READ-OUTS ARE FILTERS, NOT FIREWALLS: MANIFOLD HEALTH SETS THE CEILING OF MEMBERSHIP INFERENCE PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Membership inference attacks (MIAs) exploit distinct channels of a trained model: representations, gradients, outputs, decision boundaries, and the per-sample training loss. Why do defenses that survive adaptive auditing on one dataset collapse on another? We attribute the surviving signal to a three-stage pipeline: *data fingerprintability* creates memorizable training samples; the representation encodes them; and the *read-out* decides how visibly the memorization surfaces in the observable scalar. One geometric property, the manifold health of the class-conditional distributions, splits datasets into three regimes and drives two degradation paths. In the *fingerprint regime* (scattered distributions: high intrinsic dimension, few samples per class), leakage resurfaces through any read-out as the shadow budget grows: on Location30 a density head read at 0.58 AUC with four shadows is read at 0.79 with eight. In the *low-signal regime* (mixed distributions), utility collapses faster than any defense can act: head-only DP-SGD on Texas100 falls from 61.9% to 12.6% accuracy. Within the *manifold regime* the read-out is decisive: linear logit heads carry a per-sample parameter imprint that LiRA amplifies (0.939 at 64 shadows vs. 0.832 with no defense), while a frozen-anchor density head suppresses the same channel to chance (0.5213 AUC at 94.9% accuracy on CIFAR-10; 0.4984 at 88.1% on PathMNIST) at zero ε cost. Read-outs are filters, not firewalls. Coarsening Location's features from 446 to 112 dimensions confirms the fingerprint mechanism causally: leakage drops ($0.885 \rightarrow 0.811$) as fingerprintability falls. We contribute a pre-deployment manifold audit (intrinsic dimension, class separation, samples per class) and a mechanism-matched, noise-matched, multi-seed protocol under which the three regimes (and the defense each one demands) become a measurement problem rather than trial and error.

1 INTRODUCTION

Membership inference attacks (MIAs) (Shokri et al., 2017) determine whether a given sample was used to train a model. As large models memorize more (Carlini et al., 2021; Feldman, 2020), MIA is the standard empirical yardstick for training-data leakage, and the likelihood-ratio attack (LiRA) (Carlini et al., 2022) is its strongest instantiation. Defenses fall into two camps: formal differential privacy (DP) (Abadi et al., 2016), which bounds leakage but pays a utility cost that grows with data scarcity (Bagdasaryan et al., 2019), and empirical structural defenses (adversarial regularization (Nasr et al., 2018), relaxation (Chen et al., 2022), distillation ensembles (Tang et al., 2022), high-entropy labels (Chen & Pattabiraman, 2024)), which preserve utility but are routinely broken by adaptive LiRA (Aerni et al., 2024). This paper asks a sharper question than "which defense works": *why does the same defense succeed on one dataset and collapse on another, and what decides it?*

One geometric property, two degradation paths. Our answer is a three-stage leakage pipeline: *data fingerprintability* creates memorizable training samples; the representation encodes them; and the *read-out* (the component that turns a representation into an observable scalar) decides how visibly the memorization surfaces. The pipeline's first stage is governed by one measurable geometric property: the *manifold health* of the class-conditional distributions, quantified before any model is trained from three attack-free statistics (intrinsic dimension, class separation, samples per class).

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.

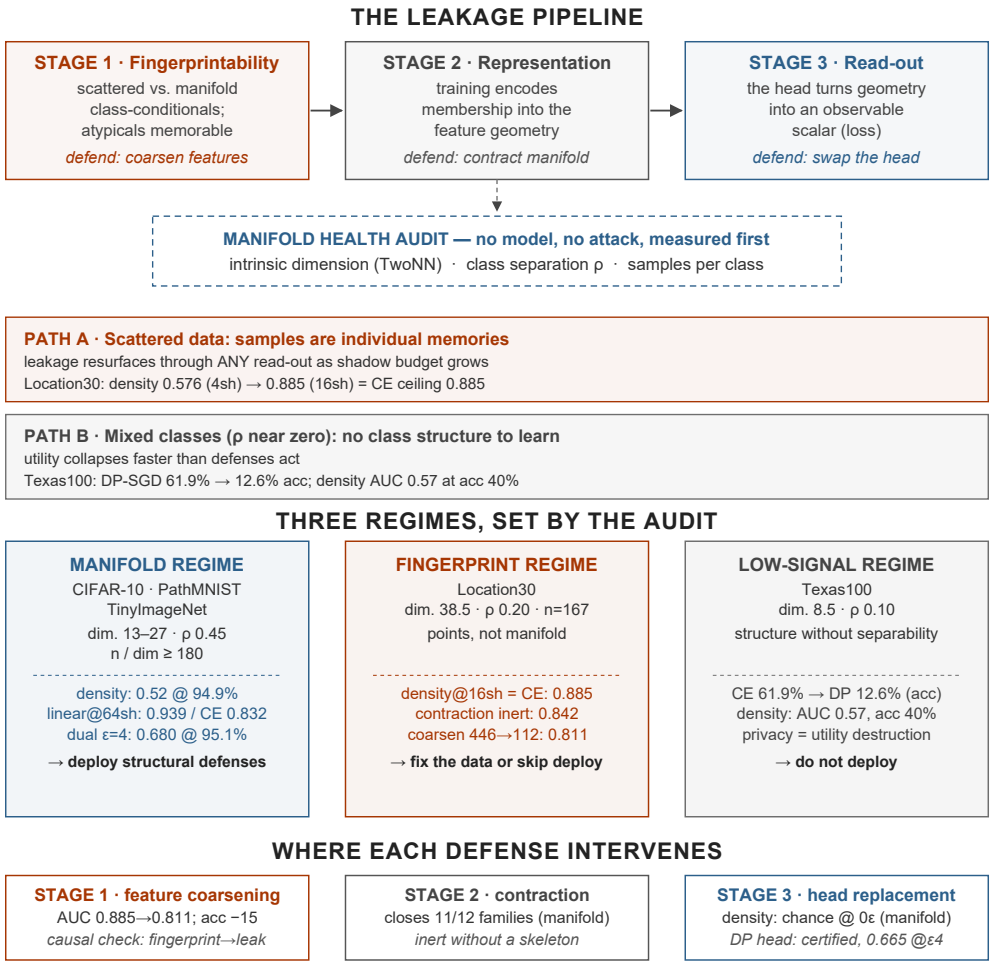


Figure 1: The leakage pipeline and its geometry. Leakage flows through three stages (fingerprintability, representation memorization, read-out visibility); one pre-attack-measurable property (manifold health) splits datasets into three regimes and activates two degradation paths: scattered data lets leakage resurface through *any* read-out as the shadow budget grows (Location30), mixed classes collapse utility faster than defenses act (Texas100). Each defense intervenes at a different stage; only the data-stage intervention moves the fingerprint-regime ceiling.

These statistics split our five evaluation datasets into three regimes, and each regime activates a different degradation path. In the *fingerprint regime* (Location30: intrinsic dimension ≈ 40 against 167 samples per class), samples are scattered points rather than manifold elements, so they are individually memorable; the memorization then resurfaces through *any* read-out as the adversary's shadow budget grows (Figure 3): a density head that reads 0.58 AUC under four shadows is read at 0.79 under eight. In the *low-signal regime* (Texas100: class separation near zero), the problem inverts: classes are so mixed that utility collapses faster than any defense can act, and head-only DP-SGD falls from 61.9% to 12.6% accuracy. Both paths share a root cause in the data geometry, and both are invisible to protocol-only fixes. A causal check confirms the mechanism: coarsening Location's features from 446 to 112 dimensions reduces fingerprintability and, with it, leakage ($0.885 \rightarrow 0.811$ AUC).

Read-outs are filters, not firewalls. Within the manifold regime (CIFAR-10, PathMNIST, TinyImageNet: intrinsic dimension 13–27, healthy class separation), the pipeline's third stage becomes decisive. Linear logit heads carry a per-sample parameter imprint ($\mathbf{W} \propto \sum_i \nabla_{\mathbf{W}} \ell_i$) that LiRA amplifies with shadow budget: with 64 shadows, every linear-head configuration we test is read out above its no-defense level (dual track 0.939 vs. CE-only 0.832), so tightening the representa-

tion cannot save a linear head. The opposite extreme is a kernel-density head whose anchors are synthetic, manifold-aware, and *frozen* at initialization; it reads out a distribution rather than an individual. Under the identical strict protocol LiRA falls to chance (0.5213 AUC at 94.9% accuracy on CIFAR-10; 0.4984 at 88.1% on PathMNIST), at zero ε cost. But a density head is a filter, not a firewall: it removes the linearly separable imprint while the memorization itself remains, and the fingerprint regime shows exactly where it fails. In manifold-regime data the filter amounts to privacy by architecture; elsewhere, the ceiling is set by the data.

Channel decomposition of MIA. We evaluate a 12-family attack matrix on feature-level classifiers: black-box loss/confidence, white-box gradients, label-only boundary distance (Choquette-Choo et al., 2021), reference-model (Ye et al., 2022), intermediate-feature, ensemble (Wang et al., 2025), and five geometry-aware attacks against our own defense's statistics. The pattern is consistent: under geometric manifold contraction (CRISP) (Anonymous, 2026), *eleven families stay at chance*, and the *single surviving family* is LiRA, which detects a residual signal in the training-loss distribution (AUC 0.826, 16 mechanism-matched shadows): a scalar the contraction, acting on the feature manifold, cannot touch. This decomposition localizes the leak in the pipeline and motivates the read-out analysis.

Dual-track defense. The complementary fix trains the classification head with DP-SGD: per-example clipping and Gaussian noise are concentrated exactly on the surviving channel, while the geometric track keeps operating on the feature manifold. The interaction is *dataset-dependent*: on CIFAR-10 the dual track matches head-only DP (0.680 vs 0.665 AUC); on PathMNIST it is better (0.587 vs 0.751). Head-only DP there is worse than no defense, which we attribute to DP-noise distortion of a less compact loss landscape.

Evaluation methodology as a contribution. We document three pitfalls that invalidate published-style numbers, as a reusable protocol: (i) *mechanism-matched shadows are necessary*: CE-only shadows evaluate DP-trained targets with systematically inverted scores (raw AUC < 0.5); the inversion is a calibration artifact an adversary exploits by flipping, not a defense; (ii) *noise-matched shadows are necessary for DP*: shadows at the same ε on 50% subsets receive a different noise multiplier, biasing the likelihood ratio; matching the multiplier (the standard of Nasr et al. (2023)) removes the bias; (iii) *single-seed claims are unstable*: across 10 seeds our headline moves in $[0.564, 0.783]$, so all results here are multi-seed with deviations. Under this protocol, four heuristic loss-channel defenses from the literature's lineage (MMD alignment, pseudo-nonmember injection, loss flooring, moment matching) fail: their apparent single-run success does not survive mechanism-matched auditing, consistent with Aerni et al. (2024).

Threat-model clarity. Finally, the loss channel *cannot be hidden*: cross-entropy loss is a deterministic function of the output probability, so any probability-returning API leaks it; hard-label APIs leave the boundary-distance channel; and mechanism secrecy is obscurity that adaptive adversaries defeat by score flipping. We evaluate under the strongest Kerckhoffs assumption (full mechanism knowledge); what survives is the mechanism itself, not its secrecy.

Scalable geometry at 200 classes. A class-sampled audit ($K{=}16$ classes receive the full topological audit; the rest a cheap closed-form prior) carries the dual track to TinyImageNet: 84.4% accuracy against 78.0% for head-only DP at LiRA AUC $0.730{\pm}0.002$. As classes grow, each class memorizes fewer atypical samples (Feldman, 2020), and the geometry prior shifts from privacy to utility.

Contributions. (1) A leakage pipeline (fingerprintability $\rightarrow$ representation memorization $\rightarrow$ read-out visibility) governed by one attack-free measurable property, whose three regimes explain where every defense we test succeeds or fails, with a coarsening intervention ($446 \rightarrow 112$ dims: $0.885 \rightarrow 0.811$) confirming the fingerprint mechanism causally. (2) Within the manifold regime, the read-out decides: linear heads amplify the per-sample imprint, while a frozen-anchor density head suppresses the channel to chance at zero ε cost. (3) A 12-family decomposition showing contraction suppresses all MIA families except the training loss: the channel the density head closes outright. (4) A strict mechanism-matched, noise-matched, multi-seed protocol under which four heuristic defenses fail, a threat-model argument that the loss channel cannot be hidden, and a class-sampled audit extending the dual track to 200 classes.

2 Related Work

Membership inference. Shokri et al. (Shokri et al., 2017) formalized shadow-model attacks; Carlini et al. (Carlini et al., 2022) introduced LiRA, whose train-only protocol we adopt; Choquette-Choo et al. (Choquette-Choo et al., 2021) showed decision boundaries leak through label-only queries; Ye et al. (Ye et al., 2022) unified attacks in a hypothesis-testing frame; Wang et al. (Wang et al., 2025) showed attacks exploit disjoint vulnerable subsets, motivating multi-family evaluation.

Defenses. DP-SGD (Abadi et al., 2016) is the formal gold standard; last-layer DP fine-tuning (De et al., 2022; Tramèr & Boneh, 2021) confines noise to the head, the setting we build on. Structural defenses (AdvReg (Nasr et al., 2018), RelaxLoss (Chen et al., 2022), SELENA (Tang et al., 2022), HAMP (Chen & Pattabiraman, 2024), MemGuard (Jia et al., 2019)) were shown by Aerni et al. (2024) to underperform tuned DP-SGD under adaptive LiRA; our four-defenses failure study reaches the same verdict with a mechanism-matched protocol. Our geometric track builds on CRISP (Anonymous, 2026); related geometric work includes Inf^{2}Guard (Noorbakhsh et al., 2024) and the manifold memorization hypothesis (Ross et al., 2025).

Read-out mechanisms and density-based classifiers. The privacy literature has largely treated the classifier head as a detail. Our decomposition shows it is not: the read-out mechanism determines whether the loss channel exists at all. Generative-model memorization studies (Ross et al., 2025) establish that geometric structure drives sample memory; we show the converse, that reading out through frozen synthetic anchors suppresses the membership signal that geometric structure otherwise enables. Alignment–uniformity analysis (Wang & Isola, 2020) gives the axes: linear heads align the decision boundary with the training set (recording membership), while density heads preserve uniformity of the read-out. Density-based classification is standard; our contribution is the mechanism analysis and the strict-protocol measurement of its privacy effect.

Auditing. Tight DP auditing (Nasr et al., 2023) establishes noise-matched shadow canaries as the standard, and Jayaraman and Evans (Jayaraman & Evans, 2019) showed that inference attacks on practical DP-SGD succeed far below the formal ε; we add the mechanism-matching requirement for defense evaluation. Appendix Table 11 positions existing defenses on the channel map: each addresses one or two channels, none targets the loss channel directly, and the dual track is, to our knowledge, the first to close the full spectrum explicitly.

3 Method

Design philosophy: relocate memory, do not erase it. Deep networks memorize atypical samples to generalize (Feldman, 2020); erasing the memory trace is therefore in tension with utility, which is exactly the cost DP pays. The dual-track design instead *relocates* memory so that it becomes indistinguishable to a membership oracle: the geometric track moves memorized samples into the dense core of their class manifold (feature space), and the DP head adds calibrated noise at the output so the scalar loss no longer discriminates membership (output space). These are two successive information bottlenecks. Under Kerckhoffs' principle the defense must not rely on hiding the mechanism; under Goodhart's law, an evaluation that can be gamed will be gamed: single-run AUC is optimized by heuristics that collapse when the evaluation tightens. Our protocol is the anti-Goodhart stance: mechanism-matched shadows (the adversary replicates the defense), noise-matched shadows (the DP audit standard of Nasr et al. (2023)), and multi-seed reporting (no favorable draws). The four heuristic loss-channel defenses we test fail precisely when these three conditions hold.

3.1 Leakage pipeline and manifold health

We model membership leakage as a three-stage pipeline. (i) *Fingerprintability*: samples whose features are far from any low-dimensional structure are individually memorable; the memorization hypothesis (Feldman, 2020) predicts atypical samples are stored. (ii) *Representation memorization*: training encodes membership into the feature geometry. (iii) *Read-out visibility*: the classifier converts geometry into scalars (logits, losses, probabilities) that an attacker tests. Stages (i)–(ii) are properties of the data and the featurizer; stage (iii) is a design choice. A defense can intervene at any stage: input coarsening and DP at stages (i)–(ii), contraction at (ii), read-out design at (iii). The intervention must match the stage that carries the signal.

Stage (i) is quantified *before any model is trained* by three attack-free statistics we call the *manifold health audit*: class-averaged intrinsic dimension $\bar{d}_{\mathrm{ID}}$ (TwoNN estimator (Facco et al., 2017) on raw features), class-separation ratio ρ (nearest-centroid distance over mean intra-class pairwise distance), and samples per class n_c. Healthy manifolds show moderate $\bar{d}_{\mathrm{ID}}$, $\rho \geq 0.4$, and $n_c \gg \bar{d}_{\mathrm{ID}}$; Table 1 reports the audit, whose statistics alone predict where every defense we test succeeds or fails (Section 4.1).

3.2 Geometric manifold contraction (track 1)

We briefly restate the geometric track; full details are in Appendix A. The feature extractor is a frozen ViT-B/16 (Dosovitskiy et al., 2021); a contraction network Φ_ϕ maps 768-d features $\mathbf{x}$ to a d-dimensional representation $\mathbf{z} = \Phi_\phi(\mathbf{x})$, and a linear head Cls maps $\mathbf{z}$ to logits. A single-shot *persistent homology audit* (PMG) estimates per-class geometry: anchors $\mathcal{A}_c$, tangent bases $\mathbf{U}_c$, intrinsic dimension $\hat{d}_c$, hole counts, and hull volumes. For datasets with many classes the audit uses a class-sampled variant that runs the full topology pipeline on K representative classes and a default geometry on the rest (Section 4.7); at small class counts every class is audited fully. During warmup the model trains with cross-entropy only; afterwards the PMG geometry is locked and the loss becomes

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \gamma(\mathcal{L}_{\mathrm{CE}}) \cdot \mathcal{L}_{\mathrm{geom}}, \tag{1}$$

where γ is an adaptive temperature from the CE loss and $\mathcal{L}_{\mathrm{geom}}$ combines (i) anchor attraction $\|\mathbf{z} - \mathbf{a}_y\|_2^2$ projected onto the tangent space, (ii) KDE-density elevation $-\log p_y(\mathbf{z})$ at the class core, and (iii) topology-adaptive scaling $\lambda_c \propto (1 + 2\nu_c)$ with ν_c the per-class hole count, which contracts non-convex classes harder. Volume conservation (residual connections and a non-degenerate $\hat{d}_c$) prevents collapse; on CIFAR-10 the contracted manifold reaches 95.4% accuracy with all holes closed.

3.3 Why geometry cannot close the loss channel

The geometric loss acts on $\mathbf{z}$ and is class-aggregated; the scalar cross-entropy $\ell = -\log p_y$ is determined by the head on top of $\mathbf{z}$. Contraction moves members and non-members onto the same dense core, erasing the representation-level signatures that black-box, white-box, geometry-aware and boundary attacks exploit. But the fact that the head was *optimized on member samples* remains encoded in the per-sample loss: members have systematically lower loss, and LiRA's per-sample Gaussian likelihood ratio is precisely the statistic that detects it. Empirically, geometry alone does not reduce LiRA (AUC 0.826 vs 0.779 for a plain CE baseline), confirming the channel decomposition at the mechanism level.

3.4 Dual track: DP-SGD on the head (track 2)

The head is a single linear layer $\theta \in \mathbb{R}^{d \times C}$. We train it with DP-SGD (Abadi et al., 2016): per-example gradient clipping with norm $C = 1$ and Gaussian noise σ fixed by the Rényi accountant for target (ε, δ) ($\delta = 10^{-5}$, $\varepsilon \in \{1, 2, 4, 8\}$):

$$\theta_{t+1} = \theta_t - \eta\big(\tfrac{1}{B}\textstyle\sum_{i \in \mathcal{B}} \mathrm{clip}_C(\nabla_\theta \ell_i) + \mathcal{N}(0, \sigma^2 C^2 I)\big), \tag{2}$$

while the geometric track continues without noise. The DP head therefore provides the formal guarantee of DP-SGD *on the output layer* (the only layer the MIA signal survives), at negligible utility cost because $d \times C$ parameters carry little information compared with the contraction network. The geometric prior simultaneously structures the features so the noisy head fits with less gradient signal.

3.5 Density read-out head (the filter)

The density head replaces the linear read-out with a frozen kernel-density classifier. At initialization (epoch 0), we run the manifold-aware anchor construction on the latent space: per class, a GaussianKDE with bandwidth h fits the latent samples; manifold-aware sampling draws anchors from a mixture of core samples and tangent-space perturbations (3% of the class, $K{=}30$ anchors, diversity weight 0.9); the per-class intrinsic dimension is estimated and anchors are *frozen* for the remainder of training. The classifier scores a latent $\mathbf{z}$ by per-class kernel density over its frozen anchors,

Dataset	n	n_c	$\bar{d}_{\mathrm{ID}}$	ρ	Regime
CIFAR-10	50k	5000	23.8	0.45	Manifold
PathMNIST	90k	10000	13.0	0.45	Manifold
TinyImageNet	100k	500	27.4	0.46	Manifold
Location30	5k	167	38.5	0.20	Fingerprint
Texas100	67k	673	8.5	0.10	Low-signal

Table 1: Manifold health audit (raw features, no training, no attack): class-averaged TwoNN intrinsic dimension $\bar{d}_{\mathrm{ID}}$, class-separation ratio ρ (nearest-centroid distance / mean intra-class distance), samples per class n_c. The statistics alone predict where every defense we test succeeds or fails.

$\hat{p}_c(\mathbf{z}) \propto \sum_{a \in \mathcal{A}_c} \exp(-\|\mathbf{z} - a\|^2/2h^2)$, and predicts $\arg\max_c \hat{p}_c$. Training uses a manifold density loss (α=50, β=1.0) that shapes the encoder toward anchor neighborhoods; the anchors themselves receive no gradient.

Three structural properties make the read-out a filter rather than an amplifier: (i) *no per-sample parameter imprint*: the frozen anchors carry no per-sample training signal, unlike a linear head whose weights accumulate per-sample gradients; (ii) *Lipschitz flattening*: removing one member shifts the density score by $O(1/n)$, whereas a linear head's loss shifts by $O(1)$; (iii) *synthetic frozen anchors*: the read-out is a function of the distribution, not of any individual training point. Within the manifold regime this suppresses the loss channel to chance at zero ε cost; in the fingerprint regime (Section 4.1) the same construction fails exactly as the pipeline predicts, because memorization lives in the scattered samples themselves.

3.6 Evaluation protocol

Protocol P (fixed train-only). 80% of training samples are members, 20% non-members; 20k queries; 5-fold calibration; metrics are AUC, calibrated AUC $\max(\mathrm{AUC}, 1 - \mathrm{AUC})$, and TPR@$0.1\%$FPR.

Mechanism-matched shadows. Shadow models replicate the full target mechanism; shadows trained without it produce systematically inverted LiRA scores. The inversion is a calibration artifact an adversary exploits by flipping, not a defense.

Noise-matched shadows. For DP targets, shadows are trained at the target's *noise multiplier* rather than at the same ε: 50%-subset shadows at equal ε receive a larger multiplier, biasing the likelihood ratio; matching follows the auditing standard of Nasr et al. (2023).

Multi-seed reporting. Every headline number is a mean $\pm$ std over seeds; single-seed numbers appear only in ablations and are marked.

4 Experiments

Setup. Datasets: CIFAR-10 (50k, 10 classes), PathMNIST (90k, 9 classes), TinyImageNet (100k, 200 classes), Texas (67k, 100 classes, 6169-d sparse), Location (5k, 30 classes, 446-d sparse), all with ViT-B/16 features where applicable. Model: contraction network $768 \to 512 \to 256 \to 64$, linear head. Hardware: single RTX-class GPU per run; each dual-track run (25 epochs + 4 matched shadows) takes $\sim$40–55 minutes. Full configurations and the 12-family attack definitions are in Appendix B.

4.1 Manifold health regimes

Table 1 and Figure 2 report the pre-deployment audit (Section 3.1). Three regimes emerge from the statistics alone.

Fingerprint regime (Location30). Intrinsic dimension ≈ 40 against 167 samples per class: samples are scattered points, not manifold elements, and memorization is individual. The prediction is that *no read-out filters this*: the leakage ceiling is set by the data. Figure 3 (curves and numbers) confirms it exactly. The density head that reads at chance on CIFAR-10 grows with the shadow budget ($0.576 \to$

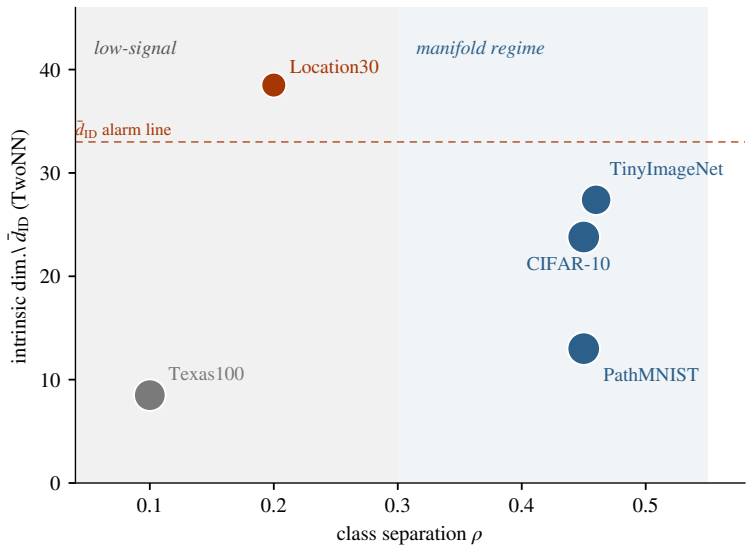


Figure 2: The audit's two informative axes: class separation ρ (x) against class-averaged intrinsic dimension (y); point area encodes samples per class. The five datasets separate into three regimes without any training or attack.

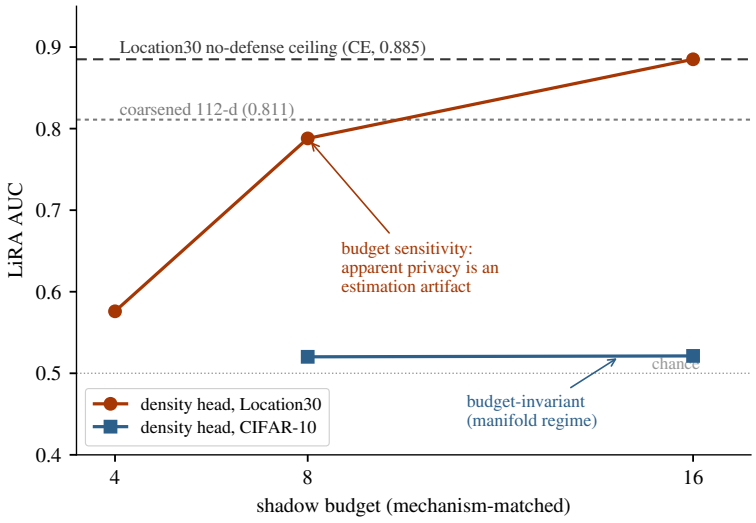


Figure 3: Shadow-budget sensitivity separates the regimes. On Location30 the density head's apparent privacy at small budgets is an estimation artifact; as shadows grow, LiRA reads it up to the no-defense ceiling, while on CIFAR-10 the same architecture stays at chance. The coarsened-feature baseline shows the data-stage fix lowers the ceiling itself.

$0.788 \rightarrow 0.885$ at 4/8/16 shadows) and lands *precisely at the CE baseline level* (0.885 at 54.5% accuracy, baseline 0.885 at 65.6%): under sufficient auditing, the density head offers no privacy beyond what the data already grants. The same evidence shows the causal check: coarsening the features from 446 to 112 dimensions (max-pooling every 4 coordinates) removes fingerprintability and, with it, leakage ($0.885 \rightarrow 0.811$), at the cost of 15 points of accuracy. Fingerprint removal is a data-stage intervention with a data-stage price. Geometric contraction on Location likewise cannot beat the ceiling: the track that closes 11 of 12 families on CIFAR-10 moves Location's LiRA by 0.045 at a cost of 2.7 accuracy points. On scattered data there is no manifold skeleton to contract.

Table 2: Noise-matched mechanism-matched LiRA AUC (mean ± std over seeds). v1 = geometric track alone. TinyImageNet (2 seeds) uses the class-sampled audit (K=16).

Method	Seeds	CIFAR-10	PathMNIST	TinyImgNet
CE baseline (no defense)	1	0.779	0.725	0.891
v1 geometric (no DP)	1	0.826	—	—
Head-only DP (ε=4)	4	0.665 ± 0.029	0.751 ± 0.005	0.615 ± 0.000
Dual track (ε=4)	10/4	$\mathbf{0.680 \pm 0.083}$	$\mathbf{0.587 \pm 0.105}$	0.730 ± 0.002

Low-signal regime (Texas100). With $\rho = 0.10$ the classes are so mixed that the problem inverts: the density head can hold LiRA near chance (0.566 at 8 shadows) but only by collapsing accuracy to 40.1% (vs. 61.9% for CE), and head-only DP-SGD collapses harder still (12.6%). When classes are not separable, there is no utility left to defend. The correct decision is to not deploy rather than to pick a defense.

Manifold regime (CIFAR-10, PathMNIST, TinyImageNet). Healthy statistics: moderate intrinsic dimension, $\rho \approx 0.45$, samples per class exceeding the intrinsic dimension by two orders of magnitude. In this regime alone the defensive machinery of the remaining sections operates as designed: channel decomposition, read-out filtering, and DP-augmented dual tracks all hold under the strict protocol.

4.2 Channel decomposition: geometry suppresses 11 of 12 families

The 12-family matrix (Appendix Table 8) shows the same structure on CIFAR-10 under the geometric track (v1): *eleven families at chance* (calibrated AUC ≤ 0.527) and a *single survivor*: adaptive LiRA at 0.826 (16 mechanism-matched shadows), the one statistic that depends on the scalar training loss rather than the feature geometry. Contraction collapses members and non-members onto the same dense class core (t-SNE in Appendix Figure 4), erasing the representation-level signatures; the loss gap survives because it is computed by the head, not stored in the manifold. This decomposition localizes the leak at stage 3 of the pipeline (Section 3.1) and motivates the read-out analysis below.

4.3 Dual-track results

Table 2 reports noise-matched mechanism-matched LiRA AUC. **(1) The DP head carries the loss channel on CIFAR-10**: dual track 0.680 ± 0.083 (10 seeds) is statistically equivalent to head-only DP 0.665 ± 0.029 (4 seeds). **(2) Geometry–DP synergy is dataset-dependent**: on PathMNIST the dual track attains 0.587 ± 0.105 vs 0.751 ± 0.005 for head-only DP (a 0.16 reduction), and head-only DP is worse than the no-defense CE baseline (0.725), which we attribute to DP-noise distorting a less compact loss landscape. **(3) The variance is intrinsic**: across 10 seeds the CIFAR dual-track AUC spans $[0.564, 0.783]$, but the mean is stable to 0.01 between 5- and 10-seed averages. **(4) At 200 classes the frontier is a trade-off** (Section 4.7): head-only DP reaches the lowest AUC (0.615) but pays a 6-point accuracy cost (78.0%), while the dual track sits at 0.730 ± 0.002 with 84.4%. The class-sampled audit makes the trade-off explicit (Appendix Figure 5).

4.4 Where the geometric track matters beyond the loss channel

Even where the loss channel is DP-dominated, the geometric track adds three guarantees: **(i) attack-spectrum coverage**: head-only DP leaves representation-level families exposed whenever features are released or transferred; the geometric track suppresses them to ≤ 0.527 (Appendix Table 8); **(ii) utility recovery**: at equal ε=4, the dual track reaches 95.1% accuracy vs 94.2% for head-only DP on CIFAR-10; **(iii) interpretable auditing**: the PMG outputs (δ, $\kappa_{\max}$, $\hat{d}_c$) provide a deployment-time check that parameter-level defenses lack.

4.5 Heuristic loss-channel defenses fail under the protocol

We evaluate four heuristics in the loss-channel lineage (MMD alignment with pseudo-nonmembers (v2), loss flooring with threshold τ (v3), and first/second-order moment matching (v4)) under

mechanism-matched shadows. All fail: v2 attains 0.926 (worse than no defense), v3's best configuration 0.893 (its single-run 0.793 does not reproduce), and v4's 3-seed mean is 0.797 with single-seed spread $[0.727, 0.838]$. This reproduces the verdict of Aerni et al. (2024) with a different protocol; full tables are in Appendix D.

4.6 Ablations: DP strength, geometric scale, and the synergy mechanism

The full ε/scale sweep (Appendix Table 9) shows two patterns: DP strength does not monotonically improve privacy (ε=1 gives 0.674 vs 0.680 at ε=4; extreme noise distorts the loss distribution until LiRA's Gaussian fit misbehaves), and the geometric scale is U-shaped ($\{0.5, 1.0, 2.0, 3.0\} \rightarrow \{0.676, 0.704, \mathbf{0.680}, 0.690\}$). Over-contraction degrades the feature structure the DP head relies on. The moderate-ε, scale-2.0 optimum is where the geometry prior is strongest without collapse.

The dataset dependence of the synergy has a mechanical account (Appendix Table 10): PathMNIST's manifold is flatter and more compact than CIFAR-10's ($\kappa_{\max} = 0.003$ vs 0.0067; $\bar{d}_c = 7.0$ vs 9.1), so contraction closes it nearly perfectly and the DP head fits on a well-structured representation, shrinking the loss gap LiRA detects; without the prior, PathMNIST's loss landscape is so poorly self-regularized that DP noise alone is itself membership-signaling (0.751 vs the 0.725 baseline). This makes a falsifiable prediction: compact, low-curvature manifolds should show the largest synergy. The PMG audit provides the fingerprint to test it before deployment.

4.7 Scalability: class-sampled audit at 200 classes

The geometry audit of Section 3 runs a topological analysis per class; its cost grows linearly in the class count, confining prior evaluation to ≤ 10 classes. A *class-sampled audit* removes this limit: the full topology pipeline runs on a representative subset K of classes, the rest receive a default geometry (the anchor loop, cheap and per-class, still runs on every class). On TinyImageNet (K=16, Table 2 last column) the audit completes within a single epoch, and the dual track keeps 84.4% accuracy (vs. 83.9% CE, 78.0% head-only DP). The geometric track's utility contribution survives the class explosion intact.

The 200-class regime completes the picture: the un-audited classes receive only the default contraction, so LiRA sits between the CE baseline and the small-class dual track, but the geometric regularization keeps the DP head within one point of the no-defense baseline while head-only DP loses 6 points. Consistent with Feldman (2020), as classes grow, each class memorizes fewer atypical samples, and the geometry prior shifts from privacy to utility.

5 Conclusion

We traced membership leakage through a three-stage pipeline (fingerprintability, representation memorization, read-out visibility) governed by one pre-attack-measurable property: the manifold health of the class-conditional distributions. In the *manifold regime*, the read-out is decisive: linear heads amplify the per-sample imprint, while a frozen-anchor density head suppresses the loss channel to chance at zero ε (0.5213 AUC at 94.9% on CIFAR-10), and the DP-augmented dual track adds the certified complement under mechanism-matched, noise-matched, multi-seed auditing that exposes four heuristic defenses as artifacts. In the *fingerprint regime*, the ceiling is set by the data: the same density head is read back to the no-defense level, contraction is inert, and only a data-stage intervention moves it. In the *low-signal regime*, there is nothing to defend. The class-sampled audit extends the machinery to 200 classes, where geometry shifts from privacy to utility preservation (84.4% vs. 78.0% at equal ε). Defense selection becomes a measurement problem: run the audit first, then pick the stage to intervene.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM CCS*, 2016.

Michael Aerni, Jie Zhang, and Florian Tramèr. Evaluations of machine learning privacy defenses are misleading. *ACM CCS*, 2024.

Anonymous. Crisp: Compact representation for intrinsic structural privacy. In *Under review*, 2026.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *NeurIPS*, 2019.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security*, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (S&P)*, 2022.

Dingfan Chen, Ning Yu, and Mario Fritz. Relaxloss: Defending membership inference attacks without losing utility. In *ICLR*, 2022.

Zitao Chen and Karthik Pattabiraman. Overconfidence is a dangerous thing: Mitigating membership inference attacks by enforcing less confident prediction. In *NDSS*, 2024.

Christopher A. Choquette-Choo, Florian Tramèr, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *ICML*, 2021.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Elena Facco, Maria d'Errico, Alessandro Rodriguez, and Antonio Laio. Estimating the intrinsic dimension of hilbert and banach spaces. *Physical Review E*, 95(4):042208, 2017.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *STOC*, 2020.

Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *USENIX Security Symposium (USENIX Security 19)*, pp. 1895–1912, 2019.

Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *ACM CCS*, 2019.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *ACM CCS*, 2018.

Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In *USENIX Security*, 2023.

Sayedeh Leila Noorbakhsh, Binghui Zhang, Yuan Hong, and Binghui Wang. Inf2guard: An information-theoretic framework for learning privacy-preserving representations against inference attacks. In *USENIX Security*, 2024.

Brendan Leigh Ross, Hamidreza Kamkari, Tongzi Wu, Rasa Hosseinzadeh, Zhaoyan Liu, George Stein, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. A geometric framework for understanding memorization in generative models. In *ICLR*, 2025.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P)*, 2017.

Xinyu Tang, Saeed Mahloujifar, Liwei Song, Virat Shejwalkar, Milad Nasr, Amir Houmansadr, and Prateek Mittal. Mitigating membership inference attacks by self-distillation through a novel ensemble architecture. In *USENIX Security Symposium*, 2022.

Marlon Tobaben, Hibiki Ito, Joonas Jälkö, Yuan He, and Antti Honkela. Impact of dataset properties on membership inference vulnerability of deep transfer learning. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations (ICLR)*, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, 2020.

Zhiqi Wang, Chengyu Zhang, Yuetian Chen, Nathalie Baracaldo, Swanand Kadhe, and Lei Yu. Membership inference attacks as privacy tools: Reliability, disparity and ensemble. In *ACM CCS*, 2025.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *ACM CCS*, 2022.

A Geometric Contraction Details

The contraction network Φ_ϕ is a ResMLP: $\Phi_\phi(\mathbf{x}) = \mathrm{Proj}_d \circ \mathrm{ResBlock} \circ \mathrm{Proj}_{D_1}(\mathbf{x})$ with $D_1 = 512$, hidden 256, and output $d = 64$ (CIFAR-10, PathMNIST) or $d = 256$ (Texas, Location, TinyImageNet). The single-shot PMG audit estimates per-class anchors (3% density-sampled), tangent bases via local PCA, intrinsic dimension via a tight local estimator, H_1 hole counts via Ripser, and convex-hull volumes. The geometric loss is

$$\mathcal{L}_{\mathrm{geom}} = \sum_c \Big[\alpha_c \, \mathrm{dist}_{\mathrm{anchor}}(\mathbf{z}, \mathcal{A}_c) + \beta_c \, (-\log p_c(\mathbf{z})) + \lambda_c \, \mathrm{contraction}_c(\mathbf{z}) \Big], \tag{3}$$

with topology-adaptive coefficients $\lambda_c \propto (1 + 2\nu_c)$ where ν_c is the per-class hole count; $\gamma(\mathcal{L}_{\mathrm{CE}})$ is a temperature that anneals the geometric term as CE stabilizes. Training: 25 epochs (4 warmup) for image data, 35–50 for tabular, Adam with 10^{-5} weight decay and StepLR.

B Experimental Setup and Datasets

Table 3: Datasets. All image datasets use ViT-B/16 [CLS] features (768-d).

Dataset	Samples	Classes	Feature
CIFAR-10	50,000	10	ViT-B/16 768-d
PathMNIST	89,996	9	ViT-B/16 768-d
TinyImageNet	100,000	200	ViT-B/16 768-d
Texas	67,330	100	sparse 6169-d
Location	5,010	30	sparse 446-d

Attack families (12). (1) Black-box loss/confidence/entropy with logistic-regression classifier over 128 shadows; (2) white-box gradient norms (head, feature, network) plus geometric statistics; (3) adaptive LiRA (online, per-sample Gaussian LR); (4) label-only boundary distance (PGD ℓ_∞/ℓ_2 perturbation to flip the label); (5) reference-model attack (population likelihood under 128 reference models); (6) gradient-based white-box (augmented features); (7)–(10) geometry-aware attacks on our own statistics: anchor-distance $\min_{\mathbf{a}\in\mathcal{A}_y} \|\mathbf{z} - \mathbf{a}\|_2$, KDE-density $-\log p_y(\mathbf{z})$, tangent-residual $\|(\mathbf{I} - \mathbf{U}_y\mathbf{U}_y^\top)(\mathbf{z} - \mathbf{a}_y)\|_2$, contraction-magnitude; (11) intermediate-feature attack (penultimate features with the standard loss/confidence score); (12) ensemble MIA (multi-instance stability + multi-attack union at 10^{-6} FPR).

Protocol. Train-only split (80/20), 20k queries (10k members + 10k non-members), 5-fold calibration, metrics AUC / calibrated AUC / TPR@0.1%FPR. Shadows: 16 for LiRA/dual-track rows (mechanism-matched), 128 for classifier-based families; DP shadows are noise-matched to the target's noise multiplier.

C Multi-Seed Tables

Table 4: CIFAR-10 dual track (ε=4, geometric scale 2.0), noise-matched mechanism-matched LiRA AUC, 10 seeds.

Seed	0	1	2	3	4	5	6	7	8	9
AUC	0.583	0.719	0.564	0.761	0.737	0.714	0.783	0.567	0.646	0.722

Mean 0.680, std 0.083, range $[0.564, 0.783]$.

Table 5: Head-only DP (ε=4), noise-matched, 4 seeds.

Seed	0	1	2	3
CIFAR-10 AUC	0.699	0.631	0.655	0.675
PathMNIST AUC	0.757	0.747	0.747	0.753

CIFAR-10 mean 0.665 ± 0.029; PathMNIST mean 0.751 ± 0.005.

Table 6: PathMNIST dual track (ε=4, geometric scale 1.0), 4 seeds.

Seed	0	1	2	3
AUC	0.543	0.743	0.510	0.552

Mean 0.587 ± 0.105.

D Heuristic Defenses Fail Under the Protocol

Interpretation: v2's MMD term sharpens decision boundaries and amplifies the loss gap; v3's clamping narrows the member-loss variance, which LiRA's Gaussian fit exploits (consistent with the moment-matching collapse at strong σ); v4's mean-alignment is ineffective because LiRA uses the full per-sample distribution shape, not its first moments. All four operate on the loss channel without making it membership-independent: the property only DP provides.

E Evaluation Pitfalls We Documented

(1) Train/val split mismatch. If the attack's member/non-member partition is drawn independently (e.g., numpy permutation) from the target's training split (torch random split), $\sim$20% of "member" queries are actually unseen and $\sim$80% of "non-member" queries were trained on, contaminating every AUC. We pass the target's exact split into the attack suite and verify with a 21-check protocol smoke test (train/val completeness, disjointness, member-query containment, shadow in-ratio ≈ 0.5).

(2) CE-shadow inversion against DP targets. CE-only shadows evaluate DP-trained targets with systematically inverted LiRA scores (raw AUC 0.16–0.40); the inversion is a loss-calibration artifact (DP noise shifts the whole loss scale), exploited by flipping. We always report raw + calibrated AUC and treat mechanism-matched shadows as the primary evidence.

(3) Equal-ε shadow bias. 50%-subset shadows at equal ε receive a larger noise multiplier than the 80% target, biasing the likelihood ratio (visible at ε=2, AUC 0.778 vs the noise-matched 0.661). We match the noise multiplier following Nasr et al. (2023).

(4) Single-seed instability. The CIFAR dual-track AUC spans $[0.564, 0.783]$ over 10 seeds; any single-seed claim (including our own earlier 0.536) is not a statement about the method. All headline numbers are multi-seed.

Table 7: Loss-channel heuristics under mechanism-matched shadows (CIFAR-10). v2 = MMD alignment + pseudo-nonmembers; v3 = loss flooring (threshold τ); v4 = moment matching (mean μ, variance σ). Single-seed numbers that do not reproduce are marked †.

Defense	Config	LiRA AUC	Seeds
No defense (CE)	—	0.779	1
v1 geometric	—	0.826	1
v2 MMD + pseudo-NM	λ_{mmd}	0.926	1
v3 loss flooring	τ=1.0	0.893 (0.793†)	1
v3 loss flooring	τ=1.2	0.923	1
v4 moment matching	μ=1, σ=0.5	0.797	3 (spread $[0.727, 0.838]$)
v4 moment matching	μ=2, σ=1	0.920	1
Dual track (ours)	ε=4	$\mathbf{0.680 \pm 0.083}$	10

F Applicability Boundary

Class count. The per-class persistent-homology audit does not terminate at $\geq$ 30 classes within a practical budget (Location, 30 classes: 44 CPU cores, 45 minutes, no convergence). The class-sampled audit of Section 4.7 removes this limit: with K=16 of 200 classes receiving the full topology pipeline and the rest a default geometry, TinyImageNet (100k samples) completes the audit within a single epoch and the dual track reaches 0.730 ± 0.002 LiRA AUC (2 seeds) at 84.4% accuracy, versus 0.615 and 78.0% for head-only DP at equal ε=4 (Section 4.7). Texas (100 classes) still collapses in utility for any head-only DP configuration (accuracy $61.9\% \rightarrow 12.6\%$), so the class-sampled track cannot rescue datasets whose no-defense accuracy is already below the DP-utility threshold.

Accuracy threshold. Head-only DP preserves utility only when the no-defense accuracy is high ($\gtrsim 90\%$): Location CE accuracy $65.6\% \rightarrow 31.1\%$; Texas $61.9\% \rightarrow 12.6\%$. CIFAR-10 (95.4%) and PathMNIST (96.4%) are in the useful regime.

Per-class sample count. Consistent with Tobaben et al. (2025), leakage follows samples-per-class rather than class count: TinyImageNet (200 classes, $\sim$500/class) CE LiRA AUC $0.891 >$ CIFAR-10 CE (0.779).

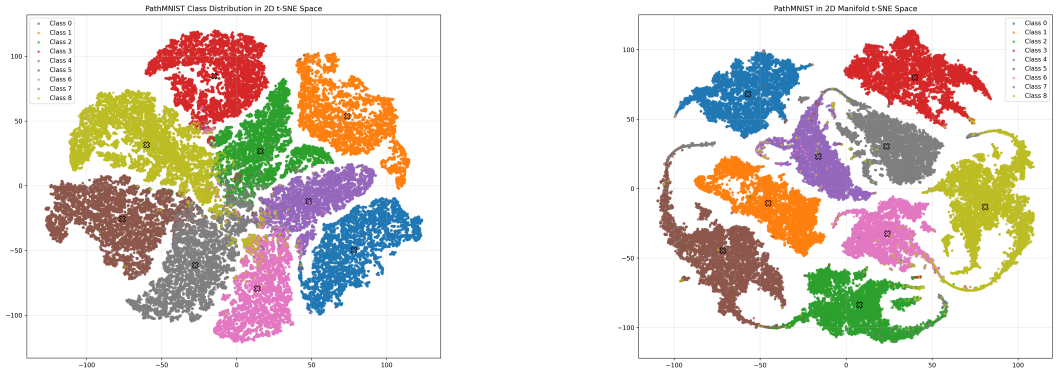
G Computational Cost

Single dual-track run (25 epochs + 4 matched shadows) on one RTX-class GPU: $\sim$40–55 minutes for CIFAR-10; the PMG audit is 92.6s per run on CIFAR-10 and scales super-linearly with per-class sample count (PathMNIST $\sim$803s). The DP head adds negligible overhead (linear layer, per-example clipping on $d \times C$ parameters).

H Moved Evidence: Full Tables and Figures

Table 8: Twelve-family attack matrix on CIFAR-10, geometric track only (v1), train-only protocol, mechanism-matched shadows. Cal. AUC $= \max(\text{AUC}, 1 - \text{AUC})$.

Attack family	**AUC**	**Cal. AUC**	**TPR@0.1%**
Black-box (loss/confidence)	0.512	0.512	0.0015
White-box (gradient + geometric)	0.514	0.514	0.0009
Adaptive LiRA	**0.826**	**0.826**	**0.0878**
Label-only (boundary distance)	0.491	0.509	0.0001
Reference-model	0.494	0.506	0.0000
Gradient-based white-box	0.514	0.514	0.0009
Anchor-distance	0.484	0.516	0.0001
KDE-density	0.485	0.515	0.0000
Tangent-residual	0.483	0.517	0.0000
Contraction-magnitude	0.497	0.503	0.0001
Intermediate-feature	0.512	0.512	0.0016
Ensemble MIA	0.527	0.527	0.5699



(a) Before contraction

(b) After (holes closed)

Figure 4: t-SNE of the PathMNIST feature manifold before and after geometric contraction: the manifold collapses onto dense cores, erasing the representation-level signatures that 11 of 12 attack families exploit; LiRA survives because it depends on the scalar output, not the manifold.

Table 9: Ablation over DP strength (ε) and geometric scale (CIFAR-10, noise-matched). † single seed; bold is the multi-seed headline.

Config	**Noise mult.**	**LiRA AUC**
ε=1, g1.0†	2.29	0.674
ε=2, g1.0†	1.35	0.661
ε=4, g0.5†	0.92	0.676
ε=4, g1.0†	0.92	0.704
ε=4, **g2.0**	0.92	$\mathbf{0.680 \pm 0.083}$
ε=4, g3.0†	0.92	0.690
ε=8, g2.0†	0.70	0.773
Head-only DP, ε=4	0.92	0.665 ± 0.029

Table 10: PMG geometry fingerprints and the geometry–DP synergy. δ: class separation margin; $\kappa_{\max}$: maximum PCA curvature; $\bar{d}_c$: mean intrinsic dimension. Δ = dual-track AUC minus head-only DP AUC (noise-matched).

Dataset	δ	$\kappa_{\max}$	$\bar{d}_c$	**Dual track**	Δ vs DP
CIFAR-10	11.97	0.0067	9.1	0.680 ± 0.083	-0.015
PathMNIST	10.84	0.0030	7.0	0.587 ± 0.105	$\mathbf{-0.164}$

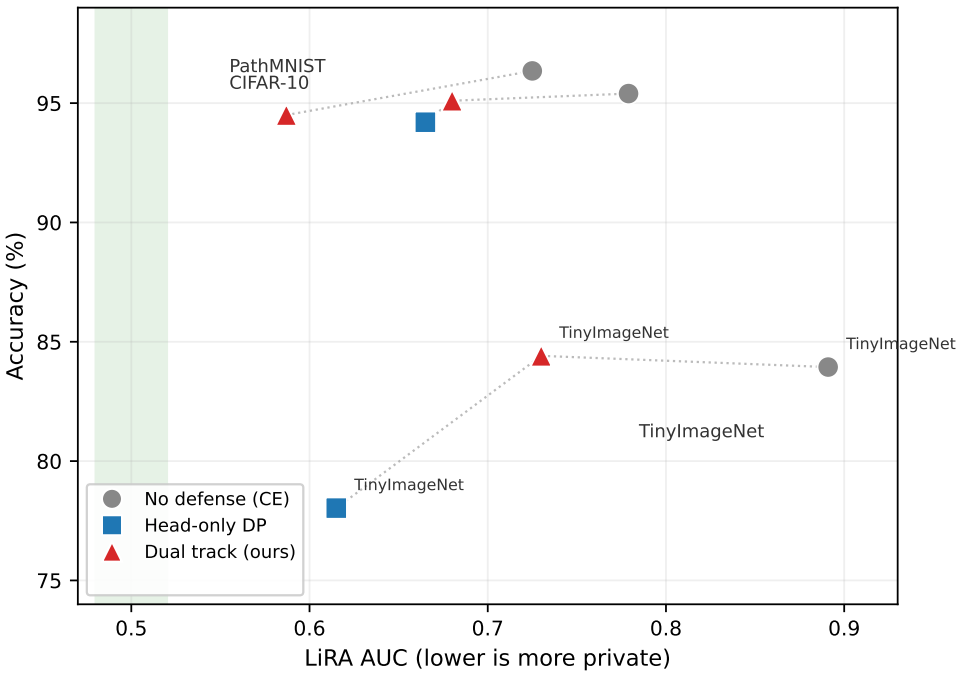


Figure 5: Privacy–utility frontier (accuracy vs LiRA AUC, lower is better). The dual-track defense sits at the frontier: on CIFAR-10 it matches head-only DP on the loss channel while keeping 95.1% accuracy; on PathMNIST it dominates (star). On TinyImageNet the frontier is a trade-off: head-only DP reaches the lowest AUC (0.615) but loses 6 points of accuracy, while the dual track keeps 84.4% accuracy at 0.730. The green band marks the random region.

Table 11: Existing defenses and the channels they address. Checkmarks denote the channel the defense targets; the loss channel (LiRA) is the surviving one our method closes.

Defense	Repr.	Grad.	Output	Boundary	**Loss**
AdvReg (Nasr et al., 2018)	✓				
RelaxLoss (Chen et al., 2022)			✓		
SELENA (Tang et al., 2022)			✓		
HAMP (Chen & Pattabiraman, 2024)			✓		
MemGuard (Jia et al., 2019)			✓		
DP-SGD (Abadi et al., 2016)		✓			(✓)
Geometric (ours, track 1)	✓	✓		✓	
Dual track (ours)	✓	✓		✓	✓