The Privacy Peak

Non-Monotonic Leakage in Federated Learning

WORKSHOP ON DATA PRIVACY MANAGEMENT · DPM 2026

A reproducible, post-hoc reconstruction study of privacy leakage across federated heterogeneity regimes

Hamza Sajjad, Xavier Lessage, Gabriele Costa, Philippe Massonet

IMT School for Advanced Studies Lucca, Italy · CETIC asbl, Belgium

hamza.sajjadahmed@imtlucca.it

DPM 2026 co-located with ESORICS 2026 · Rome, Italy

Agenda

01 **Motivation**
Why FL privacy is not solved by design

02 **Threat model**
Post-hoc, white-box, model-only adversary

03 **Method**
Anchor-guided reconstruction framework

04 **Setup**
CIFAR-10 across four heterogeneity regimes

05 **Results**
The non-monotonic “privacy peak” at E3

06 **Discussion**
Why moderate skew is the risky regime

Federated Learning Was Supposed to Protect Privacy

The promise

- Clients keep raw data local; only model updates are shared with a central server
- Widely adopted for healthcare, mobile, and personalized services precisely because it avoids centralizing sensitive data
- Federated Averaging (FedAvg) aggregates client updates into one global model

The catch

- Model updates and the final trained model still encode information about the data that produced them
- Membership inference, attribute inference, and reconstruction attacks all show FL alone is not privacy
- Most prior reconstruction work assumes access to gradients or client updates — rarely the deployed model alone

This paper asks: does FL alone protect privacy once training is done — with only the final model in an adversary's hands?

An Open Question: Does Heterogeneity Help or Hurt Privacy?

Client data is rarely IID in practice. Data heterogeneity is known to affect accuracy and convergence — but its effect on privacy leakage is far less understood.

1

Limited heterogeneity ranges

Most studies test only IID or one non-IID level, not a full spectrum

2

Optimization-success bias

Reconstruction quality is judged by convergence, not semantic consistency or stability

3

Training-time assumptions

Few works consider a strictly post-hoc adversary with only the final global model

Our goal: systematically measure post-hoc reconstruction leakage across the full IID → extreme non-IID spectrum.

Research Questions & Contributions

RQ1 How does data heterogeneity affect the vulnerability of a federated model to post-hoc reconstruction?

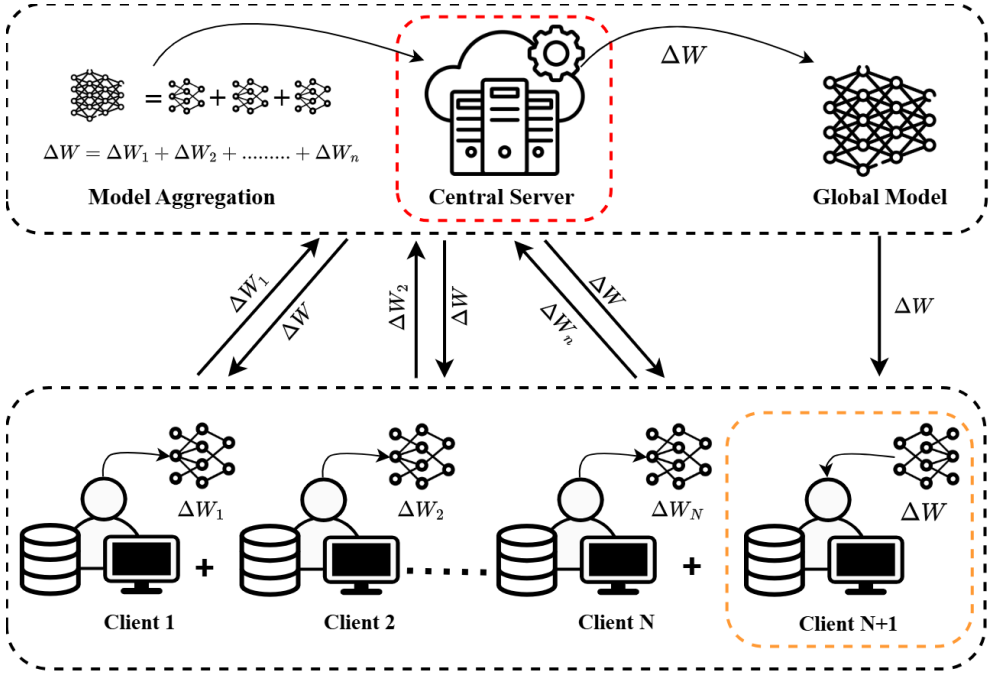
RQ2 Can anchor-guided initialization improve inversion without gradients, updates, or training dynamics?

RQ3 Is the heterogeneity – privacy relationship monotonic, as commonly assumed?

Contributions

- A systematic study of leakage across IID → extreme non-IID regimes, revealing a non-monotonic reconstruction trend
- An anchor-guided reconstruction framework using only the final global model (no gradients, updates, or training dynamics)
- Statistical validation, ablations, and class-wise analysis showing federated models retain recoverable information even under strong distribution shift

A Strictly Post-Hoc, Model-Only Adversary



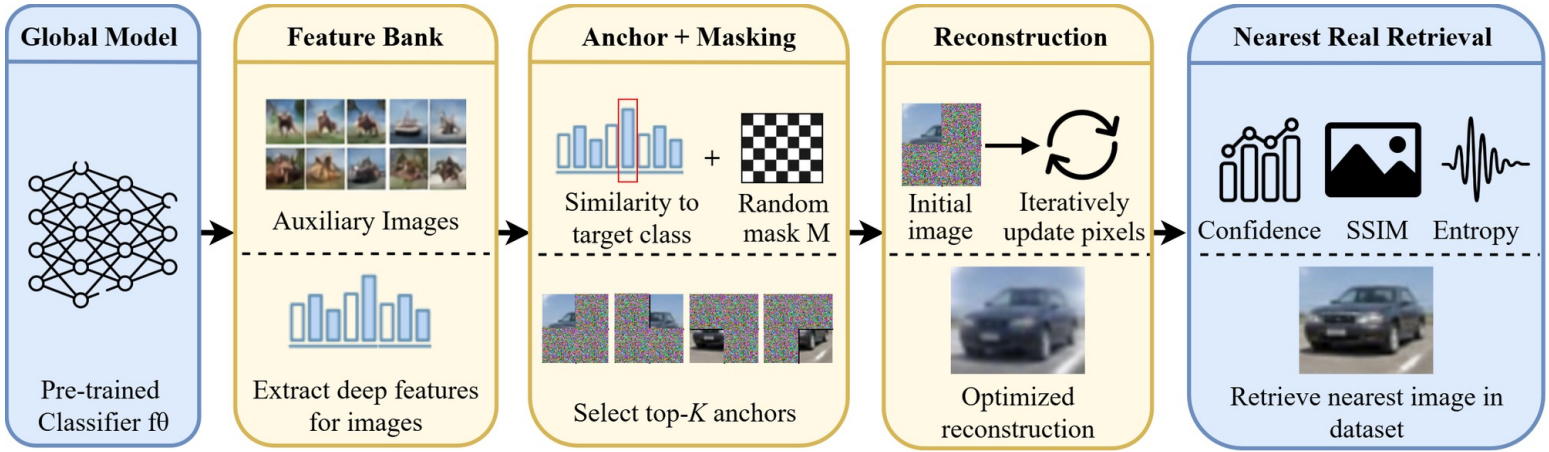
Adversary knows

- Final trained global model fθ — all parameters, BatchNorm statistics
- Can query the model for output logits on arbitrary inputs

Adversary does NOT know

- Client datasets, gradients, local updates, or optimizer states
- Any intermediate training-time information

Anchor-Guided Reconstruction Framework



How the Pipeline Works

- Feature bank → anchor + masking → iterative reconstruction → nearest-real check
- Touches only the frozen global model fθ — the exact post-hoc adversary from the threat model

Why This Matters

- This anchor-guided attack is the paper's core technical contribution
- Every result in the rest of this talk comes from running this exact pipeline

Partial-Anchor Initialization & Multi-Objective Loss

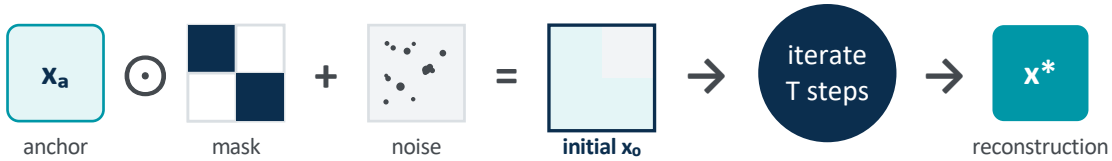
Partial-anchor initialization

- x0 = M ⊙ xa + (1−M) ⊙ xn — anchor xa preserves partial semantic content, mask M and noise xn add diversity
- Optimize a reparameterized variable u with x = σ(u), mapping the latent variable into valid image space
- Multiple random restarts mitigate sensitivity to local minima

Multi-objective loss

- L = Lce + λbn·Lbn + λf·Lfeat + λp·Lperc + λtv·Ltv + λc·Lcolor + λa·Lanchor
- Enforces classification alignment, BatchNorm-statistic matching, feature consistency, perceptual similarity, smoothness, color regularity, and anchor preservation
- Final reconstruction per sample = restart with maximum target-class confidence under fθ

From equation to pixels



Experimental Setup

Design

- Dataset: CIFAR-10, identical architecture/optimization across all settings
- Heterogeneity via Dirichlet partitioning ($\alpha = \infty$, 0.5, 0.3, 0.1)
- 100 reconstructions per experiment (10 classes × multiple anchors)

Metrics

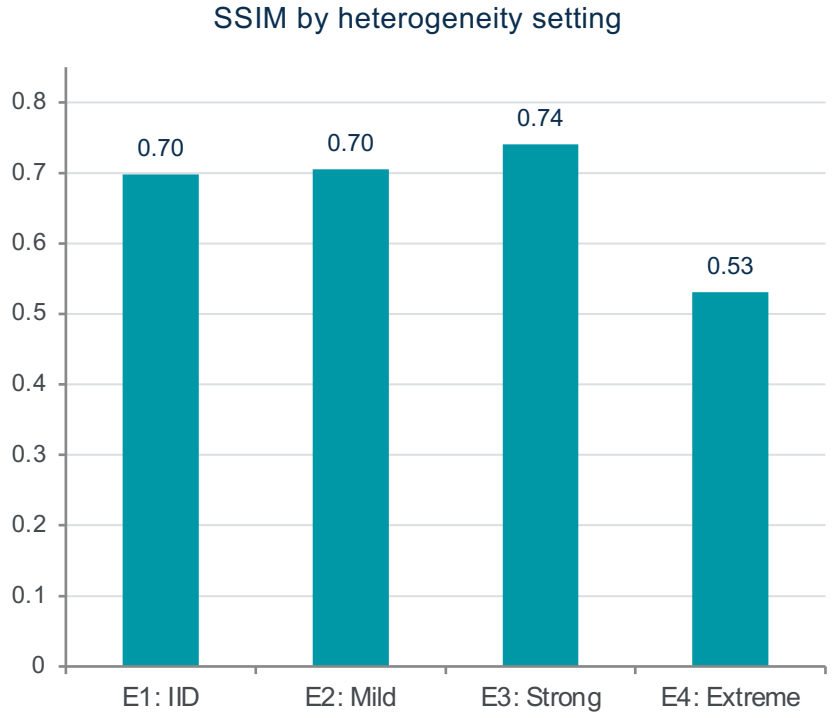
- Pixel-level: MSE, SSIM
- Semantic: penultimate-layer feature distance
- Privacy proxy: unique nearest-neighbour training images retrieved

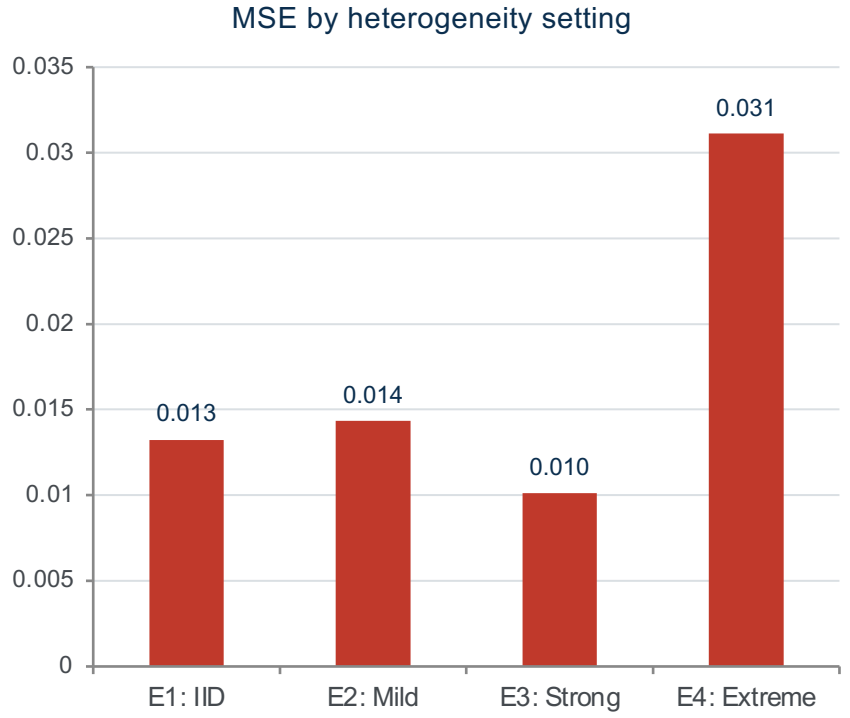
Heterogeneity regimes & global-model accuracy

E1	IID	**94.11%**
E2	Mild non-IID	**93.18%**
E3	Strong non-IID	**89.51%**
E4	Extreme non-IID	**65.39%**

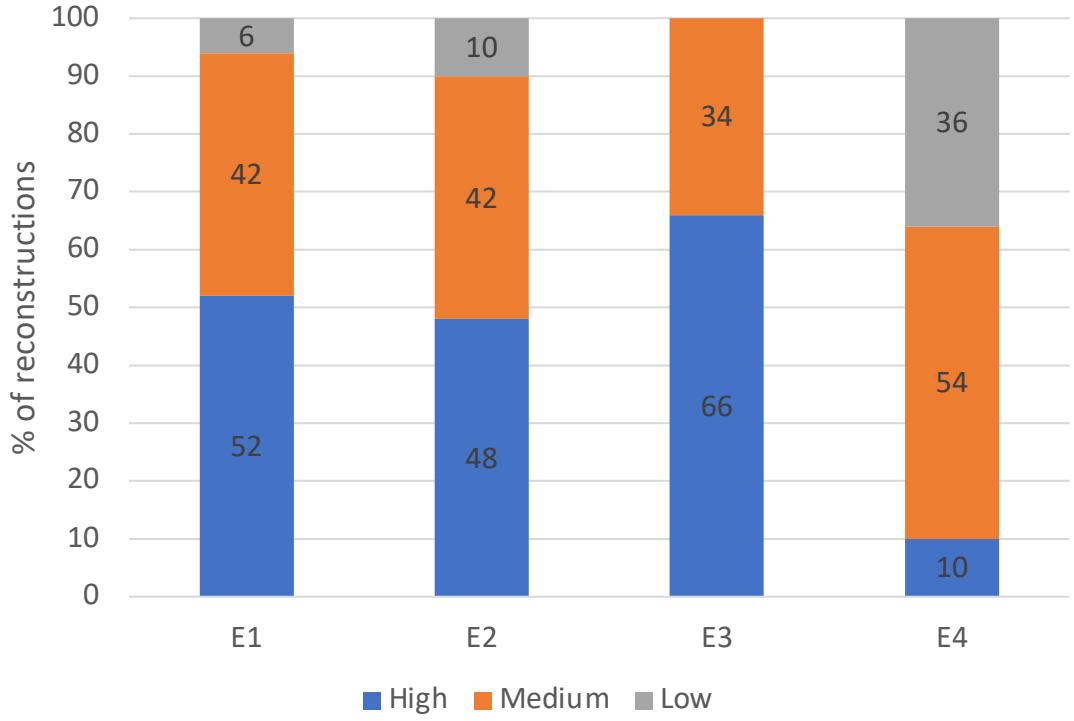
accuracy ↓ sharply only at E4 (extreme skew)

Reconstruction Fidelity Peaks at Strong Non-IID (E3)





Leakage-Level Classification Across Settings



Reading the peak

- E3 (strong non-IID): 66% high-leakage, 0% low — the most concentrated, stable reconstructions
- E4 (extreme non-IID): only 10% high-leakage; shifts to medium/low as optimization becomes unstable
- Thresholds (SSIM, unique-NN, feature distance) applied uniformly across settings for comparability

E3's Advantage Is Statistically Significant

Welch's t-test on per-sample SSIM and feature distance (N = 100 per setting, α = 0.05)

Comparison	Metric	p-value
E3 vs E1	SSIM	**0.004**
E3 vs E1	Feature Distance	**< 0.001**
E3 vs E2	SSIM	**0.018**
E3 vs E2	Feature Distance	**< 0.001**

All differences are significant ($p < 0.05$): E3's higher SSIM and lower feature distance are unlikely to be sampling noise.

Vulnerability Is Also Class-Dependent

Setting	Hardest class	Feat. dist.	Easiest class	Feat. dist.
E1 — IID	bird	0.4145	ship	0.2561
E2 — Mild non-IID	dog	0.6890	cat	0.2609
E3 — Strong non-IID	airplane	0.3012	horse	0.1664
E4 — Extreme non-IID	bird	0.3046	frog	0.0873

- Structurally complex classes (airplane, deer, bird) are more variable; simple, prototypical classes (horse) stay stable
- Under extreme non-IID (E4), visually complex classes (bird, dog) degrade most — heterogeneity amplifies existing disparities

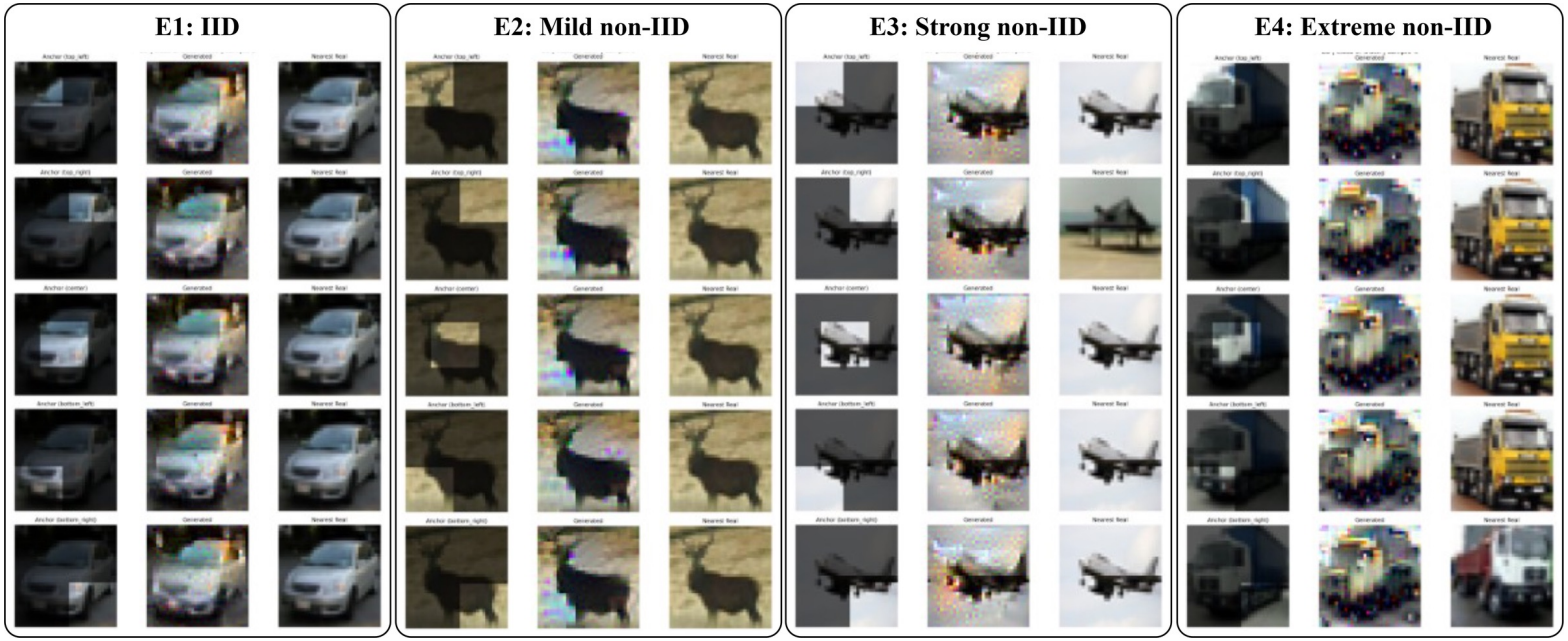
Ablation: Every Component Contributes

Setting	Method	SSIM	Feat. Dist.
E3	Random Init	0.0907	1.4547
E3	No Feature Bank	0.1794	2.5508
E3	**Full Method**	**0.6385**	**1.4755**

Takeaways (consistent across all four settings E1–E4)

- Random-noise initialization is a weak baseline — SSIM stays below 0.11 everywhere
- Removing the feature bank hurts fidelity and inflates feature distance, despite similar SSIM to random init
- The full anchor-guided method with feature-bank guidance gives the largest, most consistent gains

Qualitative Reconstructions Across Heterogeneity



Each panel: anchor → reconstruction → nearest real CIFAR-10 image. Reconstructions stay class-consistent and stable across anchor positions, especially at E3.

Why a “Privacy Peak” at Moderate Skew?

Representation stability drives leakage

- At E3, the global model learns smooth, class-consistent representations — ideal gradients for reconstruction
- Leakage tracks representation stability and concentration, not heterogeneity magnitude itself

Extreme skew is not better privacy

- E4's weaker reconstructions come from degraded model utility (65.39% accuracy), not from stronger privacy
- Feature-space similarity remains relatively strong even at E4 — semantic information is still partly recoverable

Neither extreme keeps you safe: privacy risk is a curve, not a straight line — and its peak sits in the middle.

Conclusion

> *Federated models retain semantically meaningful, recoverable information about their training data — even under strong distribution shift and even when adversaries never see a single gradient or client update.*

1 What We Found

- Reconstruction fidelity peaks at strong — not extreme — non-IID conditions: a non-monotonic “privacy peak.”
- Holds even for the strictest adversary: no gradients, no client updates, only the final deployed model.

2 What It Means

- Benchmarking privacy only at IID or extreme skew can underestimate real-world risk.
- A vulnerable “moderate skew” regime deserves explicit attention in FL privacy evaluations.

3 What's Next

- Testing generalization beyond CIFAR-10 — CIFAR-100, Tiny-ImageNet, and other domains.
- Exploring data-free variants and extending the framework to other model families.

Thank You

Questions & Discussion

Hamza Sajjad

hamza.sajjadahmed@imtlucca.it

github.com/imhamzasajjad/anchor_guided_fl_reconstruction

Supported by RDS PTR 25–27 (Cybersecurity) and SPF BOSA project AIDE