WHEN BACKDOORS DISAGREE: SOURCE-LEVEL FEDERATED UNLEARNING WITH ASYMMETRIC CONSENSUS

ID #58405

ABSTRACT

Federated unlearning removes the influence of departing clients from a trained model without retraining it, and it is usually evaluated with a single deletion source. A real request may cover several groups of clients, each of which planted a different backdoor. Their deletion objectives can conflict: when the sources pull in opposing directions, one shared update may leave a source without progress, and serving all of them takes longer steps the more they cancel. We call this tension the *coverage bind*. A surviving source is also easy to miss, since the mean attack success over sources can be low while one backdoor persists, so we test every source separately. We propose *Federated Asymmetric Consensus* (FAC), whose two design choices respond to the bind: each client keeps its own objective, and the disagreement between clients and the shared model is carried across rounds rather than lost at every aggregation, and the server penalizes changes to the parameters that retained behavior relies on. The server coordinates unlearning using the forget/retain partition, without trigger specifications or source labels. In a source-scaling study with up to seven sources, FAC is the only one of sixteen approximate unlearning methods that leaves no source measurably above retraining at every source count tested. Once there are several sources, it also retains the most accuracy among the methods that do so. With three sources, FAC leaves no survivor in eight of nine residual-network runs across three datasets; on a vision transformer, neither FAC nor SalUn removes more sources on every dataset, and FAC retains more accuracy. Across 144 endpoint runs with three to seven sources, 58% show a gap of at least 20 points between worst-source and mean attack success, so averaged scores routinely hide surviving backdoors.

1 INTRODUCTION

Federated learning trains one model across parties that keep their data private (McMahan et al., 2017), which also lets a compromised party implant a backdoor through its local updates (Bagdasaryan et al., 2020). Suppose a visual-inspection model trained across factories turns out to carry backdoors from three supplier networks: a sticker passes a cracked part, a checker pattern hides a missing screw, and a paint mark relabels an overheated component. The owner must remove all three while keeping what the honest factories contributed, and without repeating training, which is the task of federated unlearning (Liu et al., 2021b; Zhang et al., 2023). Two questions follow. How can we tell whether every backdoor has been removed? And can one unlearning run remove several different backdoors at once? An average attack success rate can conceal a surviving backdoor, and with several sources, deletion objectives can conflict.

We begin with the first question. Existing methods (Pan et al., 2025; Khalil et al., 2025; Li et al., 2026; Weng et al., 2026; Fan et al., 2024; Jia et al., 2023) usually summarize deletion with one attack success rate (ASR). In the factory example, a model that no longer responds to the sticker or the paint mark may still obey the checker pattern, and one number over all three hides it; in our experiments, a low average often hides a surviving backdoor. We therefore test each source on its own clients' held-out triggered probes and call it a *survivor* while its attack success stays measurably above the level that retraining reaches; a request is met only when no source survives.

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.

The second question is harder, and Proposition 1 shows why. If the sources' normalized forget directions are summed into one update, a source whose direction is on balance opposed by the others gets no first-order descent from it; and any step that descends every source by a fixed margin must lengthen as their directions cancel, and does not exist when they cancel exactly. We call this tension the *coverage bind*. These geometric observations motivate two design choices: carrying client–consensus disagreement across rounds and constraining changes to parameters that support retained behavior. We hypothesize that the sources left behind in practice are those the aggregate of source updates does not cover, and that carrying client–consensus disagreement across rounds lets such sources still be removed.

We propose *Federated Asymmetric Consensus* (FAC; Figure 1) to incorporate these two choices. FAC casts unlearning as an ADMM-style consensus problem (Boyd et al., 2011). Departing clients minimize a suppression loss and the others their task loss. The server carries client–consensus disagreement across rounds and uses it to correct both the local targets and the model that enters consensus. A contrast between forget- and retain-client displacements then determines the spatial penalty: selected positive-score coordinates are drawn toward zero, and the remaining coordinates are anchored to the pre-unlearning model. The server coordinates unlearning using the requested forget/retain partition and scores the returned client updates, without trigger specifications or source labels; as in FedOSD (Pan et al., 2025), departing clients take part in unlearning on their local data, to which they apply their own trigger.

Our main finding is that most methods leave exactly one source behind once a request has several: from three sources on, GA, Sequential, FedOSD, FedCARE, RFE and FedRazor usually remove all but one source, whose attack success stays near 100%. Across one, three, five and seven sources, FAC is the only one of sixteen approximate unlearning methods that leaves no survivor at every count, and from three sources on, no other survivor-free method keeps as much accuracy. The evidence supports our hypothesis: in all 30 runs of the baselines that descend a forget objective at five sources, the source with non-positive mean coverage survives, and replacing FAC's corrected target with the shared model, or removing its disagreement state, brings survivors back. With three sources on a residual network, FAC leaves no survivor on CIFAR-10 or SVHN and one in one of three CIFAR-100 seeds; at seven sources it leaves one or two. On a vision transformer, FAC and SalUn differ in deletion outcomes across datasets, while FAC retains more accuracy. Permuting the sensitivity scores within each tensor, with their distribution and selected-coordinate count unchanged, leaves two sources in every seed, so deletion depends on which coordinates the penalty selects, not only on how many.

Contributions. (i) We formulate multi-source federated unlearning and identify the coverage bind between covering every source and the step length this requires. (ii) We propose FAC, which carries client–consensus disagreement across rounds and discourages changes to the parameters retained clients rely on, through iterative consensus updates. (iii) We evaluate deletion source by source across one to seven sources, three datasets and three architectures, and show that averaged attack success misses survivors that per-source tests detect: in 58% of 144 endpoint runs, the worst source exceeds the mean by at least 20 points.

2 Related Work

Federated unlearning. Federated unlearning recalibrates or perturbs a deletion-aware training trajectory (Liu et al., 2021b; Zhang et al., 2023) or changes the trained model directly through weight negation in NoT (Khalil et al., 2025) and the sparse adapters of FUSED (Zhong et al., 2025); FOUL (Nguyen et al., 2026) and GDFA (Weng et al., 2026) address client heterogeneity. Closest to our setting, FedOSD (Pan et al., 2025), FedCARE (Li et al., 2026) and FedRazor (Hu et al., 2026) reduce forget–retain conflicts through update directions, gradient projection or trimming, and constrained recovery. A multi-source request adds a second kind of conflict, among the forget objectives themselves; we evaluate all three as baselines and test each source separately (further work in Appendix H).

Conflicting objectives and selective forgetting. PCGrad, MGDA and CAGrad (Yu et al., 2020; Sener & Koltun, 2018; Liu et al., 2021a) resolve conflicts among current gradients and take one combined step per iteration, which must then cover every source at once, the step that Proposition 1(ii)

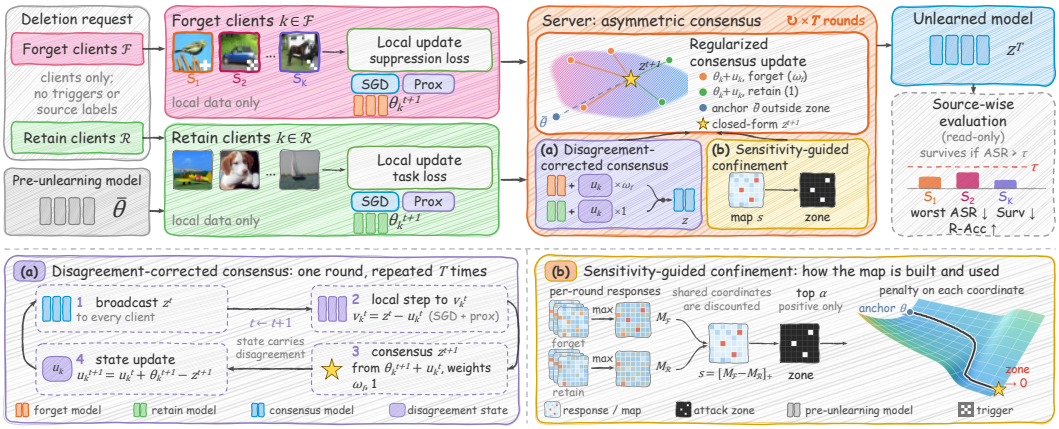

Figure 1: **Overview of FAC.** The request names only forget and retain clients. Each round, clients optimize their local losses with proximal correction when enabled, and the server combines the returned models through a coordinate-wise update; after T rounds, each source S_i is tested against τ_i. (a) The state u_k carries disagreement across rounds. (b) The sensitivity map s selects the attack zone. The recovery stage is not shown.

makes long when sources cancel; we evaluate PCGrad and MGDA as multi-source baselines. Consensus ADMM (Boyd et al., 2011) and FedADMM (Wang et al., 2022) maintain client-level dual states, while FedProx (Li et al., 2020) and FedDyn (Acar et al., 2021) use proximal or dynamic corrections to counter drift among clients that share one task. In FAC the client objectives oppose each other (suppression versus task loss), and the states carry that disagreement across rounds. Its map follows saliency- and sparsity-based unlearning (Fan et al., 2024; Jia et al., 2023), contrasting forget and retain updates to assign a spatial penalty.

Verification and evaluation. VeriFi (Gao et al., 2024) checks markers that a departing client planted, Proof of Unlearning (Weng et al., 2024) authenticates the execution of a prescribed procedure, and IAM (Wang et al., 2025) and RULI (Naderloui et al., 2025) assess completeness and privacy risk per sample. Our source-wise evaluation (Section 3.4) assesses deletion at the returned endpoint.

3 Method: Federated Asymmetric Consensus

FAC takes the pre-unlearning global model $\bar{\theta}$ and a client-level deletion request that partitions the N participating clients into a forget set $\mathcal{F}$ and a retain set $\mathcal{R}$, and it returns a single consensus model. Section 3.1 defines the local objectives and the geometric difficulty they create, Sections 3.2 and 3.3 give the two mechanisms that respond to it, and Section 3.4 specifies the source-wise evaluation against which every run is judged. Figure 1 and Algorithm 1 (Appendix A) summarize the procedure.

3.1 Asymmetric Local Objectives and the Coverage Bind

The forget set is the union of K sources, $\mathcal{F} = \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_K$ ($S_1, \ldots, S_K$ in Figure 1), and the clients in $\mathcal{F}_i$ implant one backdoor, a trigger–target pair (t_i, y_i) that differs across sources (Gu et al., 2019). The server learns which clients must leave, not what their triggers look like. FAC's optimization uses only the partition $(\mathcal{F}, \mathcal{R})$; the source membership $\{\mathcal{F}_i\}$ enters only the evaluation of Section 3.4. Following FedOSD (Pan et al., 2025), forget clients perform local unlearning with its bounded unlearning cross-entropy while retain clients continue task training. With $p_\theta(y \mid x)$ the predicted probability of label y, the local objectives are

$$f_k(\theta) = \begin{cases} \mathbb{E}_{(x,y)\sim\mathcal{D}_k}\big[-\log\big(1 - p_\theta(y \mid x)/2\big)\big], & k \in \mathcal{F}, \\ \mathbb{E}_{(x,y)\sim\mathcal{D}_k}\big[-\log p_\theta(y \mid x)\big], & k \in \mathcal{R}. \end{cases} \tag{1}$$

Minimizing the first loss lowers the probability of the client's own labels, so heterogeneous deletion pressure is present in the local problems rather than created by averaging at the server.

Heterogeneous sources can nevertheless demand incompatible changes. Let $L_i = \sum_{k \in \mathcal{F}_i} f_k$, $g_i = \nabla L_i(\bar{\theta})$ (assumed nonzero), $\hat{g}_i = g_i/\|g_i\|$, $G = [\langle \hat{g}_i, \hat{g}_j \rangle]_{i,j}$, $d = \sum_i \hat{g}_i$, and $c = 1 - \|d\|^2/K$, which is 0 for orthogonal sources and approaches 1 as they cancel.

Proposition 1 (Coverage bind). *(i) If $d \neq 0$, the normalized aggregate $\hat{d} = d/\|d\|$ covers source i by $\langle \hat{d}, \hat{g}_i \rangle = (G\mathbf{1})_i/\sqrt{K(1-c)}$, so a source with a non-positive Gram row sum receives no first-order descent from it. (ii) Any step v with $\langle v, \hat{g}_i \rangle \geq \kappa > 0$ for every source satisfies $\|v\| \geq \kappa\sqrt{K/(1-c)}$, and no such step exists when $d = 0$.*

A shared direction can therefore leave an opposed source uncovered, while covering every source requires a step whose length grows as the sources cancel. The proposition characterizes local update geometry; retained utility is evaluated empirically (proof and scope in Appendix A.3). Part (i) motivates carrying client–consensus disagreement across rounds, so that a source the shared direction leaves uncovered can still make progress (Section 3.2); part (ii) motivates keeping the possibly long step away from coordinates that support retained behavior (Section 3.3).

FAC keeps the local objectives separate and reconciles them through one deployable model. For a fixed nonnegative sensitivity map $s \in \mathbb{R}^P_{\geq 0}$ over the P model parameters, the per-round reference problem is

$$\min_{\{\theta_k\},z} \sum_{k=1}^{N} f_k(\theta_k) + \lambda R_s(z; \bar{\theta}) \qquad \text{s.t.} \qquad \theta_k = z, \quad k = 1, \ldots, N, \tag{2}$$

where the regularizer R_s (Section 3.3) limits where z may move away from $\bar{\theta}$. The map is fixed within a round and refreshed between rounds, so Problem (2) specifies one server update rather than a static objective for the whole trajectory.

3.2 Disagreement-Corrected Consensus

FAC couples local optimization through an ADMM-inspired consensus model (Boyd et al., 2011) (Figure 1a). At round t, client k starts from the broadcast model z^t and takes local SGD steps on f_k. Its accumulated disagreement u_k^t defines the proximal target $v_k^t = z^t - u_k^t$. After a local step produces $\widehat{\theta}_{k,e+1}$, the client applies the proximal map

$$\theta_{k,e+1} = \frac{\widehat{\theta}_{k,e+1} + \eta_t \rho_{\mathrm{p}} v_k^t}{1 + \eta_t \rho_{\mathrm{p}}}, \tag{3}$$

where η_t is the learning rate and $\rho_{\mathrm{p}} \geq 0$ the local coupling. States start at zero; the first round is a warm start without correction. Clients that share a broadcast can thus solve toward different targets.

The server reconciles the returned models, which pull the consensus in different directions (Figure 1, top), under the regularizer R_{s^t} of Section 3.3. With $\omega_k = \omega_{\mathrm{f}} > 0$ for forget clients, $\omega_k = 1$ for retain clients, and the round's sensitivity map held fixed, the consensus and the states are updated as

$$z^{t+1} = \arg\min_z \ \frac{\rho}{2} \sum_{k=1}^{N} \omega_k \left\| z - (\theta_k^{t+1} + u_k^t) \right\|_2^2 + \lambda R_{s^t}(z; \bar{\theta}), \tag{4}$$

$$u_k^{t+1} = u_k^t + \theta_k^{t+1} - z^{t+1}, \tag{5}$$

where $\rho > 0$ is the server coupling, independent of ρ_{p}, and ω_{f} sets the consensus mass of each forget client without reweighting its local loss. The state enters both the next local target and the server objective, so a client's disagreement with the consensus is carried across rounds instead of being discarded at each aggregation. Setting $\rho_{\mathrm{p}} = 0$ removes the local correction while keeping the state in the server update. The states then enter only through their weighted sum $\sum_k \omega_k u_k$, a single aggregate state; with $\rho_{\mathrm{p}} > 0$, each state also shifts its client's local target. Section 4.3 ablates the corrected target.

3.3 SENSITIVITY-GUIDED CONFINEMENT

FAC penalizes changes to coordinates estimated to support retained behavior (Figure 1b), estimating them from the client displacements returned each round,

$$q_k^t = \frac{z^t - \theta_k^{t+1}}{\eta_t}, \tag{6}$$

which summarize each client's net local update, including its loss and any proximal correction, without access to its examples. Let $\bar{q}_{\mathcal{F}}^t$ and $\bar{q}_{\mathcal{R}}^t$ be their means over forget and retain clients. Following magnitude-based sensitivity estimation in unlearning (Fan et al., 2024), we accumulate element-wise peak responses and contrast the two groups:

$$M_{\mathcal{F}}^t = M_{\mathcal{F}}^{t-1} \vee |\bar{q}_{\mathcal{F}}^t|, \qquad M_{\mathcal{R}}^t = M_{\mathcal{R}}^{t-1} \vee |\bar{q}_{\mathcal{R}}^t|, \qquad s^t = \left[M_{\mathcal{F}}^t - M_{\mathcal{R}}^t\right]_+, \tag{7}$$

with both accumulators starting at zero and $\vee$, $[\cdot]_+$ acting coordinate-wise. A positive score marks a coordinate whose forget-side peak exceeds its retain-side peak; the maximum keeps earlier evidence when later updates shrink, and the contrast discounts coordinates retained learning also uses.

The attack zone $\mathcal{A}^t$ is formed by a rank threshold on s^t with nominal fraction α, the rank being capped at the number of positive scores so that the threshold stays positive. If no score is positive, the zone is empty. Ties at the threshold are kept and the realized zone size is recorded each round (Appendix A.2). Set $\widetilde{\theta}_j^t = 0$ inside $\mathcal{A}^t$ and $\widetilde{\theta}_j^t = \bar{\theta}_j$ outside it, and with small $\varepsilon, \delta > 0$ define

$$a_j^t = \frac{s_j^t}{\max_\ell s_\ell^t + \varepsilon}, \qquad b_j^t = \frac{1}{a_j^t + \delta}, \qquad R_{s^t}(z; \bar{\theta}) = \sum_{j=1}^{P} \Big[a_j^t |z_j| + b_j^t \big(z_j - \widetilde{\theta}_j^t\big)^2\Big]. \tag{8}$$

Inside the zone both terms pull toward zero, releasing the pre-unlearning values; outside it, small scores give weak shrinkage and strong anchoring to $\bar{\theta}$. The mask thus selects where the original model is released, and the continuous weights set how strongly each coordinate is regularized.

With these quantities fixed, the server subproblem (4) separates by coordinate. Let $h^t = \sum_k \omega_k(\theta_k^{t+1} + u_k^t)$ and $W = \sum_k \omega_k$; then

$$z_j^{t+1} = \mathcal{S}_{\lambda a_j^t / D_j^t}\left(\frac{\rho h_j^t + 2\lambda b_j^t \widetilde{\theta}_j^t}{D_j^t}\right), \qquad D_j^t = \rho W + 2\lambda b_j^t, \tag{9}$$

where $\mathcal{S}_\xi(x) = \mathrm{sign}(x)(|x| - \xi)_+$. This gives the unique solution of the server subproblem for the current sensitivity map. The map needs no additional client message. When internal recovery is enabled, forget clients then leave and retain clients continue with clean FedAvg updates while the final zone is held at zero; otherwise FAC returns the last consensus directly (Algorithm 1).

3.4 SOURCE-WISE EVALUATION

We evaluate each source separately using source annotations and its clients' held-out triggered probes (Figure 1, right). A source survives if its ASR exceeds τ_i, defined as its matched Retrain mean plus three pooled standard deviations. Worst-source attack success and the number of surviving sources (Surv) are the primary deletion outcomes, reported with retained-client accuracy. The survivor test and a residual audit use disjoint probe halves.

The displacements q_k^t that build the sensitivity map also carry the source geometry. Given the membership $\{\mathcal{F}_i\}$, an observer forms the normalized source displacements $\widehat{q}_i^t \propto \sum_{k \in \mathcal{F}_i} q_k^t$, their Gram matrix $G_{ij}^t = \langle \widehat{q}_i^t, \widehat{q}_j^t \rangle$, and $d^t = \sum_i \widehat{q}_i^t$, and records

$$\gamma_i^t = \frac{(G^t \mathbf{1})_i}{\|d^t\|_2}, \qquad c^t = 1 - \frac{\|d^t\|_2^2}{K}. \tag{10}$$

For nonzero source displacements and a nonzero aggregate, these quantities apply the Gram identity of Proposition 1(i) to local displacements: they describe how the sources' local updates relate, not the coverage of the final server update. The observer feeds nothing back into optimization and uses source annotations that the optimization never sees (Appendix B.3); whether a source is deleted is decided by the outcome tests, not by its geometry.

Table 1: **Source-level forgetting as the number of heterogeneous sources grows** (CNN, CIFAR-10, non-IID). Entries are $a_{\pm b}$ with mean a and sample SD b over three seeds; worst-source ASR and R-Acc in %; Surv is the mean number of sources above their Retrain threshold. No Unlearning and Retrain are anchors, not ranked. **Blue** marks the unique lowest ASR per column and every survivor-free Surv entry. *Project adaptation. ‡With the reported coefficient and learning-rate schedule, L1-Sparse yields results close to retain-only fine-tuning.

Method	$K=1$			$K=3$			$K=5$			$K=7$		
	ASR↓	Surv↓	R-Acc↑	ASR↓	Surv↓	R-Acc↑	ASR↓	Surv↓	R-Acc↑	ASR↓	Surv↓	R-Acc↑
No Unlearning	$92.38_{\pm 0.26}$	1.00	$55.17_{\pm 0.51}$	$97.15_{\pm 0.53}$	3.00	$54.87_{\pm 0.45}$	$96.37_{\pm 0.58}$	5.00	$52.43_{\pm 0.45}$	$96.88_{\pm 0.36}$	7.00	$47.93_{\pm 0.70}$
Retrain	$4.65_{\pm 1.48}$	0.00	$55.30_{\pm 0.60}$	$4.41_{\pm 0.46}$	0.00	$57.07_{\pm 0.51}$	$6.63_{\pm 1.05}$	0.00	$57.03_{\pm 0.78}$	$10.23_{\pm 1.10}$	0.00	$52.57_{\pm 0.31}$
FT	$91.61_{\pm 0.80}$	1.00	$56.23_{\pm 0.31}$	$95.71_{\pm 0.52}$	3.00	$56.57_{\pm 0.31}$	$95.16_{\pm 0.80}$	5.00	$54.53_{\pm 0.21}$	$96.66_{\pm 0.44}$	7.00	$50.10_{\pm 0.53}$
NegGrad+	$89.87_{\pm 0.48}$	1.00	$56.23_{\pm 0.25}$	$95.71_{\pm 0.52}$	3.00	$56.63_{\pm 0.35}$	$95.12_{\pm 0.92}$	5.00	$54.77_{\pm 0.31}$	$96.88_{\pm 0.37}$	7.00	$49.90_{\pm 0.66}$
L1-Sparse‡	$91.61_{\pm 0.80}$	1.00	$56.20_{\pm 0.30}$	$95.71_{\pm 0.52}$	3.00	$56.57_{\pm 0.31}$	$95.16_{\pm 0.80}$	5.00	$54.50_{\pm 0.20}$	$96.66_{\pm 0.44}$	7.00	$50.10_{\pm 0.53}$
GA	$0.23_{\pm 0.00}$	**0.00**	$48.97_{\pm 0.49}$	$99.61_{\pm 0.17}$	1.00	$47.03_{\pm 0.64}$	$99.59_{\pm 0.19}$	1.00	$44.40_{\pm 1.00}$	$99.81_{\pm 0.00}$	1.00	$39.53_{\pm 0.40}$
AggSum	$0.23_{\pm 0.00}$	**0.00**	$49.13_{\pm 0.64}$	$53.44_{\pm 6.38}$	1.67	$44.93_{\pm 1.18}$	$98.42_{\pm 1.10}$	2.00	$43.57_{\pm 0.25}$	$98.89_{\pm 0.27}$	2.00	$42.83_{\pm 0.49}$
Sequential	$0.23_{\pm 0.00}$	**0.00**	$48.97_{\pm 0.49}$	$99.65_{\pm 0.27}$	1.00	$44.73_{\pm 0.64}$	$99.66_{\pm 0.30}$	1.00	$41.27_{\pm 0.65}$	$100.00_{\pm 0.00}$	1.00	$42.57_{\pm 0.40}$
PCGrad	$0.23_{\pm 0.00}$	**0.00**	$48.97_{\pm 0.49}$	$74.08_{\pm 11.54}$	1.67	$44.90_{\pm 0.44}$	$91.04_{\pm 2.21}$	2.00	$45.20_{\pm 0.70}$	$96.75_{\pm 0.57}$	2.00	$42.67_{\pm 0.70}$
MGDA	$0.15_{\pm 0.27}$	**0.00**	$49.60_{\pm 0.61}$	$76.65_{\pm 3.26}$	2.00	$47.40_{\pm 1.08}$	$97.59_{\pm 1.10}$	4.33	$48.50_{\pm 1.42}$	$96.27_{\pm 0.83}$	5.00	$45.17_{\pm 0.85}$
FedOSD	$2.74_{\pm 0.68}$	**0.00**	$55.07_{\pm 0.21}$	$98.15_{\pm 0.44}$	1.00	$55.60_{\pm 0.10}$	$98.55_{\pm 0.55}$	1.00	$53.43_{\pm 0.25}$	$99.67_{\pm 0.12}$	1.00	$48.77_{\pm 0.75}$
FedCARE	$1.52_{\pm 0.13}$	**0.00**	$53.90_{\pm 0.78}$	$99.05_{\pm 0.44}$	1.00	$54.20_{\pm 0.61}$	$98.83_{\pm 0.78}$	1.00	$52.60_{\pm 0.56}$	$99.74_{\pm 0.11}$	1.00	$47.57_{\pm 0.40}$
RFE*	$8.23_{\pm 0.60}$	**0.00**	$55.57_{\pm 0.15}$	$98.44_{\pm 0.38}$	1.00	$55.70_{\pm 0.46}$	$98.14_{\pm 1.79}$	1.33	$54.07_{\pm 0.32}$	$99.74_{\pm 0.11}$	1.00	$49.27_{\pm 0.57}$
FedRazor	$0.08_{\pm 0.13}$	**0.00**	$48.47_{\pm 0.61}$	$99.72_{\pm 0.15}$	1.00	$43.97_{\pm 1.71}$	$99.82_{\pm 0.17}$	1.00	$41.60_{\pm 2.00}$	$99.81_{\pm 0.00}$	1.00	$37.57_{\pm 0.59}$
GDFA-U	$0.15_{\pm 0.13}$	**0.00**	$49.63_{\pm 0.55}$	$95.82_{\pm 0.27}$	3.00	$56.13_{\pm 0.35}$	$95.60_{\pm 0.89}$	5.00	$54.30_{\pm 0.30}$	$96.83_{\pm 0.70}$	7.00	$49.83_{\pm 0.80}$
NoT	$5.86_{\pm 2.54}$	**0.00**	$50.53_{\pm 0.50}$	$9.45_{\pm 5.52}$	1.00	$52.20_{\pm 0.56}$	$9.71_{\pm 2.35}$	1.33	$50.17_{\pm 0.23}$	$9.83_{\pm 1.47}$	**0.00**	$45.47_{\pm 0.42}$
SalUn	$0.08_{\pm 0.13}$	**0.00**	$48.70_{\pm 0.66}$	$4.32_{\pm 1.75}$	**0.00**	$46.70_{\pm 1.54}$	$29.47_{\pm 1.58}$	3.67	$47.00_{\pm 1.04}$	$50.16_{\pm 14.46}$	6.33	$45.60_{\pm 0.95}$
FAC (ours)	$2.37_{\pm 0.95}$	**0.00**	$50.47_{\pm 1.59}$	**$0.61_{\pm 0.13}$**	**0.00**	$53.57_{\pm 0.40}$	**$1.98_{\pm 1.38}$**	**0.00**	$50.83_{\pm 0.32}$	**$1.60_{\pm 0.42}$**	**0.00**	$47.17_{\pm 0.47}$

4 EXPERIMENTS

Benchmarks. We evaluate on two benchmarks and run every setting with three seeds. The source-scaling benchmark follows the CIFAR-10 (Krizhevsky, 2009) protocol of FedOSD (Pan et al., 2025): 30 clients under a Dirichlet partition (concentration 0.5) train a two-layer CNN, and the request deletes $K \in \{1, 3, 5, 7\}$ sources of three clients each, that is, 3 to 21 of the 30 clients; every source has implanted its own BadNets-style trigger–target pair (Gu et al., 2019). Since more sources also means more deleted clients, from 10% to 70%, Section 4.3 adds a control that fixes the request size. The cross-dataset benchmark has ten clients and deletes three sources of one client each, on CIFAR-10, CIFAR-100 and SVHN (Netzer et al., 2011), with NF-ResNet-18 and an ImageNet-pretrained ViT-S/16. With the same ten clients on NF-ResNet-18 and CIFAR-10, we also vary K over $\{1, 3, 5, 7\}$, deleting one to seven clients. Within a setting, every approximate method starts from the same poisoned model and receives the same 100 communication rounds, and Retrain repeats federated training without the deleted clients. The local correction adds one model-sized download per client and round; Appendix B.3 accounts for communication payloads and disagreement-state storage.

Scoring. Each source is tested on its own, as in Section 3.4: it survives when its attack success on its clients' triggered probes exceeds its Retrain threshold τ_i. We report the worst-source attack success rate (ASR), the number of surviving sources (Surv) and the retained clients' mean clean accuracy (R-Acc), by default at round 100, except for the round-50 diagnostic in Section 4.3; an R-Acc below 30% counts as a collapse, whose deletion is not ranked. Because Retrain levels differ across datasets and backbones, there we also report the normalized forgetting progress (NFP) of the worst source, $\min_i(\mathrm{ASR}_i^{\text{No Unl.}} - \mathrm{ASR}_i)/(\mathrm{ASR}_i^{\text{No Unl.}} - \mathrm{ASR}_i^{\text{Retrain}})$, which is 0 at No Unlearning and 100% at Retrain.

Baselines. We compare fifteen approximate methods, grouped by how they handle the forget objective. FT fine-tunes on the retained clients, and NegGrad+ (Kurmanji et al., 2023) and L1-Sparse (Jia et al., 2023) add a small forget term or a sparsity step to it; these three test whether training on retained data removes the backdoors. GA, AggSum and Sequential compare different ways of combining or scheduling forget updates. PCGrad (Yu et al., 2020) and MGDA (Sener & Koltun, 2018) additionally resolve conflicts among source updates through gradient projection and adaptive weighting, respectively. FedOSD (Pan et al., 2025), FedCARE (Li et al., 2026), RFE (Cheng et al., 2026a) and FedRazor (Hu et al., 2026) are federated unlearning methods that keep forgetting from conflicting with retained learning, and GDFA-U is our GDFA (Weng et al., 2026) reimplementa-

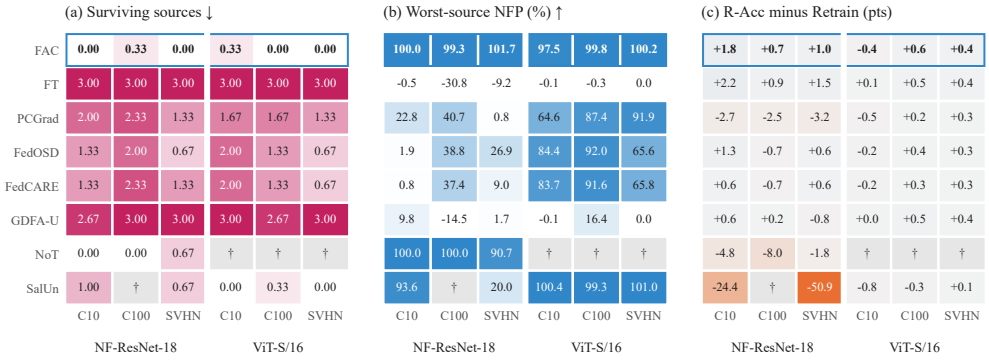

Figure 2: **Across datasets and backbones** ($K = 3$, ten clients; means over three seeds). The same seven representative baselines and FAC are shown in all six blocks. (a) Surviving sources; (b) worst-source NFP, with 100% marking the matched Retrain level; (c) R-Acc minus Retrain R-Acc. †All three seeds have R-Acc below 30%; these entries are not ranked. The NF-ResNet-18/CIFAR-100 FAC block has one surviving source in one of three seeds; FAC and SalUn differ in which ViT blocks have survivors. Appendix D reports the full matrix with sample standard deviations.

tion with unanimous sign agreement for merging forget-client updates. NoT (Khalil et al., 2025) negates weights instead of descending, and SalUn (Fan et al., 2024) confines forgetting to salient weights, as FAC's attack zone does. The cross-dataset benchmark compares all fifteen baselines on NF-ResNet-18 and seven representatives on ViT-S/16: FT, PCGrad, FedOSD, FedCARE, GDFA-U, NoT and SalUn. Baselines keep their published or project-default settings, while FAC's coefficients are set per protocol; staged baselines split the budget into 50 unlearning and 50 retain-only recovery rounds, and FAC omits recovery on the larger backbones. Appendix B lists all settings.

4.1 Source Scaling

Table 1 asks whether a method still removes every source as the request grows. With one source, thirteen of the sixteen methods leave no survivor; from three sources on, only FAC does so at every count, and the only other survivor-free entries are SalUn at $K = 3$ and NoT at $K = 7$. FedOSD, FedCARE, RFE, FedRazor, GA and Sequential usually remove all but one source, whose attack success remains near 100%. SalUn and MGDA leave more sources as K grows, and FT, NegGrad+, L1-Sparse and, for $K \geq 3$, GDFA-U leave every source. Among the survivor-free methods, FAC keeps the highest R-Acc for every $K \geq 3$; at $K = 1$ several keep more (55.57% for RFE against 50.47%). FAC is also survivor-free at every K on an IID partition (Appendix C). This holds for survivor thresholds from two to four pooled standard deviations or ten points above the Retrain mean, and every per-source ASR of FAC stays at or below that mean (Table A5).

4.2 Across Datasets and Backbones

Figure 2 fixes three sources and changes the dataset and backbone. On NF-ResNet-18, FAC leaves no surviving source on CIFAR-10 or SVHN. On CIFAR-100, one checker source survives in one seed (1.09% ASR), with a three-seed worst-source ASR of 0.66%. R-Acc remains within 1.82 points of Retrain across the three datasets. Among the fifteen baselines, only NoT is survivor-free on both CIFAR-10 and CIFAR-100, at 6.59 and 8.76 points lower R-Acc than FAC; every baseline leaves a source in at least one SVHN seed. The comparison changes on ViT-S/16, where FAC runs without the local proximal step ($\rho_{\mathrm{p}} = 0$). SalUn has lower worst-source ASR on CIFAR-10 and SVHN and is survivor-free on both, whereas FAC leaves one source in one CIFAR-10 seed. On CIFAR-100 the order reverses: FAC is survivor-free, while SalUn leaves one source in one seed. FAC retains 0.33–0.90 points more R-Acc than SalUn across the three datasets.

More sources on NF-ResNet-18. The $K = 3$ point reuses the cross-dataset configuration; the other counts use the settings in Table A1. On NF-ResNet-18 and CIFAR-10, FAC removes every source at $K = 1$, 3 and 5, keeping 6.5 to 11.4 points more R-Acc than NoT, which is also survivor-free at these counts; at $K = 7$, where seven of the ten clients (70%) are deleted, it leaves one or two sources (Surv = 1.33), while NoT leaves 0.33 at 6.1 points lower R-Acc (Table A9).

4.3 Component Ablations

On the CNN benchmark, we ask whether deletion needs the disagreement state across source counts and data partitions (IID and non-IID), and whether the spatial penalty depends on which coordinates it selects or only on how many. Figure 3 and Table 2(a) evaluate the complete 100-round procedure, including 50 recovery rounds. Table 2(b) tests score assignment at round 50, before recovery. Appendix E gives all ablations and sweeps.

Disagreement correction. Carrying disagreement in the server update alone does not recover FAC's deletion results. At $K = 5$, replacing the local target $z - u_k$ with the shared consensus z raises worst-source ASR from 1.98% to 14.64%; removing the proximal step raises it to 15.32%. Both retain the server-side correction. Merging the forget-side states at $K = 5$, by comparison, leaves no survivor and retains similar accuracy on both partitions. Setting $u_k \equiv 0$ instead removes the state from both updates and leaves 2.33 sources at 23.11% ASR. At non-IID $K = 7$, the no-state control leaves four sources at 58.20% ASR, against FAC's 1.60%.

Across both partitions, FAC and the three controls in Figure 3 leave no survivor at $K = 1$; the shared-target, no-proximal-step and no-state controls leave survivors at every larger tested K, whereas FAC leaves none.

The distinction persists on an IID control with the deleted roster fixed at 15 clients: the shared-target and no-proximal variants leave sources when that roster carries three or five backdoors, while FAC removes them all, at 1.10–3.11 points lower R-Acc.

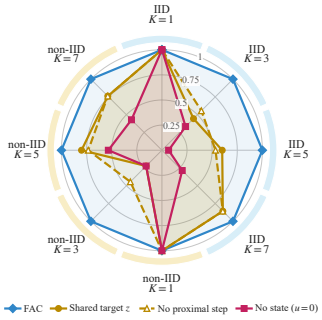

Figure 3: **State variants across source counts.** Mean $1 - \mathrm{Surv}/K$ after 100 rounds, over three seeds.

Other spatial choices. In the complete 100-round procedure, all four tested nominal zone fractions from 0.01 to 0.50 leave no survivor with the contrastive score, while R-Acc decreases by 3.20 points across that range. The regularizer weight matters more: on non-IID data, $\lambda \leq 10$ leaves survivors, whereas $\lambda \in \{50, 100, 200\}$ leaves none. EMA accumulation also leaves no survivor. Forget-only saliency and a forget/retain ratio collapse three and two seeds, respectively. Expanding the zone to all coordinates lowers R-Acc to 15.40%; this control also projects every parameter to zero during recovery.

Score assignment to parameters. We replay each reference FAC run's sensitivity scores with a fixed permutation within each weight or bias tensor, preserving its score distribution and selected-coordinate count at every matching round. Table 2(b) evaluates the returned model at round 50, before recovery.

Worst-source ASR rises from 0.92% to 94.13% under permutation. Each permuted run retains checker and diagonal, both above 92% ASR; the original assignment leaves no survivor and has higher R-Acc (49.85% versus 48.60%). Preserving score distributions and zone sizes alone therefore does not reproduce the original assignment's deletion outcome in this setting.

Table 2: **Component ablations** (CNN, non-IID, $K = 5$; format as in Table 1). †At least one seed collapses.

Variant	ASR↓	Surv↓	R-Acc↑
(a) Complete procedure, round 100, with recovery			
FAC	$1.98_{\pm 1.38}$	0.00	$50.83_{\pm 0.32}$
Merged forget state	$2.14_{\pm 1.49}$	0.00	$50.77_{\pm 0.29}$
Shared target z	$14.64_{\pm 1.06}$	1.00	$52.93_{\pm 0.40}$
No proximal step	$15.32_{\pm 1.65}$	1.33	$52.93_{\pm 0.35}$
No state ($u = 0$)	$23.11_{\pm 3.02}$	2.33	$53.23_{\pm 0.32}$
All-coordinate zone†	$100.00_{\pm 0.00}$	1.00	$15.40_{\pm 0.00}$
EMA, $\beta = 0.9$	$1.45_{\pm 0.47}$	0.00	$51.03_{\pm 0.29}$
Forget-only saliency†	$48.59_{\pm 11.16}$	2.67	$17.47_{\pm 4.91}$
Forget/retain ratio†	$37.16_{\pm 17.85}$	3.33	$28.87_{\pm 2.80}$
(b) Scores permuted in each tensor, round 50			
Original FAC	$0.92_{\pm 0.60}$	0.00	$49.85_{\pm 0.46}$
Permuted in tensor	$94.13_{\pm 0.60}$	2.00	$48.60_{\pm 0.33}$

4.4 Source-Wise Analysis

We compare mean-source and worst-source ASR during unlearning, then relate source-wise coverage γ_i^t (Section 3.4) to endpoint survival.

Averages hide survivors. A falling mean ASR can conceal a source whose attack success remains high. By round 50, GA, Sequential, FedOSD, FedCARE and FedRazor reduce mean-source ASR to

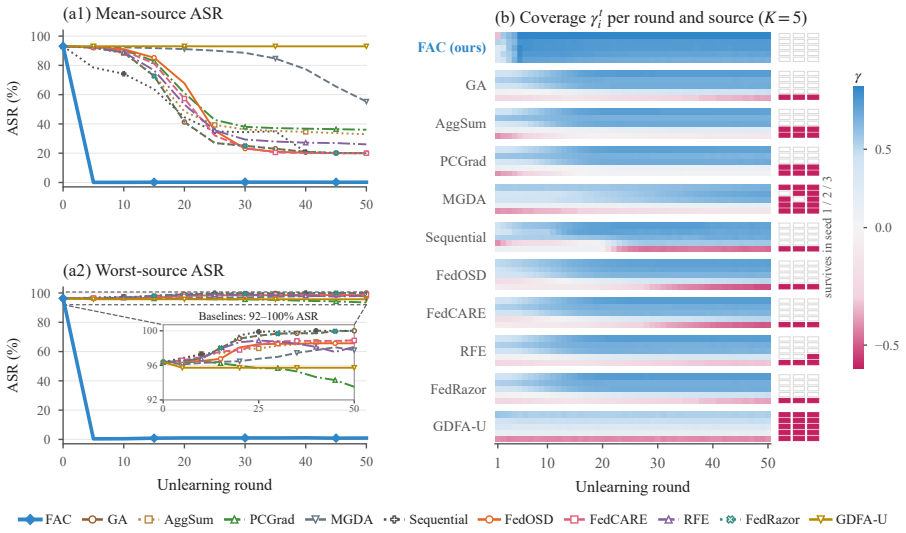

Figure 4: **ASR trajectories and source-wise coverage.** CNN, CIFAR-10, non-IID, $K = 5$; FAC and ten baselines that descend a forget objective. (a1) Mean-source and (a2) worst-source ASR over the first 50 rounds, measured every five rounds. Curves average three seeds; the worst-source ASR is computed within each seed before averaging. The inset enlarges the 92–100% baseline region. (b) Coverage γ_i^t over the same unlearning phase. Within each run, sources are ranked by mean coverage, lowest at the bottom of each method block; each row averages the same rank over three seeds. Filled squares identify surviving sources in seeds 1–3 after the complete 100-round procedure. All sixteen methods and source counts: Appendix F.

about 20%, while their worst-source ASR remains at 98.55–100% (Figure 4a). FAC reduces both, reaching 0.18% mean-source and 0.92% worst-source ASR. The gap also appears across source counts: among the 144 endpoint runs with $K \in \{3, 5, 7\}$, 83 (58%) have a worst-minus-mean gap of at least 20 percentage points; 31 of the 48 runs with mean ASR below 20% still retain a source (Appendix F, Figure A3). Thus low averaged attack success does not establish that every source has been removed.

Coverage and source survival. Figure 4b records coverage over the 50 unlearning rounds and marks survival after the complete 100-round procedure. In each of the 30 descent-baseline runs, exactly one source has non-positive mean coverage, and all 30 such sources survive. By comparison, 29 of the 120 positively covered source instances survive. FAC has positive mean coverage for all 15 source instances and leaves none.

Coverage is diagnostic rather than a universal deletion criterion. SalUn has positive mean coverage for all 15 sources yet leaves 11, whereas NoT deletes a source with non-positive mean coverage (Appendix F).

5 Conclusion

We study multi-source federated unlearning and show that a shared update direction can leave conflicting sources uncovered. FAC combines disagreement-corrected consensus with a sensitivity-guided penalty that limits changes to parameters used by retained clients. It removes every source at all tested source counts on the CNN benchmark; on larger backbones, the outcome depends on the backbone, dataset and request size. Clean continuation can increase source-wise ASR after unlearning (Appendix G). Source-wise evaluation reveals surviving backdoors that averaged attack success conceals, and coverage analysis helps diagnose these failures in the tested descent baselines. These results support evaluating each source separately and accounting for conflicts among forget objectives, alongside the trade-off between deletion and retained utility. Extending source-wise evaluation to other trigger types, to departing clients that do not cooperate, and to larger federations is left to future work.

AI USE STATEMENT

Generative AI tools were used to assist with experiment orchestration, software-code editing, artifact organization, result-interpretation support, manuscript structuring, and language editing. AI-assisted outputs were checked against source code, run-level records, and regenerated aggregate results. The authors take responsibility for the final text, claims, code, and artifacts.

ETHICS STATEMENT

The experiments use standard vision benchmarks and involve no human subjects or newly collected personal data. Backdoor attacks are evaluated only to expose and verify unlearning failures. Released artifacts should carry clear defensive-use context and omit operational guidance for attacking deployed systems.

REPRODUCIBILITY STATEMENT

Appendix A gives the complete algorithm and the proof of Proposition 1, and Appendix B gives the FAC coefficients of every protocol, the probe construction, the metric definitions and the run records. The lightweight anonymous artifact records the frozen configurations, three-seed structured aggregates, aggregation scripts, audits, and generated tables and figures; large datasets, checkpoints, and path-bearing raw logs remain outside the upload. The anonymous supplement describes the boundary between lightweight evidence reproduction and a full GPU rerun.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=B7v4QMR6Z9w.

Sadia Asif and Mohammad Mohammadi Amiri. Ofmu: Optimization-driven framework for machine unlearning. In *International Conference on Learning Representations*, volume 2026, pp. 88811–88836, 2026.

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, pp. 2938–2948. PMLR, 2020.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Jingpu Cheng, Ping Liu, Qianxiao Li, and Chi Zhang. Machine unlearning under retain–forget entanglement. In *International Conference on Learning Representations (ICLR)*, 2026a. arXiv:2603.26569.

Xinwen Cheng, Zhehao Huang, Wenxing Zhou, Zhengbao He, Ruikai Yang, Yingwen Wu, and Xiaolin Huang. Remaining-data-free machine unlearning by suppressing sample contribution. In *International Conference on Learning Representations*, volume 2026, pp. 27521–27549, 2026b.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, volume 2024, pp. 53643–53673, 2024.

Xiangshan Gao, Xingjun Ma, Jingyi Wang, Youcheng Sun, Bo Li, Shouling Ji, Peng Cheng, and Jiming Chen. VeriFi: Towards verifiable federated unlearning. *IEEE Transactions on Dependable and Secure Computing*, 21(6):5720–5736, 2024.

Hanlin Gu, Hong Xi Tae, Lixin Fan, and Chee Seng Chan. Towards privacy-guaranteed label unlearning in vertical federated learning: few-shot forgetting without disclosure. In *International Conference on Learning Representations*, volume 2026, pp. 80775–80801, 2026.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.

Yanxin Hu, Xiaoman Liu, Yan Huang, Junjie Pang, Chao Cheng, and Gang Liu. Fedrazor: Two-stage federated unlearning via representation divergence and gradient conflict trimming. *Information*, 17(2):146, 2026.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. *Advances in neural information processing systems*, 36:51584–51605, 2023.

Weipeng Jiang, Juan Zhai, Shiqing Ma, Ziyan Lei, Xiaofei Xie, Yige Wang, and Chao Shen. From chaos to clarity: A knowledge graph-driven audit dataset generation framework for llm unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31337–31345, 2026.

Yasser H. Khalil, Leo Brunswic, Soufiane Lamghari, Xu Li, Mahdi Beitollahi, and Xi Chen. Not: Federated unlearning via weight negation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25759–25769, 2025.

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.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 2, pp. 429–450, 2020.

Yue Li, Mingmin Chu, Xilei Yang, Da Xiao, Ziqi Xu, Wei Shao, Qipeng Song, and Hui Li. FedCARE: Federated unlearning with conflict-aware projection and relearning-resistant recovery. *arXiv preprint arXiv:2601.22589*, 2026.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 18878–18890, 2021a.

Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. FedEraser: Enabling efficient client-level data removal from federated learning models. In *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQoS)*, pp. 1–10, 2021b. doi: 10.1109/IWQOS52092.2021.9521274.

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

Nima Naderloui, Shenao Yan, Binghui Wang, Jie Fu, Wendy Hui Wang, Weiran Liu, and Yuan Hong. Rectifying privacy and efficacy measurements in machine unlearning: A new inference attack perspective. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 5545–5564. USENIX Association, 2025.

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://research.google/pubs/reading-digits-in-natural-images-with-unsupervised-feature-learning/.

Minh-Duong Nguyen, Senura Wanasekara, Le-Tuan Nguyen, Quoc-Viet Pham, Ken-Tye Yong, Nguyen H. Tran, and Dung D. Le. Computation and communication efficient federated unlearning via on-server gradient conflict mitigation and expression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3347–3357, 2026.

Zibin Pan, Zhichao Wang, Chi Li, Kaiyan Zheng, Boqi Wang, Xiaoying Tang, and Junhua Zhao. Federated unlearning with gradient descent and conflict mitigation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19):19804–19812, 2025. doi: 10.1609/aaai.v39i19.34181.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Xinyi Sheng, Wei Bao, Hequn Wang, Yuqin Liu, and Sen Fu. Retaliatory attacks against federated unlearning via data leakage. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (30):25321–25329, 2026. doi: 10.1609/aaai.v40i30.39725.

Haoran Tang and Rajiv Khanna. Sharpness-aware machine unlearning. In *International Conference on Learning Representations*, volume 2026, pp. 79941–79985, 2026.

Cheng-Long Wang, Qi Li, Zihang Xiang, Yinzhi Cao, and Di Wang. Towards lifecycle unlearning commitment management: Measuring sample-level unlearning completeness. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 6481–6500, 2025.

Han Wang, Siddartha Marella, and James Anderson. FedADMM: A federated primal–dual algorithm allowing partial participation. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 287–294, 2022.

Siyuan Wen, Meng Zhang, Yang Yang, and Ningning Ding. FedShard: Federated unlearning with efficiency fairness and performance fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(32):26841–26848, 2026. doi: 10.1609/aaai.v40i32.39895.

Jiasi Weng, Shenglong Yao, Yuefeng Du, Junjie Huang, Jian Weng, and Cong Wang. Proof of unlearning: Definitions and instantiation. *IEEE Transactions on Information Forensics and Security*, 19:3309–3323, 2024. doi: 10.1109/TIFS.2024.3358993.

Xiuting Weng, Ruizhi Pu, Yuanhang Yao, Kun Yue, Zhiwen Tang, and Lixing Yu. GDFA: Geometry-driven federated unlearning with directional task vector alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10346–10356, 2026.

Wenhan Wu, Zhili He, Huanghuang Liang, Yili Gong, Jiawei Jiang, Chuang Hu, and Dazhao Cheng. ReMisVFU: Vertical federated unlearning via representation misdirection for intermediate output feature. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(32):26983–26991, 2026. doi: 10.1609/aaai.v40i32.39911.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Lefeng Zhang, Tianqing Zhu, Haibin Zhang, Ping Xiong, and Wanlei Zhou. FedRecovery: Differentially private machine unlearning for federated learning frameworks. *IEEE Transactions on Information Forensics and Security*, 18:4732–4746, 2023. doi: 10.1109/TIFS.2023.3297905.

Zhengyi Zhong, Weidong Bao, Ji Wang, Shuai Zhang, Jingxuan Zhou, Lingjuan Lyu, and Wei Yang Bryan Lim. Unlearning through knowledge overwriting: Reversible federated unlearning via selective sparse adapter. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30661–30670, 2025.

Shiji Zhou, Tianbai Yu, Zhi Zhang, Heng Chang, Xiao Zhou, Dong Wu, and Han Zhao. Efficient utility-preserving machine unlearning with implicit gradient surgery. *Advances in Neural Information Processing Systems*, 38:4531–4569, 2026.

Appendix overview. Appendix A gives the algorithm, derivations and proof; Appendix B gives protocols, baselines and metrics; Appendices C–F follow Sections 4.1–4.4; Appendix G reports recovery schedules and clean re-entry; and Appendix H discusses further related work.

A Algorithm and Derivations

A.1 Complete algorithm

Algorithm 1 Federated Asymmetric Consensus (FAC)

Require: Pre-unlearning model $\bar{\theta}$; client sets $\mathcal{F}, \mathcal{R}$; objectives $\{f_k\}_{k=1}^N$; total rounds T; unlearning rounds $1 \le T_{\mathrm{u}} \le T$; local steps E; $\eta_t, \eta_{\mathrm{r}}, \rho_{\mathrm{p}}, \rho, \lambda, \alpha, \omega_{\mathrm{f}}$; client sizes $\{n_k\}$; observer groups $\{\mathcal{F}_i\}$
Ensure: Consensus model z^T and audit trace $\mathcal{Q}$

$z^0 \leftarrow \bar{\theta}$; $u_k^0 \leftarrow 0$ for all k; $M_{\mathcal{F}}^{-1}, M_{\mathcal{R}}^{-1} \leftarrow 0$; $\mathcal{Q} \leftarrow \varnothing$
for $t = 0, \dots, T_{\mathrm{u}} - 1$ **do**
  Broadcast z^t to the participating clients
  for all $k \in \mathcal{F} \cup \mathcal{R}$ in parallel **do**
    $v_k^t \leftarrow z^t - u_k^t$; $\theta_{k,0} \leftarrow z^t$
    for $e = 0, \dots, E-1$ **do**
      $\widehat{\theta}_{k,e+1} \leftarrow \theta_{k,e} - \eta_t \nabla f_k(\theta_{k,e})$
      if $t = 0$ **then**
        $\theta_{k,e+1} \leftarrow \widehat{\theta}_{k,e+1}$ {warm start}
      else
        $\theta_{k,e+1} \leftarrow (\widehat{\theta}_{k,e+1} + \eta_t \rho_{\mathrm{p}} v_k^t)/(1 + \eta_t \rho_{\mathrm{p}})$
      end if
    end for
    Upload $\theta_k^{t+1} \leftarrow \theta_{k,E}$ and the local loss trace ℓ_k^t
  end for
  Compute q_k^t by Eq. (6); average over $\mathcal{F}$ and $\mathcal{R}$ separately
  Update $M_{\mathcal{F}}^t, M_{\mathcal{R}}^t, s^t$ by (7); apply the positive-score selection rule and form $\mathcal{A}^t$, $\widetilde{\theta}^t$, a^t, and b^t by (8)
  Form ω_k (ω_{f} on $\mathcal{F}$, 1 on $\mathcal{R}$); update z^{t+1} by Eq. (9)
  Update every u_k^{t+1} by (5)
  Append $\{\ell_k^t\}_{k\in\mathcal{F}}$, $\{\|\theta_k^{t+1} - z^{t+1}\|_2\}_{k=1}^N$, $\rho\|z^{t+1} - z^t\|_2$, and $\mathcal{A}^t$ to $\mathcal{Q}$
  Append source Gram, coverage, and cancellation from Eqs. (6)–(10) to $\mathcal{Q}$
end for
$\mathcal{M} \leftarrow \mathcal{A}^{T_{\mathrm{u}}-1}$
Optional internal recovery when $T_{\mathrm{u}} < T$:
for $t = T_{\mathrm{u}}, \dots, T-1$ **do**
  Remove $\mathcal{F}$; broadcast z^t to retain clients
  for all $k \in \mathcal{R}$ in parallel **do**
    Starting from z^t, take E SGD steps on f_k with learning rate η_{r}
    Upload the recovered model θ_k^{t+1}
  end for
  $z^{t+1} \leftarrow \big(\sum_{k\in\mathcal{R}} n_k \theta_k^{t+1}\big)/\big(\sum_{k\in\mathcal{R}} n_k\big)$
  $[z^{t+1}]_{\mathcal{M}} \leftarrow 0$ {hard zero projection}
  Append the recovery loss and $\mathcal{M}$ to $\mathcal{Q}$
end for
return $z^T, \mathcal{Q}$

The observer groups are used only for recording. When $T_{\mathrm{u}} = T$, the recovery loop is empty. The source Gram computation adds $O(K^2P)$ arithmetic per recorded round; sensitivity construction and consensus use $O(NP)$ arithmetic. The client-indexed states require $O(NP)$ working memory, excluding the stored audit history.

A.2 Derivations and Numerical Conventions

Local proximal correction. For a fixed target $v = z^t - u_k^t$ and a local SGD output $\widehat{\theta}$, the quadratic proximal correction solves

$$\theta^+ = \arg\min_x \left\{ \frac{1}{2\eta_t}\|x - \widehat{\theta}\|_2^2 + \frac{\rho_{\mathrm{p}}}{2}\|x - v\|_2^2 \right\}. \tag{11}$$

Its first-order condition is $(x-\widehat{\theta})/\eta_t + \rho_{\mathrm{p}}(x-v)=0$. Rearranging gives $\theta^+ = (\widehat{\theta}+\eta_t\rho_{\mathrm{p}}v)/(1+\eta_t\rho_{\mathrm{p}})$, as in Eq. (3). This is the exact proximal map of the quadratic term after the SGD step, not an exact solution of the entire local nonconvex objective. The first round omits this correction; when $\rho_{\mathrm{p}}=0$, it reduces to the uncorrected SGD output.

Attack-zone selection and numerical conventions. Let P be the number of parameters and let $\mathcal{I}^t=\{j: s_j^t>0\}$. If $\mathcal{I}^t\neq\varnothing$, set

$$\begin{aligned}&\kappa_t=\min\{\max(1,\lfloor\alpha P\rfloor),|\mathcal{I}^t|\}, \qquad \xi_t=\text{the }\kappa_t\text{-th largest entry of } s^t,\\ &\mathcal{A}^t=\{j\in\mathcal{I}^t: s_j^t\geq\xi_t\}.\end{aligned} \tag{12}$$

Here $\alpha\in(0,1]$ is nominal. The rank cap keeps the threshold positive, and ties at the threshold are retained. Thus αP is not a strict cardinality bound for all inputs. If $\mathcal{I}^t$ is empty, set $\mathcal{A}^t=\varnothing$ and keep the scores unchanged. The realized zone fraction is recorded per round.

For the coverage observer, source sums are normalized by their Euclidean norm plus 10^{-12}. When the norm of the sum of normalized source directions is at most 10^{-12}, per-source coverage is logged as zero. These are numerical recording conventions; the unit-vector identities below are stated for nonzero vectors with exact normalization.

Coordinate-wise consensus solution. Fix the round's sensitivity weights and masked anchor. For coordinate j, discarding terms independent of x reduces the server objective to

$$\begin{aligned}\phi_j(x)&=\frac{\rho}{2}\sum_k \omega_k\big(x-\theta_{k,j}^{t+1}-u_{k,j}^t\big)^2+\lambda a_j^t|x|+\lambda b_j^t(x-\widetilde{\theta}_j^t)^2\\&=\frac{D_j^t}{2}x^2-B_j^t x+\lambda a_j^t|x|+C_j,\end{aligned} \tag{13}$$

where $D_j^t=\rho W+2\lambda b_j^t>0$, $B_j^t=\rho h_j^t+2\lambda b_j^t\widetilde{\theta}_j^t$, and C_j does not depend on x. Completing the square gives

$$\phi_j(x)=\frac{D_j^t}{2}\left(x-\frac{B_j^t}{D_j^t}\right)^2+\lambda a_j^t|x|+C_j'. \tag{14}$$

The optimality condition is $0\in D_j^t x-B_j^t+\lambda a_j^t\partial|x|$. Consequently,

$$x^*=\begin{cases}(B_j^t-\lambda a_j^t)/D_j^t, & B_j^t>\lambda a_j^t,\\ 0, & |B_j^t|\leq\lambda a_j^t,\\ (B_j^t+\lambda a_j^t)/D_j^t, & B_j^t<-\lambda a_j^t,\end{cases} \tag{15}$$

which is exactly the soft-threshold update in Eq. (9). This derivation establishes the unique solution of each fixed-map server subproblem. The changing sensitivity map and independent local coupling coefficient do not turn the full training trajectory into exact scaled ADMM for a single static objective.

A.3 Proof of Proposition 1

The proposition is stated in Section 3.1 for unit vectors $\hat{g}_1,\dots,\hat{g}_K$ with $G_{ij}=\langle\hat{g}_i,\hat{g}_j\rangle$, $d=\sum_i\hat{g}_i$ and $c=1-\|d\|^2/K$. It is a local geometric statement and does not assert global convergence, permanent deletion, or a retained-loss lower bound.

Proof. For part (i),

$$\langle\hat{d},\hat{g}_i\rangle=\frac{1}{\|d\|}\left\langle\sum_{j=1}^K\hat{g}_j,\hat{g}_i\right\rangle=\frac{\sum_{j=1}^K G_{ij}}{\|d\|}=\frac{(G\mathbf{1})_i}{\sqrt{K(1-c)}}, \tag{16}$$

where the last equality uses $\|d\|^2=K(1-c)$.

For part (ii), summing the K coverage constraints gives

$$\langle v, d\rangle = \sum_{i=1}^{K} \langle v, \hat{g}_i\rangle \geq K\kappa. \tag{17}$$

If $d = 0$, the left-hand side is zero, contradicting $K\kappa > 0$. Otherwise, Cauchy–Schwarz yields

$$K\kappa \leq \langle v, d\rangle \leq \|v\|\,\|d\|, \tag{18}$$

and substituting $\|d\| = \sqrt{K(1-c)}$ proves the result. □

Interpretation and scope. Part (i) concerns one direction at one model state. A non-positive value rules out a first-order decrease from that particular step, but gradients may rotate after an update. Part (ii) concerns an unnormalized step that simultaneously satisfies explicit per-source linear coverage constraints. Its norm lower bound is purely geometric. Converting that bound into a retained-utility guarantee would require an additional signed-alignment assumption between v and the retain gradients; FAC does not make that assumption, and the manuscript treats retained utility as an empirical outcome.

B Experimental Setup

This appendix gives the settings behind Section 4 that the main text omits.

B.1 Protocols and Training

Table A1 lists the settings of each protocol. Each participating client runs one local epoch per round, without subsampling the active client roster of that method. The staged baselines spend 50 rounds unlearning and 50 in retain-only recovery; NoT negates weights once and then trains on the retained clients for all 100 rounds; FAC follows T_{u} in Table A1 and omits recovery on the larger backbones (Appendix G). FAC's coefficients in Table A1 were fixed per protocol in development runs on each benchmark (per K for the NF-ResNet-18 source-count study; e.g., $K = 5$ uses $\eta_0 = 0.02$ after 0.03 collapsed one seed). Appendix E reports their sensitivity on the CNN benchmark only; we do not provide a retrain-free rule for selecting them. FAC's CNN coefficients and the baseline budgets were fixed in a written protocol before the formal CNN runs, and the non-IID runs of Table 1 reuse them unchanged. Baseline configurations follow the published or project-default settings listed below.

Table A1: **Protocol settings.** Clients, partition, batch size, the round budget T and the base/Retrain rate apply to every method; T_{u} (unlearning rounds), the FAC rate, λ, ω_{f}, α and ρ_{p} are FAC's. All protocols use the contrastive score of Eq. (7) with running-maximum accumulation. *Scored at round 100; the learning rate decays per round, so this equals a 100-round run.

Protocol	Clients	Per source	Dirichlet	Batch	T	T_{u}	η_0 Base/Retrain	η_0 FAC	λ	ω_{f}	α	ρ_{p}
CNN, scaling and ablations	30	3	0.5 (IID: 100)	200	100	50	0.02	0.01	100	3	0.05	0.05
NF-ResNet-18, cross-dataset ($K = 3$)	10	1	0.5	128	100	100	0.02	0.01	1	1	0.05	0.05
ViT-S/16, cross-dataset	10	1	0.5	128	100*	100	0.003	0.001	1	3	0.05	0
NF-ResNet-18, $K = 1, 7$	10	1	0.5	128	100	100	0.02	0.03	1	5	0.05	0.05
NF-ResNet-18, $K = 5$	10	1	0.5	128	100	100	0.02	0.02	1	5	0.05	0.05

Optimization and learning rates. The FAC configurations in Table A1 use $\rho = 10$, $\delta = 0.1$ and $\varepsilon = 10^{-12}$, with SGD, zero momentum and no weight decay. The learning rate at zero-based round t is $\eta_t = \eta_0\, 0.999^t$. For staged methods, including FAC on the CNN, the recovery phase replaces η_0 with 10^{-6} and retains the global round index. Table A1 lists the initial rates. Coefficient ablations change the named coefficient only.

The NF-ResNet-18 source-count study reuses the cross-dataset $K = 3$ runs, including $\omega_{\mathrm{f}} = 1$; its $K = 1, 5, 7$ runs use $\omega_{\mathrm{f}} = 5$. The default initial unlearning rate for the staged baselines is 4×10^{-4}. Table A6 reruns four baselines at higher rates. NoT uses 0.01 on CNN and NF-ResNet-18, and 0.001 on ViT; its retain-only training uses the corresponding recovery-rate switch where specified by the staged protocol.

Backdoor construction and model training. Poisoned-base training applies each source's trigger and target label to 50% of its clients' local training examples. The CNN source order is square$\rightarrow 0$, cross$\rightarrow 1$, checker$\rightarrow 2$, diagonal$\rightarrow 3$, anti-diagonal$\rightarrow 4$, corners$\rightarrow 5$, and border$\rightarrow 6$, with the first K pairs assigned to three clients each. The cross-dataset studies use cross$\rightarrow 1$, checker$\rightarrow 2$, and diagonal$\rightarrow 0$, one client per source. The NF-ResNet-18 source-count study extends that order with square$\rightarrow 4$, anti-diagonal$\rightarrow 5$, corners$\rightarrow 6$, and border$\rightarrow 3$, using the first K pairs. Named 3×3 patterns are expanded by a factor of three and placed at the bottom right of the input image before any ViT resizing.

Base and Retrain use 300 FedAvg rounds, one local epoch and full participation of the eligible clients; Retrain excludes the deleted clients. CNN and NF-ResNet-18 start from random initialization, while ViT-S/16 uses ImageNet initialization and resizes inputs to 112×112. During the unlearning task, each forget client's local examples receive its assigned trigger and target, while retain clients use clean data. The three run seeds are 1–3. The non-IID CNN and NF-ResNet-18 partitions use seed 20260826; the IID CNN and ViT CIFAR partitions reuse their stored client pools, while ViT SVHN uses partition seed 20260826. The configuration summary in the accompanying package, `configs/protocols.json`, records these distinctions.

B.2 Baselines

Table A2 lists how each baseline runs under the shared federated protocol; every staged baseline recovers with retain-only FedAvg. GDFA-U denotes our GDFA reimplementation with a unanimous-sign merge. Its sign-agreement merge keeps about 80% of the coordinates in the first unlearning round at $K = 1$, under 3% at $K = 3$ and about 0.1% at $K \geq 5$. With the original paper's merge rule, which elects each coordinate's sign from the summed updates and averages the entries that share it, the separate CNN non-IID reproduction leaves exactly one source in every seed for $K \geq 3$ (99.72–99.89% worst-source ASR, 43.43–44.97% R-Acc) and none at $K = 1$.

Table A2: Baselines under the shared federated protocol

Method	Unlearning step under the shared protocol	Coefficients
FT	FedAvg on the retained clients only	—
NegGrad+	Server combines the mean retain-client update and the mean forget-client update on the suppression loss of Eq. (1)	weights 0.95 / 0.05
L1-Sparse	Retain-only FedAvg; the global model is soft-thresholded once per round	$\lambda = 5 \times 10^{-4}$
GA	Server steps along the mean forget-client update	—
AggSum	Server steps along the sum of the per-source forget updates, each normalized to unit length	—
Sequential	One source per phase of the unlearning budget, in a fixed order	—
PCGrad	Gradient surgery among the per-source forget updates, then their sum	—
MGDA	Min-norm convex combination of the per-source forget updates	—
FedOSD	Original federated algorithm	defaults
FedCARE	Mean forget update projected against the retained updates; relearning-resistant recovery; its model-inversion generator is replaced by the retained clients' data	defaults
RFE*	Augmented-Lagrangian stage, then W2-guided projection, on the mean forget and retain updates	$\alpha = 0.5$
FedRazor	Reimplemented from its description: normalized mean forget step, then recovery with the conflicting components removed	—
GDFA-U	Five retain-only FedAvg rounds, then the forget clients' updates, merged on coordinates where all signs agree, are subtracted	—
NoT	Negates the selected layers once, then retain-only FedAvg	—

Method	Unlearning step under the shared protocol	Coefficients
SalUn	Saliency mask from the mean forget update; forget clients fine-tune on random labels inside the mask; unlike the original joint objective, the retain term enters through the recovery phase	—

Baselines under the shared federated protocol, with their published default coefficients unless listed. *Project adaptation (Table 1).

B.3 Metrics, Coverage Observer and Communication

Probes and survivor thresholds. For each deleted client, the held-out examples whose clean label differs from its source's target are split into two disjoint halves, stratified by clean label with a seed fixed by the run seed and the client index; both receive the trigger and the target label. ASR is measured on the first half, averaged over a source's clients; the second half serves only the deletion residual. With x_{im} the ASR of source i under Retrain seed m, $\tau_i = \bar{x}_i + 3\sigma$, where $\bar{x}_i$ is the mean over seeds and $\sigma^2 = \sum_i \sum_m (x_{im} - \bar{x}_i)^2/(2K)$. All methods in a setting share these thresholds. Worst-source NFP is the minimum normalized progress across sources, computed per seed against that seed's No Unlearning and Retrain runs and then averaged over seeds.

Auxiliary diagnostics (Appendix F). The deletion residual is the forget loss of Eq. (1) on the second probe half, averaged over clients and sources. D2R-KL averages $\mathrm{KL}(p_{\text{unlearned}} \,\|\, p_{\text{Retrain}})$ over the retained clients' held-out examples. The membership-inference AUC scores examples by negative cross-entropy, with the deleted clients' clean training examples as members and the retained clients' held-out examples as non-members.

Coverage observer. The observer of Section 3.4 adds one Gram product per round. It was attached to FAC, FedOSD, GDFA-U and NoT at $K \in \{3, 5, 7\}$ and to every other Table 1 method at $K = 5$; all 75 probed runs reproduce the endpoint digests of their Table 1 runs. It needs the source membership of the deleted clients, which our experiments take from annotations that the optimization never sees.

Communication and state storage. Let P be the number of communicated model parameters and b their storage size in bytes. We count dense parameter payloads per participating client per round, excluding scalar metadata, non-parameter buffers and transport overhead. A model download and a local-model upload together require $2bP$ bytes. When the local correction is active, a client also needs its proximal target $z^t - u_k^t$. With the disagreement state kept on the server and this target sent separately, the payload is $3bP$ bytes: $2bP$ downstream and bP upstream. When the correction is disabled, including our ViT configuration, the target is unnecessary and the payload is $2bP$. The sensitivity map and selected zone are computed from returned models and require no additional client upload.

These are analytical payload counts for the stated state placement, not measured network traffic: the experiments simulate clients in one process. The client-indexed disagreement states require bNP bytes for N registered clients, in addition to shared model vectors, sensitivity accumulators and temporary tensors. This state-storage count is not a measurement of peak device memory. We do not infer distributed latency or wall-clock speedup from the payload counts.

C Source Scaling

This section supports Section 4.1. Table A3 repeats Table 1 on an IID partition (Dirichlet concentration 100); FAC again leaves no survivor at any K, with worst-source ASR between 0.73% and 2.05%.

Worst-source progress. Table A4 gives the worst-source NFP of both partitions. From $K = 5$ on, every baseline other than NoT and SalUn is at most 10%, so its least-removed source keeps about its No Unlearning attack success; at $K = 3$, AggSum, PCGrad and MGDA still reach 21–72%. SalUn

Table A3: **Source-level forgetting on the IID partition** (CNN, CIFAR-10, Dirichlet concentration 100). Entries are $a_{\pm b}$ with mean a and sample SD b over three seeds; worst-source ASR and R-Acc in %; Surv counts sources above their Retrain threshold. No Unlearning and Retrain are anchors, not ranked. Blue marks the unique lowest ASR per column and every survivor-free Surv entry. *Project adaptation.

Method	$K=1$			$K=3$			$K=5$			$K=7$		
	ASR↓	Surv↓	R-Acc↑	ASR↓	Surv↓	R-Acc↑	ASR↓	Surv↓	R-Acc↑	ASR↓	Surv↓	R-Acc↑
No Unlearning	$91.81_{\pm1.50}$	$1.00_{\pm0.00}$	$59.78_{\pm0.35}$	$93.84_{\pm1.58}$	$3.00_{\pm0.00}$	$58.34_{\pm0.30}$	$94.27_{\pm0.81}$	$5.00_{\pm0.00}$	$56.73_{\pm0.61}$	$94.98_{\pm1.12}$	$7.00_{\pm0.00}$	$53.82_{\pm0.85}$
Retrain	$6.33_{\pm1.86}$	$0.00_{\pm0.00}$	$60.15_{\pm0.32}$	$6.78_{\pm1.22}$	$0.00_{\pm0.00}$	$59.92_{\pm0.45}$	$6.79_{\pm0.80}$	$0.00_{\pm0.00}$	$59.57_{\pm0.54}$	$9.60_{\pm1.77}$	$0.00_{\pm0.00}$	$59.24_{\pm1.34}$
FT	$90.04_{\pm1.28}$	$1.00_{\pm0.00}$	$59.76_{\pm0.35}$	$93.47_{\pm1.30}$	$3.00_{\pm0.00}$	$58.49_{\pm0.28}$	$93.32_{\pm0.46}$	$5.00_{\pm0.00}$	$56.71_{\pm0.47}$	$94.36_{\pm1.33}$	$7.00_{\pm0.00}$	$54.31_{\pm0.44}$
NegGrad+	$87.23_{\pm1.71}$	$1.00_{\pm0.00}$	$59.74_{\pm0.29}$	$93.55_{\pm1.23}$	$3.00_{\pm0.00}$	$58.45_{\pm0.23}$	$93.25_{\pm0.36}$	$5.00_{\pm0.00}$	$56.69_{\pm0.43}$	$94.29_{\pm1.46}$	$7.00_{\pm0.00}$	$54.33_{\pm0.41}$
L1-Sparse	$90.04_{\pm1.28}$	$1.00_{\pm0.00}$	$59.75_{\pm0.35}$	$93.47_{\pm1.30}$	$3.00_{\pm0.00}$	$58.50_{\pm0.27}$	$93.32_{\pm0.46}$	$5.00_{\pm0.00}$	$56.71_{\pm0.48}$	$94.36_{\pm1.33}$	$7.00_{\pm0.00}$	$54.31_{\pm0.44}$
GA	$0.07_{\pm0.13}$	$\mathbf{0.00}_{\pm0.00}$	$52.62_{\pm0.05}$	$98.09_{\pm0.15}$	$1.00_{\pm0.00}$	$50.14_{\pm0.98}$	$99.93_{\pm0.13}$	$1.00_{\pm0.00}$	$43.81_{\pm0.37}$	$99.93_{\pm0.13}$	$1.00_{\pm0.00}$	$45.29_{\pm0.46}$
AggSum	$0.07_{\pm0.13}$	$\mathbf{0.00}_{\pm0.00}$	$52.61_{\pm0.06}$	$30.82_{\pm5.95}$	$1.00_{\pm0.00}$	$41.74_{\pm0.50}$	$96.49_{\pm2.18}$	$2.00_{\pm0.00}$	$44.94_{\pm0.91}$	$96.39_{\pm3.29}$	$2.00_{\pm0.00}$	$42.97_{\pm0.73}$
Sequential	$0.07_{\pm0.13}$	$\mathbf{0.00}_{\pm0.00}$	$52.62_{\pm0.05}$	$98.89_{\pm0.46}$	$1.00_{\pm0.00}$	$48.63_{\pm0.88}$	$100.00_{\pm0.00}$	$1.00_{\pm0.00}$	$40.69_{\pm0.92}$	$99.93_{\pm0.13}$	$1.00_{\pm0.00}$	$42.96_{\pm1.52}$
PCGrad	$0.07_{\pm0.13}$	$\mathbf{0.00}_{\pm0.00}$	$52.62_{\pm0.05}$	$53.62_{\pm15.38}$	$1.00_{\pm0.00}$	$44.15_{\pm0.70}$	$93.88_{\pm1.66}$	$2.00_{\pm0.00}$	$47.32_{\pm0.38}$	$92.10_{\pm2.25}$	$2.00_{\pm0.00}$	$44.05_{\pm0.71}$
MGDA	$0.07_{\pm0.13}$	$\mathbf{0.00}_{\pm0.00}$	$52.62_{\pm0.03}$	$32.64_{\pm4.62}$	$2.00_{\pm0.00}$	$42.58_{\pm0.93}$	$84.67_{\pm2.45}$	$5.00_{\pm0.00}$	$50.47_{\pm1.60}$	$89.78_{\pm3.83}$	$7.00_{\pm0.00}$	$48.92_{\pm2.84}$
FedOSD	$0.58_{\pm0.26}$	$\mathbf{0.00}_{\pm0.00}$	$59.05_{\pm0.24}$	$95.82_{\pm0.32}$	$1.00_{\pm0.00}$	$58.02_{\pm0.31}$	$98.38_{\pm0.85}$	$1.00_{\pm0.00}$	$55.73_{\pm0.70}$	$98.83_{\pm0.65}$	$1.00_{\pm0.00}$	$53.25_{\pm0.89}$
FedCARE	$2.49_{\pm0.32}$	$\mathbf{0.00}_{\pm0.00}$	$59.71_{\pm0.16}$	$94.95_{\pm0.79}$	$1.33_{\pm0.58}$	$58.35_{\pm0.35}$	$97.80_{\pm0.80}$	$1.33_{\pm0.58}$	$56.40_{\pm0.58}$	$98.39_{\pm0.73}$	$1.67_{\pm0.58}$	$53.54_{\pm0.54}$
RFE*	$3.53_{\pm0.48}$	$\mathbf{0.00}_{\pm0.00}$	$59.08_{\pm0.22}$	$95.90_{\pm0.26}$	$1.00_{\pm0.00}$	$57.98_{\pm0.35}$	$98.31_{\pm0.93}$	$1.00_{\pm0.00}$	$55.59_{\pm0.57}$	$98.91_{\pm0.62}$	$2.33_{\pm0.58}$	$52.88_{\pm0.62}$
FedRazor	$0.07_{\pm0.13}$	$\mathbf{0.00}_{\pm0.00}$	$52.24_{\pm0.07}$	$98.24_{\pm0.56}$	$1.00_{\pm0.00}$	$47.77_{\pm2.08}$	$100.00_{\pm0.00}$	$1.00_{\pm0.00}$	$39.51_{\pm0.19}$	$99.93_{\pm0.13}$	$1.00_{\pm0.00}$	$44.32_{\pm0.71}$
GDFA-U	$0.15_{\pm0.26}$	$\mathbf{0.00}_{\pm0.00}$	$52.64_{\pm0.10}$	$90.02_{\pm1.38}$	$3.00_{\pm0.00}$	$57.01_{\pm1.12}$	$93.87_{\pm1.15}$	$5.00_{\pm0.00}$	$56.74_{\pm0.34}$	$94.84_{\pm1.19}$	$7.00_{\pm0.00}$	$54.03_{\pm0.60}$
NoT	$5.91_{\pm1.34}$	$\mathbf{0.00}_{\pm0.00}$	$55.65_{\pm0.89}$	$8.19_{\pm2.45}$	$0.33_{\pm0.58}$	$54.76_{\pm1.07}$	$8.12_{\pm1.84}$	$0.33_{\pm0.58}$	$53.46_{\pm1.46}$	$8.06_{\pm3.08}$	$\mathbf{0.00}_{\pm0.00}$	$50.26_{\pm1.33}$
SalUn	$\mathbf{0.00}_{\pm0.00}$	$\mathbf{0.00}_{\pm0.00}$	$52.28_{\pm0.78}$	$11.22_{\pm5.26}$	$0.33_{\pm0.58}$	$47.78_{\pm1.47}$	$31.15_{\pm2.32}$	$4.00_{\pm1.00}$	$48.65_{\pm0.35}$	$47.29_{\pm4.05}$	$6.00_{\pm1.00}$	$49.11_{\pm0.29}$
FAC	$2.05_{\pm0.62}$	$\mathbf{0.00}_{\pm0.00}$	$55.15_{\pm0.88}$	$\mathbf{1.48}_{\pm0.47}$	$\mathbf{0.00}_{\pm0.00}$	$56.38_{\pm0.48}$	$\mathbf{1.94}_{\pm0.91}$	$\mathbf{0.00}_{\pm0.00}$	$54.36_{\pm0.49}$	$\mathbf{0.73}_{\pm0.26}$	$\mathbf{0.00}_{\pm0.00}$	$51.56_{\pm0.34}$

falls from 93–99% at $K = 3$ to 46–52% at $K = 7$, while NoT and FAC stay between 94% and 105% at every K.

Table A4: **Worst-source NFP on the CNN benchmark** (CIFAR-10, %, means over three seeds). For each seed, NFP is the smallest per-source progress $(\mathrm{ASR}_{\text{No Unl.}} - \mathrm{ASR})/(\mathrm{ASR}_{\text{No Unl.}} - \mathrm{ASR}_{\text{Retrain}})$: 100 means the least-removed source is at its Retrain level, and values near 0 mean it was not removed. NFP uses seed-matched Retrain values and the survivor test against the mean plus three pooled standard deviations, so a source just above its threshold, as for NoT at $K = 3$ and 5, still has NFP near 100. †Collapsed (R-Acc below 30%). *Project adaptation.

Method	non-IID				IID			
	$K=1$	3	5	7	$K=1$	3	5	7
FT	1	-3	-2	-3	2	0	1	0
GA	105	-3	-5	-4	107	-6	-7	-10
AggSum	105	45	-2	-6	107	72	-4	-5
PCGrad	105	25	0	-3	107	46	-1	-1
MGDA	105	21	-2	-2	107	69	10	6
Sequential	105	-3	-10	-7	107	-7	-9	-10
FedOSD	102	-1	-6	-4	107	-3	-5	-9
FedCARE	104	-2	-6	-4	105	-2	-5	-8
RFE*	96	-1	-5	-4	103	-3	-5	-9
FedRazor	105	-3	-6	-4	107	-6	-7	-10
NoT	99	94	95	97	101	98	96	97
NegGrad+	3	-3	-1	-4	5	0	1	-1
GDFA-U	105	-2	-1	-3	107	3	0	0
SalUn	105	99	67	46	107	93	70	52
L1-Sparse	1	-3	-2	-3	2	0	1	0
FAC	103	103	100	100	105	103	102	103

Survivor threshold. The survivor test uses the matched Retrain mean plus three pooled standard deviations (Section 3.4); with three Retrain seeds, the pooled standard deviation is 0.71–1.48 points on the CNN benchmark. Table A5 recounts the Table 1 runs with two and four pooled standard deviations and with a fixed margin of ten points above the Retrain mean; the three-SD column reproduces Table 1. FAC is survivor-free at every K under all four rules: in every seed and at every K, each of its per-source ASRs is at or below the matched Retrain mean. No baseline is survivor-free at every K under any rule; the closest is NoT, which clears three of the four source counts with the ten-point margin.

Baseline learning rates. The staged baselines use the protocol's default unlearning rate 4×10^{-4}, while FAC uses 10^{-2} on the CNN; because the update rules differ, a shared nominal rate is not by

Table A5: **Survivor-threshold sensitivity** (CNN, CIFAR-10, non-IID; the Table 1 runs). Thresholds are the matched Retrain mean plus 2, 3 or 4 pooled standard deviations, or plus 10 points. Clean K: source counts $K \in \{1, 3, 5, 7\}$ with no survivor in any seed; ΣSurv: mean survivor count summed over the four K. *Project adaptation.

Method	2σ		3σ (paper)		4σ		$+10$ points	
	clean K	ΣSurv	clean K	ΣSurv	clean K	ΣSurv	clean K	ΣSurv
FT	0/4	16.00	0/4	16.00	0/4	16.00	0/4	16.00
NegGrad+	0/4	16.00	0/4	16.00	0/4	16.00	0/4	16.00
L1-Sparse	0/4	16.00	0/4	16.00	0/4	16.00	0/4	16.00
GA	1/4	3.00	1/4	3.00	1/4	3.00	1/4	3.00
AggSum	1/4	5.67	1/4	5.67	1/4	5.33	1/4	5.00
Sequential	1/4	3.00	1/4	3.00	1/4	3.00	1/4	3.00
PCGrad	1/4	5.67	1/4	5.67	1/4	5.67	1/4	5.33
MGDA	1/4	11.33	1/4	11.33	1/4	11.33	1/4	11.00
FedOSD	1/4	3.00	1/4	3.00	1/4	3.00	1/4	3.00
FedCARE	1/4	3.00	1/4	3.00	1/4	3.00	1/4	3.00
RFE*	0/4	4.33	1/4	3.33	1/4	3.33	1/4	3.33
FedRazor	1/4	3.00	1/4	3.00	1/4	3.00	1/4	3.00
GDFA-U	1/4	15.00	1/4	15.00	1/4	15.00	1/4	15.00
NoT	1/4	3.00	2/4	2.33	2/4	1.67	3/4	0.33
SalUn	1/4	10.67	2/4	10.00	2/4	9.67	2/4	9.00
FAC	4/4	0.00	4/4	0.00	4/4	0.00	4/4	0.00

itself a fairness criterion. Table A6 therefore reruns FedOSD, PCGrad and SalUn at 2×10^{-3}, 10^{-2} (FAC's rate) and 5×10^{-2}, and NoT at 3×10^{-3} and 3×10^{-2} (its default is 10^{-2}), at $K = 5$ with every other setting of Table 1. At FAC's rate, FedOSD and PCGrad still leave one and three sources; only SalUn becomes survivor-free, and at that rate it also clears $K = 1$ and 3 but leaves 2.33 sources at $K = 7$, with lower R-Acc than FAC at every K. At 5×10^{-2}, FedOSD, PCGrad and SalUn collapse. This test varies only the unlearning rate on the CNN benchmark and is not an exhaustive retuning of method-specific hyperparameters.

D Datasets and Backbones

This section gives the full results behind Section 4.2. At three sources, NF-ResNet-18 is evaluated with all fifteen baselines (Table A7); ViT-S/16 is evaluated with FT, PCGrad, FedOSD, FedCARE, GDFA-U, NoT and SalUn (Table A8). Each dataset–backbone setting uses matched No Unlearning and Retrain anchors and three seeds. NFP is the minimum per-source normalized progress within a seed, followed by the mean and sample standard deviation across seeds; it is not the average progress across sources.

NF-ResNet-18. Across the nine NF-ResNet-18 runs with three sources, FAC leaves one surviving checker source in one CIFAR-100 run (1.09% ASR, 6 of 552 probes, against a threshold of 0.82%); the other eight runs are survivor-free, so the reported mean survivor count on CIFAR-100 is 0.33. SalUn reaches 6.10% worst-source ASR on CIFAR-10 but leaves one source in every seed and lowers R-Acc to 50.42%, against FAC's 76.59%. On CIFAR-100, SalUn's 22.03% R-Acc falls below the collapse threshold in every seed; on SVHN, its 62.31% worst-source ASR and 40.29% R-Acc show that low survivor counts alone need not imply low attack success or high retained accuracy. GA, AggSum, MGDA, Sequential, RFE, FedRazor, NegGrad+ and L1-Sparse leave a source in at least one seed of every dataset. NoT is the only baseline with no survivor in any CIFAR-10 or CIFAR-100 seed, but its R-Acc is lower than FAC's. On SVHN, FAC is survivor-free in all three runs, while every baseline leaves a source in at least one run.

ViT-S/16. SalUn has 0.30% and 0.09% worst-source ASR on CIFAR-10 and SVHN, compared with FAC's 3.20% and 0.94%. It removes all sources in both settings; FAC leaves one in one CIFAR-10 seed. On CIFAR-100, FAC has 0.35% worst-source ASR and no survivor, whereas SalUn has 1.05% and one survivor in one seed. FAC has the higher mean retained accuracy in all three com-

Table A6: **Baseline learning rates** (CNN, CIFAR-10, non-IID; $a_{\pm b}$ over three seeds, %). Top: $K = 5$, each baseline at its default rate (as in Table 1) and at the listed rates, all other settings unchanged; a rerun of the default FedOSD configuration reproduces the corresponding Table 1 run (seed 1) exactly. Bottom: SalUn at $\eta_0 = 10^{-2}$ against FAC at every K. Blue as in Table 1. †R-Acc below 30% in at least one seed; deletion scores are not ranked.

Configuration ($K = 5$)	Worst ASR↓	Surv↓	R-Acc↑
FedOSD, $\eta_0 = 0.0004$ (Table 1)	$98.55_{\pm 0.55}$	1.00	$53.43_{\pm 0.25}$
FedOSD, $\eta_0 = 0.002$	$77.72_{\pm 5.06}$	2.00	$54.00_{\pm 0.17}$
FedOSD, $\eta_0 = 0.01$	$69.17_{\pm 12.46}$	1.00	$54.20_{\pm 0.44}$
FedOSD, $\eta_0 = 0.05$	$92.67_{\pm 2.25}$	3.00	$28.47^{\dagger}_{\pm 2.15}$
PCGrad, $\eta_0 = 0.0004$ (Table 1)	$91.04_{\pm 2.21}$	2.00	$45.20_{\pm 0.70}$
PCGrad, $\eta_0 = 0.002$	$95.00_{\pm 1.97}$	1.00	$50.03_{\pm 0.59}$
PCGrad, $\eta_0 = 0.01$	$96.59_{\pm 1.42}$	3.00	$48.47_{\pm 0.67}$
PCGrad, $\eta_0 = 0.05$	$95.44_{\pm 3.63}$	4.00	$17.90^{\dagger}_{\pm 1.31}$
SalUn, $\eta_0 = 0.0004$ (Table 1)	$29.47_{\pm 1.58}$	3.67	$47.00_{\pm 1.04}$
SalUn, $\eta_0 = 0.002$	$18.23_{\pm 0.84}$	1.00	$49.00_{\pm 0.96}$
SalUn, $\eta_0 = 0.01$	$2.84_{\pm 1.00}$	**0.00**	$49.37_{\pm 0.90}$
SalUn, $\eta_0 = 0.05$	$2.67_{\pm 3.87}$	0.33	$22.33^{\dagger}_{\pm 4.18}$
NoT, $\eta_0 = 0.003$	$9.29_{\pm 1.39}$	1.00	$44.60_{\pm 0.20}$
NoT, $\eta_0 = 0.01$ (Table 1)	$9.71_{\pm 2.35}$	1.33	$50.17_{\pm 0.23}$
NoT, $\eta_0 = 0.03$	$11.22_{\pm 4.37}$	1.33	$56.57_{\pm 0.35}$
FAC (Table 1)	**1.98**$_{\pm 1.38}$	**0.00**	$50.83_{\pm 0.32}$

Method	$K = 1$		$K = 3$		$K = 5$		$K = 7$	
	Surv	R-Acc	Surv	R-Acc	Surv	R-Acc	Surv	R-Acc
SalUn, $\eta_0 = 0.01$	**0.00**	$48.70_{\pm 0.78}$	**0.00**	$48.73_{\pm 0.81}$	**0.00**	$49.37_{\pm 0.90}$	2.33	$42.83_{\pm 0.25}$
FAC (Table 1)	**0.00**	$50.47_{\pm 1.59}$	**0.00**	$53.57_{\pm 0.40}$	**0.00**	$50.83_{\pm 0.32}$	**0.00**	$47.17_{\pm 0.47}$

Table A7: **NF-ResNet-18 across datasets** ($K = 3$, ten clients). Entries are $a_{\pm b}$ over three seeds; worst-source ASR, R-Acc and worst-source NFP in %; Surv counts sources above the matched Retrain thresholds. No Unlearning and Retrain are anchors, not ranked. Blue marks the unique lowest ASR per column and every survivor-free Surv entry among ranked methods. †R-Acc below 30% in every seed; deletion scores are reported but not ranked.

Method	CIFAR-10				CIFAR-100				SVHN			
	ASR↓	Surv↓	R-Acc↑	NFP↑	ASR↓	Surv↓	R-Acc↑	NFP↑	ASR↓	Surv↓	R-Acc↑	NFP↑
No Unlearning	$95.51_{\pm 0.91}$	$3.00_{\pm 0.00}$	$75.56_{\pm 0.10}$	$0.00_{\pm 0.00}$	$76.32_{\pm 2.09}$	$3.00_{\pm 0.00}$	$44.34_{\pm 0.23}$	$0.00_{\pm 0.00}$	$94.22_{\pm 0.74}$	$3.00_{\pm 0.00}$	$91.03_{\pm 0.11}$	$0.00_{\pm 0.00}$
Retrain	$4.85_{\pm 1.90}$	$0.00_{\pm 0.00}$	$74.78_{\pm 0.90}$	$100.00_{\pm 0.00}$	$0.50_{\pm 0.10}$	$0.00_{\pm 0.00}$	$46.19_{\pm 0.53}$	$100.00_{\pm 0.00}$	$22.61_{\pm 5.19}$	$0.00_{\pm 0.00}$	$91.16_{\pm 0.13}$	$100.00_{\pm 0.00}$
FT	$95.98_{\pm 0.98}$	$3.00_{\pm 0.00}$	$76.99_{\pm 0.26}$	$-0.51_{\pm 0.29}$	$69.18_{\pm 1.10}$	$3.00_{\pm 0.00}$	$47.11_{\pm 0.16}$	$-30.81_{\pm 53.72}$	$92.34_{\pm 1.27}$	$3.00_{\pm 0.00}$	$92.61_{\pm 0.07}$	$-9.19_{\pm 4.64}$
NegGrad+	$97.16_{\pm 0.53}$	$3.00_{\pm 0.00}$	$76.98_{\pm 0.21}$	$-1.81_{\pm 0.48}$	$71.87_{\pm 4.00}$	$3.00_{\pm 0.00}$	$47.11_{\pm 0.13}$	$-17.57_{\pm 17.20}$	$89.19_{\pm 1.36}$	$2.67_{\pm 0.58}$	$92.66_{\pm 0.09}$	$-14.21_{\pm 5.44}$
L1-Sparse	$95.98_{\pm 0.98}$	$3.00_{\pm 0.00}$	$76.98_{\pm 0.27}$	$-0.51_{\pm 0.29}$	$69.18_{\pm 1.10}$	$3.00_{\pm 0.00}$	$47.12_{\pm 0.15}$	$-30.81_{\pm 53.72}$	$92.34_{\pm 1.27}$	$3.00_{\pm 0.00}$	$92.61_{\pm 0.07}$	$-9.19_{\pm 4.64}$
GA	$98.11_{\pm 0.67}$	$1.67_{\pm 1.15}$	$74.24_{\pm 0.68}$	$-2.83_{\pm 0.48}$	$45.85_{\pm 32.26}$	$2.00_{\pm 0.00}$	$43.85_{\pm 0.42}$	$35.96_{\pm 46.07}$	$95.48_{\pm 0.81}$	$1.33_{\pm 0.58}$	$90.94_{\pm 0.77}$	$-9.98_{\pm 9.22}$
AggSum	$97.10_{\pm 0.67}$	$1.00_{\pm 0.00}$	$72.40_{\pm 0.82}$	$-1.74_{\pm 0.80}$	$64.64_{\pm 37.87}$	$1.67_{\pm 0.58}$	$43.57_{\pm 0.25}$	$13.17_{\pm 48.99}$	$93.72_{\pm 1.69}$	$1.67_{\pm 0.58}$	$91.11_{\pm 0.24}$	$-16.73_{\pm 6.83}$
Sequential	$98.23_{\pm 0.31}$	$1.33_{\pm 0.58}$	$74.67_{\pm 0.19}$	$-2.97_{\pm 0.73}$	$58.09_{\pm 37.29}$	$1.00_{\pm 0.00}$	$44.15_{\pm 0.16}$	$-90.42_{\pm 228.99}$	$95.59_{\pm 0.99}$	$1.67_{\pm 0.58}$	$91.45_{\pm 0.15}$	$-16.43_{\pm 13.54}$
PCGrad	$74.23_{\pm 40.64}$	$2.00_{\pm 0.00}$	$72.08_{\pm 0.53}$	$22.81_{\pm 44.42}$	$40.00_{\pm 30.12}$	$2.33_{\pm 0.58}$	$43.74_{\pm 0.33}$	$40.74_{\pm 45.96}$	$87.12_{\pm 12.60}$	$1.33_{\pm 0.58}$	$87.93_{\pm 1.49}$	$0.79_{\pm 11.27}$
MGDA	$95.92_{\pm 1.51}$	$2.00_{\pm 0.00}$	$74.04_{\pm 1.53}$	$-0.48_{\pm 2.62}$	$38.81_{\pm 26.16}$	$2.00_{\pm 0.00}$	$43.89_{\pm 0.45}$	$47.61_{\pm 34.98}$	$95.57_{\pm 0.87}$	$1.67_{\pm 0.58}$	$89.88_{\pm 1.15}$	$-1.47_{\pm 1.20}$
FedOSD	$93.85_{\pm 7.73}$	$1.33_{\pm 0.58}$	$76.09_{\pm 0.26}$	$1.89_{\pm 7.57}$	$44.03_{\pm 32.82}$	$2.00_{\pm 0.00}$	$45.50_{\pm 0.13}$	$38.80_{\pm 45.56}$	$66.16_{\pm 52.24}$	$0.67_{\pm 0.58}$	$91.72_{\pm 0.31}$	$26.93_{\pm 66.54}$
FedCARE	$94.80_{\pm 5.48}$	$1.33_{\pm 0.58}$	$75.37_{\pm 0.36}$	$0.83_{\pm 5.08}$	$45.12_{\pm 31.03}$	$2.33_{\pm 0.58}$	$45.50_{\pm 0.17}$	$37.39_{\pm 43.07}$	$78.24_{\pm 31.28}$	$1.33_{\pm 0.58}$	$91.72_{\pm 0.21}$	$8.96_{\pm 36.02}$
RFE	$95.98_{\pm 4.20}$	$1.33_{\pm 0.58}$	$76.26_{\pm 0.33}$	$-0.47_{\pm 3.66}$	$46.55_{\pm 24.02}$	$2.33_{\pm 0.58}$	$45.81_{\pm 0.20}$	$33.35_{\pm 35.93}$	$88.12_{\pm 13.66}$	$1.33_{\pm 0.58}$	$92.13_{\pm 0.09}$	$1.11_{\pm 22.60}$
FedRazor	$97.99_{\pm 0.57}$	$1.33_{\pm 0.58}$	$73.34_{\pm 1.46}$	$-2.71_{\pm 0.55}$	$38.87_{\pm 32.85}$	$1.67_{\pm 0.58}$	$42.97_{\pm 0.51}$	$45.66_{\pm 46.28}$	$95.73_{\pm 0.22}$	$1.00_{\pm 0.00}$	$88.69_{\pm 1.80}$	$-10.42_{\pm 9.95}$
GDFA-U	$86.41_{\pm 16.64}$	$2.67_{\pm 0.58}$	$75.42_{\pm 1.07}$	$9.82_{\pm 18.22}$	$75.58_{\pm 3.91}$	$3.00_{\pm 0.00}$	$46.37_{\pm 0.22}$	$-14.49_{\pm 24.07}$	$92.67_{\pm 0.33}$	$3.00_{\pm 0.00}$	$90.37_{\pm 0.09}$	$1.69_{\pm 1.11}$
NoT	$2.94_{\pm 1.56}$	**0.00**$_{\pm 0.00}$	$70.00_{\pm 0.21}$	$100.00_{\pm 0.00}$	**0.18**$_{\pm 0.18}$	**0.00**$_{\pm 0.00}$	$38.18_{\pm 0.37}$	$100.00_{\pm 0.00}$	$19.52_{\pm 15.42}$	$0.67_{\pm 1.15}$	$89.33_{\pm 0.16}$	$90.74_{\pm 12.37}$
SalUn	$6.10_{\pm 2.76}$	$1.00_{\pm 0.00}$	$50.42_{\pm 1.23}$	$93.61_{\pm 2.92}$	$2.90_{\pm 2.06}$	$1.33_{\pm 0.58}$	$22.03^{\dagger}_{\pm 0.89}$	$96.33_{\pm 2.79}$	$62.31_{\pm 37.62}$	$0.67_{\pm 0.58}$	$40.29_{\pm 1.81}$	$20.03_{\pm 55.46}$
FAC	**1.12**$_{\pm 0.10}$	**0.00**$_{\pm 0.00}$	$76.59_{\pm 0.06}$	$100.00_{\pm 0.17}$	$0.66_{\pm 0.38}$	$0.33_{\pm 0.58}$	$46.94_{\pm 0.29}$	$99.34_{\pm 0.40}$	**1.83**$_{\pm 0.87}$	**0.00**$_{\pm 0.00}$	$92.17_{\pm 0.11}$	$101.68_{\pm 1.85}$

parisons. These results show a backbone-dependent comparison with SalUn, rather than uniform dominance by either method. NoT's zero ASR on the two CIFAR datasets coincides with collapse; on SVHN, collapse instead coexists with 100% worst-source ASR.

Source count on NF-ResNet-18. The separate CIFAR-10 source-count study compares FT, PCGrad, FedOSD, FedCARE, GDFA-U, NoT and FAC at $K = 1, 5, 7$, alongside the $K = 3$ point reused from the cross-dataset study (Table A9). FAC uses $(\eta_0, \omega_{\mathrm{f}}) = (0.01, 1)$ at $K = 3$, $(0.02, 5)$ at $K = 5$, and $(0.03, 5)$ at $K = 1, 7$; these are per-setting results rather than a sweep holding every coefficient fixed. FAC removes every source at $K = 1, 3, 5$ but leaves one or two at $K = 7$, where

Table A8: **ViT-S/16 across datasets** ($K = 3$, ten clients; format as in Table A7). †R-Acc below 30% in every seed; deletion scores are reported but not ranked.

Method	CIFAR-10				CIFAR-100				SVHN			
	ASR↓	Surv↓	R-Acc↑	NFP↑	ASR↓	Surv↓	R-Acc↑	NFP↑	ASR↓	Surv↓	R-Acc↑	NFP↑
No Unlearning	$98.87_{\pm 0.43}$	$3.00_{\pm 0.00}$	$95.63_{\pm 0.21}$	$0.00_{\pm 0.00}$	$97.43_{\pm 0.80}$	$3.00_{\pm 0.00}$	$82.77_{\pm 0.25}$	$0.00_{\pm 0.00}$	$99.36_{\pm 0.15}$	$3.00_{\pm 0.00}$	$93.57_{\pm 0.06}$	$0.00_{\pm 0.00}$
Retrain	$2.12_{\pm 1.57}$	$0.00_{\pm 0.00}$	$95.90_{\pm 0.62}$	$100.00_{\pm 0.00}$	$0.58_{\pm 0.48}$	$0.00_{\pm 0.00}$	$82.50_{\pm 0.36}$	$100.00_{\pm 0.00}$	$3.46_{\pm 0.84}$	$0.00_{\pm 0.00}$	$93.17_{\pm 0.15}$	$100.00_{\pm 0.00}$
FT	$98.95_{\pm 0.30}$	$3.00_{\pm 0.00}$	$96.00_{\pm 0.40}$	$-0.08_{\pm 0.14}$	$97.14_{\pm 0.69}$	$3.00_{\pm 0.00}$	$83.00_{\pm 0.10}$	$-0.25_{\pm 0.65}$	$99.21_{\pm 0.18}$	$3.00_{\pm 0.00}$	$93.60_{\pm 0.10}$	$0.03_{\pm 0.05}$
PCGrad	$36.13_{\pm 17.28}$	$1.67_{\pm 0.58}$	$95.43_{\pm 0.45}$	$64.64_{\pm 17.33}$	$12.29_{\pm 9.00}$	$1.67_{\pm 0.58}$	$82.73_{\pm 0.21}$	$87.35_{\pm 9.35}$	$9.28_{\pm 7.36}$	$1.33_{\pm 1.53}$	$93.47_{\pm 0.21}$	$91.86_{\pm 8.34}$
FedOSD	$17.11_{\pm 7.46}$	$2.00_{\pm 0.00}$	$95.70_{\pm 0.35}$	$84.44_{\pm 7.58}$	$7.76_{\pm 7.17}$	$1.33_{\pm 1.15}$	$82.87_{\pm 0.21}$	$91.99_{\pm 7.40}$	$34.63_{\pm 56.17}$	$0.67_{\pm 0.58}$	$93.47_{\pm 0.12}$	$65.58_{\pm 57.28}$
FedCARE	$17.63_{\pm 8.63}$	$2.00_{\pm 0.00}$	$95.70_{\pm 0.35}$	$83.75_{\pm 9.07}$	$8.13_{\pm 7.61}$	$1.33_{\pm 1.15}$	$82.83_{\pm 0.25}$	$91.60_{\pm 7.85}$	$34.46_{\pm 56.33}$	$0.67_{\pm 0.58}$	$93.50_{\pm 0.10}$	$65.76_{\pm 57.44}$
GDFA-U	$98.96_{\pm 0.27}$	$3.00_{\pm 0.00}$	$95.90_{\pm 0.40}$	$-0.14_{\pm 0.24}$	$80.04_{\pm 22.16}$	$2.67_{\pm 0.58}$	$83.03_{\pm 0.31}$	$16.37_{\pm 22.71}$	$99.26_{\pm 0.19}$	$3.00_{\pm 0.00}$	$93.57_{\pm 0.06}$	$0.00_{\pm 0.00}$
NoT	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$11.93^{\dagger}_{\pm 0.81}$	$100.44_{\pm 0.40}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$1.30^{\dagger}_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$22.80^{\dagger}_{\pm 0.00}$	$-0.72_{\pm 0.21}$
SalUn	$\mathbf{0.30}_{\pm 0.35}$	$\mathbf{0.00}_{\pm 0.00}$	$95.10_{\pm 0.69}$	$100.37_{\pm 0.47}$	$1.05_{\pm 0.59}$	$0.33_{\pm 0.58}$	$82.17_{\pm 0.21}$	$99.26_{\pm 0.55}$	$\mathbf{0.09}_{\pm 0.09}$	$\mathbf{0.00}_{\pm 0.00}$	$93.27_{\pm 0.15}$	$100.98_{\pm 0.40}$
FAC	$3.20_{\pm 5.42}$	$0.33_{\pm 0.58}$	$95.47_{\pm 0.47}$	$97.54_{\pm 5.12}$	$\mathbf{0.35}_{\pm 0.46}$	$\mathbf{0.00}_{\pm 0.00}$	$83.07_{\pm 0.15}$	$99.78_{\pm 0.23}$	$0.94_{\pm 1.48}$	$\mathbf{0.00}_{\pm 0.00}$	$93.60_{\pm 0.10}$	$100.17_{\pm 1.46}$

both FAC and NoT leave survivors in at least one seed and deletion and retained accuracy must be read together. This study does not include SalUn at $K = 1, 5, 7$, so it establishes the reported FAC–NoT trade-off without ranking those unmeasured settings.

Table A9: **Source count on NF-ResNet-18** (CIFAR-10, ten clients; format as in Table A7). The $K = 3$ block repeats the cross-dataset runs of Table A7.

Method	$K = 1$			$K = 3$			$K = 5$			$K = 7$		
	ASR↓	Surv↓	R-Acc↑	ASR↓	Surv↓	R-Acc↑	ASR↓	Surv↓	R-Acc↑	ASR↓	Surv↓	R-Acc↑
No Unlearning	$95.95_{\pm 1.07}$	$1.00_{\pm 0.00}$	$76.27_{\pm 0.12}$	$95.51_{\pm 0.91}$	$3.00_{\pm 0.00}$	$75.56_{\pm 0.10}$	$94.92_{\pm 0.72}$	$5.00_{\pm 0.00}$	$76.30_{\pm 0.35}$	$95.69_{\pm 0.33}$	$7.00_{\pm 0.00}$	$72.77_{\pm 0.60}$
Retrain	$0.21_{\pm 0.09}$	$0.00_{\pm 0.00}$	$75.90_{\pm 0.56}$	$4.85_{\pm 1.90}$	$0.00_{\pm 0.00}$	$74.78_{\pm 0.90}$	$8.59_{\pm 2.32}$	$0.00_{\pm 0.00}$	$76.23_{\pm 0.59}$	$7.20_{\pm 0.20}$	$0.00_{\pm 0.00}$	$72.87_{\pm 0.65}$
FT	$95.64_{\pm 1.22}$	$1.00_{\pm 0.00}$	$77.00_{\pm 0.36}$	$95.98_{\pm 0.98}$	$3.00_{\pm 0.00}$	$76.99_{\pm 0.26}$	$95.80_{\pm 0.71}$	$5.00_{\pm 0.00}$	$78.17_{\pm 0.51}$	$95.58_{\pm 0.00}$	$7.00_{\pm 0.00}$	$76.17_{\pm 0.50}$
PCGrad	$1.47_{\pm 0.87}$	$1.00_{\pm 0.00}$	$70.87_{\pm 0.23}$	$74.23_{\pm 40.64}$	$2.00_{\pm 0.00}$	$72.08_{\pm 0.53}$	$94.03_{\pm 0.81}$	$4.00_{\pm 1.00}$	$75.70_{\pm 0.56}$	$98.69_{\pm 0.33}$	$3.33_{\pm 0.58}$	$72.97_{\pm 0.83}$
FedOSD	$0.00_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$	$75.60_{\pm 0.20}$	$93.85_{\pm 7.73}$	$1.33_{\pm 0.58}$	$76.09_{\pm 0.26}$	$96.37_{\pm 3.46}$	$1.33_{\pm 0.58}$	$77.10_{\pm 0.26}$	$88.65_{\pm 0.47}$	$2.00_{\pm 0.00}$	$73.87_{\pm 0.60}$
FedCARE	$0.00_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$	$75.37_{\pm 0.12}$	$94.80_{\pm 5.48}$	$1.33_{\pm 0.58}$	$75.37_{\pm 0.36}$	$96.24_{\pm 3.49}$	$1.33_{\pm 0.58}$	$76.90_{\pm 0.17}$	$89.47_{\pm 0.58}$	$2.00_{\pm 0.00}$	$73.70_{\pm 0.92}$
GDFA-U	$1.89_{\pm 0.99}$	$1.00_{\pm 0.00}$	$70.50_{\pm 0.10}$	$86.41_{\pm 16.64}$	$2.67_{\pm 0.58}$	$75.42_{\pm 1.07}$	$96.22_{\pm 0.57}$	$5.00_{\pm 0.00}$	$77.33_{\pm 0.91}$	$96.06_{\pm 0.16}$	$7.00_{\pm 0.00}$	$75.60_{\pm 0.70}$
NoT	$0.00_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$	$65.77_{\pm 0.12}$	$2.94_{\pm 1.56}$	$\mathbf{0.00}_{\pm 0.00}$	$70.00_{\pm 0.21}$	$6.89_{\pm 4.60}$	$\mathbf{0.00}_{\pm 0.00}$	$68.17_{\pm 0.68}$	$\mathbf{6.73}_{\pm 6.01}$	$0.33_{\pm 0.58}$	$66.43_{\pm 0.32}$
FAC	$0.00_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$	$77.20_{\pm 1.00}$	$\mathbf{1.12}_{\pm 0.10}$	$\mathbf{0.00}_{\pm 0.00}$	$76.59_{\pm 0.06}$	$\mathbf{0.49}_{\pm 0.19}$	$\mathbf{0.00}_{\pm 0.00}$	$74.67_{\pm 0.50}$	$16.81_{\pm 3.62}$	$1.33_{\pm 0.58}$	$72.57_{\pm 0.91}$

E Component Ablations

This section supports Section 4.3. Tables A10–A13 use the CNN benchmark, three seeds, and the complete procedure of 50 unlearning and 50 recovery rounds; the position diagnostic of Table A14 stops at round 50, before recovery. Worst ASR is the worst-source ASR; Surv uses matched Retrain thresholds.

Disagreement across source counts. Table A10 separates where the state enters the update. The shared-target control keeps a proximal step toward z and the correction in server aggregation; the no-proximal control removes that local step but keeps the server correction. The no-state control sets every u_k to zero throughout. At $K = 1$, all four variants remove the single source on both partitions. For $K \in \{3, 5, 7\}$, each control leaves survivors, while FAC leaves none. At $K = 7$, removing the state leaves 4.00 sources on non-IID data and 5.00 on IID data. The non-IID control retains 49.10% accuracy against FAC's 47.17%, so its additional retained accuracy accompanies incomplete deletion.

Fixed request size. In source scaling, the request contains $M = 3K$ clients, so both source count and the deleted fraction increase. Table A11 fixes an IID roster of 15 deleted clients out of 30 and allocates it to one, three or five sources, keeping the attacker fraction and poison budget fixed. With one source, FAC and both controls leave no survivor; with three or five, the shared-target and no-proximal controls leave 2.00–2.33 sources while FAC leaves none. FAC has lower worst-source ASR in all 18 seed-matched comparisons with those controls, at 1.10–3.11 points lower R-Acc. This control separates source allocation from request size; it does not compare client-indexed states with a merged state.

Spatial selection and state merging. Table A12 gives the remaining component changes at $K = 5$. Merging the forget-side states changes worst-source ASR by at most 0.53 points for FAC and the shared-target and no-proximal controls on non-IID data. The merged variant initializes one forget-side state at zero and uses it in every forget client's local target. The server combines the

Table A10: **State variants across source counts** (CNN; $a_{\pm b}$ over three seeds, %). The IID partition tests the same variants with more uniform client data. No seed collapses.

Variant	non-IID			IID		
	ASR↓	Surv↓	R-Acc↑	ASR↓	Surv↓	R-Acc↑
$K = 1$, FAC	$2.37_{\pm 0.95}$	$0.00_{\pm 0.00}$	$50.47_{\pm 1.59}$	$2.05_{\pm 0.62}$	$0.00_{\pm 0.00}$	$55.15_{\pm 0.88}$
$K = 1$, Shared target z	$3.81_{\pm 0.52}$	$0.00_{\pm 0.00}$	$52.40_{\pm 0.61}$	$2.93_{\pm 0.82}$	$0.00_{\pm 0.00}$	$56.49_{\pm 0.85}$
$K = 1$, No proximal step	$3.82_{\pm 0.35}$	$0.00_{\pm 0.00}$	$52.33_{\pm 0.59}$	$2.78_{\pm 1.07}$	$0.00_{\pm 0.00}$	$56.61_{\pm 0.87}$
$K = 1$, No state $u = 0$	$0.92_{\pm 0.61}$	$0.00_{\pm 0.00}$	$49.87_{\pm 0.64}$	$1.24_{\pm 0.76}$	$0.00_{\pm 0.00}$	$53.89_{\pm 0.83}$
$K = 3$, FAC	$0.61_{\pm 0.13}$	$0.00_{\pm 0.00}$	$53.57_{\pm 0.40}$	$1.48_{\pm 0.47}$	$0.00_{\pm 0.00}$	$56.38_{\pm 0.48}$
$K = 3$, Shared target z	$13.49_{\pm 4.89}$	$2.33_{\pm 0.58}$	$55.03_{\pm 0.21}$	$11.85_{\pm 2.28}$	$1.67_{\pm 0.58}$	$57.64_{\pm 0.31}$
$K = 3$, No proximal step	$14.56_{\pm 6.08}$	$1.67_{\pm 0.58}$	$55.07_{\pm 0.32}$	$11.81_{\pm 3.34}$	$1.33_{\pm 0.58}$	$57.69_{\pm 0.23}$
$K = 3$, No state $u = 0$	$14.40_{\pm 3.52}$	$2.33_{\pm 0.58}$	$55.20_{\pm 0.20}$	$13.71_{\pm 2.81}$	$2.00_{\pm 1.00}$	$57.56_{\pm 0.51}$
$K = 5$, FAC	$1.98_{\pm 1.38}$	$0.00_{\pm 0.00}$	$50.83_{\pm 0.32}$	$1.94_{\pm 0.91}$	$0.00_{\pm 0.00}$	$54.36_{\pm 0.49}$
$K = 5$, Shared target z	$14.64_{\pm 1.06}$	$1.00_{\pm 0.00}$	$52.93_{\pm 0.40}$	$18.87_{\pm 3.75}$	$2.00_{\pm 0.00}$	$56.31_{\pm 0.49}$
$K = 5$, No proximal step	$15.32_{\pm 1.65}$	$1.33_{\pm 0.58}$	$52.93_{\pm 0.35}$	$18.30_{\pm 4.82}$	$2.33_{\pm 0.58}$	$56.21_{\pm 0.54}$
$K = 5$, No state $u = 0$	$23.11_{\pm 3.02}$	$2.33_{\pm 0.58}$	$53.23_{\pm 0.32}$	$26.29_{\pm 2.50}$	$4.67_{\pm 0.58}$	$56.26_{\pm 0.37}$
$K = 7$, FAC	$1.60_{\pm 0.42}$	$0.00_{\pm 0.00}$	$47.17_{\pm 0.47}$	$0.73_{\pm 0.26}$	$0.00_{\pm 0.00}$	$51.56_{\pm 0.34}$
$K = 7$, Shared target z	$29.80_{\pm 13.30}$	$1.67_{\pm 1.15}$	$48.13_{\pm 0.23}$	$11.12_{\pm 2.79}$	$1.00_{\pm 1.00}$	$53.10_{\pm 0.61}$
$K = 7$, No proximal step	$31.02_{\pm 14.36}$	$1.67_{\pm 0.58}$	$48.13_{\pm 0.21}$	$12.51_{\pm 4.79}$	$1.00_{\pm 1.00}$	$52.95_{\pm 0.56}$
$K = 7$, No state $u = 0$	$58.20_{\pm 3.81}$	$4.00_{\pm 0.00}$	$49.10_{\pm 0.30}$	$41.37_{\pm 3.56}$	$5.00_{\pm 1.00}$	$53.61_{\pm 0.91}$

Table A11: **Fixed deleted roster** (CNN, IID, 15 of 30 clients deleted; $a_{\pm b}$ over three seeds, %). No seed collapses. The no-state variant was not run at $K = 1$ or 3; at $K = 5$ it coincides with Table A10.

Variant	ASR↓	Surv↓	R-Acc↑
$K = 1$, FAC	$0.70_{\pm 0.30}$	$0.00_{\pm 0.00}$	$51.02_{\pm 1.17}$
$K = 1$, Shared target z	$2.27_{\pm 0.69}$	$0.00_{\pm 0.00}$	$53.71_{\pm 1.39}$
$K = 1$, No proximal step	$2.31_{\pm 0.63}$	$0.00_{\pm 0.00}$	$53.77_{\pm 1.35}$
$K = 3$, FAC	$1.86_{\pm 1.99}$	$0.00_{\pm 0.00}$	$54.57_{\pm 0.84}$
$K = 3$, Shared target z	$13.26_{\pm 3.98}$	$2.00_{\pm 1.00}$	$56.22_{\pm 0.35}$
$K = 3$, No proximal step	$14.12_{\pm 3.45}$	$2.33_{\pm 0.58}$	$56.22_{\pm 0.52}$
$K = 5$, FAC	$1.94_{\pm 0.91}$	$0.00_{\pm 0.00}$	$54.36_{\pm 0.49}$
$K = 5$, Shared target z	$18.87_{\pm 3.75}$	$2.00_{\pm 0.00}$	$56.31_{\pm 0.49}$
$K = 5$, No proximal step	$18.30_{\pm 4.82}$	$2.33_{\pm 0.58}$	$56.21_{\pm 0.54}$

mean forget-client model with this state, preserving the total forget-side aggregation weight, then updates the shared state by the difference between that mean model and the new consensus. Retain clients keep their individual states. The all-coordinate zone collapses every seed on both partitions; its recovery projection sets every parameter to zero. Forget-only saliency and the forget/retain ratio collapse three and two seeds, respectively, on each partition. Replacing running-maximum accumulation with an EMA leaves no survivor on either partition. All four tested support fractions, $\alpha \in \{0.01, 0.05, 0.20, 0.50\}$, are survivor-free on both partitions, while R-Acc decreases by 3.20 points on non-IID data and 3.43 on IID data. Only positive-score coordinates can enter the zone, so $\alpha = 0.50$ does not require half the parameters to be selected.

Table A12: **Component changes at** $K = 5$ (CNN; $a_{\pm b}$ over three seeds, %). Coll. counts seeds with R-Acc below 30%; their deletion scores are descriptive and not ranked.

Variant	non-IID				IID			
	ASR↓	Surv↓	R-Acc↑	Coll.	ASR↓	Surv↓	R-Acc↑	Coll.
FAC	$1.98_{\pm 1.38}$	$0.00_{\pm 0.00}$	$50.83_{\pm 0.32}$	0/3	$1.94_{\pm 0.91}$	$0.00_{\pm 0.00}$	$54.36_{\pm 0.49}$	0/3
Merged forget state	$2.14_{\pm 1.49}$	$0.00_{\pm 0.00}$	$50.77_{\pm 0.29}$	0/3	$2.02_{\pm 0.87}$	$0.00_{\pm 0.00}$	$54.38_{\pm 0.45}$	0/3
Shared target	$14.64_{\pm 1.06}$	$1.00_{\pm 0.00}$	$52.93_{\pm 0.40}$	0/3	$18.87_{\pm 3.75}$	$2.00_{\pm 0.00}$	$56.31_{\pm 0.49}$	0/3
Shared target, merged	$14.72_{\pm 1.17}$	$1.33_{\pm 0.58}$	$52.90_{\pm 0.26}$	0/3	$17.63_{\pm 4.12}$	$2.33_{\pm 0.58}$	$56.32_{\pm 0.21}$	0/3
No proximal step	$15.32_{\pm 1.65}$	$1.33_{\pm 0.58}$	$52.93_{\pm 0.35}$	0/3	$18.30_{\pm 4.82}$	$2.33_{\pm 0.58}$	$56.21_{\pm 0.54}$	0/3
No proximal step, merged	$15.85_{\pm 2.39}$	$1.33_{\pm 0.58}$	$52.87_{\pm 0.31}$	0/3	$17.65_{\pm 3.00}$	$2.33_{\pm 0.58}$	$56.30_{\pm 0.40}$	0/3
No state	$23.11_{\pm 3.02}$	$2.33_{\pm 0.58}$	$53.23_{\pm 0.32}$	0/3	$26.29_{\pm 2.50}$	$4.67_{\pm 0.58}$	$56.26_{\pm 0.37}$	0/3
$\alpha = 0.01$	$5.18_{\pm 3.33}$	$0.00_{\pm 0.00}$	$51.93_{\pm 0.49}$	0/3	$3.81_{\pm 2.50}$	$0.00_{\pm 0.00}$	$55.83_{\pm 0.49}$	0/3
$\alpha = 0.20$	$1.60_{\pm 0.60}$	$0.00_{\pm 0.00}$	$49.17_{\pm 0.21}$	0/3	$1.79_{\pm 0.73}$	$0.00_{\pm 0.00}$	$52.73_{\pm 0.82}$	0/3
$\alpha = 0.50$	$2.21_{\pm 1.26}$	$0.00_{\pm 0.00}$	$48.73_{\pm 0.90}$	0/3	$1.74_{\pm 0.95}$	$0.00_{\pm 0.00}$	$52.40_{\pm 0.92}$	0/3
All-coordinate zone	$100.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$15.40_{\pm 0.00}$	3/3	$100.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$9.49_{\pm 0.04}$	3/3
EMA, $\beta = 0.9$	$1.45_{\pm 0.47}$	$0.00_{\pm 0.00}$	$51.03_{\pm 0.29}$	0/3	$1.09_{\pm 0.42}$	$0.00_{\pm 0.00}$	$54.47_{\pm 1.04}$	0/3
Forget-only	$48.59_{\pm 11.16}$	$2.67_{\pm 1.15}$	$17.47_{\pm 4.91}$	3/3	$30.81_{\pm 10.69}$	$2.33_{\pm 0.58}$	$22.52_{\pm 2.91}$	3/3
Forget/retain ratio	$37.16_{\pm 17.85}$	$3.33_{\pm 0.58}$	$28.87_{\pm 2.80}$	2/3	$24.34_{\pm 12.15}$	$2.67_{\pm 0.58}$	$27.82_{\pm 5.66}$	2/3

Coefficients. Table A13 varies the regularizer weight and forget mass separately, keeping all other coefficients at the FAC setting. On non-IID data, $\lambda \in \{0.1, 1, 10\}$ leaves 1.33–3.00 sources, while $\lambda \in \{50, 100, 200\}$ leaves none, with R-Acc ranging from 50.30% to 50.83%. The IID sweep has an exception at $\lambda = 200$, which leaves one source in one seed. Every tested forget mass $\omega_{\mathrm{f}} \in \{1, 3, 5, 10\}$ leaves no survivor on either partition. On non-IID data, increasing it from 1 to 10 lowers worst-source ASR from 2.82% to 1.22% and R-Acc from 51.00% to 49.63%. These sweeps describe the CNN setting; they do not establish the same coefficient range for the larger backbones.

Table A13: **Coefficient sweeps at** $K = 5$ (CNN; $a_{\pm b}$ over three seeds, %). Each partition uses its own matched reference. No seed collapses.

Variant	non-IID			IID		
	ASR↓	Surv↓	R-Acc↑	ASR↓	Surv↓	R-Acc↑
$\lambda = 0.1$	$64.10_{\pm 15.29}$	$3.00_{\pm 0.00}$	$47.97_{\pm 2.55}$	$63.34_{\pm 8.08}$	$2.00_{\pm 1.00}$	$48.34_{\pm 3.80}$
$\lambda = 1$	$67.84_{\pm 16.54}$	$2.67_{\pm 0.58}$	$47.57_{\pm 2.81}$	$73.92_{\pm 10.08}$	$1.67_{\pm 0.58}$	$50.21_{\pm 2.79}$
$\lambda = 10$	$85.05_{\pm 19.35}$	$1.33_{\pm 0.58}$	$46.27_{\pm 1.45}$	$94.15_{\pm 5.09}$	$1.00_{\pm 0.00}$	$49.05_{\pm 1.33}$
$\lambda = 50$	$0.99_{\pm 0.35}$	$0.00_{\pm 0.00}$	$50.30_{\pm 0.79}$	$0.79_{\pm 0.12}$	$0.00_{\pm 0.00}$	$53.14_{\pm 1.04}$
$\lambda = 100, \omega_f = 3$ (FAC)	$1.98_{\pm 1.38}$	$0.00_{\pm 0.00}$	$50.83_{\pm 0.32}$	$1.94_{\pm 0.91}$	$0.00_{\pm 0.00}$	$54.36_{\pm 0.49}$
$\lambda = 200$	$2.52_{\pm 1.42}$	$0.00_{\pm 0.00}$	$50.80_{\pm 0.36}$	$4.89_{\pm 5.10}$	$0.33_{\pm 0.58}$	$54.31_{\pm 0.47}$
$\omega_f = 1$	$2.82_{\pm 1.39}$	$0.00_{\pm 0.00}$	$51.00_{\pm 0.46}$	$4.84_{\pm 3.40}$	$0.00_{\pm 0.00}$	$54.62_{\pm 0.43}$
$\omega_f = 5$	$1.60_{\pm 0.82}$	$0.00_{\pm 0.00}$	$50.43_{\pm 0.46}$	$1.43_{\pm 0.79}$	$0.00_{\pm 0.00}$	$53.92_{\pm 0.74}$
$\omega_f = 10$	$1.22_{\pm 0.70}$	$0.00_{\pm 0.00}$	$49.63_{\pm 0.49}$	$0.51_{\pm 0.13}$	$0.00_{\pm 0.00}$	$52.72_{\pm 1.05}$
Per-coordinate mass	$2.52_{\pm 1.72}$	$0.00_{\pm 0.00}$	$50.87_{\pm 0.60}$	$2.96_{\pm 1.39}$	$0.00_{\pm 0.00}$	$54.34_{\pm 0.48}$

Position assignment before recovery. We pair three original FAC runs with three runs that permute their sensitivity scores on CNN, CIFAR-10, non-IID, at $K = 5$. All six runs start from the matched poisoned checkpoints and stop after 50 unlearning rounds. The initial learning rate is 0.01 with a per-round factor of 0.999; $\lambda = 100$, $\omega_{\mathrm{f}} = 3$, $\alpha = 0.05$, $\rho_{\mathrm{p}} = 0.05$, $\rho = 10$, $\delta = 0.1$, and $\varepsilon = 10^{-12}$. Local correction follows Algorithm 1, including the first-round warm start, and remains enabled thereafter.

For each seed, the original run computes and saves the complete sensitivity score vector at every round. Its paired run replays that same sequence after a fixed random permutation within each weight tensor and, separately, within each bias tensor. The permutation is held fixed across rounds and generated independently of training randomness; the parameters and pre-unlearning anchor are not permuted. Scores from the paired run's own updates do not replace the reference sequence. Thus the two runs have identical per-tensor score distributions, positive-score counts, and selected-coordinate counts at each corresponding round; their model trajectories evolve separately.

The three training seeds each use one independent permutation. Initial models and first-round client updates match within each pair. Saving and reloading the original scores reproduces every reference server update exactly, and the score distributions and zone sizes match across all 150 paired rounds. Table A14 gives the source-level and endpoint results using the same Retrain thresholds as the CNN source-scaling experiment. In every permuted run, checker and diagonal ASR exceeds 92%. The original assignment gives lower worst-source ASR and higher retained accuracy. This diagnostic tests the score-to-coordinate correspondence while holding score distributions and zone sizes fixed; it does not equate the realized update magnitudes.

Fixed-map replay. To separate the state from the map it induces, we rerun the three state variants of Table A10 at $K = 5$ for 50 unlearning rounds while replaying, round by round, the sensitivity scores saved from the reference FAC runs of Table A14; the first round, which has no correction, matches the reference exactly. Table A15 compares each variant under its own map and under FAC's map. Without the state, the variant's own map leaves 2.33 sources, while FAC's map removes every source but lowers R-Acc to 36.47%, 13.4 points below FAC; the map therefore does not substitute for the state. The shared-target and no-proximal variants are survivor-free at round 50 under either map, so the survivors they leave at round 100 (Table A10) arise during recovery.

Table A14: **Position assignment at round 50, before recovery** (CNN, non-IID, $K = 5$; %; columns are seeds 1–3). Source rows give each source's ASR against its unchanged threshold τ_i; the last rows give worst-source ASR, the number of surviving sources, and R-Acc.

Source	τ_i	Original FAC			Permuted in tensor		
		1	2	3	1	2	3
Square	9.020	1.149	1.370	0.230	0.460	0.460	0.000
Cross	4.891	0.000	0.000	0.000	0.000	0.000	0.000
Checker	6.811	0.000	0.000	0.000	93.886	92.304	93.458
Diagonal	5.804	0.000	0.000	0.000	92.204	93.695	94.813
Anti-diagonal	2.721	0.000	0.000	0.000	0.243	0.000	0.000
Worst-source ASR	–	1.149	1.370	0.230	93.886	93.695	94.813
Surv	–	0	0	0	2	2	2
R-Acc	–	49.899	49.372	50.289	48.942	48.569	48.282

Table A15: **Fixed-map replay** (CNN, CIFAR-10, non-IID, $K = 5$; round 50, before recovery; $a_{\pm b}$ over three seeds, %). Own: the variant's own sensitivity map (round-50 records of the Table A10 runs); FAC's: the reference FAC scores replayed every round.

Variant	Map	Worst ASR↓	Surv↓	R-Acc↑
FAC	own	$0.92_{\pm 0.60}$	0.00	$49.85_{\pm 0.46}$
Shared target z	own	$1.60_{\pm 0.99}$	0.00	$48.87_{\pm 0.29}$
	FAC's	$1.22_{\pm 0.70}$	0.00	$50.01_{\pm 0.39}$
No proximal step	own	$1.83_{\pm 1.21}$	0.00	$48.87_{\pm 0.12}$
	FAC's	$1.22_{\pm 0.70}$	0.00	$50.05_{\pm 0.42}$
No state ($u = 0$)	own	$20.05_{\pm 2.51}$	2.33	$52.93_{\pm 0.51}$
	FAC's	$0.08_{\pm 0.13}$	0.00	$36.47_{\pm 0.19}$

F Source-Wise Analysis

This section supports Section 4.4.

Dynamics. Figure A1 follows every Table 1 method at $K = 5$ through 100 rounds. GA, Sequential, FedOSD, FedCARE and FedRazor lower mean-source ASR to about 20%, while their worst-source ASR remains high throughout the recorded trajectory. Thus a low mean does not imply that all sources have been removed. FAC's mean worst-source ASR falls below 1% by round 5 and ends at 1.98%.

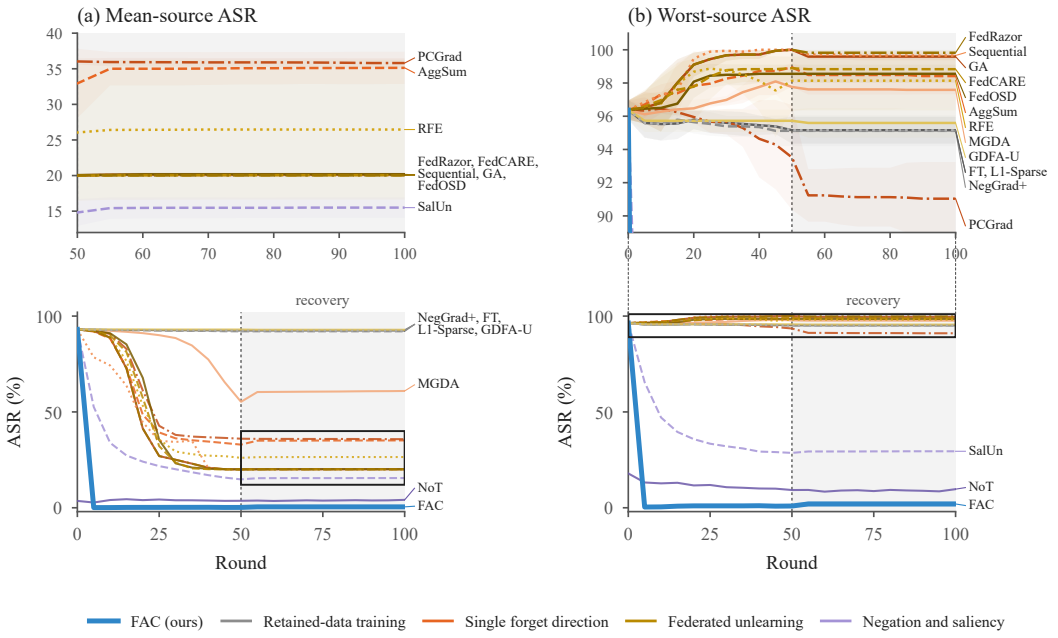

Figure A1: **Deletion dynamics** (CNN, non-IID, $K = 5$). (a) Mean-source and (b) worst-source ASR every five rounds; lines are means over three seeds, and each baseline has a shade of its group's hue (groups of Section 4). The upper panels enlarge the boxed regions and show one-standard-deviation bands; every curve is labelled at its end. The shaded half is the retain-only recovery of the staged methods and of FAC; NoT trains on the retained clients throughout. GA, Sequential, FedOSD, FedCARE and FedRazor retain high worst-source ASR throughout the recorded trajectory.

Aggregate diagnostics. Figure A2 compares four run-level diagnostics with the survivor count over the 192 approximate-method runs of Table 1. Mean-source ASR and the deletion residual correlate with Surv but overlap between runs with and without a survivor at small values; D2R-KL and the membership-inference AUC barely separate them.

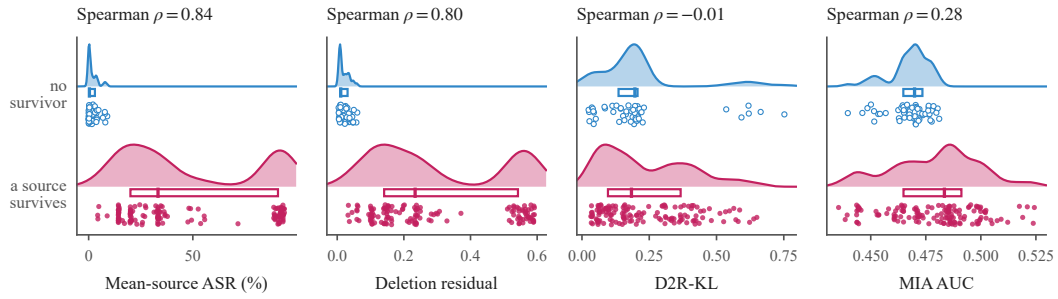

Figure A2: **Aggregate diagnostics against surviving sources** over the 192 approximate-method runs of Table 1, split by whether any source survives; each group is shown as a density, a quartile box with the median, and the individual runs. ρ is the Spearman correlation with Surv. No diagnostic separates the two groups.

Endpoint comparison across source counts. Figure A3 compares mean-source and worst-source ASR after the complete 100-round procedure for all sixteen methods at $K \in \{3, 5, 7\}$, with three seeds per setting. Of these 144 runs, 83 (58%) have a worst-minus-mean gap of at least 20 percentage points, and 31 of the 48 runs below 20% mean ASR retain a source. Source identity also varies: in 18 of the 48 method-by-K settings, the worst source differs across seeds. For example, NoT's mean Surv of 1.00 at $K = 3$ combines runs retaining `square`→`0`, no source, and both `square`→`0`

and `checker→2`. Neither a mean ASR nor an averaged survivor count identifies which source remains.

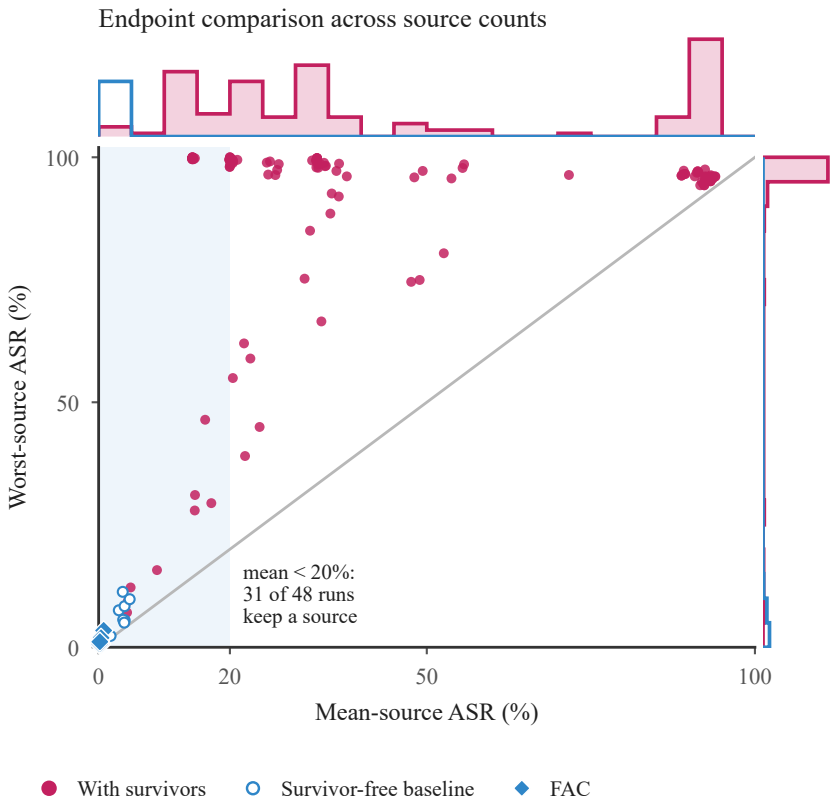

Figure A3: **Mean-source and worst-source ASR at the endpoint.** CNN, CIFAR-10, non-IID; sixteen methods, $K \in \{3, 5, 7\}$ and three seeds, totaling 144 runs. Each point is one run after 100 rounds. Filled pink circles denote runs with surviving sources; open blue circles denote survivor-free baseline runs; blue diamonds denote FAC. The diagonal marks equality. In the shaded region (mean ASR below 20%), 31 of 48 runs retain a source. Marginal histograms show raw run counts in five-percentage-point bins: pink for runs with survivors and blue for survivor-free runs, including FAC. Survival uses the source-specific thresholds of Appendix B.

Coverage of every method. Figure A4 places each of the 15 source instances of every Table 1 method at $K = 5$ by its mean coverage. In thirteen methods, one source per run has non-positive coverage, and it survives in all 39 runs; FT, NegGrad+, L1-Sparse and GDFA-U also keep their positively covered sources, since they barely change the model. NoT's only non-positively covered source is deleted, SalUn keeps all 15 sources positively covered yet leaves 11, and FAC keeps all 15 positively covered and deletes them. At $K = 3$ and $K = 7$, all 14 non-positively covered sources of FedOSD and GDFA-U survive, and FAC keeps every source positively covered.

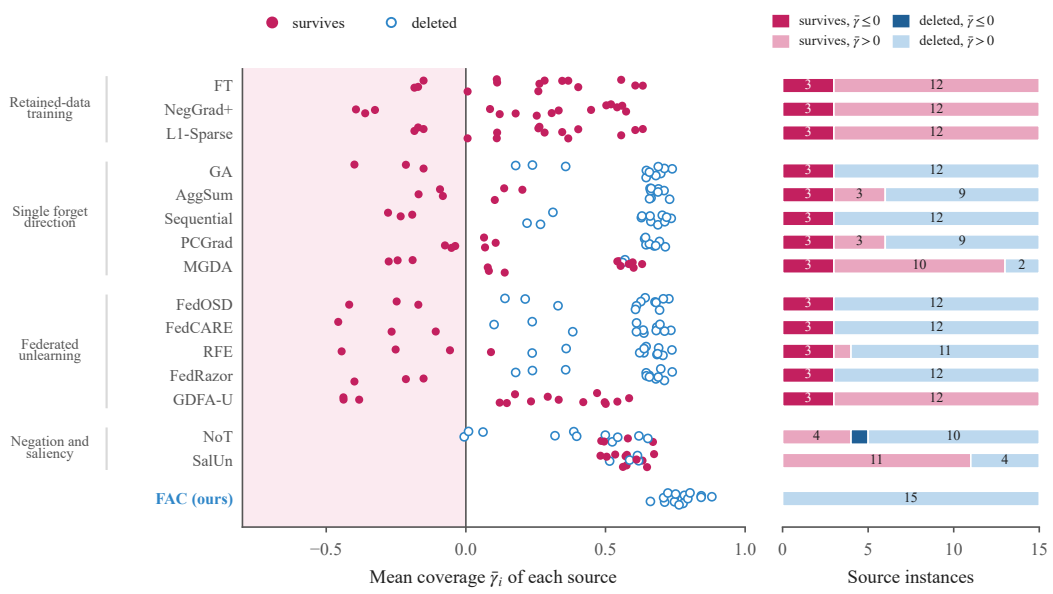

Figure A4: **Coverage and survival of every source** ($K = 5$). Left: each dot is one source in one run, placed at its coverage averaged over the recorded rounds (for NoT, its 100 retained-training rounds); filled dots survive. Right: the same 15 instances per method, counted by outcome and coverage sign. Every non-positively covered source survives except one of NoT's, and positive coverage does not imply deletion (SalUn).

G Recovery Schedules and Clean Re-entry

Recovery on large backbones. The CNN schedule, 50 unlearning rounds followed by 50 retain-only rounds that hold the attack zone at zero, lowers FAC's R-Acc on ViT-S/16 to 12.67–21.97%, against 83.07–95.47% without recovery (Table A16). This schedule comparison changes both the stage allocation and projection.

Table A16: **Recovery schedule on ViT-S/16** ($K = 3$; $a_{\pm b}$ over three seeds; ASR and R-Acc in %). All recovery-schedule runs collapse; their deletion scores are not ranked. The schedules differ in both stage allocation and projection.

Dataset	100 unlearning			50 unlearning + 50 recovery		
	ASR	Surv	R-Acc	ASR	Surv	R-Acc
CIFAR-10	$3.20_{\pm 5.42}$	$0.33_{\pm 0.58}$	$95.47_{\pm 0.47}$	$37.10_{\pm 35.08}$	$1.00_{\pm 0.00}$	$12.67_{\pm 2.28}$
CIFAR-100	$0.35_{\pm 0.46}$	$0.00_{\pm 0.00}$	$83.07_{\pm 0.15}$	$0.31_{\pm 0.54}$	$0.00_{\pm 0.00}$	$12.70_{\pm 11.71}$
SVHN	$0.94_{\pm 1.48}$	$0.00_{\pm 0.00}$	$93.60_{\pm 0.10}$	$63.77_{\pm 47.69}$	$1.00_{\pm 0.00}$	$21.97_{\pm 0.78}$

Rejoining clients. We further evaluate the returned models after the deleted clients rejoin with clean data and all clients continue FedAvg training for 20 rounds. Table A17 reports worst-source ASR and retained accuracy during this continuation for two CNN requests, which delete nine or twenty-one clients (the latter 70% of the federation), and for the three-source NF-ResNet-18 endpoints on CIFAR-10 and CIFAR-100, each with its own matched endpoints and Retrain reference. On NF-ResNet-18/SVHN, retained accuracy falls sharply after clean re-entry for most models, including Retrain (6.78% at the first round in all three seeds), so attack success there is not interpreted as persistence.

Table A17: **Clean re-entry.** Three-seed means in percent; r counts continuation rounds, and $r = 0$ repeats the corresponding endpoint values.

Method	Worst ASR (%)			R-Acc (%)	
	$r = 0$	$r = 5$	$r = 20$	$r = 0$	$r = 20$
CNN, non-IID, $K = 3$					
FAC	0.61	2.58	5.19	53.57	53.64
NoT	9.45	9.00	9.07	52.20	51.92
SalUn	4.32	59.16	83.87	46.70	55.06
FedOSD	98.15	97.49	96.32	55.60	55.40
CNN, non-IID, $K = 7$					
FAC	1.60	3.13	6.70	47.17	48.17
NoT	9.83	9.90	9.93	45.47	45.55
SalUn	50.16	70.56	79.05	45.60	48.40
FedOSD	99.67	98.62	97.70	48.77	48.34
NF-ResNet-18, CIFAR-10					
FAC	1.12	11.20	32.31	76.59	77.52
NoT	2.94	36.48	38.39	70.00	51.84
FedOSD	93.85	92.11	90.64	76.09	76.82
FT	95.98	92.49	91.05	76.99	76.88
NF-ResNet-18, CIFAR-100					
FAC	0.66	4.41	8.64	46.94	46.93
NoT	0.18	0.37	0.24	38.18	40.29
FedOSD	44.03	65.53	58.21	45.50	45.70
FT	69.18	70.07	62.48	47.11	45.77

H Additional Related Work

Federated unlearning. FedShard (Wen et al., 2026) studies efficiency and performance fairness across leaving clients. In vertical federated learning, where parties own different features, unlearning removes labels or party representations (Gu et al., 2026; Wu et al., 2026). Backdoor probes are a security-relevant test of residual client influence: ReMisVFU (Wu et al., 2026) evaluates client removal through backdoor attack success, and retaliatory attacks (Sheng et al., 2026) show that the unlearning procedure itself can expose information or be manipulated to resist deletion. When several client groups encode distinct trigger–target associations, a low average attack success rate does not establish that every association has been removed.

Utility-preserving unlearning. Implicit gradient surgery (Zhou et al., 2026) optimizes forgetting subject to a bound on the increase of the retained loss, RFE (Cheng et al., 2026a) addresses retain–forget entanglement with an augmented-Lagrangian forgetting phase followed by projected recovery, OFMU (Asif & Mohammadi Amiri, 2026) uses a bilevel formulation that prioritizes removal while restoring retained utility, Sharp MinMax (Tang & Khanna, 2026) separates retain and forget signals through complementary sharpness objectives, and MU-Mis (Cheng et al., 2026b) suppresses the sensitivity associated with individual sample contributions without accessing the remaining data. These methods constrain unlearning by parameter structure, retained utility, or the relation between retained and forgotten information, and most set one forget objective against one retain objective.

Evaluation. LUCID (Jiang et al., 2026) shows that audit coverage and retain–forget redundancy can materially change conclusions about removal. With the procedure-level and sample-level checks of Section 2, these lines verify different units: that a prescribed procedure was executed, completeness for individual samples, or markers that one departing client planted itself. The source-wise evaluation of Section 3.4 tests the unit that a multi-source request adds.