Supplementary Material FedXDS: Leveraging Model Attribution Methods to Counteract Data Heterogeneity in Federated Learning

Maximilian Andreas Hoefler[1] * Karsten Mueller[1] Wojciech Samek[1,2,3]
[1]Fraunhofer Heinrich Hertz Institute, Berlin, Germany
[2]Technical University of Berlin, Berlin, Germany
[3]Berlin Institute for the Foundations of Learning and Data (BIFOLD)

1. Implementation Details

All models use a ResNet8 architecture and are trained with stochastic gradient descent (SGD) with a momentum of 0.9 and a learning rate of 0.01. Attribution computations are performed using the Zennit library [4]. For evaluation, we use top-1 accuracy as the primary metric.

FEMNIST For our implementation, we use the FEMNIST dataset as organized in Caldas et al. (2019). This dataset, derived from EMNIST, includes 62 classes (digits and both uppercase and lowercase letters). Each client's data comes from a single writer—roughly 200–300 samples per client. We work with 100 clients in total and select 10 clients per communication round.

CelebA We evaluate our method on CelebA, following the setup of [35] and [7]. The dataset is structured as a binary classification task, distinguishing between "smiling" and "not smiling." We utilize ten clients with a 50% participation rate per round. This dataset is important from the vantage point of privacy. We show below reconstruction attacks on faces, trying to re-identify faces visually.

2. Extended Discussion on Task-Relevant Features and Privacy

In our work, we build upon the finding that sharing features derived from raw client data helps counteract data heterogeneity in federated learning. However, this approach inherently poses privacy risks. While differential privacy (DP) could in principle be applied directly to the raw client data, the high dimensionality of image data would necessitate substantial noise addition, significantly decreasing utility. ε-metric privacy. This method allows us to maintain strong privacy guarantees while preserving the utility benefits of feature sharing. Unlike standard differential privacy, which defines neighboring datasets by a single-entry difference, we define privacy based on input similarity under a metric d_X, allowing a more natural treatment of image data where task-relevant and potentially sensitive features (e.g., facial attributes, backgrounds) are often inseparable.

We emphasize that our approach treats all retained pixels after masking as equally sensitive and applies our privacy mechanism to every pixel that remains after masking. The attribution-based method identifies these crucial regions dynamically without predefined sensitive regions, ensuring adaptive privacy without manual supervision. Notably, we do **not** assume that the **retained** pixels are "less private" than those discarded; rather, our ε-metric mechanism uniformly bounds the disclosure risk of whichever features are deemed relevant. By removing uninformative regions altogether and applying noise only to the most discriminative subset, we reduce the dimensionality of data requiring protection while limiting an adversary's ability to infer private details from the shared features. This approach provides better theoretical guarantees while maintaining utility. Our aim is to protect against membership inference attacks (MIA) and feature inversion when sharing these masked features, which we demonstrate empirically in . Below we detail the technical arguments that ensure metric privacy.

2.1. Why Task-Relevant is equivalent with Potentially Sensitive Features

Many real-world image tasks involve visual features that are both *critical for classification* and *potentially identifying* (e.g., facial landmarks, unique physical traits, or distinctive backgrounds). While it might be tempting to assume that discarding "sensitive regions" a priori would suffice, this often fails in practice if the model itself relies on exactly those regions for

*Correspondence to maximilian.andreas.hoefler@hhi.fraunhofer.de

accurate predictions. Our method *does not* rely on any manual definition of what constitutes "sensitive" or "background." Instead, we let an *attribution-based mask* discover the truly *discriminative* subset of pixels. Because these retained pixels can naturally include highly private elements (for example, a key part of the face), we *treat all retained pixels as equally sensitive* and apply our privacy mechanism to every pixel that survives the mask.

2.2. Adaptive Attribution Masking Without Manual Annotations

The binary mask $\mathbf{m}$ arises from an attribution method that identifies pixels most influential to the model's output. This ensures an *adaptive*, data-driven selection of features, mitigating the need for *pre-labeled* sensitive regions (e.g., bounding boxes for faces). By defaulting to the viewpoint that any discriminative pixel is potentially private, our approach sidesteps the risk of *missing* an unintuitive but privacy-revealing region. All masked-out (zeroed) pixels are completely removed from subsequent sharing, effectively *reducing* the dimensionality of data exposed to the privacy mechanism.

2.3. Uniform Privacy Guarantee Over Retained Pixels

One might worry that by selectively *retaining* the most discriminative features, the mask could inadvertently highlight the *most private* parts of the image (such as a person's unique facial contour). However, under our ε-metric privacy model, the mechanism adds *Gaussian noise* calibrated to a *worst-case sensitivity* $\Delta_f = 1$. Concretely, for any two images that differ in potentially identifying ways, the *distribution* of the noisy output remains within e^{ε} of each other, bounding an adversary's ability to infer whether a specific individual's pixels were present. Hence, even if a user's exact facial features remain in the retained region, the probability of distinguishing their face from another's in the shared representation remains formally constrained by ε.

2.4. Practical Advantages of Non-Expansive Masking

As detailed in Section 4.2, our function $f(\mathbf{x}) = \mathbf{x} \odot \mathbf{m}$ is *non-expansive* under the ℓ_2 norm, yielding a bounded global sensitivity $\Delta_f \leq 1$. This bounded sensitivity is crucial: it means the *noise scale* we add does not have to grow arbitrarily for large or varied image spaces. By retaining only the task-relevant pixels and zeroing out the rest, we effectively:

- Remove vast amounts of non-essential background or other context that might be re-identifying in unexpected ways.
- Limit how much noise is needed to preserve privacy, thereby improving the *utility–privacy tradeoff*.

2.5. Extensions and Future Directions

Region-Specific Noise While we apply a *single* noise scale σ to all retained pixels, the same framework could be extended to *varying* noise scales for different groups of pixels if domain knowledge indicates that some regions (e.g., near eyes) are more sensitive. This would require refining the sensitivity analysis (e.g., weighting the mask) but is straightforwardly encompassed by the metric privacy formulation.

Combining with Other Privacy Filters In high-stakes settings where certain features (like license plates or medical indicators) are strictly off-limits, one could combine our attribution mask with *rule-based filters* that forcibly zero out specific known identifiers before applying noise. This hybrid approach ensures that critical known-sensitive areas are always removed, while other discriminative regions receive the DP-based protection automatically.

Adapting the Metric d_X We adopt a straightforward ℓ_2 norm over image space for simplicity, but the metric d_X can be customized (e.g., ℓ_1, embedding-based distances, or perceptual metrics). The choice of metric can further align with specific privacy or robustness goals, such as controlling visually perceptible differences.

3. Sensitivity for Differential Privacy

3.1. Mathematical Formulation

In traditional differential privacy, the sensitivity of a function $f : \mathcal{D} \to \mathbb{R}^m$ measures how much the output can change when one record in the dataset changes. For continuous data domains like images, a more appropriate notion is the following senstivity choice as used in prior works [9, 13, 27] as follows:

Definition 1 (Sensitivity)**.** *For a function $f : X \to \mathbb{R}^m$ where (X, d_X) is a metric space, the Lipschitz sensitivity of f is defined as:*

$$\Delta_f = \sup_{x,x' \in X, x \neq x'} \frac{\|f(x) - f(x')\|_2}{d_X(x, x')}$$

This definition captures the maximum rate of change of the function f with respect to changes in the input space. A function with this sensitivity Δ_f satisfies:

$$\|f(x) - f(x')\|_2 \leq \Delta_f \cdot d_X(x, x')$$

for all $x, x' \in X$. In the context of image data, $X = \mathbb{R}^{H \times W \times C}$ typically represents the space of images with height H, width W, and C channels, and $d_X(x, x') = \|x - x'\|_2$ is the Euclidean distance between images.

3.2. Privacy Mechanism

Using this sensitivity definition, we can construct a privacy mechanism by adding calibrated Gaussian noise:

$$\mathcal{M}(x) = f(x) + \mathcal{N}(0, \sigma^2 I_m) \tag{1}$$

where $\sigma = \frac{\Delta_f \cdot \sqrt{2 \ln(1.25/\delta)}}{\varepsilon}$ for (ε, δ)-differential privacy. This mechanism satisfies:

Theorem 1. *For a function $f : X \to \mathbb{R}^m$ with Lipschitz sensitivity Δ_f, the mechanism $\mathcal{M}(x) = f(x) + \mathcal{N}(0, \sigma^2 I_m)$ with σ as defined above is (ε, δ)-differentially private with respect to the metric d_X.*

3.3. Practical Advantages of our Sensitivity Choice

Our choise of sensitivity provides a natural way to reason about privacy for continuous domains like images, where small changes in input should correspond to proportionally small privacy losses, unlike traditional sensitivity which is designed for discrete changes. This approach exhibits scale invariance, automatically adjusting to the scale of the data—if inputs are scaled by some factor, the privacy guarantees remain consistent without needing to recalibrate the privacy mechanism.

A significant benefit is tighter noise calibration. For many applications, Lipschitz sensitivity allows adding noise proportional to the actual "size" of the input difference rather than adding a fixed amount of noise based on worst-case scenarios, leading to better utility. In image data, where the distance between inputs has semantic meaning (similar images are close in pixel space), this choice of sensitivity respects this semantic structure in the privacy guarantee.

For unbounded domains like $\mathbb{R}^d$, traditional global sensitivity might be infinite, rendering standard mechanisms unusable, while the sensitivity can remain finite. This approach also aligns well with stable machine learning algorithms that are already designed so that small input changes cause small output changes, making them naturally compatible with our sensitivity frameworks.

3.4. Graduated Protection Based on Image Similarity

Unlike binary notions of privacy, Lipschitz sensitivity provides a continuous spectrum of protection that aligns with visual similarity. This is formalized through the concept of indistinguishability:

Proposition 1 (Graduated Protection)**.** *For two images x, x' with distance $d_X(x, x') = d$, and a mechanism $\mathcal{M}$ with Lipschitz sensitivity Δ_f, the privacy loss is bounded by $\varepsilon \cdot d$.*

This means that visually similar images (small d) are more difficult to distinguish even with access to the private output, while images that are perceptually very different receive appropriately scaled protection.

3.4.1. Natural Handling of Feature Correlations

Images exhibit strong spatial correlations between pixels, which traditional independent noise addition fails to respect. Lipschitz sensitivity naturally accommodates these correlations:

Proposition 2 (Correlation Preservation)**.** *For spatially correlated features in images, Lipschitz-based mechanisms preserve the relative importance of correlated structures while providing privacy guarantees.*

This property ensures that important visual structures (edges, textures, shapes) receive appropriate protection without being disproportionately distorted by the privacy mechanism.

3.5. Applications to Federated Learning

In the context of federated learning with image data, Lipschitz sensitivity provides a theoretical foundation for:

1. Bounding information leakage from shared features
2. Ensuring that small variations in sensitive attributes cannot be recovered
3. Providing guarantees against membership inference attacks
4. Enabling privacy-utility tradeoffs that scale with the semantic importance of features

The use of attribution-based masking in combination with Lipschitz sensitivity, as employed in our method, further enhances these properties by focusing the privacy protection on the most task-relevant features, ensuring that noise addition is maximally efficient in preserving utility while maintaining privacy guarantees.

3.6. Comparison with Traditional Sensitivity

Traditional global sensitivity for a function $f : \mathcal{D} \to \mathbb{R}^m$ is defined as:

$$\Delta_f^{global} = \max_{D, D' \text{ adjacent}} \|f(D) - f(D')\|_2$$

For unbounded continuous domains like images, this sensitivity can be infinite, making standard differential privacy mechanisms unusable. In contrast, Lipschitz sensitivity remains bounded as long as the function f is Lipschitz continuous, which is the case for many practical feature extraction methods in computer vision.

Additionally, for high-dimensional data, traditional additive noise mechanisms require noise scaling with $\sqrt{d}$ for d-dimensional outputs, whereas Lipschitz sensitivity allows for noise calibration that depends only on the privacy parameters and the Lipschitz constant, not the dimensionality of the data.

3.7. Sensitivity Analysis and Dimensionality Reduction

An interesting observation regarding our sensitivity bound $\Delta_f \leq 1$ is that this bound equals the sensitivity of the identity function. Indeed, for the identity function $\text{id}(x) = x$, we have $\Delta_{\text{id}} = \max_{x,x'} \frac{\|x-x'\|}{\|x-x'\|} = 1$. This raises the question: If both our masking approach and the unmasked (identity) approach have the same formal sensitivity, what concrete privacy advantage does our method provide?

The key insight lies in the *dimensionality* of the space requiring noise addition. Consider an image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ with $d = H \times W \times 3$ total dimensions. With the standard approach (no masking), noise would be added to all d dimensions according to $\mathcal{M}_{\text{id}}(\mathbf{x}) = \mathbf{x} + \mathcal{N}(0, \sigma^2 \mathbf{I})$. In contrast, our approach first applies a sparsifying mask $\mathbf{m}$ that retains only a fraction $\alpha = \frac{\|\mathbf{m}\|_0}{d}$ of the dimensions, where $\|\mathbf{m}\|_0$ counts the number of non-zero elements in $\mathbf{m}$.

For a given privacy budget ε, both approaches require the same per-dimension noise scale σ. However, the total expected squared ℓ_2 distortion differs significantly:

$$\mathbb{E}[\|\mathcal{M}_{\text{id}}(\mathbf{x}) - \mathbf{x}\|_2^2] = d\sigma^2 \tag{2}$$

$$\mathbb{E}[\|\mathcal{M}(\mathbf{x}) - \mathbf{x}\|_2^2] = \alpha d\sigma^2 + \|(\mathbf{1} - \mathbf{m}) \odot \mathbf{x}\|_2^2 \tag{3}$$

The first term in our approach's distortion ($\alpha d\sigma^2$) represents the noise added to the retained dimensions, while the second term represents the distortion from setting masked dimensions to zero. Crucially, the masked dimensions are specifically those deemed less relevant to the task by our attribution method. This means that for a fixed privacy budget ε, our approach concentrates the noise budget on fewer, more task-relevant dimensions rather than spreading it across all dimensions.

Furthermore, by applying our attribution-guided masking, we explicitly remove potentially identifying background features entirely, rather than simply perturbing them with noise. This dual effect—reducing dimensionality while focusing on task-relevant features—allows our approach to maintain higher utility at equivalent privacy levels, as demonstrated empirically in the main text.

This analysis highlights a fundamental principle in privacy-preserving machine learning: when limited to a fixed privacy budget, selective disclosure of the most task-relevant features often yields better utility than indiscriminate disclosure of all features with uniformly distributed noise. Our attribution-guided masking approach provides a principled method for implementing this selective disclosure while maintaining formal privacy guarantees.

4. Inversion Attacks

We conduct inversion attacks on three different methods. Specifically we show the denoising inversion attack for our method FedXDS Figure 1, for the raw and DP protected features Figure 2, and for FedFed in Figure 3 which shares DP distilled features [30]. The results shows the superiority of our method in finding the most crucial representations and protecting crucial features needed for generalization. In contrast, raw features heavily leak privacy under the same DP constraints. This speaks to the validity of our privacy argument, i.e., that the norm is significantly reduced such that we get a stronger DP guarantee at the same noise level. In addition, FedFed also appears to be prone to the denoising attack. Although not as severe as in sharing raw features, it is nevertheless possible to reconstruct a considerable amount from the original images.

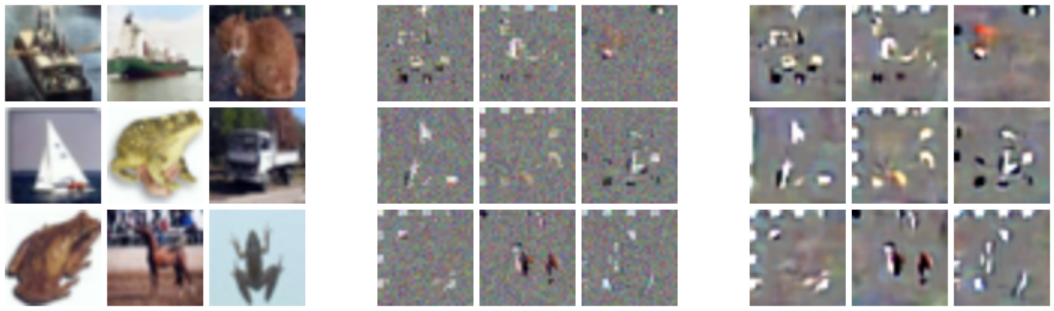


Figure 1. Inversion Attack on FedXDS

Figure 2. Inversion Attack on Raw Features

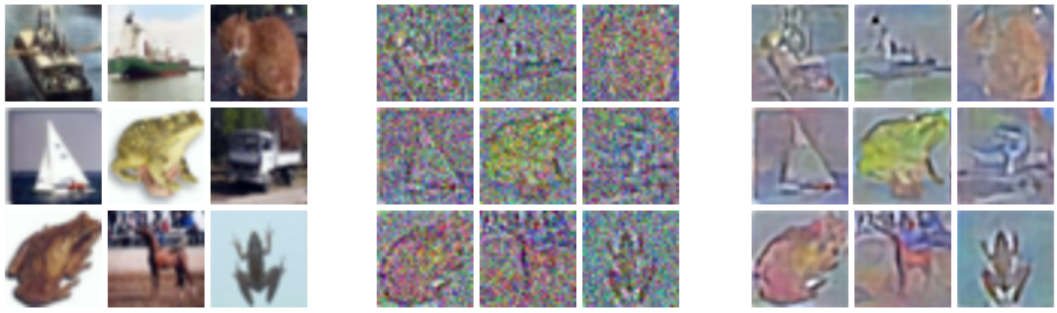


Figure 3. Inversion Attack on FedFed

5. Privacy Outlook and Considerations in Federated Learning

Privacy remains a critical challenge in federated learning systems [11, 17]. Recent investigations have revealed significant vulnerabilities to various privacy attacks, particularly model inversion attacks [10] and membership inference attacks [23],

which can potentially reconstruct private training data or determine if specific samples were used during training. These concerns are particularly relevant in our context, as our method introduces additional data sharing mechanisms that could potentially expand the attack surface.

Several promising approaches exist to enhance privacy guarantees in federated systems. Differential privacy (DP) [1, 8] has emerged as a leading framework, offering mathematical guarantees for privacy preservation. While traditional DP mechanisms often involve adding calibrated noise to model updates [2], recent advances in adaptive noise scaling [32] and personalized privacy budgets [29] provide more nuanced trade-offs between model utility and privacy protection.

To strengthen the privacy guarantees of our methodology, we propose several potential extensions:

- Integration of secure aggregation protocols [6] combined with local differential privacy to protect individual client contributions
- Implementation of gradient pruning and compression techniques [21] to minimize potential information leakage while maintaining performance
- Adoption of privacy-preserving attribution mechanisms that provide interpretability without compromising sensitive information [26]

Future work should investigate the empirical privacy-utility trade-offs of these approaches in our context, particularly focusing on how they interact with our attribution-based sampling strategy. Additionally, exploring privacy-preserving techniques specifically designed for handling interpretation mechanisms could provide valuable insights for the broader federated learning community.

6. Extended Discussion on Attribution Methods and Sparsification

Motivation and Setup. In our main text, we evaluated how different *attribution methods* perform under increasing sparsification of the input image. Specifically, we applied a thresholded binary mask $\mathbf{m}$ to retain only the top-$s\%$ most salient pixels, where s ranges from 60 to 85. We observed that **FedXDS maintains stable performance** across all attribution techniques (LRP, Gradient, SmoothGrad, Integrated Gradients), despite removing a significant fraction of pixels at higher s values.

These findings suggest that **selecting truly discriminative pixels** is crucial for accurate predictions and that *how* those pixels are identified can strongly influence final model performance under extreme sparsification.

6.1. Why LRP is Particularly Robust at High Sparsification

Structured vs. Isolated Features. Our experiments consistently show that LRP [5, 14, 25] outperforms gradient-based approaches. Unlike methods such as SmoothGrad or basic gradients—which often highlight *isolated* pixels deemed locally important—LRP is designed to *propagate* relevance backward through each layer of the network. This layer-wise relevance propagation produces *coherent, contiguous* attributions. Consequently, at high sparsity (e.g., $s = 85$), the retained pixel regions form *semantically consistent* patches rather than scattered points, preserving enough context to maintain robust classification performance.

Numerical Stability via the ε-Rule. LRP's use of the ε-rule helps mitigate saturation effects, ensuring that small variations in activation do not excessively amplify or nullify the attributions. In contrast, gradient-based methods can encounter vanishing or exploding gradients, particularly in deeper networks, leading to unstable attributions. When such attributions are thresholded, the final retained pixels may be *suboptimal* for preserving discriminative features. By contrast, LRP's numerical stability and hierarchical propagation framework preserve more *functionally relevant* pixels under high-threshold masking.

Saliency vs. Structural Relevance. While gradient methods excel at detecting certain "hotspot" pixels, they sometimes overlook broader *structural* cues. By design, LRP aims to reveal how individual neurons (and layers) contribute to the network's decision in a cumulative manner, capturing important spatial patterns. At high sparsification, *having a coherent region* of relevant pixels (rather than discrete points) can maintain classification accuracy. Our results confirm that this approach is particularly advantageous in federated scenarios, where *masked, privacy-preserving data* must still retain sufficient representational power for shared training.

6.2. Implications for FedXDS

Overall, these experiments confirm that the *efficacy* of FedXDS in federated scenarios hinges on:

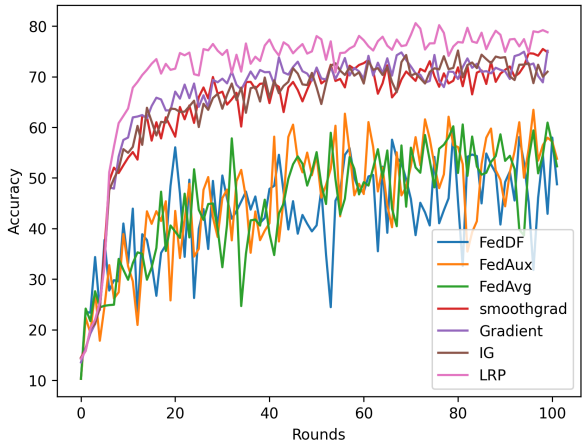
1. **Effectively identifying task-relevant regions** (i.e., robust attributions), so that the most informative pixels remain under high sparsification.

2. **Preserving** enough spatially coherent context to enable generalization despite removing most of the input.

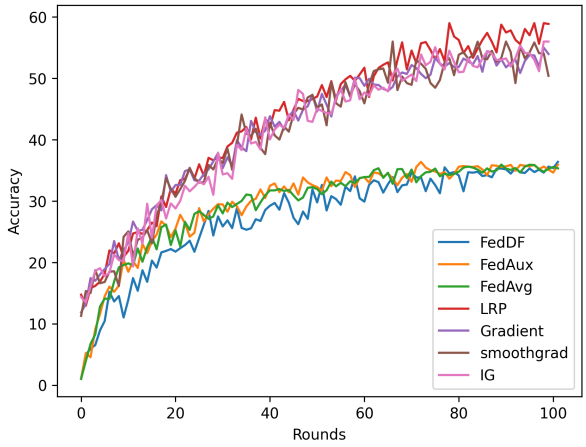
The superior performance of LRP underscores the importance of *structural coherence* in saliency maps when only a small fraction of the image is shared. Combined with FedXDS's privacy-preserving pipeline, LRP's ability to highlight *contiguous* and *semantically meaningful* areas ensures that even heavily masked images can support effective model training.

7. Training Porgression

We plot the training progression for the CIFAR10 and CIFAR100 datasets of FedXDS versus some common baselines. It shows that the convergence speed and overall progression is faster and smoother for our method.



(a) Training Progression of FedXDS with CIFAR10, K=10 and $\alpha = 0.1$.



(b) Training Progression of FedXDS with CIFAR100, K=10 and $\alpha = 0.1$.

8. Performance and Number of Samples

We also conduct experiments on how the performance of our method depends on the number of samples shared. Specifically, we share decreasing fractions of the original dataset size of each client and track the performance. The results are shown in Table 1:

Table 1. Test accuracy (%) for different sample sizes across various attribution methods.

Fraction of Original Samples	Gradient	SmoothGrad	Int. Gradients	LRP
0.5	70.12	69.85	71.23	**77.61**
0.6	71.34	71.02	72.45	**78.91**
0.7	72.15	71.88	73.12	**79.05**
0.8	73.45	73.21	74.56	**80.38**
0.9	74.82	74.53	75.89	**81.61**
1.0	76.12	75.89	77.23	**83.67**

The table shows classification accuracy for different sample sizes from 0.5 to 1.0 across four attribution methods. LRP consistently achieves the best performance, with accuracy ranging from 77.61% to 83.67%, while the other methods (Gradient, SmoothGrad, and Integrated Gradients) perform 6-8% worse but improve similarly with increasing sample size.

9. Significance Tests on Improvements

To rigorously validate the performance gains of our proposed method, we conducted statistical significance tests comparing `FedXLRP` to all other baselines. We performed a one-tailed t-test for each experimental setting, with the null hypothesis

Table 2. Statistical significance of FedXLRP's accuracy improvements on CelebA and FEMNIST. Table shows p-values from a one-tailed t-test comparing FedXLRP against each baseline. The reference performance for FedXLRP is 91.55 ± 0.48 (CelebA) and 89.03 ± 0.35 (FEMNIST).

Baseline Method	**CelebA (p-value)**	**FEMNIST (p-value)**
FedAvg	< 0.001	< 0.001
FedProx	< 0.001	< 0.001
SCAFFOLD	< 0.001	< 0.001
FedDyn	< 0.001	< 0.001
FedSAM	< 0.001	0.002
FedDISCO	0.001	0.001
FedFed	0.009	0.007
FedFTG	0.006	0.004
FedGen	0.002	< 0.001
FedAux	0.001	< 0.001
FedDF	< 0.001	< 0.001

Table 3. Statistical significance (p-values) of FedXLRP's accuracy improvement over other federated learning methods. P-values are derived from a one-tailed t-test for each experimental setting. "n.s." denotes a non-significant result ($p \geq 0.05$) where FedXLRP did not outperform the baseline.

Dataset	CIFAR-10				CIFAR-100				Tiny-ImageNet			
	K=10		K=100		K=10		K=100		K=10		K=100	
	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$
FedAvg	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
FedProx	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.002	< 0.001	0.003	< 0.001	< 0.001	0.015
FedDyn	< 0.001	0.002	< 0.001	< 0.001	0.005	< 0.001	0.001	< 0.001	0.002	0.004	< 0.001	0.021
SCAFFOLD	< 0.001	< 0.001	< 0.001	< 0.001	0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.005	< 0.001	0.011
FedSAM	< 0.001	< 0.001	< 0.001	< 0.001	0.003	< 0.001	0.001	< 0.001	0.001	0.003	< 0.001	0.018
FedDISCO	< 0.001	< 0.001	< 0.001	< 0.001	0.002	< 0.001	< 0.001	< 0.001	< 0.001	0.002	< 0.001	0.019
FedFed	0.012	0.025	0.021	0.018	n.s.	0.011	0.024	0.004	0.015	0.009	0.011	n.s.
FedFTG	< 0.001	0.003	< 0.001	0.002	< 0.001	0.004	< 0.001	< 0.001	0.001	< 0.001	< 0.001	0.025
FedGen	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.006	0.007	< 0.001	0.017
FedAux	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.013
FedDF	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.001	< 0.001	0.012
FedXIG	< 0.001	< 0.001	< 0.001	< 0.001	0.008	0.009	0.003	0.001	0.004	0.003	< 0.001	0.020
FedXGrad	< 0.001	< 0.001	< 0.001	< 0.001	0.011	0.005	0.002	0.001	0.003	0.001	< 0.001	0.014
FedXSG	< 0.001	< 0.001	< 0.001	< 0.001	0.023	0.008	0.001	0.001	0.004	0.002	< 0.001	0.016

being that `FedXLRP`'s performance is not superior to the baseline. As demonstrated by the consistently low p-values, the improvements achieved by `FedXLRP` are statistically significant across all datasets and heterogeneity configurations. A comprehensive summary of these p-values for our main experiments is presented in Table 3 in the appendix. Similarly, the significance of our method's substantial gains on the CelebA and FEMNIST datasets is detailed in Appendix Table 2. These results confirm that the observed superiority of `FedXLRP` is not due to random chance but represents a consistent and meaningful advantage.

10. Related Work

Federated Learning (FL) provides a framework for training models on decentralized data without explicit data sharing. The foundational algorithm, FedAvg [22], aggregates locally trained models to update a global model. However, its performance degrades significantly when client data is not independently and identically distributed (non-IID). Many subsequent works have sought to address this challenge.

10.1. Client Drift Correction and Regularization

One major line of research focuses on mitigating "client drift," where local models diverge due to heterogeneous data. FedProx [18] introduces a proximal term to the local objective function, restricting local updates from moving too far from the global model. Similarly, methods like SCAFFOLD [12] and the work of [15] introduce control variates to correct for the variance introduced by non-IID data, leading to improved convergence and stability. FedDyn [3] addresses client heterogeneity by dynamically regularizing local objectives based on historical updates, aligning local and global optima. Other specialized approaches include FedBN [19], which maintains local batch normalization parameters to handle feature shifts, FedNova [28], which normalizes local updates to ensure fair client contributions, and MOON [16], which uses model-contrastive learning to align local and global model representations. [24] attempts to regularize the loss landscape of clients, whereas [31] uses a discrepancy aware approach.

10.2. Data Sharing and Knowledge Distillation

Another prominent line of work aims to directly tackle data heterogeneity through various forms of data or knowledge sharing. FedDF [20] aggregates knowledge from client models into a global model by using ensemble distillation on an unlabeled public dataset, unifying disparate local knowledge without sharing the private data itself. Some methods leverage generative models to create shareable synthetic data. FedGen [35] synthesizes class-conditional feature-space representations to align distributions between clients and the server. FedFTG [33] transfers knowledge specifically at the feature level by generating pseudo-data and using it in an ensemble-distillation setup. FedFed [30] combines feature distillation with a variational auto-encoder to generate data under differential privacy constraints. FedAux [34] shares differentially private model predictions in a distillation framework, though it also requires a public dataset.

While powerful, these data-sharing approaches often introduce significant computational and communication overhead from generating and transmitting data. Furthermore, as we show in our experiments, generator-based methods may not guarantee strong empirical privacy. This highlights the central challenge that we seek to address: enabling effective knowledge sharing that mitigates heterogeneity while simultaneously preserving privacy and maintaining computational efficiency. Our work, `FedXDS`, diverges from these approaches by leveraging XAI to extract and share only the most salient feature importances—a highly compact and abstract representation that is inherently more private and efficient than sharing raw data, features, or predictions.

11. Warmup and Computational Efficiency

11.1. Warmup Rounds

We investigate the effect of varying the number of warmup rounds—used to initialize the model before applying attribution-based feature selection—on downstream model performance. Warmup is performed using standard FedAvg for $R_{\text{warmup}} \in \{0, 5, 10, 15\}$ rounds. The results, demonstrate a consistent improvement in test accuracy with increasing warmup. Across all datasets (CIFAR-10, CIFAR-100, Tiny-ImageNet), performance improves sharply from 0 to 10 rounds and then begins to plateau.

This behavior aligns with the need for stable and meaningful attributions: early in training, the model has not yet learned reliable patterns, and attribution maps may reflect noise or spurious correlations. A few rounds of pretraining help the model to focus on task-relevant features, producing more semantically meaningful relevance maps for masking. Beyond 10 rounds, marginal accuracy gains diminish, suggesting that longer warmup adds little benefit but incurs extra communication overhead. Thus, we fix $R_{\text{warmup}} = 10$ as a good trade-off between accuracy and efficiency in our main experiments.

11.2. Computational Efficiency

FedXDS is designed with federated settings in mind, where client devices often have limited compute and memory. Our approach offers key efficiency advantages over prior methods that rely on data generation:

- **One-time attribution pass:** Each client computes attribution maps using a single backward pass through a pretrained (warmup) model for each sample. This step is comparable in cost to one training epoch and is performed only once.
- **No generators or decoders:** Unlike FedGen and FedFTG, which require training and updating GANs or feature generators at every communication round, FedXDS relies on simple masking and additive noise. This eliminates the need for high-dimensional generation pipelines and avoids mode collapse or instability issues.
- **No latent encoders:** FedFed uses variational autoencoders (VAEs) for feature distillation, which requires both encoder and decoder networks, often doubling the model size and computation. Moreover, VAEs require tuning of KL regularization terms and are known to be sensitive to data heterogeneity.

In contrast, FedXDS operates with minimal model modifications and negligible per-round overhead. Once attribution-based masks are computed and privacy noise is added, the resulting masked dataset can be reused throughout training. This design not only reduces client-side resource requirements but also improves communication efficiency by accelerating convergence (as shown in Table 2 of the main paper).

We examine computation times (seconds per round) for CIFAR-10 ($\alpha = 0.1$, 10 clients, 50% participation, 100 rounds), in for generator training are: FedFed (79), FedGen (34), FedFTG (37), and FedXDS (32). FedFed requires 15 rounds to train the generator, while FedGen and FedFTG update every round. Our method requires a single backward pass per sample one time only, limiting computation to a single round.

In summary, FedXDS strikes a favorable balance between computational cost, model performance, and privacy—making it particularly well-suited for real-world federated learning deployments where resources are constrained.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016. 6

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, page 308–318, New York, NY, USA, 2016. Association for Computing Machinery. 6

[3] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021. 9

[4] Christopher J. Anders, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. Software for dataset-wide xai: From local explanations to global insights with zennit, corelay, and virelay, 2023. 1

[5] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 2015. 6

[6] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017. 6

[7] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018. 1

[8] Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. Now Publishers Inc, 2014. 6

[9] Oluwaseyi Feyisetan and Shiva Kasiviswanathan. Private release of text embedding vectors. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 15–27, 2021. 2

[10] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients–how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, pages 16937–16947, 2020. 5

[11] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021. 5

[12] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020. 9

[13] Fragkiskos Koufogiannis, Shuo Han, and George J. Pappas. Optimality of the laplace mechanism in differential privacy, 2015. 2

[14] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019. 6

[15] Bo Li, Mikkel N Schmidt, Tommy S Alstrøm, and Sebastian U Stich. On the effectiveness of partial variance reduction in federated learning with heterogeneous data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3964–3973, 2023. 9

[16] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021. 9

[17] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020. 5

[18] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. 9

[19] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fed{bn}: Federated learning on non-{iid} features via local batch normalization. In *International Conference on Learning Representations*, 2021. 9

[20] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems*, 33:2351–2363, 2020. 9

[21] Tao Lin, Lingchen Kong, Shaofeng Zhu, and Wei Liu. Privacy-preserving gradient sparsification for federated learning. In *International Conference on Machine Learning*, pages 13228–13242, 2022. 6

[22] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2017. 8

[23] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE Symposium on Security and Privacy (SP)*, pages 739–753, 2019. 5

[24] Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *International conference on machine learning*, pages 18250–18280. PMLR, 2022. 9

[25] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021. 6

[26] Abhishek Singh, Ayush Chopra, Vivek Sharma, Pin-Yu Chen, and Krishna P Gummadi. Attribution privacy: Bridging the gap between attribution and privacy in federated learning. In *AAAI Conference on Artificial Intelligence*, pages 5859–5866, 2020. 6

[27] Abhishek Singh, Praneeth Vepakomma, Vivek Sharma, and Ramesh Raskar. Posthoc privacy guarantees for collaborative inference with modified propose-test-release. *Advances in Neural Information Processing Systems*, 36:26438–26451, 2023. 2

[28] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 9

[29] Zheng Wang, Xiaoliang Li, Dongqing Yuan, Longxiang Gao, and Laurence T Yang. Personalized privacy preservation with multiple objectives in federated learning. *IEEE Transactions on Industrial Informatics*, 2023. 6

[30] Zhiqin Yang, Yonggang Zhang, Yu Zheng, Xinmei Tian, Hao Peng, Tongliang Liu, and Bo Han. Fedfed: Feature distillation against data heterogeneity in federated learning. *Advances in Neural Information Processing Systems*, 36, 2024. 5, 9

[31] Rui Ye, Mingkai Xu, Jianyu Wang, Chenxin Xu, Siheng Chen, and Yanfeng Wang. Feddisco: Federated learning with discrepancy-aware collaboration. In *International Conference on Machine Learning*, pages 39879–39902. PMLR, 2023. 9

[32] Jianxin Zhang, Chen Chen, Bo Li, Liehuang Wu, and Mingzhi Li. Adaptive differential privacy for federated learning. In *IEEE International Conference on Computer Communications*, pages 1–10, 2022. 6

[33] Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10174–10183, 2022. 9

[34] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018. 9

[35] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12878–12889. PMLR, 2021. 1, 9