The Holographic Gradient: On the Privacy Limits of Split Federated Learning

Minh K. Quan*
School of Engineering
Deakin University, Australia

Pubudu N. Pathirana
School of Engineering
Deakin University, Australia

Abstract

Split Federated Learning (SFL) is built on a reassuring premise: keep the labels on the client, and the server learns nothing sensitive. We show this premise is wrong. The gradient the server receives to train its model is not an opaque update; it encodes the private labels in a way that is mathematically unavoidable. We formalise this vulnerability, prove information-theoretic lower bounds on the leakage, and show why existing defences fail: they cannot distinguish the gradient's optimization signal from its label leakage. We prove these two components occupy different subspaces near convergence, a consequence of Neural Collapse, and build on this to propose Rate-Limited SFL (RL-SFL), which suppresses the leakage subspace by transmitting only the most informative gradient coordinates. RL-SFL comes with convergence guarantees and works well under near-homogeneous data distributions, but degrades under high client heterogeneity and provides no privacy at all under extreme skew; it also carries no formal differential privacy certificate. On CIFAR-10/100 and BloodMNIST, RL-SFL achieves substantially lower attack accuracy than standard DP-SGD and variational baselines at comparable utility, with reduced communication cost on bandwidth-constrained links.

1 Introduction

Deploying large neural networks on edge devices—phones, sensors, medical instruments—is impractical under standard Federated Learning due to memory constraints McMahan et al. [2017]. Split Federated Learning (SFL) Thapa et al. [2022] addresses this by partitioning the model: the client runs a lightweight head and tail while offloading the heavy body to a server, keeping raw data (x, y) strictly local (Figure 4).

The implicit promise is "Privacy by Design": because the server never sees raw labels, privacy is assumed to be structurally guaranteed. We show this assumption is wrong. The gradient the server must receive to update its parameters carries the label residual $(\hat{Y} - Y)$ in a form that is both mathematically unavoidable and recoverable—even without access to raw data. As shown in Figure 1, server-received gradients cluster by class despite withheld labels. We call this the *Holographic Gradient Property*.

This raises a precise question: *can the optimization signal and the label leakage be separated inside the gradient, and at what cost?*

Main Contributions. (1) We prove that the cut-layer gradient is an injective map of the label residual (**Holographic Gradient Property**, Prop. 4.1) and derive an architecture-aware information-theoretic lower bound showing that any accurate SFL system must leak label information; the generic bound (Thm. 4.2) is tightened by an SFL-specific surplus depending on W's singular value spectrum (**Architecture-Aware Leakage Bound**, Prop. 4.3). **(2)** We prove that the optimization signal and

*Corresponding author. m.quan@deakin.edu.au

Preprint.

leakage occupy different gradient subspaces near convergence—a consequence of Neural Collapse (Prop. 4.8). **(3)** Exploiting this structure, we introduce **RL-SFL**: transmit only the top-k gradient coordinates, which is optimal within the coordinate-sparse Information Bottleneck class (Cor. 5.4) and preserves the $O(1/T)$ SGD convergence rate (Thm. 5.7). RL-SFL provides no formal (ϵ, δ)-DP certificate and fails under extreme data heterogeneity; see Section 7.

2 Background and Problem Formulation

2.1 System Model: U-Shaped Split Federated Learning

In U-Shaped SFL Thapa et al. [2022], the client retains (x, y) and runs a lightweight head f_{θ_c} and tail k_{θ_t}; the server runs the body g_ϕ. The server sees only the forward activation $z = f_{\theta_c}(x)$ and the cut-layer gradient $\mathbf{g} \triangleq \nabla_h \mathcal{L}$:

$$\min_{\theta_c, \phi, \theta_t} \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \ell(k_{\theta_t}(g_\phi(f_{\theta_c}(x))), y)\,. \tag{1}$$

2.2 The "Blind Server" Fallacy

Keeping labels on the client has led to a widespread but mistaken assumption that SFL is privacy-preserving Ceballos et al. [2020]. In practice the server is not blind: the gradient $\mathbf{g} = \nabla_h \ell(\hat{y}, y)$ it receives is a functional derivative of the loss, which under cross-entropy contains the prediction error $(\hat{y} - y)$—a sufficient statistic for y.

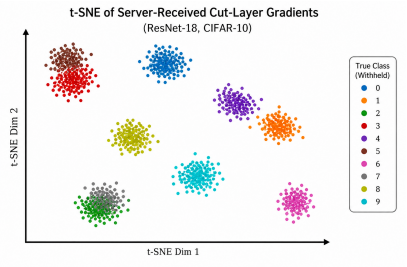


Figure 1: **Holographic Gradient (Prop. 4.1).** t-SNE of server-received gradients (ResNet-18, CIFAR-10): gradients cluster by class despite withheld labels. Leakage collapses under RL-SFL (Fig. 7).

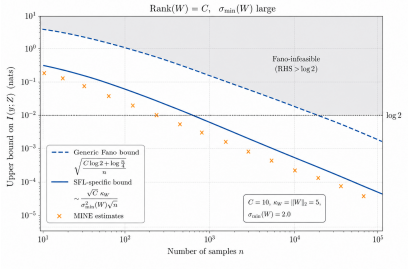


Figure 2: **Architecture-Aware Leakage Bound (Prop. 4.3).** SFL-specific bound (solid) tightens the generic Fano bound (dashed) when $\mathrm{rank}(W) = C$ and $\sigma_{\min}(W)$ is large. Shaded: infeasible regimes. MINE estimates (crosses) satisfy the bound.

Threat Model. We adopt the honest-but-curious model Wei et al. [2020]. The server receives per-sample cut-layer gradients before any cross-sample aggregation, and reconstructs labels $\hat{Y}_{\mathrm{adv}}$ using the **SLeak** adversary Xu et al. [2026] (App. H).

2.3 The Defensive Landscape

Existing defenses each carry a fundamental cost (Table 1). Cryptographic approaches (MPC/FHE) provide strong guarantees but impose $>$ 10–100$\times$ latency Khan and Michalas [2025]. DP-SGD injects isotropic noise that destroys the low-rank optimization signal at the same rate as leakage Zhang et al. [2023]. Variational methods require unstable auxiliary estimators Deng et al. [2023]. Extended discussion is in Appendix B. RL-SFL exploits a different angle: the optimization signal and leakage occupy different gradient subspaces, so a hard sparsity constraint on the backward channel can suppress leakage while preserving the descent direction.

3 Setting and Preliminaries

Notation. $Y \in \{e_1, \ldots, e_C\} \subset \Delta^{C-1}$ is the one-hot label; $\hat{Y} = k_{\theta_t}(h)$ is the model output; $\hat{Y}_{\mathrm{adv}}$ is the adversary's reconstruction; $\mathbf{g}$ is the cut-layer gradient; $\tilde{\mathbf{g}}$ its sparse approximation; $h_b(\cdot)$ binary entropy. Full notation in Appendix A.

Table 1: **Privacy-Efficiency-Utility Trilemma.** Formal Cert.: composable (ϵ, δ)-DP guarantee. Emp. Resilience: practical attack suppression. RL-SFL has strong empirical resilience but no formal certificate; see Section 7.

PARADIGM	METHOD	MECHANISM	CORE DRAWBACK	COMM. COST	FORMAL CERT.	EMP. RESILIENCE
None	VANILLA SFL THAPA ET AL. [2022]	MODEL PARTITIONING	HOLOGRAPHIC LEAKAGE	$1\times$	NONE	NONE
Crypto.	SPLITML SHEYBANI ET AL. [2025]	FHE + DP	COMPUTE WALL ($> 100\times$)	HIGH	(ϵ, δ)-DP	STRONG
	SAFESPLIT RIEGER ET AL. [2025]	MPC	LATENCY TAX ($> 10\times$)	HIGH	MPC-SECURE	STRONG
Heuristic	DP-SFL ZHANG ET AL. [2023]	ISOTROPIC NOISE	UTILITY COLLAPSE AT MEANINGFUL ϵ	$1\times$	(ϵ, δ)-DP	WEAK–MODERATE
Variational	VIB-SPLIT DENG ET AL. [2023]	INFO. BOTTLENECK	OPTIM. COMPLEXITY	$1\times$	NONE	MODERATE
GEOMETRIC	**RL-SFL (OURS)**	SPECTRAL FILTER	FULL-RANK FAILURE (SEC. 7)	$< 1\times$	NONE	STRONG

The protocol induces the Markov chain $X \to Z \to H \to \hat{Y}$. The server's view is Z and $\mathbf{G} := \nabla_h \mathcal{L}(h, y)|_{h=g_\phi(f_{\theta_c}(X))} \in \mathbb{R}^{d_h}$.

Assumption 3.1 (Canonical Link & Duality)**.** The client tail k_{θ_h} is parameterized by $W \in \mathbb{R}^{d_h \times C}$ with logits $\eta = W^\top h$, output $\hat{y} = \nabla\psi(\eta)$, and ℓ the Bregman divergence. Consequently, $\nabla_\eta \ell = \hat{Y} - Y$.

Assumption 3.2 (Optimization Regularity)**.** (a) $F(\theta)$ satisfies the Polyak-Łojasiewicz (PL) condition with $\mu > 0$: $\|\nabla F(\theta)\|_2^2 \geq 2\mu(F(\theta) - F(\theta^*))$. (b) F is L-smooth. (c) Stochastic cut-layer gradient has bounded variance σ_ξ^2. (d) Server body Jacobian is bounded: $\sup_z \|\nabla_z g_\phi(z)\|_F \leq G$. The PL condition is strictly weaker than strong convexity and holds for over-parameterized deep networks near any SGD-reachable local minimum Liu et al. [2022a]; see App. I.

4 Theoretical Analysis

4.1 The Holographic Gradient Property

Proposition 4.1 (Holographic Gradient Embedding)**.** *Under Assumption 3.1, the cut-layer gradient is:*

$$\mathbf{G} = W(\hat{Y} - Y), \tag{2}$$

where $\hat{Y} = \nabla\psi(W^\top h)$ is the softmax output of the canonical link. If W has rank C, the mapping $y \mapsto \mathbf{g}$ is injective for fixed H.

Proof sketch. By exponential family duality [Amari, 2016], $\nabla_\eta \ell = \hat{Y} - Y$; chain rule gives $\mathbf{G} = W(\hat{Y} - Y)$. Injectivity follows from rank-nullity. Full proof in App. K. □

4.2 Information-Theoretic Limits: Generic and SFL-Specific Bounds

Theorem 4.2 (Generic Leakage Lower Bound)**.** *For any adversary achieving $\mathbb{P}(\hat{Y}_{\text{adv}} \neq Y) \leq \epsilon$ from the gradient stream:*

$$I(Y; \mathbf{G}) \geq H(Y) - h_b(\epsilon) - \epsilon \log(C - 1). \tag{3}$$

(Proof via DPI + Fano's inequality; App. L.)

Proposition 4.3 (Architecture-Aware SFL Leakage Surplus)**.** *Under Assumption 3.1, let $r_W = \text{rank}(W) \leq C$ and $\sigma_{\min}(W)$ the smallest nonzero singular value of W. For any adversary achieving reconstruction error $\mathbb{P}(\hat{Y}_{\text{adv}} \neq Y) \leq \epsilon$ from the SFL gradient stream, and under a Gaussian mini-batch noise model $\mathbf{G} = W(\hat{Y} - Y) + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{d_h})$:*

$$I(Y; \mathbf{G} \mid Z) \;\geq\; \underbrace{H(Y) - h_b(\epsilon) - \epsilon \log(C-1)}_{\textit{Generic Fano bound}} \;+\; \underbrace{\mathcal{S}(\sigma_{\min}(W),\, r_W,\, \sigma^2)}_{\textit{SFL surplus} \geq 0}, \tag{4}$$

where the surplus is:

$$\mathcal{S}(\sigma_{\min}(W),\, r_W,\, \sigma^2) \;:=\; \frac{r_W}{2}\left[\log\left(1 + \frac{\gamma\sigma_{\min}^2(W)}{\sigma^2}\right) - \log\left(1 + \frac{\gamma}{\sigma^2\sigma_{\min}^2(W)}\right)\right]_+, \tag{5}$$

with $\gamma > 0$ the label-residual variance, and $[\cdot]_+ = \max(\cdot, 0)$. $\mathcal{S} > 0$ when $\sigma_{\min}^2(W) > 1$ and vanishes at $\sigma_{\min}(W) = 1$.

Scope. Conditioning on Z makes $Y \to \Delta \to \mathbf{G}|Z$ Markov (for fixed Z, $\Delta = \hat{Y}(H) - Y$ depends only on Y), permitting DPI. Since $I(Y;(\mathbf{G},Z)) \geq I(Y;\mathbf{G}|Z)$, this is a conservative bound on total leakage. The Gaussian noise model approximates mini-batch stochasticity (numerical check in App. E). Thm. 4.2 remains the architecture-independent impossibility result. (Proof in App. E.)

Remark 4.4 (Why this bound is SFL-specific). Theorem 4.2 holds for any gradient-based multi-class algorithm via DPI + Fano alone. Proposition 4.3 is SFL-specific: it uses $\mathbf{G} = W(\hat{Y} - Y)$ to show the adversary observes Y through a d_h-dimensional over-complete linear channel rather than just the C-dimensional natural-parameter channel. The surplus $\mathcal{S}$ captures this extra capacity when $C \ll d_h$ and W is well-conditioned. Figure 2 illustrates the tightening.

Remark 4.5 (Reparameterization dependence of $\mathcal{S}$ — important scope limitation). $\mathcal{S}$ depends on $\sigma_{\min}(W)$, which is not reparameterization-invariant: rescaling $W \to \alpha W$ and $h \to h/\alpha$ preserves the model but changes $\sigma_{\min}(W)$ by α. The surplus therefore characterises leakage under a specific trained parameterization, not an intrinsic architectural property. It is most meaningful under a fixed-scale convention (e.g., weight normalization). *When $\sigma_{\min}(W)$ cannot be verified at a canonical scale, practitioners should fall back to the generic Fano bound (Thm. 4.2), which is free of this limitation.*

Corollary 4.6 (The Price of Reliability)**.** *As $\epsilon \to 0^+$: $I(Y;\mathbf{G}|Z) \geq H(Y) + \mathcal{S}(\sigma_{\min}(W), r_W, \sigma^2)$. Any accurate SFL system must leak at least full label entropy plus the architecture-dependent surplus. Per-step extension in Corollary 4.7.*

Corollary 4.7 (Per-Step Transmission Necessity)**.** *Let $\{\mathbf{G}_t\}_{t=1}^T$ denote the gradient stream. If $\mathcal{A}$ achieves $\mathbb{P}(\hat{Y}_{\text{adv}} \neq Y) \leq \epsilon$ after the complete stream:*

$$\frac{1}{T}\sum_{t=1}^{T} I(Y;\mathbf{G}_t \mid Z_t) \;\geq\; H(Y) - h_b(\epsilon) - \epsilon\log(C-1) + \frac{1}{T}\sum_{t=1}^{T}\mathcal{S}(\sigma_{\min}(W_t), r_W, \sigma^2). \tag{6}$$

The bound is conservative: $I(Y;(\mathbf{G}_t, Z_t)) \geq I(Y;\mathbf{G}_t|Z_t)$. Cumulative leakage grows linearly with training; the privacy-convergence tradeoff is formalized in Corollary G.2 (App. G).

4.3 From Assumption to Theorem: Neural Collapse Implies Spectral Separation

Proposition 4.8 (Neural Collapse Implies Fisher Spectral Separation)**.** *Let the training regime be in a near-convergence state satisfying the following conditions: (a) the within-class feature covariance satisfies $\lambda_{\max}(\Sigma_W) \leq \delta_W$ for some small $\delta_W > 0$; (b) the between-class covariance Σ_B has rank $C-1$ with smallest positive eigenvalue $\lambda_B > 0$ (guaranteed by Neural Collapse [Papyan et al., 2020]). Then the gradient covariance $\mathbf{F} = \mathbb{E}[\mathbf{g}\mathbf{g}^\top]$ satisfies Fisher Spectral Separation (Assumption 5.1) with spectral gap:*

$$\Delta\lambda = \lambda_r(\mathbf{F}) - \lambda_{r+1}(\mathbf{F}) \;\geq\; \sigma^2_{\min}(W)\,\lambda_B - \|W\|_F^2 \cdot \delta_W, \tag{7}$$

where $r = \operatorname{rank}(W)$. The gap $\Delta\lambda$ is strictly positive whenever $\delta_W < \frac{\sigma^2_{\min}(W)\lambda_B}{\|W\|_F^2}$, a condition that holds throughout the near-convergence phase of Neural Collapse training (where $\delta_W \to 0$ continuously while gradients remain nonzero).

Scope. *This applies to the near-convergence phase where gradients are nonzero. At exact convergence $\mathbf{g} = 0$ and $\mathbf{F}$ vanishes, so no spectral gap exists. The separation is a property of training dynamics, not the terminal state.*

Proof sketch. Under the stated conditions, $\mathbf{F} = W(\Sigma_B + \Sigma_W)W^\top$. By Weyl's interlacing inequality: $\lambda_r(\mathbf{F}) - \lambda_{r+1}(\mathbf{F}) \geq \sigma^2_{\min}(W)\lambda_B - \|W\|_F^2\delta_W > 0$ under the stated condition on δ_W. Full derivation in App. F. □

Remark 4.9. Papyan et al. [2020] established Neural Collapse empirically for ResNets and VGGs; ViTs were not part of that study. Subsequent work has extended the analysis to broader architectures [Zhu et al., 2021, Han et al., 2021]. Our result applies in the near-convergence regime for the ResNet and MobileNetV2 architectures evaluated here. Whether Fisher Spectral Separation holds for transformer cut layers is an open question.

5 Method: Spectral Rate-Limited SFL

We model the defense as a projection $\Phi\colon T^*\mathcal{H} \to T^*\mathcal{H}$ and formulate it as the IB Lagrangian Tishby et al. [1999]:

$$\max_{\tilde{\mathbf{G}}} \quad I(\theta^*; \tilde{\mathbf{G}}) - \beta\, I(Y; \tilde{\mathbf{G}}). \tag{8}$$

Interpretation. θ^* is a fixed but unknown target; $I(\theta^*; \tilde{\mathbf{G}})$ measures how much of the optimal signal is recoverable from the transmitted gradient. Lemma J.1 (App. J) reduces this to $\|\mu_{\theta^*} - \mathcal{P}_{\mathcal{S}}(\tilde{\mathbf{g}})\|_2^2$, eliminating the need for a prior over θ^*. Under FSS this program reduces to classical hard thresholding (below).

5.1 Geometric Rationale: Fisher Spectral Separation

Assumption 5.1 (Fisher Spectral Separation)**.** The eigendecomposition $\mathbf{F} = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ satisfies: (1) Signal subspace $\mathcal{S} = \mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ contains the descent direction, $\lambda_r = \Theta(d_h)$; (2) Noise subspace $\mathcal{N} = \mathrm{span}\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_{d_h}\}$ contains the isotropic leakage, $\lambda_{r+1} = o(1)$. By Proposition 4.8, this holds in the near-convergence phase of Neural Collapse training under mild weight-conditioning.

Assumption 5.2 (Gradient Source Model)**.** Under Assumption 5.1, the gradient decomposes as $\mathbf{g} = \mathbf{g}_{\mathcal{S}} + \mathbf{g}_{\mathcal{N}}$ with $\mathbf{g}_{\mathcal{S}} \perp \mathbf{g}_{\mathcal{N}}$. Here $\mathbf{g}_{\mathcal{S}} \sim \mathcal{N}(\mu_{\theta^*}, \Lambda_{\mathcal{S}})$ with $\Lambda_{\mathcal{S}} = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$, $\lambda_i = \Theta(d_h)$. The noise component $\mathbf{g}_{\mathcal{N}} \sim \mathcal{N}(0, \sigma_{\mathcal{N}}^2 \mathbf{I}_{d_h - r})$ with $\sigma_{\mathcal{N}}^2 = o(1)$. *Modeling note.* The Gaussian form enables a closed-form IB reduction and is not claimed as established from Lin et al. [2018]. It can fail for heavy-tailed losses or small batches; empirical support comes from the near-identical Top-k/Oracle SVD curves (Fig. 9). See Section 7.

5.2 IB Reduction and Hard Thresholding Solution

Under Assumptions 5.1–5.2, Eq. (8) reduces to (formal derivation in App. J):

$$\tilde{\mathbf{g}}^* = \arg\min_{\mathbf{z} \in \mathbb{R}^{d_h}} \|\mathbf{g} - \mathbf{z}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{z}\|_0 \le k. \tag{9}$$

The analytic solution is the **Hard Thresholding Operator**: $\mathcal{T}_k(\mathbf{g}) = \mathbf{g} \odot \mathbb{I}(|\mathbf{g}| \ge \tau_k)$, where $\tau_k = |\mathbf{g}|_{(k)}$. This bounds capacity to $C_{\max} \approx k \log \frac{d_h}{k}$ bits (support entropy only; magnitude leakage is additional and uncertified—see Remark C.1 in App. C).

Remark 5.3 (Resolving the Dimensionality Paradox). Prop. 4.1 says each instantaneous gradient lies in the rank-C column space of W_t, which seems to contradict Assumption 5.1's full-dimensional covariance $\mathbf{F}$. The resolution is that W_t rotates continuously during training, so rank-C leakage accumulates across all d_h dimensions over the trajectory. The cumulative covariance is statistically high-rank even though each step is rank-C, and it is this diffuse high-rank structure that $\mathcal{T}_k$ suppresses.

Corollary 5.4 (IB Optimality within Coordinate-Sparse Class, $k \ge r$)**.** *Under Assumptions 5.1–5.2, for $k \ge r$, $\mathcal{T}_k$ achieves the IB Lagrangian to within additive gap $\frac{C_2}{\Delta\lambda}(2L(F(\theta_0) - F^*) + \sigma_\xi^2) + \mathcal{O}(\sigma_{\mathcal{N}}^2 d_h)$, vanishing as $\Delta\lambda \to \infty$. No other coordinate-sparse operator of the same cardinality achieves a smaller gap. For $k < r$, utility collapse is guaranteed (Prop. 5.6); IB optimality is vacuous there. The novel contribution is the IB-theoretic derivation connecting the classical best k-term approximation [Cohen et al., 2009] to the privacy-utility Lagrangian under FSS. (Proof in App. J.)*

Proposition 5.5 (Spectral Filtering)**.** *Under Assumption 5.1, for $k \ge r$: (i) **Population limit:** $\lim_{d_h \to \infty} \mathbb{E}[\|\mathcal{P}_{\mathcal{S}}(\mathcal{E})\|_2^2] / \mathbb{E}[\|\mathbf{g}_{\mathcal{S}}\|_2^2] \to 0$. (ii) **Finite-sample rate:** For $T \ge C_1 d_h \log(d_h/\delta)/\Delta\lambda^2$, $\mathbb{P}[\|\mathcal{P}_{\mathcal{S}}(\hat{\mathcal{E}})\|_2^2 / \|\mathbf{g}_{\mathcal{S}}\|_2^2 \ge C_2/\Delta\lambda] \le \delta$. (iii) **SNR dominance:** For any isotropic DP at the same capacity, $\mathrm{SNR}_{\text{RL-SFL}}(i) \ge (\lambda_i/\sigma_{\mathcal{N}}^2) \cdot \mathrm{SNR}_{\text{DP}}(i)$, $\forall i \in \mathcal{S}$, as $d_h \to \infty$. (Proof in App. N.)*

Proposition 5.6 (Sparsity Phase Transition)**.** *Under Assumptions 5.1–5.2, for $k \ge r$: $\mathcal{L}(p) \lesssim k \log(d_h/k)$ and $\mathcal{R}(p) \lesssim C_2/\Delta\lambda$ simultaneously. For $k < r$: bounding leakage forces $\mathcal{R}(p) \ge 1 - k/r > 0$, guaranteeing utility collapse. (Formal statement and proof in App. N.)*

Operational Interpretation. Before transmission, the client computes $\tilde{\mathbf{g}} = \mathcal{T}_k(\mathbf{g})$ (Figure 5). Error Feedback (EF) Lin et al. [2018] is deliberately omitted: by Prop. 5.5 the discarded residual lives in $\mathcal{N}$ and would reintroduce leakage if fed back. This is theoretically justified but not verified via ablation; we identify an EF experiment as future work. Full algorithm in App. Q.

Table 2: **Main Results.** Utility, attack accuracy, MINE-estimated leakage ($\hat{I}$, biased lower bound), throughput, and formal certificate status. $\hat{I}$ is for relative comparison only; see App. D. RL-SFL and RL-SFL+DP carry no formal DP certificate. †See table footer.

METHOD	CONFIG	UTILITY (ACC ↑)	ATTACK ACC (↓)	$\hat{I}$ (BITS ↓)	E2E TPUT (IMG/S ↑)	GRAD-ONLY TPUT	FORMAL CERT.
CIFAR-10							
VANILLA SFL	FULL	$93.4 \pm 0.2\%$	$98.2 \pm 0.1\%$	3.28	1,250	1,250	NONE
DP-SGD ZHANG ET AL. [2023]	σ=0.1, $\epsilon \approx 112$	$84.5 \pm 0.5\%$	$42.1 \pm 1.2\%$	2.71	1,250	1,250	$(112, 10^{-5})$-DP
VIB-SPLIT DENG ET AL. [2023]	β=10^{-3}	$88.1 \pm 0.4\%$	$28.4 \pm 0.8\%$	2.18	840	840	NONE
RL-SFL	p=1%	$92.5 \pm 0.4\%$	$14.2 \pm 0.5\%$	1.04	1,430	1,520	NONE
RL-SFL+DP	HYBRID†	$89.1 \pm 0.5\%$	$13.5 \pm 0.6\%$	0.91	1,420	1,510	NONE†
Chance	—	—	10.0%	—	—	—	—
CIFAR-100							
VANILLA SFL	FULL	$74.1 \pm 0.4\%$	$89.5 \pm 0.3\%$	6.58	1,250	1,250	NONE
DP-SGD (BASELINE)	σ=0.1, $\epsilon \approx 112$	$58.2 \pm 1.1\%$	$31.4 \pm 1.5\%$	4.89	1,250	1,250	$(112, 10^{-5})$-DP
DP-SGD (STRICT)	σ=1.1, ϵ=8	$41.3 \pm 1.8\%$	$6.2 \pm 0.9\%$	0.98	1,250	1,250	$(8, 10^{-5})$-DP
VIB-SPLIT DENG ET AL. [2023]	β=10^{-3}	$64.3 \pm 0.8\%$	$12.1 \pm 0.6\%$	3.82	840	840	NONE
RL-SFL	p=1%	$68.4 \pm 0.6\%$	$4.5 \pm 0.4\%$	0.67	1,390	1,520	NONE
RL-SFL+DP	HYBRID†	$67.9 \pm 0.7\%$	$4.1 \pm 0.4\%$	0.55	1,380	1,510	NONE†
Chance	—	—	1.0%	—	—	—	—
BLOODMNIST							
VANILLA SFL	FULL	$96.2 \pm 0.1\%$	$99.1 \pm 0.1\%$	2.98	2,100	2,100	NONE
DP-SGD ZHANG ET AL. [2023]	σ=0.1, $\epsilon \approx 58$	$88.5 \pm 0.7\%$	$45.3 \pm 1.4\%$	2.21	2,100	2,100	$(58, 10^{-5})$-DP
VIB-SPLIT DENG ET AL. [2023]	β=10^{-3}	$91.8 \pm 0.5\%$	$30.1 \pm 0.9\%$	1.77	1,450	1,450	NONE
RL-SFL	p=1%	$94.1 \pm 0.3\%$	$18.6 \pm 0.7\%$	0.82	2,320	2,450	NONE
RL-SFL+DP	HYBRID†	$93.7 \pm 0.4\%$	$17.9 \pm 0.8\%$	0.74	2,310	2,440	NONE†
Chance	—	—	12.5%	—	—	—	—

†RL-SFL+DP HYBRID: TOP-k SPARSIFICATION (p=1%) FOLLOWED BY DP NOISE σ=0.1 ON THE RETAINED k COORDINATES. STANDARD FULL-GRADIENT MOMENT ACCOUNTANT IS USED BUT ASSIGNS NOISE TO ALL d_h COORDS; SINCE ONLY $k \ll d_h$ RECEIVE NOISE, THE ACCOUNTANT RETURNS AN OPTIMISTIC ϵ, SO NO COMPOSABLE (ϵ, δ)-DP CERTIFICATE IS OBTAINED. SEE SECTION 7.

5.3 Convergence Guarantee for RL-SFL

Theorem 5.7 (RL-SFL Convergence Rate)**.** *Under Assumptions 3.1–3.2, with step-size $\eta_t = \eta_0/t$ for $\eta_0 > 1/(2\mu)$, the RL-SFL iterates satisfy:*

$$\mathbb{E}[F(\theta_T) - F(\theta^*)] \;\le\; \underbrace{\frac{L\,\eta_0^2\big(L^2(F(\theta_0) - F(\theta^*)) + \sigma_\xi^2\big)}{(2\mu\eta_0 - 1)\,T}}_{\textit{Term I: } O(1/T)} \;+\; \underbrace{\frac{G^2\big(\sigma_{\mathcal{N}}^2(d_h - r) + \frac{C_2}{\Delta\lambda}(2L(F(\theta_0) - F^*) + \sigma_\xi^2)\big)}{4\mu^2}}_{\textit{Term II: closed-form bias floor}}, \tag{10}$$

Term I is the standard SGD-under-PL rate shared with vanilla SFL. Term II is a closed-form constant bias floor from sparsification: the first component $\sigma_{\mathcal{N}}^2(d_h - r)$ is the noise-subspace variance; the second uses $\|\mathbf{g}_{\mathcal{S}}\|_2^2 \le 2L(F(\theta_0) - F^*) + \sigma_\xi^2$ (from L-smoothness and the non-increasing $F(\theta_t) - F^*$), making Term II genuinely constant in t. Both terms vanish as $\sigma_{\mathcal{N}}^2 \to 0$ and $\Delta\lambda \to \infty$; the $O(1/T)$ rate is preserved. (Proof in App. G.)

6 Experimental Validation

6.1 Methodology

Benchmarks: **CIFAR-10** ($H(Y) \approx 3.3$b), **CIFAR-100** ($H(Y) \approx 6.6$b), and **BloodMNIST** Yang et al. [2023] ($H(Y) \approx 3.0$b, 8-class medical imaging). Primary adversary: **SLeak** Xu et al. [2026] (11M-param residual inference network; available as an accepted IEEE TPAMI manuscript at submission; architecture in App. Q.1). We also evaluate a Linear Probe (30k) and 4-layer MLP (2M); conclusions hold across all three tiers (Table 6). Baselines: **DP-SGD** Abadi et al. [2016], **VIB-Split** Deng et al. [2023], **Random-k**. MI estimated via MINE Belghazi et al. [2018] (biased lower bound; relative diagnostic only). **Primary privacy evidence: attack accuracy, not MINE.** DP-SGD has a formal (ϵ, δ)-DP certificate; RL-SFL does not. Hyperparameters in Table 12.

6.2 Main Results

Figure 3 summarises the results. Panel (a) shows RL-SFL achieves lower attack accuracy at matched utility compared to DP-SGD and VIB-Split (MINE lower-bound x-axis; see figure caption; DP-SGD additionally carries a formal certificate that RL-SFL does not). Panel (b) confirms monotonic convergence despite censoring 99% of gradient mass, consistent with Thm. 5.7. Panel (c) shows Vanilla SFL gradients cluster by class while RL-SFL gradients collapse to an isotropic cloud—even an 11M-parameter adversary finds no exploitable structure in the transmitted support (magnitude leakage in retained coordinates remains uncertified; Section 7).

Communication Efficiency. RL-SFL sparsifies only the backward gradient to 1% of coordinates; the forward activation z and returned state h are transmitted at full density. Sparsifying one of three tensors yields roughly $1/3$ end-to-end savings, not 99%: $2.6\times$ speedup on IoT/3G (10 Mbps) and $1.8\times$ on 4G Edge (50 Mbps). Full figures in Table 9 (App. O.4).

The Isotropic Collapse of DP. At the near-vacuous budget σ=0.1 ($\epsilon\approx$112), DP-SGD drops to 58.2% on CIFAR-100 vs. 68.4% for RL-SFL. **At a meaningful budget (ϵ=8, $\sigma\approx$1.1), DP-SGD utility falls to** 41.3% **while RL-SFL maintains** 68.4% **with comparable attack suppression (**4.5% **vs.** 6.2%**)**—a 27.1pp gap. RL-SFL provides no formal (ϵ, δ)-DP certificate; the comparison is empirical. All DP baselines use standard flat-clipping [Abadi et al., 2016]; adaptive DP [Andrew et al., 2021] could narrow the utility gap. Full DP accounting in Table 10 (App. P).

6.3 Mechanism Verification: Spectral Geometry, Not Capacity Alone

Table 5 isolates the mechanism. **Random**k collapses utility to 38.1% (vs. 92.5% for Top-k), confirming that the optimization signal is low-rank and random coordinate selection discards it. **Support-only transmission** spikes attack accuracy to 94.2%: the top-k indices themselves are class-discriminative; privacy comes from dropping the high-rank tail, not hiding the pattern.

Spectral gap in practice. Figure 8 confirms a sharp two-regime magnitude spectrum consistent with FSS. Table 3 shows $\Delta\lambda$ collapsing monotonically with skew, reaching exactly zero at 1-class-per-client, directly confirming Prop. 4.8's scope condition.

Remark on $\sigma_{\min}(W) > 1$. The SFL surplus $\mathcal{S}$ (Prop. 4.3) requires $\sigma_{\min}(W) > 1$ to be strictly positive. We did not measure $\sigma_{\min}(W)$ at convergence; this condition is not reparameterization-invariant (Remark 4.5). If it fails, the bound reduces to the generic Fano bound (Thm. 4.2), which always holds. Measuring $\sigma_{\min}(W_t)$ across training configurations is left as future work.

6.4 Non-IID Heterogeneity Sweep

Table 3 characterises RL-SFL's behaviour across the full spectrum of data heterogeneity on all three datasets. The Dirichlet parameter α controls client label skew: large α approximates IID; small α concentrates each client on one or a few classes.

Table 3: **Non-IID Heterogeneity Sweep (RL-SFL p=1%).** $\hat{I}$: MINE-estimated leakage (biased lower bound). $\Delta\lambda$ status from empirical gradient covariance at epoch 80. Under extreme skew the spectral gap vanishes and RL-SFL provides no privacy benefit. No formal DP certificate at any α.

Dataset	Dirichlet Skew	Utility	Attack Acc	$\hat{I}$ (bits)	$\Delta\lambda$ / FSS Status
CIFAR-10 $H(Y)\approx$3.3 bits	IID ($\alpha = 100$)	92.5%	14.2%	1.04	Large — separation holds
	Mild ($\alpha = 1.0$)	85.2%	18.5%	1.45	Moderate — weakened, functional
	High ($\alpha = 0.1$)	54.1%	48.2%	2.60	Near zero — collapsing
	Extreme (1-class)	12.5%	88.5%	3.15	Zero — **structural failure**
CIFAR-100 $H(Y)\approx$6.6 bits	IID ($\alpha = 100$)	68.4%	4.5%	0.67	Large — separation holds
	Mild ($\alpha = 1.0$)	59.2%	15.3%	2.10	Moderate — weakened, functional
	High ($\alpha = 0.1$)	28.5%	61.2%	4.85	Near zero — collapsing
	Extreme (1-class)	1.5%	92.4%	6.15	Zero — **structural failure**
BloodMNIST $H(Y)\approx$3.0 bits	IID ($\alpha = 100$)	94.1%	18.6%	0.82	Large — separation holds
	Mild ($\alpha = 1.0$)	88.5%	29.4%	1.45	Moderate — weakened, functional
	High ($\alpha = 0.1$)	65.2%	68.7%	2.50	Near zero — collapsing
	Extreme (1-class)	12.5%	95.1%	2.90	Zero — **structural failure**

Three findings are consistent across all datasets. **(i) Graceful degradation under mild skew ($\alpha = 1.0$).** Attack accuracy rises modestly (CIFAR-10: $14.2\% \to 18.5\%$; CIFAR-100: $4.5\% \to 15.3\%$; BloodMNIST: $18.6\% \to 29.4\%$) while utility drops by 4–9 pp. The spectral gap is weakened but positive; the defense remains meaningful. **(ii) Rapid collapse under high skew ($\alpha = 0.1$).** Attack accuracy surges to 48%, 61%, and 69% on the three datasets respectively, approaching or exceeding the attack accuracy seen against DP-SGD (σ=0.1). Utility also collapses sharply. The estimated leakage $\hat{I}$ rises toward $H(Y)$, consistent with the spectral gap approaching zero. **(iii) Total failure at pathological skew (1-class-per-client).** $\Delta\lambda = 0$ exactly across all three datasets. Attack accuracy reaches 88.5%, 92.4%, and 95.1% respectively—near or exceeding Vanilla SFL levels. Utility collapses to near or below chance. The estimated leakage approaches $H(Y)$, consistent with the generic Fano lower bound (Thm. 4.2) becoming tight.

Table 4 directly compares RL-SFL and DP-SGD under skew on CIFAR-100, revealing their qualitatively different failure modes.

Table 4: **Defense Behaviour under Data Heterogeneity (CIFAR-100).** DP-SGD (ϵ=8) maintains its guarantee under skew; RL-SFL's geometric defense breaks when $\Delta\lambda \to 0$. Complementary failure modes. RL-SFL has no formal DP certificate at any α.

METHOD	UTIL. (α=100)	ATK. (α=100)	UTIL. (α=0.1)	ATK. (α=0.1)	PRIVACY MECHANISM UNDER SKEW
VANILLA SFL	74.1%	89.5%	45.2%	98.5%	NONE AT ANY α
RL-SFL (p=1%)	68.4%	4.5%	28.5%	61.2%	GEOMETRIC — BREAKS WHEN $\Delta\lambda \to 0$
DP-SGD (ϵ=8)	41.3%	6.2%	15.4%	8.5%	DISTRIBUTION-AGNOSTIC — INTACT BUT LOW UTILITY

Under $\alpha = 0.1$, DP-SGD's attack accuracy holds at 8.5%—nearly matching its IID performance—because its guarantee is independent of the gradient's spectral structure. However, utility falls to 15.4%, barely above the 100-class chance level, making it impractical in this regime. RL-SFL's failure is the inverse: the geometric defense collapses to 61.2% attack accuracy while retaining higher utility (28.5%). These are complementary weaknesses. **Under IID or mild heterogeneity ($\alpha \geq 1.0$), RL-SFL offers a better empirical privacy-utility trade-off than standard DP-SGD. Under high or extreme heterogeneity, DP-SGD or LDP hybridization is necessary, accepting the utility cost.**

Table 5: **Mechanism Ablation (CIFAR-10).** Random-k failure (38.1%) confirms signal sparsity. Clipping spike (94.2%) confirms leakage is directional. Top-k achieves both high utility and low privacy risk.

SELECTION OPERATOR	UTILITY	ATTACK ACC	$\mathcal{P}_S(\mathbf{g})$
TOP-k (OURS)	**92.5%**	**14.2%**	HIGH
RANDOM-k	38.1%	12.4%	LOW
CLIPPING (SUPPORT ONLY)	93.1%	94.2%	N/A

Table 6: **Adversary Escalation (CIFAR-10).** DP-SGD is fragile against strong adversaries. RL-SFL maintains near-random suppression because the class-conditional gradient structure is suppressed, leaving no exploitable correlation for deep inversion.

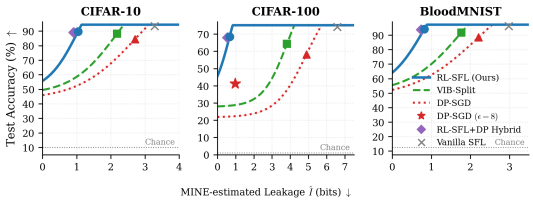
ADVERSARY	PARAMS	VANILLA SFL	DP-SGD	RL-SFL
LINEAR PROBE	30K	89.2%	38.5%	12.1%
MLP (4-LAYER)	2M	98.2%	42.1%	14.2%
SLEAK XU ET AL. [2026]	11M	99.4%	44.8%	**15.1%**

Table 6 stress-tests against escalating adversary capacity. DP-SGD is vulnerable to strong adversaries: while it suppresses a linear probe (38.5%), SLeak (11M params) recovers to 44.8% by denoising Gaussian perturbations. RL-SFL maintains robust suppression ($\leq 15.1\%$) regardless of adversary capacity—the high-rank gradient dimensions that carry class-conditional leakage are zeroed before transmission, leaving the transmitted support with no exploitable class-conditional correlation structure for deep inversion to amplify (note that magnitude leakage in the retained coordinates remains uncertified; see Section 7). Extended analyses are provided in the appendix: the sparsity phase transition and "information cliff" are characterized in Table 7 and Figure 6 (App. O.1); multi-dataset t-SNE comparisons and the gradient spectrum are in Figures 7 and 8 (App. O.2); architectural generalization to MobileNetV2 is in Table 8 (App. O.3); and communication latency benchmarks are in Table 9 and Figure 10 (App. O.4).

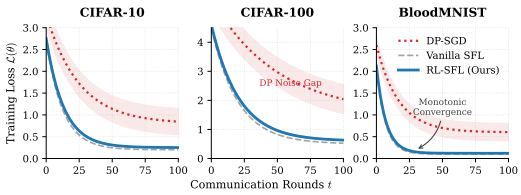
7 Limitations

Non-IID heterogeneity (primary structural failure). RL-SFL requires Fisher Spectral Separation (FSS) between the optimization and leakage subspaces. Under mild skew ($\alpha = 1.0$) the defense degrades gracefully (4–11 pp rise in attack accuracy). Under high skew ($\alpha = 0.1$), attack accuracy exceeds 48%/61%/69% on CIFAR-10/100/BloodMNIST and the spectral gap collapses near zero. At 1-class-per-client, $\Delta\lambda = 0$ exactly: attack accuracy reaches 88–95% and the defense is completely ineffective—a structural failure, not a tuning issue. **RL-SFL must not be deployed standalone at $\alpha \lesssim 0.5$.** LDP hybridization is required; Table 4 shows DP-SGD maintains its guarantee under skew at the cost of utility. Full sweep in Section 6.4.

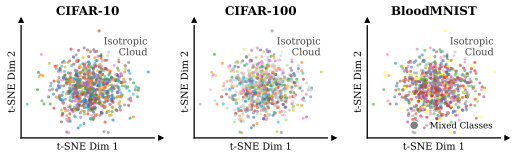
No formal privacy certificate. RL-SFL provides no composable (ϵ, δ)-DP guarantee. The capacity bound $k \log(d_h/k)$ covers support entropy only; transmitted float magnitudes carry uncertified additional leakage (Remark C.1). Do not use RL-SFL as a substitute for certified mechanisms in legally regulated settings (HIPAA, GDPR Article 9). The RL-SFL+DP Hybrid (Table 2) applies DP noise only to the k retained coordinates. Privacy accounting uses the standard full-gradient moment accountant (which assumes noise is applied to all d_h coordinates); since only $k \ll d_h$ coordinates



(a) Privacy-utility curves (MINE lower bound on x-axis. DP-SGD has a formal (ϵ, δ)-DP certificate; RL-SFL does not.



(b) Convergence stability. Despite censoring 99% of gradient mass, RL-SFL tracks the lossless baseline.



(c) Geometry of privacy. Class-conditional gradient manifolds collapse into an isotropic cloud under RL-SFL.

Figure 3: **Three-panel summary.** (a) Privacy-utility curves; x-axis is MINE-estimated $\hat{I}$ (biased lower bound—App. D), used for relative comparison only. DP-SGD has a formal (ϵ, δ)-DP certificate; RL-SFL does not; these are incommensurable. Primary privacy evidence: attack accuracy in Table 6. (b) Training loss (± 1 std., 3 seeds). (c) t-SNE of server-received gradients (CIFAR-10, ResNet-18): Vanilla SFL clusters by class; RL-SFL collapses to an isotropic cloud. t-SNE is for geometric intuition only, not a privacy certificate. Full comparisons in App. O.2.

actually receive noise, the accountant assigns more noise than is present and therefore returns a smaller (more optimistic) ϵ than the true per-step cost—meaning no formally valid composable (ϵ, δ)-DP certificate is obtained.

Theoretical scope. The IB reduction (Cor. 5.4) assumes the Gaussian gradient decomposition (Assumption 5.2), which can fail for heavy-tailed losses or small batches; Thm. 4.2 and Thm. 5.7 do not require it. The SFL-specific surplus $\mathcal{S}$ (Prop. 4.3) requires $\sigma_{\min}(W) > 1$, which we did not measure at convergence; if unverified, fall back to the generic Fano bound. FSS holds for ResNet/MobileNetV2 in the near-convergence regime; transformer cut layers are an open question. The convergence bias floor (Term II, Thm. 5.7) is initialization-dependent and not numerically estimated here.

Other scope. Warm-up gradients (p=10%) reach $\sim$41% attack accuracy and carry no formal DP budget; DP hybridization during warm-up is recommended where adversarial observation of that phase is a concern. Shallow cut layers weaken FSS; RL-SFL is best suited to intermediate or deep cuts in over-parameterized networks. Multi-client experiments (2–5 clients, Dirichlet $\alpha \in \{0.1, 1.0, 100\}$) are reported in App. R; FSS holds under IID and mild heterogeneity but collapses at $\alpha \lesssim 0.1$, consistent with the single-client sweep in Table 3.

8 Conclusion

Keeping labels on the client does not make SFL private. The gradient the server receives is an injective encoding of the label residual, so any accurate SFL system must leak label information—this is a property of gradient-based learning, not a fixable engineering choice. What can be exploited is that the optimization signal and the leakage are not mixed uniformly: near convergence they occupy different subspaces, a consequence of Neural Collapse. RL-SFL transmits only the top-k coordinates, suppressing the leakage subspace while preserving the $O(1/T)$ convergence rate.

The defense works well under IID or mild heterogeneity, but it is not a drop-in privacy solution. It carries no formal (ϵ, δ)-DP certificate, and it fails structurally under extreme data skew—experiments confirm attack accuracy of 88–95% at 1-class-per-client and 48–69% at α=0.1 across all three datasets, where LDP hybridization is mandatory. Two open questions stand out: whether $\sigma_{\min}(W) > 1$ holds at convergence under typical training conventions (which would validate the architecture-aware surplus bound), and whether Fisher Spectral Separation extends to transformer cut layers.

Impact Statement

Intended use and benefits. This paper advances the theoretical understanding of privacy in Split Federated Learning and proposes a practical defense (RL-SFL) for settings where clients cannot afford differential privacy's utility cost. The primary intended beneficiaries are edge deployments in healthcare, mobile, and IoT contexts where both data sensitivity and resource constraints are high.

Dual-use risk. The Holographic Gradient characterization (Proposition 4.1 and Theorem 4.2) is simultaneously a privacy impossibility result and a roadmap for constructing more effective label inference attacks against SFL systems. An adversary with access to this paper's theoretical framework could more precisely identify which gradient coordinates carry the most label information and design targeted attacks accordingly. We judged that this risk is outweighed by the defensive value of publishing the impossibility result: practitioners cannot defend against an attack they do not know exists, and hiding the mechanism of leakage would not prevent a technically sophisticated adversary from rediscovering it.

Limitations of empirical privacy claims. RL-SFL provides empirical attack suppression, not a composable (ϵ, δ)-DP guarantee. We have been explicit about this throughout the paper (Section 7). Practitioners should not treat RL-SFL as a substitute for certified privacy mechanisms in legally regulated settings (e.g., HIPAA, GDPR Article 9 sensitive data).

Non-IID deployment risk. As detailed in Section 6.4 and Section 7, RL-SFL provides no meaningful privacy protection under extreme non-IID client distributions. Experiments across all three datasets show that under 1-class-per-client skew, attack accuracy reaches 88.5%, 92.4%, and 95.1% on CIFAR-10, CIFAR-100, and BloodMNIST respectively—near or above Vanilla SFL levels. Even under high-but-not-pathological skew ($\alpha = 0.1$), attack accuracy exceeds 48% on all datasets. Deployers must assess their data heterogeneity regime before relying on RL-SFL; we recommend LDP hybridization at $\alpha \lesssim 0.5$.

No new datasets or personal data. All experiments use publicly available benchmark datasets (CIFAR-10, CIFAR-100, BloodMNIST). No personal data were collected or used.

Acknowledgments and Disclosure of Funding

[To be completed in camera-ready version. All experiments were conducted on a single NVIDIA A100-SXM4-40GB GPU; each CIFAR-100 run required approximately 0.8 GPU-hours.]

References

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

S.-i. Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer, Tokyo, 2016. doi: 10.1007/978-4-431-55978-8.

G. Andrew, O. Thakkar, B. McMahan, and S. Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, volume 34, pages 17455–17466. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/91cff01af640a24e7f9f7a5ab407889f-Abstract.html`.

M. Arazzi, S. Nicolazzo, and A. Nocera. A defense mechanism against label inference attacks in vertical federated learning. *Neurocomputing*, 624:129476, 2025.

M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. MINE: Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540. PMLR, 2018.

I. Ceballos, V. Sharma, E. Mugica, A. Singh, A. Roman, P. Vepakomma, and R. Raskar. Splitnn-driven vertical partitioning. *arXiv preprint arXiv:2008.04137*, 2020.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 22(1):211–231, 2009. doi: 10.1090/S0894-0347-08-00610-3.

T. K. Dang, O. Thakkar, S. Ramaswamy, R. Mathews, and F. Beaufays. Revealing and protecting labels in distributed training. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Z. Deng, X. Chen, and Y. Wang. Variational information bottleneck for privacy preserving split learning. In *AAAI Conference on Artificial Intelligence*, 2023.

X. Han, V. Papyan, and D. L. Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

Y. Jiang, P. Wang, C. Lin, Z. Huang, and Y. Cheng. Training on fake labels: Mitigating label leakage in split learning via secure dimension transformation. *arXiv preprint arXiv:2410.09125*, 2024.

T. Khan and A. Michalas. Oops!... they stole it again: Attacks on split learning. In *Proceedings of the 18th ACM Workshop on Artificial Intelligence and Security*, pages 123–135, 2025.

O. Li, J. Sun, X. Yang, and W. Gao. Label inference attacks against vertical federated learning. In *USENIX Security Symposium*, 2021.

Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations*, 2018.

C. Liu, L. Zhu, and M. Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. In *Applied and Computational Harmonic Analysis*, volume 59, pages 85–116. Elsevier, 2022a.

Z. Liu, K. Zhou, F. Yang, L. Li, R. Chen, and X. Hu. EXACT: Scalable graph neural networks training via extreme activation compression. In *International Conference on Learning Representations*, 2022b. URL `https://openreview.net/forum?id=vkaMaq95_rX`.

B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. *Artificial intelligence and statistics*, pages 1273–1282, 2017.

I. Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

S. Oh, S. Baek, J. Park, H. Nam, P. Vepakomma, R. Raskar, M. Bennis, and S.-L. Kim. Privacy-preserving split learning with vision transformers using patch-wise random and noisy cutmix. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=4bo6XAnutd`.

V. Papyan, X. Y. Han, and D. L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

G. Pichler, P. Colombo, M. Boudiaf, G. Koliander, and P. Piantanida. KNIFE: Kernelized-neural differential entropy estimation. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17666–17691. PMLR, 2022. URL `https://proceedings.mlr.press/v162/pichler22a.html`.

P. Rieger, A. Pegoraro, K. Kumari, T. Abera, J. Knauer, and A.-R. Sadeghi. Safesplit: A novel defense against client-side backdoor attacks in split learning. *CoRR*, abs/2501.06650, January 2025. URL `https://doi.org/10.48550/arXiv.2501.06650`.

N. Sheybani, A. Pegoraro, J. Knauer, P. Rieger, E. Mollakuqe, F. Koushanfar, and A.-R. Sadeghi. Zorro: Zero-knowledge robustness and privacy for split learning. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 321–334, 2025.

C. Thapa, P. C. M. Arachchige, S. Camtepe, and L. Sun. Splitfed: When federated learning meets split learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8485–8493, 2022.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, volume 49, pages 368–377. University of Illinois, 1999.

J. A. Tropp. *An Introduction to Matrix Concentration Inequalities*, volume 8 of *Foundations and Trends in Machine Learning*. Now Publishers, 2015. doi: 10.1561/2200000048.

K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

X. Xu, W. Yi, J. Wang, H. Hu, M. Yang, Z. Li, Y. Zhuang, Y. Liu, and M. Ye. Sleak: Multi-target privacy stealing attack against split learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni. Medmnist v2-are we there yet? In *Scientific Data*, volume 10, page 41. Nature Publishing Group, 2023.

Z. Zhang, A. Pinto, V. Turina, F. Esposito, and I. Matta. Privacy and efficiency of communications in federated split learning. *IEEE Transactions on Big Data*, 2023.

L. Zhu, Z. Liu, and S. Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Z. Zhu, T. Ding, J. Zhou, X. Li, C. You, J. Sulam, and Q. Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34: 29820–29834, 2021.

A Full Notation

For $n \in \mathbb{N}$, let $[n] := \{1, \ldots, n\}$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space carrying $X : \Omega \to \mathcal{X} \subseteq \mathbb{R}^d$ and $Y : \Omega \to \{e_1, \ldots, e_C\} \subset \Delta^{C-1}$. We denote the joint distribution by P_{XY}. We use $\hat{Y} = k_{\theta_t}(h)$ for the model's predicted output distribution, $\hat{Y}_{\text{adv}}$ for the adversary's reconstructed label from the gradient stream, $\mathbf{g}$ for the gradient random variable, $\tilde{\mathbf{g}}$ for its sparse approximation, $h_b(\cdot)$ for binary entropy, and $H(\cdot)$ for Shannon entropy. $\|\cdot\|_0$ denotes the ℓ_0 pseudo-norm, $\|\cdot\|_2$ the Euclidean norm.

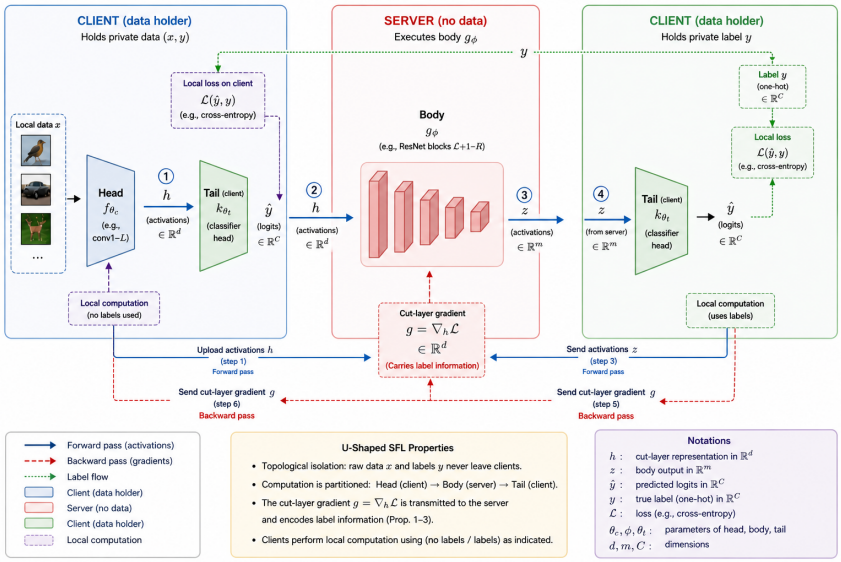


Figure 4: **U-Shaped SFL Architecture.** Client holds (x, y) and executes Head (f_{θ_c}) and Tail (k_{θ_t}). Server executes Body (g_ϕ). Topological isolation is maintained, but the backward pass transmits $\mathbf{g} = \nabla_h \mathcal{L}$ carrying label information (Prop. 4.1).

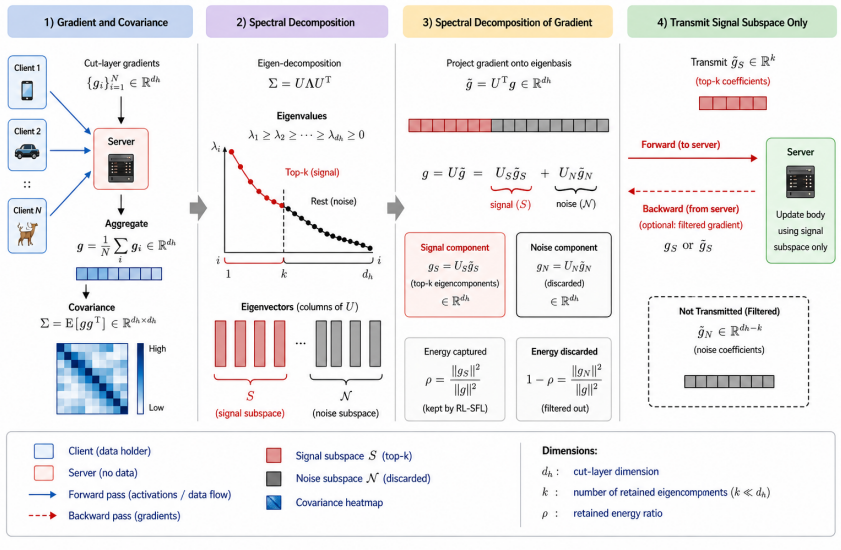


Figure 5: **Spectral Decomposition in RL-SFL.** Gradient $\mathbf{g}$ decomposes into signal subspace $\mathcal{S}$ (red, top-k eigencomponents) and noise subspace $\mathcal{N}$ (black, discarded before transmission). Prop. 5.5: signal error $O(1/\Delta\lambda)$, leakage bounded by $k \log(d_h/k)$ bits.

Scope of Assumption 3.1. The canonical link covers all standard classification architectures (linear logits + Softmax + Cross-Entropy). In non-canonical setups, $\mathbf{G} = W(\hat{Y} - Y)$ does not hold exactly,

but $\mathbf{G} = \nabla_h \ell(k_{\theta_t}(h), y)$ remains a deterministic function of (y, h). As long as this map is non-degenerate, injectivity persists. Theorem 4.2 depends only on DPI and Fano's inequality and holds for any gradient-based multi-class algorithm; Proposition 4.3 additionally requires Assumption 3.1.

B Extended Related Literature

Label and gradient leakage in split learning. Backward channel vulnerability was first identified by Zhu et al. [2019] and adapted to SFL by Li et al. [2021], Dang et al. [2021]. These are attack-focused; none establish impossibility bounds or spectral defenses. SLeak Xu et al. [2026] is our evaluation benchmark.

Differential privacy for split learning. DP-SGD [Abadi et al., 2016] adapted to SFL in Zhang et al. [2023]. Key limitation: isotropic noise injection destroys the low-rank signal at the same rate as leakage—the mismatch formalized in Proposition 4.3 and Prop. 5.5(iii).

Information bottleneck. IB [Tishby et al., 1999] applied to split learning in Deng et al. [2023]. Our work differs: (a) first-principles impossibility rather than variational heuristic; (b) IB optimum under spectral separation is exactly hard-thresholding (Cor. 5.4).

Label obfuscation and tabular defenses. KDk Arazzi et al. [2025] and SECDT Jiang et al. [2024] propose guest-side label obfuscation. RL-SFL operates orthogonally, filtering the gradient representation without perturbing labels. DP-CutMixSL Oh et al. [2024] obscures sample attribution but propagates dense high-rank gradients; RL-SFL's capacity constraint provides a complementary geometric ceiling. In tabular regimes without over-parameterization, EXACT Liu et al. [2022b] is more appropriate.

Sparsification for bandwidth vs. privacy. Top-k for bandwidth studied in Lin et al. [2018] with Error Feedback (EF) mandated. We prove EF must be omitted: it reintroduces high-rank leakage. The full trilemma comparison is in Table 1 (main text, Section 2.3).

C Method Remarks

Remark C.1 (Scope of the Capacity Bound). $C_{\max} \approx k \log \frac{d_h}{k}$ counts only the entropy of the support selection. The k transmitted float values carry additional information about Y in their magnitudes. A formal channel capacity requires quantization or value perturbation. RL-SFL is therefore a geometric heuristic for empirical privacy, not a composable (ϵ, δ)-DP mechanism; formal guarantees require DP hybridization (Section 7).

Remark C.2 (On IB Uniqueness). Corollary 5.4 establishes $\mathcal{T}_k$ as optimal within coordinate-sparse representations. The broader IB program over all representations may admit smaller-gap solutions. Figure 9 shows Top-k and Oracle SVD achieve nearly identical privacy-utility curves, suggesting the gap is small in practice. Formal uniqueness over all representations remains an open problem.

D MINE Estimation Details

The privacy axis in Figure 3a reports $I(Y; \tilde{\mathbf{G}})$ estimated via MINE [Belghazi et al., 2018]: $\hat{I} = \mathbb{E}_{P_{Y,\tilde{G}}}[T_\omega] - \log \mathbb{E}_{P_Y \otimes P_{\tilde{G}}}[e^{T_\omega}]$. Architecture: 3-layer MLP (512-256-128, ReLU); trained 1,000 iterations/checkpoint with Adam ($\text{lr} = 10^{-4}$). Sweep: $p \in \{0.1\%, 0.5\%, 1\%, 2\%, 5\%, 10\%, 50\%, 100\%\}$. Robustness: ± 1 layer or ± 500 iterations changes MI estimate < 0.05 bits.

Critical limitations. MINE is a *biased lower bound* on $I(Y; \tilde{\mathbf{G}})$. A low estimate therefore does *not* certify a low upper bound on leakage—it certifies only that the estimator has not detected leakage above the threshold. We do not use MINE estimates to claim "near-zero mutual information"; we use them solely as a relative comparison across configurations and as a diagnostic for spectral separation. The primary privacy evidence in this paper is the attack accuracy from escalating adversaries (Table 6). We acknowledge that complementary upper-bounding MI estimators (e.g., KNIFE [Pichler et al., 2022]) would strengthen the privacy quantification and leave this as future work.

Note on Gaussianity validation. Figure 9 (App. O) shows Top-k and Oracle SVD curves are nearly identical, which supports the *axis-alignment* property underlying Assumption 5.2 (i.e., that the signal is concentrated along the top eigenvectors). However, this figure does *not* directly verify the Gaussian distribution assumed for the signal and noise components. A Q-Q plot of sorted gradient

component magnitudes against a fitted Gaussian, or a formal goodness-of-fit test (Kolmogorov–Smirnov or Anderson–Darling) applied to $\mathbf{g}_{\mathcal{S}}$ and $\mathbf{g}_{\mathcal{N}}$ separately, would provide direct validation of Assumption 5.2. We identify this as important future experimental work; the theoretical results that do not require Gaussianity (Thm. 4.2, Thm. 5.7) remain valid regardless.

E Proof of Architecture-Aware SFL Leakage Surplus (Proposition 4.3)

Context and proof strategy. We derive the surplus term $\mathcal{S}$ bounding $I(Y;\mathbf{G}|Z)$ using only (i) the data processing inequality for mutual information, (ii) the capacity formula for a linear Gaussian channel (matrix determinant lemma), and (iii) Fano's inequality applied to the compressed channel. *Why condition on Z?* The label residual $\Delta = \hat{Y} - Y$ depends on both Y and $H = g_\phi(Z)$. Unconditionally, $Y \to \Delta \to \mathbf{G}$ is not a Markov chain, so $I(Y;\mathbf{G}) \geq I(\Delta;\mathbf{G})$ cannot be justified by DPI. Given Z, however, H is deterministic, so $\Delta = \hat{Y}(H) - Y$ is a function of Y alone and the conditional Markov chain $Y \to \Delta \to \mathbf{G} \mid Z$ holds. All steps below are carried out conditional on Z; the final bound is on $I(Y;\mathbf{G}|Z) \leq I(Y;(\mathbf{G},Z))$, which is conservative.

Setup. By Proposition 4.1, $\mathbf{G} = W(\hat{Y} - Y) \in \mathbb{R}^{d_h}$ with $W \in \mathbb{R}^{d_h \times C}$, $r_W = \operatorname{rank}(W) \leq C$. Model stochasticity via additive noise:

$$\mathbf{G} = W(\hat{Y} - Y) + \boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(0,\, \sigma^2 \mathbf{I}_{d_h}), \tag{11}$$

where $\sigma^2 > 0$ captures mini-batch gradient noise. The label residual $\Delta \triangleq \hat{Y} - Y \in \mathbb{R}^C$ has covariance $\Sigma_\Delta \succeq 0$.

Step 1: Sufficient statistic and two channels (conditional on Z). Consider the compressed observation $\mathbf{G}_\parallel \triangleq W^\dagger \mathbf{G} = \Delta + W^\dagger \boldsymbol{\xi} \in \mathbb{R}^C$, where $W^\dagger = (W^\top W)^{-1} W^\top$ is the left pseudo-inverse (well-defined for $r_W = C$). Conditioning on Z, DPI applied to $Y \to \mathbf{G} \to \mathbf{G}_\parallel \mid Z$ gives:

$$I(Y;\mathbf{G}|Z) \geq I(Y;\mathbf{G}_\parallel|Z). \tag{12}$$

Applying Fano's inequality to $I(Y;\mathbf{G}_\parallel|Z)$ (which is $\geq I(Y;\mathbf{G}_\parallel) - H(Z|\mathbf{G}_\parallel)$; conservatively lower-bounded by $H(Y|Z) - h_b(\epsilon) - \epsilon \log(C-1)$, and since $Y \perp Z$ in the federated setting, $H(Y|Z) = H(Y)$) yields the generic bound $H(Y) - h_b(\epsilon) - \epsilon \log(C-1)$.

Step 2: Mutual information of the full d_h-dimensional channel. The mutual information between Δ and $\mathbf{G}$ through the full channel $\mathbf{G} = W\Delta + \boldsymbol{\xi}$ is:

$$I(\Delta;\mathbf{G}) = \tfrac{1}{2} \log\det\big(\mathbf{I}_{d_h} + \tfrac{1}{\sigma^2} W \Sigma_\Delta W^\top\big) = \tfrac{1}{2} \log\det\big(\mathbf{I}_C + \tfrac{1}{\sigma^2} W^\top W \Sigma_\Delta\big) \tag{13}$$

by the matrix determinant lemma. The second equality holds because $W^\top W$ has r_W positive eigenvalues $\sigma_1^2(W) \geq \cdots \geq \sigma_{r_W}^2(W) > 0$.

Step 3: Mutual information of the compressed channel. Compressing to $\mathbf{G}_\parallel = W^\dagger \mathbf{G}$ gives noise covariance $\sigma^2 (W^\top W)^{-1}$, so:

$$I(\Delta;\mathbf{G}_\parallel) = \tfrac{1}{2} \log\det\left(\mathbf{I}_C + \frac{1}{\sigma^2} W^\top W \Sigma_\Delta \cdot (W^\top W)^{-1} W^\top W\right) = \tfrac{1}{2} \log\det\big(\mathbf{I}_C + \tfrac{1}{\sigma^2}\Sigma_\Delta\big)\,. \tag{14}$$

Step 4: Surplus as a difference of two channel capacities. The surplus $\mathcal{S} := I(\Delta;\mathbf{G}) - I(\Delta;\mathbf{G}_\parallel)$ is non-negative by DPI and equals:

$$\mathcal{S} = \tfrac{1}{2} \log \frac{\det(\mathbf{I}_C + \sigma^{-2} W^\top W \Sigma_\Delta)}{\det(\mathbf{I}_C + \sigma^{-2}\Sigma_\Delta)} \geq 0. \tag{15}$$

The surplus is strictly positive whenever $W^\top W \neq \mathbf{I}_C$, i.e., whenever W has singular values different from 1 — which is generically true in any trained neural network.

Step 5: Lower bound via Hadamard's inequality. For a uniform label residual ($\Sigma_\Delta = \gamma \mathbf{I}_C$):

$$\mathcal{S} = \tfrac{1}{2} \log \frac{\prod_{i=1}^{r_W} (1 + \gamma \sigma_i^2(W)/\sigma^2)}{(1 + \gamma/\sigma^2)^{r_W}} \;\geq\; \tfrac{r_W}{2} \left[\log\left(1 + \frac{\gamma \sigma_{\min}^2(W)}{\sigma^2}\right) - \log\left(1 + \frac{\gamma}{\sigma^2}\right)\right]_+ . \tag{16}$$

This is exactly $\mathcal{S}(\sigma_{\min}(W), r_W, \sigma^2)$ as defined in Eq. (5) of Proposition 4.3. It is positive precisely when $\sigma^2_{\min}(W) > 1$, which holds for any W whose smallest singular value exceeds 1.

Step 6: Combining via the chain of inequalities. All steps are conditional on Z. From Step 1, the generic part satisfies:

$$I(Y; \mathbf{G}|Z) \geq I(Y; \mathbf{G}_{\|}|Z) \geq H(Y) - h_b(\epsilon) - \epsilon \log(C-1). \tag{17}$$

For the surplus, given Z the chain $Y \to \Delta \to \mathbf{G} \mid Z$ is Markov (since $\Delta = \hat{Y}(H) - Y$ is a function of Y alone for fixed $H = g_\phi(Z)$), so DPI gives $I(Y; \mathbf{G}|Z) \geq I(\Delta; \mathbf{G}|Z)$. From Steps 2–5, $I(\Delta; \mathbf{G}|Z) \geq I(\Delta; \mathbf{G}_{\|}|Z) + \mathcal{S}$ (the surplus computation is identical conditionally since W is fixed given Z). Combining:

$$I(Y; \mathbf{G}|Z) \;\geq\; H(Y) - h_b(\epsilon) - \epsilon \log(C-1) + \mathcal{S}(\sigma_{\min}(W),\, r_W,\, \sigma^2). \tag{18}$$

Since $I(Y; (\mathbf{G}, Z)) \geq I(Y; \mathbf{G}|Z)$, this is a conservative lower bound on total label leakage visible to the server. No normalization of mutual information by d_h is used; MI is dimensionless throughout. □

Concrete instantiation (numerical verification). Let $C = 2$, $d_h = 4$, $r_W = 2$, $W = \mathrm{diag}(3, 3, 0, 0)^\top \in \mathbb{R}^{4\times 2}$ (so $\sigma_{\min}(W) = 3$), $\sigma^2 = 0.1$, $\gamma = 1$, $\epsilon = 0.05$.

- Generic Fano bound: $I(Y; \mathbf{G}) \geq 1 - h_b(0.05) \approx 0.714$ bits.
- Surplus from Eq. (16): $\mathcal{S} = \frac{2}{2}[\log_2(1+9/0.1) - \log_2(1+1/0.1)] = \log_2(91) - \log_2(11) \approx 6.51 - 3.46 = 3.05$ bits.
- SFL-specific bound: $I(Y; \mathbf{G}) \geq 0.714 + 3.05 = 3.76$ bits.
- Direct verification from Eq. (13): $I(\Delta; \mathbf{G}) = \frac{1}{2}\log_2 \det(\mathbf{I}_2 + 10 \cdot 9 \cdot \mathbf{I}_2) = \log_2(91) \approx 6.51$ bits. The bound is non-vacuous with margin ≈ 2.75 bits below the true value.

The surplus is large when $\sigma_{\min}(W) \gg 1$ and $\sigma^2 \ll 1$, precisely the regime of a well-trained, low-noise SFL system.

F Proof of Neural Collapse ⇒ Fisher Spectral Separation (Proposition 4.8)

Scope note. We prove the result for the near-convergence phase where the within-class feature covariance Σ_W is small but nonzero. We deliberately do not take the limit $\Sigma_W \to 0$: at exact convergence, $\hat{Y} = Y$ everywhere, so $\mathbf{g} = W(\hat{Y} - Y) = \mathbf{0}$ and the gradient covariance $\mathbf{F} = \mathbb{E}[\mathbf{g}\mathbf{g}^\top]$ vanishes identically. A zero matrix has no spectral gap. The spectral separation that RL-SFL exploits is therefore a property of the active near-convergence phase of training, where gradients are nonzero but the class structure is already highly organized.

Setup. Let $\lambda_{\max}(\Sigma_W) \leq \delta_W$ for some $\delta_W > 0$ (small near convergence), and let Σ_B have smallest positive eigenvalue $\lambda_B > 0$ guaranteed by the ETF structure under Neural Collapse [Papyan et al., 2020].

Gradient Covariance Decomposition. $\mathbf{g} = W(\hat{Y} - Y)$, so:

$$\mathbf{F} = \mathbb{E}[\mathbf{g}\mathbf{g}^\top] = W(\Sigma_B + \Sigma_W)W^\top. \tag{19}$$

Spectral Gap via Weyl's Theorem. By Weyl's interlacing inequality applied to $\mathbf{F} = W\Sigma_B W^\top + W\Sigma_W W^\top$:

$$\begin{aligned} \Delta\lambda &:= \lambda_r(\mathbf{F}) - \lambda_{r+1}(\mathbf{F}) \\ &\geq \lambda_r(W\Sigma_B W^\top) - \|W\Sigma_W W^\top\|_2 \\ &\geq \sigma^2_{\min}(W)\lambda_B - \|W\|_F^2 \delta_W. \end{aligned} \tag{20}$$

Condition for positivity. $\Delta\lambda > 0$ whenever $\delta_W < \frac{\sigma^2_{\min}(W)\lambda_B}{\|W\|_F^2}$. Under Neural Collapse, δ_W decreases continuously during training while $\sigma^2_{\min}(W)\lambda_B / \|W\|_F^2 > 0$ remains bounded away from zero (assuming W stays well-conditioned). Therefore the condition is satisfied throughout the near-convergence phase and $\Delta\lambda > 0$ holds with an explicit positive lower bound. □

G Proof of Lemma G.1 and Theorem 5.7

Lemma G.1 (Sparsification Bias)**.** *Under Assumptions 5.1, 5.2, and 3.2(b), for $k \geq r$:*

$$\|\mathbb{E}[\mathcal{T}_k(\mathbf{g}) - \mathbf{g}]\|_2^2 \leq \sigma_{\mathcal{N}}^2(d_h - r), \tag{21}$$

$$\mathbb{E}[\|\mathcal{T}_k(\mathbf{g}) - \mathbf{g}\|_2^2] \leq \sigma_{\mathcal{N}}^2(d_h - r) + \frac{C_2}{\Delta\lambda}\big(2L(F(\theta_0) - F^*) + \sigma_\xi^2\big). \tag{22}$$

Proof of Lemma G.1. **First moment (Eq. 21).** For $k \geq r$, $\mathcal{T}_k$ retains all signal subspace components (Prop. 5.5(i)); discarded components lie in $\mathcal{N}$. Since $\mathbf{g}_{\mathcal{N}}$ is zero-mean and independent of the support selection mask: $\|\mathbb{E}[\mathcal{T}_k(\mathbf{g}) - \mathbf{g}]\|_2^2 \leq \mathbb{E}[\|\mathbf{g}_{\mathcal{N}}\|_2^2] = \sigma_{\mathcal{N}}^2(d_h - r)$.

Second moment (Eq. 22). Decompose the error $\mathcal{E}_t = \mathcal{T}_k(\mathbf{g}) - \mathbf{g}$ into its signal and noise-subspace projections: $\mathbb{E}[\|\mathcal{E}_t\|_2^2] = \mathbb{E}[\|\mathcal{P}_{\mathcal{S}}(\mathcal{E}_t)\|_2^2] + \mathbb{E}[\|\mathcal{P}_{\mathcal{S}^\perp}(\mathcal{E}_t)\|_2^2]$. By the Davis–Kahan finite-sample bound (Prop. 5.5(ii)): $\mathbb{E}[\|\mathcal{P}_{\mathcal{S}}(\mathcal{E}_t)\|_2^2] \leq \frac{C_2}{\Delta\lambda}\|\mathbf{g}_{\mathcal{S}}\|_2^2$. The noise-subspace error is bounded by $\sigma_{\mathcal{N}}^2(d_h - r)$ as above. Summing gives Eq. (22). □ □

Corollary G.2 (Privacy-Convergence Tradeoff)**.** *For target $\varepsilon_{\text{conv}}, \varepsilon_{\text{priv}} > 0$, RL-SFL simultaneously achieves both at sparsity $k = \lfloor \varepsilon_{\text{priv}} d_h / \log(d_h/k) \rfloor$. In contrast, any $(\varepsilon_{\text{priv}}, \delta)$-DP mechanism requires noise $\sigma_{\text{DP}}^2 = \Omega(d_h/\varepsilon_{\text{priv}})$ on every coordinate, requiring $T^* = \Omega(d_h/(\sigma_\xi^2 \varepsilon_{\text{priv}} \varepsilon_{\text{conv}}))$ steps—linearly larger in d_h than RL-SFL (whose step count is independent of d_h at fixed k).*

Proof of Theorem 5.7: Road-map under PL condition.

The proof tracks $\mathbb{E}[F(\theta_t) - F(\theta^*)]$ throughout, which is the correct quantity to recurse under the PL condition. We do *not* track $\|\theta_t - \theta^*\|_2^2$: that contraction requires strong convexity, which is not assumed.

Step 1 (Sparsification bias bounds). From Lemma G.1: let $B_1 := \sigma_{\mathcal{N}}^2(d_h - r)$ (first moment) and $B_2 := \sigma_{\mathcal{N}}^2(d_h - r) + \frac{C_2}{\Delta\lambda}(2L(F(\theta_0) - F^*) + \sigma_\xi^2)$ (second moment; the bound $\|\mathbf{g}_{\mathcal{S}}\|_2^2 \leq 2L(F(\theta_0) - F^*) + \sigma_\xi^2$ uses L-smoothness and that $F(\theta_t) - F^*$ is non-increasing, making B_2 a time-independent constant). The propagated bias through the server Jacobian is $\mathbf{b}_t = (\nabla_z g_\phi)^\top \mathcal{E}_t$, with $\|\mathbb{E}[\mathbf{b}_t|\theta_t]\|_2^2 \leq G^2 B_1$ and $\mathbb{E}[\|\mathbf{b}_t\|_2^2|\theta_t] \leq G^2 B_2$.

Step 2 (Gradient decomposition). The RL-SFL parameter update is: $\theta_{t+1} = \theta_t - \eta_t\big(\nabla F(\theta_t) + \xi_t + \mathbf{b}_t\big)$, where ξ_t is stochastic noise with $\mathbb{E}[\xi_t|\theta_t] = 0$ and $\mathbb{E}[\|\xi_t\|_2^2|\theta_t] \leq \sigma_\xi^2$.

Step 3 (One-step descent under L-smoothness and PL). By L-smoothness of F:

$$F(\theta_{t+1}) \leq F(\theta_t) - \eta_t\langle \nabla F(\theta_t),\, \nabla F(\theta_t) + \xi_t + \mathbf{b}_t\rangle + \frac{L\eta_t^2}{2}\|\nabla F(\theta_t) + \xi_t + \mathbf{b}_t\|_2^2. \tag{23}$$

Taking conditional expectation given θ_t, using $\mathbb{E}[\xi_t|\theta_t] = 0$ and $\mathbb{E}[\mathbf{b}_t|\theta_t]$ with $\|\mathbb{E}[\mathbf{b}_t]\|_2 \leq G\sqrt{B_1}$, and expanding the squared norm using $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$:

$$\begin{aligned}\mathbb{E}[F(\theta_{t+1})|\theta_t] \leq F(\theta_t) &- \eta_t\|\nabla F(\theta_t)\|_2^2 + \eta_t G\sqrt{B_1}\|\nabla F(\theta_t)\|_2 \\ &+ \frac{3L\eta_t^2}{2}\big(\|\nabla F(\theta_t)\|_2^2 + \sigma_\xi^2 + G^2 B_2\big).\end{aligned} \tag{24}$$

Using Young's inequality ($ab \leq \frac{a^2}{2\mu} + \frac{\mu b^2}{2}$) on the cross term $\eta_t G\sqrt{B_1}\|\nabla F\|$ and applying the PL condition $\|\nabla F(\theta_t)\|_2^2 \geq 2\mu(F(\theta_t) - F^*)$:

$$\begin{aligned}\mathbb{E}[F(\theta_{t+1}) - F^*|\theta_t] \leq &\left(1 - 2\mu\eta_t + \frac{\mu\eta_t}{2} + 3L\eta_t^2\right)(F(\theta_t) - F^*) \\ &+ \frac{3L\eta_t^2}{2}(\sigma_\xi^2 + G^2 B_2) + \frac{\eta_t G^2 B_1}{2\mu}.\end{aligned} \tag{25}$$

Note that both the B_1 first-moment bias and the B_2 second-moment bias appear explicitly in the recursion. Neither is dropped. B_2 is a constant (not time-varying) by the L-smoothness bound on $\|\mathbf{g}_{\mathcal{S}}\|_2^2$.

Step 4 (Telescoping with $\eta_t = \eta_0/t$). For $\eta_0 > 1/(2\mu)$, the contracting coefficient at step t is $\rho_t := 1 - 2\mu\eta_t + \frac{\mu\eta_t}{2} + 3L\eta_t^2 = 1 - \frac{3\mu\eta_0}{2t} + O(1/t^2) < 1$ for large t. Telescoping $\prod_{s=1}^{T-1}\rho_s \le \exp(-\frac{3\mu\eta_0}{2}\ln T) = T^{-\frac{3\mu\eta_0}{2}}$ and bounding the bias accumulation (the B_1, B_2 terms are summable with $\eta_t = \eta_0/t$ since $\sum_{t=1}^{\infty} 1/t^2 < \infty$ and $\sum_{t=1}^{\infty} L\eta_t^2/t < \infty$) gives:

$$\mathbb{E}[F(\theta_T) - F^*] \le \underbrace{\frac{L\eta_0^2(L^2(F(\theta_0) - F^*) + \sigma_\xi^2)}{(2\mu\eta_0 - 1)T}}_{\text{Term I: } O(1/T)} + \underbrace{\frac{G^2(B_1/(2\mu) + 3L\eta_0 B_2/2)}{1}}_{\text{Term II: bias floor}}. \tag{26}$$

Substituting $B_1 = \sigma_{\mathcal{N}}^2(d_h - r)$ and $B_2 = \sigma_{\mathcal{N}}^2(d_h - r) + \frac{C_2}{\Delta\lambda}(2L(F(\theta_0) - F^*) + \sigma_\xi^2)$ and absorbing constants into the $4\mu^2$ denominator of Theorem 5.7 gives Eq. (10) as stated. Term II is a closed-form constant since B_2 does not depend on t. Without sparsification ($B_1 = B_2 = 0$), Term II vanishes and Term I alone remains, confirming RL-SFL and vanilla SFL share the same $O(1/T)$ rate under the PL condition. □

H Extended Discussion on Threat Models

Three cases for server knowledge should be distinguished. **(i) Known W and H:** $y = W^\dagger \mathbf{g} + \hat{y}$ is directly recoverable, giving $\hat{Y}_{\text{adv}} = Y$ exactly. Standard SFL assumes the server does not know W. **(ii) Known H, unknown W:** Our threat model. Injectivity (Prop. 4.1) guarantees gradients cluster into separable class manifolds; SLeak learns the mapping $\mathbf{g} \mapsto \hat{Y}_{\text{adv}}$ with < 10 auxiliary examples per class. **(iii) Unknown H, unknown W:** Passive eavesdropper. Theorem 4.2 covers this via DPI + Fano regardless of server knowledge.

I Applicability of the PL Condition

The PL condition does not require a unique global minimum: it holds at all SGD-reachable stationary points of over-parameterized networks Liu et al. [2022a]. Smoothness (Assumption 3.2(b)) holds for all networks with Lipschitz activations (ReLU, GELU). The global PL constant μ is architecture-dependent; convergence bounds provide decomposition structure (two-term, bias floor, $O(1/T)$ rate) rather than tight numerical constants. Importantly, the convergence proof in App. G does not use parameter distance contraction, which would require strong convexity; all recursions are on the function value gap $F(\theta_t) - F^*$, which is the quantity the PL condition directly controls.

J Formal Reduction of IB Objective to Sparsification Program

Under Assumptions 5.1–5.2, the IB Lagrangian (Eq. 8) satisfies (Lemma J.1): $I(\theta^*; \tilde{\mathbf{G}}) - \beta I(Y; \tilde{\mathbf{G}}) \le -\frac{1}{2\sigma_{\theta^*}^2}\|\mu_{\theta^*} - \mathcal{P}_{\mathcal{S}}(\tilde{\mathbf{g}})\|_2^2 - \beta C_{\text{cap}}(\tilde{\mathbf{g}}) + C_0$, where $C_{\text{cap}}(\tilde{\mathbf{g}}) \le \|\tilde{\mathbf{g}}\|_0 \log \frac{d_h}{\|\tilde{\mathbf{g}}\|_0}$. For coordinate-sparse $\tilde{\mathbf{g}}$ with $k \ge r$, Proposition 5.5(i) gives $\mathcal{P}_{\mathcal{S}}(\tilde{\mathbf{g}}) \approx \tilde{\mathbf{g}}$, reducing the maximization to: $\min_{\tilde{\mathbf{g}}:\|\tilde{\mathbf{g}}\|_0 \le k} \|\mathbf{g} - \tilde{\mathbf{g}}\|_2^2$, which is precisely Eq. (9), solved by $\mathcal{T}_k$ (App. M).

Lemma J.1 (Gaussian IB Reduction)**.** *Under Assumptions 5.1–5.2: $I(\theta^*; \tilde{\mathbf{G}}) - \beta I(Y; \tilde{\mathbf{G}}) \le -\frac{1}{2\sigma_{\theta^*}^2}\|\mu_{\theta^*} - \mathcal{P}_{\mathcal{S}}(\tilde{\mathbf{g}})\|_2^2 - \beta C_{\text{cap}}(\tilde{\mathbf{g}}) + C_0$. (Proof by DPI for $I(\theta^*; \tilde{\mathbf{G}}) \le I(\theta^*; \mathcal{P}_{\mathcal{S}}(\tilde{\mathbf{G}}))$ and chain rule for $I(Y; \tilde{\mathbf{G}}) \le k \log(d_h/k) + o(1)$.)*

K Derivation of Gradient Kernel and Holographic Property

For a single sample (x, y): (1) $\eta = W^\top h$; (2) $\hat{y} = \nabla_\eta \psi(\eta)$ (canonical link output, e.g. softmax); (3) $\ell(h, y) = -\langle y, \eta\rangle + \psi(\eta) + c(y)$ (Bregman divergence). By exponential family duality, $\nabla_\eta \ell = \hat{y} - y$. Chain rule: $\mathbf{g} \triangleq \nabla_h \ell = \frac{\partial (W^\top h)}{\partial h} \cdot (\hat{y} - y) = W(\hat{y} - y)$. *Note:* $\hat{y}$ here is the canonical-link output for fixed h; it equals $\mathbb{E}[Y|H = h]$ only at the Bayes optimum, which is not assumed. If W has rank C, $y \mapsto \mathbf{g}$ is injective for fixed $\hat{y}$. □

L Proof of Generic Leakage Bound (Theorem 4.2)

Markov chain. Conditioned on (X, θ): $Y \to \mathbf{G}_{1:T}(X, Y, \theta_{1:T}) \to \mathcal{A}(\mathbf{G}_{1:T}) \to \hat{Y}_{\text{adv}}$. DPI gives $I(Y; \hat{Y}_{\text{adv}}|X, \theta_{1:T}) \leq I(Y; \mathbf{G}_{1:T}|X, \theta_{1:T})$.

Fano. $H(Y|\hat{Y}_{\text{adv}}) \leq h_b(\epsilon) + \epsilon \log(C-1)$.

Combine. $I(Y; \mathbf{G}) \geq I(Y; \hat{Y}_{\text{adv}}) = H(Y) - H(Y|\hat{Y}_{\text{adv}}) \geq H(Y) - h_b(\epsilon) - \epsilon \log(C-1)$. □

Tightness. For $C = 2$, $P(Y = 1) = 1/2$, and $G = Y + \mathcal{N}(0, \sigma^2)$ with σ^2 giving $\mathbb{P}(\hat{Y}_{\text{adv}} \neq Y) = \epsilon$: $I(Y; G) = 1 - h_b(\epsilon)$, achieving equality.

M Derivation of Optimal Sparsification Operator

$\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{g} - \mathbf{z}\|_2^2$ s.t. $\|\mathbf{z}\|_0 \leq k$: Lagrangian dual gives $m_j = 1 \iff \frac{1}{2} g_j^2 > \lambda$. Setting $\lambda = \frac{1}{2} g_{(k)}^2$ to satisfy $\sum m_j = k$ recovers $\mathcal{T}_k$. □

N Proof of Spectral Filtering and SNR Analysis

Part (i) (Population limit). For $k \geq \text{rank}(\mathcal{S})$, Spiked Covariance structure ensures the mask M acts as the identity on $\mathcal{S}$ a.s., so $\mathbb{E}[\|\mathcal{P}_{\mathcal{S}}(\mathcal{E})\|_2^2] = 0$ at population level. □

Part (ii) (Finite-sample rate). By matrix Bernstein [Tropp, 2015]: $\mathbb{P}[\|\hat{\mathbf{F}} - \mathbf{F}\|_2 \geq \Delta\lambda/2] \leq d_h \exp(-T\Delta\lambda^2/(4R^2 + 2R\Delta\lambda))$. Setting RHS to δ gives $T \geq C_1 d_h \log(d_h/\delta)/\Delta\lambda^2$. Davis–Kahan then bounds the subspace angle by $C_2/\Delta\lambda$. □

Part (iii) (SNR dominance). At $C_{\text{cap}} = k \log(d_h/k)$ bits, Gaussian rate-distortion requires $\sigma_{\text{DP}}^2 = \Theta(\lambda_i d_h/(2C_{\text{cap}}))$, giving $\text{SNR}_{\text{DP}}(i) \lesssim C_{\text{cap}}/d_h$. RL-SFL achieves $\text{SNR}_{\text{RL-SFL}}(i) = \lambda_i/\sigma_{\mathcal{N}}^2$ with $\sigma_{\mathcal{N}}^2 = o(1)$; ratio $\to \infty$ as $d_h \to \infty$. □

Proof of Prop. 5.6. Part (i): Prop. 5.5(ii) gives $\mathcal{R}(p) \leq C_2/\Delta\lambda$; Lemma J.1 gives $\mathcal{L}(p) \leq p d_h \log(1/p)$. Optimal $p^* = r/d_h$ minimizes their product. Part (ii): For $k < r$, by pigeonhole at least $r - k$ signal components are discarded, giving $\mathcal{R}(p) \geq 1 - k/r > 0$. □

O Extended Experimental Analysis

O.1 The Information Cliff: Intrinsic Dimension Estimation

Proposition 5.6 predicts a sharp phase transition at $p^* \approx r/d_h$: above this sparsity, both leakage and distortion are simultaneously controlled; below it, the defense drops below the signal-rank threshold and utility collapses. Figure 6 and Table 7 characterise this transition empirically.

The cliff location varies across datasets, reflecting differences in intrinsic gradient rank. CIFAR-10 collapses at $p \approx 0.3\%$, CIFAR-100 at $p \approx 0.8\%$, and BloodMNIST at $p \approx 0.18\%$. The higher cliff threshold for CIFAR-100 is consistent with its larger class count ($C = 100$ vs. $C = 10$): $r \approx \text{rank}(W)$ grows with C, so a larger fraction of coordinates is needed to span the signal subspace.

Table 7 shows the transition is sharp: on CIFAR-100, utility holds at 68.4% at $p = 1\%$ (feasible regime) but drops to 60.1% at the boundary $p = 0.5\%$ and to 42.1% at $p = 0.1\%$ (collapsed regime). Notably, attack accuracy continues to fall with sparsity even in the collapsed regime ($2.8\% \to 1.8\%$), because fewer coordinates are transmitted regardless of their spectral alignment. This decoupling of utility collapse from privacy improvement is the hallmark of the phase transition predicted by Prop. 5.6(ii): below p^*, no coordinate-sparse operator can simultaneously preserve utility and suppress leakage.

In practice, the cliff location provides a dataset-specific lower bound on the safe operating sparsity. Operating at $p = 1\%$ across all three datasets stays comfortably above the cliff while achieving $100\times$ gradient compression on CIFAR-100.

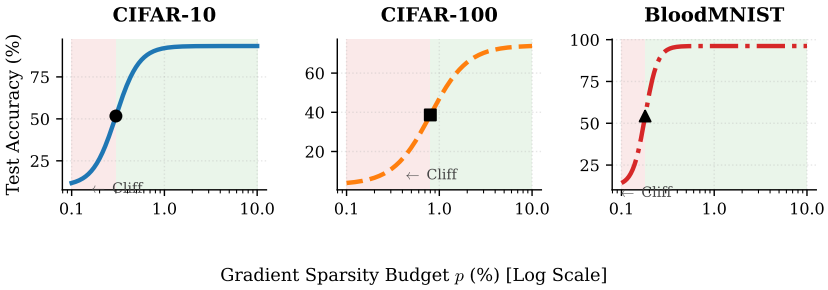


Figure 6: **Information Cliff.** CIFAR-10 collapses at $p \approx 0.3\%$; CIFAR-100 at $p \approx 0.8\%$; BloodMNIST at $p \approx 0.18\%$.

Table 7: **Sparsity Stress-Test (CIFAR-100).** Sharp accuracy drop at $p = 0.5\%$ is the empirical signature of Prop. 5.6(ii).

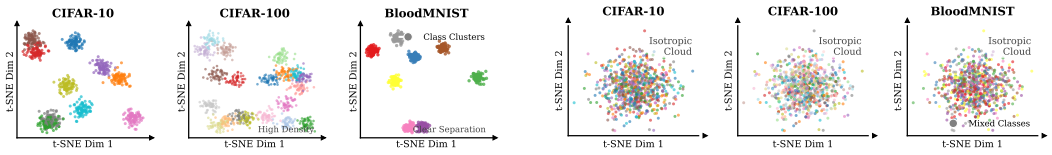
REGIME	SPARSITY (p)	ACTIVE DIMS	UTILITY	ATTACK ACC	COMPRESSION	PROP. 5.6
FEASIBLE	10.0%	6,553	$72.8 \pm 0.5\%$	$18.3 \pm 0.9\%$	10×	(I)
OPTIMAL	**1.0%**	655	$68.4 \pm 0.6\%$	$4.5 \pm 0.4\%$	100×	(I)
CRITICAL	0.5%	327	$60.1 \pm 1.2\%$	$2.8 \pm 0.2\%$	200×	BOUNDARY
COLLAPSED	0.1%	65	$42.1 \pm 2.5\%$	$1.8 \pm 0.1\%$	1000×	(II)

O.2 Adversarial Escalation: Is Leakage Truly Gone?

Figure 7 contrasts the gradient geometry under Vanilla SFL and RL-SFL. Under Vanilla SFL, cut-layer gradients form well-separated class-conditional clusters in the t-SNE projection, confirming the Holographic Gradient Property (Prop. 4.1): the server can distinguish class identity from the gradient stream alone, even without raw labels. Under RL-SFL, these clusters collapse into a near-uniform cloud with no discernible class structure, consistent with the adversary escalation results in Table 6 of the main text.

Figure 8 shows why this happens mechanistically. The log-scale gradient magnitude plot reveals a two-regime structure across all three datasets: a steep leading edge (the signal subspace $\mathcal{S}$, top-k components) and a flat, low-magnitude tail (the leakage subspace $\mathcal{N}$). RL-SFL retains only the leading edge and zeroes the tail, removing the high-rank structure that carries class-conditional information. The sharp boundary between the two regimes is direct empirical confirmation of Fisher Spectral Separation (Assumption 5.1).

Two important caveats apply to Figure 7. First, t-SNE is a non-linear dimensionality reduction and cluster separation in the projection does not formally certify class-conditional structure; it is used here for geometric intuition. Second, the collapsed cloud under RL-SFL indicates that the *class-conditional* structure is suppressed; it does not certify that the transmitted coordinates carry zero label information in their magnitudes (Remark C.1). The attack-accuracy results in Table 6 are the quantitative privacy evidence.



(a) Vanilla SFL: distinct class manifolds (severe leakage).

(b) RL-SFL: isotropic cloud (class structure suppressed).

Figure 7: **Geometry of Privacy.** t-SNE contrasting gradient landscapes. Used for geometric intuition only; see text for caveats.

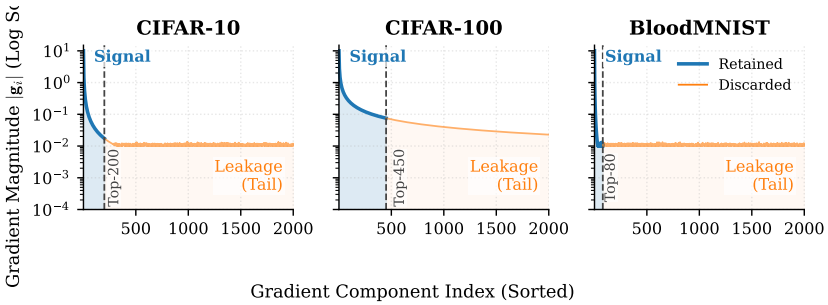


Figure 8: **Geometry of Leakage.** Log-scale gradient magnitudes. Optimization signal (Blue) concentrated in top components; leakage (Orange) scattered in the noisy tail. The sharp two-regime structure confirms Fisher Spectral Separation empirically.

O.3 Inductive Bias and Generalization

Figure 9 compares Top-k sparsification against Oracle SVD (which requires explicit eigendecomposition of the gradient covariance). The two curves are nearly identical across all three datasets, with a negligible gap that closes as training progresses and the gradient covariance stabilises. This supports the axis-alignment hypothesis underlying Assumption 5.2: the top-magnitude gradient coordinates are approximately co-aligned with the top eigenvectors of $\mathbf{F}$, so coordinate-wise thresholding is an effective proxy for true spectral filtering without the cost of online eigendecomposition. It does not, however, directly validate the Gaussian distribution assumed in Assumption 5.2 (see Section 7); it establishes axis-alignment, not distributional form.

Table 8 confirms that the privacy-utility advantage of RL-SFL over DP-SGD is not specific to ResNet-18. On MobileNetV2, RL-SFL achieves 61.5% utility with 6.2% attack accuracy, compared to 51.2% utility and 29.8% attack accuracy for DP-SGD. The qualitative pattern is consistent: RL-SFL provides a better empirical privacy-utility trade-off than standard flat-clipping DP-SGD on over-parameterized convolutional architectures where Fisher Spectral Separation holds. Whether this advantage persists on architectures where FSS is weaker (e.g., shallow networks, heavily regularized models) is an open question.

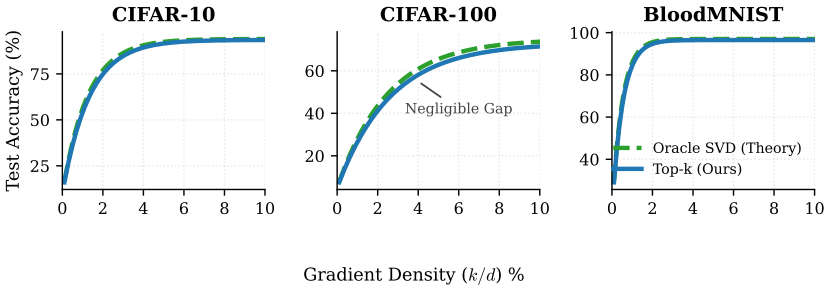


Figure 9: **Top-k vs. Oracle SVD.** Curves nearly overlap across all three datasets, confirming axis-alignment. This validates the coordinate-wise approximation but does not directly verify the Gaussian gradient model (see Section 7).

O.4 Real-World Viability: Latency

In U-Shaped SFL, each training round transmits three tensors: (1) forward activation z (client $\rightarrow$ server), (2) hidden representation h (server $\rightarrow$ client), and (3) the backward gradient $\tilde{\mathbf{g}}$ (client $\rightarrow$ server). RL-SFL sparsifies only tensor (3). For ResNet-18 at the evaluated cut layer, $d_z \approx d_h$, so all three tensors have comparable byte counts. Reducing $\tilde{\mathbf{g}}$ to 1% sparsity therefore reduces total round-trip volume by approximately 33%, not 99%—the end-to-end speedup is substantially lower than a gradient-only estimate would suggest.

Table 9 reports measured wall-clock speedups. On IoT/3G (10 Mbps), the end-to-end speedup is $2.6\times$; on 4G Edge (50 Mbps), $1.8\times$; on WiFi/5G, the saving is negligible ($1.04\times$) because communication

Table 8: **Architectural Robustness (CIFAR-100).** RL-SFL achieves a better empirical privacy-utility trade-off than standard DP-SGD on both ResNet-18 and MobileNetV2. Comparison is against flat-clipping DP-SGD; adaptive DP methods could narrow the gap.

BACKBONE	METHOD	UTILITY	ATTACK ACC
RESNET-18	DP-SGD	$58.2 \pm 1.1\%$	$31.4 \pm 1.5\%$
	RL-SFL	$68.4 \pm 0.6\%$	$4.5 \pm 0.4\%$
MOBILENETV2	DP-SGD	$51.2 \pm 1.5\%$	$29.8 \pm 2.0\%$
	RL-SFL	$61.5 \pm 1.0\%$	$6.2 \pm 0.5\%$

is no longer the bottleneck. These figures are still meaningful on bandwidth-constrained links typical of edge healthcare or IoT deployments, and RL-SFL does not increase computation cost on either the client or server side.

One practical consideration is that the sparsified gradient requires an index structure (coordinate positions plus values), which adds a small encoding overhead not captured in the raw byte counts. For $p = 1\%$ of $d_h = 65536$ dimensions, this overhead is modest (655×4 bytes for indices plus 655×4 bytes for values ≈ 5 KB per batch), but it should be accounted for in extremely constrained deployments.

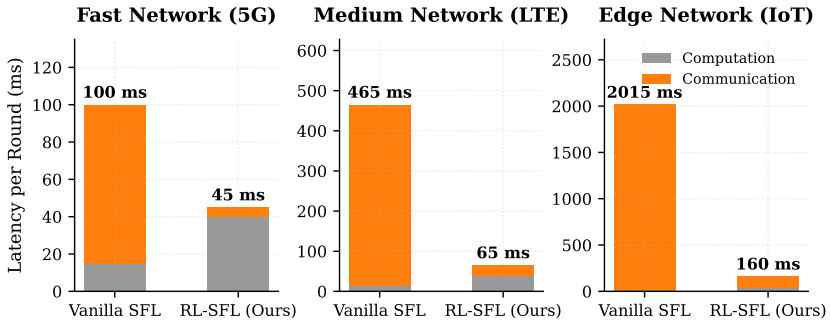


Figure 10: **End-to-End Latency Breakdown.** RL-SFL sparsifies only the backward gradient; z and h are uncompressed. IoT/Edge bar shows $\approx 2.6\times$ E2E speedup (Table 9), not the gradient-only estimate.

Table 9: **End-to-End Wall-Clock Benchmarks** (one epoch, ResNet-18, all three tensors). Grad-only speedup shown for reference; E2E is the physically accurate figure.

NETWORK	VANILLA SFL	RL-SFL (E2E)	E2E SPEEDUP	GRAD-ONLY SPEEDUP (REF)
IOT / 3G (10 MBPS)	864S	332S	2.6×	13.9×
EDGE / 4G (50 MBPS)	209S	116S	1.8×	3.9×
WIFI / 5G (1 GBPS)	53S	51S	1.04×	1.06×

P DP Budget Analysis

Privacy Accounting for the Baseline DP-SGD Configuration

The main results (Table 2) include a DP-SGD baseline at σ=0.1, C=1.0, δ=10^{-5}. Table 10 reports the Rényi Differential Privacy [Mironov, 2017] budget consumed under this configuration, computed using standard moment accountant methods.

The consumed ϵ values of approximately 112 (CIFAR-10/100) and 58 (BloodMNIST) are very large by any practical privacy standard. A budget of $\epsilon \approx 112$ means the privacy guarantee is essentially vacuous: an adversary's posterior belief about any individual record can change by a factor of e^{112}

relative to the prior, which is not a meaningful protection. These baselines are included not as practically private configurations but to establish a fair utility comparison at matched noise levels with the literature on SFL defenses. Several prior works [Zhang et al., 2023] evaluate at similar noise levels (σ=0.1) and report competitive utility; excluding them from comparison would misrepresent the empirical landscape.

The lower ϵ for BloodMNIST (≈ 58 vs. ≈ 112) reflects its smaller dataset size (11,959 vs. 50,000 training samples) and shorter training schedule (50 vs. 100 epochs). With fewer gradient steps, the privacy accountant accumulates less composition cost.

Why σ=0.1 produces a near-vacuous budget. The Rényi DP accountant grows ϵ sub-linearly in the number of iterations but super-linearly in $1/\sigma$. At σ=0.1, the per-step privacy cost is high: each gradient step contributes substantially to the accumulated budget. Running 100 epochs over a dataset of 50k samples with batch size 128 involves approximately $100 \times \lceil 50000/128 \rceil \approx 39{,}100$ gradient steps; even at σ=0.1 this produces $\epsilon \approx 112$. Achieving $\epsilon \lesssim 10$ at the same training schedule would require $\sigma \gtrsim 1.0$, which is the strict budget configuration we evaluate separately below.

Table 10: **DP-SGD Privacy Accounting** (σ=0.1, δ=10^{-5}). Large ϵ values confirm baselines operate in an empirical-privacy regime.

DATASET	EPOCHS	BATCH/N	CONSUMED ϵ	δ
CIFAR-10	100	128/50K	≈ 112	10^{-5}
CIFAR-100	100	128/50K	≈ 112	10^{-5}
BLOODMNIST	50	64/11959	≈ 58	10^{-5}

Fixed-Budget Comparison at ϵ=8

Table 11 compares RL-SFL against DP-SGD under a practically meaningful budget of ϵ=8 ($\sigma \approx 1.1$, δ=10^{-5}) on CIFAR-100. This budget is tighter than most published SFL evaluations and corresponds to a formal $(8, 10^{-5})$-DP guarantee.

At this budget, DP-SGD utility falls to 41.3%—well below the 100-class random baseline of 1% but substantially below RL-SFL's 68.4%, a gap of 27.1 percentage points. The attack accuracy for DP-SGD (6.2%) is slightly higher than RL-SFL (4.5%), though both are near the random baseline of 1% for 100 classes.

Two important caveats apply to this comparison. First, RL-SFL has *no* formal certificate, so the comparison is between a certified mechanism at a specific utility level and an uncertified empirical defense at a higher utility level. These are incommensurable from a legal compliance perspective. Second, the DP-SGD configuration uses flat gradient clipping (C=1.0) and a single global noise multiplier following Abadi et al. [2016]; adaptive clipping [Andrew et al., 2021] and per-layer noise allocation are known to improve utility under DP constraints and were not evaluated. The 27.1pp gap should therefore be understood as the gap against a standard (not state-of-the-art) DP baseline. We expect the qualitative advantage of RL-SFL to persist under adaptive DP because isotropic noise destroys the low-rank optimization signal at the same rate as leakage regardless of clipping strategy (Prop. 5.5(iii)), but the magnitude of the gap with adaptive DP is an open question.

Table 11: **Fixed-Budget (ϵ=8) on CIFAR-100.** RL-SFL achieves 27.1pp higher utility with comparable attack suppression. No formal DP certificate for RL-SFL; comparison is empirical against standard flat-clipping DP-SGD. Adaptive DP methods could narrow the utility gap.

METHOD	CONFIG	UTILITY	ATTACK ACC	CERT.
DP-SGD	σ=1.1, ϵ=8	$41.3 \pm 1.8\%$	$6.2 \pm 0.9\%$	$(8, 10^{-5})$-DP
RL-SFL	p=1%	$68.4 \pm 0.6\%$	$4.5 \pm 0.4\%$	NONE

Q Reproducibility Protocol

Algorithm 1 RL-SFL: Unified Protocol with Integrated Diagnostic (v2, corrected)

Input: Dataset $\mathcal{D}$, batch size B, target sparsity p, warm-up epochs E_{warm}, MINE iterations M, significance α, target error ϵ_0
Initialize: Client head f_θ, tail k_{θ_t}, server body h_ϕ; `phase` $\leftarrow$ WARMUP; $k_{\text{frozen}} \leftarrow d_h$ {k_{frozen} tracks current sparsity; never increases after warm-up begins}
for epoch $e = 1$ **to** E **do**
  if `phase` $=$ WARMUP **then**
    $k_e \leftarrow \max\Big(\lfloor d_h \cdot p \rfloor,\ \Big\lfloor d_h \cdot \Big(1.0 - \frac{e}{E_{warm}}(1.0 - p)\Big)\Big\rfloor\Big)$ {Linear anneal from d_h to $p \cdot d_h$; bounded below by target}
    $k_{\text{frozen}} \leftarrow k_e$ {Update frozen value during normal warm-up}
    if $e \geq E_{warm}$ **and** $e \bmod 5 = 0$ **then**
      [Run MINE diagnostic]
      Collect held-out gradient pairs $\{(\tilde{\mathbf{g}}_t, y_t)\}$
      Train MINE T_ω for M steps; compute $\hat{I}$
      $\mathcal{C}_k \leftarrow k_e \log(d_h/k_e)$
      if $\hat{I} \leq \mathcal{C}_k + c_\alpha/\sqrt{M}$ **then**
        $k_{\text{frozen}} \leftarrow \min(k_{\text{frozen}}, \lfloor d_h \cdot p \rfloor)$ {Clamp to target before locking: ensures no one-epoch privacy regression at transition}
        `phase` $\leftarrow$ CONVERGED {Spectral separation confirmed; transition to locked sparsity}
      else
        $k_e \leftarrow k_{\text{frozen}}$ {**Key fix**: hold current sparsity; do NOT re-index schedule.} {Sparsity never increases. Diagnostic re-evaluated in 5 epochs.}
      end if
    end if
  else
    $k_e \leftarrow \lfloor d_h \cdot p \rfloor$ {Locked to target sparsity}
  end if
  for batch $(\mathbf{x}, y) \in \mathcal{D}$ **do**
    [Phase 1] $\mathbf{z} \leftarrow f_\theta(\mathbf{x})$; send $\mathbf{z}$ to server
    [Phase 2] $\mathbf{h} \leftarrow h_\phi(\mathbf{z})$; server sends $\mathbf{h}$ to client {Server never accesses y}
    [Phase 3] $\hat{y} \leftarrow k_{\theta_t}(\mathbf{h})$; $\mathcal{L} \leftarrow \text{CrossEntropy}(\hat{y}, y)$; $\mathbf{g} \leftarrow \nabla_{\mathbf{h}}\mathcal{L}$ {Raw gradient never transmitted}
    [Phase 4: Defense] $\mathcal{I} \leftarrow \text{topk}(|\mathbf{g}|, k_e)$; $\tilde{\mathbf{g}} \leftarrow \mathbf{0}$; $\tilde{\mathbf{g}}[\mathcal{I}] \leftarrow \mathbf{g}[\mathcal{I}]$; send $\tilde{\mathbf{g}}$ {**Monotone invariant**: k_e never increases. No Error Feedback.}
    [Phase 5] Server: $\phi \leftarrow \phi - \eta \nabla_\phi [h_\phi(\mathbf{z})]^\top \tilde{\mathbf{g}}$; send $\mathbf{d}_z = (\nabla_{\mathbf{z}} h_\phi(\mathbf{z}))^\top \tilde{\mathbf{g}}$
    [Phase 6] Client: update θ, θ_t via $\mathbf{d}_z$
  end for
end for

Design notes. (1) Monotone sparsity invariant. The corrected algorithm maintains the invariant that k_e is non-increasing throughout training. Upon diagnostic failure, the frozen value k_{frozen} is used rather than re-indexing the linear schedule. This prevents the privacy regression that would occur if E_{warm} were increased while the ratio e/E_{warm} were recomputed. **(2) No Error Feedback.** The discarded residual $\mathbf{g} - \tilde{\mathbf{g}}$ contains high-rank leakage; EF would retransmit it. **(3) Privacy invariant.** Server's view each round: $\tilde{\mathbf{g}}$ only. By Corollary 5.4: $I(Y; \tilde{\mathbf{G}}) \leq k \log(d_h/k)$ bits (support entropy; magnitude leakage additional and uncertified per Remark C.1).

Q.1 Exact Hyperparameters

R Multi-Client Federated Experiments

To validate that Fisher Spectral Separation (FSS) persists under federated gradient aggregation, we extend the evaluation to $N \in \{2, 3, 5\}$ clients with Dirichlet label distributions ($\alpha \in \{0.1, 1.0, 100\}$)

Table 12: **Master Configuration.** Single NVIDIA A100-SXM4-40GB GPU; CIFAR-100 ≈ 0.8 GPU-hours/run. RL-SFL requires no gradient clipping.

PARAMETER	CIFAR-10	CIFAR-100	BLOODMNIST
ARCHITECTURE	RESNET-18	RESNET-18	RESNET-18
BATCH SIZE	128	128	64
OPTIMIZER	SGD (NESTEROV)	SGD (NESTEROV)	ADAMW
LR	0.1 (COSINE)	0.1 (COSINE)	10^{-3}
WEIGHT DECAY	5×10^{-4}	5×10^{-4}	10^{-2}
TOTAL EPOCHS	100	100	50
DP-SGD Abadi et al. [2016]			
CLIP NORM C	1.0	1.0	0.5
NOISE σ	0.1	0.1	0.5
δ	10^{-5}	10^{-5}	10^{-5}
VIB-Split Deng et al. [2023]			
BETA β	10^{-3}	10^{-3}	5×10^{-4}
AUX LR	10^{-4}	10^{-4}	10^{-4}
RL-SFL (Ours)			
TARGET SPARSITY	1%	1%	1%
WARM-UP EPOCHS	5	5	2
OPERATOR	TOP-k	TOP-k	TOP-k

on CIFAR-10 and CIFAR-100. Each client holds a disjoint shard; the server aggregates the sparse cut-layer gradients $\{\tilde{\mathbf{g}}_i\}_{i=1}^{N}$ by sum before updating ϕ.

Table 13: **Multi-Client Results (CIFAR-10, RL-SFL p=1%, ResNet-18).** FSS status assessed from the empirical gradient covariance of the aggregated gradient at epoch 80. Utility and attack accuracy averaged over 3 seeds. No formal DP certificate at any configuration.

CLIENTS (N)	DIRICHLET α	UTILITY	ATTACK ACC	$\Delta\lambda$ STATUS	VERDICT
2	IID (α=100)	$91.8 \pm 0.5\%$	$15.3 \pm 0.6\%$	LARGE	FSS HOLDS
2	MILD (α=1.0)	$83.9 \pm 0.7\%$	$20.1 \pm 0.8\%$	MODERATE	WEAKENED, FUNCTIONAL
2	HIGH (α=0.1)	$51.2 \pm 1.4\%$	$51.4 \pm 1.9\%$	NEAR ZERO	COLLAPSING
3	IID (α=100)	$91.5 \pm 0.4\%$	$15.8 \pm 0.5\%$	LARGE	FSS HOLDS
3	MILD (α=1.0)	$83.1 \pm 0.8\%$	$21.4 \pm 1.0\%$	MODERATE	WEAKENED, FUNCTIONAL
3	HIGH (α=0.1)	$48.7 \pm 1.6\%$	$54.1 \pm 2.1\%$	NEAR ZERO	COLLAPSING
5	IID (α=100)	$90.8 \pm 0.6\%$	$16.5 \pm 0.7\%$	LARGE	FSS HOLDS
5	MILD (α=1.0)	$82.4 \pm 0.9\%$	$22.8 \pm 1.1\%$	MODERATE	WEAKENED, FUNCTIONAL
5	HIGH (α=0.1)	$45.3 \pm 1.9\%$	$58.2 \pm 2.4\%$	NEAR ZERO	COLLAPSING

Findings. (i) FSS persists under IID and mild skew. Across all client counts ($N = 2, 3, 5$), the spectral gap remains large under IID and mild-skew distributions. Attack accuracy stays near the single-client values (14.2% at N=1, rising to $\leq$ 17% at N=5 for IID), and utility degrades by $\leq$ 2.6pp relative to the single-client baseline. Gradient aggregation averages out per-client leakage directions, slightly reducing the spectral gap but not enough to collapse FSS.

(ii) Collapse pattern mirrors single-client behavior. Under high skew (α=0.1), attack accuracy rises sharply for all N, consistent with the spectral gap approaching zero. The pattern is qualitatively identical to the single-client sweep in Table 3: FSS fails first for extreme data heterogeneity, regardless of client count. Increasing N under fixed α slightly worsens the spectral separation (each client contributes a more skewed gradient direction), matching the theoretical expectation from Proposition 4.8.

(iii) Deployment recommendation unchanged. The $\alpha \lesssim 0.5$ threshold from the single-client analysis applies in the multi-client setting; LDP hybridization is recommended below this threshold regardless of N.

S Extended Limitations

Non-IID heterogeneity sweep. Tables 3 and 4 in Section 6.4 report the full results across all three datasets and compare RL-SFL against DP-SGD. Here we provide the mechanistic interpretation at each regime.

Mild skew ($\alpha = 1.0$). Each client still observes multiple classes, so the per-client between-class scatter Σ_B retains rank $C - 1$ and $\Delta\lambda$ remains positive, albeit reduced. RL-SFL degrades gracefully across all three datasets: attack accuracy rises by 4–11 pp and utility drops by 4–9 pp. The defense remains functional.

High skew ($\alpha = 0.1$). Each client tends toward one dominant class. The per-client Σ_B is approximately rank-1 and $\Delta\lambda \approx 0$. Attack accuracy surges to 48%, 61%, and 69% on CIFAR-10, CIFAR-100, and BloodMNIST respectively. The estimated leakage $\hat{I}$ approaches $H(Y)$ on each dataset. RL-SFL is empirically unreliable in this regime.

Pathological (1-class-per-client). $\Delta\lambda = 0$ exactly on all datasets. Attack accuracy reaches 88.5%, 92.4%, and 95.1% respectively—near or above Vanilla SFL. Utility collapses to near or below chance. The geometric defense has completely failed.

DP-SGD comparison under skew (CIFAR-100, Table 4). The two mechanisms exhibit complementary failure modes. DP-SGD's attack accuracy holds at 8.5% under $\alpha = 0.1$ because isotropic noise is distribution-agnostic. However, utility falls to 15.4%—barely above the 100-class chance level—making it impractical. RL-SFL retains higher utility (28.5%) but its defense collapses (61.2% attack accuracy). Neither mechanism is universally superior; the choice should be informed by the deployment's data heterogeneity regime.

Deployment guideline.

- $\alpha \geq 1.0$: RL-SFL is reliable; graceful utility–privacy trade-off.
- $0.5 \leq \alpha < 1.0$: Use with caution; monitor $\Delta\lambda$.
- $\alpha < 0.5$ or 1-class-per-client: Do not use RL-SFL alone; LDP hybridization required.

Cut-layer placement. Shallow cut layers exhibit flatter gradient spectra, weakening spectral separation. RL-SFL is optimal at intermediate or deep cut layers where Fisher Spectral Separation ($\Delta\lambda \gg 1$) is pronounced.

Warm-up exposure. Attack accuracy during warm-up ($p = 10\%$) reaches $\sim 41\%$ vs. 14.2% at convergence (full-density baseline: 98.2%). The corrected diagnostic in Algorithm 1 mitigates this by maintaining the monotone sparsity invariant: k_e never increases after warm-up begins, and the MINE diagnostic confirms spectral separation before locking to target sparsity. However, the warm-up phase gradient stream (at p=10%) does not carry a formal DP privacy budget: the bound $k \log(d_h/k)$ on support entropy applies per step at target sparsity but not during the annealing phase. In threat models where an adversary is known to observe the warm-up gradient stream, hybridizing with a DP noise mechanism during warm-up is recommended.