

A ADDITIONAL RESULTS

A.1 PROPORTION OF SENSITIVE DATA

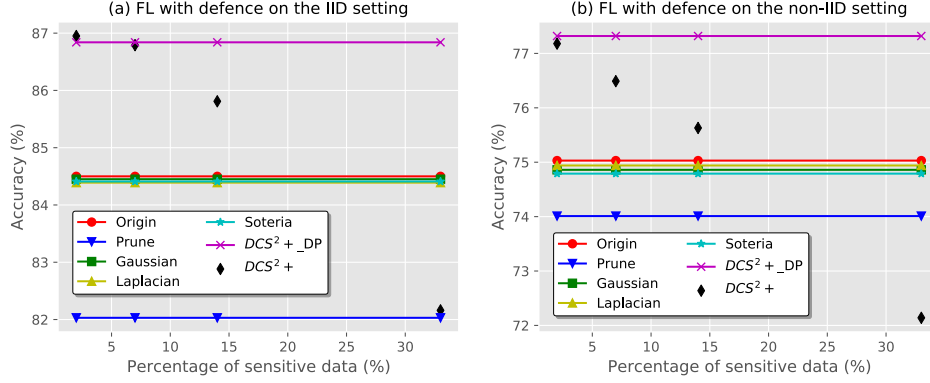


Figure 3: FL performance with defenses under the different proportion of sensitive data.

Figure 3 illustrates the performance of FL with defenses against DLG and GS attacks for varying proportions of the sensitive data. The attack and the defence settings are identical to those in Table 1. When the proportion of sensitive data is less than 30%, our method can even improve the performance of FL compared to other defenses. Combining our defence with differential privacy and adding Gaussian noise offers the best protection for all data points, and shows the highest FL performance.

A.2 COMPARISON WITH OTHER DEFENCES AND ATTACKS

Table 5: Performance on CelebA gender classification.

Defense	GGL		Imprint		FL
	PSNR↓	SSIM↓	PSNR↓	SSIM↓	Non-IID↑
-	11.59	0.27	155.79	1.00	74.69
Prune	10.56	0.24	140.86	1.00	78.36
Gaussian	10.56	0.24	38.64	0.99	78.71
Laplacian	10.69	0.22	36.87	0.98	76.68
Soteria	11.50	0.27	155.79	1.00	77.95
DCS ² + (ours)	8.27	0.16	19.13	0.72	79.85

Table 6: Compared with PRECODE and ATS on CIFAR10.

Defence	GS		Imprint		FL	
	PSNR↓	SSIM↓	PSNR↓	SSIM↓	IID↑	Non-IID↑
PRECODE	5.06	0.07	120.02	0.85	67.06	28.44
ATS	17.50	0.47	49.97	0.48	60.46	37.89
DCS ² + (ours)	6.43	0.11	14.14	0.49	68.05	41.86

We follow the settings in (Li et al., 2022) to compare with Generative Gradient Leakage (GGL) attack (Li et al., 2022) on CelebA, and the setting in (Balunović et al., 2021) to compare with PRECODE (Scheliga et al., 2022) and ATS (Gao et al., 2021). Results shown in Table 5 and Table 6 suggest that our defence provides the best protection with minimal drop in FL performance.

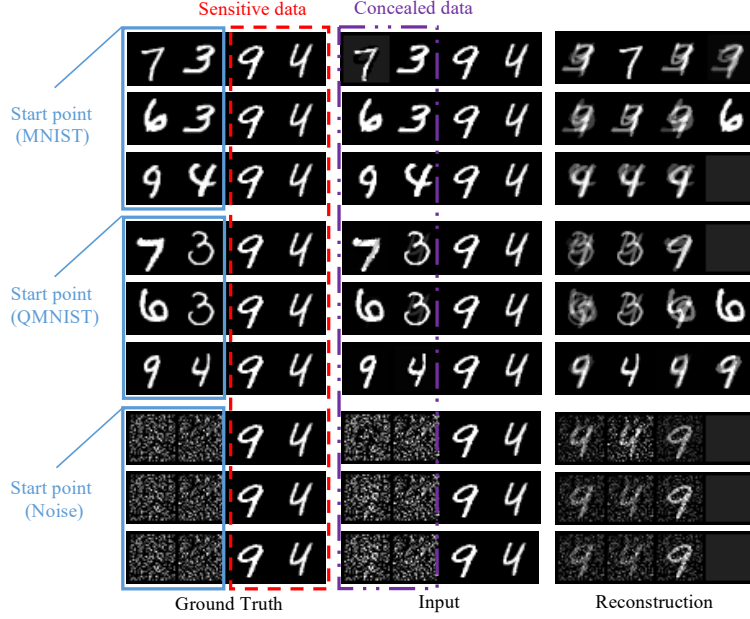


Figure 4: Examples of reconstructions for defending against the Imprint attack on MNIST when using different start points to craft the concealed data. Images in the red dashed box and purple dashed box are the sensitive data and the concealed data, respectively.

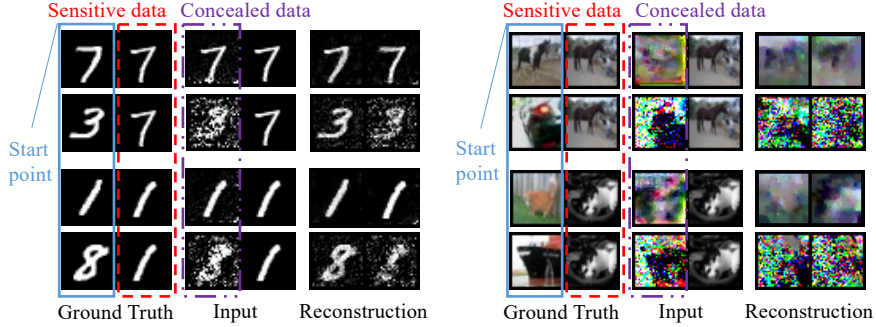


Figure 5: Examples of the reconstructions for defending against the GS attack when concealed data are computed from data with the same or different labels as the sensitive data. Images in the red dashed box and purple dashed box are the sensitive data and the concealed data, respectively.

A.3 EXAMPLES OF DIFFERENT START POINTS

Figure 4 and Figure 5 show examples of the reconstructions for defending against the Imprint attack and GS attack with different start points to craft the concealed data. To obfuscate the gradients from the sensitive data with the concealed data, generating the concealed data with starting point data sampled from different distributions is better than from the identical distributions as the sensitive data. If the start data points for computing the concealed data are sampled from the identical distributions as the sensitive data, they are likely to lie on the same side of the decision boundary and have the same gradient directions. Modifying these data points to approach the sensitive data while being visually dissimilar, therefore, becomes a challenge. As shown in Figure 5 our method cannot withstand the GS attack when the concealed data is modified from the data having the same label as the sensitive data. On the contrary, when the concealed data is crafted from the data having different labels from the sensitive data, our method can effectively defend against attacks.

A.4 ABLATION ANALYSIS ON CONCEALED SAMPLES

Here we vary the number of concealed data points.

Number of concealed data In Table 7 we report the defence results against the GS attack with different number of concealed samples for each sensitive data. The FL performance in the Non-IID setting without defences is about 68.03%, as shown in Table 7 when our defence method uses more concealed data points to imitate the sensitive data, the performance of defence against attacks is better while maintaining the FL performance.

Table 7: Defence against the GS attack with different number of concealed samples on MNIST. k denotes the number of concealed data for each sensitive data and m denotes the number of sensitive data in a mini-batch.

	$k = 1$			$k = 2$			$k = 4$		
	PSNR↓	SSIM↓	FL↑	PSNR↓	SSIM↓	FL↑	PSNR↓	SSIM↓	FL↑
$m = 1$	10.53	0.12	72.46	10.05	0.11	72.40	9.24	0.10	72.46
$m = 2$	10.40	0.09	72.40	10.21	0.09	72.40	9.75	0.09	72.26
$m = 4$	10.31	0.10	72.42	9.99	0.09	72.21	9.15	0.08	71.94

Compute Overhead In Table 8 we report the additional memory required to generate one concealed data in Table 1 and Table 2 using one GeForce RTX 3090 GPU.

Table 8: Overhead for crafting one concealed data against the attacks.

	GS attack		Imprint attack
	MNIST ($28 \times 28 \times 1$)	CIFAR10($32 \times 32 \times 3$)	CelebA($224 \times 224 \times 3$)
Time (s)	+7.7	+25.8	+8.8
Memory (MB)	+50	+90	+1072

A.5 EXAMPLES OF RECONSTRUCTIONS

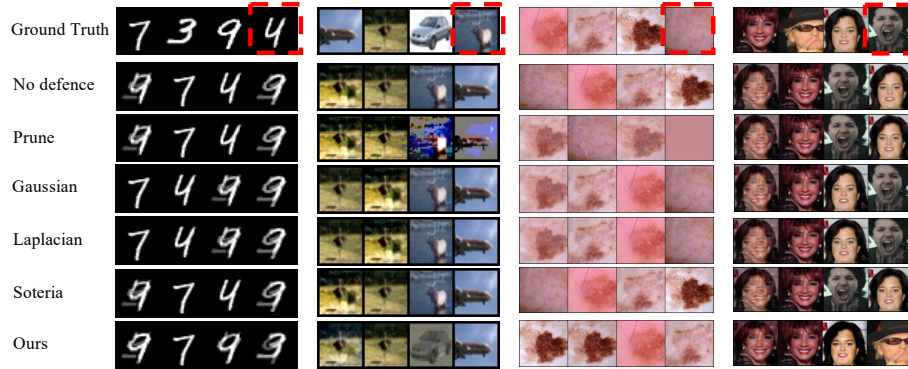


Figure 6: Examples of reconstructions for defending against the Imprint attack on MNIST, CIFAR10, HAM10000 and CelebA, respectively. The parameters for defenses are the same as those in Table 1 and Table 2. Images in red dashed box are sensitive data.

B FURTHER IMPLEMENTATION DETAILS

B.1 PSEUDOCODE OF THE PROPOSED DEFENCE METHOD

Pseudocode 1 Defense by Concealing Sensitive Samples (DCS² and DCS²+)

```

1: procedure GRADIENT OBFUSCATION
2:   initialize the start point for constructing the concealed data  $\mathbf{x}_c \leftarrow \mathbf{x}_0, \mathbf{y}_c \leftarrow \mathbf{y}_0$ ;
3:   get the concealed sample  $\mathbf{x}_c \leftarrow \text{Eq. (6)}$ ;
4:   compute the new gradient  $\mathbf{g}_c \leftarrow \text{Eq. (7)}$ ;
5: procedure GRADIENT PROJECTION
6:   get the gradient from the original batch
    $\mathbf{g} \leftarrow \nabla_{\theta} \mathcal{L}(f_{\theta}(\mathbf{X}), \mathbf{Y})$ ;
7:   if  $\langle \mathbf{g}, \mathbf{g}_c \rangle < 0$  then
8:     get the solution  $\mathbf{v}^* \leftarrow \text{Eq. (9)}$ ;
9:     project the new gradient to the closest gradient  $\hat{\mathbf{g}}_c = \mathbf{g}\mathbf{v}^* + \mathbf{g}_c$ .

```

B.2 MODEL ARCHITECTURES

Table 9: Model architectures for different datasets.

MNIST	CIFAR10	HAM10000 / CelebA
5×5 Conv, 12	5×5 Conv, 32	7×7 Conv, 64
5×5 Conv, 12	$\{5 \times 5 \text{ Conv}, 64\} \times 2$	3×3 MaxPool
5×5 Conv, 12	$\{5 \times 5 \text{ Conv}, 128\} \times 3$	$\left\{ \begin{array}{l} 3 \times 3 \text{ Conv}, 64 \\ 3 \times 3 \text{ Conv}, 64 \end{array} \right\} \times 2$
5×5 Conv, 12	3×3 MaxPool	$\left\{ \begin{array}{l} 3 \times 3 \text{ Conv}, 128 \\ 3 \times 3 \text{ Conv}, 128 \end{array} \right\} \times 2$
FC-10	$\{5 \times 5 \text{ Conv}, 128\} \times 3$	$\left\{ \begin{array}{l} 3 \times 3 \text{ Conv}, 256 \\ 3 \times 3 \text{ Conv}, 256 \end{array} \right\} \times 2$
	3×3 MaxPool	$\left\{ \begin{array}{l} 3 \times 3 \text{ Conv}, 512 \\ 3 \times 3 \text{ Conv}, 512 \end{array} \right\} \times 2$
	FC-10	7×7 AveragePool
		FC-7 (HAM10000) / FC-2 (CelebA)

Details of the models used in this study are shown in Table 9. The activation layers of the model for MNIST are Sigmoid, and for CIFAR10 and HAM10000, CelebA are ReLU.

B.3 PARAMETERS

Table 10: Parameters of different defenses against model inversion attacks.

	MNIST		CIFAR10		HAM/CelebA	CelebA (32)
Defense	DLG/GS	Imprint	DLG/GS	Imprint	Imprint	GGL
Prune	$p = 70\%$	$p = 30\%$	$p = 70\%$	$p = 70\%$	$p = 50\%$	$p = 20\%$
Gaussian	$\sigma = 1e-2$	$\sigma = 1e-2$	$\sigma = 1e-3$	$\sigma = 1e-3$	$\sigma = 1e-1$	$\sigma = 1e-1$
Laplacian	$\sigma = 1e-2$	$\sigma = 1e-2$	$\sigma = 1e-3$	$\sigma = 1e-3$	$\sigma = 1e-1$	$\sigma = 1e-1$
Soteria	$p = 5\%$	$p = 1\%$	$p = 90\%$	$p = 90\%$	-	$p = 10\%$

Attacks and defenses We build on the repository using the official implementation of the DLG, GS, GGL and the Imprint attack methods. For Soteria, Prune and DP, we build on the repository from the study (Sun et al., 2021). For ATS and PRECODE, we build upon the repository from the study (Balunović et al., 2021). Details of parameters can be found in Table 10. We set the mean and the variance of the noise distribution from the defense DP as 0 and σ , respectively. We set the pruning rate of the models’ gradients from the defense Prune and the defense Soteria as p . For our

defense method, we set $\lambda = 0.3, \alpha = 0.1, \beta = 0.001, T = 1000$ when defending against the DLG or GS attack, and $\lambda = 0.3, \alpha = 30.0, \beta = 100.0, T = 100$ when defending against the Imprint attack, $\lambda = 0.3, \alpha = 1.0, \beta = 10.0, T = 1000$ for the GGL attack.

Federated learning We build the Federated Learning (FL) framework based on the Flower (Beutel et al., 2020) platform and the FedAvg (McMahan et al., 2017a) algorithm. The details of the federated learning are shown in Table 11. For the Independent and Identically Distributed (IID) setting, the server randomly selects five from 10 clients in each round. Each client has 2000 samples for MNIST and CIFAR10, 200 samples for HAM10000 randomly sampled from the train set. For the Not Independent and Identically Distributed (Non-IID) setting, the server updates the model using gradients from the ten clients. Each client only has 400 samples for MNIST, and 4000 samples for CIFAR10 with two labels. Each label has 200 samples for MNIST, and 2000 samples for CIFAR10. For HAM10000, each client has 214, 958, 958, 594, 214, 958, 258, 214, 214, 958 samples, respectively in the Non-IID setting. For CelebA, from client 1 to client 10, each one has 170, 190, 109, 210, 151, 174, 209, 194, 235, 193 samples, respectively in the Non-IID setting. And each client has samples from 10 identities, each identity has about 0-30 images. The performance is evaluated on 10,000, 10,000, 1103 and 19962 test samples for MNIST, CIFAR10, HAM10000 and CelebA, respectively. The optimizer is SGD, and the batch size is 256 for MNIST and CIFAR10, 32 for HAM10000 and CelebA for each client, and the maximum number of training rounds is 100. Experiments about the different proportion of sensitive data are evaluated with 128 images for each client, the batch size is 32 and the FL train for 300 rounds.

Table 11: Details of the federated learning on different datasets. $|C|$, $|\tilde{C}|$, $|D_c|$, $|y_c|$ and $|B_c|$ denote the total number of clients, the number of clients selected in each round, the number of training data, the number of labels (identities for CelebA) and the batch size in each client, respectively. η and T denote the learning rate and the number of training rounds, respectively.

	Dataset	$ C $	$ \tilde{C} $	$ D_c $	$ y_c $	$ B_c $	η	T
IID	MNIST	10	5	2,000	10	256	0.01	100
	CIFAR10	10	5	2,000	10	256	0.01	100
Non-IID	MNIST	10	10	400	2	256	0.01	100
	CIFAR10	10	10	4,000	2	256	0.001	100
	HAM10000	10	10	214*958	2	32	0.001	100
	CelebA	10	10	109*235	10	32	0.001	100

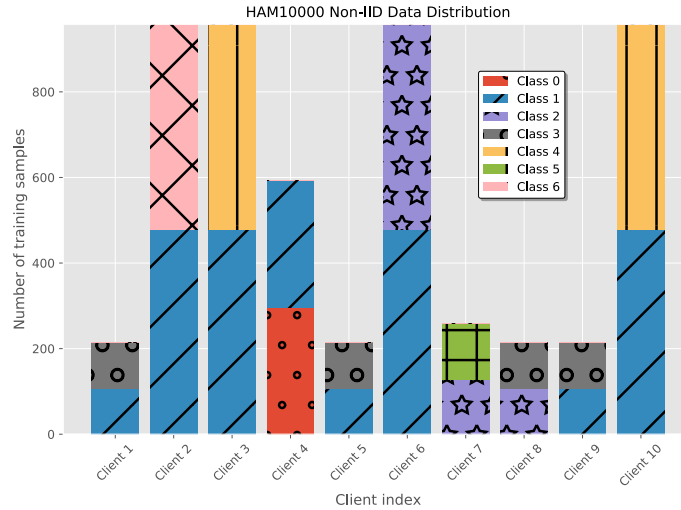


Figure 7: Non-IID data distribution on HAM10000.