ADAPTIVE SPARSE SUPPORT AS A CALIBRATION-DEPENDENT LABEL CHANNEL IN SPLIT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse gradients are usually discussed in terms of the values they retain. In split learning, however, the coordinates chosen for transmission can themselves reveal the client's label. We study when this support channel is exploitable and how it interacts with the aligned values. Across our experiments, label decoding is strong when the server has labeled calibration messages from the same checkpoint or training lineage, but it largely collapses across independent initializations. Our experiments also overturn our initial theoretical explanation. On Diamonds, same-label cross-fit fidelity is negative in every seed even though ordinary support decoding reaches .795; the attack is instead driven by class-conditional coordinate frequencies. A decoder-free test based on one auxiliary-selected coordinate gives 95% lower bounds of .140, .094, and .334 bits on Digits, Synthetic-10, and Diamonds. At epoch 48 on compact CIFAR-10, support decoding remains .919 and joint decoding .998. Data-independent supports close the support channel at small measured utility cost, yet values remain predictive in every useful-utility trainer we test. Support is therefore an additional privacy risk, not the whole leakage story. Our audit covers support and values for the declared row-wise sparse backward message, not the complete training transcript.

1 INTRODUCTION

Split learning partitions a model so that labels and a private prediction tail remain on the client while an intermediate body is executed by a server. Label inference from split representations and gradients is established (Li et al., 2022; Kariyappa & Qureshi, 2023), and sparsification is widely used to reduce communication (Zheng et al., 2023; Oh et al., 2025). Their interaction creates a distinct privacy question: the wire object is the pair (M, V) of selected coordinates and coordinate-aligned values, not merely the retained values.

Sensitive sparse positions are not new. Prior work studies access-pattern leakage, private top-k, joint protection of sparse values and indices, and coordinate-hiding secure aggregation for sparsified updates (Kato et al., 2023; Vithana & Ulukus, 2023; 2024; Silva Carvalho et al., 2020; Qiao et al., 2021; Gillenwater et al., 2022; Zhu & Wang, 2022; Sami & Güler, 2024). Our narrower contribution is label-specific and protocol-specific. We ask when adaptive *cut-gradient* supports transfer within a class, how that transfer constrains information leakage, and when a calibration-capable server can exploit it.

The paper makes four contributions. First, we prove an independent same-label cross-fit implication and then falsify it as an explanation of our strongest attack: the larger Diamonds gap is negative. This is a sufficient population pathway, not the empirical attack mechanism. Second, we identify class-conditional coordinate frequency as that mechanism and certify it without a fitted decoder: an auxiliary-selected coordinate, data processing, and simultaneous exact binomial intervals give a finite-sample information lower bound on untouched test masks. Third, we identify two operational boundaries: calibration is strong within a checkpoint lineage but fails across independent initializations, and cumulative leakage requires linkable rows or informative false matches. Fourth, we audit support, aligned values, and their joint observation for the declared row-wise sparse backward message against data-independent selectors, a support-DP selector, TPSL, conventional gradient features, exact sparsity endpoints, compressed training, heterogeneous clients, multiple cuts, a real-large task, and a 48-epoch modern-image check.

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

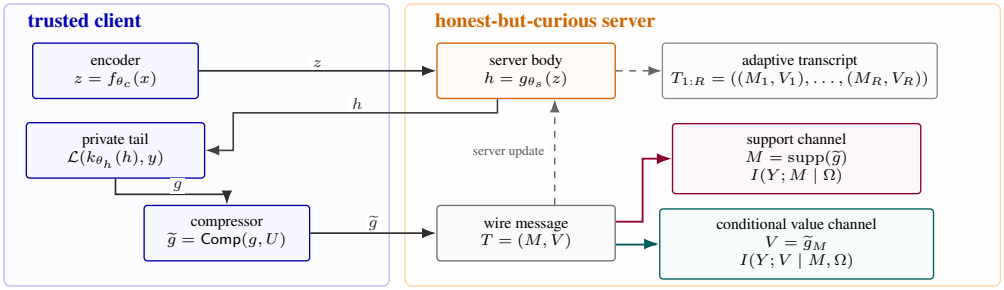


Figure 1: A sparse backward message exposes a support channel and a conditional value channel. Repeated releases form a transcript. In the row-wise protocol studied here, a minibatch message is a $B \times d$ tensor with separately recoverable example rows; flattening or averaging before compression defines a different wire object.

We organize the evidence into three levels. *Population mechanism* means transferable support fidelity implies information. *Fixed-audit certificate* means a frozen calibrated decoder certifies information for one declared checkpoint, auxiliary set, class weighting, and held-out population. *Operational transfer* asks whether that decoder remains exploitable when lineage, tail, class coverage, or row linkage changes. No result at one level is presented as evidence for another.

Three findings organize the paper. Adaptive dependence creates a graded privacy–distortion frontier rather than a binary "sparse/private" distinction. Class-conditional selection frequency, not same-label energy transfer, explains the strongest mask attack. Finally, support leakage is not inevitable: for a binary linear softmax tail in exact arithmetic, all nonzero per-example gradients are scalar multiples of one vector, so deterministic magnitude support is constant. The practical claim is precise: adaptive support is an additional label channel for a calibration-capable server, while the aligned value channel can already dominate and must be protected independently.

2 Protocol, Context, and Threat Model

Let $Y \in \{1, \ldots, C\}$ be private. A client encoder produces Z, a server body returns H, and a private client tail evaluates loss $\mathcal{L}$. The cut gradient is $G = \partial\mathcal{L}/\partial H \in \mathbb{R}^d$. A randomized sparse compressor releases

$$\widetilde{G} = \mathsf{Comp}(G, U), \qquad M = \mathrm{supp}(\widetilde{G}), \qquad V = \widetilde{G}_M. \tag{1}$$

The honest-but-curious server observes (M, V, Ω). The public context Ω contains the complete server-known state relevant to the message distribution: architecture and cut, all server-known checkpoint parameters and optimizer state, round, client identity, prior transcript, compressor configuration, public randomness, and any row-linkage metadata. When a quantity is written conditional on $\Omega = \pi$, the trained instance and this state are fixed.

Adversaries and auxiliary access. A white-box diagnostic knows a linear private tail and prediction vector. A learned auxiliary decoder receives labeled messages generated by the same protocol; a prototype decoder uses only classwise support frequencies. This is an *audit-capable* server model, stronger than the standard honest-but-curious server that observes only unlabeled traffic. Plausible calibration sources include public or consented examples processed by the target checkpoint, operational calibration traffic with labels, or colluding clients, but none is assumed to be automatic. The decoder does not receive private-tail weights. We test auxiliary sample size, class absence, preprocessing mismatch, checkpoint staleness, private-tail mismatch, and independent initialization. The strongest practical claim is therefore about a calibration-capable server with target-checkpoint or same-lineage messages, not a server with architecture knowledge or unlabeled traffic alone.

Generating a labeled calibration gradient and decoding it are distinct capabilities: the gradient-generating client, audit service, or colluding party must execute a compatible private tail, although the fitted decoder need not receive those tail weights. One concrete non-target sequence is: consenting or colluding clients use the shared checkpoint family on labeled records; the server joins their labels to the resulting messages, fits a decoder, and applies it to a different client's later unlabeled

messages. The target does not cooperate and the server need not observe tail weights, but shared lineage and calibration-class coverage remain essential. Table 2 specifies the other scenarios and their H/V/U-shaped boundaries.

For repeated-message attacks, the target is the same private example observed at several checkpoints. Perfect linkage can arise from stable minibatch order, persistent sample or session identifiers, or protocol metadata. It is not automatic: we separately randomize earlier-row identities and sweep the probability of correct linkage.

Because (M, V) losslessly represents the sparse message,

$$I(Y; M, V \mid \Omega) = I(Y; M \mid \Omega) + I(Y; V \mid M, \Omega). \tag{2}$$

The joint decoder therefore receives $[M, \widetilde{G}]$. Sorting values and discarding coordinate identity is retained only as an ablation, not as an estimate of $I(Y; V \mid M, \Omega)$. We report prior-adjusted support advantage $\mathrm{Adv} = (a - p)/(1 - p)$, where a is attack accuracy and p is the majority prior under the same context.

3 Mechanism Bounds and Information Certificates

3.1 Cross-fitting removes example-specific self-fit

Let G, G' be conditionally i.i.d. given (Y, Ω), and draw $M \sim Q(\cdot \mid G, \Omega)$ from the first gradient. For $S \subseteq [d]$, define normalized omitted energy

$$\ell_{\mathrm{om}}(g, S) = 1 - \frac{\|P_S g\|_2^2}{\|g\|_2^2} \in [0, 1], \qquad \ell_{\mathrm{om}}(0, S) = 0. \tag{3}$$

For $\mathcal{S}_k = \{S : |S| \le k\}$ and fixed context π, let

$$D_{\mathrm{cf}}(\pi) = \mathbb{E}[\ell_{\mathrm{om}}(G', M) \mid \Omega = \pi], \tag{4}$$

$$D_0(\pi) = \inf_{q(\cdot \mid \pi) : S \perp\!\!\!\perp Y \mid \Omega} \mathbb{E}[\ell_{\mathrm{om}}(G', S) \mid \Omega = \pi]. \tag{5}$$

The objective is linear in q, so a deterministic label-independent support attains D_0.

Theorem 3.1 (Cross-fitted privacy–distortion bound)**.** *For any label prior and randomized support mechanism,*

$$I(Y; M \mid \Omega) \ge \frac{2}{\ln 2} \mathbb{E}_\Omega\big[(D_0(\Omega) - D_{\mathrm{cf}}(\Omega))_+^2\big] \quad \text{bits}. \tag{6}$$

The proof uses the marginal support law as a feasible label-independent comparator, the variational characterization of total variation, Pinsker, and Jensen (Appendix B.1). It assumes neither Gaussian gradients nor deterministic top-k. The independent target G' is the substantive modeling step: evaluating on G itself could reward arbitrary per-example support memorization. Conditional i.i.d. sampling defines a fixed-context audit population; it is not an assumption that consecutive training gradients are independent. Within-class heterogeneity can make the transferable gap small even when a flexible attack uses other distributional features. The exact construction in Appendix B.4 quantifies both effects.

Proposition 3.2 (Mixture-selector bounds with an arbitrary label-independent branch)**.** *Fix context $\Omega = \pi$. Let Q_b be any label-independent support law with cross-fitted distortion $D_b(\pi)$, let Q_1 be any adaptive selector with cross-fitted distortion $D_1(\pi)$, and draw $B \sim \mathrm{Bernoulli}(\lambda)$ independently of (Y, G, G'). Release $M_\lambda \sim Q_B$. Then*

$$D_\lambda(\pi) = (1 - \lambda) D_b(\pi) + \lambda D_1(\pi). \tag{7}$$

The cross-fit theorem gives the lower bound

$$\frac{2}{\ln 2}[D_0(\pi) - D_\lambda(\pi)]_+^2 \le I(Y; M_\lambda \mid \Omega = \pi), \tag{8}$$

and convexity of mutual information in the channel gives the separate upper bound

$$I(Y; M_\lambda \mid \Omega = \pi) \le \lambda I(Y; M_1 \mid \Omega = \pi). \tag{9}$$

If Q_b attains $D_0(\pi)$, Equation (8) *simplifies to $2\lambda^2[D_0(\pi) - D_1(\pi)]_+^2 / \ln 2$. In all empirical mixture attacks the branch bit B is sampled independently per row and is* not *transmitted to the decoder; the plotted and tabulated lower curve is Equation* (8) *with the measured D_b. If B were revealed, then $I(Y; M_\lambda, B \mid \Omega = \pi) = \lambda I(Y; M_1 \mid \Omega = \pi)$. Proofs appear in Appendix B.2.*

Proposition 3.3 (Dimension-free finite-sample cross-fit bound)**.** *At fixed context, let* $q_i \in [0,1]^d$ *be normalized coordinate-energy vectors with* $\sum_j q_{ij} = 1$*, let* $\widehat{F} = \max_{|S|\le k} n^{-1} \sum_i \sum_{j\in S} q_{ij}$ *and* $\widehat{D}_0 = 1 - \widehat{F}$*, and let* $\widehat{D}_{\mathrm{cf}}$ *be the mean of* n *independent cross-fitted losses in* $[0,1]$*. With* $\epsilon_n = \sqrt{\log(2/\alpha)/(2n)}$*, with probability at least* $1-\alpha$*,*

$$I(Y; M \mid \Omega) \ge \frac{2}{\ln 2} \left[\widehat{D}_0 - \widehat{D}_{\mathrm{cf}} - 2\epsilon_n \right]_+^2 . \tag{10}$$

Unlike a coordinatewise union bound, Equation (10) is independent of d and k: replacing one simplex-valued sample changes $\widehat{F}$ by at most $1/n$. It is nevertheless zero on every evaluated primary dataset, a sample-complexity limitation of the present certificate rather than evidence of privacy (Appendices B.3 and B.4). The three cross-fit quantities have distinct status: Theorem 3.1 is a population implication; the reported $D_0 - D_{\mathrm{cf}}$ plug-in is an oracle descriptive diagnostic because it optimizes on evaluation targets; and Proposition 3.3 is statistically valid but vacuous here. Neither is used to explain the separate frequency-DPI or frozen-decoder Fano certificates.

3.2 A FREQUENCY CERTIFICATE FOR THE OBSERVED ATTACK MECHANISM

Choose one coordinate $j^\star$ using only labeled auxiliary masks and freeze it before test access. Let $B = M_{j^\star}$, let $p_y = \Pr(B = 1 \mid Y = y, \Omega)$, and fix audit weights w_y.

Proposition 3.4 (Decoder-free coordinate-frequency certificate)**.** *On an independent stratified test set, let* $[\ell_y, u_y]$ *be simultaneous* $(1-\alpha)$ *Clopper–Pearson intervals for every* p_y*. Then, with probability at least* $1-\alpha$*,*

$$I(Y; M \mid \Omega) \ge I(Y; B \mid \Omega) \ge \frac{2}{\ln 2} \min_{p_y \in [\ell_y, u_y]} \sum_y w_y \left(p_y - \sum_c w_c p_c \right)^2 . \tag{11}$$

The first inequality is data processing; the second applies Pinsker to each Bernoulli class-conditional law. We use Bonferroni-exact two-sided intervals, so coordinate search costs no test multiplicity because it occurs only on the auxiliary split. Unlike cross-fitted distortion, this certificate targets class-conditional selection frequency directly, uses no fitted decoder, and is positive in every seed on Digits, Synthetic-10, and Diamonds (Appendix B.5).

3.3 FINITE-SAMPLE FANO CERTIFICATE AND ITS SCOPE

Fix $\Omega = \pi$, auxiliary data, and a decoder φ before touching an independent test sample. If $e_U(\pi)$ is a one-sided upper confidence bound on its audit-population error, then

$$I(Y; M \mid \Omega = \pi) \ge \max\left\{ 0, \min_{0 \le u \le e_U(\pi)} H(Y \mid \pi) - h_2(u) - u \log_2(C-1) \right\}. \tag{12}$$

The 95% statement covers held-out sampling for this fixed audit instance. It does *not* provide 95% coverage over training initialization, split choice, architecture selection, or client populations. We therefore compute the certificate separately for every trained seed and report between-seed mean and standard deviation as a distinct uncertainty source. If context varies within an audit, it is supplied to the decoder and weighting is performed within context strata (Appendix B.6).

A complementary support-list certificate is retained for settings where a low-overlap class codebook exists. It is weaker in our data and is presented in Appendix B.7, including zero outcomes; it is not the headline quantitative measure.

Proposition 3.5 (Value perturbation and transcript monotonicity)**.** *Let* M *be selected before an arbitrary stochastic value release* $\bar{V} \sim K(\cdot \mid V, M, \Omega)$*. Then* $I(Y; M \mid \Omega)$ *is unchanged by* K*, and*

$$I(Y; M, \bar{V} \mid \Omega) \ge I(Y; M \mid \Omega). \tag{13}$$

For a transcript $T_{1:R}$*,* $I(Y; T_{1:R} \mid \Omega)$ *is nondecreasing in* R *and equals* $\sum_{r=1}^{R} I(Y; T_r \mid T_{<r}, \Omega)$*.*

This proposition does not supply a composed privacy guarantee; it formalizes why post-selection value noise cannot repair a support leak and why repeated-message evaluation is necessary.

Proposition 3.6 (Linkability with correlated false matches)**.** *Let $L \in \{0,1\}$ indicate whether an earlier row is correctly linked, and let the observed row be $O = M$ when $L = 1$ and $O = U$ otherwise. Without assuming that false matches are label-independent,*

$$I(Y;O \mid \Omega) \le I(Y;L \mid \Omega) + \Pr(L=1)I(Y;M \mid L=1,\Omega) \tag{14}$$

$$+ \Pr(L=0)I(Y;U \mid L=0,\Omega). \tag{15}$$

If $L \perp\!\!\!\perp Y \mid \Omega$, the first term vanishes. If additionally $U \perp\!\!\!\perp Y \mid \Omega$, this reduces to the idealized gate $I(Y;O \mid \Omega) \le pI(Y;M \mid \Omega)$. Correlated streams, client skew, or class-structured sessions can make the U term nonzero; the theorem therefore does not equate an incorrect link with privacy. See Appendix E.2.

Proposition 3.7 (Row-wise batch invariance)**.** *For the row-wise protocol studied here, let $G_B \in \mathbb{R}^{B\times d}$ and $\mathsf{Comp}_B(G_B) = (\mathsf{Comp}(G_1),\dots,\mathsf{Comp}(G_B))$. The support released for row i is identical to the support obtained by processing example i alone, up to the common positive scaling induced by mean-loss reduction. Flattened-tensor, blockwise, microbatch-aggregated, or example-averaged compression produces a different wire object and is outside this proposition.*

Proposition 3.8 (Binary linear-tail support null)**.** *In exact arithmetic, for a binary linear softmax tail with class-weight columns w_0, w_1, every finite-logit per-example cross-entropy gradient has the form*

$$G = a(X,Y)(w_1 - w_0), \qquad a(X,Y) \neq 0. \tag{16}$$

Hence any deterministic selector that depends only on the rank ordering of $|G_j|$ and uses fixed tie-breaking releases a constant support, so $I(Y;M \mid \Omega) = 0$ when the tail is fixed in Ω. The retained values may still reveal Y through the signed scalar $a(X,Y)$.

The nonzero condition is load-bearing: finite-precision softmax saturation or cancellation can create zero/residual rows whose implementation-level ranks are not covered. This proposition explains the late-residual LFW null; a separate Cancer diagnostic exposes exactly this finite-precision boundary (Appendices E.5 and E.6).

4 Experimental Design

Controlled mechanism suite. Breast Cancer, Iris, Wine, Digits, and Synthetic-10 use disjoint 60/20/20 model-training, auxiliary, and test partitions and seeds $\{1, 7, 19, 31, 43\}$. These inexpensive datasets isolate mechanisms and uncertainty; they are not production-scale benchmarks. The primary split MLPs use a *linear* private softmax tail. For multiclass Digits and Synthetic-10 this permits label-dependent supports; for binary Cancer, exact arithmetic predicts a constant mask. Its apparent float32 leakage is therefore removed from the headline table and audited separately as saturation/cancellation, not counted as evidence against Proposition 3.8. Nonlinear tails, a three-cut convolutional Digits model, residual/TinyViT controls, and LFW test architectural boundaries.

Non-degenerate utility stress test. To measure a setting where compression changes task performance, we add the public Diamonds dataset (Wickham, 2016): 53,940 rows, eight clarity classes, and mixed numerical/categorical covariates. A split MLP with cut dimension 128 is trained for eight epochs with compression in every backward pass. At $k/d \in \{1/8, 1/4\}$ we compare seven stateless selectors over five seeds; at $1/4$ we also sweep a sequential exponential-mechanism selector over $\varepsilon_M \in \{1, 2, 4, 8, 16, 32\}$. Its one-message guarantee covers the exact-k support only; aligned values and repeated training remain unprotected.

Limited modern-image check. A compact residual CNN is trained on CIFAR-10 with augmentation and a nonlinear private tail behind a 128-dimensional pooled cut. Dense, top-k, and random-k use 45,000 training images and equal sparse wire budgets in the six-epoch comparison. Separate three-seed top-k runs use terminal 24- and 48-epoch cosine schedules; the latter audits epochs $24, 36, 48$ rather than extrapolating or restarting the former (Appendix C.4).

Selectors and attacks. We evaluate deterministic top-k, explicit mixtures using either uniform random-k or an auxiliary-estimated fixed support as the label-independent branch, randomized top-k, magnitude-proportional sampling, adaptive dropout, Gumbel top-k, fixed public support, independent random-k, one-message ε_M-DP support selection with unprotected values, error-feedback

Table 1: Five-seed top-k audits at $k/d = 1/4$. Support is ordinary accuracy (Diamonds balanced support is .861). DPI LCB is the decoder-free 95% bound from one auxiliary-frozen coordinate; Sign is a conventional calibration-matched sign-vector attack.

Dataset	Task	Support	Joint	DPI LCB (bits)	Sign
Digits	.966±.005	.852±.038	.879±.033	.140±.046	.997±.000
Synthetic-10	.875±.008	.748±.027	.843±.022	.094±.027	.984±.003
Diamonds	.528±.005	.796±.022	.851±.030	.334±.029	1.000±.000

top-k, and TPSL's binary discrete GradPerturb primitive before top-k (Yang et al., 2022). Parameter grids range from data-independent to deterministic. Post-training audits isolate the wire channel; training-time studies match sparsity, bit, epoch, and optimizer budgets. Learned support and joint decoders use only the auxiliary partition; hyperparameters are selected on an auxiliary-only validation split and frozen before test access. We compare logistic, prototype, random-forest, and shallow-MLP support decoders; the joint decoder receives $[M, \widetilde{G}]$. The decoder-free coordinate certificate freezes $j^\star$ on auxiliary masks. Conventional baselines use dense gradients, norms, and sign vectors under the same calibration split. Metrics include task utility, support and joint attacks, conditional value gain, information lower bounds, omitted energy, and bits. Main deployment-facing tables use literal 32-bit values plus explicit coordinate indices:

$$B_{\mathrm{literal}}(d,k) = 32k + k\lceil \log_2 d \rceil, \qquad B_{\mathrm{enum}}(d,k) = 32k + \left\lceil \log_2 \binom{d}{k} \right\rceil . \tag{17}$$

The second quantity is an information-theoretic lossless lower encoding, reported separately. All attacks assume the declared row-wise $B \times d$ protocol; no prevalence claim is made for systems that flatten, aggregate, or block-compress the batch.

5 Results

5.1 Unified five-seed audit and conditional certificates

The larger Diamonds cross-fit audit uses 5,394 disjoint pairs per seed and has gap $-.0168$ (negative in all five seeds) despite .795 ordinary mask accuracy under the natural class proportions. Therefore transferable energy fidelity is not the mechanism used by the attack. Per-coordinate class frequencies diagnose the alternative: a single auxiliary-selected mask bit has test plug-in information .528, .271, and .482 bits on Digits, Synthetic-10, and Diamonds, and Equation (11) remains positive after simultaneous exact intervals in every seed. The cross-fit route is retained as a candidate pathway falsified as an explanation of these audits; the coordinate-frequency and Fano bounds are the quantitative certificates for the observed audits (Appendices B.5 and B.8).

5.2 A real-large task makes utility non-degenerate

Dense task accuracy is $.546 \pm .006$. At $k/d = 1/4$, seven stateless selectors span .527–.541 task accuracy and .121–.861 *balanced* support accuracy (chance .125); the raw top-k score in Table 1 uses the natural prior .242. Random-k dominates top-k on both axes in every Diamonds seed, but this is dataset-specific: on CIFAR-10, top-k retains 2.1 task points over equal-budget random-k. The support-DP sweep is operationally indistinguishable from or worse than free random-k: random-k gives task/support/joint .540/.121/.459, while $\varepsilon_M = 1$–32 gives .541–.543/.126–.171/.457–.505. Its benefit is the formal one-message support guarantee, not an empirical frontier improvement; values remain open. Literal communication is 1,248 bits throughout (Appendix D.3).

5.3 A representative existing defense under the sparse-message audit

TPSL's discrete GradPerturb primitive randomly replaces the true-label gradient by its counterfactual-label gradient before transmission (Yang et al., 2022). We apply the same draw to the private-tail update, then top-k, on a five-seed 25,323-row Diamonds SI1-versus-VS2 task. A symmetric flip model predicts support accuracy $a_{\mathrm{flip}} = (1-p)a_0 + p(1-a_0)$ from the same-seed

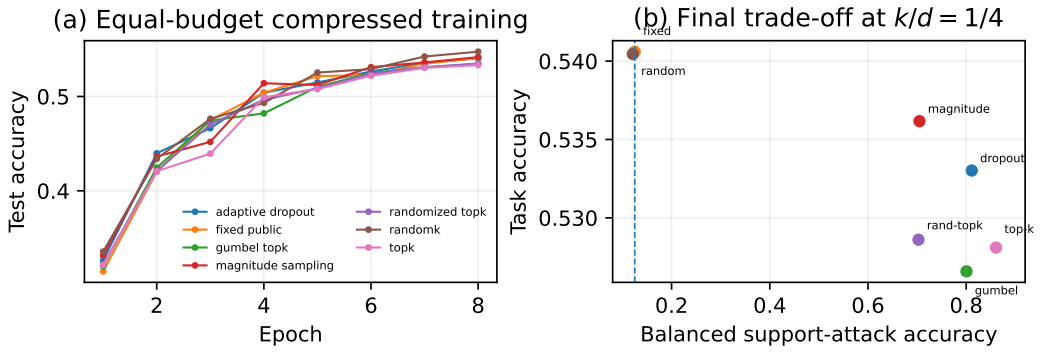


Figure 2: Diamonds compressed training. (a) Representative equal-budget learning curves at $k/d = 1/4$; all selectors are active in every backward pass. (b) Five-seed final means: task accuracy varies while balanced support inference spans chance to .861. The vertical line marks eight-class chance .125.

unprotected score a_0. At $\varepsilon = 0, 1, 2, 4$, measured support exceeds this arithmetic prediction by $.071, .079, .054, .013$, respectively. Thus most of the high-ε behavior is the definition of GradPerturb, while the positive residual is the representation-dependent result. At $\varepsilon = 2$, task/joint are $.752/.887$ versus $.765/1.000$ unprotected. Eight epochs require composition; we claim neither the original one-pass TPSL transcript theorem nor end-to-end privacy (Appendix D.4).

5.4 Graded selectors, calibration, and transcripts

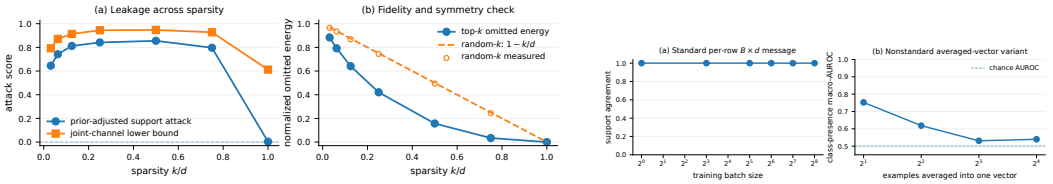
The explicit random-k/top-k family instantiates Proposition 3.2 with an arbitrary label-independent branch Q_b: its distortion is affine between D_b and D_1, and its correct lower bound is $2[D_0 - D_\lambda]_+^2 / \ln 2$, not the sharper λ^2 special case. A fixed auxiliary support gives mean distortion .638, near the *evaluation-target-optimized oracle descriptive* D_0 plug-in .633, versus .752 for random-k. Mean-grid nondominance is uncertainty-sensitive near the random endpoint: paired-bootstrap nondominance is .703 at $\lambda = .1$, .965 at .25, and 1.0 for $\lambda \geq .5$. We therefore call only the latter region a robust competitive reference frontier (Appendices D.1 and D.5).

Calibration remains the main exploitability boundary. On Digits, same-checkpoint accuracy is .852, a 32-epoch same-lineage transfer is .825, and independent-initialization transfer is .106. Synthetic-10 independently gives .749 matched, .711 across a 24-epoch gap, and .105 across initializations. Two labeled Digits messages per class give $.653 \pm .023$, while missing classes or a replaced private tail reduce transfer. With disjoint shared-checkpoint clients, retaining $25, 50, 75, 100\%$ of calibration clients gives $.453, .653, .740, .778$ accuracy as class coverage rises from .755 to one. Table 2 maps these tests to authorized replay, centrally versioned cross-client calibration, and retrospective outcome joins, including label ownership, tail synchronization, checkpoint coordination, row-wise wire requirements, and the measured decoder horizon. Personalized independent tails or asynchronously diverged lineages remain outside the demonstrated threat.

Perfectly linked eight-mask transcripts increase the available attack from .852 to .930. Under a non-oracle session model, false rows are drawn from the same public client/session after Dirichlet class skew, without using the target label. Across 5–20 sessions and $\alpha \in \{.05, .1, .3, 1\}$, false-link-only transcript decoders remain below the single-message envelope; gains appear only at high correct-link probability. An oracle same-label replacement can exceed this envelope, consistent with Proposition 3.6; it is retained only as a dependence stress test.

5.5 Architecture-dependent support null and value-channel persistence

The late residual LFW cut releases exactly one deterministic top-k mask at epochs $1, 3, 6$, and 18, with support accuracy .5 even when test loss is .642 and gradient norm is nonzero. This is not a vanishing-gradient or perfect-fit artifact: Proposition 3.8 predicts it exactly for a binary linear tail. Early/middle residual and all TinyViT cuts have nonlinear effective tails and retain strong support



(a) Full support-size sweep.

(b) Row-wise versus averaged batching.

Figure 3: Endpoint and wire-semantics tests. Random-k follows the exact $1 - k/d$ omitted-energy identity, top-k support inference returns to prior at $k = d$, and row-wise batching leaves each per-example mask unchanged. The averaged-vector curve is a different protocol.

inference. At the null cut, the joint decoder still reaches perfect accuracy, confirming that closing support alone does not protect values.

Post-selection value noise provides the complementary control: it drives conditional value gain to zero but leaves support accuracy unchanged at .852. Together, the two experiments show that support and value channels can be closed independently, and neither result constitutes a complete defense.

6 Protocol Semantics and Endpoint Validation

Two protocol checks are necessary before interpreting an attack. First, the support channel must vanish at $k = d$: every example releases the same mask, even though the dense values may still leak. Second, the row-wise protocol studied here must preserve separately recoverable example rows. The invariance result is conditional on that wire format: flattened-tensor sparsification, accumulation before compression, channel/block compression, and microbatch aggregation can alter both the support object and row linkability and require separate audits. These checks are visualized in Figure 3.

At six epochs, CIFAR-10 dense/top-k/random-k task accuracy is .751/.724/.703 and top-k/random-k support accuracy is .968/.099. Top-k calibration transfers from the half-training checkpoint at .834 but falls to .143 across independent initializations. On the independent 48-epoch schedule, task accuracy rises from .832 $\pm$ .024 at epoch 24 to .902 $\pm$ .002 at the terminal learning rate; support declines from .956 $\pm$ .005 to .919 $\pm$ .002 but remains over nine times the .1 prior, joint accuracy is .998 $\pm$ .000, and the decoder-free certificate is .404 $\pm$ .034 bits (Appendix C.4). This measures continued decay rather than extrapolating persistence, but does not establish prevalence in deeper or production systems.

7 Calibration and Selector Diagnostics

The calibration curve saturates near 20 labeled messages per class, and the same-lineage curve remains above .82 across a 32-epoch gap. By contrast, independent initialization gives .106, close to the ten-class prior. This separation is the main operational qualification of the attack: it is strong for calibration-capable servers, not architecture-only observers. Decoder ablations preserve the ordering: on Digits, top-k support accuracy is .787–.859 across prototype, logistic, random forest, and shallow MLP, whereas random-k is .096–.098; Synthetic-10 shows the same separation. Auxiliary-only model selection, NLL/ECE diagnostics, per-class recall, and confusion summaries are reported in Appendix C.1. The session-level false-link audit further shows that public client/session correlation alone does not improve the available attack at low linkage in the tested non-oracle models.

Across four selector families, lower omitted energy tracks stronger support inference. The explicit mixture verifies affine distortion; random-k uses the general D_b bound, while an auxiliary-fixed branch approximates the D_0 special case. Paired nondominance is robust for $\lambda \geq .5$, not at .1 (Appendix D.1 and Figure 6).

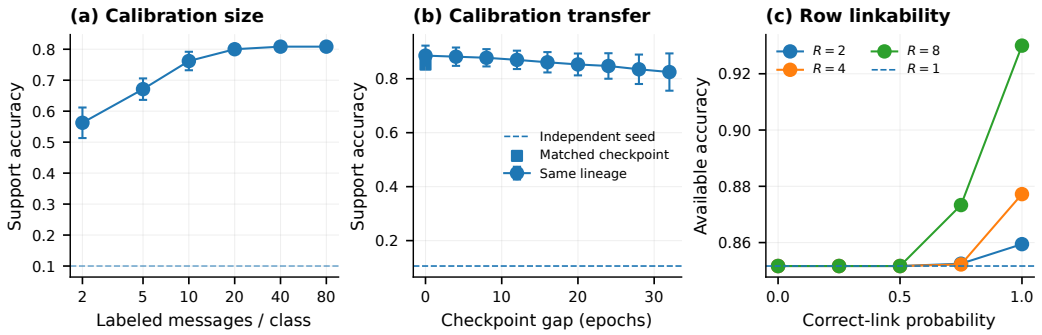


Figure 4: Calibration and linkage boundaries on Digits (five seeds). (a) Support inference is sample-efficient under target-checkpoint calibration. (b) It transfers within one training lineage but collapses across independent initializations. (c) Transcript gains require sufficiently reliable row linkage; the plotted attack can always ignore earlier masks.

8 Channel Separation and Design Consequences

Support is not the dominant label channel in every setting. CIFAR-10 dense training has a constant mask (support .100) but joint accuracy .996; fixed public support on Diamonds has balanced support .125 but joint .606. Thus closing support is necessary for a complete defense but is demonstrably insufficient, and the support contribution is best understood as an exposed object that existing value-only reasoning misses. The value-noise and binary-tail experiments form complementary interventions. Post-selection noise modifies V while holding M fixed; the late binary linear tail makes M constant while retaining informative values. Their crossed outcomes empirically validate the decomposition in Equation (2): neither channel is a proxy for the other, and a successful defense must address both.

The constructive output is an audit contract rather than a private trainer: selection needs its own guarantee, value privacy must condition on support, joint-message privacy must cover (M, V), and training privacy must compose over linkable rounds. For the Diamonds sweep, basic per-record composition over eight observed training releases is $8\varepsilon_M$; the experiment is therefore a one-message support control, not private training. No trained configuration at useful task utility brings the joint decoder near its class prior: on Diamonds, fixed public support *without* added value noise already leaves joint accuracy .606 against balanced support .125 (Table 20). The only joint-private construction we evaluate is the one-message public-support Gaussian diagnostic, which we run on Digits alone and at prohibitive distortion (Appendix D.6); we do not report it on Diamonds. Population bounds, fixed-instance certificates, retraining stability, utility, and operational audits are reported as distinct evidence layers.

9 Scope and Conclusion

Adaptive support alone is a label channel under lineage-shared calibration; data-independent supports close it at 0–2 measured task points, while values remain open regardless. The negative Diamonds cross-fit result shows that the population theorem is a distinct sufficient mechanism, not an explanation of the attack. The auxiliary-frozen coordinate certificate instead identifies and certifies class-conditional selection frequency. Independent lineages, full support, and exact binary linear-tail geometry expose precise nulls. No useful-utility trainer tested here protects the joint message. Evidence includes non-degenerate utility and a 48-epoch compact CIFAR-10 ResNet, not prevalence in deeper vision, language, or production systems. The row-wise protocol is supported by a small open-source code survey but remains a scope condition rather than a prevalence estimate (Appendix A.4).

REFERENCES

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 394–403, 2018.

Jennifer Gillenwater, Matthew Joseph, Andres Munoz, and Monica Ribero Diaz. A joint exponential mechanism for differentially private top-k. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7570–7582, 2022.

Yukun Jiang, Peiran Wang, Chengguo Lin, Ziyue Huang, and Yong Cheng. Training on fake labels: Mitigating label leakage in split learning via secure dimension transformation. *arXiv preprint arXiv:2410.09125*, 2024.

Kousha Kalantari, Oliver Kosut, and Lalitha Sankar. Information-theoretic privacy with general distortion constraints. *arXiv preprint arXiv:1708.05468*, 2017.

Sanjay Kariyappa and Moinuddin K. Qureshi. ExPLoit: Extracting private labels in split learning. In *IEEE Conference on Secure and Trustworthy Machine Learning*, 2023. URL `https://openreview.net/forum?id=0nrrpSTPKq3`.

Fumiyuki Kato, Yang Cao, and Masatoshi Yoshikawa. OLIVE: Oblivious federated learning on trusted execution environment against the risk of sparsification. *Proceedings of the VLDB Endowment*, 16(10):2404–2417, 2023.

Oscar Li, Jiankai Sun, Xin Yang, Weihao Gao, Hongyi Zhang, Junyuan Xie, Virginia Smith, and Chong Wang. Label leakage and protection in two-party split learning. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=cOtBRgsf2fO`.

Jiachun Liao, Oliver Kosut, Lalitha Sankar, and Flavio P. Calmon. Privacy under hard distortion constraints. *arXiv preprint arXiv:1806.00063*, 2018.

Yongjeong Oh, Jaeho Lee, Christopher G. Brinton, and Yo-Seb Jeon. Communication-efficient split learning via adaptive feature-wise compression. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):10844–10858, 2025.

Gang Qiao, Weijie Su, and Li Zhang. Oneshot differentially private top-k selection. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8672–8681, 2021.

Xinchi Qiu, Ilias Leontiadis, Luca Melis, Alexandre Sablayrolles, and Pierre Stock. Evaluating privacy leakage in split learning. *arXiv preprint arXiv:2305.12997*, 2023.

Hasin Us Sami and Başak Güler. Sparsity-based secure gradient aggregation for resource-constrained federated learning. In *2024 IEEE International Symposium on Information Theory*, pp. 2718–2723, 2024. doi: 10.1109/ISIT57864.2024.10619102.

Ricardo Silva Carvalho, Ke Wang, Lovedeep Gondara, and Chunyan Miao. Differentially private top-k selection via stability on unknown domain. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1109–1118, 2020.

Sajani Vithana and Sennur Ulukus. Rate–privacy–storage tradeoff in federated learning with top-r sparsification. In *IEEE International Conference on Communications*, pp. 3127–3132, 2023.

Sajani Vithana and Sennur Ulukus. Private read–update–write with controllable information leakage for storage-efficient federated learning with top-r sparsification. *IEEE Transactions on Information Theory*, 70(5):3669–3692, 2024.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

Xiaoyang Xu, Wenzhe Yi, Juan Wang, Hongxin Hu, Mengda Yang, Ziang Li, Yong Zhuang, Yaxin Liu, and Mang Ye. SLeak: Multi-target privacy stealing attack against split learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(5):5879–5891, 2026. doi: 10.1109/TPAMI.2026.3654092.

Xin Yang, Jiankai Sun, Yuanshun Yao, Junyuan Xie, and Chong Wang. Differentially private label protection in split learning. *arXiv preprint arXiv:2203.02073*, 2022.

Dixi Yao and Baochun Li. Is split learning privacy-preserving for fine-tuning large language models? *IEEE Transactions on Big Data*, 2024. doi: 10.1109/TBDATA.2024.3524101.

Fei Zheng, Chaochao Chen, Lingjuan Lyu, Xinyi Fu, Xing Fu, Weiqiang Wang, Xiaolin Zheng, and Jianwei Yin. Protecting split learning by potential energy loss. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5590–5598, 2024. doi: 10.24963/ijcai.2024/618.

Fei Zheng et al. Reducing communication for split learning by randomized top-k sparsification. *arXiv preprint arXiv:2305.18469*, 2023.

Yuqing Zhu and Yu-Xiang Wang. Adaptive private-k-selection with adaptive k and application to multi-label PATE. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 5622–5635, 2022.

A Protocol, Scope, and Deployment

A.1 Notation and Complete Protocol

Symbol	Meaning
X, Y	private input and class label
Z, H	client-encoder activation and server-body output
$G \in \mathbb{R}^d$	one-example cut gradient $\partial\mathcal{L}/\partial H$
$G_B \in \mathbb{R}^{B\times d}$	row-wise audit tensor, one separately recoverable row per example
Comp, U	sparse compressor and its randomness
M, V	support and coordinate-aligned retained values
Ω	full server-known context: architecture/cut, known checkpoint and optimizer state, round, client identity, previous transcript, mechanism parameters, and public randomness
ℓ_{om}	normalized omitted energy; $\mathcal{L}$ denotes training loss
C, d, k	class count, cut dimension, and retained coordinates

One U-shaped split-learning step is:

1. the client sends $Z = f_{\theta_c}(X)$;
2. the server returns $H = g_{\theta_s}(Z)$;
3. the private tail computes predictions, loss, and $G = \partial\mathcal{L}/\partial H$;
4. the client sends $T = (M, V)$ corresponding to $\widetilde{G} = \mathsf{Comp}(G, U)$;
5. the server reconstructs the sparse tensor and updates its parameters; the client updates its private parameters.

For a batch, the leading dimension is B. The evaluated protocol sparsifies each row and transmits B supports and value vectors. This is a declared scope condition, not a prevalence claim: a system that flattens, aggregates, or block-compresses across rows exposes a different object. In each reported conditional certificate, Ω is fixed to one trained checkpoint and audit configuration. Cross-checkpoint and cross-seed experiments deliberately change components of Ω.

A.2 ASSUMPTION LEDGER

Result	Assumptions	Not assumed
Cross-fitted bound	G, G' conditionally i.i.d. given (Y, Ω); $M \sim Q(\cdot \mid G, \Omega)$; bounded distortion	Gaussianity, linear tails, deterministic selection, balanced classes
Mixture-selector bounds	label-independent comparator branch; independent Bernoulli branch choice; fixed context	a specific decoder, a linear empirical curve, or privacy of the adaptive branch
Conditional empirical Fano	model, checkpoint, auxiliary data and decoder fixed; independent test; declared audit weights	coverage over training initialization or architecture selection
Coordinate-frequency DPI	coordinate selected only on auxiliary masks; independent stratified test; simultaneous binomial intervals; fixed audit weights	fitted decoder, cross-fit fidelity, or dimension-dependent search on test
Hard support list	released support belongs to or is covered by the true-class list, with miss probability δ	optimal decoder or model linearity
Post-selection invariance	value channel is changed without changing the already selected M	independence of V and M
Transcript monotonicity	ordinary mutual-information chain rule	independent rounds or a DP composition theorem
Row-wise batch result	cut message retains per-example rows; compressor acts row-wise	example-averaged communication
Random-k identity	uniform exact-k support independent of nonzero g	exchangeable gradient coordinates
One-message support DP	normalized scores in $[0, 1]$; k composed exponential-mechanism draws; one support	privacy of retained values or transcript composition
One-message Gaussian baseline (Digits only)	fixed public support, declared clipping and sensitivity	repeated-round composition, private training, or transfer to other datasets
TPSL primitive comparison	binary labels; one discrete GradPerturb draw applied to cut and tail gradients before top-k	the original online-use transcript theorem or eight-epoch composition
Binary linear-tail null	exact arithmetic; finite logits; nonzero realized gradient; rank-only selector; fixed tie-breaking	robustness to softmax saturation, cancellation, or a numeric zero-row policy

A.3 CALIBRATION CAPABILITY MATRIX

Table 2: Concrete pathways from labeled records to compatible calibration messages. Every viable row requires the declared per-example row-wise sparse backward wire. "Horizon" reports the tested model-version evidence, not wall-clock validity or prevalence. H/V/U denote horizontal, vertical, and U-shaped split learning.

Pathway	Who has labels	Compatible gradient/tail	Version coordination	Wire and server view	Measured horizon/boundary
Authorized target replay	target client or authorized curator	exact live private tail and preprocessing execute consented records; weights need not leave client	checkpoint is pinned and version-tagged during replay; rerun after update	server receives labeled $B \times d$ sparse rows from the ordinary backward interface	exact-checkpoint evidence; useful until update or detected drift; requires target/label-party cooperation (direct in U, V through label holder)
Centrally versioned cross-client audit	consenting clients or sites own labeled records	centrally distributed tail/checkpoint family is shared or synchronized; calibration/target clients are disjoint	coordinator enforces version IDs; short asynchrony stays within one lineage	server joins labels from consenting clients to per-example rows, then decodes other clients	.825 after 32 same-lineage epochs on Digits; class coverage/turnover matter; personalized independent tails are unsupported
Retrospective outcome join	application or outcome-log owner (e.g., delayed click or clinical outcome)	the recorded live client already generated the message with its own tail; no copied weights needed	message IDs retain the exact checkpoint version used when the outcome later arrives	authorized join maps delayed labels to stored per-example sparse rows	same-checkpoint by construction for calibration rows; decoder must be revalidated across model turnover; requires legal/operational label access and linkage
Authorized audit sandbox	public/consented-data curator	sandbox runs the versioned compatible client tail and preprocessing and returns labeled messages	audit API pins a release version; target traffic must carry the same lineage tag	server sees labeled sparse rows, not tail weights; privileged interface, not ordinary unlabeled traffic	exact/stale-checkpoint evidence only; feasible in H/U and in V only through a label-side service
Personalized / async boundary	no compatible labeled-message provider	client-specific independent tails or full-model initializations	versions diverge without a shared lineage contract	ordinary unlabeled sparse rows only	replaced-tail transfer .321 and independent-lineage transfer .106 on Digits; the supervised attack is not demonstrated

The experiments instantiate the second row with disjoint clients and stress class coverage, turnover, and preprocessing. Exact- and stale-checkpoint transfer instantiate the first four rows; replacing the private tail and changing the full-model seed test the final boundary. TPSL's two-party formulation independently documents the per-example backward-gradient interface and label-holding party in vertical split learning (Yang et al., 2022); this paper does not estimate how frequently any pathway

occurs in production. None turns calibration into an automatic capability of an unlabeled honest-but-curious server.

A.4 Open-Source Wire-Format Spot Check

To test whether row preservation is merely hypothetical, we inspected three immutable public code snapshots. This convenience sample is not a prevalence estimate and none implements our exact sparse compressor. The survey establishes that row-preserving tensor transport exists in public implementations, while also narrowing our claim: we found no surveyed system that applies exact top-k independently to those rows. Deployment of the specific sparse wire format remains an assumption to verify system by system.

Table 3: Code-level wire semantics. "Rows" means the leading minibatch dimension survives communication; it does not imply stable cross-round identity.

Snapshot	Observed exchange	Boundary
splitlearning	Socket sends 64-row cut activations and returns the corresponding batched tensor gradient.	Rows preserved; dense, not top-k.
SplitLearning	RPC sends batched activations; distributed autograd propagates the batched gradient across the RPC graph.	Rows preserved by the framework; no explicit sparse compressor.
Raspberry-Pi SL	Socket sends four-row activations and returns `client_grad` with the leading batch dimension.	Rows preserved, but labels are also sent, so it is outside the label-hidden threat.

A.5 Conditional Uncertainty, Calibration, and Evaluation Hierarchy

Let Θ denote the random trained instance induced by initialization, model-training data order, and checkpoint selection. For a fixed realization $\Theta = \theta$ and fixed auxiliary set, the Fano interval controls only the independent test-sampling error in

$$I_\theta = I(Y; M \mid \Omega = \pi_\theta). \tag{18}$$

It does not cover the distribution of I_Θ under retraining. We therefore do not pool test examples across seeds to obtain an artificially narrow interval. The artifact reports (i) one conditional certificate per seed and (ii) the mean and standard deviation of attack metrics across five independently trained instances. With five seeds, the latter is a stability summary rather than a high-confidence estimate of a deployment population.

The transfer experiments distinguish three contexts. Same-checkpoint evaluation fixes π_θ; nearby-checkpoint transfer changes the round and parameters while retaining one training lineage; cross-seed transfer changes the lineage. The observed support accuracies $.852$, $.825$ at a 32-epoch gap, and $.106$ across independent seeds show that the strongest learned channel is calibration-dependent. The practical adversary is therefore a calibration-capable server possessing target-checkpoint or same-lineage labeled messages, not a server with architecture knowledge alone.

A privacy claim about a selector should pass four increasingly strong tests. First, endpoint identities and batching semantics must match the wire protocol. Second, support, aligned values, and their joint observation must be attacked under the same public context. Third, task utility, transmitted bits, and leakage must be measured while the mechanism participates in training. Fourth, any formal upper guarantee must cover support selection, values, and repeated composition. The experiments address the first three levels; the public-support Gaussian construction, evaluated on Digits only, addresses the fourth for one message. No result is presented as composed private training.

B Theory, Certificates, and Proofs

B.1 Proof of the Cross-Fitted Bound

Fix $\Omega = \pi$ and suppress π. Let $P_y = P_{M|Y=y}$ and $P_M = \sum_y p_y P_y$. Define $a_y(S) = \mathbb{E}[\ell_{\mathrm{om}}(G', S) \mid Y = y] \in [0, 1]$. The marginal law P_M is label-independent and is therefore

Table 4: Digits boundary measurements for the auxiliary-data and transcript threat model (five seeds).

Condition	Support acc.	Value gain	Interpretation
2 auxiliary messages/class	.653	–	detectable with very limited labels
20 auxiliary messages/class	.834	–	near saturation
Matched checkpoint	.852	–	primary calibration threat
Same run, 32-epoch gap	.825	–	transfers along training lineage
Independent model seed	.106	–	does not transfer across lineage
One observed mask	.852	–	single-message exposure
Eight correctly linked masks	.930	–	cumulative exploitability
Eight masks, at most 50% linked	.852	–	no gain beyond the single-mask decoder
Value noise ≥ 1 retained SD	.852	.000	values suppressed; mask unchanged

feasible in the optimization defining D_0. Hence

$$D_0 - D_{\mathrm{cf}} \leq \sum_y p_y \left(\mathbb{E}_{S \sim P_M} a_y(S) - \mathbb{E}_{S \sim P_y} a_y(S) \right) \tag{19}$$

$$\leq \sum_y p_y \, \mathrm{TV}(P_y, P_M). \tag{20}$$

The second inequality is the variational characterization of total variation for functions in $[0, 1]$. Pinsker in bits gives $\mathrm{TV}(P_y, P_M)^2 \leq (\ln 2/2) D_{\mathrm{KL}}^{(2)}(P_y \| P_M)$. Jensen then yields

$$(D_0 - D_{\mathrm{cf}})_+^2 \leq \left(\sum_y p_y \, \mathrm{TV}(P_y, P_M) \right)^2 \tag{21}$$

$$\leq \sum_y p_y \, \mathrm{TV}(P_y, P_M)^2 \tag{22}$$

$$\leq \frac{\ln 2}{2} \sum_y p_y D_{\mathrm{KL}}^{(2)}(P_y \| P_M) \tag{23}$$

$$= \frac{\ln 2}{2} I(Y; M \mid \Omega = \pi). \tag{24}$$

Rearranging and averaging over Ω proves Theorem 3.1. The square remains inside the expectation because the conditional statement is applied separately at each public context value.

Why the independent same-label target matters. Replacing G' with G permits a mechanism to select coordinates that fit sample-specific noise. Such self-fit can reduce reconstruction error while producing no class-regular support law. Evaluating the support selected from G on $G' \sim P_{G|Y,\Omega}$ retains only distortion improvement that transfers within class. This comparator construction, rather than Pinsker itself, is the substantive modeling contribution.

B.2 Proof of the Mixture-Selector Bounds

Fix $\Omega = \pi$. Let P_b be the label-independent channel induced by Q_b, and let $P_1(\cdot \mid y)$ be the channel induced by Q_1. The mixture channel is

$$P_\lambda(\cdot \mid y) = (1 - \lambda) P_b + \lambda P_1(\cdot \mid y). \tag{25}$$

Linearity of expectation gives $D_\lambda = (1 - \lambda) D_b + \lambda D_1$. Applying Theorem 3.1 directly yields $2[D_0 - D_\lambda]_+^2 / \ln 2 \leq I(Y; M_\lambda \mid \pi)$. If $D_b = D_0$, then $D_0 - D_\lambda = \lambda(D_0 - D_1)$, which gives the stated λ^2 special case. For the upper bound, the marginal is $P_\lambda = (1 - \lambda) P_b + \lambda P_1$. Joint convexity

of KL gives

$$I(Y; M_\lambda \mid \pi) = \sum_y p_y D_{\mathrm{KL}}^{(2)}(P_\lambda(\cdot \mid y) \| P_\lambda) \tag{26}$$

$$\leq \lambda \sum_y p_y D_{\mathrm{KL}}^{(2)}(P_1(\cdot \mid y) \| P_1) = \lambda I(Y; M_1 \mid \pi). \tag{27}$$

If B is observed, $B \perp\!\!\!\perp Y \mid \pi$ and the $B = 0$ branch is label-independent, hence $I(Y; M_\lambda, B \mid \pi) = \lambda I(Y; M_1 \mid \pi)$.

B.3 Dimension-Free Finite-Sample Cross-Fit Bound

Let q_i be the normalized coordinate-energy vector of target gradient i; q_i lies in the probability simplex. Define

$$\widehat{F} = \max_{|S| \leq k} \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in S} q_{ij}, \qquad \widehat{D}_0 = 1 - \widehat{F}. \tag{28}$$

Replacing one sample changes every fixed-support average by at most $1/n$, and therefore changes their maximum $\widehat{F}$ by at most $1/n$. Moreover, if $F = \max_{|S| \leq k} \mathbb{E} \sum_{j \in S} q_j$, then $F \leq \mathbb{E}\widehat{F}$. McDiarmid's lower-tail inequality yields, with probability at least $1 - \alpha/2$,

$$D_0 = 1 - F \geq \widehat{D}_0 - \sqrt{\frac{\log(2/\alpha)}{2n}}. \tag{29}$$

Hoeffding gives the simultaneous upper bound

$$D_{\mathrm{cf}} \leq \widehat{D}_{\mathrm{cf}} + \sqrt{\frac{\log(2/\alpha)}{2n}}. \tag{30}$$

A union bound and Theorem 3.1 prove Proposition 3.3. This argument is dimension- and sparsity-independent, unlike the coordinatewise union bound. At $\alpha = .05$, the fixed-split five-seed quantity Δ_{fix} is $.033, .108, .056, .064, .101$ for Cancer, Digits, Iris, Synthetic-10, and Wine, respectively, but the two confidence penalties exceed every gap. A larger independent-pair Diamonds audit uses every auxiliary example as a source and a disjoint same-label test example as its target, yielding 5,394 pairs per seed. Its five plug-in gaps range from $-.0213$ to $-.0047$ (mean $-.0168$), so its certificate is also zero for a substantive rather than merely sample-size reason. We therefore do not use this theorem as quantitative attack evidence; the frequency-DPI and frozen-decoder Fano bounds are the operational fixed-audit certificates. Table 9 reconciles Δ_{fix} with the two other descriptive gaps reported below.

B.4 Exact Two-Coordinate Interpretation Example

Let Y be uniform on $\{0, 1\}$, $d = 2$, and $k = 1$. A same-label gradient places normalized energy $p > 1/2$ on coordinate Y, except that an independent within-class flip $E \sim \mathrm{Bernoulli}(\eta)$ reverses the two coordinates. Top-1 releases the high-energy coordinate. The cross-fit target has an independent flip E'. Direct calculation gives

$$I(Y; M) = 1 - h_2(\eta), \qquad D_0 - D_{\mathrm{cf}} = (p - \tfrac{1}{2})(1 - 2\eta)^2, \tag{31}$$

so the theorem lower bound is $2(p - \frac{1}{2})^2(1 - 2\eta)^4 / \ln 2$. This separates selector information from same-class transfer: the gap contains a squared heterogeneity attenuation, and the information bound contains its square again.

Thus the bound is informative but not tight at $(p, \eta) = (1, 0)$, and can be arbitrarily loose when the coordinate ranking reveals the class but the energy contrast $p - \frac{1}{2}$ is small. In 5,000 Monte Carlo audit repetitions at $p = .9, \eta = .2$, the valid finite-sample certificate is positive in .691 of runs at $n = 500$ and .999 at $n = 1{,}000$. At $\eta = .35$, the frequencies are .137 at $n = 5{,}000$ and .891 at $n = 10{,}000$. These simulations use fresh independent (G, G') pairs and include the empirical optimization defining $\widehat{D}_0$; they illustrate sample demand rather than modify the theorem.

Table 5: Exact construction. n_{pen} is the first integer for which the population gap exceeds the two $\alpha = .05$ bounded-difference penalties; it is a planning threshold, not a random-sample guarantee.

p	η	$I(Y;M)$	$D_0 - D_{\mathrm{cf}}$	Bound	n_{pen}
1.00	.00	1.000	.500	.721	30
.90	.00	1.000	.400	.462	47
.55	.00	1.000	.050	.007	2,952
.90	.10	.531	.256	.189	113
.90	.20	.278	.144	.060	356
.90	.35	.066	.036	.004	5,693
.90	.50	.000	.000	.000	∞

B.5 Decoder-Free Coordinate-Frequency Certificate

The auxiliary rule evaluates the d binary contingency tables and freezes

$$j^\star = \arg\max_j \widehat{I}_{\mathrm{aux}}(Y; M_j), \tag{32}$$

using the smallest coordinate index for a tie. The test set is untouched during this choice. Conditional on its realized stratified class counts, $S_y = \sum_{i:Y_i=y} M_{ij^\star}$ is binomial with parameter p_y. We form two-sided Clopper–Pearson intervals with tail probability $\alpha/(2C)$; a union bound covers all C intervals simultaneously.

For a fixed feasible vector p, data processing gives $I(Y; M \mid \Omega) \geq I(Y; M_{j^\star} \mid \Omega)$. Applying Pinsker in bits to each Bernoulli conditional and its mixture gives

$$I(Y; M_{j^\star} \mid \Omega) = \sum_y w_y D^{(2)}_{\mathrm{KL}}(\mathrm{Ber}(p_y) \| \mathrm{Ber}(\bar{p})) \geq \frac{2}{\ln 2} \sum_y w_y (p_y - \bar{p})^2, \tag{33}$$

where $\bar{p} = \sum_y w_y p_y$. Minimizing the convex weighted variance over the simultaneous interval box proves Proposition 3.4. The implementation reduces the optimization to one scalar center: $\min_q \sum_y w_y [\mathrm{clip}(q, [\ell_y, u_y]) - q]^2$.

Table 6: Single-coordinate frequency diagnosis (five seeds). "Plug-in" is descriptive test mutual information; the LCB is the decoder-free 95% certificate. Selection frequency reports the full class/seed range.

Dataset	Test n	Plug-in bits	95% LCB bits	Class-frequency range
Digits	360	$.528 \pm .077$	$.140 \pm .046$	$[.000, 1.000]$
Synthetic-10	1,000	$.271 \pm .048$	$.094 \pm .027$	$[.000, .879]$
Diamonds	10,788	$.482 \pm .048$	$.334 \pm .029$	$[.000, 1.000]$

Every per-seed LCB is positive. This directly reconciles the negative Diamonds cross-fit gap with strong decoding: a coordinate can be selected nearly always for one class and almost never for another without concentrating disproportionate energy in an independent same-label gradient. All 140 class/seed frequencies, counts, and simultaneous interval endpoints are released in `dpi_frequency_class_frequencies.csv`; the table compresses them to their range.

Conventional gradient baselines. To establish marginal rather than absolute significance, the same auxiliary/test messages are attacked using the cut-gradient norm, full sign vector, and full dense gradient. These are calibration-matched conventional gradient-feature baselines related to direction/sign leakage (Li et al., 2022; Kariyappa & Qureshi, 2023), not a claim to reproduce every optimization detail of ExPLoit. The pre-existing white-box linear-tail distance test also scores 1.000 on the two controlled multiclass tasks. The sign and joint results show that labels are not protected when the mask is closed. The support result instead identifies an additional exposed channel that remains observable even if values are later hidden.

Table 7: Conventional gradient attacks under the same five splits.

Dataset	Norm	Dense learned	Sign learned	White-box linear tail
Digits	.118	.354	.997	1.000
Synthetic-10	.101	.624	.984	1.000
Diamonds	.248	1.000	1.000	n/a

B.6 Empirical Fano Confidence Bound

Fix $\Omega = \pi$, the trained checkpoint, auxiliary set, decoder, and audit distribution. Let class weights be w_y and $H_w(Y) = -\sum_y w_y \log_2 w_y$. On an independent stratified test set, class y contributes n_y Bernoulli errors with empirical rate $\widehat{e}_y$. For weights fixed independently of the class-conditional errors, weighted Hoeffding gives

$$e_U = \min\left\{1, \widehat{e} + \sqrt{\frac{\log(1/\alpha)}{2}\sum_y \frac{w_y^2}{n_y}}\right\} \tag{34}$$

as a one-sided $(1-\alpha)$ upper bound on $e = \sum_y w_y e_y$. The conditional certificate is

$$I(Y; M \mid \Omega = \pi) \geq \max\left\{0, \min_{0 \leq u \leq e_U} [H_w(Y) - h_2(u) - u \log_2(C-1)]\right\}. \tag{35}$$

The inner minimization handles the non-monotone region. For every headline and tabled empirical-Fano result, we condition on the realized stratified test class counts and then set $w_y = n_y/n$. Hence the weights are fixed by the declared benchmark split for the conditional statement, $\sum_y w_y^2/n_y = 1/n$, and the implemented radius is exactly $\sqrt{\log(1/\alpha)/(2n)}$. The target is the resulting *empirical-mixture audit population*; it is not a confidence statement for unknown deployment prevalence, and the weights are not post-hoc tuned using decoder errors. If predetermined population weights are desired, they must be declared independently and the general weighted radius above applies. The separate support-list experiment uses a uniform class prior and is labeled accordingly.

Table 8: Empirical-mixture support Fano results for the two retained controlled tasks at $k/d = 1/4$. Each 95% lower bound conditions on realized held-out class counts, one fixed checkpoint, auxiliary set, decoder, and audit population; averaging over five seeds does not confer retraining coverage.

Dataset	Mean point estimate	Mean 95% lower bound
Digits	2.256	1.908
Synthetic-10	1.713	1.535

B.7 Hard Support Codes and Auxiliary-Only List Selection

For threshold τ and coverage tolerance η, define

$$\mathcal{A}_y(\tau, \eta) = \{S \in \mathcal{S}_k : \mathbb{P}(\ell_{\mathrm{om}}(G', S) \leq \tau \mid Y = y) \geq 1 - \eta\}, \tag{36}$$

and $m(\tau, \eta) = \max_S |\{y : S \in \mathcal{A}_y(\tau, \eta)\}|$.

Theorem B.1 (Support-list information bound)**.** *If $M \in \mathcal{A}_Y(\tau, \eta)$ almost surely, then*

$$I(Y; M) \geq H(Y) - \log_2 m(\tau, \eta). \tag{37}$$

If a frozen empirical codebook misses the true class with probability δ, then

$$I(Y; M) \geq H(Y) - h_2(\delta) - (1-\delta)\log_2 m - \delta \log_2 C. \tag{38}$$

The first statement follows because an observed support leaves at most m possible classes. The second conditions on the codebook-miss event. Candidate list sizes are selected on an auxiliary build/validation split and frozen before test evaluation. Under the declared uniform audit prior, the selected-list 95% bounds are .258 bits on Digits and .394 on Synthetic-10 and zero elsewhere. A zero is inconclusive, not evidence of privacy.

B.8 Exact Distortion, Random-k, and the D_0/D_{cf} Plug-In

The implemented metric is exactly

$$\widehat{\ell_{\mathrm{om}}}(g,S)=\begin{cases}1-\sum_{j\in S}g_j^2/\sum_{j=1}^d g_j^2, & \|g\|_2>0,\\ 0, & \|g\|_2=0,\end{cases} \tag{39}$$

with float64 accumulation and no additive denominator constant. For a uniform exact-k subset S independent of fixed $g\neq 0$,

$$\mathbb{E}_S\ell_{\mathrm{om}}(g,S)=1-\frac{k}{d}. \tag{40}$$

This identity and the constant-support endpoint at $k=d$ are executable unit tests.

For the descriptive cross-fit plug-in, each support selected from G_i is evaluated on a within-class derangement G_i'. The empirical label-independent optimum selects the k largest marginal normalized coordinate energies. Because this optimization reuses evaluation targets, it is an oracle descriptive plug-in; Appendix B.3 gives a valid conservative alternative.

Table 9: Reconciliation of every reported cross-fit gap. All three controlled-suite quantities use within-class derangements and optimize the fixed comparator on the evaluation targets, so they are descriptive unless the finite-sample penalty is applied. "Before/after" specifies the order relative to averaging over instances.

Symbol	Population/empirical	Sample and optimization	Pairing/aggregation	Instances and why values differ
Δ_{fix}	empirical plug-in; finite penalty applied separately	one fixed 60/20/20 split (split seed 2027); D_0 optimized on each test target set	within-class test derangement; compute gap per seed, then average; square per seed	five early-32 checkpoints per dataset; values .033, .108, .056, .064, .101 in Cancer, Digits, Iris, Synthetic-10, Wine order
Δ_{part}	empirical oracle plug-in in Table 10	partition changes with model seed; D_0 optimized on each test target set	within-class test derangement; average D_0, D_{cf} over seeds, subtract, then square	five early-32 checkpoints; changed partitions and aggregation give .017, .096, .067, .077, .095 in the same dataset order
Δ_{grid}	empirical model-instance diagnostic	original multi-architecture and cut audit; D_0 optimized on each instance's targets	within-class derangement; apply $[\cdot]_+$ per instance, then average and bootstrap instances	10 instances for Cancer, Iris, and Wine and 15 for Digits and Synthetic-10; architecture mix and clipping give .000, .082, .013, .045, .024
Δ_{dia}	empirical independent-pair plug-in with valid penalty	source auxiliary split and disjoint test targets; D_0 optimized on test energies	no derangement or reused row; fixed per-class schedule; compute per seed, then average	five 128-dimensional Diamonds checkpoints, 5,394 pairs each; mean $-.0168$, valid certificate zero

Table 10: Partition-varying descriptive inputs Δ_{part} at $k/d=1/4$; the final column squares the difference of the seed-averaged D_0 and D_{cf}.

Dataset	D_0	D_{cf}	Plug-in bits
Breast Cancer	.628	.611	.0009
Iris	.537	.469	.0135
Wine	.618	.524	.0260
Digits	.624	.528	.0269
Synthetic-10	.671	.594	.0181

The small plug-in values and larger Fano certificates measure different sufficient pathways and do not imply tightness. The population theorem establishes a mechanism implication, not a numerical estimator of the attack-based certificate. A model-instance bootstrap gives the following descriptive uncertainty; it is not a fixed-instance confidence interval. This scale analysis explains why the valid finite-sample cross-fit certificate is zero at current n without conflating that result with the much larger decoder-based Fano certificate. An empirical resampling planning audit makes the scale concrete: over fixed-checkpoint empirical pair distributions, the dimension-free certificate is

Table 11: Model-grid quantity Δ_{grid} across trained instances. The positive part is taken within each instance before averaging. n_+ is the approximate sample size at which the dimension-free penalty would become smaller than the observed mean; it is a planning diagnostic, not a guarantee.

Dataset	Instances	Gap mean $\pm$ sd	Bootstrap 95%	n_+
Cancer	10	$.000 \pm .000$	$[.000, .000]$	–
Digits	15	$.082 \pm .022$	$[.071, .092]$	1,085
Iris	10	$.013 \pm .019$	$[.003, .026]$	41,036
Synthetic-10	15	$.045 \pm .015$	$[.038, .053]$	3,575
Wine	10	$.024 \pm .022$	$[.011, .037]$	12,608

positive in 92% of Digits replicates at $n = 1{,}000$ and 100% at $n = 2{,}000$; Synthetic-10 reaches 57% at $n = 2{,}000$ and 100% at $n = 5{,}000$. This bootstrap-style extrapolation diagnoses sample demand only; it is not a new confidence statement or a substitute for independently collected messages.

B.9 Batch Semantics and Proof of Proposition 3.7

For an additive minibatch loss $\mathcal{L}_B = B^{-1}\sum_i \mathcal{L}_i$, automatic differentiation returns a tensor G_B with one row per example (possibly scaled by $1/B$). Scaling a row by a common positive constant does not change top-k support. If Comp_B is defined row-wise, its ith output is a deterministic/randomized function of G_i and row-specific public randomness only; processing other rows cannot change M_i. This proves the proposition.

The experiment verifies the implementation directly by comparing each support obtained (i) in a batch and (ii) from the same row processed alone. Agreement is exactly 1.000 at $B = 1, 8, 32, 64, 128, 256$. A separate experiment forms $\bar{G} = B^{-1}\sum_i G_i$ and sends one sparse vector. That protocol loses row identity; flattened, blockwise, accumulated, and microbatch-compressed variants likewise fall outside Proposition 3.7.

Averaged-vector batch size	2	4	8	16
Class-presence macro-AUROC	.752	.618	.531	.540

C Audit Implementations and Models

C.1 Attack Implementations and Channel Interpretation

Decoder selection without test leakage. Input support is $M \in \{0,1\}^d$; joint input is $[M, \widetilde{G}] \in \mathbb{R}^{2d}$ with coordinate identity retained. The auxiliary partition is split 70/30, stratified, for decoder selection. Logistic regression chooses $C \in \{.03, .1, .3, 1, 3, 10\}$; random forests choose between $(\text{depth}, \text{leaf}) = (12, 1)$ and $(\text{None}, 3)$ with 100 trees; shallow MLPs choose $\alpha \in \{10^{-4}, 10^{-2}\}$ with one 64-unit hidden layer and auxiliary-only early stopping. The selected model is refit on all auxiliary messages before the untouched test set is evaluated. Joint continuous features are standardized using auxiliary statistics only. The prototype decoder uses classwise auxiliary mean masks and has no fitted hyperparameter.

Demonstrated-attack interpretation. We report the explicit joint decoder and $\max\{a_{\text{support}}, a_{\text{joint}}\}$ as a demonstrated lower envelope so finite-sample decoder capacity is not misread as an information ordering. This is not an estimator of mutual-information ordering. Coordinate-sorted values are an index-removal ablation only.

Per-class recall and confusion summaries are released. On Digits the logistic top-k minimum recall averages $.594$; the joint logistic minimum recall averages $.726$. Thus high aggregate accuracy is not produced by one majority class, although some classes remain harder. Probability diagnostics are also released. On Digits top-k, random forests obtain mean NLL $.457$ and top-label ECE $.088$, while logistic, prototype, and MLP decoders have NLL $.693, .961, .974$; under random-k, all families remain near the uniform ten-class NLL (2.30–2.52). Accuracy ordering is therefore not an artifact of one solver, although probability calibration varies across families.

Table 12: Decoder-family ablation at $k/d = 1/4$; mean $\pm$ sd over five Digits and three Synthetic-10 lineages. Every hyperparameter is selected using auxiliary data only.

Dataset	Decoder	Top-k support	Random-k support	Top-k joint	Random-k joint
Digits	Prototype	$.787 \pm .043$	$.098 \pm .026$	–	–
Digits	Logistic	$.849 \pm .045$	$.097 \pm .024$	$.885 \pm .030$	$.153 \pm .020$
Digits	Random forest	$.859 \pm .032$	$.096 \pm .012$	–	–
Digits	Shallow MLP	$.807 \pm .029$	$.097 \pm .010$	$.809 \pm .042$	$.109 \pm .023$
Synthetic-10	Prototype	$.701 \pm .031$	$.101 \pm .004$	–	–
Synthetic-10	Logistic	$.747 \pm .009$	$.103 \pm .002$	$.853 \pm .003$	$.222 \pm .008$
Synthetic-10	Random forest	$.759 \pm .028$	$.095 \pm .003$	–	–
Synthetic-10	Shallow MLP	$.743 \pm .018$	$.102 \pm .006$	$.801 \pm .033$	$.199 \pm .014$

Metrics and transfer. Accuracy, balanced accuracy, macro-AUROC where defined, per-class recall, minimum recall, Adv, explicit joint accuracy, and conditional Fano quantities are released per seed. Primary auxiliary and test messages use the same checkpoint. Separate experiments vary labeled messages per class, remove calibration classes, perturb auxiliary preprocessing, replace the private tail, transfer across nearby checkpoints, and transfer across independent model seeds.

C.2 Datasets, Models, and Training

Dataset	Examples	Classes	Input dimension	Natural majority
Breast Cancer	569	2	30	.632
Iris	150	3	4	.333
Wine	178	3	13	.389
Digits	1,797	10	64	.100
Synthetic-10	5,000	10	64	.101
Diamonds	53,940	8	mixed	.242
CIFAR-10	60,000	10	$32\times32\times3$	.100

Synthetic-10 is regenerated per seed with `make_classification`: 5,000 observations, 64 features, 32 informative and 16 redundant features, 10 classes, one cluster per class, class separation 1.55, label-flip probability .015, and random state $1000 +$ seed. Primary 60/20/20 splits are stratified and preprocessing is fitted only on model-training data.

Diamonds uses clarity as the eight-class target. Its natural test majority is .242; balanced accuracy instead weights all eight classes equally and has chance level .125. Numerical variables are standardized; cut and color are one-hot encoded using model-training categories. The split MLP has cut dimension 128 and is trained for eight epochs. Each mechanism is used in every server-to-client backward message. The released CSV records all five seeds, task accuracy, balanced accuracy, macro-AUROC, support and joint attacks, omitted energy, and bits.

For CIFAR-10, 45,000 official training images train the split model, 5,000 class-balanced training images are reserved for the auxiliary decoder, and the 10,000 official test images remain untouched. Training uses random crops, horizontal flips, channel normalization, SGD with momentum, and a cosine schedule. A residual client stem and residual server body end in a 128-dimensional pooled cut; the client-side private tail is a two-layer nonlinear MLP. Dense, top-k, and random-k use seeds $\{1, 19, 43\}$ for the six-epoch comparison. Separate top-k runs use the same seeds and data partition with terminal 24- and 48-epoch cosine schedules; each reaches learning rate zero only at its declared terminal epoch.

The primary split MLPs use five independently initialized checkpoints. Nonlinear private tails provide an architecture control. A three-cut convolutional Digits model exposes dimensions 1024, 512, and 256; five-seed support accuracies are .862, .869, and .867.

C.3 Architecture, Cut, and Real-Image Controls

The procedural RGB stress test isolates dimension and architecture. Residual cuts have dimensions 4096, 2048, and 1024; TinyViT has three cuts of dimension 2048. The reported 300 evaluations are 2 architectures $\times 3$ cuts $\times 5$ seeds $\times 2$ sparsities $\times 5$ mixture probabilities. Because task accuracy is near saturated, this experiment is used only for support scaling, not utility conclusions.

The packaged LFW subset contains 100 face and 100 non-face images with disjoint 60/20/20 base-image splits. At $k/d = 1/16$, support accuracy is $.900, .955, .500$ across residual early/middle/late cuts and $.930, .930, .895$ across TinyViT cuts. The late residual cut uses a binary linear tail. Across epochs $1, 3, 6, 18$, it releases one unique mask, support accuracy remains $.5$, test loss changes from $.642$ to $.050$, and mean gradient norm changes from $.093$ to $.010$ without zero gradients. Thus the null is explained by Proposition 3.8, not by perfect fitting or vanishing gradients. The coordinate-aligned values still leak, so this is a support-only mitigation lever under a restricted geometry.

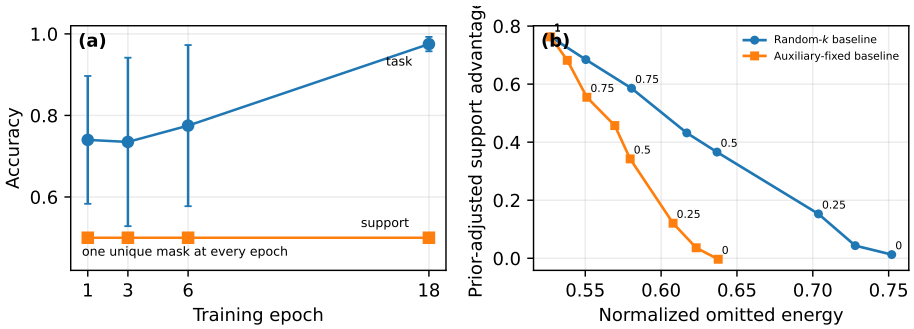


Figure 5: (a) The late binary linear tail has one mask at every checkpoint despite nonzero loss and gradients. (b) Explicit mixtures with uniform random-k and auxiliary-fixed label-independent branches; labels are adaptive-branch probabilities. The former uses the general D_b bound, while the latter approximates the D_0-optimal special case.

C.4 Limited CIFAR-10 Split-ResNet Check

Compression is active in every training cut message rather than applied to a frozen dense model. The residual body uses batch normalization and image augmentation, and the private nonlinear tail makes the cut-gradient geometry distinct from the tabular MLPs and Digits convolutional control. Top-k

Table 13: CIFAR-10 after six training epochs (three seeds). Top-k and random-k have identical exact-k communication at $k/d = 1/4$; Bits uses literal 32-bit values plus explicit 7-bit indices.

Training mechanism	Task	Support	Joint	Bits
Dense	$.751 \pm .015$	$.100 \pm .000$	$.996 \pm .001$	4,096
Top-k	$.724 \pm .008$	$.968 \pm .001$	$1.000 \pm .000$	1,248
Random-k	$.703 \pm .004$	$.099 \pm .001$	$.922 \pm .012$	1,248

retains 2.1 task-accuracy points over random-k at equal communication, but exposes a nearly deterministic label support code. Random support closes that code while its retained coordinate-aligned values remain highly predictive. Dense is a training reference, not a communication-matched competitor.

The same-lineage/independent-initialization separation therefore replicates in a second image dataset as well as Synthetic-10. The residual independent score is above exact ten-class chance but far below matched calibration.

Table 14: CIFAR-10 top-k calibration transfer. The stale source is epoch 3 and the target is epoch 6; independent entries are six off-diagonal source/target seed pairs.

Calibration source	Accuracy	Balanced accuracy
Same checkpoint	$.968 \pm .001$	.968
Same lineage, half-training gap	$.834 \pm .015$	.834
Independent initialization	$.143 \pm .047$	.143

Table 15: Matched-checkpoint top-k audits on independent terminal schedules (three seeds). The 48-epoch schedule is not a continuation of the 24-epoch run; only each terminal row has learning rate zero.

Schedule	Epoch	Task	Support	Joint	DPI LCB
24 terminal	24	$.880 \pm .002$	$.930 \pm .002$	$1.000 \pm .000$	–
48 terminal	24	$.832 \pm .024$	$.956 \pm .005$	$1.000 \pm .000$	$.384 \pm .009$
48 terminal	36	$.887 \pm .002$	$.928 \pm .002$	$.999 \pm .000$	$.392 \pm .038$
48 terminal	48	$.902 \pm .002$	$.919 \pm .002$	$.998 \pm .000$	$.404 \pm .034$

On the 48-epoch schedule, support inference decays by 3.8 points from epoch 24 to 48 while task utility gains 7.0 points. It remains far above the .1 prior, every one of the nine coordinate certificates is positive, and the sign-vector baseline is 1.000 at the terminal checkpoint. This addresses the requested longer horizon within the compact model; it does not establish prevalence in deeper ResNets, CIFAR-100, ImageNet, transformers, or production client heterogeneity.

D Selectors, Utility, and Defenses

Table 16: Orthogonal channel controls. "Low" means no detected support advantage under the stated audit, not a joint privacy guarantee.

Control	Support	Conditional values	Consequence
Deterministic top-k	high	informative	both channels are exposed
Value noise after selection	high	suppressed	support remains exploitable
Binary linear tail	low/constant	informative	cut geometry closes support only
Random-k	low	informative	data-independent support is insufficient
Support DP (V open)	formal; reduced	informative	selection privacy changes ranking, not joint privacy
Public support + Gaussian (Digits only)	low	protected once	complete one-message baseline, but high distortion

D.1 Adaptive Selector Definitions and Results

All stochastic mechanisms release exactly k coordinates. Randomized top-k assigns 80% probability mass uniformly to the deterministic top-k set and 20% to the remainder before sampling without replacement. Magnitude sampling uses probabilities proportional to $|g_j|$. Gumbel top-k selects the largest $|g_j| + \zeta_j$, with Gumbel scale .35 times the empirical standard deviation of magnitudes. Adaptive dropout uses sigmoid probabilities centered at the kth magnitude, with temperature .35 times the row standard deviation, followed by exact-k sampling. These are evaluation controls, not proposed defenses.

Synthetic-10 has the same qualitative ordering. Full results for $k/d \in \{1/8, 1/4, 1/2\}$ are released.

Mixture alignment. The empirical random-k/top-k curve uses a uniform random-k baseline, so it is evaluated with the arbitrary-branch form of Proposition 3.2. A second curve uses one fixed support selected from auxiliary normalized energies without labels. The table reports descriptive cross-fitted endpoint plug-ins; D_0 is the oracle empirical optimum on the evaluation targets and is not a finite-sample certificate. Across all seeds and mixture probabilities, the mean absolute difference between Monte Carlo mixture distortion and the affine endpoint prediction is $.0040$. The fixed auxiliary

Table 17: Digits post-training selectors at $k/d = 1/4$ (five seeds).

Mechanism	Adv	Joint	Omitted energy
Top-k	.836	.879	.325
Adaptive dropout	.641	.686	.446
Randomized top-k	.508	.577	.486
Magnitude sampling	.251	.341	.582
Gumbel top-k	.026	.150	.713
Random-k	-.001	.139	.750
Fixed support	-.003	.231	.633

Table 18: Digits mixture alignment at $k/d = 1/4$ (five seeds). "LB@.5" is $2[D_0 - D_{.5}]_+^2 / \ln 2$ using the appropriate D_b.

Baseline branch	D_0	D_b	D_1	LB@.5 (bits)	Attack@.5
Uniform random-k	.633	.752	.527	.0000	.429
Auxiliary fixed	.633	.638	.527	.0076	.408

branch nearly matches the empirical optimum, whereas uniform random-k does not; this is exactly why the general D_b expression is required.

Parameter sensitivity. To separate mechanism behavior from arbitrary control settings, we sweep four selector parameters on Digits over five checkpoints. Randomized top-k assigns total mass $q \in \{.25, .4, .6, .8, .95, 1\}$ to the deterministic top set; $q = .25$ is uniform when $k/d = .25$, and $q = 1$ recovers deterministic top-k. Gumbel scale is swept over $\{0, .05, .1, .2, .35, .7, 1.4\}$ row-magnitude standard deviations; adaptive-dropout temperature over $\{.05, .1, .2, .35, .7, 1.4, 3\}$; and magnitude-sampling power over $\{0, .25, .5, 1, 2, 4\}$. The same frozen prototype decoder is used throughout each curve.

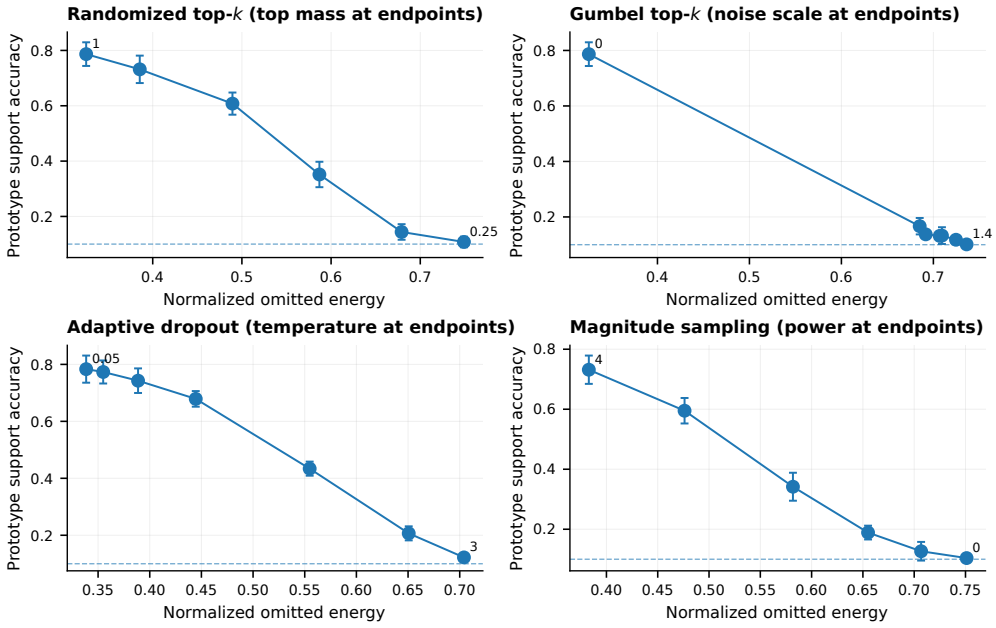


Figure 6: Five-seed selector sensitivity at $k/d = 1/4$. Each panel sweeps one mechanism from strongly magnitude-adaptive toward nearly data-independent selection. Across all four families, lower omitted energy is accompanied by higher support exploitability. Dashed lines mark the class prior.

D.2 Compressed-Training Utility, Privacy, and Communication

On Digits, compression is applied at every backward step for 24 epochs with equal batch size, bit budget, and optimizer schedule. Task accuracy changes little, so this is a controlled optimization check rather than evidence that compression is free. The complete grid includes all selector families used in the post-training audit. The support ordering therefore survives training-time use and

Table 19: Training-time mechanisms on Digits at $k/d = 1/4$ (five seeds, equal 280-bit messages).

Mechanism	Task	Support	Joint	Omitted energy
Top-k	$.964 \pm .004$	$.918 \pm .015$	.949	.284
Error-feedback top-k	$.966 \pm .002$	$.885 \pm .026$	.921	.298
Adaptive dropout	$.966 \pm .003$	$.709 \pm .040$	.753	.428
Randomized top-k	$.964 \pm .003$	$.646 \pm .039$	.682	.469
Magnitude sampling	$.969 \pm .004$	$.388 \pm .031$	.463	.573
Gumbel top-k	$.966 \pm .002$	$.139 \pm .028$	.198	.687
Random-k	$.963 \pm .006$	$.088 \pm .015$	.184	.747
Fixed public support	$.957 \pm .004$	$.097 \pm .000$	.472	.663

is not an artifact of applying selectors to a frozen model. Error-feedback residuals are keyed by stable training-row IDs; support and joint attacks use disjoint final-checkpoint rows with zero initial residual, so that row is a fresh-message audit rather than a transcript claim. Representative 24-epoch curves show no late reversal. Under an equal three-seed schedule, AdamW task accuracy is $.967/.967/.966$ for top-k/randomized/random-k; SGD gives $.967/.963/.950$, so the qualitative conclusion is not specific to one optimizer.

The Diamonds experiment supplies the non-degenerate utility axis and extends the training-time comparison to seven stateless exact-k selectors; its curves and five-seed endpoints appear once in Figure 2. Dense task accuracy is $.546 \pm .006$ with macro-AUROC .904. Mechanism choice changes

Table 20: Diamonds compressed training (five seeds). "Bal. sup." is balanced support-attack accuracy; Bits uses literal 32-bit values plus explicit 7-bit indices. All rows transmit unperturbed values at the selected coordinates; no value noise is added in any Diamonds configuration.

Mechanism	k/d	Task	Bal. sup.	Joint	Omitted energy	Bits
Adaptive dropout	$1/8$	$.526 \pm .006$	.790	.805	.590	624
Fixed public	$1/8$	$.530 \pm .011$	.125	.561	.745	624
Gumbel top-k	$1/8$	$.523 \pm .007$	.749	.785	.584	624
Magnitude sampling	$1/8$	$.530 \pm .006$	.557	.719	.759	624
Randomized top-k	$1/8$	$.521 \pm .006$	.700	.721	.538	624
Random-k	$1/8$	$.536 \pm .005$	.125	.416	.875	624
Top-k	$1/8$	$.521 \pm .006$	.845	.845	.452	624
Adaptive dropout	$1/4$	$.533 \pm .006$	.811	.810	.427	1248
Fixed public	$1/4$	$.541 \pm .001$	.125	.606	.531	1248
Gumbel top-k	$1/4$	$.527 \pm .005$	.801	.797	.370	1248
Magnitude sampling	$1/4$	$.536 \pm .006$	.705	.777	.545	1248
Randomized top-k	$1/4$	$.529 \pm .006$	.703	.712	.349	1248
Random-k	$1/4$	$.540 \pm .004$	.121	.459	.750	1248
Top-k	$1/4$	$.528 \pm .005$	.861	.851	.246	1248

task performance, balanced support leakage, joint leakage, and omitted energy under identical budgets. Fixed or random support closes the support channel but not the value channel; deterministic top-k minimizes omitted energy while giving the strongest support attack. The natural-distribution majority baseline is .242, raw random-k attack accuracy is .105–.109, and balanced chance is .125; these are deliberately reported as different quantities.

The table prioritizes the literal encoding used by ordinary sparse-tensor serialization. The ideal enumerative totals are 579 and 1,125 bits at $k/d = 1/8$ and $1/4$; privacy conclusions are invariant to the lossless code, but deployment cost is not.

D.3 One-Message DP Support Sweep with Unprotected Values

To isolate what a support-only guarantee can and cannot accomplish, define $s_j(g) = |g_j|/\max_\ell |g_\ell|$ with $s(0) = 0$. Every score lies in $[0, 1]$, so $\max_j |s_j(g) - s_j(g')| \leq 1$ for arbitrary sub-

stitute gradients g, g'. We select an ordered sequence of k distinct coordinates by k conditional exponential-mechanism draws, each with privacy parameter ε_M/k and probability proportional to $\exp\{\varepsilon_M s_j/(2k)\}$. Adaptive composition makes the ordered sequence ε_M-DP; discarding order leaves the exact-k mask ε_M-DP. Gumbel top-k with scale $2k/\varepsilon_M$ implements the same Plackett–Luce draw. This is a standard private-selection construction specialized to the cut-gradient score (Qiao et al., 2021; Gillenwater et al., 2022).

The selector is active in every Diamonds backward step at $k/d = 1/4$. Every point sends the same 1,248 literal bits. The original values at selected coordinates are deliberately transmitted: the guarantee covers M for one message, not V, (M, V), or repeated training. The equal-budget random-k

Table 21: Diamonds one-message support-DP sweep (five seeds). Support is balanced attack accuracy; $8\varepsilon_M$ is basic per-record composition over eight training releases, not a joint-message guarantee.

ε_M	Task	Support	Joint	Omitted energy	$8\varepsilon_M$
1	$.542 \pm .011$	$.126 \pm .005$	$.457 \pm .014$	.749	8
2	$.541 \pm .012$	$.127 \pm .005$	$.458 \pm .012$	.748	16
4	$.542 \pm .011$	$.126 \pm .005$	$.462 \pm .009$	.746	32
8	$.542 \pm .012$	$.126 \pm .004$	$.466 \pm .012$	.742	64
16	$.542 \pm .013$	$.138 \pm .007$	$.474 \pm .010$	.735	128
32	$.543 \pm .013$	$.171 \pm .007$	$.505 \pm .011$	.719	256

reference has task/support/joint $.540/.121/.459$. None of the six private selectors improves both task and either attack axis over that reference: $\varepsilon_M \leq 8$ is effectively the same operating point within five-seed variation, and larger ε_M leaks more. The formal support guarantee is still meaningful, but it does not buy an empirical frontier improvement in this task. The full stream contains 302,064 row messages, but each training record appears once per epoch and therefore has eight linkable releases under full participation. Parallel composition within an epoch and basic sequential composition across epochs give the illustrative $8\varepsilon_M$ column. This accounting still excludes V. On a 2,048-by-128 CPU microbenchmark, private selection takes 3.54–3.60 ms per batch versus 1.86 ms for top-k (1.90–1.93×); it is not the dominant limitation here. The sweep therefore sharpens the negative design conclusion: support protection is effective for M, but the coordinate-aligned values prevent a private-trainer claim at every operating point.

D.4 TPSL GradPerturb Under Top-k

TPSL is formulated for binary two-party split learning. Its discrete primitive draws $u \sim \mathrm{Bernoulli}(p)$ and releases

$$\widetilde{g}_y = g_y + u(g_{1-y} - g_y), \qquad \varepsilon = \log \frac{1-p}{p}. \tag{41}$$

For cross-entropy, gradients are linear in the target distribution. Replacing the one-hot label by its counterfactual when $u = 1$ therefore applies exactly the same GradPerturb draw to both the cut gradient and private-tail parameter gradient. We do so during every training step, then apply deterministic top-k to the perturbed cut gradient. Since top-k is post-processing, the one-message label-DP primitive covers both the resulting support and values under TPSL's fixed-feature neighboring-label definition (Yang et al., 2022).

The task filters public Diamonds to its two most frequent clarity grades, SI1 and VS2 (25,323 rows), uses a 70/10/20 split, a nonlinear private tail, $d = 128$, $k = 32$, eight epochs, and five seeds. The support decoder receives M; the value-only control receives the sorted retained scalar multiset with indices removed; the joint decoder receives $[M, \widetilde{G}]$. The label prior is .516. At $\varepsilon = 0$, message-only accuracy can remain above this prior because the perturbed gradient also depends on the forward representation, which is correlated with the label; TPSL's guarantee is conditional on the feature record. The flip prediction explains much of the high-ε curve and is therefore a sanity check, not the finding. The positive excess at every protected setting is the substantive result and is consistent with forward-representation correlation outside a symmetric flip channel. This is a representative sparse-message comparison, not a claim that TPSL generally fails: the original transcript theorem

Table 22: TPSL discrete GradPerturb primitive before top-k (five-seed means). p is the counterfactual-label probability. "Flip" is $(1-p)a_0 + p(1-a_0)$ from the same-seed unprotected support score; excess is measured support minus this analytic sanity check.

Per-message ε	p	Task	Support	Flip	Excess	Joint
0	.500	$.513 \pm .028$	$.571 \pm .004$	.500	.071	$.571 \pm .004$
1	.269	$.701 \pm .014$	$.695 \pm .012$	.617	.079	$.765 \pm .010$
2	.119	$.752 \pm .011$	$.746 \pm .008$	.692	.054	$.887 \pm .003$
4	.018	$.764 \pm .007$	$.756 \pm .008$	.743	.013	$.977 \pm .001$
Unprotected	0	$.765 \pm .008$	$.752 \pm .003$	.752	.000	$1.000 \pm .000$

also perturbs model updates under its online-use assumptions. Here every record appears in eight epochs; basic message composition is 8ε, and no composed or end-to-end guarantee is claimed.

D.5 Paired Frontier Uncertainty

All Diamonds mechanisms use the same five declared training seeds, enabling paired task differences against dense and seedwise dominance checks. The intervals below are percentile intervals from 20,000 paired seed resamples; Pareto frequencies use 50,000 paired resamples. They characterize these five trained instances and are not high-confidence coverage statements for a deployment population. Random-k dominates top-k, randomized top-k, magnitude sampling, Gumbel top-k,

Table 23: Diamonds $k/d = 1/4$ paired uncertainty. Lower support attack and higher task accuracy are preferred.

Mechanism	Task Δ dense (95%)	Support (95%)	Pareto frequency
Adaptive dropout	$-.013$ $[-.016, -.011]$	.811 [.805, .816]	.000
Fixed public	$-.006$ $[-.010, -.002]$	.125 [.125, .125]	.560
Gumbel top-k	$-.020$ $[-.023, -.016]$	.801 [.787, .815]	.000
Magnitude sampling	$-.010$ $[-.013, -.007]$	.705 [.690, .719]	.000
Randomized top-k	$-.018$ $[-.022, -.014]$	.703 [.693, .715]	.000
Random-k	$-.006$ $[-.009, -.003]$	.121 [.117, .125]	.987
Top-k	$-.018$ $[-.021, -.015]$	.861 [.850, .871]	.000

and adaptive dropout in all five seeds; fixed support and random-k are incomparable in three seeds and each dominates once. Complete pair counts are released.

For the separate post-training Digits frontier, the claim being tested is whether an intrinsic selector simultaneously has no greater distortion and no greater prior-adjusted leakage than an explicit random-k/top-k mixture. The $\lambda = .1$ mixture is nondominated in only one of five individual seeds and .703 of paired-bootstrap means; $\lambda = .25$ is nondominated in four of five seeds and .965 of resamples; all points with $\lambda \geq .5$ are nondominated in every seed and all 50,000 resamples. This replaces the previous unqualified mean-frontier wording.

D.6 Value Perturbation and a One-Message Joint-Private Baseline

Post-selection value noise. After deterministic top-k selects M, Gaussian noise with standard deviation λ times the retained-value standard deviation is added to V. Across five Digits checkpoints, support accuracy is exactly .8522 for every $\lambda \in \{0, .25, .5, 1, 2, 4, 8\}$. Conditional cross-entropy gain from values beyond the mask falls from .0564 to zero by $\lambda = 1$, while the joint lower envelope remains at least the support score.

Public-support Gaussian construction. Fix support S_0 independently of the example, clip selected values to ℓ_2 radius one, and add Gaussian noise for replace-one sensitivity $\Delta_2 \leq 2$. Analytic calibration (Balle & Wang, 2018) gives one-message (ε, δ)-DP. On Digits at $\delta = 10^{-5}$, support accuracy is .097, while relative message distortion is $37.42, 11.97, 4.35, 1.95$ for $\varepsilon = 4, 8, 16, 32$. This is a formal baseline, not a practical or composed defense. It is run on Digits only; we report no joint-

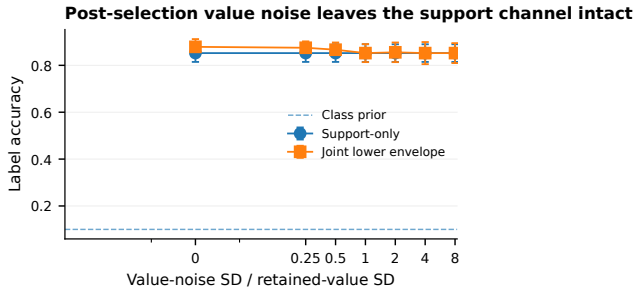


Figure 7: Post-selection Gaussian noise on retained values. Support accuracy is invariant because the mask is selected before perturbation; conditional value gain vanishes while the joint observation remains at least as informative as the support.

decoder score for it and do not evaluate it on Diamonds, where every fixed-support configuration in Table 20 transmits unperturbed values.

E Robustness and Transcript Controls

E.1 Threat-Model and Architecture Controls

Auxiliary availability and preprocessing. Across five lineages, mean Digits support accuracy is $.653, .746, .793, .834, .853$ with $2, 5, 10, 20, 40$ labeled messages per class. Applying affine scaling plus calibration-only Gaussian input noise of $0, .1, .3, .5, 1.0$ standardized units yields $.850, .857, .841, .850, .841$, so this perturbation is not the principal boundary.

Class coverage, checkpoint, and private-tail transfer. Removing 2, 4, or 6 classes from calibration reduces overall accuracy to $.728, .548, .367$ because unseen classes cannot be decoded; accuracy restricted to covered classes remains $.907$–$.917$. The $.852$ matched-checkpoint number is the fixed final checkpoint used by the primary audit; the $.886$ zero-gap number averages the diagonal of a separate nine-snapshot within-lineage grid. That grid declines to $.825$ at a 32-epoch gap. An independently initialized private tail attached to the same encoder/server lowers transfer to $.321 \pm .078$, and independent full-model seeds average $.106$. Thus the learned support code is checkpoint- and tail-family specific.

Second lineage replication. The same predeclared protocol on Synthetic-10 uses five 32-epoch lineages and top-k at $k/d = 1/4$. Matched-final-checkpoint calibration obtains $.749 \pm .024$ accuracy; calibration from epochs 24, 16, and 8 to the final checkpoint obtains $.752 \pm .026$, $.747 \pm .024$, and $.711 \pm .028$, respectively. The 20 off-diagonal final-checkpoint source/target seed pairs average $.105 \pm .035$. The exact values differ from Digits, but the same-lineage versus independent-initialization boundary replicates on a second dataset.

Calibration from disjoint clients. We assign auxiliary and test rows to 10 or 20 clients using the same Dirichlet class-mixture draw, train the support decoder only on alternating calibration clients, and evaluate on the disjoint client half under one shared checkpoint. This models a server obtaining consented labels or collusion from some clients rather than from the target client. The result makes calibration access more operational but not automatic: it depends on a shared checkpoint family and sufficient class coverage among calibration clients. It does not contradict near-prior transfer across independently initialized models.

Client turnover and client-specific preprocessing. Starting from ten labeled calibration clients and ten disjoint targets at Dirichlet $\alpha = .3$, we randomly retain $25, 50, 75, 100\%$ of calibration clients over five checkpoints and 12 retention draws per checkpoint. Turnover hurts mainly through message count and lost class coverage. Giving each calibration client an independent affine/noise preprocessing perturbation of severity 0–1 changes attack accuracy only from $.778$ to $.761$–$.771$ in

Table 24: Cross-client calibration transfer on Digits (five seeds).

Clients	Dirichlet α	Target prior	Attack	Adv
10	.1	.242	.564	.425
10	.3	.167	.729	.671
10	1.0	.146	.794	.759
20	.1	.168	.688	.623
20	.3	.173	.769	.721
20	1.0	.150	.790	.752

Table 25: Operational calibration stress. Coverage is the fraction of target labels represented among retained calibration clients.

Retained clients	Messages	Coverage	Attack	Balanced
25% (2)	37	.755	$.453 \pm .086$	.475
50% (5)	89	.971	$.653 \pm .085$	.674
75% (8)	142	1.000	$.740 \pm .062$	.768
100% (10)	178	1.000	$.778 \pm .057$	.802

this standardized-input model; this robustness is protocol-specific and does not negate the much stronger private-tail or independent-lineage failures.

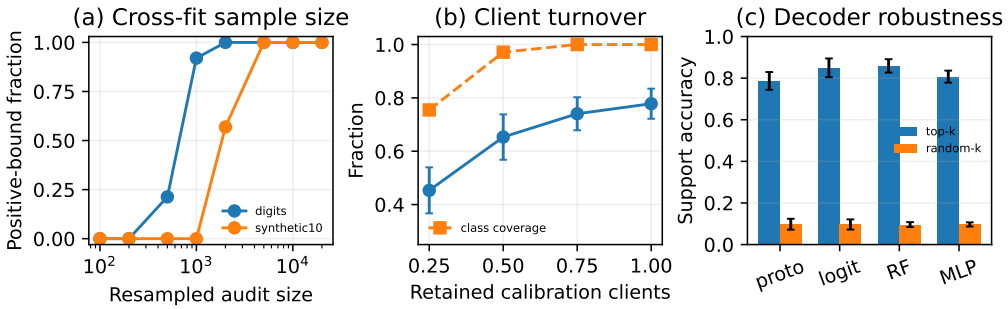


Figure 8: Additional validation. (a) Empirical resampling frequency of a positive dimension-free cross-fit certificate; this is a planning diagnostic, not a coverage guarantee. (b) Attack degradation and class coverage under calibration-client turnover. (c) Top-k/random-k ordering across four auxiliary-selected support decoders on Digits.

Repeated messages and row linkage. The same-example experiment uses fixed row identities across checkpoints, corresponding to stable order or persistent identifiers. Concatenating $R = 1, 2, 4, 8$ linked supports gives .852, .859, .877, .930 accuracy. For $R = 8$, the available attack lower bound is .852, .852, .852, .873, .930 when each earlier row is correctly linked with probability $0, .25, .5, .75, 1$, respectively. Thus cumulative improvement is a linkability-dependent threat, not an automatic consequence of repeated training or a composed DP parameter.

Nonlinear tails, clients, and permutation. A nonlinear private tail retains .842 support accuracy versus a .101 prior. Across 5–50 heterogeneous clients, attacks condition on client identity and retain .75–.88 prior-adjusted advantage. Permuting auxiliary labels returns inference near prior; the signed-rank value is reported as the minimum attainable 2^{-10} for ten matched pairs.

E.2 Proof and Scope of the Linkability Bound

Augmenting the observation cannot reduce information:

$$I(Y;O \mid \Omega) \le I(Y;O,L \mid \Omega) \tag{42}$$
$$= I(Y;L \mid \Omega) + I(Y;O \mid L,\Omega) \tag{43}$$
$$= I(Y;L \mid \Omega) + \Pr(L=1)I(Y;M \mid L=1,\Omega) \tag{44}$$
$$+ \Pr(L=0)I(Y;U \mid L=0,\Omega), \tag{45}$$

which proves Equation (15). If $L \perp\!\!\!\perp Y \mid \Omega$, the first term is zero. If false matches are also independent of the target label, the last term is zero and the idealized p-gate follows. This extra term is essential: under client-level class skew, temporal correlation, class-structured sessions, or same-class replacement, a wrongly linked row can remain informative about Y. For multiple earlier messages, the inequality applies conditionally at each history.

E.3 Correlated and Session-Level False-Link Audits

The oracle stress test replaces an incorrectly linked row by a different row with the same label; it reaches .998 at zero correct links and is not a practical threat model. A non-oracle audit instead assigns rows to 5, 10, or 20 public sessions using Dirichlet class mixtures with $\alpha \in \{.05, .1, .3, 1\}$ and draws false matches from the same session without using the target label. Across five seeds, session-only prediction is near prior, and false-link-only transcript decoders score .678–.794, below the frozen single-message accuracy .852. At $p = .5$, the strongest session setting reaches .847 and still does not improve the available attack; gains emerge at $p = .75$ or perfect linkage. Thus realistic client-skew correlation can preserve signal in false rows, but the measured non-oracle laws do not beat the single-message envelope at low linkage.

Table 26: Representative 20-session, $\alpha = .05$ non-oracle false-link result. "Available" is the better of the transcript and frozen one-mask decoder.

p_{link}	0	.25	.50	.75	1
Transcript	.769	.798	.847	.909	.930
Available	.852	.852	.858	.909	.930

E.4 Transcript Identities and Empirical Cumulative Attack

The chain rule always gives

$$I(Y;M_{1:R} \mid \Omega) = \sum_{r=1}^{R} I(Y;M_r \mid M_{<r},\Omega). \tag{46}$$

If, at each realized history $h_r = (M_{<r},\Omega)$, one can generate conditionally i.i.d. same-label gradients $G_r, G_r' \mid (Y,h_r)$ and estimate a conditional cross-fit gap $\Delta_r(h_r)$, then applying Theorem 3.1 conditionally yields

$$I(Y;M_{1:R} \mid \Omega) \ge \frac{2}{\ln 2}\sum_{r=1}^{R} \mathbb{E}[\Delta_r(h_r)^2]. \tag{47}$$

Ordinary consecutive gradients are dependent and nonstationary, so the conditional cross-fit gaps in this optional bound are not estimated from one run. The empirical transcript experiment trains a decoder on concatenated supports for the same example and reports $R = 1, 2, 4, 8$ accuracies $.852, .859, .877, .930$. A separate identity-randomization audit shows no gain beyond the single-message decoder when correct linkage is at most .5. These experiments measure exploitability under stated linkage, not a privacy composition parameter.

E.5 Linear-Tail Geometry and Gaussian Sufficiency

For logits $W^\top H + b$, softmax prediction p, and one-hot label e_Y, $G = W(p - e_Y)$. In the binary case, write $W = [w_0, w_1]$. If $Y = 0$, then $p - e_0 = p_1(-1, 1)$; if $Y = 1$, then $p - e_1 = -p_0(-1, 1)$.

Therefore

$$G = a(X, Y)(w_1 - w_0), \tag{48}$$

with nonzero scalar a for finite logits. Absolute coordinate ranks depend only on $|w_1 - w_0|$, proving Proposition 3.8. The values retain the scalar's sign and magnitude, so constant support does not imply joint privacy.

For $\widetilde{G} = W\Delta + \xi$ with fixed full-column-rank W and isotropic Gaussian ξ, the orthogonal component $(I - P_W)\widetilde{G} = (I - P_W)\xi$ is independent of both Δ and $P_W\widetilde{G}$. Since $P_W\widetilde{G} = W(W^\dagger\widetilde{G})$,

$$I(\Delta; \widetilde{G} \mid W) = I(\Delta; W^\dagger\widetilde{G} \mid W). \tag{49}$$

E.6 Binary-Tail Finite-Precision Boundary

The primary Cancer MLP uses the same binary *linear* softmax tail as Proposition 3.8; it is not nonlinear. Its original float32 mask result is therefore an implementation-precision phenomenon. At the highly fitted checkpoints, softmax saturation and cancellation produce exact zeros and non-collinear residuals: 3.9% of test rows are exactly zero, nonzero rows have median relative orthogonal residual $.294$, and only $.188$ of masks agree with the exact direction mask. Top-k consequently emits 4–8 numerical masks and a decoder scores $.967$.

Reconstructing the support from the theorem's exact direction $|w_1 - w_0|$ emits one mask and returns the $.623$ prior in every seed. We therefore exclude Cancer from the headline mechanism/certificate table and retain the raw float32 result only as a warning: exact-arithmetic architecture nulls need an explicit numeric realization (stable analytic direction, non-saturating loss, or a declared zero-row policy) before they become system guarantees. The LFW late cut independently realizes the proposition on the actual float32 wire with one mask and nonzero gradients.

F Verification and Claim Boundaries

F.1 Computational and Statistical Checks

The artifact tests the exact random-k identity, the constant-support endpoint, explicit inclusion of M in joint features, row-wise batch agreement through $B = 256$, the exact two-coordinate formulas, zero gradients, codebook overlap, Fano turning points, class-reweighted populations, selector endpoints, wire-bit counts, and separation of fixed-instance confidence from across-seed variation. The finite-channel suites were re-executed rather than inferred from paper tables: 100,000 cross-fit, 50,000 arbitrary-mixture, 50,000 generalized linkability, 50,000 conditional-transcript, 50,000 hard-code, 20,000 list-Fano, and 100,000 binary-tail cases have zero violations. The reviewer self-check independently recomputes all 25 fixed-audit Fano rows, all 25 finite-cross-fit penalties, five independent Diamonds rows, 15 coordinate-frequency certificates, 20 TPSL flip predictions, five exact-direction binary controls, the corrected mixture endpoint, and nine CIFAR-48 rows before hashing the exact outputs. All numerical tables consume released CSV files; the manifest records file sizes and SHA-256 digests.

F.2 Closest-Work Boundary

Work	Relevant object	Boundary relative to this paper
Label leakage/ExPLoit (Li et al., 2022; Kariyappa & Qureshi, 2023)	Split gradients/representations	Establishes norm, direction, and learned label attacks; our same-message norm, dense, and sign baselines isolate the additional support object.
Marvell (Li et al., 2022)	Optimized pre-selection gradient noise	Directly relevant value-and-support perturbation; our Gumbel selector is only a generic pre-selection-noise control and TPSL is the representative training-time defense. We do not label either as a Marvell reproduction.

TPSL/SecDT (Yang et al., 2022; Jiang et al., 2024)	Private split-learning transcript	Provides defense mechanisms/guarantees; our one-message baseline is narrower and our main results are leakage lower bounds.
OLIVE and sparse-update privacy (Kato et al., 2023; Vithana & Ulukus, 2023; 2024)	Sparse positions and values	Establishes that sensitive indices are not novel; different protocol and label-specific question.
Sami–Güler (Sami & Güler, 2024)	Coordinates of sparsified federated updates	Shows auxiliary coordinate information can reveal private user data and proposes coordinate-hiding sparse secure aggregation; its aggregated FL protocol is not split-learning label inference.
Private top-k (Silva Carvalho et al., 2020; Qiao et al., 2021; Gillenwater et al., 2022; Zhu & Wang, 2022)	Differentially private selection	Protects support under defined sensitivity; exact selected values and repeated split-learning messages require separate treatment.
Privacy–distortion (Kalantari et al., 2017; Liao et al., 2018)	General release mechanisms	General information/distortion formulations; no same-label cross-fit or sparse cut-gradient audit.
Qiu et al. (Qiu et al., 2023)	Split representations and gradients	Systematic leakage/mitigation evaluation; does not isolate sparse support or calibration-lineage transfer.
Potential-energy loss (Zheng et al., 2024)	Forward embeddings	Suppresses embedding label information; complementary to backward support/value analysis.
SLeak (Xu et al., 2026)	Smashed data and server model	Multi-target extraction with same-domain public data; broader targets and different substitute-client attack.
LLM split fine-tuning (Yao & Li, 2024)	Large-model intermediate states	Shows scale- and model-dependent privacy behavior, motivating our explicit non-generalization to language models.
This paper	Adaptive sparse cut gradients	Same-label mechanism bound plus conditional support/value audits, calibration, linkage, and wire semantics.

F.3 Claim Ledger and Explicit Non-Claims

Ethics and responsible use. This is a defensive audit of messages already visible to the declared protocol server. The architectural control uses `skimage.data.lfw_subset`, documented as 100 LFW faces and 100 background non-faces; these are human images, not a non-human benchmark. We use the copy supplied through the upstream package and require users to obtain it independently and follow the upstream scikit-image/LFW access and use terms. We redistribute no images, names, identities, per-example gradients, masks, or linkable transcripts. Released files contain code and aggregate/per-run metrics only. Documentation limits intended use to authorized auditing and warns against re-identification or attacks on non-consenting users. No new human data were collected and no attempt was made to identify individuals.

- **Claimed:** class-transferable support fidelity implies support information but does not explain the negative Diamonds cross-fit result; an auxiliary-frozen coordinate yields positive decoder-free fixed-audit certificates for class-conditional selection frequency; frozen held-out decoders yield empirical-mixture Fano certificates conditional on realized class counts; a real-large task exhibits measurable utility costs; target-checkpoint or same-lineage calibration enables strong attacks through the pathways in Table 2; support leakage remains far above prior through a separate 48-epoch compact CIFAR-10 schedule; TPSL support accuracy contains an analytic label-swap component plus a measured residual; repeated messages increase exploitability when rows are correctly linkable or false matches remain label-correlated; the exact-arithmetic binary linear-tail null identifies a support-only architecture boundary.
- **Not claimed:** support leakage itself is new; top-k always leaks; the cross-fit bound is tight or quantitatively explains the Fano values; row-wise linkable supports are the dominant deployment format; a zero list cer-

tificate proves privacy; decoders transfer across independent model seeds; one-message support DP with unprotected values is private training; or sparse compression is utility-free.

- **Outside current evidence:** CIFAR-100, BloodMNIST, ImageNet-scale or production vision, production-scale transformers or language models, prevalence of calibration pathways, composed transcript privacy, and direct matched-utility reproduction of full TPSL, Marvell, potential-energy loss, or SecDT.

F.4 Large Language Model Use Statement

Large language model tools were used for language editing, code assistance, experiment orchestration, and figure/layout refinement. The authors checked the theorem statements and proofs, experiment protocols, data partitions, executed outputs, tables, citations, and scientific claims and take full responsibility for the submission. No language model is treated as an author.

F.5 Reproducibility Checklist

- Seeds are $\{1, 7, 19, 31, 43\}$; model-training, auxiliary, and test partitions are disjoint.
- Preprocessing and model fitting use no auxiliary/test examples; decoder choices are frozen before test evaluation.
- CSV rows record context/checkpoint, dataset, seed, dimension, k, selector, task metrics, prior, support/joint metrics, omitted energy, bits, and certificate components.
- `primary_audit.py`, `sparsity_sweep.py`, `utility_stress.py`, `selector_sensitivity.py`, `mixture_frontier.py`, `calibration_audit.py`, `linkability_audit.py`, and `cut_null_audit.py` reproduce the original analyses. Reviewer-requested additions are reproduced by `reviewer_private_selector.py`, `reviewer_cifar_convergence.py`, `reviewer_diamonds_independent_crossfit.py`, `reviewer_tpsl_sparse_audit.py`, `reviewer_tpsl_flip_analysis.py`, `reviewer_dpi_frequency_certificate.py`, and `reviewer_binary_tail_precision.py`.
- `reviewer_artifact_selfcheck.py` recomputes all 25 fixed-audit Fano certificates, all 25 finite-cross-fit penalties, the five independent Diamonds certificates, all 15 coordinate-frequency certificates, 20 TPSL flip predictions, five exact binary-tail controls, the corrected mixture endpoint, the nine 48-epoch CIFAR rows, and SHA-256 hashes for theorem outputs and core scripts/CSVs.
- Figure scripts consume released CSVs and emit vector PDFs; computational falsification counts and zero-violation results appear in Appendix F.1. These tests supplement, not replace, proofs.