A Supplementary material

A.1 Theoretical proof

This section presents all the missing theoretical analyses appeared in the manuscript orderly.

Proposition 1. *For any given input* $\mathbf{x}$ *and shared model* $\mathbf{W}$*, the distance between the recovered data* $\mathbf{x}'$ *and the real data* $\mathbf{x}$ *is bounded by:*

$$||\mathbf{x}-\mathbf{x}'||_2 \geq \frac{||\nabla\mathbf{W}-\mathbf{g}||_2}{||\partial\varphi(\mathbf{x},\mathbf{W})/\partial\mathbf{x}||_2},$$

Proof. By definition, we know

$$\nabla\mathbf{W}-\mathbf{g}=\varphi(\mathbf{x},\mathbf{W})-\varphi(\mathbf{x}',\mathbf{W}). \tag{1}$$

Apply the first-order Taylor expansion to Eq. (1), it is easy to find

$$\begin{aligned}||\nabla\mathbf{W}-\mathbf{g}||_2 &= ||\varphi(\mathbf{x},\mathbf{W})-\varphi(\mathbf{x}',\mathbf{W})||_2\\ &\approx ||(\partial\varphi(\mathbf{x},\mathbf{W})/\partial\mathbf{x})(\mathbf{x}-\mathbf{x}')||_2\\ &\leq ||(\partial\varphi(\mathbf{x},\mathbf{W})/\partial\mathbf{x})||_2||(\mathbf{x}-\mathbf{x}')||_2.\end{aligned}$$

Hence, we have

$$||\mathbf{x}-\mathbf{x}'||_2 \geq \frac{||\nabla\mathbf{W}-\mathbf{g}||_2}{||\partial\varphi(\mathbf{x},\mathbf{W})/\partial\mathbf{x}||_2}.$$

□

Theorem 1. *For any* (ε,δ) *optimization attack, under the presence of* DGP*, it will be degenerated to* $(\varepsilon+\sqrt{\gamma_2}||\nabla\mathbf{W}||_2,\delta)$*-attack if* $\mathcal{D}$ *is measured by Euclidean distance, and degenerated to* $(1-(1-\sqrt{\gamma_2})(1-\varepsilon),\delta)$*-attack if* $\mathcal{D}$ *is measured by cosine distance.*

Proof. If $\mathcal{D}$ is measured by Euclidean distance, by the definition of (ε,δ)-attack, the attacker can achieve the following estimation

$$\mathbb{E}||\nabla\mathbf{W}^*-\nabla\mathbf{W}||_2 \leq \varepsilon,$$

where $\nabla\mathbf{W}^*$ is the attacker's optimized gradients of the ground-truth gradients $\mathbf{W}$. When DGP or ADGP is used, from the bi-Lipschitz assumption (i.e., Assumption 1), we know

$$\sqrt{\gamma_1}||\nabla\mathbf{W}||_2 \leq ||\mathrm{DGP}(\nabla\mathbf{W})-\nabla\mathbf{W}||_2 \leq ||\mathrm{ADGP}(\nabla\mathbf{W})-\nabla\mathbf{W}||_2 \leq \sqrt{\gamma_2}||\nabla\mathbf{W}||_2. \tag{2}$$

Then, when FL aggregation is protected by DGP, the attacker's optimized gradients is based on the observation of DGP($\nabla\mathbf{W}$) and this modified observation will degrade the attacker's capability in optimizing $\nabla\mathbf{W}$ because

$$\begin{aligned}\mathbb{E}||\nabla\mathbf{W}^*-\nabla\mathbf{W}||_2 &= \mathbb{E}||\nabla\mathbf{W}^*-\mathrm{DGP}(\nabla\mathbf{W})+\mathrm{DGP}(\nabla\mathbf{W})-\nabla\mathbf{W}||_2\\ &\leq \varepsilon+||\mathrm{DGP}(\nabla\mathbf{W})-\nabla\mathbf{W}||_2\\ &\leq \varepsilon+\sqrt{\gamma_2}||\nabla\mathbf{W}||_2.\end{aligned}$$

Hence, the first part of this theorem is true.

Similarly, when $\mathcal{D}$ is measured by cosine distance, the definition of (ε,δ)-attack reveals

$$\mathbb{E}\left[1-\frac{<\nabla\mathbf{W}^*,\nabla\mathbf{W}>}{||\nabla\mathbf{W}^*||_2||,\nabla\mathbf{W}||_2}\right]\leq\varepsilon.$$

Then, we can obtain

$$\begin{aligned}\mathbb{E}\left[\frac{<\nabla\mathbf{W}^*,\nabla\mathbf{W}>}{||\nabla\mathbf{W}^*||_2||\nabla\mathbf{W}||_2}\right] &= \mathbb{E}\left[\frac{<\nabla\mathbf{W}^*,\nabla\mathbf{W}-\mathrm{DGP}(\nabla\mathbf{W})+\mathrm{DGP}(\nabla\mathbf{W})>}{||\nabla\mathbf{W}^*||_2||\nabla\mathbf{W}||_2}\right]\\ &\overset{(a)}{=} \mathbb{E}\left[\frac{<\nabla\mathbf{W}^*,\mathrm{DGP}(\nabla\mathbf{W})>}{||\nabla\mathbf{W}^*||_2||\nabla\mathbf{W}||_2}\right]\\ &= \mathbb{E}\left[\frac{<\nabla\mathbf{W}^*,\mathrm{DGP}(\nabla\mathbf{W})>}{||\nabla\mathbf{W}^*||_2||\mathrm{DGP}(\nabla\mathbf{W})||_2}\frac{||\mathrm{DGP}(\nabla\mathbf{W})||_2}{||\nabla\mathbf{W}||_2}\right]\\ &\overset{(b)}{\geq} (1-\sqrt{\gamma_2})\mathbb{E}\left[\frac{<\nabla\mathbf{W}^*,\mathrm{DGP}(\nabla\mathbf{W})>}{||\nabla\mathbf{W}^*||_2||\mathrm{DGP}(\nabla\mathbf{W})||_2}\right]\\ &\geq (1-\sqrt{\gamma_2})(1-\varepsilon),\end{aligned} \tag{3}$$

where (a) is based on the fact that the all non-zero elements of $(\nabla\mathbf{W}-\mathrm{DGP}(\nabla\mathbf{W}))$ are pruned in DGP so $\mathbb{E}(\nabla\mathbf{W}^*,(\nabla\mathbf{W}-\mathrm{DGP}(\nabla\mathbf{W})))=0$, and (b) is the direct application of Eq. (2). Based on Eq. (3), it is easy to conclude

$$\mathbb{E}\left[1-\frac{<\nabla\mathbf{W}^*,\nabla\mathbf{W}>}{||\nabla\mathbf{W}^*||_2||\nabla\mathbf{W}||_2}\right]\le 1-(1-\sqrt{\gamma_2})(1-\varepsilon), \tag{4}$$

which completes the proof. $\square$

Lemma 1. *Let* $\mathbf{e}^t=\sum_{i=1}^N \mathbf{e}_i^t/N$ *be the averaged accumulated error among all users at iteration* t*, the expectation of the norm of* $\mathbf{e}^t$ *is bounded, i.e.,*

$$\mathbb{E}||\mathbf{e}^t||_2^2\le\frac{\gamma_2}{2}(\frac{2+\gamma_2}{1-\gamma_2})^2(G^2+\sigma^2).$$

Proof. To use the theoretical tools of SGD, we set up the following dummy matrix $\mathbf{V}$:

$$\mathbf{V}^{t+1}=\mathbf{V}^t-\eta\nabla\mathbf{W}^t.$$

Since $\mathbf{W}^0=\mathbf{V}^0$, $\mathbf{e}^0=0$, it is easy to find

$$\mathbf{V}^t-\mathbf{W}^t=\eta\mathbf{e}^t. \tag{5}$$

Under Assumption 1, we have

$$\begin{aligned}||\mathbf{X}-\mathrm{ADGP}(\mathbf{X})||_2^2&\le\gamma_2||\mathbf{X}||_2^2,\\ ||\mathbf{X}-\mathrm{DGP}(\mathbf{X})||_2^2&\ge\gamma_1||\mathbf{X}||_2^2.\end{aligned} \tag{6}$$

Under Assumption 3, we have

$$\mathbb{E}||\nabla\mathbf{W}_i^t||_2^2\le G^2+\sigma^2, \tag{7}$$

$$\mathbb{E}||\nabla\mathbf{W}^t||_2^2\le G^2+\frac{\sigma^2}{N}. \tag{8}$$

By definition of $\mathbf{e}^t$, we know

$$||\mathbf{e}^t||_2^2\le\frac{\sum_{i=1}^N||\mathbf{e}_i^t||_2^2}{N},$$

and the $||\mathbf{e}_i^t||_2^2$ is also bounded because

$$\begin{aligned}||\mathbf{e}_i^t||_2^2&=||\nabla\mathbf{W}_i^{t-1}+\mathbf{e}_i^{t-1}-\mathrm{ADGP}(\nabla\mathbf{W}_i^{t-1}+\mathbf{e}_i^{t-1})||_2^2\\ &\overset{(6)}{\le}\gamma_2||\nabla\mathbf{W}_i^{t-1}+\mathbf{e}_i^{t-1}||_2^2\\ &\overset{(c)}{\le}\gamma_2\left((1+\frac{1}{a})||\nabla\mathbf{W}_i^{t-1}||_2^2+(1+a)||\mathbf{e}_i^{t-1}||_2^2\right).\end{aligned}$$

where (c) is based on the variant of Young's inequality $||x+y||_2^2\le(1+a)||x||_2^2+(1+\frac{1}{a})||y||_2^2$. Set $1+a=\frac{2+\gamma_2}{3\gamma_2}$, it is concluded that

$$\mathbb{E}||\mathbf{e}_i^t||_2^2\overset{(8)}{\le}\frac{\gamma_2}{2}(\frac{2+\gamma_2}{1-\gamma_2})^2(G^2+\sigma^2), \tag{9}$$

$$\mathbb{E}||\mathbf{e}^t||_2^2\le\frac{\gamma_2}{2}(\frac{2+\gamma_2}{1-\gamma_2})^2(G^2+\sigma^2). \tag{10}$$

$\square$

Theorem 2. *The averaged norm of the full gradient* $\nabla l(\mathbf{W}^t)$ *derived from centralized training is correlated with the our algorithm as follows:*

$$\frac{\sum_{t=0}^{T-1}\mathbb{E}||\nabla l(\mathbf{W}^t)||_2^2}{T}\le 4\frac{K^0-l^*}{\eta T}+4\eta^2K^2\frac{\gamma_2}{2}(\frac{2+\gamma_2}{1-\gamma_2})^2(G^2+\sigma^2)+2K\eta(G^2+\frac{\sigma^2}{N}),$$

where l^0 *is the initialization of the objective* l*, and* η *is the learning rate.*

Proof. Under Assumption 3, we have

$$||\nabla l(\mathbf{V}^t) - \nabla l(\mathbf{W}^t)|| \leq K||\mathbf{V}^t - \mathbf{W}^t||, \tag{11}$$

and

$$\begin{aligned} l(\mathbf{V}^{t+1}) &\leq l(\mathbf{V}^t) + < \nabla l(\mathbf{V}^t), \mathbf{V}^{t+1} - \mathbf{V}^t > + \frac{K}{2}||\mathbf{V}^{t+1} - \mathbf{V}^t||_2^2 \\ &= l(\mathbf{V}^t) - \eta < \nabla l(\mathbf{V}^t), \nabla \mathbf{W}^t > + \frac{K\eta^2}{2}||\nabla \mathbf{W}^t||_2^2. \end{aligned} \tag{12}$$

Taking expectation on both sides of Eq. (12), we can get

$$\begin{aligned} \mathbb{E}(l(\mathbf{V}^{t+1})) &\leq \mathbb{E}(l(\mathbf{V}^t)) - \eta\mathbb{E}(< \nabla l(\mathbf{V}^t), \nabla l(\mathbf{W}^t) >) + \frac{K\eta^2}{2}\mathbb{E}||\nabla \mathbf{W}^t||_2^2 \\ &\overset{(d)}{=} \mathbb{E}(l(\mathbf{V}^t)) - \frac{\eta}{2}\left[\mathbb{E}(||\nabla l(\mathbf{V}^t)||_2^2 + ||\nabla l(\mathbf{W}^t)||_2^2) - \mathbb{E}||\nabla l(\mathbf{V}^t) - \nabla l(\mathbf{W}^t)||_2^2\right] + \frac{K\eta^2}{2}\mathbb{E}||\nabla \mathbf{W}^t||_2^2 \\ &\leq \mathbb{E}(l(\mathbf{V}^t)) - \frac{\eta}{2}\mathbb{E}(||\nabla l(\mathbf{V}^t)||_2^2) + \frac{\eta}{2}\mathbb{E}||\nabla l(\mathbf{V}^t) - \nabla l(\mathbf{W}^t)||_2^2 + \frac{K\eta^2}{2}\mathbb{E}||\nabla \mathbf{W}^t||_2^2 \\ &\overset{(11)}{\leq} \mathbb{E}(l(\mathbf{V}^t)) - \frac{\eta}{2}(\mathbb{E}||\nabla l(\mathbf{V}^t)||_2^2) + \frac{\eta K^2}{2}\mathbb{E}||\mathbf{V}^t - \mathbf{W}^t||_2^2 + \frac{K\eta^2}{2}\mathbb{E}||\nabla \mathbf{W}^t||_2^2 \\ &\overset{(5)}{\leq} \mathbb{E}(l(\mathbf{V}^t)) - \frac{\eta}{2}(\mathbb{E}||\nabla l(\mathbf{V}^t)||_2^2) + \frac{\eta^3 K^2}{2}\mathbb{E}||\mathbf{e}^t||_2^2 + \frac{K\eta^2}{2}\mathbb{E}||\nabla \mathbf{W}^t||_2^2 \\ &\overset{(8)}{\leq} \mathbb{E}(l(\mathbf{V}^t)) - \frac{\eta}{2}(\mathbb{E}||\nabla l(\mathbf{V}^t)||_2^2) + \frac{\eta^3 K^2}{2}\mathbb{E}||\mathbf{e}^t||_2^2 + \frac{K\eta^2}{2}(G^2 + \frac{\sigma^2}{N}), \end{aligned}$$

where (d) is based on the fact $< x, y >= \frac{1}{2}(||x||^2 + ||y||^2 - ||x - y||^2)$. Base on the deduction above, we can further calculate

$$\begin{aligned} \frac{\eta}{2}(\mathbb{E}||\nabla l(\mathbf{V}^t)||_2^2) &\leq \mathbb{E}(l(\mathbf{V}^t)) - \mathbb{E}(l(\mathbf{V}^{t+1})) + \frac{\eta^3 K^2}{2}\mathbb{E}||\mathbf{e}^t||_2^2 + \frac{K\eta^2}{2}(G^2 + \frac{\sigma^2}{N}), \\ (\frac{\sum_0^{T-1}\mathbb{E}||\nabla l(\mathbf{V}^t)||_2^2}{T}) &\leq \frac{2(l^0 - l^*)}{\eta T} + \eta^2 K^2 \mathbb{E}||\mathbf{e}^t||_2^2 + K\eta(G^2 + \frac{\sigma^2}{N}). \end{aligned} \tag{13}$$

According to Eq. (11), it can be found that

$$\begin{aligned} &||\nabla l(\mathbf{W}^t)|| \leq K||\mathbf{V}^t - \mathbf{W}^t|| + ||\nabla l(\mathbf{V}^t)||, \\ &||\nabla l(\mathbf{W}^t)||_2^2 \leq 2K^2||\mathbf{V}^t - \mathbf{W}^t||_2^2 + 2||\nabla l(\mathbf{V}^t)||_2^2. \end{aligned} \tag{14}$$

Combining Eq. (10), Eq. (13) and Eq. (14), it is concluded

$$\begin{aligned} \mathbb{E}||\nabla l(\mathbf{W}^t)||_2^2 &\leq \frac{4(l^0 - l^*)}{\eta T} + 4\eta^2 K^2 \mathbb{E}||\mathbf{e}^t||_2^2 + 2K\eta(G^2 + 2\frac{\sigma^2}{N}) \\ &\leq \frac{4(l^0 - l^*)}{\eta T} + 4\eta^2 K^2 \frac{\gamma_2}{2}(\frac{2 + \gamma_2}{1 - \gamma_2})^2(G^2 + \sigma^2) + 2K\eta(G^2 + 2\frac{\sigma^2}{N}). \end{aligned}$$

Set $\eta = \sqrt{\frac{l^0 - l^*}{KT(\frac{\sigma^2}{N} + G^2)}}$, we have

$$\frac{\sum_0^{T-1}\mathbb{E}||\nabla l(\mathbf{W}^t)||_2^2}{T} \leq 6\sqrt{\frac{l^0 - l^*}{KT(\frac{\sigma^2}{N} + G^2)}} + \mathcal{O}(\frac{1}{T}).$$

Hence, the theorem is true. □

A.2 More experimental results

We run our experiments on balanced datasets, and train models with batchsize=32. We use SGD optimizer with momentum of 0.9, and train LeNet (Zhu) on CIFAR10 and CIFAR100 with decay 5e-4, train VGG13_bn and ResNet18 on CIFAR10 and CIFAR100 with decay 1e-4. In this section, we present more experimental results, such as the computational cost. To make privacy evaluation more comprehensive, we implement gradient inversion attacks with different batches on different datasets. We also set different send rates k and hyperparameters p to observe their effect on privacy protection and accuracy.

A.2.1 Computation cost

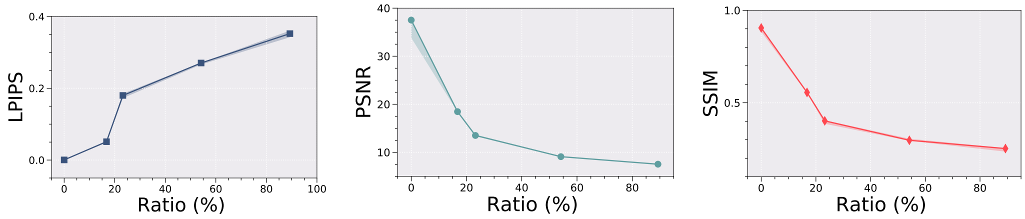
Table 1 shows the computation cost comparison of gradient parameter searching for one iteration. Although the average computation cost of our method is slightly higher than Top-k because we need to search for both large and small parameters, this computation cost is trivial considering the reduced communication cost. And our method is obviously better than Soteria, because Soteria requires a lot of computation on gradients, which leads to expensive computation cost.

Table 1: Comp. cost of gradient perturbation (ms)

Method	Top-k	Soteria	Ours
ResNet18	10.6843	5051.2300	14.7884
VGG13	7.6973	2493.5627	9.7330
LeNet (Zhu)	2.3255	388.1032	3.8247

A.2.2 Effect of relative gradient distance on recovery quality

We give a specific example about Proposition 1. In particular, we plot the recovery results of IVG attack (in terms of LPIPS, PSNR, SSIM metrics) under various relative gradient distance $\frac{||\nabla \mathbf{W}-\mathbf{g}||_2}{||\nabla \mathbf{W}||_2}$ (measured in ratio). As shown in Fig. 1, it is clear experimental results aligns with the analytic results of Proposition 1.



(a) LPIPS (b) PSNR (c) SSIM

Figure 1: Relationship between relative gradient distance $\frac{||\nabla \mathbf{W}-\mathbf{g}||_2}{||\nabla \mathbf{W}||_2}$ and reconstructed data quality under IVG attack, CIFAR10 with LeNet (Zhu).

A.2.3 Defense under attacks with different batch size

In this section, to better evaluate privacy protection, we implement IVG attack and Rob attack with different batches on different datasets. Figs. 2-10 and Table 2 show that our method protect the data privacy against IVG and Rob attacks better than recent works. In particular, our method can comprehensively defend against gradient inversion attacks, while Top-k cannot defend against IVG attack, and Soteria, ATS, Precode cannot defend against Rob attack.

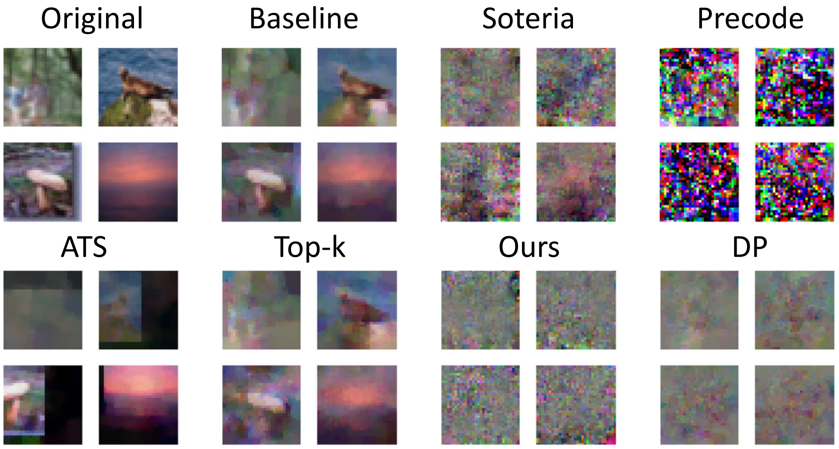


Figure 2: Visualization of the reconstructed data under IVG attack with batchsize=4, CIFAR100 with ResNet18.

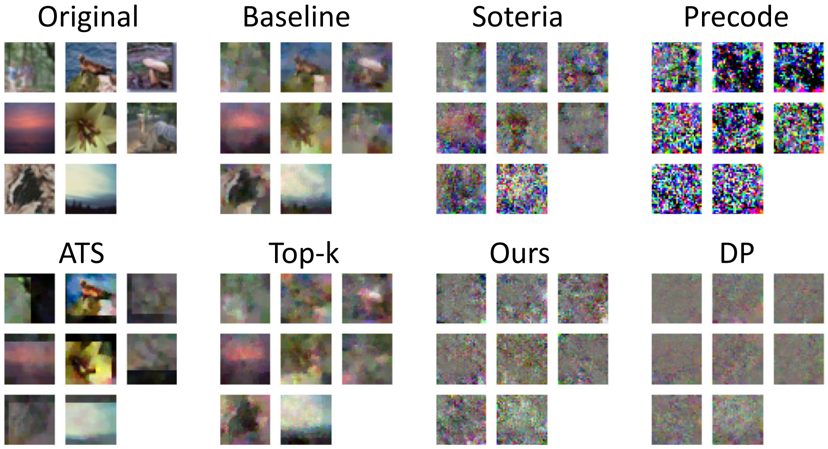


Figure 3: Visualization of the reconstructed data under IVG attack with batchsize=8, CIFAR100 with ResNet18.

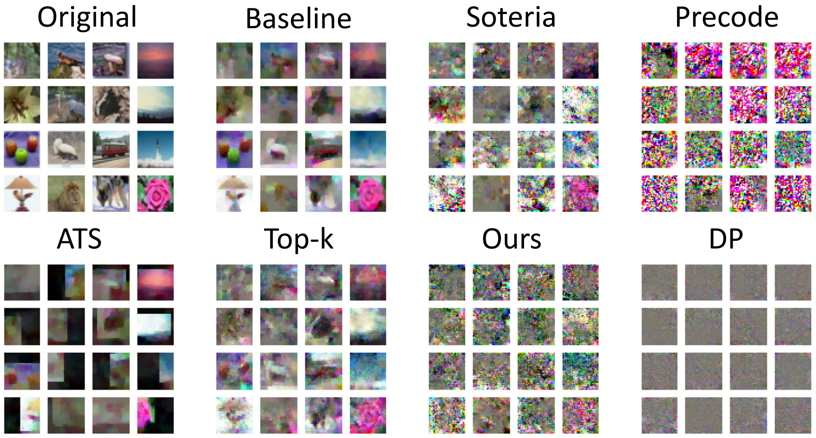


Figure 4: Visualization of the reconstructed data under IVG attack with batchsize=16, CIFAR100 with ResNet18.

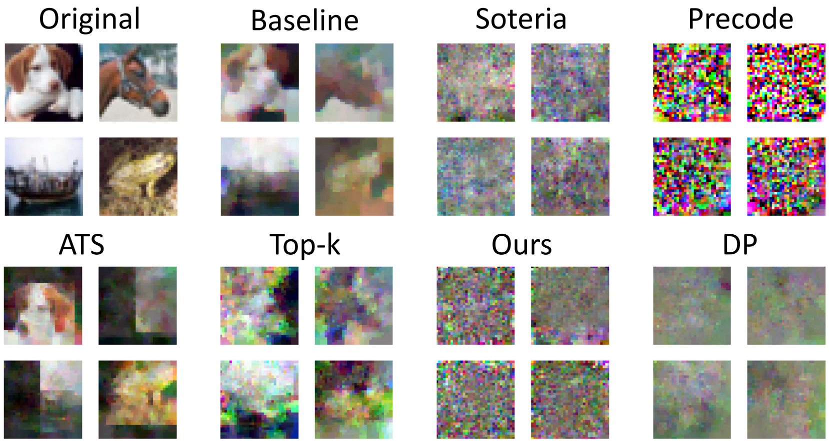


Figure 5: Visualization of the reconstructed data under IVG attack with batchsize=4, CIFAR10 with ResNet18.

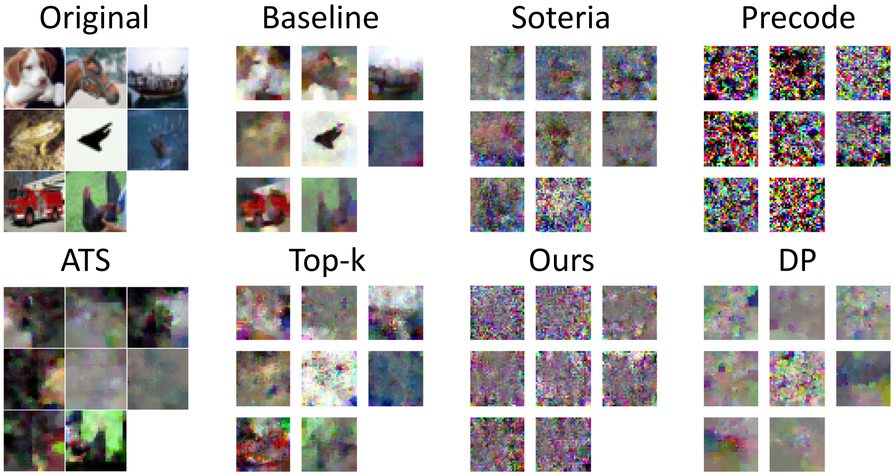


Figure 6: Visualization of the reconstructed data under IVG attack with batchsize=8, CIFAR10 with ResNet18.

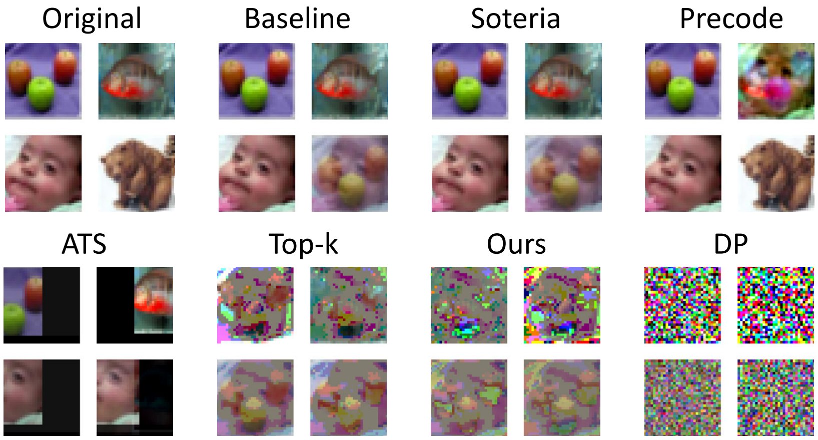


Figure 7: Visualization of the reconstructed data under Rob attack with batchsize=4, CIFAR100 with LeNet (Zhu).

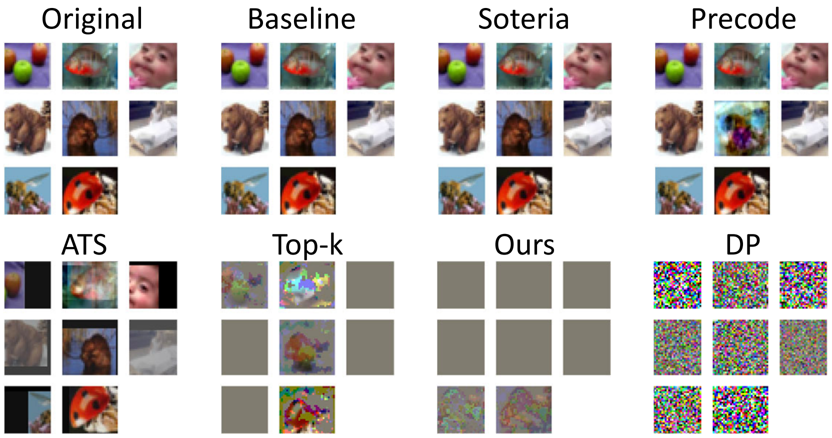


Figure 8: Visualization of the reconstructed data under Rob attack with batchsize=8, CIFAR100 with LeNet (Zhu).

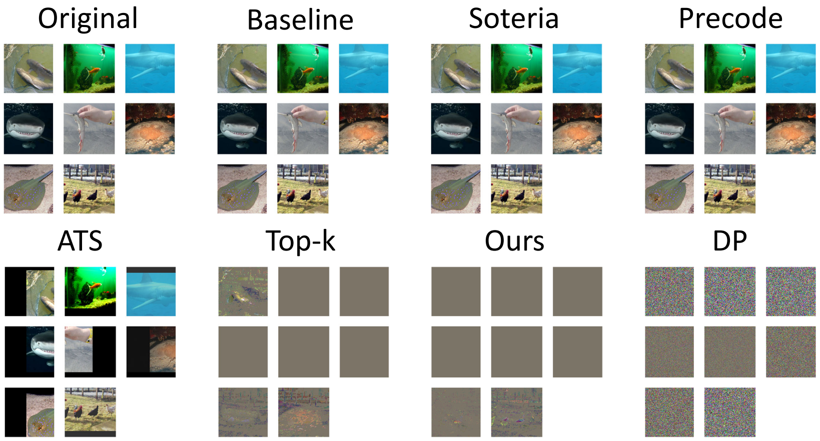


Figure 9: Visualization of the reconstructed data under Rob attack with batchsize=8, ImageNet with ResNet18.

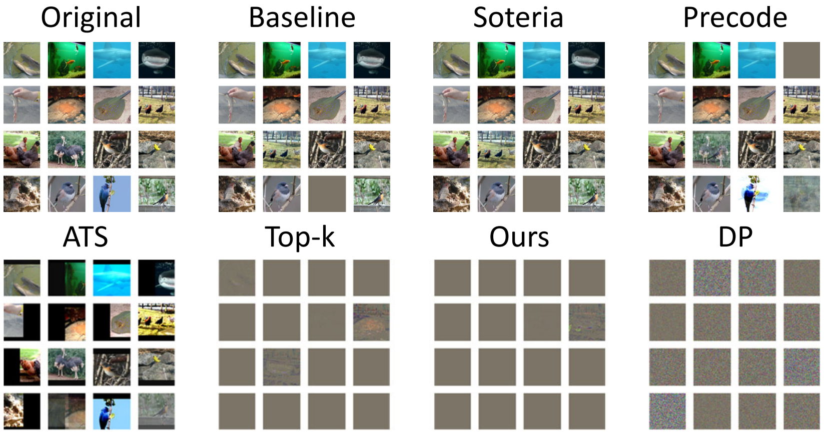


Figure 10: Visualization of the reconstructed data under Rob attack with batchsize=16, ImageNet with ResNet18.

Table 2: Evaluation of defense performance under the Rob attack with varying batchsize and datasets.

Dataset	**Method**	**Baseline**	**Top-k**	**DP**	**Soteria**	**ATS-I**	**ATS-II**	**Precode**	**Ours**
ImageNet	ResNet18, Batchsize= 8								
	PSNR (↓)	136.5906	12.7156	**9.4024**	134.6119	9.5952	112.8973	135.3928	12.7786
	LPIPS (↑)	5.74E-8	0.8469	**1.2870**	4.47E-8	0.6859	0.1099	5.30E-8	0.8970
	SSIM (↓)	1.0000	0.1062	0.2055	1.0000	0.2229	0.8709	1.0000	**0.0527**
	Best SSIM (↓)	1.0000	0.3266	**0.2485**	1.0000	0.2791	1.0000	1.0000	0.2499
	ResNet18, Batchsize= 16								
	PSNR (↓)	102.8838	13.0685	**8.7491**	101.7651	9.6166	115.9886	109.6553	13.0804
	LPIPS (↑)	0.0960	0.8920	**1.3434**	0.0960	0.6410	0.0486	0.1488	0.9184
	SSIM (↓)	0.8969	0.0428	0.2064	0.8969	0.2545	0.9490	0.8440	**0.0229**
	Best SSIM (↓)	1.0000	0.2665	0.2602	1.0000	0.3590	1.0000	1.0000	**0.2478**
CIFAR100	LeNet (Zhu), Batchsize= 8								
	PSNR (↓)	148.6047	12.0199	**8.9393**	146.8654	9.5952	128.5156	115.3705	11.7075
	LPIPS (↑)	1.73E-13	0.4469	0.4605	3.24E-12	**0.6859**	0.0178	0.0312	0.4645
	SSIM (↓)	1.0000	0.1366	0.2280	1.0000	0.2229	0.9303	0.9168	**0.1246**
	Best SSIM (↓)	1.0000	0.4138	**0.2494**	1.0000	0.2791	1.0000	1.0000	0.3921
	LeNet (Zhu), Batchsize= 16								
	PSNR (↓)	136.2503	12.1314	**8.2597**	136.9068	9.6166	115.9886	118.5187	12.1234
	LPIPS (↑)	0.0105	0.4390	0.4249	0.0161	**0.6410**	0.0486	0.0479	0.4821
	SSIM (↓)	0.9644	0.1240	0.2436	0.9587	0.2545	0.9490	0.8732	**0.0782**
	Best SSIM (↓)	1.0000	0.4878	**0.2942**	1.0000	0.3590	1.0000	1.0000	0.3056

A.2.4 More experiments under different pruning rates

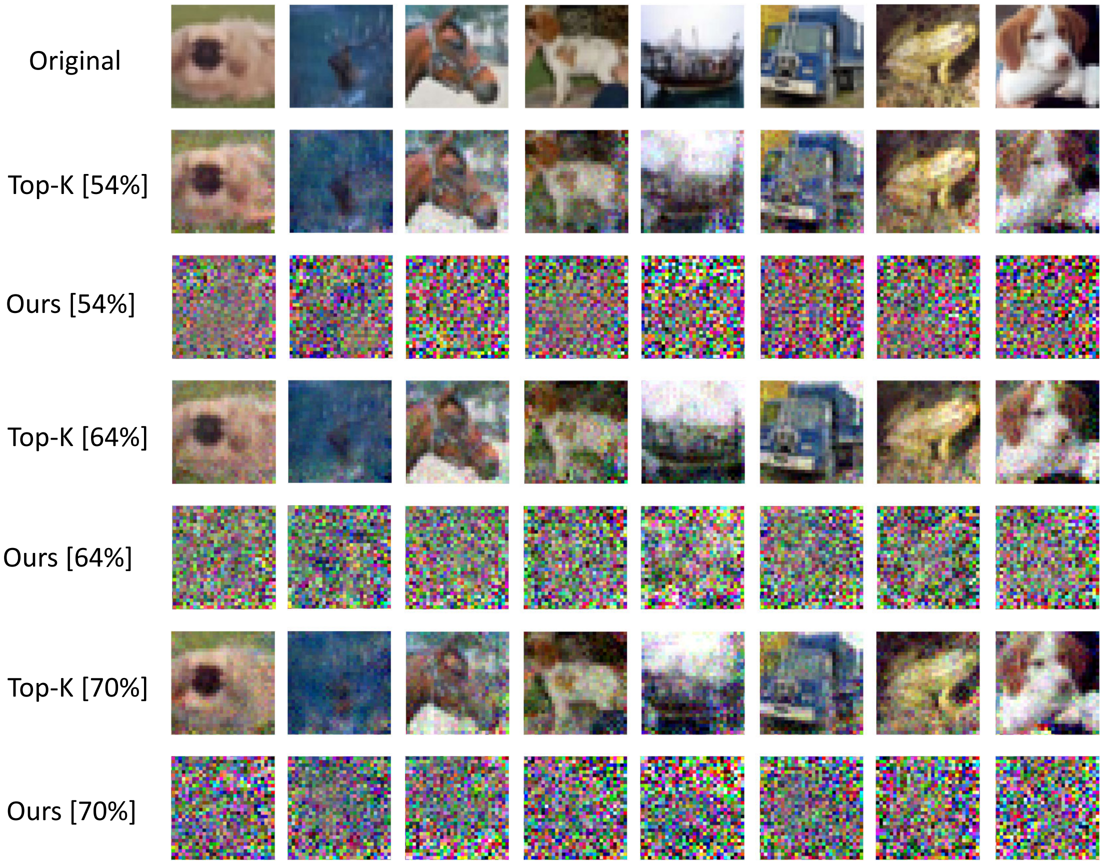


Figure 11: Visualization of the reconstructed data under IVG attack with different pruning rates, CIFAR10 with ResNet18.

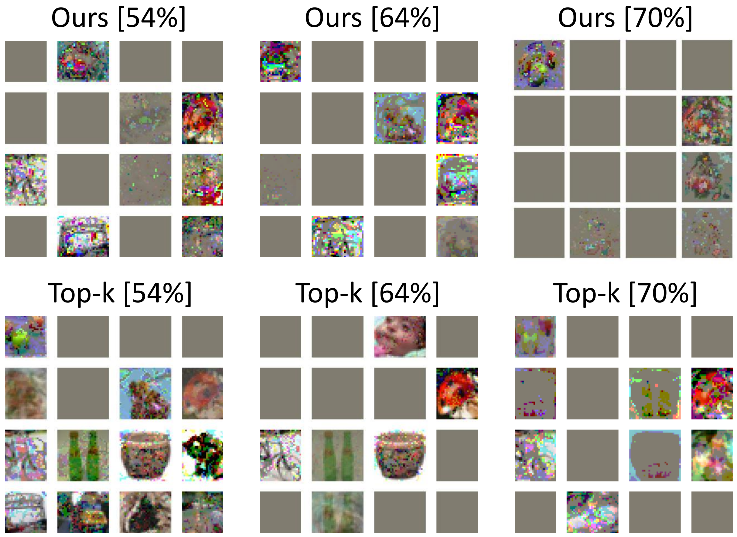
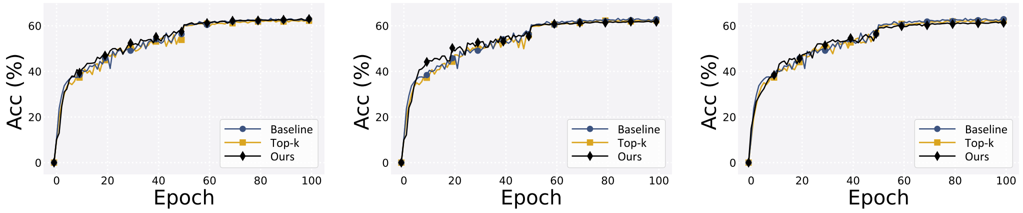


Figure 12: Visualization of the reconstructed data under Rob attack with different pruning rates, CIFAR100 with LeNet (Zhu).



(a) Pruning rate: 54% (b) Pruning rate: 64% (c) Pruning rate: 70%

Figure 13: The effect of different pruning rates on accuracy with LeNet (Zhu) on CIFAR10.

We select different sending rates k=46%, 36%,30%, *i.e.*, the pruning rates (100%-k)=54%, 64%, 70%, and evaluate the privacy protection and model performance of different pruning rates under IVG attack and Rob attack. Figs. 11-12 show that a high pruning rate is more privacy-preserving. According to Fig. 13, it can be found that the effect of high pruning rates on accuracy is not obvious under the correction of the error feedback mechanism.

A.2.5 More experiments under different p

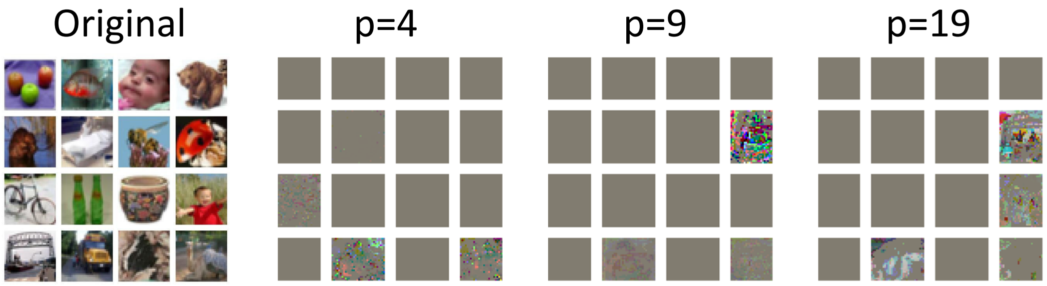


Figure 14: Visualization of the reconstructed data under Rob attack with different hyperparameter p, batchsize=16, CIFAR100 with LeNet (Zhu).

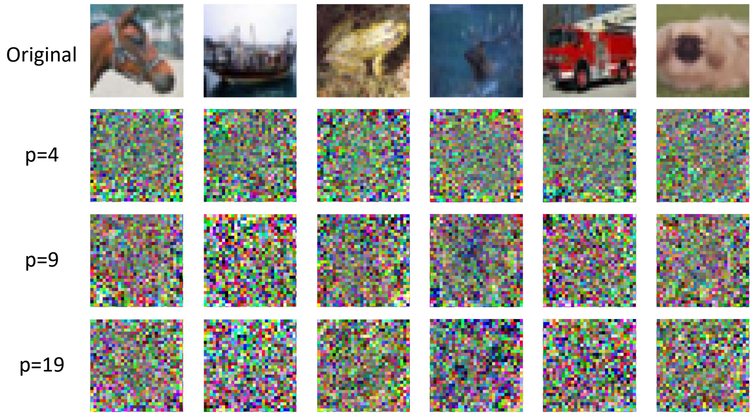
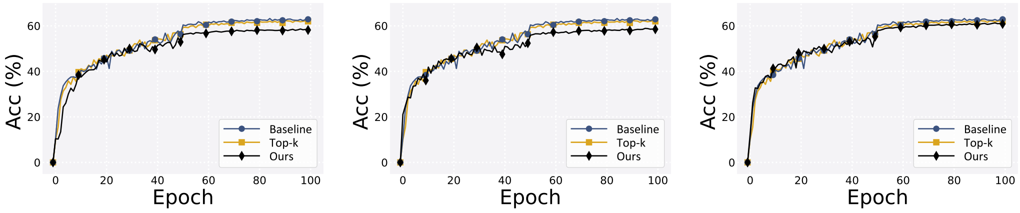


Figure 15: Visualization of the reconstructed data under IVG attack with different hyperparameter p, batchsize=1, CIFAR10 with ResNet18.



(a) p=4 (b) p=9 (c) p=19

Figure 16: The effect of different p on accuracy with LeNet (Zhu) on CIFAR10.

As show in Figs. 15-16, we set the hyperparameter p as 4, 9, 19 and evaluate privacy protection and accuracy under this settings. It is easy to find that p is a trade-off between privacy and accuracy, which will reduce the model accuracy when enhancing privacy.

A.3 System model

As shown in Fig. 17, ADGP is achieved by randomly selecting a user, who broadcasts binary matrix $\mathcal{I}$ to all other users. Each user then only transmits gradient parameters whose locations belong to $\mathcal{I}$.

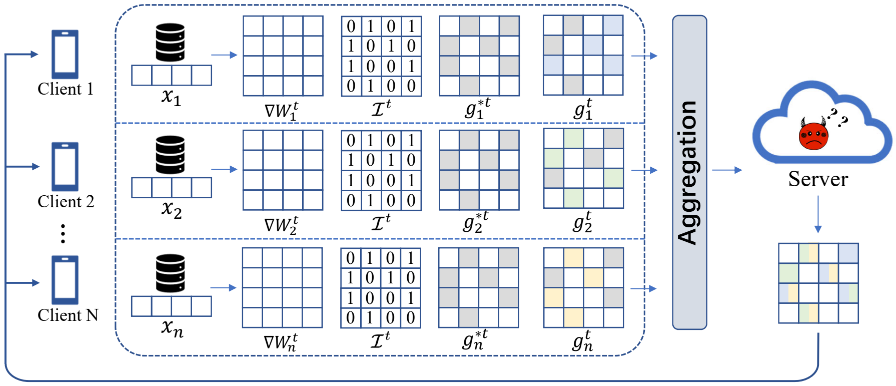


Figure 17: The t-th iteration model update process, where $\mathbf{g}^*$ represents the gradient parameters whose position belong to $\mathcal{I}$.