Appendix

Our Appendix is organized as follows: In section A, we complete the proof of Theorem 1. In section B, we detail the experimental setup not covered in our main paper. In section C, we show our experimental results with the Adam optimizer. In section D, we present our experiments on additional types of non-IID data splits. In section E, we provide additional defense results. In section F, we report the impact of local steps E. In section G, we present more gradient matching loss analysis.

A. Proof of Theorem 1

We analyze FedAvg with uniformly distributed random local learning rates in this section. Motivated by Li *et al.* [7], we first define several additional notations. Let W_t^k be the model parameter maintained in the $k-th$ client at the $t-th$ step and I_E be the set of communication steps, i.e., $I_E = \{nE \mid n = 1, 2, \cdots\}$. Then the update with full clients active can be described as

$$V_{t+1}^k = W_t^k - \eta_t^k \nabla \mathcal{L}_k(W_t^k, \xi_t^k), \tag{15}$$

$$W_{t+1}^k = \begin{cases} V_{t+1}^k & \text{if } t+1 \notin I_E, \\ \sum_{k=1}^N p_k V_{t+1}^k & \text{if } t+1 \in I_E, \end{cases} \tag{16}$$

where V_{t+1}^k denotes the immediate result of one step SGD update from W_t^k.

Besides the randomness of (1) learning rate perturbation introduced by our defense, there are two sources of randomness in FedAvg: (2) stochastic gradients and (3) random sampling of clients. We use the notations $\mathbb{E}_{\eta_t^k}(\cdot)$, $\mathbb{E}_{\xi_t^k}(\cdot)$, and $\mathbb{E}_{S_t}(\cdot)$ to denote the corresponding expectations.

We also define two virtual sequences $\overline{V}_t = \sum_{k=1}^N p_k V_t^k$ and $\overline{W}_t = \sum_{k=1}^N p_k W_t^k$. In addition, we define $\overline{\mathcal{G}}_t = \sum_{k=1}^N p_k \eta_t \nabla \mathcal{L}_k(W_t^k)$ and $\mathcal{G}_t = \sum_{k=1}^N p_k \eta_t^k \nabla \mathcal{L}_k(W_t^k, \xi_t^k)$. ξ_t^k and η_t^k are independent with each other. We have $\overline{V}_{t+1} = \overline{W}_t - \mathcal{G}_t$ and $\mathbb{E}(\mathcal{G}_t) = \overline{\mathcal{G}}_t$.

A.1. Lemmas

Here, we provide several Lemmas which are useful for our proof. The proof of this Lemmas can be found in A.3.

Lemma 1. *Assume Assumption 1 and 2. If $\eta_t \le \frac{1}{4L}$, we have*

$$\begin{aligned} \mathbb{E}\left\|\overline{V}_{t+1} - W^\star\right\|^2 \le & (1-\eta_t\mu)\mathbb{E}\left\|\overline{W}_t - W^\star\right\|^2 \\ & + \mathbb{E}\left\|\mathcal{G}_t - \overline{\mathcal{G}}_t\right\|^2 + 6L\eta_t^2\Gamma \\ & + 2\mathbb{E}\sum_{k=1}^N p_k \left\|\overline{W}_t - W_k^t\right\|^2. \end{aligned}$$

Lemma 2. *Assume Assumption 3 holds. It follows that*

$$\mathbb{E}\left\|\mathcal{G}_t - \overline{\mathcal{G}}_t\right\|^2 \le \eta_t^2 \sum_{k=1}^N p_k^2(\sigma_k^2 + \frac{1}{3}G^2).$$

Lemma 3. *Assume Assumption 4, that η_t is non-increasing and $\eta_t \le 2\eta_{t+E}$ for all $t \ge 0$. It follows that*

$$\mathbb{E}\left[\sum_{k=1}^N p_k \left\|\overline{W}_t - W_k^t\right\|^2\right] \le 16\eta_t^2(E-1)^2G^2.$$

A.2. Details of Convergence Analysis

Here, we first generate the convergence guarantee for the case when full clients participate. A similar proof can be found in **A.3** of [7], except for the difference of constant B. For coherence of the proof, we show the detailed procedure. When full clients participate, we have $\overline{W}_{t+1} = \overline{V}_{t+1}$. Let $\Delta_t = \mathbb{E}\left\|\overline{W}_t - W^\star\right\|^2$. Assuming Lemma 1, Lemma 2 and Lemma 3 hold, it follows that

$$\Delta_{t+1} \le (1-\eta_t\mu)\Delta_t + \eta_t^2 B, \tag{17}$$

where

$$B = \sum_{k=1}^N p_k^2(\sigma_k^2 + \frac{1}{3}G^2) + 6L\Gamma + 32(E-1)^2G^2.$$

For a learning rate expectation, $\eta_t = \frac{\beta}{t+\gamma}$ for some $\beta > \frac{1}{\mu}$ and $\gamma > 0$ such that $\eta_1 \le \min\{\frac{1}{\mu}, \frac{1}{4L}\} = \frac{1}{4L}$ and $\eta_t \le 2\eta_{t+E}$. It can be proved by induction that $\Delta_t \le \frac{v}{\gamma+t}$ where $v = \max\left\{\frac{\beta^2 B}{\beta\mu - 1}, (\gamma+1)\Delta_1\right\}$.

Firstly, $\Delta_t \le \frac{v}{\gamma+t}$ holds for $t = 1$. Assume it also holds for some t, it follows that

$$\begin{aligned} \Delta_{t+1} &\le (1-\eta_t\mu)\Delta_t + \eta_t^2 B \\ &\le \left(1 - \frac{\beta\mu}{t+\gamma}\right)\frac{v}{t+\gamma} + \frac{\beta^2 B}{(t+\gamma)^2} \\ &= \frac{t+\gamma-1}{(t+\gamma)^2}v + \left[\frac{\beta^2 B}{(t+\gamma)^2} - \frac{\beta\mu - 1}{(t+\gamma)^2}v\right] \\ &\le \frac{v}{t+\gamma+1}. \end{aligned}$$

Then by the L-smoothness of $\mathcal{L}(\cdot)$,

$$\mathbb{E}[\mathcal{L}(\overline{W}_t)] - \mathcal{L}^* \le \frac{L}{2}\Delta_t \le \frac{L}{2}\frac{v}{\gamma+t}.$$

Specifically, if we choose $\beta = \frac{2}{\mu}, \gamma = \max\{8\frac{L}{\mu}, E\} - 1$ and denote $\kappa = \frac{L}{\mu}$, then $\eta_t = \frac{2}{\mu}\frac{1}{\gamma+t}$ such that η_t satisfies $\eta_t \le 2\eta_{t+E}$ for $t \ge 1$. Then, we have

$$\begin{aligned} v &= \max\left\{\frac{\beta^2 B}{\beta\mu - 1}, (\gamma+1)\Delta_1\right\} \\ &\le \frac{\beta^2 B}{\beta\mu - 1} + (\gamma+1)\Delta_1 \le \frac{4B}{\mu^2} + (\gamma+1)\Delta_1, \end{aligned}$$

and

$$\mathbb{E}[\mathcal{L}(\overline{W}_t)] - \mathcal{L}^* \leq \frac{L}{2}\frac{v}{\gamma + t} \leq \frac{\kappa}{\gamma + t}\left(\frac{2B}{\mu} + \frac{\mu(\gamma+1)}{2}\Delta_1\right).$$

Partial Client Participation. For full clients participation, we always have $\overline{W}_{t+1} = \overline{V}_{t+1}$, while it does not always hold when partial clients participate. We need another Lemma to complete our proof.

Lemma 4. *Assume Assumption 5 holds, it follows that*

$$\mathbb{E}_{S_t}\left\|\overline{V}_{t+1} - \overline{W}_{t+1}\right\|^2 \leq \frac{N-K}{N-1}\frac{16}{K}\eta_t^2 E^2 G^2.$$

When partial clients participate,

$$\begin{aligned}\left\|\overline{W}_{t+1} - W^*\right\|^2 &= \left\|\overline{W}_{t+1} - \overline{V}_{t+1} + \overline{V}_{t+1} - W^*\right\|^2 \\ &= \underbrace{\left\|\overline{W}_{t+1} - \overline{V}_{t+1}\right\|^2}_{A_1} + \underbrace{\left\|\overline{V}_{t+1} - W^*\right\|^2}_{A_2} \\ &\quad + \underbrace{2\langle\overline{W}_{t+1} - \overline{V}_{t+1}, \overline{V}_{t+1} - W^*\rangle}_{A_3}.\end{aligned}$$

Assume Assumption 5 holds, $\mathbb{E}_{S_{t+1}}\overline{W}_{t+1} = \overline{V}_{t+1}$. Thus, A_3 vanishes when we take the expectation $\mathbb{E}_{S_{t+1}}(\cdot)$. And from Lemma 4, it follows that

$$\begin{aligned}\mathbb{E}\left\|\overline{W}_{t+1} - W^*\right\|^2 &= \mathbb{E}(\left\|\overline{W}_{t+1} - \overline{V}_{t+1}\right\|^2 + \left\|\overline{V}_{t+1} - W^*\right\|^2) \\ &\leq \mathbb{E}\left\|\overline{V}_{t+1} - W^*\right\|^2 + \frac{N-K}{N-1}\frac{16}{K}\eta_t^2 E^2 G^2.\end{aligned}$$

We use Lemma 1, Lemma 2, and Lemma 3. Then

$$\begin{aligned}\mathbb{E}\left\|\overline{W}_{t+1} - W^*\right\|^2 \leq &(1 - \eta_t\mu)\mathbb{E}\left\|\overline{W}_t - W^*\right\|^2 \\ &+ \eta_t^2(B + C),\end{aligned} \tag{18}$$

where

$$C = \frac{N-K}{N-1}\frac{16}{K}\eta_t^2 E^2 G^2.$$

Since the difference between eqn. (18) and eqn. (17) is the constant C, we can apply the same porcess with case where full clients participate and obtain

$$\mathbb{E}[\mathcal{L}(\overline{W}_t)] - \mathcal{L}^* \leq \frac{\kappa}{\gamma + t}\left(\frac{2B + C}{\mu} + \frac{\mu(\gamma+1)}{2}\Delta_1\right),$$

where

$$\Delta_1 = \mathbb{E}\left\|\overline{W}_1 - W^*\right\|^2.$$

A.3. Proofs of Lemmas

Proof of Lemma 1. Since $\overline{V}_{t+1} = \overline{W}_t - \mathcal{G}_t$,

$$\begin{aligned}\left\|\overline{V}_{t+1} - W^\star\right\|^2 &= \left\|\overline{W}_t - \mathcal{G}_t - W^\star - \overline{\mathcal{G}}_t + \overline{\mathcal{G}}_t\right\|^2 \\ &= \underbrace{\left\|\overline{W}_t - W^\star - \overline{\mathcal{G}}_t\right\|^2}_{A_1} + \left\|\mathcal{G}_t - \overline{\mathcal{G}}_t\right\|^2 \\ &\quad + \underbrace{2\left\langle\overline{W}_t - W^\star - \overline{\mathcal{G}}_t, \overline{\mathcal{G}}_t - \mathcal{G}_t\right\rangle}_{A_2}\end{aligned}$$

Notice that $\mathbb{E}(\overline{\mathcal{G}}_t - \mathcal{G}_t) = 0$, so that $\mathbb{E}A_2 = 0$. By applying **eqn. (18)** in [7], we obtain

$$\begin{aligned}A_1 \leq &(1 - \eta_t\mu)\left\|\overline{W}_t - W^\star\right\|^2 \\ &+ 6L\eta_t^2\Gamma + 2\sum_{k=1}^N p_k\left\|\overline{W}_t - W_k^t\right\|^2.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left\|\overline{V}_{t+1} - W^\star\right\|^2 &= \mathbb{E}(A_1 + A_2 + \left\|\mathcal{G}_t - \overline{\mathcal{G}}_t\right\|^2) \\ &\leq (1 - \eta_t\mu)\mathbb{E}\left\|\overline{W}_t - W^\star\right\|^2 \\ &\quad + \mathbb{E}\left\|\mathcal{G}_t - \overline{\mathcal{G}}_t\right\|^2 + 6L\eta_t^2\Gamma \\ &\quad + 2\mathbb{E}\sum_{k=1}^N p_k\left\|\overline{W}_t - W_k^t\right\|^2.\end{aligned}$$

Proof of Lemma 2. We start from

$$\begin{aligned}&\mathbb{E}\left\|\mathcal{G}_t - \overline{\mathcal{G}}_t\right\|^2 \\ &= \mathbb{E}\left\|\sum_{k=1}^N p_k(\eta_t^k\nabla\mathcal{L}_k(W_t^k, \xi_t^k) - \eta_t\nabla\mathcal{L}_k(W_t^k))\right\|^2 \\ &= \sum_{k=1}^N p_k^2\mathbb{E}\left\|(\eta_t^k\nabla\mathcal{L}_k(W_t^k, \xi_t^k) - \eta_t\nabla\mathcal{L}_k(W_t^k))\right\|^2 \\ &= \sum_{k=1}^N p_k^2(\mathbb{E}\left\|\eta_t^k\nabla\mathcal{L}_k(W_t^k, \xi_t^k)\right\|^2 \\ &\quad + \mathbb{E}\left\|\eta_t\nabla\mathcal{L}_k(W_t^k)\right\|^2 \\ &\quad - 2\mathbb{E} < \eta_t^k\nabla\mathcal{L}_k(W_t^k, \xi_t^k), \eta_t\nabla\mathcal{L}_k(W_t^k) >).\end{aligned}$$

Since ξ_t^k and η_t^k are independent with each other. We first erase the randomness of η_t^k and obtain

$$\begin{aligned}\mathbb{E}\left\|\mathcal{G}_t - \overline{\mathcal{G}}_t\right\|^2 &= \sum_{k=1}^N p_k^2(\frac{4}{3}\mathbb{E}_{\xi_t^k}\left\|\eta_t\nabla\mathcal{L}_k(W_t^k, \xi_t^k)\right\|^2 \\ &\quad + \mathbb{E}_{\xi_t^k}\left\|\eta_t\nabla\mathcal{L}_k(W_t^k)\right\|^2 \\ &\quad - 2\mathbb{E}_{\xi_t^k} < \eta_t\nabla\mathcal{L}_k(W_t^k, \xi_t^k), \eta_t\nabla\mathcal{L}_k(W_t^k) >) \\ &= \sum_{k=1}^N p_k^2\mathbb{E}_{\xi_t^k}(\frac{1}{3}\left\|\eta_t\nabla\mathcal{L}_k(W_t^k, \xi_t^k)\right\|^2 \\ &\quad + \left\|\eta_t(\nabla\mathcal{L}_k(W_t^k, \xi_t^k) - \nabla\mathcal{L}_k(W_t^k))\right\|^2).\end{aligned}$$

Assume Assumptions 3 and 4 hold, then

$$\mathbb{E}\left\|\mathcal{G}_t - \overline{\mathcal{G}}_t\right\|^2 \le \eta_t^2 \sum_{k=1}^{N} p_k^2(\sigma_k^2 + \frac{1}{3}G^2).$$

Proof of Lemma 3. In a local training round, for any $t \ge 0$, there exists a $t_0 \le t$, such that $t - t_0 \le E - 1$ and $W_{t_0}^k = \overline{W}_{t_0}$ for all $k = 1, 2, \cdots, N$. When the learning rate expectation η_t is non-increasing and $\eta_{t_0} \le 2\eta_t$ for all $t - t_0 \le E - 1$, then

$$\begin{aligned}
&\mathbb{E}\sum_{k=1}^{N} p_k \left\|\overline{W}_t - W_t^k\right\|^2 \\
&= \mathbb{E}\sum_{k=1}^{N} p_k \left\|(W_t^k - \overline{W}_{t_0}) - (\overline{W}_t - \overline{W}_{t_0})\right\|^2 \\
&\le \mathbb{E}\sum_{k=1}^{N} p_k \left\|(W_t^k - \overline{W}_{t_0})\right\|^2 \\
&\le \sum_{k=1}^{N} p_k \mathbb{E}\left\|\sum_{t=t_0}^{t-1} \eta_t^k \nabla\mathcal{L}_k(W_t^k, \xi_t^k)\right\|^2 \\
&\le \sum_{k=1}^{N} p_k \mathbb{E}\sum_{t=t_0}^{t-1} (E-1)(\eta_t^k)^2 \left\|\nabla\mathcal{L}_k(W_t^k, \xi_t^k)\right\|^2 \\
&\le \sum_{k=1}^{N} p_k \mathbb{E}\sum_{t=t_0}^{t-1} 4(E-1)\eta_t^2 \left\|\nabla\mathcal{L}_k(W_t^k, \xi_t^k)\right\|^2 \\
&\le \sum_{k=1}^{N} p_k \sum_{t=t_0}^{t-1} 4(E-1)\eta_t^2 G^2 \\
&\le \sum_{k=1}^{N} p_k 4(E-1)^2\eta_{t_0}^2 G^2 \\
&\le 16(E-1)^2\eta_{t_0}^2 G^2
\end{aligned}$$

Here in the first inequality, we use $\mathbb{E}\|X - \mathbb{E}X\|^2 \le \mathbb{E}\|X\|^2$ where $X = W_t^k - \overline{W}_{t_0}$ with probability p_k. The third inequality is obtained by Jensen inequality:

$$\begin{aligned}
&\left\|\sum_{t=t_0}^{t-1} \eta_t^k \nabla\mathcal{L}_k(W_t^k, \xi_t^k)\right\|^2 \\
&\le \sum_{t=t_0}^{t-1} (t - t_0)(\eta_t^k)^2 \left\|\nabla\mathcal{L}_k(W_t^k, \xi_t^k)\right\|^2.
\end{aligned}$$

In the fourth inequality, we use $\eta_t^k \le 2\eta_t$. In the fifth inequality, we use $\mathbb{E}\left\|\nabla\mathcal{L}_k(W_t^k, \xi_t^k)\right\|^2 \le G^2$ (i.e., Assumption 4). And in the last inequality, we use $\eta_t \le \eta_{t_0} \le 2\eta_{t_0+E} \le 2\eta_t$ for $t_0 \le t \le t_0 + E$.

Proof of Lemma 4. Assume Assumption 5 holds, we let $S_{t+1} = \{i_1, \cdots, i_K\}$ denote the multiset of chosen indexes. Then we have $W_{t+1} = \frac{1}{K}\sum_{l=1}^{K} V_{t+1}^{i_l}$ and

$$\begin{aligned}
&\mathbb{E}_{S_t}\left\|\overline{W}_{t+1} - \overline{V}_{t+1}\right\|^2 \\
&= \mathbb{E}_{S_t}\left\|\frac{1}{K}\sum_{i\in S_{t+1}} V_{t+1}^i - \overline{V}_{t+1}\right\|^2 \\
&= \frac{1}{K^2}\mathbb{E}_{S_t}\left\|\sum_{i=1}^{N} \mathbb{I}\{i \in S_t\}(V_{t+1}^i - \overline{V}_{t+1})\right\|^2 \\
&= \frac{1}{K^2}\left[\sum_{i\in[N]} \mathbb{P}\left(i \in S_{t+1}\right)\left\|V_{t+1}^i - \overline{V}_{t+1}\right\|^2\right] \\
&+ \frac{1}{K^2}\left[\sum_{i\neq j} \mathbb{P}\left(i, j \in S_{t+1}\right)\langle V_{t+1}^i - \overline{V}_{t+1}, V_{t+1}^j - \overline{V}_{t+1}\rangle\right] \\
&= \frac{1}{KN}\sum_{i=1}^{N}\left\|V_{t+1}^i - \overline{V}_{t+1}\right\|^2 \\
&\quad + \sum_{i\neq j}\frac{K-1}{KN(N-1)}\langle V_{t+1}^i - \overline{V}_{t+1}, V_{t+1}^j - \overline{V}_{t+1}\rangle \\
&= \frac{1}{K(N-1)}\left(1 - \frac{K}{N}\right)\sum_{i=1}^{N}\left\|V_{t+1}^i - \overline{V}_{t+1}\right\|^2.
\end{aligned}$$

In the third equality, we use $\mathbb{P}\left(i \in S_{t+1}\right) = \frac{K}{N}$ and $\mathbb{P}\left(i, j \in S_{t+1}\right) = \frac{K(K-1)}{N(N-1)}$ for all $i \neq j$. In the last equality, we use $\sum_{i\in[N]}\left\|V_{t+1}^i - \overline{V}_{t+1}\right\|^2 + \sum_{i\neq j}\langle V_{t+1}^i - \overline{V}_{t+1}, V_{t+1}^j - \overline{V}_{t+1}\rangle = 0$.

Then we bound $\mathbb{E}_{S_t}\left\|V_{t+1}^i - \overline{V}_{t+1}\right\|^2$ using the same argument for proving Lemma 3. For any $t + 1 \ge 0$, there exists a $t_0 \le t + 1$, such that $t + 1 - t_0 \le E - 1$ and $W_{t_0}^k = \overline{W}_{t_0}$ for all $k = 1, 2, \cdots, N$, then

$$\begin{aligned}
\mathbb{E}_{S_t}\left\|V_{t+1}^i - \overline{V}_{t+1}\right\|^2 &= \left\|(V_{t+1}^k - \overline{W}_{t_0}) - (\overline{V}_{t+1} - \overline{W}_{t_0})\right\|^2 \\
&\le \left\|V_{t+1}^k - \overline{W}_{t_0}\right\|^2 \\
&\le \left\|\sum_{t=t_0}^{t} \eta_t^k \nabla\mathcal{L}_k(W_t^k, \xi_t^k)\right\|^2 \\
&\le \sum_{t=t_0}^{t} E(\eta_t^k)^2 \left\|\nabla\mathcal{L}_k(W_t^k, \xi_t^k)\right\|^2 \\
&\le 16\eta_t^2 E^2 G^2.
\end{aligned}$$

The first inequality is obtained from $\mathbb{E}\|X - \mathbb{E}X\|^2 \le \mathbb{E}\|X\|^2$ where $X = V_{t+1}^k - \overline{W}_{t_0}$ with probability p_k. In the last inequality, we use Jensen inequality, $\eta_t^k \le 2\eta_t$, $\mathbb{E}\left\|\nabla\mathcal{L}_k(W_t^k, \xi_t^k)\right\|^2 \le G^2$, and $\eta_t \le \eta_{t_0} \le 2\eta_t$.

Thus, we obtain

$$
\begin{aligned}
&\mathbb{E}\left\|V_{t+1}^i - \overline{V}_{t+1}\right\|^2 \\
&= \frac{1}{K(N-1)}\left(1-\frac{K}{N}\right)\mathbb{E}\left[\sum_{i=1}^{N}\left\|V_{t+1}^i - \overline{V}_{t+1}\right\|^2\right] \\
&\le \frac{N-K}{N-1}\frac{16}{K}\eta_t^2 E^2 G^2.
\end{aligned}
$$

B. Experimental Setup

B.1. Model Architecture

We conduct our experiments of section 4.1 and 5.4 on a logistic regression. The loss function is given by

$$\mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^{n}\text{CrossEntropy}(f(W, x_i), y_i) + \lambda\|W\|_2^2.$$

This is a convex optimization problem. The regularization parameter λ is set to 10^{-4} [7].

We also conduct the experiments of section 4.1 on a multilayer perceptron (MLP) and a CNN. For the MLP, there is one hidden layer with 64 units and dropout with a probability of p = 0.5 in the first layer. The activation function is ReLu. For the CNN, there are two 5×5 convolution layers (the first with 10 channels, the second with 20 channels, each followed with 2×2 max pooling), a fully connected layer with 50 units and dropout with a probability of p = 0.5, and the activation function is ReLu.

In section 5, we use a CNN on the MNIST dataset with the same architecture as that in section 4.1, except the channel sizes of the convolution layers and the units of the fully connected layer are 32, 64 and 512, respectively and there is no dropout. For the ImageNet experiments, we use a RexNet architecture instead of a vision transformer because we find the vision transformer architecture does not perform well in the non-IID FL case on ImageNet.

For experiments against gradient inversion attacks, we use the same models as in the original papers [9, 11]. In detail, we use a ResNet-18 architecture for defending against the GGL attack. For DLG and iDLG attacks, the architectures are shown in Table 4.

Table 4: Model architectures for *DLG* attack and *iDLG* attack.

MNIST	*CIFAR-100*	*LFW*
5× 5 Conv 1-12	5× 5 Conv 3-12	5× 5 Conv 3-12
5× 5 Conv 12-12	5× 5 Conv 12-12	5× 5 Conv 12-12
5× 5 Conv 12-12	5× 5 Conv 12-12	5× 5 Conv 12-12
5× 5 Conv 12-12	5× 5 Conv 12-12	5× 5 Conv 12-12
FC–10	FC–100	FC–5749

B.2. Additional Hyperparameter Setup

The initialized learning rate expectations are chosen from the set $\{0.1, 0.01, 0.001\}$ evaluated by the reserved validation dataset. For the convergence analysis in section 5.4, the learning rate expectations decay following $\eta_t = \frac{\eta_0}{1+t/50}$ and $\eta_0 = 0.01$.

For the experiments of Table 1 in section 4.1, we set the number of local steps E as 25, the batch size B as 32, and weight decay as 1e-4. For the experiments of Table 2 in section 5.2, we train models with the momentum of 0.5 and weight decay of 1e-4, 5e-4, and 1e-5 for the shallow CNN, ResNet-18, and RexNet-130, respectively.

C. LRP with Adam Optimizer

Since FedAvg is developed based on SGD, most of our experiments use the SGD optimizer. We also conduct the experiments of section 4.1 using the Adam optimizer with the same MLP. The mean and standard deviation across three trials are reported in Table 5.

Table 5: Comparision of different learning rate combinations with the Adam optimizer.

η_1	η_2	**Test Accuracy %**	
		Without LRP	**LRP**
0.001	0.001	82.76 ± 1.00	82.05 ± 1.69
0.001	0.002	88.05 ± 0.04	87.65 ± 0.69
0.001	0.0005	61.84 ± 2.20	62.46 ± 3.18
0.002	0.001	57.87 ± 9.35	62.24 ± 0.55
0.002	0.002	81.86 ± 4.36	76.08 ± 5.42
0.002	0.0005	**21.43** ± 0.21	**21.69** ± 0.46
0.0005	0.001	**88.69** ± 0.21	**88.73** ± 0.26
0.0005	0.002	88.47 ± 0.51	88.47 ± 0.55
0.0005	0.0005	84.71 ± 0.24	84.35 ± 0.74

Table 5 shows that learning rate perturbation does not incur much accuracy fluctuation with the Adam optimizer, except for the cases where the learning rate combinations are $\{0.002, 0.001\}$ and $\{0.002,0.002\}$. For these two cases, the standard deviation obtained by repeated trials is quite high. Therefore, the accuracy fluctuation is mainly caused by allocating over-large learning rates to the first client, not LRP. Besides these two cases, the accuracy of trained models with and without learning rate perturbation does not change much. In addition, we get the best test-set accuracy with the learning rate combination of $\{0.0005, 0.001\}$ and the worst with $\{0.002,0.0005\}$, and scaling clients' learning rates expectation significantly impacts the test-set accuracy, which is consistent with our observation in section 4.1.

Table 6: Additional results of test accuracy (%) on MNIST, FEMNIST, CIFAR-10, and CIFAR-100 for FedAvg and LRP under non-IID settings, including the mean and standard deviation of test accuracy across 3 runs.

Algorithm	**Local learning rate**	**MNIST**	**FEMNIST**	**CIFAR-10**	**CIFAR-100**
FedAvg	$\eta_k = \eta$	99.06 ± 0.01	99.24 ± 0.01	92.09 ± 0.45	71.86 ± 0.32
LRP	$\eta_k \sim \mathcal{U}(0, 2p_k N\eta)$	99.07 ± 0.03	99.24 ± 0.01	92.32 ± 0.21	71.89 ± 0.17

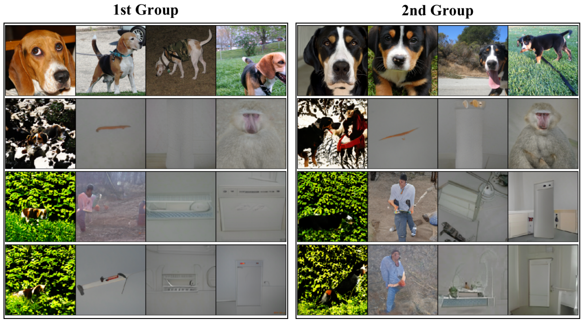


Figure 6: Image reconstructions by GGL with $B = 4$, $E = 4$: original images (1st row) and their reconstructions with $\theta_k = 0.5$ (2nd row), 1.0 (3rd row), and 2.0 (4th row) under LRP.

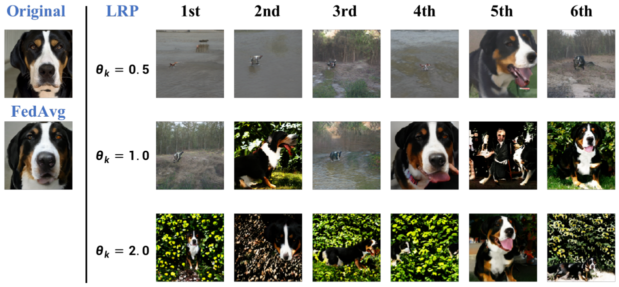


Figure 7: Image reconstructions by GGL with $B = 1$, $E = 4$, and $\theta_k = 0.5, 1.0, 2.0$.

D. Additional Accuracy Results

In our main paper, we report the accuracy results of non-IID cases formed by our dataset sharding. We include experimental results of more types of non-IID data splits in this section.

We compare the accuracy of models trained with LRP and FedAvg on the MNIST, FEMNIST [1], CIFAR-10, and CIFAR-100 datasets. For MNIST, CIFAR-10, and CIFAR-100, we assign a proportion of the data points of each label according to Dirichlet distribution and the concentration parameter β is 0.5. For FEMNIST, we divide the dataset into shards, each shard has data points of digits from a single writer. Then we allocate the shards to each client randomly and equally, following [6]. Since each client has data points from different writers, the feature distributions are not identical among the clients. We conducted experiments on FEMNIST using the same model architecture and hyperparameters as those employed for MNIST. The experimental setup is the same as in section 5, except that we set the momentum to 0 for MNIST and FEMNIST experiments. From the results in Table 6, it can be seen that LRP does not cause obvious accuracy loss in all the cases.

E. Additional Defense Results

E.1. Additional Results of LRP against GGL

For fair comparisons with previous works, we only show the results of experiments against GGL when batch size $B = 1$, the number of local steps $E = 1$, and $\mathbb{E}(\eta_t^k)/\mathbb{E}(\eta_t) = 2.0$ in our main paper. Here we present the reconstructions of other cases.

Figure 6 shows the reconstructions generated by GGL when the client trains its local model with $B = 4$, $E = 4$ and there are four samples in the client's dataset. We set the initialized learning rate to be 1e-4 and decay as $\eta_t = \eta_0/t$ to avoid one-shot training [7]. For convenience, we define $\theta_k = \mathbb{E}(\eta_t^k)/\mathbb{E}(\eta_t)$, which denotes the degree of learning rate scaling. We sample two groups of images from the validation set of ImageNet and set θ_k to be 0.5, 1.0, and 2.0, respectively. The indexes of the samples are {8112, 8113, 8114, 8115} and {11943, 11944, 11945, 11946}. It can be seen that the client's private information is well protected under our defense.

We also conduct experiments where the client only has one data point with $E = 4$ and repeat the experiments six times for every θ_k. The results are shown in Figure 7. In this setting, we admit that it is possible for the adversary to reveal information about the raw data, especially when $\theta_k = 1$. This is because the adversary can analytically extract the true label from the gradients [10] and utilize the label as prior information to generate the images. Since the GAN used by the adversary to generate reconstructions is trained on the training set of ImageNet, the label can significantly improve the reconstructions. However, this experimental case is mostly not realistic and the adversary cannot precisely extract the labels when there is more than one data point in the client's dataset, as shown in Figure 6.

E.2. Comparison of LRP with Additional Privacy Defenses

Besides the five baseline defenses in our main paper, some other methods, which are not specifically designed for privacy preservation, may also have the property to defend

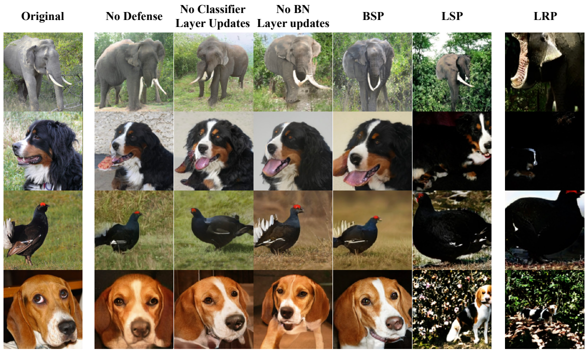


Figure 8: Comparison of our method with additional privacy defenses against GGL.

against privacy attacks, such as not sharing the BN layer updates [5, 8] or classifier layer updates [2]. Moreover, to investigate whether hiding local hyperparameters except for learning rates can also degrade gradient inversion attacks, we evaluate the performance of perturbations on the number of local steps (LSP) and batch size (BSP) against GGL.

The reconstructions generated by GGL under these additional defense settings are shown in Figure 8. We set $B = 4$, $E = 4$, $\theta_k = 2.0$, and learning rate to be 1e-4. The original images are randomly sampled from the validation set of ImageNet and their corresponding labels are provided to the attacker. It can be seen that merely hiding the BN layer or classifier layer updates cannot prevent GGL from recovering information about clients' data. In addition, LRP and LSP deteriorate the quality of images generated by GGL. Conversely, BSP exhibits limited impact on the attack in this case. Consequently, except for LRP, perturbing the number of local training steps to obscure this hyperparameter from the server can also mitigate this attack. However, introducing perturbations to the number of local steps might increase local computing overhead. Furthermore, since the possible number of local steps is limited, adversaries could evade this defense by iteratively deploying attacks with different numbers of local steps and selecting the reconstructions with the highest quality.

E.3. Defense Results against IG Attack

In addition to DLG, iDLG, and GGL attacks, we also evaluate the effectiveness of our method against one other image-based gradient inversion attack [4], which exploits a magnitude-invariant loss based on cosine similarity as the gradient matching loss and total variation as an image prior. Like [9], we call this attack as IG.

Figure 9, shows the data reconstructions generated by IG when batch size $B = 4$ and local steps $E = 4$. We use a ResNet-18 architecture. The learning rate is set to be 1e-2 and the images are randomly sampled from the test set of the CIFAR-10 dataset. It can be seen that IG generates similar images to the origin data when no defense is employed. In contrast, with the application of our approach, the quality of generated images is significantly declined, which prevents privacy leakage.

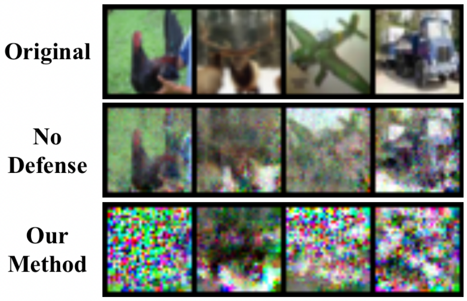


Figure 9: Defense results against IG attack.

E.4. Protection against Gradient Inversion Attack on Text

As discussed in section 6, LRP can also be applied to enhance privacy preservation for NLP FL tasks (e.g., text classification). Tabel 7 shows the results of our experiments against TAG attack [3], which is a gradient inversion attack on transformer-based language models. We choose the ideal case for the adversary with the number of local step $E = 1$. It can be seen that the attack reveals much less information about the raw data when LRP is applied.

Table 7: Defense against TAG attack on COLA.

	Example 1	**Example 2**
Original	One more pseudo generalization and i'm giving up.	The more we study verbs, the crazier they get.
FedAvg	**i m'**5 oneization more **giving** five **pseudo up general** and.	They **study** the **verbs** get **we**,. " 46. **they**
LRP	20 **general**, organize a " interaction 3 and.. ization 184	anderson 20,, the 23 20 more **verbs** 15 we 266 the

F. Impact of Number of Local Steps

In FL tasks, overhead is mainly caused by local training and communication. Intuitionally, more local training steps E means heavier local training burden but less communication overhead. However, Theorem 1 indicates that to achieve a certain accuracy, the required communication rounds T/E is a hyperbolic function of E. Therefore, it first decreases and then increases as E increases. We empirically observe this phenomenon in Figure 10. The experimental setting is the same as that in section 5.4. We set the target global training-set accuracy as 87% and 85.5% for IID and non-IID settings, respectively. A similar phenomenon can be found in the work of [7], except the local learning rates are not perturbed. Hence, increasing local training steps does not always reduce the communication burden. Overlarge E may cause high communication overhead.

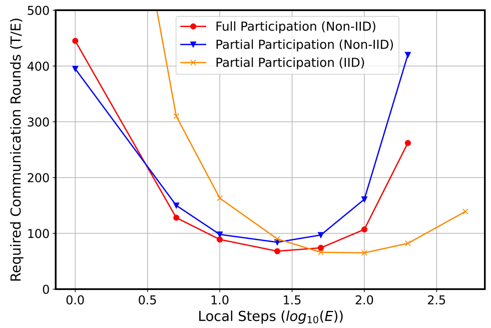
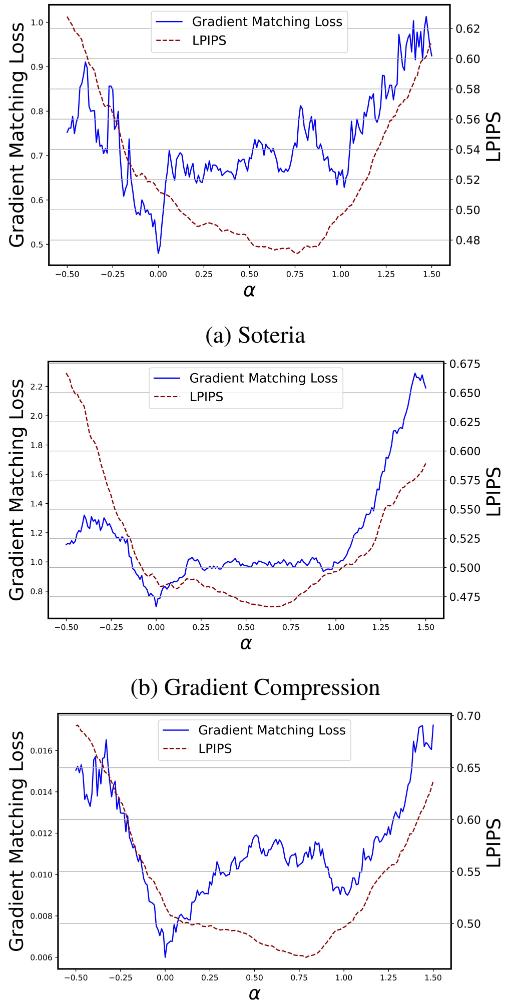


Figure 10: The impact of E.

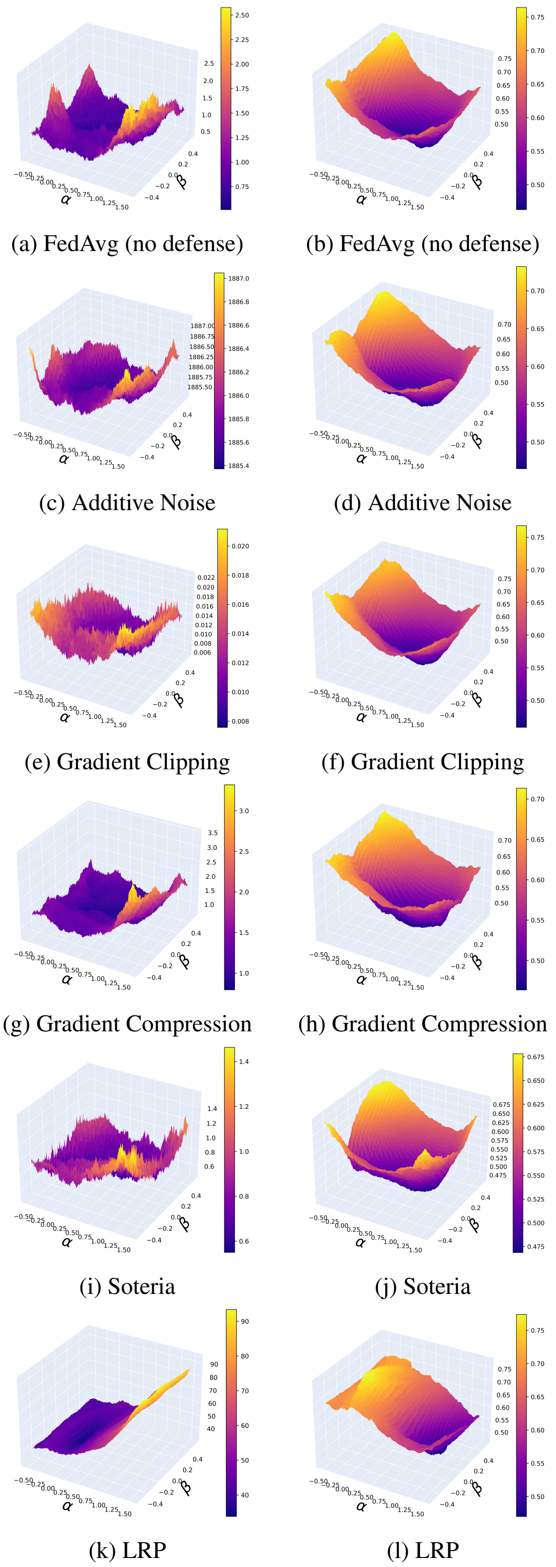


(a) Soteria

(b) Gradient Compression

(c) Gradient Clipping

Figure 11: Curves of gradient matching loss and LPIPS.



(a) FedAvg (no defense)

(b) FedAvg (no defense)

(c) Additive Noise

(d) Additive Noise

(e) Gradient Clipping

(f) Gradient Clipping

(g) Gradient Compression

(h) Gradient Compression

(i) Soteria

(j) Soteria

(k) LRP

(l) LRP

Figure 12: Landscapes of the ℓ_2 gradient matching loss (left) and LPIPS (right) under various defensive settings.

G. More Gradient Matching Loss Analysis

In our main paper, we only include the curves of gradient matching loss and LPIPS when no defense, additive noise, and LRP are applied (Figure 4b). Here we present the full results under the other baseline defenses in Figure 11. We also extend the curves to surfaces by adding a second random vector z_3: $z(\alpha, \beta) = z_1 + \alpha(z_2 - z_1) + \beta z_3$, where the latent vectors z_1 is found by GGL, z_2 is from GAN inversion and z_3 is normalized according to $z_2 - z_1$. The raw data is indexed by 11943 in the ImageNet validation set. The results are shown in Figure 12.

A general observation is that non of the baseline defenses obviously reform the curves or surfaces, compared with the results where no defense is applied, while our LRP defense changes the shape of the curves and surfaces such that latent space vectors with low gradient matching loss do not generate images with low LPIPS. This explains why LRP outperforms the baselines.

References

[1] 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. 5

[2] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR, 2021. 6

[3] Jieren Deng, Yijue Wang, Ji Li, Chao Shang, Hang Liu, Sanguthevar Rajasekaran, and Caiwen Ding. TAG: Gradient attack on transformer-based language models. *arXiv preprint arXiv:2103.06819*, 2021. 6

[4] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020. 6

[5] Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in Neural Information Processing Systems*, 34:7232–7241, 2021. 6

[6] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-IID data silos: An experimental study. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 965–978. IEEE, 2022. 5

[7] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-IID data. *arXiv preprint arXiv:1907.02189*, 2019. 1, 2, 4, 5, 7

[8] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021. 6

[9] Zhuohang Li, Jiaxin Zhang, Luyang Liu, and Jian Liu. Auditing privacy defenses in federated learning via generative gradient leakage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10132–10142, 2022. 4, 6

[10] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. iDLG: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020. 5

[11] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in Neural Information Processing Systems*, 32, 2019. 4