined if the common evaluation set contains that class. This distinction matters when label-skewed clients are evaluated over the entire label space.

The paired clean control is evaluated on the same intended distribution to form $\Delta u_{ic} = u_{ic}^C - u_c^A$. Because the control dispatches the same global model, u_c^A has no client index. By contrast, standalone models are separately trained on individual clients' data. Their accuracies must not be confused with the clean federated control. A method can outperform all standalone models while underperforming a shared model for every client.

We report macro-F1 as a percentage (100 times its zero-to-one value). Its usual definition averages $2\mathrm{TP}_c/(2\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c)$ over declared evaluation classes, where the counts come from a global-model confusion matrix. We retain the macro-F1 outputs of our evaluation implementation. Recomputing them requires its zero-denominator convention as well as the per-class confusion counts; aggregate accuracy alone is insufficient.

B.2 ACROSS-CLIENT ALIGNMENT

For $x_i = c_i$ and $y_i = u_i$, define the sample Pearson correlation

$$\rho(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}. \tag{22}$$

This statistic measures a linear relation across the ten clients within one run. It is undefined when either vector is constant.

Spearman correlation applies Equation (22) to average ranks of x and y. We use average ranks for ties in both vectors, including contribution ranks. Exact recomputation requires full-precision scores and the evaluator tie tolerance.

We also compare contribution and utility orderings through pairwise agreement, excluding contribution ties and assigning half credit to utility ties. Shared-model methods therefore score 0.5. This statistic uses the same contribution proxy as Pearson and Spearman; none of the three independently validates that proxy.

Pearson, Spearman, and pairwise agreement need not rank methods identically. Pearson is sensitive to the spacing of utility values; rank measures ignore that spacing. Very small accuracy differences can create an apparently good ranking without a practically meaningful separation of benefits. Conversely, high utility near the accuracy ceiling can leave little room for a stable ranking. We consider these possible influences when interpreting SVHN, but do not attribute its measured ordering to them without a direct sensitivity experiment.

B.3 WITHIN-CLIENT CLASS ALIGNMENT

Let $\mathcal{C}_i$ contain classes with finite target and evaluation accuracy. An explicit within-client statistic is

$$\rho_i^{\mathrm{class}} = \mathrm{Corr}_{c\in\mathcal{C}_i}(d_{ic}, \Delta u_{ic}). \tag{23}$$

This correlation is defined only with at least two eligible classes and nonzero variance in both vectors. We distinguish undefined correlations from zero and interpret the reduction over eligible clients

and classes. The eligible counts and missing-class conventions are part of the evaluator specification needed for exact recomputation.

We take the arithmetic mean of per-client class utility alignments within each seed, then report the equal-weight mean and sample SD across three seeds. This preserves the within-client interpretation, rather than flattening all client–class entries into a single correlation.

This metric compares outcomes with the *same target used to design the reward*. It measures target realization, not independent validation that d_i is the correct normative or causal contribution measure. Independent contribution valuation would require a separate reference criterion.

Correlation also discards a level shift. If $\Delta u_i = a d_i - b\mathbf{1}$ with $a > 0$, nonconstant d_i, and sufficiently large b, then $\mathrm{Corr}(d_i, \Delta u_i) = 1$ although every component of Δu_i is negative. This analytical example is not a measured experiment. It shows why a class-alignment score must be accompanied by absolute gains, mean dispatch utility, and preferably the distribution of class losses. Our mean dispatch results quantify the average cost; they do not identify which individual clients or classes incurred it.

B.4 Individual-rationality statistic and aggregation over runs

We use the following operational statistic for individual-rationality comparisons:

$$\widehat{\mathrm{IR}}_{\mathrm{cross}} = \frac{1}{N} \sum_i \mathbf{1}\{u_i(\mathcal{T}) \geq c_i(\mathcal{V})\}, \tag{24}$$

where the two sides are evaluated on different splits. This comparison does not establish individual rationality on one common distribution, let alone a statement about utility net of participation cost. We obtain a value of one for SICAR in all nine settings. This means that each received-model test accuracy exceeds its standalone validation reference, not that each client improves over clean FedAvg. We have not evaluated a participation guarantee using common-distribution utility net of costs.

For n independent run-level observations x_s, the required descriptive summaries are

$$\bar{x} = \frac{1}{n} \sum_s x_s, \qquad s_x = \sqrt{\frac{1}{n-1} \sum_s (x_s - \bar{x})^2}. \tag{25}$$

The standard deviation is undefined for $n = 1$. For paired comparisons, first compute $d_s = x_s - y_s$ and then apply Equation (25). In general s_d cannot be recovered from s_x and s_y without covariance. Client observations, classes, and communication rounds do not become independent training runs. We report sample SDs over three seeds and compute paired differences within each seed before aggregation.

C Complete algorithm

We expand the construction and cancellation equations below. We fix each perturbation after reference collection and describe its inputs in Appendix E. Our accompanying CPU routines implement the algebraic construction.

C.1 Reward construction

For history $U \in \mathbb{R}^{M \times L}$, let $U = L_U \Sigma_U V_U^\top$. Retain columns indexed by $\mathcal{J}_U = \{j : \sigma_j \geq 10^{-3}\sigma_1\}$, capped at 64, and set $Q = (L_U)_{:,\mathcal{J}_U}$. The cap and threshold are experimental choices. For each class-conditioned probe, form the sample mean in Equation (4). A mean of batch means agrees only when the batch sizes are equal. Batching and precision are part of the implementation record needed for numerical reproduction.

The weighted projection can be computed without an $M \times M$ matrix:

$$B_R = R^\top W^{1/2}, \qquad H = B_R - Q(Q^\top B_R). \tag{26}$$

If $H = L_H \Sigma_H V_H^\top$, retain the leading $r \le \min(K-1, \operatorname{rank} H)$ nonzero directions and orthogonalize them. Rank deficiency reduces the available response directions; arbitrary vectors must not be inserted to reach $K-1$. We treat the numerical rank as an implementation parameter; reproducing a particular carrier requires its saved basis or the rank threshold used in that run.

For $A_G = G^\top W G + \beta^{-1} I_r$, a Cholesky factor $L_G L_G^\top = A_G$ gives the equivalent solution

$$L_G y(v) = G^\top W v, \qquad L_G^\top z(v) = y(v), \qquad \delta_i = a_i P z(d_i). \tag{27}$$

This factorization solves the stated objective and can be reused because G, W, β are fixed. Algebraically $z(a_i d_i) = a_i z(d_i)$; the stored tensor is authoritative if separately rounded computations differ.

Algorithm 2 Full reward construction in equations

Input: n_{ic}, c_i, $\theta_\star$, labeled V_c, clean history u_{it}, $m_{\mathrm{ref}} > 0$

$n_i \leftarrow \sum_c n_{ic}$; $w_i \leftarrow n_i / \sum_j n_j$; $q_c \leftarrow \sum_i w_i n_{ic}/n_i$.
Require $q_c > 0$ and $|V_c| > 0$ for every class.
$\widetilde{p}_{ic} \leftarrow (n_{ic} + \alpha q_c)/(n_i + \alpha)$; $s_i \leftarrow C_K \log(\widetilde{p}_i / q)$.
$d_i \leftarrow s_i / \max(\|s_i\|, \varepsilon)$; $\rho_i \leftarrow \operatorname{rank}_0(c_i)/(N-1)$; $f_i \leftarrow 1 + 0.25(1 - 2\rho_i)$.
$Q \leftarrow$ retained left singular vectors of $U = [u_{it}]$.
for $c = 1, \dots, K$ **do**
    $m_c \leftarrow |V_c|^{-1} \sum_{x \in V_c} [f_c(x; \theta_\star) - (K-1)^{-1} \sum_{k \neq c} f_k(x; \theta_\star)]$.
    $R_{c,:} \leftarrow (\nabla_{E^\top \theta} m_c(\theta_\star))^\top$.
end for
$W \leftarrow \operatorname{diag}((q + 10^{-8})^{-1})$; $H \leftarrow R^\top W^{1/2} - Q(Q^\top R^\top W^{1/2})$.
$P \leftarrow \operatorname{orth}(H_r)$; $G \leftarrow RP$; $A_G \leftarrow G^\top W G + \beta^{-1} I_r$.
for $i = 1, \dots, N$ **do**
    Solve $A_G z_i = G^\top W d_i$; $h_i \leftarrow \|P z_i\|$.
end for
$h_{\mathrm{med}} \leftarrow \operatorname{median}_i(f_i h_i)$; require $h_{\mathrm{med}} > 0$; $a_0 \leftarrow m_{\mathrm{ref}} / h_{\mathrm{med}}$.
$a_i \leftarrow s a_0 f_i$, $s = 0.5$; $t_i \leftarrow a_i d_i$; solve $A_G z_i = G^\top W t_i$; $\delta_i \leftarrow P z_i$.
Save complete δ_i in the classifier packing order; freeze q, R, P, δ_i.

C.2 Local training and cancellation

The packing order for `classifier.weight` and `classifier.bias` defines E. No other parameters are directly perturbed. All parameters are optimized according to Equation (12). After the client's local epoch, the trained model is uploaded and the server subtracts its saved classifier tensor:

$$x_{it,J_i} = \theta_t + E\delta_i - \sum_{j=0}^{J_i - 1} \eta_t v_{it,j+1}, \tag{28}$$

$$\widetilde{\phi}_{it} = x_{it,J_i} - E\delta_i = \theta_t - \sum_{j=0}^{J_i - 1} \eta_t v_{it,j+1}, \tag{29}$$

$$\theta_{t+1} = \theta_t - \sum_i w_i \sum_{j=0}^{J_i - 1} \eta_t v_{it,j+1}. \tag{30}$$

The expansion identifies the saved offset removed by cancellation and the local velocity sum retained for aggregation. It implements the definition in Section 3.3.

Under full participation, server-side and client-side subtraction agree because

$$\sum_i w_i (x_{it,J_i} - E\delta_i) = \sum_i w_i x_{it,J_i} - E \sum_i w_i \delta_i. \tag{31}$$

This requires the same weights and saved tensors. Subtracting a reconstructed or stale tensor need not match subtraction of the saved injection. Since $\delta_i \in \mathrm{range}(P)$, projecting the injection onto P is redundant in exact arithmetic; projecting the trained update is a different operation. Optimizer reset between rounds does not remove momentum propagation within a round.

C.3 Controls and clean-twin diagnostics

The experimental branches have distinct recursion rules:

$$\begin{aligned}\theta^A_{t+1} &= \sum_i w_i F_{it}(\theta^A_t),\\ \theta^B_{t+1} &= \sum_i w_i F_{it}(\theta^B_t + E\delta_i),\\ \theta^C_{t+1} &= \sum_i w_i [F_{it}(\theta^C_t + E\delta_i) - E\delta_i].\end{aligned} \tag{32}$$

A is the clean control, B retains the reward, and C applies direct cancellation. Their starting states coincide at the reference point but subsequently diverge. Endpoint C-minus-B is a trajectory contrast. It is not a measurement of a local residual at a common state.

A clean twin instead holds the current C state fixed:

$$\tau_{it} = F_{it}(\theta^C_t + E\delta_i) - F_{it}(\theta^C_t), \qquad e_{it} = \tau_{it} - E\delta_i. \tag{33}$$

Both evaluations must share minibatches, augmentation, dropout, initial optimizer state, and other randomness. We pair local batches in our clean-twin diagnostics. The clean A trajectory and the common-state twin have distinct roles; only the C recursion is required for deployment.

C.4 Degenerate inputs and implementation boundaries

A globally absent class makes the CLR and class probe undefined. If every useful class response is projected away, H may have zero rank; if enough client contrasts vanish, the median in Equation (10) may be zero. We exclude these undefined construction regimes rather than substituting an uncalibrated carrier or default perturbation. We use average ranks for standalone ties. The normalized-rank formula also excludes $N = 1$, outside our collaborative setting.

Our final implementation uses the class perturbation alone at global strength $s = 0.5$. We reuse the reference state, partitions, and saved perturbations within each paired comparison. Appendix E distinguishes our algebraic reference implementation from the artifacts needed to reproduce training.

D Derivations for the reward and cancellation

The derivations establish the properties used by the mechanism: relative class coordinates, a unique regularized response fit, calibrated reward scale, and a conditional cancellation criterion. Their conclusions concern these mathematical quantities under the stated assumptions.

D.1 Centered log-ratios and the class target

Let $C = I_K - K^{-1}\mathbf{1}\mathbf{1}^\top$ be the centering projector. For positive p, q, the contrast in Equation (2) is $s = C(\log p - \log q)$. It lies in the $(K-1)$-dimensional zero-sum subspace because $\mathbf{1}^\top C = 0$. For every positive scalar a, $C\log(ap) = C\log p$; a common multiplicative scale vanishes. The contrast therefore records relative ratios. For any two classes,

$$s_c - s_k = \log \frac{p_c/q_c}{p_k/q_k}. \tag{34}$$

The sign of s_c compares p_c/q_c to the geometric mean of all ratios, not directly to one. Describing every positive coordinate as simply "more than the global frequency" would be imprecise. The normalized target inherits this ordering when its denominator is positive.

The pseudocount construction is a convex combination,

$$\widetilde{p}_i = \frac{n_i}{n_i + \alpha} p_i + \frac{\alpha}{n_i + \alpha} q. \tag{35}$$

It shrinks smaller clients more strongly toward the federation distribution. This is a prior-like smoothing rule, not an empirical estimate of missing-class accuracy. If $q_c = 0$, smoothing does not create positive mass in class c; both the log-ratio and a class-conditioned probe may be undefined. The method as described therefore requires global class support.

Normalization removes most magnitude information from s_i. Two positively proportional contrasts with both norms at least ε have the same d_i, even if one client has much stronger label skew. The subsequent amplitude does depend on rank and response norm, but it does not restore a causal interpretation of class counts. An identical class profile with different image quality can yield the same target but different standalone rank. The class direction and its rank-dependent amplitude therefore encode different inputs to the same perturbation. We interpret the direction as relative class participation, not absolute sample contribution. A client can contribute many examples of a class while having a negative target coordinate for it.

For sensitivity, suppose two positive distributions p, p' have each coordinate at least $b > 0$. The scalar mean-value theorem gives $|\log p_c - \log p'_c| \le |p_c - p'_c|/b$. Since $\|C\|_2 = 1$,

$$\|C \log p - C \log p'\|_2 \le b^{-1} \|p - p'\|_2 . \tag{36}$$

For nonzero s, s' with norms at least $a > 0$, a triangle inequality yields

$$\left\| \frac{s}{\|s\|} - \frac{s'}{\|s'\|} \right\| \le \frac{2}{a} \|s - s'\| . \tag{37}$$

Thus rare classes and nearly zero contrasts can make normalized directions sensitive. Our stabilized map $s/\max(\|s\|, \varepsilon)$ is continuous at zero and equals zero there; its slope can nevertheless be as large as $1/\varepsilon$ near zero. Smoothing and stabilization ensure defined coordinates but do not establish robustness to count noise. We use actual partition counts in these experiments and have not measured this sensitivity under perturbed counts.

D.2 Existence, uniqueness, and shrinkage of the response fit

Set $\kappa = \beta^{-1} > 0$. The objective is

$$J(z) = (Gz - v)^\top W (Gz - v) + \kappa z^\top z. \tag{38}$$

Because W is positive definite, for every nonzero x,

$$x^\top (G^\top W G + \kappa I) x = \left\| W^{1/2} G x \right\|^2 + \kappa \|x\|^2 > 0. \tag{39}$$

The Hessian is twice this positive-definite matrix, so the unique global minimizer satisfies $(G^\top W G + \kappa I) z = G^\top W v$. This gives Equation (8) even if G is rank deficient. With an orthonormal carrier P, $\|Pz\| = \|z\|$, so penalizing coordinates also penalizes the classifier perturbation norm.

Let $B = W^{1/2} G = U_B \Sigma V_B^\top$ be a thin singular value decomposition. Then

$$z(v) = V_B \operatorname{diag}\left(\frac{\sigma_j}{\sigma_j^2 + \kappa} \right) U_B^\top W^{1/2} v, \tag{40}$$

$$W^{1/2} G z(v) = U_B \operatorname{diag}\left(\frac{\sigma_j^2}{\sigma_j^2 + \kappa} \right) U_B^\top W^{1/2} v. \tag{41}$$

Unreachable components of $W^{1/2} v$ are discarded. Reachable components are shrunk unequally, except in special spectral configurations. The fitted response is therefore not generally collinear with v. The spectral factors identify which class directions are retained most strongly by the response fit.

The scalar function $\sigma/(\sigma^2 + \kappa)$ reaches its maximum at $\sigma = \sqrt{\kappa}$, with value $1/(2\sqrt{\kappa})$. Hence

$$\|z(v) - z(v')\| \le \frac{1}{2\sqrt{\kappa}} \left\| W^{1/2}(v - v') \right\|, \qquad \|z(v)\| \le \frac{\sqrt{\beta}}{2} \left\| W^{1/2} v \right\| . \tag{42}$$

This is conditional on fixed G, W. Changes to probe data, the history basis, or numerical rank change the operator itself. Near a singular-value cutoff, the selected subspace can change discontinuously; Equation (42) alone does not bound that effect. Likewise, a small q_c produces a large entry in W, increasing weighted sensitivity rather than eliminating class imbalance.

D.3 Calibration and historical orthogonality

Linearity gives $\delta_i = a_i P z(d_i)$, so $\|\delta_i\| = 0.5 a_0 f_i h_i$ because $a_0, f_i \geq 0$. If the median denominator in Equation (10) is positive, homogeneity of the median under positive rescaling yields

$$\operatorname{median}_i \|\delta_i\| = 0.5 a_0 \operatorname{median}_i(f_i h_i) = 0.5 m_{\mathrm{ref}}. \tag{43}$$

This fixes a central norm scale; it does not fix every client's norm or realized margin. A high-rank client has a smaller f_i but may still have a larger norm because h_i depends on its target. We therefore call f_i a rank-dependent factor: it neither guarantees an ordering of final perturbation norms nor ensures that a larger factor improves received utility.

For an exact orthonormal Q, $Q^\top(I - QQ^\top) = 0$, so every column of H lies in the orthogonal complement of the retained history. Any basis P selected wholly from that range satisfies $Q^\top P = 0$. For a future local training Jacobian J, however, JEP need not stay in that complement or even in the classifier coordinates. For example, in two dimensions, take a carrier along the first coordinate and a linear map that sends the first coordinate to a vector with a nonzero second coordinate. Historical orthogonality is unchanged, yet the response leaves the original subspace. This illustrates the distinction between a construction constraint and invariance under local training.

The final injection satisfies $\delta_i \in \operatorname{range}(P)$ and is orthogonal to Q in exact arithmetic. Its local-training influence need not satisfy either property. Saved-vector subtraction removes δ_i itself; it does not project away the trained update or recover influence outside the original carrier.

D.4 Cancellation criterion and a scalar operating boundary

For Equation (18), expand

$$\|\tau - a\|^2 - \|\tau\|^2 = \|a\|^2 - 2\langle \tau, a \rangle. \tag{44}$$

Taking the sign proves both directions of the equivalence, including the equality boundary. Applying the same identity to $S\tau, Sa$ proves the response version. The response norm is only a seminorm on full parameter space if S has a nontrivial nullspace; improvement under that seminorm need not imply a smaller full-model discrepancy.

Consider $\ell(x) = \frac{1}{2} h x^2$ and one gradient step with $h > 0$, $F(x) = (1 - \eta h)x$. An injection $a \neq 0$ produces $\tau = (1 - \eta h)a$ and residual $e = -\eta h a$. For $0 \leq \eta h \leq 1$, subtraction reduces discrepancy exactly when

$$\eta h < 1 - \eta h \quad \Longleftrightarrow \quad \eta h < \tfrac{1}{2}. \tag{45}$$

At $\eta h = 0.6$, ordinary gradient descent remains contractive, yet subtracting the original injection increases its residual magnitude from $0.4|a|$ to $0.6|a|$. At $\eta h = 1$, training itself erases the offset and subsequent subtraction reintroduces a nonzero error. These analytical examples delimit the useful-cancellation condition independently of local convergence.

D.5 Multiple SGD steps and within-round momentum

For deterministic paired minibatch losses and plain SGD, let x_j be the clean trajectory and x'_j the perturbed trajectory, $\xi_j = x'_j - x_j$, $\xi_0 = a$. Then

$$\xi_{j+1} = \xi_j - \eta_j \left[\nabla \ell_j(x'_j) - \nabla \ell_j(x_j)\right], \qquad e = \xi_J - a = -\sum_{j=0}^{J-1} \eta_j \left[\nabla \ell_j(x'_j) - \nabla \ell_j(x_j)\right]. \tag{46}$$

For $\eta_j \geq 0$, if each minibatch gradient is L_j-Lipschitz on the relevant segment,

$$\|e\| \leq \left[\prod_{j=0}^{J-1} (1 + \eta_j L_j) - 1\right] \|a\|. \tag{47}$$

To prove this, first use $\|\xi_{j+1}\| \le (1+\eta_j L_j)\,\|\xi_j\|$ and iterate. Substitute the resulting product bound in the sum in Equation (46); the products telescope. This worst-case upper bound can be loose or large. It does not guarantee improvement relative to leaving the perturbation in place.

For momentum, let $b_j = v'_j - v_j$, with $b_0 = 0$. Under the convention of Equation (12),

$$b_{j+1} = \mu b_j + \nabla \ell_j(x'_j) - \nabla \ell_j(x_j), \tag{48}$$

$$\xi_{j+1} = \xi_j - \eta_j b_{j+1}. \tag{49}$$

Subtracting a after J steps gives $\xi_J - a$, which depends on every gradient difference and momentum propagation. Resetting momentum between communication rounds sets the next round's initial b_0 to zero, but does not cancel its effect within the preceding round. We therefore analyze momentum propagation separately from the single-step case.

Full-model Hessian cross terms explain why a classifier injection can affect a backbone. Partition parameters into head h and backbone b. A first-order gradient difference includes a backbone component $H_{bh}a_h$. Subtracting only a_h cannot remove that component. Restricting diagnostics to the head can therefore understate total influence even if the head residual is small.

D.6 Aggregation and the distinction between one step and a trajectory

At a common server state, let $a_i^{\text{full}} = E\delta_i$ and $e_i = F_i(\theta + a_i^{\text{full}}) - a_i^{\text{full}} - F_i(\theta)$. The corrected aggregated update differs from the common-start clean update by

$$e_{\text{agg}} = \sum_i w_i e_i, \qquad \|e_{\text{agg}}\| \le \sum_i w_i \,\|e_i\|\,. \tag{50}$$

Signed cancellation between clients can make the aggregate smaller than individual residuals, but is not ensured by sample weights or centered class targets. In particular, $\sum_i w_i d_i$ need not be zero: the nonlinear log transform, smoothing, and per-client normalization prevent that inference from $q = \sum_i w_i p_i$.

Let $A_t(\theta) = \sum_i w_i F_{it}(\theta)$ be the clean round map and $C_t(\theta) = A_t(\theta) + e_t(\theta)$ the corrected map under paired local randomness. If A_t is L_t-Lipschitz and $\left\|e_t(\theta_t^C)\right\| \le \epsilon_t$, then

$$\left\|\theta_{t+1}^C - \theta_{t+1}^A\right\| \le L_t \left\|\theta_t^C - \theta_t^A\right\| + \epsilon_t. \tag{51}$$

Unrolling yields a weighted sum of past ϵ_t values, multiplied by subsequent Lipschitz constants. These constants and residual bounds are not established for the experiments. A small local residual, even if measured, does not by itself guarantee that the full long-run trajectory remains close.

D.7 Relating margin response and parameter distance to predictions

Margins in Equation (4) compare one logit with an average of the others. Classification instead compares that logit with their maximum. A mean-logit margin can increase while a competing class increases enough to change the argmax. Further, the probe averages can improve while some individual examples worsen. Exact realization of $R\delta = v$ at a warm-up point would not establish per-example decisions at later models.

For a separate sufficient condition, suppose every logit is L_f-Lipschitz in parameters on a neighborhood. If an example's original winning-logit gap exceeds $2L_f\,\|e\|$, its predicted class is unchanged under parameter perturbation e. This follows by bounding the winner's decrease and the competitor's increase separately by $L_f\,\|e\|$. The condition preserves the old prediction, which may itself be incorrect. We use this condition to distinguish prediction stability from accuracy improvement; we do not estimate L_f or the gap distribution in our experiments.

Target alignment evaluates realization of the allocation design. Strategic participation and distributional fairness require additional objectives and assumptions, as discussed in Appendix G.

E EXPERIMENTAL DETAILS

E.1 DATASETS, MODELS, AND PARTITIONS

CIFAR-100, Tiny-ImageNet, and SVHN cover object and digit classification (Krizhevsky, 2009; Le & Yang, 2015; Netzer et al., 2011). We use ten fully participating clients, label-Dirichlet concentration 0.5, and seeds 0, 1, and 2. The architectures are ResNet18, VGG16, and DenseNet121, each adapted with two-group GroupNorm (He et al., 2016; Simonyan & Zisserman, 2015; Huang et al., 2017; Wu & He, 2018).

We evaluate all nine dataset–architecture settings, with complete A/B/C endpoints and nine adapted baselines per setting. This gives 270 formal method–setting–seed records. Our source data include the per-seed endpoint metrics and record hashes. Tiny-ImageNet requires a labeled evaluation split because its released test images lack public labels. Reproducing the data pipeline additionally requires our sample manifests, transformations, partition-generation procedure, and SVHN extra-data policy; we do not infer these identities from aggregate scores.

E.2 REFERENCE COLLECTION AND CONSTRUCTION PARAMETERS

We use clean sample-weighted FedAvg to obtain the reference model $\theta_\star$ and retained clean-update history $\mathcal{H}$ (McMahan et al., 2017). For each recorded client update, $\phi_{it} = F_{it}(\theta_t)$ and $u_{it} = E^\top(\phi_{it} - \theta_t)$; the retained classifier updates form U.

The history SVD retains at most 64 directions with relative singular value at least 10^{-3}. The requested response rank is $K - 1$, truncated to numerical rank. The pseudocount mass is $\alpha = 5$, assigning class c mass $5q_c$; response weights use $q_c + 10^{-8}$. The ridge coefficient is $\beta = 0.000225$, with penalty $1/\beta$. The rank factor uses $\lambda = 0.25$, strength is $s = 0.5$, and the CLR stabilizer is $\varepsilon = 10^{-8}$. The construction is fixed after reference collection.

We calibrate amplitude using a fixed, seed-specific norm reference m_{ref} retained from our earlier CLR calibration. For the revised SVHN protocol, we regenerate this reference at its reference state. Equation (10) then sets the median final perturbation norm to sm_{ref}. The norm reference is an external calibration input to the present method, not a quantity determined by class proportions alone. Reproducing our absolute perturbation strengths requires these saved calibration values. On a new task, the scale must be specified and calibrated using training or server-validation data before final evaluation; our experiments do not establish a universal choice.

E.3 LOCAL TRAINING AND ENDPOINT EVALUATION

We train with cross-entropy and SGD, using one local epoch, momentum 0.9, and zero weight decay. Each client initializes a fresh optimizer and retains momentum between local minibatches. We collect the reference during the first 12 communication rounds, construct the carrier once, and begin perturbed training afterward. Total budgets include this reference phase: CIFAR-100 / ResNet18 and VGG16, all three Tiny-ImageNet models, and SVHN / ResNet18 use 200 rounds; CIFAR-100 / DenseNet121 and SVHN / DenseNet121 use 100; SVHN / VGG16 uses 120. The corresponding perturbed phases contain 188, 88, or 108 rounds.

We use learning rate 0.01 in every setting except SVHN / VGG16, which uses 0.0033. CIFAR-100 uses batch sizes 128, 128, and 64 for ResNet18, VGG16, and DenseNet121, respectively. Tiny-ImageNet uses 64, 64, and 48 in the same architecture order; SVHN uses 128, 128, and 64. These settings are fixed across all three seeds within each experiment. Our validation-only strength scans use 22 total rounds, including the same 12-round reference phase; their endpoint and evaluation split therefore differ from the formal test experiments.

We evaluate each formal run at its fixed endpoint. We subsequently reevaluate saved CFFL and CGSV models to recover client vectors, preserving their original global accuracy and mean utility. Those additional evaluations are further test accesses, not new training runs. We also used earlier SVHN diagnostics during protocol development. We consequently treat SVHN as development-stage evidence rather than an untouched confirmation set.

E.4 Data access and cross-method comparisons

We construct SICAR from training class counts, a labeled server-validation probe, and standalone validation scores. Counts determine p_i, q, w_i, d_i; probe labels determine R; scores determine the rank-dependent amplitude factor. We use validation inputs for the final construction and calibration. Verification of split separation requires the sample and score-cache identities in addition to these declared roles.

Some of our adapted baselines retain historical final-test contribution scores. Recomputing their final alignment against a common validation-based reference does not undo those training-time accesses. This is a limitation of our adaptations, not of the published methods; for example, native CGSV estimates contribution from updates without an auxiliary validation set (Xu et al., 2021).

We use unequal cross-method budgets in several settings. On CIFAR-100 / DenseNet121, we use 100 rounds for SICAR and 200 for the baselines. On SVHN / ResNet18, we use 200 for SICAR and 100 for the baselines. Tiny-ImageNet / ResNet18 uses 200 for SICAR and most baselines, but 100 for Quota and FedAKD. Tiny-ImageNet / DenseNet121 uses 200 for SICAR, FedAvg, and CFFL, mixed 100/200 budgets for CGSV and ShapFed, and 100 for the remaining baselines. The other settings use equal nominal round counts, which still do not equate computational cost: Aequa trains sampled widths and FedAKD adds distillation. We use these cross-method results as descriptive context and the within-seed A/B/C contrasts for mechanism conclusions.

Our independently trained FedAvg baseline is a separate run family from paired A. On CIFAR-100 / ResNet18 they have the same nominal 200-round budget but different endpoints (54.43% and 56.61%). We have not isolated the initialization, reference-reuse, or implementation differences responsible for this gap. We therefore compare SICAR with A when attributing a change to its intervention. Our Quota adaptation is a resource surrogate returning a shared model, rather than the payment mechanism of reciprocally fair FL (Murhekar et al., 2025).

E.5 Evidence scope and reproduction materials

Our source package preserves the numerical inputs for all displayed results, their table and figure generators, and algebraic checks. It supports reproduction of the aggregate analysis. Reproducing federated training additionally requires our training implementation, split manifests, calibration tensors, score-cache lineage, and evaluation conventions. We provide endpoint seed records separately from aggregate mean/SD values and do not reconstruct individual outcomes from those summaries.

F Extended mechanism evidence and comparison context

We report 81 baseline–setting aggregates and paired SICAR evidence for nine settings, each with three seeds. We give complete A/B/C endpoints, paired utility contrasts, class alignment, and cross-method comparisons. We retain every completed setting, including weaker SVHN alignment and the variable CIFAR-100 / DenseNet121 client ranking.

F.1 Paired endpoints, recovery, and allocation

Tables 2, 3, and 4 cover all nine settings. A, B, and C follow Equation (32). Shared accuracy evaluates each server endpoint; dispatch utility evaluates the reward-bearing model before additional local training. The global and dispatch contrasts therefore measure different parts of the allocation procedure.

CIFAR-100 and Tiny-ImageNet provide the strongest shared-recovery evidence. Table 2 places the three conditions side by side, and Table 3 gives the paired differences with their directly reported SDs. Table 4 shows that the six C class-alignment means on these datasets remain between 0.765 and 0.855. We compute paired SDs directly; small discrepancies from subtracting displayed endpoints reflect rounding.

Table 2: Shared-model endpoints across all nine settings. Accuracy is a percentage; entries are mean ± sample SD over three seeds.

Model	Clean A	Retained B	SICAR C
CIFAR-100			
ResNet18	56.61 ± 0.56	55.74 ± 0.66	58.14 ± 0.72
VGG16	55.90 ± 0.59	54.09 ± 0.55	56.03 ± 0.81
DenseNet121	48.09 ± 0.41	47.80 ± 0.36	48.98 ± 0.66
Tiny-ImageNet			
ResNet18	48.40 ± 1.02	47.69 ± 0.75	48.89 ± 1.01
VGG16	46.16 ± 0.33	45.96 ± 0.89	46.57 ± 0.53
DenseNet121	49.85 ± 0.45	49.53 ± 0.36	50.45 ± 0.16
SVHN			
ResNet18	95.83 ± 0.28	95.92 ± 0.35	95.95 ± 0.25
VGG16	94.87 ± 0.16	94.87 ± 0.20	94.84 ± 0.17
DenseNet121	95.57 ± 0.14	95.60 ± 0.15	95.61 ± 0.20

Table 3: Paired changes across all nine settings, in percentage points. We compute each difference within seed before taking the mean and sample SD over three seeds.

Model	Global $C - B$	Global $C - A$	Client $C - B$	Client $C - A$
CIFAR-100				
ResNet18	$+2.400 \pm 0.416$	$+1.529 \pm 0.346$	-1.938 ± 0.923	-3.861 ± 0.894
VGG16	$+1.933 \pm 0.305$	$+0.121 \pm 0.838$	-1.098 ± 0.648	-4.482 ± 0.446
DenseNet121	$+1.183 \pm 0.757$	$+0.896 \pm 0.439$	$+0.476 \pm 0.202$	-3.252 ± 0.677
Tiny-ImageNet				
ResNet18	$+1.196 \pm 0.313$	$+0.488 \pm 0.045$	-0.850 ± 0.058	-3.518 ± 0.185
VGG16	$+0.613 \pm 0.839$	$+0.413 \pm 0.806$	-1.092 ± 0.728	-2.747 ± 0.807
DenseNet121	$+0.917 \pm 0.325$	$+0.600 \pm 0.291$	-0.429 ± 0.252	-3.153 ± 0.279
SVHN				
ResNet18	$+0.022 \pm 0.262$	$+0.114 \pm 0.201$	$+0.021 \pm 0.267$	$+0.110 \pm 0.203$
VGG16	-0.032 ± 0.074	-0.035 ± 0.024	-0.021 ± 0.046	-0.090 ± 0.019
DenseNet121	$+0.011 \pm 0.047$	$+0.045 \pm 0.063$	$+0.019 \pm 0.072$	$+0.045 \pm 0.071$

Table 4: Class allocation with and without cancellation. Fine Pearson measures within-client target realization. We average clients within each seed, then give the mean ± sample SD over three seeds.

Model	Fine B	Fine C
CIFAR-100		
ResNet18	0.447 ± 0.055	0.855 ± 0.009
VGG16	0.516 ± 0.044	0.844 ± 0.015
DenseNet121	0.835 ± 0.026	0.830 ± 0.029
Tiny-ImageNet		
ResNet18	0.665 ± 0.029	0.833 ± 0.001
VGG16	0.607 ± 0.014	0.765 ± 0.007
DenseNet121	0.766 ± 0.015	0.836 ± 0.015
SVHN		
ResNet18	0.010 ± 0.038	-0.040 ± 0.028
VGG16	0.490 ± 0.128	0.602 ± 0.112
DenseNet121	0.121 ± 0.071	0.067 ± 0.104

CIFAR-100: shared recovery across architectures. ResNet18 yields C-minus-B $+2.40 \pm 0.42$ and C-minus-A $+1.53 \pm 0.35$ percentage points. VGG16 gives $+1.93 \pm 0.30$ over B, while DenseNet121 gives $+1.18 \pm 0.76$ over B and $+0.90 \pm 0.44$ over A. The shared-model pattern is accompanied by a dispatch cost: client C-minus-A is -3.86 ± 0.89, -4.48 ± 0.45, and -3.25 ± 0.68 for ResNet18, VGG16, and DenseNet121. The VGG16 net gain over A is small relative to its dispersion, distinguishing repair of reward retention from a uniform gain over clean training.

Tiny-ImageNet: strong class allocation with architecture dependence. All three architectures combine positive mean shared recovery with strong class alignment. ResNet18 gives C-minus-B $+1.20 \pm 0.31$ and C-minus-A $+0.49 \pm 0.05$ points, alongside Fine C 0.833 ± 0.001. DenseNet121 extends this pattern with $+0.92 \pm 0.33$ over B, $+0.60 \pm 0.29$ over A, and Fine C 0.836 ± 0.015. VGG16 gives Fine C 0.765 ± 0.007 and positive mean contrasts with wider dispersion: $+0.61 \pm 0.84$ over B and $+0.41 \pm 0.81$ over A. We obtain seed-2 C-minus-B of -0.35 points, so the positive mean does not hold in every seed. Dispatch C-minus-A is -3.52 ± 0.18, -2.75 ± 0.81, and -3.15 ± 0.28 for ResNet18, VGG16, and DenseNet121, respectively.

Table 5: Client-level changes relative to clean A, in percentage points. We take the minimum and maximum over ten clients within each seed, then summarize these extrema across three seeds. The final column counts positive client–seed differences; these observations are not independent training runs.

Setting	Minimum $C - A$	Maximum $C - A$	$C > A$
CIFAR-100 / ResNet18	-5.67 ± 1.00	-1.89 ± 0.76	0/30
CIFAR-100 / VGG16	-5.95 ± 0.46	-3.11 ± 0.19	0/30
CIFAR-100 / DenseNet121	-4.60 ± 1.08	-1.76 ± 0.39	0/30
Tiny-ImageNet / ResNet18	-4.99 ± 0.16	-1.98 ± 0.37	0/30
Tiny-ImageNet / VGG16	-3.98 ± 1.14	-1.57 ± 0.89	0/30
Tiny-ImageNet / DenseNet121	-4.53 ± 0.13	-1.81 ± 0.44	0/30
SVHN / ResNet18	$+0.10 \pm 0.20$	$+0.12 \pm 0.21$	16/30
SVHN / VGG16	-0.21 ± 0.06	$+0.04 \pm 0.04$	2/30
SVHN / DenseNet121	$+0.03 \pm 0.07$	$+0.07 \pm 0.09$	20/30

SVHN: small endpoint changes. The three architectures remain close to their paired controls in shared accuracy at the chosen strength. C-minus-B means are $+0.022$, -0.032, and $+0.011$ points for ResNet18, VGG16, and DenseNet121. This descriptive endpoint stability accompanies a weaker allocation signal, detailed in Table 24; it is not a formal equivalence result.

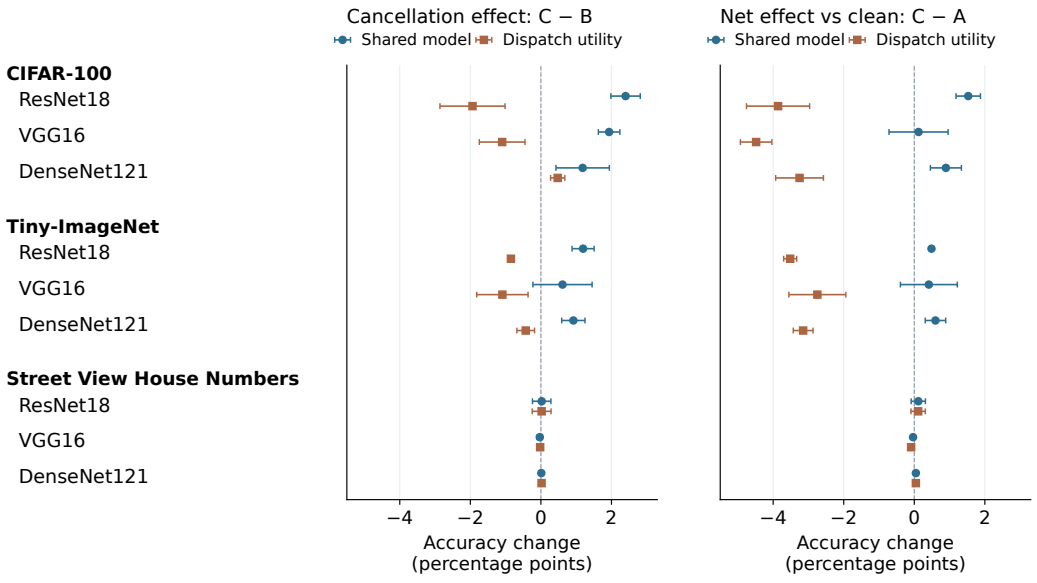

Figure 4: Paired utility contrasts across all nine settings. Blue denotes the shared model and orange the mean dispatched utility. Points are means and bars sample SDs across three seeds. The left panel isolates cancellation relative to B; the right compares the complete procedure with clean A. All architectures use GroupNorm.

Figure 4 locates the shared-recovery and dispatch-cost trade-off in the same comparison. The positive shared contrasts on CIFAR-100 and Tiny-ImageNet coexist with negative mean dispatch contrasts against clean A. SVHN clusters closer to zero on both axes. These patterns support analyzing shared training and received allocation separately, rather than treating one utility as a proxy for the other.

We report sample SDs rather than claiming statistical significance. Student-t intervals based on three seeds would require independent, approximately normal differences and would not adjust for development reuse or multiple comparisons.

F.2 Reading the cross-method inventory

We include all nine settings and all adapted methods in the following tables, together with separate client-alignment tables for CIFAR-100 and Tiny-ImageNet. We use GroupNorm throughout. Method names refer to our adaptations, interpreted with the data-access and budget qualifications in Appendix E.

F.3 CIFAR-100 / ResNet18

Table 6: Endpoint performance for CIFAR-100 / ResNet18. Accuracy and macro-F1 are percentages; entries are mean ± sample SD over three seeds.

Method	Global (%)	Macro-F1 (%)	Client (%)
FedAvg	54.43 ± 0.13	53.62 ± 0.52	54.43 ± 0.13
CFFL	30.61 ± 3.07	27.07 ± 3.41	27.25 ± 1.41
CGSV	48.60 ± 9.16	47.23 ± 10.49	50.94 ± 5.00
IAFL	41.46 ± 3.41	39.96 ± 3.93	36.16 ± 5.29
ShapFed	36.08 ± 1.63	33.64 ± 2.65	33.00 ± 0.81
Quota surrogate	55.45 ± 0.60	55.03 ± 0.64	55.45 ± 0.60
Aequa	43.83 ± 1.15	42.74 ± 1.27	43.32 ± 1.16
FedAKD	27.59 ± 0.42	25.30 ± 0.31	24.83 ± 0.33
FedRAC	45.02 ± 1.22	42.93 ± 1.03	39.51 ± 2.76
SICAR	58.14 ± 0.72	57.54 ± 0.79	52.75 ± 1.07

Table 7: Across-client alignment for CIFAR-100 / ResNet18. Values are mean ± sample SD over three seeds. IR(val-ref) compares dispatch utility with the standalone validation reference; a dash denotes an undefined correlation.

Method	Pearson	Spearman	Pairwise	IR(val-ref)
FedAvg	-	-	0.500 ± 0.000	1.000 ± 0.000
CFFL	0.851 ± 0.113	0.801 ± 0.102	0.828 ± 0.072	1.000 ± 0.000
CGSV	0.537 ± 0.337	0.471 ± 0.097	0.664 ± 0.071	1.000 ± 0.000
IAFL	0.632 ± 0.071	0.606 ± 0.301	0.682 ± 0.100	1.000 ± 0.000
ShapFed	0.547 ± 0.170	0.582 ± 0.303	0.731 ± 0.115	1.000 ± 0.000
Quota surrogate	-	-	0.500 ± 0.000	1.000 ± 0.000
Aequa	0.783 ± 0.289	0.665 ± 0.387	0.762 ± 0.179	1.000 ± 0.000
FedAKD	0.758 ± 0.095	0.746 ± 0.074	0.791 ± 0.012	1.000 ± 0.000
FedRAC	0.695 ± 0.257	0.674 ± 0.372	0.788 ± 0.163	0.900 ± 0.000
SICAR	0.731 ± 0.183	0.841 ± 0.079	0.858 ± 0.041	1.000 ± 0.000

We obtain shared accuracy $58.14 \pm 0.72\%$ and dispatch utility $52.75 \pm 1.07\%$. The separate FedAvg run has utility $54.43 \pm 0.13\%$. We use clean A, rather than this independently trained FedAvg model, for the paired mechanism comparison. CFFL and Aequa have higher Pearson means but lower mean dispatch utility; the metrics describe different aspects of allocation.

Table 8: Endpoint performance for CIFAR-100 / VGG16. Accuracy and macro-F1 are percentages; entries are mean ± sample SD over three seeds.

Method	Global (%)	Macro-F1 (%)	Client (%)
FedAvg	55.39 ± 0.65	55.04 ± 0.19	55.39 ± 0.65
CFFL	35.81 ± 1.35	33.35 ± 2.08	29.86 ± 0.10
CGSV	47.26 ± 10.18	46.10 ± 11.70	50.35 ± 4.69
IAFL	28.81 ± 1.89	25.30 ± 2.21	24.75 ± 1.40
ShapFed	40.85 ± 0.36	39.96 ± 0.45	32.95 ± 0.22
Quota surrogate	51.15 ± 1.98	50.90 ± 2.09	51.15 ± 1.98
Aequa	46.30 ± 1.61	45.70 ± 1.71	34.10 ± 0.76
FedAKD	42.03 ± 0.36	41.18 ± 0.42	28.64 ± 0.27
FedRAC	34.83 ± 0.73	32.41 ± 1.42	31.34 ± 1.98
SICAR	56.03 ± 0.81	55.70 ± 0.74	51.42 ± 0.26

Table 9: Across-client alignment for CIFAR-100 / VGG16. Values are mean ± sample SD over three seeds. IR(val-ref) compares dispatch utility with the standalone validation reference; a dash denotes an undefined correlation.

Method	Pearson	Spearman	Pairwise	IR(val-ref)
FedAvg	-	-	0.500 ± 0.000	1.000 ± 0.000
CFFL	0.354 ± 0.190	0.205 ± 0.369	0.580 ± 0.142	1.000 ± 0.000
CGSV	0.108 ± 0.292	0.185 ± 0.571	0.566 ± 0.234	1.000 ± 0.000
IAFL	0.849 ± 0.081	0.761 ± 0.231	0.847 ± 0.115	1.000 ± 0.000
ShapFed	0.397 ± 0.463	0.336 ± 0.549	0.651 ± 0.214	1.000 ± 0.000
Quota surrogate	-	-	0.500 ± 0.000	1.000 ± 0.000
Aequa	0.972 ± 0.021	0.965 ± 0.016	0.938 ± 0.013	1.000 ± 0.000
FedAKD	0.459 ± 0.234	0.335 ± 0.405	0.635 ± 0.162	1.000 ± 0.000
FedRAC	0.375 ± 0.091	0.696 ± 0.193	0.813 ± 0.098	1.000 ± 0.000
SICAR	0.826 ± 0.100	0.865 ± 0.064	0.888 ± 0.055	1.000 ± 0.000

F.4 CIFAR-100 / VGG16

We obtain shared accuracy $56.03 \pm 0.81\%$ and dispatch utility $51.42 \pm 0.26\%$. Aequa has higher across-client correlation, while we obtain higher absolute client utility. This setting illustrates why correlation and utility must be interpreted jointly.

F.5 CIFAR-100 / DENSENET121

Table 10: Endpoint performance for CIFAR-100 / DenseNet121. Accuracy and macro-F1 are percentages; entries are mean ± sample SD over three seeds.

Method	Global (%)	Macro-F1 (%)	Client (%)
FedAvg	59.45 ± 0.17	58.91 ± 0.38	59.45 ± 0.17
CFFL	34.65 ± 2.21	31.66 ± 2.65	29.52 ± 0.49
CGSV	53.15 ± 0.41	52.72 ± 0.56	53.10 ± 0.23
IAFL	41.36 ± 1.79	39.76 ± 1.86	35.76 ± 2.79
ShapFed	35.50 ± 0.44	33.65 ± 1.25	31.53 ± 0.46
Quota surrogate	53.10 ± 0.45	52.71 ± 0.59	53.10 ± 0.45
Aequa	43.75 ± 0.92	42.61 ± 1.03	42.35 ± 0.30
FedAKD	37.95 ± 0.79	36.35 ± 0.92	28.69 ± 0.50
FedRAC	47.25 ± 1.05	45.58 ± 0.69	41.59 ± 2.05
SICAR	48.98 ± 0.66	47.80 ± 0.82	44.84 ± 0.28

Table 11: Across-client alignment for CIFAR-100 / DenseNet121. Values are mean ± sample SD over three seeds. IR(val-ref) compares dispatch utility with the standalone validation reference; a dash denotes an undefined correlation.

Method	Pearson	Spearman	Pairwise	IR(val-ref)
FedAvg	-	-	0.500 ± 0.000	1.000 ± 0.000
CFFL	0.757 ± 0.010	0.798 ± 0.035	0.815 ± 0.013	1.000 ± 0.000
CGSV	0.378 ± 0.545	0.584 ± 0.375	0.726 ± 0.148	1.000 ± 0.000
IAFL	0.758 ± 0.086	0.732 ± 0.178	0.752 ± 0.093	1.000 ± 0.000
ShapFed	0.670 ± 0.090	0.611 ± 0.107	0.722 ± 0.033	1.000 ± 0.000
Quota surrogate	-	-	0.500 ± 0.000	1.000 ± 0.000
Aequa	0.915 ± 0.033	0.938 ± 0.018	0.893 ± 0.026	1.000 ± 0.000
FedAKD	0.766 ± 0.147	0.847 ± 0.081	0.830 ± 0.071	1.000 ± 0.000
FedRAC	0.561 ± 0.183	0.775 ± 0.140	0.815 ± 0.076	0.933 ± 0.058
SICAR	0.463 ± 0.491	0.505 ± 0.429	0.715 ± 0.167	1.000 ± 0.000

Our SICAR and paired controls use a shorter training budget than the independent baselines (Appendix E). We therefore interpret this inventory as descriptive context. Within the paired experiment, global $C - A$ is $+0.896 \pm 0.439$ pp and $C - B$ is $+1.183 \pm 0.757$ pp. Across-client Spearman is variable (0.505 ± 0.429), whereas Fine Pearson remains 0.830 ± 0.029.

F.6 TINY-IMAGENET / RESNET18

Table 12: Endpoint performance for Tiny-ImageNet / ResNet18. Accuracy and macro-F1 are percentages; entries are mean ± sample SD over three seeds.

Method	Global (%)	Macro-F1 (%)	Client (%)
FedAvg	48.25 ± 0.26	47.56 ± 0.18	48.25 ± 0.26
CFFL	29.29 ± 0.90	26.38 ± 1.12	24.24 ± 0.32
CGSV	42.24 ± 2.30	41.76 ± 2.02	43.52 ± 1.27
IAFL	40.12 ± 1.00	38.59 ± 1.06	35.20 ± 2.31
ShapFed	34.70 ± 0.71	33.82 ± 0.74	26.90 ± 0.73
Quota surrogate	40.97 ± 0.25	39.60 ± 0.19	40.97 ± 0.25
Aequa	42.34 ± 0.36	41.37 ± 0.37	42.23 ± 0.60
FedAKD	31.13 ± 0.74	28.69 ± 0.83	24.23 ± 0.48
FedRAC	48.88 ± 0.31	48.16 ± 0.46	45.76 ± 1.20
SICAR	48.89 ± 1.01	48.10 ± 1.00	44.88 ± 0.92

We obtain shared accuracy $48.89 \pm 1.01\%$, close to the FedRAC mean of $48.88 \pm 0.31\%$. We obtain higher mean Spearman (0.859 versus 0.781), while FedRAC has higher mean dispatch utility (45.76% versus 44.88%). These results show an allocation–utility trade-off rather than uniform dominance.

Table 13: Across-client alignment for Tiny-ImageNet / ResNet18. Values are mean ± sample SD over three seeds. IR(val-ref) compares dispatch utility with the standalone validation reference; a dash denotes an undefined correlation.

Method	Pearson	Spearman	Pairwise	IR(val-ref)
FedAvg	-	-	0.500 ± 0.000	1.000 ± 0.000
CFFL	0.764 ± 0.037	0.790 ± 0.062	0.822 ± 0.039	1.000 ± 0.000
CGSV	0.114 ± 0.839	-0.069 ± 0.757	0.456 ± 0.290	1.000 ± 0.000
IAFL	0.739 ± 0.191	0.699 ± 0.167	0.715 ± 0.093	1.000 ± 0.000
ShapFed	0.527 ± 0.363	0.459 ± 0.407	0.682 ± 0.185	1.000 ± 0.000
Quota surrogate	-	-	0.500 ± 0.000	1.000 ± 0.000
Aequa	0.030 ± 0.836	0.037 ± 0.829	0.507 ± 0.308	1.000 ± 0.000
FedAKD	0.733 ± 0.078	0.731 ± 0.145	0.785 ± 0.063	1.000 ± 0.000
FedRAC	0.761 ± 0.122	0.781 ± 0.035	0.819 ± 0.017	1.000 ± 0.000
SICAR	0.812 ± 0.121	0.859 ± 0.081	0.874 ± 0.056	1.000 ± 0.000

F.7 TINY-IMAGENET / VGG16

Table 14: Endpoint performance for Tiny-ImageNet / VGG16. Accuracy and macro-F1 are percentages; entries are mean ± sample SD over three seeds.

Method	Global (%)	Macro-F1 (%)	Client (%)
FedAvg	46.92 ± 0.35	46.22 ± 0.33	46.92 ± 0.35
CFFL	28.12 ± 0.19	25.37 ± 0.59	24.10 ± 0.20
CGSV	42.77 ± 1.68	42.38 ± 1.60	43.27 ± 0.94
IAFL	38.80 ± 1.90	37.30 ± 2.34	33.30 ± 3.27
ShapFed	31.81 ± 0.80	30.62 ± 0.84	23.90 ± 0.88
Quota surrogate	44.48 ± 0.13	44.05 ± 0.13	44.48 ± 0.13
Aequa	38.03 ± 0.31	36.81 ± 0.27	37.03 ± 0.15
FedAKD	40.08 ± 0.51	38.78 ± 0.56	27.00 ± 0.04
FedRAC	42.85 ± 0.68	41.70 ± 0.64	37.29 ± 1.28
SICAR	46.57 ± 0.53	45.96 ± 0.56	43.41 ± 0.51

Table 15: Across-client alignment for Tiny-ImageNet / VGG16. Values are mean ± sample SD over three seeds. IR(val-ref) compares dispatch utility with the standalone validation reference; a dash denotes an undefined correlation.

Method	Pearson	Spearman	Pairwise	IR(val-ref)
FedAvg	-	-	0.500 ± 0.000	1.000 ± 0.000
CFFL	0.682 ± 0.188	0.652 ± 0.052	0.729 ± 0.041	1.000 ± 0.000
CGSV	-0.224 ± 0.341	-0.419 ± 0.217	0.328 ± 0.083	1.000 ± 0.000
IAFL	0.807 ± 0.093	0.773 ± 0.091	0.778 ± 0.025	1.000 ± 0.000
ShapFed	0.508 ± 0.149	0.409 ± 0.385	0.659 ± 0.149	1.000 ± 0.000
Quota surrogate	-	-	0.500 ± 0.000	1.000 ± 0.000
Aequa	0.897 ± 0.035	0.858 ± 0.089	0.838 ± 0.047	1.000 ± 0.000
FedAKD	0.597 ± 0.094	0.522 ± 0.117	0.695 ± 0.042	1.000 ± 0.000
FedRAC	0.647 ± 0.345	0.881 ± 0.082	0.883 ± 0.065	0.967 ± 0.058
SICAR	0.855 ± 0.050	0.904 ± 0.020	0.891 ± 0.024	1.000 ± 0.000

We obtain Spearman 0.904 ± 0.020 and dispatch utility $43.41 \pm 0.51\%$. The global mean is below independent FedAvg (46.57% versus 46.92%), while the paired global $C - A$ is positive on average. The independent and paired clean references are separate runs.

F.8 TINY-IMAGENET / DENSENET121

We obtain shared accuracy $50.45 \pm 0.16\%$, dispatch utility $46.70 \pm 0.24\%$, and Spearman 0.920 ± 0.056. Paired recovery is $+0.917 \pm 0.325$ pp over B and $+0.600 \pm 0.291$ pp over A. Several

Table 16: Endpoint performance for Tiny-ImageNet / DenseNet121. Accuracy and macro-F1 are percentages; entries are mean ± sample SD over three seeds.

Method	Global (%)	Macro-F1 (%)	Client (%)
FedAvg	49.88 ± 0.08	49.16 ± 0.14	49.88 ± 0.08
CFFL	27.05 ± 0.41	23.83 ± 0.38	22.16 ± 0.02
CGSV	41.39 ± 0.70	40.27 ± 1.17	41.34 ± 0.76
IAFL	33.20 ± 1.25	30.95 ± 1.15	27.54 ± 1.57
ShapFed	29.27 ± 1.82	27.59 ± 1.18	21.13 ± 0.86
Quota surrogate	33.48 ± 0.81	31.15 ± 0.88	33.48 ± 0.81
Aequa	35.13 ± 0.17	33.19 ± 0.23	34.40 ± 0.20
FedAKD	30.90 ± 0.92	28.19 ± 1.10	23.93 ± 0.25
FedRAC	37.90 ± 0.91	36.30 ± 0.81	37.06 ± 0.30
SICAR	50.45 ± 0.16	49.70 ± 0.22	46.70 ± 0.24

Table 17: Across-client alignment for Tiny-ImageNet / DenseNet121. Values are mean ± sample SD over three seeds. IR(val-ref) compares dispatch utility with the standalone validation reference; a dash denotes an undefined correlation.

Method	Pearson	Spearman	Pairwise	IR(val-ref)
FedAvg	-	-	0.500 ± 0.000	1.000 ± 0.000
CFFL	0.798 ± 0.116	0.692 ± 0.158	0.791 ± 0.069	1.000 ± 0.000
CGSV	0.416 ± 0.509	0.342 ± 0.668	0.650 ± 0.256	1.000 ± 0.000
IAFL	0.786 ± 0.182	0.600 ± 0.217	0.708 ± 0.072	1.000 ± 0.000
ShapFed	0.617 ± 0.124	0.492 ± 0.196	0.695 ± 0.088	1.000 ± 0.000
Quota surrogate	-	-	0.500 ± 0.000	1.000 ± 0.000
Aequa	0.803 ± 0.126	0.755 ± 0.180	0.798 ± 0.081	1.000 ± 0.000
FedAKD	0.780 ± 0.164	0.803 ± 0.055	0.851 ± 0.025	1.000 ± 0.000
FedRAC	0.571 ± 0.060	0.532 ± 0.328	0.697 ± 0.126	1.000 ± 0.000
SICAR	0.880 ± 0.056	0.920 ± 0.056	0.910 ± 0.046	1.000 ± 0.000

independent baselines have shorter or mixed training budgets; these rankings provide context rather than a controlled cross-method advantage.

F.9 SVHN / RESNET18

Table 18: Endpoint performance for SVHN / ResNet18. Accuracy and macro-F1 are percentages; entries are mean ± sample SD over three seeds.

Method	Global (%)	Macro-F1 (%)	Client (%)
FedAvg	95.15 ± 0.13	94.75 ± 0.14	95.15 ± 0.13
CFFL	87.61 ± 5.42	87.29 ± 4.46	80.70 ± 3.11
CGSV	94.62 ± 0.25	94.23 ± 0.23	94.48 ± 0.22
IAFL	35.73 ± 28.37	24.78 ± 30.71	27.96 ± 19.70
ShapFed	92.42 ± 0.46	91.78 ± 0.48	84.90 ± 1.54
Quota surrogate	58.57 ± 33.85	48.32 ± 41.10	58.57 ± 33.85
Aequa	93.61 ± 0.85	93.02 ± 0.99	93.28 ± 0.84
FedAKD	19.90 ± 0.47	3.96 ± 0.98	49.87 ± 4.43
FedRAC	92.55 ± 1.92	91.87 ± 2.16	81.01 ± 5.79
SICAR	95.95 ± 0.25	95.63 ± 0.28	95.94 ± 0.25

Table 19: Across-client alignment for SVHN / ResNet18. Values are mean ± sample SD over three seeds. IR(val-ref) compares dispatch utility with the standalone validation reference; a dash denotes an undefined correlation.

Method	Pearson	Spearman	Pairwise	IR(val-ref)
FedAvg	-	-	0.500 ± 0.000	1.000 ± 0.000
CFFL	0.648 ± 0.098	0.847 ± 0.093	0.867 ± 0.059	1.000 ± 0.000
CGSV	0.725 ± 0.061	0.665 ± 0.165	0.748 ± 0.105	1.000 ± 0.000
IAFL	0.472 ± 0.135	0.637 ± 0.190	0.744 ± 0.087	0.067 ± 0.058
ShapFed	0.216 ± 0.167	0.519 ± 0.130	0.682 ± 0.068	1.000 ± 0.000
Quota surrogate	-	-	0.500 ± 0.000	0.400 ± 0.520
Aequa	0.947 ± 0.015	0.778 ± 0.210	0.811 ± 0.100	1.000 ± 0.000
FedAKD	0.753 ± 0.200	0.774 ± 0.250	0.844 ± 0.111	0.033 ± 0.058
FedRAC	0.871 ± 0.102	0.749 ± 0.114	0.785 ± 0.076	0.900 ± 0.000
SICAR	0.044 ± 0.224	0.166 ± 0.222	0.563 ± 0.091	1.000 ± 0.000

Our global and dispatch accuracies remain close, but the mean Pearson is 0.043 and Fine Pearson is -0.040. Several baseline runs have high dispersion or low accuracy under our adaptations. We retain these outcomes and give priority to paired controls when assessing cancellation.

F.10 SVHN / VGG16

Table 20: Endpoint performance for SVHN / VGG16. Accuracy and macro-F1 are percentages; entries are mean ± sample SD over three seeds.

Method	Global (%)	Macro-F1 (%)	Client (%)
FedAvg	94.90 ± 0.11	94.47 ± 0.10	94.90 ± 0.11
CFFL	89.78 ± 4.72	89.59 ± 3.83	83.73 ± 3.34
CGSV	94.81 ± 0.12	94.37 ± 0.18	94.71 ± 0.03
IAFL	94.44 ± 0.29	93.95 ± 0.30	90.32 ± 2.55
ShapFed	93.23 ± 0.40	92.66 ± 0.27	83.88 ± 2.20
Quota surrogate	92.95 ± 0.55	92.19 ± 0.68	92.95 ± 0.55
Aequa	95.15 ± 0.06	94.74 ± 0.04	94.90 ± 0.06
FedAKD	94.58 ± 0.65	94.05 ± 0.79	83.40 ± 2.75
FedRAC	94.32 ± 0.25	93.76 ± 0.31	84.75 ± 1.01
SICAR	94.84 ± 0.17	94.40 ± 0.12	94.78 ± 0.18

We observe Fine Pearson 0.602 ± 0.112, but across-client Spearman varies substantially (0.032 ± 0.701). Aequa achieves higher mean global accuracy and Pearson. These outcomes distinguish class target realization from reliable ordering across clients.

Table 21: Across-client alignment for SVHN / VGG16. Values are mean ± sample SD over three seeds. IR(val-ref) compares dispatch utility with the standalone validation reference; a dash denotes an undefined correlation.

Method	Pearson	Spearman	Pairwise	IR(val-ref)
FedAvg	-	-	0.500 ± 0.000	1.000 ± 0.000
CFFL	0.928 ± 0.040	0.906 ± 0.047	0.893 ± 0.045	1.000 ± 0.000
CGSV	0.423 ± 0.292	0.293 ± 0.503	0.615 ± 0.173	1.000 ± 0.000
IAFL	0.858 ± 0.109	0.869 ± 0.039	0.830 ± 0.061	1.000 ± 0.000
ShapFed	0.866 ± 0.068	0.818 ± 0.076	0.844 ± 0.039	1.000 ± 0.000
Quota surrogate	-	-	0.500 ± 0.000	1.000 ± 0.000
Aequa	0.912 ± 0.012	0.873 ± 0.080	0.844 ± 0.044	1.000 ± 0.000
FedAKD	0.601 ± 0.432	0.794 ± 0.202	0.844 ± 0.080	0.967 ± 0.058
FedRAC	0.649 ± 0.093	0.967 ± 0.031	0.937 ± 0.042	0.900 ± 0.000
SICAR	0.142 ± 0.600	0.032 ± 0.701	0.515 ± 0.283	1.000 ± 0.000

F.11 SVHN / DenseNet121

Table 22: Endpoint performance for SVHN / DenseNet121. Accuracy and macro-F1 are percentages; entries are mean ± sample SD over three seeds.

Method	Global (%)	Macro-F1 (%)	Client (%)
FedAvg	95.62 ± 0.23	95.27 ± 0.23	95.62 ± 0.23
CFFL	91.93 ± 3.60	91.70 ± 3.05	83.26 ± 3.53
CGSV	95.17 ± 0.27	94.79 ± 0.29	95.15 ± 0.30
IAFL	95.34 ± 0.23	94.96 ± 0.24	91.49 ± 2.25
ShapFed	93.95 ± 1.05	93.55 ± 0.86	84.40 ± 2.09
Quota surrogate	95.41 ± 0.22	95.00 ± 0.29	95.41 ± 0.22
Aequa	95.09 ± 0.06	94.66 ± 0.09	94.93 ± 0.13
FedAKD	94.77 ± 0.25	94.26 ± 0.33	85.10 ± 0.57
FedRAC	94.57 ± 0.39	94.05 ± 0.47	86.17 ± 1.02
SICAR	95.61 ± 0.20	95.29 ± 0.13	95.61 ± 0.21

Table 23: Across-client alignment for SVHN / DenseNet121. Values are mean ± sample SD over three seeds. IR(val-ref) compares dispatch utility with the standalone validation reference; a dash denotes an undefined correlation.

Method	Pearson	Spearman	Pairwise	IR(val-ref)
FedAvg	-	-	0.500 ± 0.000	1.000 ± 0.000
CFFL	0.898 ± 0.101	0.887 ± 0.123	0.889 ± 0.077	1.000 ± 0.000
CGSV	0.447 ± 0.220	0.323 ± 0.208	0.611 ± 0.069	1.000 ± 0.000
IAFL	0.791 ± 0.060	0.865 ± 0.035	0.841 ± 0.006	1.000 ± 0.000
ShapFed	0.759 ± 0.130	0.814 ± 0.145	0.844 ± 0.080	1.000 ± 0.000
Quota surrogate	-	-	0.500 ± 0.000	1.000 ± 0.000
Aequa	0.890 ± 0.089	0.858 ± 0.194	0.863 ± 0.124	1.000 ± 0.000
FedAKD	0.827 ± 0.039	0.855 ± 0.085	0.867 ± 0.039	1.000 ± 0.000
FedRAC	0.597 ± 0.059	0.464 ± 0.777	0.678 ± 0.375	0.900 ± 0.000
SICAR	0.214 ± 0.587	0.180 ± 0.598	0.563 ± 0.221	1.000 ± 0.000

Our shared accuracy is close to independent FedAvg (95.61% in both cases after rounding), and paired global $C - B$ is $+0.011 \pm 0.047$ pp. Across-client alignment and Fine Pearson remain weak or variable.

F.12 Allocation in the high-accuracy regime

Table 24 consolidates the SVHN allocation evidence alongside an Aequa reference. SICAR endpoint accuracy remains close to the paired controls, while its across-client ranking is less pronounced. For ResNet18, Pearson is 0.043 ± 0.224, compared with Aequa's 0.947 ± 0.015 in the descriptive

inventory; SICAR's Fine C is -0.040 ± 0.028. VGG16 retains positive mean within-client class alignment despite variable client ranking. These results delimit generality beyond the stronger class-allocation regime, and the unequal budgets prevent a strictly matched ranking interpretation.

Table 24: Allocation on Street View House Numbers (SVHN). The first four metrics describe SICAR. We include Aequa Pearson for context; training budgets differ on ResNet18. Values are mean $\pm$ sample SD over three seeds.

Model	Pearson	Spearman	Pairwise	Fine C	Aequa Pearson
ResNet18	0.044 ± 0.224	0.166 ± 0.222	0.563 ± 0.091	-0.040 ± 0.028	0.947 ± 0.015
VGG16	0.142 ± 0.600	0.032 ± 0.701	0.515 ± 0.283	0.602 ± 0.112	0.912 ± 0.012
DenseNet121	0.214 ± 0.587	0.180 ± 0.598	0.563 ± 0.221	0.067 ± 0.104	0.890 ± 0.089

F.13 Carrier, strength, and response diagnostics

A direct carrier comparison. Figure 3(a,b) supports selecting directions through class responses on CIFAR-100 with ResNet18. The response-aware carrier improves global C accuracy by 1.57 ± 0.28 points and the C-minus-B contrast by 3.24 ± 0.22 points over a random carrier orthogonal to clean history. Both improvements occur in all three paired seeds. We obtain global C means of 58.14% and 56.57% for the response-aware and random carriers, respectively. We match carrier rank and the clean-history orthogonality constraint. The result supports response-aware direction selection under this control, without isolating every effect of the induced response and perturbation norms. The random carrier meets the response gate in zero of three seeds. Its mean C-minus-B is approximately -0.84 points, so its mean B accuracy is approximately 57.41%, above both random C and clean A. Thus retaining an arbitrary perturbation need not be harmful; cancellation is beneficial in a response-dependent regime.

Our SVHN / VGG16 validation comparisons give C-minus-B gains of $+0.519$ and $+0.154$ points for response-aware over random carriers in seeds 1 and 2. Their random C accuracies are 88.170% and 90.321%, versus 88.496% and 90.398% with the response carrier. We exclude seed 0 from this paired endpoint comparison because its random arm did not complete a valid endpoint; its response arm also fails the response-residual gate. These development runs provide secondary evidence only.

Table 25: CIFAR-100 / ResNet18 carrier endpoints. The common clean A accuracy is $56.61 \pm 0.56\%$. For the random carrier, we reconstruct each B endpoint as $B = C - (C - B)$ from the rounded seed values, so its mean and sample SD are approximate. Response-aware values use our full aggregate summary. Global accuracies are percentages and contrasts are percentage points.

Carrier	Global B (%)	Global C (%)	$C - B$ (pp)
Random (approx.)	57.41 ± 0.41	56.57 ± 0.48	-0.84 ± 0.23
Response-aware	55.74 ± 0.66	58.14 ± 0.72	$+2.40 \pm 0.42$

Reward strength. The validation-only scan in Figure 3(c) quantifies the role of perturbation strength on CIFAR-100 / ResNet18. For $s = 0.25, 0.5, 0.75, 1.0$, the C-minus-B means are 0.40, 1.53, 2.43, and 2.68 points, with sample SDs 0.22, 0.64, 0.95, and 1.09. The corresponding C-minus-A means are 0.12, 0.13, 0.15, and -0.43 points, with sample SDs 0.28, 0.54, 0.36, and 0.29. The fixed default $s = 0.5$ gives positive mean contrasts against both controls in this scan, while the largest strength increases recovery over B at a cost relative to A. This short-run experiment describes sensitivity; it does not establish an optimal strength at the formal test endpoint. In particular, $s = 0.75$ also has positive mean C-minus-A. We retain the frozen default and do not select a new strength from these results. The 0.13-point validation C-minus-A at $s = 0.5$ and the 1.53-point formal test value compare different training endpoints and splits, so their difference cannot be attributed to either factor alone.

Diagnostic quantities. A clean twin at a common server state measures classifier influence $\tau_h = E^\top[F(\theta + E\delta) - F(\theta)]$. The diagnostic statistics are

$$D_z = \frac{\|\tau_h - \delta\|}{\|\delta\|}, \tag{52}$$

$$\eta_{\text{param}} = 1 - \frac{\|\tau_h - \delta\|}{\|\tau_h\| + \varepsilon}, \tag{53}$$

$$\eta_{\text{resp}} = 1 - \frac{\|R(\tau_h - \delta)\|}{\|R\tau_h\| + \varepsilon}, \tag{54}$$

$$\text{retention} = \frac{\left\|PP^\top \tau_h\right\|^2}{\|\tau_h\|^2}. \tag{55}$$

D_z uses injection norm as its scale; the two η statistics measure reduction relative to parameter or class-response influence. Retention is the squared fraction of classifier influence in the carrier. These diagnostics complement endpoint contrasts, with their scope determined by the observed classifier and fixed response operator.

Response-space evidence at the endpoint. We report the mean and sample SD of the within-seed median η_{resp} in Table 26. All six CIFAR-100 and Tiny-ImageNet settings have positive seed-level medians in all three seeds. This connects the observed endpoint contrasts to a local reduction in class-response discrepancy at the evaluated checkpoints. It does not assert that every client or earlier checkpoint passes: our earlier Tiny-ImageNet / DenseNet121 seed 2 and CIFAR-100 / DenseNet121 seeds 0 and 1 diagnostics include failures. On SVHN, the positive median response diagnostics coexist with small endpoint accuracy changes. We therefore interpret response reduction as a local diagnostic rather than a performance guarantee.

Table 26: Endpoint response diagnostics. We first compute the median across clients in each seed, then its mean and sample SD across three seeds.

Dataset	Architecture	Median η_{resp}
CIFAR-100	ResNet18	0.351 ± 0.025
CIFAR-100	VGG16	0.090 ± 0.047
CIFAR-100	DenseNet121	0.059 ± 0.014
Tiny-ImageNet	ResNet18	0.387 ± 0.039
Tiny-ImageNet	VGG16	0.444 ± 0.022
Tiny-ImageNet	DenseNet121	0.204 ± 0.040
SVHN	ResNet18	0.039 ± 0.023
SVHN	VGG16	0.201 ± 0.099
SVHN	DenseNet121	0.081 ± 0.062

Auxiliary contribution-proxy check. We also examine a restricted coalition game with a frozen backbone, one classifier-head update, and server-validation utility over all 1,024 client coalitions. Our auxiliary correlations vary in sign across seeds and utility definitions. We therefore regard this check as inconclusive and do not use it to validate class-wise contribution. A class-resolved marginal-contribution analysis would test a different claim from target realization and remains necessary to validate the allocation policy independently.

G Scope, Limitations, and Computational Cost

G.1 One-time construction and deployment cost

The fixed construction allows reuse of the response fit across clients and across training. Let L be the retained history length, r_Q its retained rank, and $r \le K - 1$ the carrier rank. Storing the history costs $O(ML)$ numbers, the response $O(KM)$, the carrier $O(Mr)$, and saved rewards $O(NM)$. A dense thin SVD of an $M \times L$ history costs approximately $O(ML^2)$; forming the projected weighted response costs $O(MKr_Q)$. For $M \ge K$, a dense thin SVD of the $M \times K$ response matrix costs $O(MK^2)$.

Forming $G = RP$ costs $O(KMr)$. The ridge Gram matrix requires $O(Kr^2)$ work and its factorization $O(r^3)$. Each right-hand side and solve costs $O(Kr + r^2)$, followed by $O(Mr)$ to form a classifier reward. These are one-time arithmetic costs; probe differentiation additionally depends on data cardinality, batching, and the network forward pass.

Deployment adds and subtracts an M-dimensional saved reward per client, adding $O(NM)$ arithmetic to full-model communication and local training. Clean controls and diagnostic twins are evaluation tools, not deployment requirements. Standalone training and calibration remain setup costs. We separately measure total elapsed time for our three-branch A/B/C experiments. The per-seed means are 3.70, 1.80, and 6.38 hours on CIFAR-100; 22.96, 9.34, and 81.73 hours on Tiny-ImageNet; and 5.46, 1.55, and 9.17 hours on SVHN, in ResNet18, VGG16, and DenseNet121 order. These are experimental costs for all three branches, not standalone SICAR deployment costs, and reference reuse may exclude the cost of constructing the reference. We draw no deployment speedup or energy claim from them.

G.2 Theoretical and empirical scope

Known offset cancellation exactly removes the saved additive reward. Changes induced by local optimization remain outside direct subtraction, including effects transmitted to the backbone. Equation (18) gives the condition for reduced local discrepancy; neither historical orthogonality nor a positive projected diagnostic implies exact recovery of the clean trajectory. Class alignment evaluates realization of a participation-based target, rather than causal contribution or an economic incentive guarantee.

We obtain the strongest shared-recovery and class-alignment evidence on CIFAR-100 and Tiny-ImageNet at the chosen strength. Dispatch utility carries a cost, and the SVHN allocation signal is weaker. High baseline accuracy and different response sensitivity are possible explanations, not experimentally established causes. The current protocol covers one partition concentration and full participation; varying reward strength, population composition, or data drift may change the regime.

G.3 Information, participation, and evidence boundaries

We assume reliable counts, a labeled server probe, meaningful validation scores, and compliant local execution. Counts and scores reveal information beyond model updates, and differences in standalone quality can mix class coverage, sample difficulty, and training resources. The allocation policy therefore needs an application-specific justification and does not itself provide privacy or demographic fairness.

We supply each client with its perturbed dispatch. Our setting assumes truthful class counts and compliant execution; it does not provide incentive compatibility against strategic reports. In particular, reporting a class distribution equal to q would produce a zero class direction. All clients receive the same backbone, so model exchange, bias recalibration, or classifier retraining could also weaken the intended allocation. We have not measured resistance to those actions and do not interpret our mechanism as irreversible knowledge withholding.

Our positive shared-model contrasts coexist with lower mean dispatch utility on CIFAR-100 and Tiny-ImageNet. This establishes neither individual rationality relative to clean FedAvg nor a benefit for every high-contribution client. Our client-level common-distribution comparisons in Table 5 show a negative dispatch contrast for every observed CIFAR-100 and Tiny-ImageNet client. An economic participation claim would also require modeling participation costs and outside options. Our present contribution concerns controllable class-wise allocation and the shared-training consequences of a known offset under compliant participation.

H DEFINITIONS OF NOTATIONS

Table 27 groups the notation by the data composition, reward construction, training flow, and evaluation quantities used in SICAR. Dimensions refer to column vectors unless stated otherwise. The labels A, B, and C denote experimental conditions; italic B and D denote a matrix and a parameter dimension in their explicitly defined mathematical contexts.

Table 27: Definitions of notations.

Notation	Definition
Clients and relative class composition	
N, K	Numbers of clients and classes, respectively.
i, c, k	Client index and class indices.
n_i, n_{ic}	Number of training examples at client i, and number belonging to class c.
w_i	Sample weight $n_i / \sum_j n_j$ used in aggregation.
p_i	Client class distribution in $\mathbb{R}^K$, with $p_{ic} = n_{ic}/n_i$.
q	Federation class distribution, $q = \sum_i w_i p_i$.
$\widetilde{p}_i$	Smoothed client distribution $(n_i p_i + \alpha q)/(n_i + \alpha)$.
Δ_{++}^{K-1}	Interior of the probability simplex: positive class proportions summing to one.
α	Total pseudocount mass used to smooth each client distribution toward q.
C_K	Centering matrix $I_K - K^{-1}\mathbf{1}\mathbf{1}^\top$.
$\mathrm{clr}(p)$	Centered log-ratio coordinate vector $C_K \log p$.
s_i	Relative class contrast $\mathrm{clr}(\widetilde{p}_i) - \mathrm{clr}(q)$, whose coordinates sum to zero.
d_i	Normalized class direction $s_i / \max(\|s_i\|_2, \varepsilon)$.
ε	Small positive stabilizer for normalization and diagnostic denominators.
Model and class-response construction	
θ, D	Full model parameters in $\mathbb{R}^D$, and their dimension.
M, E	Classifier dimension and coordinate embedding $E \in \mathbb{R}^{D\times M}$, with $E^\top E = I_M$.
$\theta_\star$	Reference model used to construct the fixed reward.
V_c	Labeled server-validation probe examples belonging to class c.
$f_k(x;\theta)$	Logit for class k on input x.
$m_c(\theta)$	Probe-averaged margin between class c and the mean logit of the other classes.
$m(\theta)$	Vector of the K class margins.
R	Classifier-response matrix in $\mathbb{R}^{K\times M}$; row c is the gradient of m_c at $\theta_\star$.
$\mathcal{H}$, U	Retained clean-update index set and classifier-update matrix $U \in \mathbb{R}^{M\times L}$.
u_{it}	Clean classifier update $E^\top(F_{it}(\theta_t) - \theta_t)$ recorded for the history.
L, r_Q	Number of retained history vectors and retained history rank.
Q	Orthonormal basis for retained clean-history directions, in $\mathbb{R}^{M\times r_Q}$.
W	Diagonal response-weight matrix $\mathrm{diag}((q + 10^{-8})^{-1})$.
H	Projected weighted response $(I_M - QQ^\top)R^\top W^{1/2}$.
r, P	Carrier rank and orthonormal carrier basis $P \in \mathbb{R}^{M\times r}$, with $r \le K-1$ and $Q^\top P = 0$.
G	Response restricted to the carrier, $G = RP \in \mathbb{R}^{K\times r}$.
v, $J(z;v)$	Desired class-margin vector and objective $(Gz - v)^\top W(Gz - v) + \beta^{-1} z^\top z$.
β, κ	Positive ridge parameter and corresponding penalty coefficient $\kappa = \beta^{-1}$.
$z(v)$	Unique ridge-fit coordinates $(G^\top WG + \beta^{-1} I_r)^{-1} G^\top W v$.
B, σ_j	Weighted response $B = W^{1/2}G$ and its singular values in the response analysis.
A_G, L_G	Positive-definite ridge matrix $A_G = G^\top WG + \beta^{-1} I_r$ and a Cholesky factor $L_G L_G^\top = A_G$.
Reward calibration and saved perturbation	
c_i	Fixed standalone validation accuracy used as a client-quality proxy and amplitude input.
ρ_i	Normalized ascending average rank $\mathrm{rank}_0(c_i)/(N-1)$, with average ranks for ties.
λ, f_i	Rank-based amplitude coefficient and factor $f_i = 1 + \lambda(1 - 2\rho_i)$.

Continued on the next page

Table 27 continued: definitions of notations

Notation	Definition
h_i	Unit-direction response-fit norm $\|Pz(d_i)\|_2$.
m_{ref}	Positive reference norm used for same-seed amplitude calibration.
a_0	Normalization scale $m_{\mathrm{ref}}/\operatorname{median}_i(f_i h_i)$.
s, a_i	Global reward strength and client amplitude $a_i = s a_0 f_i$.
t_i, z_i	Scaled class target $t_i = a_i d_i$ and fitted coordinates $z_i = z(t_i)$.
δ_i	Saved classifier reward $Pz_i \in \mathbb{R}^M$; its full-model embedding is $E\delta_i$.
Local training, cancellation, and analysis	
$t_\star$, t, T	Start index of reward deployment, communication index, and final endpoint index.
F_{it}	Local full-model training map for client i at communication index t.
$x_{it,j}$	Local parameters after minibatch step j, starting from $x_{it,0} = \theta_t + E\delta_i$.
J_i	Number of local minibatch steps for client i.
ℓ_{itj}	Local minibatch loss used at step j.
$v_{it,j}$	Momentum vector in the local training recursion.
η_t, μ	Local learning rate and momentum coefficient.
$\widehat{\theta}_{t+1}^{(i)}$	Uploaded trained model $F_{it}(\theta_t + E\delta_i)$.
$\widetilde{\phi}_{it}$	Corrected client model $F_{it}(\theta_t + E\delta_i) - E\delta_i$.
a	Full-model injected offset $E\delta_i$ in the cancellation analysis.
τ	Counterfactual local influence $F(\theta + a) - F(\theta)$ under matched training randomness.
e	Corrected local discrepancy $\tau - a$.
ξ_j, b_j	Differences between rewarded and clean local parameters and momentum vectors at step j.
e_{agg}	Common-start aggregate discrepancy $\sum_i w_i e_i$.
L_j, L_t, L_f	Conditional Lipschitz constants for a minibatch gradient, a clean aggregate map, and logits, respectively.
Evaluation and response diagnostics	
A, B, C	Clean control, reward retention without cancellation, and SICAR with cancellation, respectively.
u_g, u_i	Accuracy of the shared endpoint and client i's dispatched model, before additional local adaptation.
u_{ic}	Client i's dispatched-model accuracy on class c.
Δu_{ic}	Change in class accuracy relative to clean A, $u_{ic}^C - u_c^A$.
$\mathcal{T}$, $\mathcal{V}$	Endpoint evaluation and standalone validation sets.
$\mathcal{C}_i$	Eligible classes for a within-client correlation, with finite target and evaluation values.
$\rho(x, y)$	Across-client Pearson correlation between two client-level vectors.
ρ_i^{class}	Within-client class correlation $\mathrm{Corr}_{c\in\mathcal{C}_i}(d_{ic}, \Delta u_{ic})$; Fine Pearson averages it over clients.
$\widehat{\mathrm{IR}}_{\mathrm{cross}}$	Cross-split proxy $N^{-1}\sum_i \mathbf{1}\{u_i(\mathcal{T}) \geq c_i(\mathcal{V})\}$.
n, $\bar{x}$, s_x	Number of independent runs, their arithmetic mean, and sample standard deviation. The experimental summaries use three seeds.
τ_h	Classifier component of the common-start counterfactual influence, $E^\top \tau$.
D_z	Classifier residual norm divided by injected reward norm.
η_{param}	Reduction of classifier residual norm relative to the counterfactual classifier influence.
η_{resp}	Reduction of residual class-response norm measured through R.
retention	Squared fraction of classifier influence lying in the carrier: $\left\|PP^\top \tau_h\right\|^2 / \|\tau_h\|^2$.