Safeguarding Federated Learning from Data Reconstruction Attacks via Gradient Dropout

Ekanut Sotthiwat†, Chi Zhang†, Xiaokui Xiao, *Fellow, IEEE*, Liangli Zhen, *Senior Member, IEEE*

***Abstract*—Federated Learning (FL) enables collaborative model training across distributed participants without sharing raw data, offering a privacy-preserving paradigm. However, recent studies on gradient inversion attacks have demonstrated the vulnerability of FL to adversaries who can reconstruct sensitive local training data from shared gradients. To mitigate this threat, we propose Gradient Dropout, a novel defense mechanism that disrupts reconstruction attempts while preserving model utility. Specifically, Gradient Dropout perturbs gradients by randomly scaling a subset of components and replacing the remainder with Gaussian noise, thereby creating a transformed gradient space that significantly impedes reconstruction attempts. Moreover, this mechanism is applied across all layers of the model, ensuring that attackers cannot exploit any unperturbed gradients. Theoretical analysis reveals that the perturbed gradients can be kept sufficiently distant from their true values, thereby providing safety guarantees for the proposed algorithm. Furthermore, we demonstrate that this protection mechanism minimally impacts model performance, as gradient dropout and the original training dynamics remain effectively bounded under certain convexity conditions. These findings are substantiated through experimental evaluations, where we show that various attack methods yield low-quality reconstructed images while model performance is largely preserved, with less than 2% accuracy reduction relative to the baseline. As such, Gradient Dropout is presented as an effective solution for safeguarding privacy in FL, providing a balanced trade-off between privacy protection, computational efficiency, and model accuracy.**

***Keywords*—Inversion attack, deep leakage, data privacy, federated learning**

I. Introduction

Federated learning (FL) [1] is a collaborative machine learning framework in which decentralized participants collaboratively train models without disclosing their local raw data. Specifically, participants transmit only model updates, i.e., weight parameters or gradients, to a central server, rather than sharing their raw data. This framework is particularly important in privacy-sensitive domains, such as healthcare [2–4] and banking [5, 6], where regulations strictly prohibit the sharing of personal medical records or financial transactions.

Despite its privacy-preserving design, FL remains susceptible to various security and privacy threats, particularly those that exploit shared model updates. Recent studies have demonstrated several attack vectors that can compromise the confidentiality of local training data, including membership inference [7–11], property inference [12–14], class representative attacks [15], and reconstruction attacks [16–24]. Among these, reconstruction attacks pose a particularly severe threat by enabling adversaries to reconstruct private training samples directly from shared gradients. These reconstruction attacks leverage gradient-based optimization to infer local training data. Typically, these attacks initialize synthetic (or "dummy") images with random noise and iteratively refine them by optimizing the dummy gradients to match the shared gradients [16–19]. Advanced methods further leverage generative models to enhance reconstruction fidelity [20–22], often producing high-quality images that closely resemble the original training data. The success of such attacks depends critically on the consistency and completeness of gradient information [25]: gradients encode fine-grained, spatially-structured information about input data, enabling attackers to reliably match dummy gradients to true gradients across all parameter dimensions.

To mitigate these threats, it is essential to develop techniques that secure model updates by preventing the leakage of sensitive information, while maintaining the overall performance of the FL system. To protect training data from reconstruction attacks, various approaches have been proposed, including differential privacy methods [26–28] and encryption techniques [29, 30]. In particular, differential privacy protects the trained gradients or model weights by adding randomly generated noise, making the original data and the perturbed outputs indistinguishable within a defined privacy bound. This method allows participants to manage their privacy bound through a privacy budget, allowing them to balance the trade-off between privacy and model accuracy. However, our studies, along with findings from [16, 31], reveal that the level of noise required to defend against reconstruction attacks is considerable and often leads to notable degradation in model performance. Recent work has further demonstrated that even with differential privacy defenses in place, adaptive learning-based attacks can still successfully invert gradients [32]. In contrast, encryption-based methods conceal the trained gradients by ensuring their original values remain hidden from both participants and the central server [30, 33, 34]. Unlike differential privacy, encryption methods do not introduce noise, thereby preserving model performance. However, recent studies [18, 21] have shown

† The first two authors contributed equally to this work.

This work is supported by the National Research Foundation Singapore under its AI Singapore Programme (Award No.: AISG4-GC-2023-007-1B). (Corresponding author: Liangli Zhen).

E. Sotthiwat and X. Xiao are with the School of Computing, National University of Singapore, Singapore 119077, Singapore.

C. Zhang is with the Department of Mathematics, National University of Singapore, Singapore 117543, Singapore.

L. Zhen is with the Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore 138632, Singapore (e-mail: llzhen@outlook.com).

that even when individual participants' gradients are encrypted, aggregated gradients can still reveal sufficient information for attackers to reconstruct training images. Moreover, encryption-based methods impose additional computational overhead. Gradient sparsification and compression techniques [35, 36], while reducing communication costs, similarly fail to address reconstruction vulnerabilities, as the retained gradients still maintain their data-aligned structure [37].

This paper addresses privacy concerns in FL by proposing Gradient Dropout, a novel defense mechanism that protects local data from reconstruction attacks while minimizing impact on model performance and computational cost. As shown in Figure 1, Gradient Dropout perturbs gradients layer by layer by scaling a selected portion with a predefined scaling factor $(\frac{1}{p})$ and replacing the remaining gradients with random numbers drawn from a Gaussian distribution. This perturbation creates a distinct gradient space that effectively mitigates reconstruction attacks. By introducing scaling and noise, the gradients are displaced from the data-aligned subspace, reducing their alignment with specific features of the training data. Consequently, attackers attempting reconstruction are constrained to operate within this altered gradient space, resulting in distorted or inaccurate representations that do not resemble the original data.

The design of Gradient Dropout is grounded in a principled understanding of both reconstruction attack mechanics and gradient utility. Our key insight is that reconstruction attacks require consistent gradient matching across all dimensions [25]. By randomly selecting a subset of gradients and replacing them entirely with Gaussian noise sampled from the same value range as the original gradients, we inject irreducible uncertainty into the attacker's optimization objective. Unlike differential privacy's uniform additive noise, our replacement strategy ensures that the attacker cannot distinguish between true gradient values and decoy values, as both fall within the same statistical distribution. This transforms the reconstruction problem from a well-posed inverse problem into an ill-posed one with multiple plausible solutions. Furthermore, we theoretically demonstrate that under specific convexity conditions, Gradient Dropout ensures that the training dynamics remain bounded, allowing the model to maintain performance while safeguarding its training data. This strategy achieves a superior privacy-utility trade-off: strong protection through complete obscuration of a gradient subset, combined with preserved learning capacity through the scaled retention of remaining gradients.

To validate the effectiveness of our proposed defense algorithm, we present comprehensive experimental results comparing Gradient Dropout with existing defense methods against various reconstruction attacks. Additionally, we assess the impact of each defense method on model performance within a federated learning setting. The results demonstrate that Gradient Dropout effectively mitigates reconstruction attacks while preserving model performance, with only minimal degradation. Notably, our experiments include high-resolution datasets such as ImageNet and NIH Chest X-ray, demonstrating that our method effectively protects training data even in these challenging scenarios. An ablation study further highlights the robustness of our method, showing negligible impact on model performance across various scaling factors and noise levels. Overall, our method achieves strong data protection with minor performance trade-offs, successfully safeguarding against reconstruction attacks.

The key novelties and main contributions of this work are summarized as follows:

- We propose Gradient Dropout, a robust defence mechanism that perturbs gradients through selective scaling and Gaussian noise replacement. Unlike differential privacy, which applies uniform additive noise across all gradient components, our method strategically replaces a subset of gradients with statistically indistinguishable decoy values. This approach effectively safeguards training data while maintaining high model performance.
- We provide a theoretical analysis demonstrating that Gradient Dropout transforms the reconstruction problem from a well-posed to an ill-posed inverse problem. Specifically, we derive certified safety guarantees by bounding the discrepancy between original and reconstructed gradients. Moreover, we prove that under specific convexity conditions, our method ensures bounded training dynamics whilst maintaining convergence properties.
- Extensive experiments on CIFAR-10, ImageNet, and the NIH Chest X-ray dataset validate the effectiveness of our method against advanced reconstruction attacks, achieving strong privacy protection with minimal impact on model accuracy. Our method outperforms existing defenses including differential privacy and data representation perturbation in the privacy-utility trade-off, particularly in medical imaging scenarios where privacy requirements are most essential.

II. Related Work

In this section, we review different types of data leakage from local participants in centralized federated learning, with a particular focus on reconstruction attacks that allow adversaries to extract sensitive training data from shared gradients. We also examine key defense mechanisms designed to mitigate reconstruction attacks, including gradient perturbation techniques using differential privacy, gradient encryption, the Soteria framework, and regularization methods such as dropout layers. These methods aim to obscure gradient information, limiting attackers' ability to reconstruct the original data accurately. Furthermore, we analyze the effectiveness of these defenses across different scenarios, highlighting the trade-offs between privacy protection and model accuracy.

A. Reconstruction Attacks

FL is highly susceptible to reconstruction attacks since trained gradients often encode sensitive training data. Deep Leakage from Gradients (DLG) [16] demonstrates that training data can be reconstructed in federated learning (FL) frameworks by exploiting access to the global model parameters w and the gradients ∇w. It generates a dummy image x'

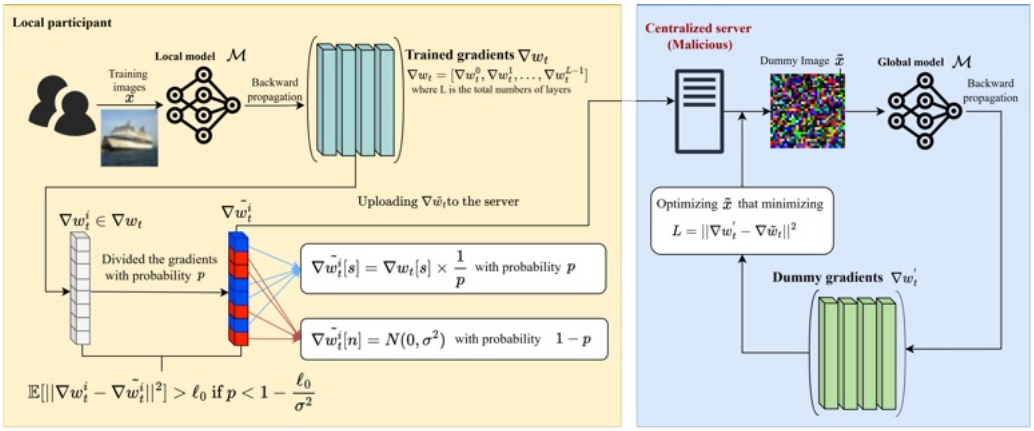


Fig. 1. Overview of our defense framework against reconstruction attacks. Scaling the blue gradients by $(1/p)$ and replacing the remaining gradients with Gaussian noise to ensure that L2 distance between trained gradients and protected gradients is greater than ℓ_0

and computes dummy gradients $\nabla\tilde{w}$ using the global model parameters w and the corresponding labels y. To reconstruct the original images x, it iteratively adjusts x' to minimize the $L2$ distance between dummy gradients $\nabla w'$ and trained gradients ∇w, as described in Eq. (1).

$$X^{'} \leftarrow \underset{X'}{\operatorname{argmin}} \; ||\nabla w' - \nabla w||_2 \tag{1}$$

A key limitation of DLG is its reliance on modifying the model architecture by replacing ReLU with Sigmoid activation functions to guarantee twice differentiability. To overcome this limitation, Inverting Gradients (IG) [18] introduces enhancements including the use of cosine similarity as a loss function, Adam as an optimizer, and a total variation regularization term. These improvements facilitated reconstruction on medium batch sizes without requiring modifications to the model architecture. However, IG faces challenges when applied to large batch sizes and high-resolution images. To address these limitations, See-Through Gradients [19] utilizes prior information, such as the mean and standard deviation from the batch normalization layers, to enhance reconstruction. Although this approach successfully reconstructs large batch sizes and high-resolution images, the dependency on batch normalization statistics raises privacy concerns, as such information is inherently private and not typically accessible through shared gradients in FL. Some studies [20–22] further demonstrate the effectiveness of using generative models, such as those from Generative Adversarial Networks (GANs), to update dummy images instead of directly adjusting pixel values.

Recent research has explored the adaptive attacks for data reconstruction from gradients [38]. These attacks are explicitly designed to remain effective even when gradients are modified by common defense mechanisms, such as noise injection, clipping, or pruning. For example, Liu *et al.* [32] proposed Mjölnir, a diffusion-based gradient leakage attack that adaptively denoises perturbed gradients through reverse diffusion, posing a direct challenge to perturbation-based defenses such as differential privacy and gradient perturbation. Similarly, advances in multi-modal federated learning have revealed additional privacy vulnerabilities. MGIA [39] exploits cross-modal information leakage in multi-modal federated learning settings, while subsequent work [40] further investigates the impact of noise injection on privacy preservation across different modalities. Zhang *et al.* [41] provided an extensive overview of the evolving landscape of gradient inversion attacks and defences. These recent developments underscore the importance of designing robust defences that can withstand adaptive adversaries, which motivates our work on Gradient Dropout that strategically replaces gradients with statistically indistinguishable decoy values rather than simply adding noise to the original gradients.

B. Defense Methodologies

1) Differential Privacy: Differential privacy (DP) is one of the most widely adopted methods for preventing information leakage in federated learning, offering strong theoretical guarantees for protecting sensitive training data. In deep learning, DP-SGD [42] ensures privacy by injecting random noise into gradients, effectively obscuring specific details and making it difficult for attackers to extract meaningful information. The noise level is controlled by a privacy budget, denoted as ϵ, which governs the trade-off between privacy and model accuracy: a smaller ϵ provides stronger privacy but introduces more disruptive noise that degrades model performance.

To address this trade-off, Xue *et al.* [28] proposed an adaptive noise mechanism that dynamically adjusts noise levels during training. While this strategy reduces unnecessary accuracy degradation compared to fixed-noise approaches, it still fundamentally relies on noise injection, which inherently limits model performance under strong privacy guarantees. Prior studies [16] and our research demonstrate that in federated learning, the substantial noise required to prevent reconstruction attacks has a significant adverse effect on model accuracy. Also, You *et al.* [31] demonstrated that local differential privacy (LDP) alone is insufficient against sophisticated reconstruction attacks. Their work shows that training samples can

be successfully reconstructed even from clipped and perturbed gradients protected by LDP, as gradient compression and sample denoising techniques can circumvent these defenses. These findings highlight that while DP provides theoretical privacy guarantees, achieving robust protection against reconstruction attacks without significantly compromising model utility remains an open challenge.

2) Multi-Party Computation: Multi-Party Computation (MPC) [33, 43] is a widely used defense mechanism in federated learning to enhance privacy and security during model training. MPC offers a robust framework that allows distributed participants to collaboratively compute arbitrary functions without revealing while keeping their private gradients. Generally, in standard federated learning frameworks without defense mechanisms, local participants upload their training parameters to a central server, which aggregates them to compute global updates. With MPC, this computation is securely performed without exposing private gradients. Specifically, MPC preserves privacy by splitting sensitive model updates from each participant into multiple secret shares, which are then distributed among other participants. This process ensures that no single participant can reconstruct the original data from the shares they receive. Each participant then aggregates the shares obtained from others with their own and transmits the aggregated shares to the central server. The server aggregates these to update the global model. At every round of this process, the original training data and model updates remain encrypted and inaccessible, ensuring privacy protection.

A key advantage of Multi-Party Computation (MPC) over methods like differential privacy is its ability to preserve model accuracy by leveraging encryption and decryption, ensuring that model updates remain precise and unaffected by noise. This eliminates the performance degradation commonly associated with differential privacy. However, MPC is computationally intensive, particularly in large-scale federated settings, where generating, transmitting, and processing secret shares across numerous participants introduce substantial communication and computation overhead. Additionally, despite its encryption-based security, MPC does not inherently prevent information leakage, as attackers may still reconstruct training data from aggregated gradients stored on the central server.

3) Gradient Perturbation and Transformation Methods: Beyond differential privacy and encryption-based approaches, various methods have been proposed that directly manipulate gradients or training processes to thwart reconstruction attacks. Sun *et al.* [44] proposed *Soteria*, which identifies fully connected layers as the primary source of sensitive data representation and selectively perturbs their input parameters while leaving other layer gradients unchanged. Although *Soteria* provides theoretical guarantees for robustness and convergence, attackers can still exploit the unperturbed gradients of other layers to reconstruct training images. He *et al.* [45] leveraged dropout layers to randomly deactivate neurons during training, thereby obscuring partial gradient information. However, Dropout Inversion Attacks [46] demonstrated that this defence can be circumvented by accessing or approximating dropout masks.

More recent approaches attempt to fundamentally decouple gradients from training data. Gao *et al.* [47] proposed training local models using statistical information rather than raw data, combined with knowledge distillation to transfer knowledge to lightweight student models, ensuring that only semantically meaningless information can be reconstructed. Zhou *et al.* [48] introduced *Shade*, which generates alternative shadow data using generative adversarial networks or diffusion models to construct surrogate models that eliminate memory of raw data. Ye *et al.* [49] analyzed the underlying causes of gradient inversion vulnerabilities and proposed a plug-and-play defense that augments training data using a designed vicinal distribution, providing privacy protection. While these methods offer stronger privacy guarantees, they introduce substantial computational overhead and implementation complexity, limiting their practicality in resource-constrained federated learning scenarios. In contrast, our proposed Gradient Dropout achieves effective protection with minimal computational cost by strategically replacing gradient subsets with statistically indistinguishable noise, without requiring generative models, knowledge distillation, or architectural modifications.

III. Our Proposed Method

A. Problem Statement

In the FL system, the objective is to collaboratively train an accurate global model $\mathcal{M}$ using distributed data from multiple local participants or devices, denoted as $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_m$. Each participant trains the model on their local dataset $\{\mathcal{X}_i \mid i \in [1, \cdots, m]\}$, ensuring that the raw data remain on their device throughout the training process. This decentralized approach is designed to enhance privacy by avoiding the transfer of sensitive data.

Although local data are never transferred directly in FL, privacy leakage remains a significant concern due to reconstruction attacks, in which shared gradients can be exploited to reconstruct sensitive information from local devices. To protect against such privacy breaches, various defense methodologies, including encryption and differential privacy (DP), have been developed. However, these methods have inherent trade-offs: encryption incurs considerable communication overhead, while DP reduces model accuracy by adding noise to safeguard training data from reconstruction attacks.

B. Gradient dropout in Federated Learning

To prevent privacy leakage, it is essential to incorporate randomness during the training process to obscure sensitive information embedded within gradients. Yet, this incorporation must not unduly compromise the overall training performance. An effective defense algorithm should therefore achieve the following key objectives:

1) Objective 1: Safeguard training data against reconstruction attacks by perturbing gradients in a manner that ensures the deviation between the original and perturbed gradients remains effectively bounded.

2) Objective 2: Maintain model performance by ensuring that the introduced perturbations have minimal impact on accuracy.

To achieve these objectives, we propose a defense framework called Gradient Dropout, which is designed to protect training data from reconstruction attacks while maintaining model performance. Specifically, after each training iteration t, the local gradients ∇w_t are obtained by backward propagation. Gradient Dropout randomly perturbs these local gradients by scaling a subset with the gradients and replacing others with random values drawn from a Gaussian distribution $N(0, \sigma^2)$.

Algorithm 1 Our Proposed Gradient Dropout

Input:

- L: Number of hidden layers in the neural network.
- $\nabla w_t = \{\nabla w_t^1, \nabla w_t^2, \ldots, \nabla w_t^L\}$: Shared gradients at round t, where ∇w_t^i is the gradient for layer i.
- $p \in (0, 1]$: Retention probability (*i.e.*, the fraction of gradient elements to be scaled).
- σ^2: Variance of the Gaussian noise distribution.

Output:

- $\nabla \tilde{w}_t = \{\nabla \tilde{w}_t^1, \nabla \tilde{w}_t^2, \ldots, \nabla \tilde{w}_t^L\}$: Protected gradients.

for $i \leftarrow 1$ **to** L **do**
  Let ∇w_t^i be the gradient for layer i
  Sample a matrix $X^i \sim \mathcal{U}(0, 1)$ with the same shape as ∇w_t^i
  Define the binary mask:

$$\text{mask}_{jk} = \begin{cases} 1, & \text{if } X_{jk}^i < p, \\ 0, & \text{otherwise,} \end{cases}$$

  for all indices (j, k).
  for each index (j, k) in ∇w_t^i:
  - **if** $\text{mask}_{jk} = 1$: Set

$$\nabla \tilde{w}_t^i[j, k] = \frac{\nabla w_t^i[j, k]}{p}.$$

  - **else** (*i.e.*, if $\text{mask}_{jk} = 0$): Set

$$\nabla \tilde{w}_t^i[j, k] \sim \mathcal{N}(0, \sigma^2).$$

end for
return $\nabla \tilde{w}_t$

Let L denote the total number of layers in the model $\mathcal{M}$, where $i \in [1, \ldots, L]$ representing the index of each layer. Our proposed defense mechanism applies to the gradients of each layer, ∇w_t^i. To protect these gradients, components of the true gradient ∇w_t^i are randomly selected with probability p, forming the set of scaling indices, denoted as $[s]$, while the remaining indices are categorized as noise indices, denoted as $[n]$. For the scaling components $\nabla w_t^i[s]$, the original gradients are scaled by a factor of $1/p$. This preserves the gradient direction, ensuring minimal impact on model performance, while effectively obscuring the gradients from attackers. Moreover, as demonstrated in Eq. (4), this choice of scaling factor ensures that the system retains gradient direction in expectation. For the remaining gradients $\nabla w_t^i[n]$, the gradient values are replaced with random samples drawn from a Gaussian distribution $N(0, \sigma^2)$. Such randomization introduces noise in both magnitude and sign, significantly increasing the difficulty of reconstruction attacks and rendering the recovery of the original gradients substantially more challenging.

Overall, the framework presented in Algorithm 1 provides a robust defense against reconstruction attacks by integrating gradient scaling with controlled noise injection. This method effectively obscures sensitive information embedded within gradients while preserving model performance and ensuring convergence guarantees, as demonstrated in the theoretical analysis that follows.

C. Gradient Dropout Preventing Gradient Leakage

We first establish that the gradient dropout strategy can effectively prevent deep leakage with an appropriate choice of p in the following Theorem.

Theorem 1 (Expected Gradient Leakage Bound)**.** *Suppose an attacker attempts to reconstruct the underlying ground-truth data (x, y) by minimizing the discrepancy between the true gradients and the surrogate gradients:*

$$\ell := \frac{1}{N} \sum_{i}^{N} \|\nabla w_t^i - \nabla \tilde{w}_t^i\|^2, \tag{2}$$

where i denotes the i-th component of the model. Then, it suffices to require

$$p < 1 - \frac{\ell_0}{\sigma^2}, \tag{3}$$

to ensure that the expected gradient matching loss is sufficiently large:

$$\mathbb{E}_{w,B}[\ell] > \ell_0.$$

Remark: In Algorithm 1, any third party has access to the protected gradient $\nabla \tilde{w}_t^i$. Theorem 1 establishes that by selecting a sufficiently small probability p, the gradient ∇w_t^i computed on the true images remains sufficiently distant from such a public gradient $\nabla \tilde{w}_t^i$, exceeding a threshold ℓ_0. Consequently, since the optimization process in deep leakage is designed to minimize the loss—often driving it toward nearly zero, as noted in [22]—the gradient dropout algorithm inherently prevents convergence toward the ground-truth images under these conditions. Note that the above analysis considers the ℓ_2 distance between gradients; analogous results for cosine similarity are provided in Theorem 4 in the Appendix.

D. Convergence Guarantee

Many existing algorithms can prevent gradient leakage by establishing results similar to those in the above Theorem; however, they often do so at the expense of reduced model performance. For example, the widely recognized Differential Privacy (DP) algorithm [50] achieves protection by continuously adding noise to the gradients, thereby safeguarding local image data. However, this approach may potentially compromise the overall model performance. In this section,

we demonstrate that the impact of gradient dropout on training performance can be effectively controlled within a small range by rigorously analyzing the dynamics of the training process.

To proceed, we first interpret the training step index t as a temporal variable [1], effectively transforming the model into a continuous-time analogue, as explored in prior studies [51–53]. Under this perspective, the original discrete updates can be approximated by a continuous function:

$$\mathbb{E}[d\tilde{w}(t)] = -\nabla\tilde{w}(t)dt - (1-p)\sigma dB(t). \tag{4}$$

This representation corresponds to a Stochastic Differential Equation (SDE), where the term $\nabla\tilde{w}(t)dt$ captures the gradient dynamics, and $(1-p)\sigma dB(t)$ introduces a stochastic drift component.

In a similar manner, we can express the original dynamics using an ordinary differential equation (ODE):

$$dw(t) = -\nabla w(t)dt. \tag{5}$$

Let $\delta(t) = \mathbb{E}[\tilde{w}(t)] - w(t)$ represent the weight difference at time t. Our goal is to analyze the dynamics of this term:

$$d\delta(t) = -\left(\nabla\tilde{w}(t) - \nabla w(t)\right)dt - (1-p)\sigma dB(t). \tag{6}$$

This expression corresponds to another SDE, incorporating the drift term $dB(t)$, which introduces stochastically into the dynamics. Then we are able to bound this term in the following Theorems.

Theorem 2 (Weight Difference Bound under General Lipschitz Gradients)**.** *Consider the continuous-time approximation of the parameter updates, where the discrete updates are modeled by the differential equations in Eqs. (4) and (5). Let $\delta(t) = \mathbb{E}[\tilde{w}(t)] - w(t)$ denote the weight difference between the stochastic and deterministic dynamics. Suppose that the gradient function ∇f is Lipschitz continuous with constant L, i.e.,*

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\| \quad \textit{for all } x, y.$$

Then the evolution of the expected squared weight difference satisfies

$$\frac{d}{dt}\mathbb{E}[\delta^2(t)] \;\le\; 2L\,\mathbb{E}[\delta^2(t)] + (1-p)^2\sigma^2.$$

Moreover, using Grönwall's inequality, we obtain the bound

$$\mathbb{E}[\delta^2(t)] \;\le\; \frac{(1-p)^2\sigma^2}{2L}\left(e^{2Lt} - 1\right).$$

Remark: Under the assumption that the gradient ∇f is Lipschitz continuous, Theorem 2 shows that the expected squared weight difference $\mathbb{E}[\delta^2(t)]$ admits an upper bound determined by the gradient dropout variance term $(1-p)^2\sigma^2$ for any fixed time t. This result holds for general smooth (possibly non-convex) objectives and thus applies directly to modern deep neural networks.

[1]Throughout this paper, the symbol i indexes neural network layers, while the variable t represents optimization time in the continuous-time limit.

Theorem 3 (Weight Difference Bound under Strong Convexity)**.** *In addition to the assumptions of Theorem 2, suppose that f is μ-strongly convex with parameter $\mu > 0$, i.e.,*

$$(\nabla f(x) - \nabla f(y))^\top (x-y) \;\ge\; \mu\|x-y\|^2 \quad \textit{for all } x, y.$$

Then the evolution of the expected squared weight difference satisfies

$$\frac{d}{dt}\mathbb{E}[\delta^2(t)] \;\le\; -2\mu\,\mathbb{E}[\delta^2(t)] + (1-p)^2\sigma^2.$$

Solving this differential inequality yields

$$\mathbb{E}[\delta^2(t)] \;\le\; \frac{(1-p)^2\sigma^2}{2\mu}\left(1 - e^{-2\mu t}\right).$$

Remark: If, in addition, the objective f is μ-strongly convex, Theorem 3 further guarantees that $\mathbb{E}[\delta^2(t)]$ converges to a steady-state bound

$$\lim_{t\to\infty}\mathbb{E}[\delta^2(t)] \le \frac{(1-p)^2\sigma^2}{2\mu}.$$

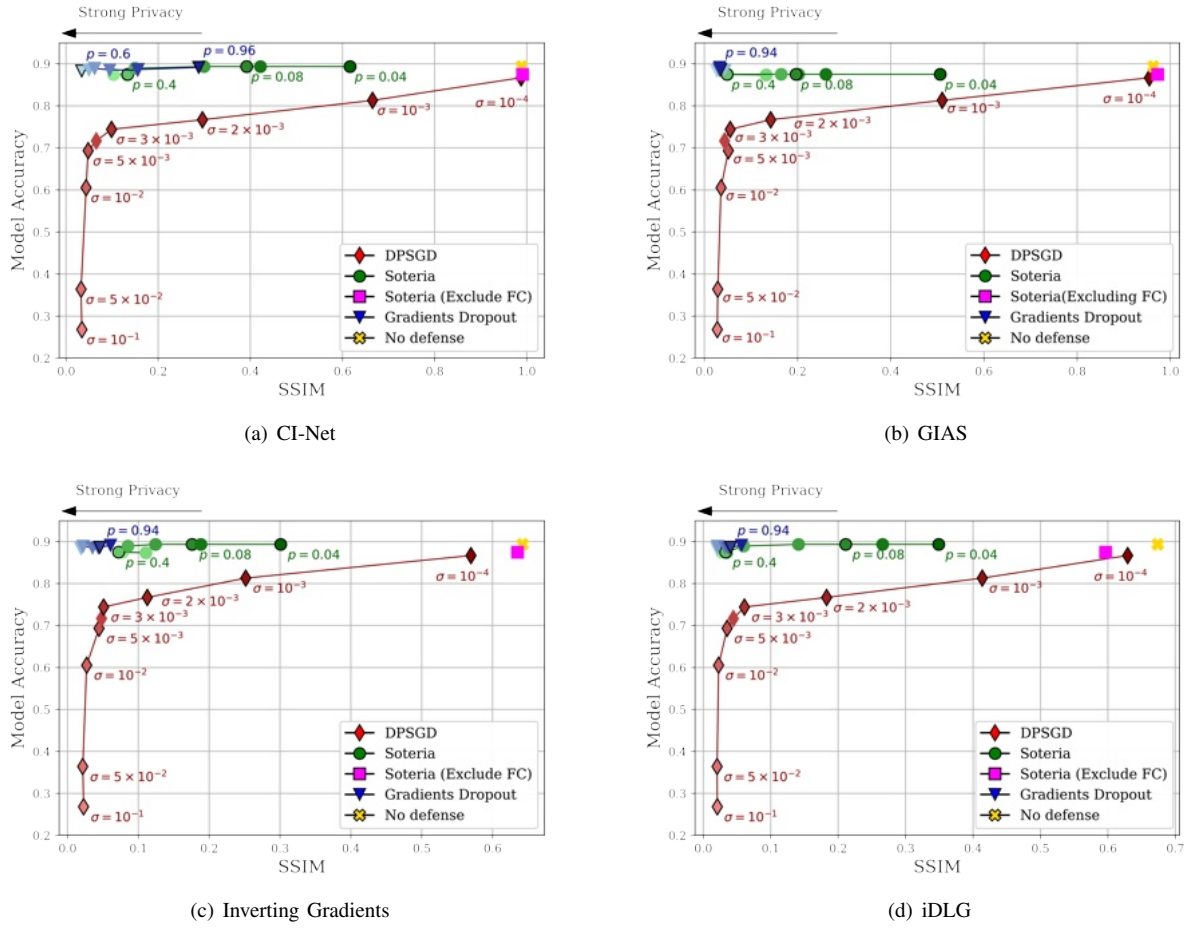
This refinement demonstrates the asymptotic stability of the dynamics under strong convexity, but we emphasize that such an assumption is *not* required for the general applicability of our method in typical federated learning settings.

In summary, the above theorems indicate that we can always bound the expected weight difference between the gradient dropout and the original training dynamics. In particular, under the strong convexity condition, the gap is bounded by $\frac{(1-p)^2\sigma^2}{2\mu}$. This implies that the weight difference does not grow uncontrollably, and instead, it converges to a steady-state value, demonstrating that gradient dropout maintains stability in the training process, even in the presence of noise introduced by the dropout mechanism.

IV. Experimental Study

To evaluate the effectiveness of our defense against reconstruction attacks, we conduct comprehensive experiments comparing state-of-the-art defense methods, including DP-SGD and Soteria, against four advanced reconstruction attacks: iDLG, IG, GIAS [20], and CI-Net [22]. In addition, to assess robustness against adaptive attacks, we conduct experiments under the gradient inversion attacks of Mjölnir [32] and Learning to Invert (LTI) [38]. We consider an honest-but-curious server as the adversary, who has white-box access to the shared gradients and the global model parameters. However, batch normalization statistics, such as the local mean and standard deviation, are not assumed to be available to the attacker, as this constitutes an unrealistically strong prior in practical federated learning settings.

Four datasets are used in our experiments, including the CIFAR-10 (32×32 px) [54], FashionMNIST [55], ImageNet (256×256 px) [56], and NIH ChestX-ray (256×256 px) [57] [57] datasets. For the evaluations, we utilize a modified ResNet-18 model in which the ReLU activation is replaced with a Sigmoid activation by following the settings



(a) CI-Net (b) GIAS

(c) Inverting Gradients (d) iDLG

Fig. 2. Comparison of training accuracy for defense methodologies and SSIM between reconstructed and original images on the CIFAR-10 dataset using ResNet-18 with a batch size (BS) of 8. For the hyperparameter settings: in DP-SGD, σ represents the standard deviation of the added noise; in Soteria, p denotes the percentage of perturbation; and in our defense, p indicates the percentage of scaling gradients. The gradient of colors represents different hyperparameter settings, where lighter colors indicate higher privacy, and darker colors indicate lower privacy levels.

in [16]. We measure the defense effectiveness with three complementary metrics: the Structural Similarity Index Measure (SSIM) [58], Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) [59]. SSIM and PSNR measure the structural and pixel-level similarity between reconstructed and original images, where higher values indicate greater reconstruction accuracy. Conversely, LPIPS captures perceptual dissimilarity, where lower values indicate higher reconstruction quality. Also, we examine the impact of the defenses on model performance to identify methods that effectively balance privacy preservation and predictive accuracy.

1) ***Peer Defense Methods****:* We compare our method with two existing defenses: DP-SGD and Soteria. DP-SGD mitigates privacy leakage by adding Gaussian noise to the training gradients, with noise configured to have a mean of 0 and varying standard deviations as shown in our experiments. Soteria perturbs intermediate representations before the fully connected layer while leaving gradients in other layers unchanged. This design restricts the defense's effect to the fully connected layer, with differences between the original and protected gradients arising primarily from modifications introduced in that layer.

2) ***Hyper-parameter setting****:* In the federated learning setup, models are trained on the CIFAR-10 dataset using three local participants unless otherwise specified, each receiving a random data split. Training consists of 100 rounds, with one local epoch per round, and SGD as the optimizer. The learning rate is set to 0.1 for the first 50 rounds and reduced to 0.001 for the remaining rounds. Reconstruction attacks are performed with a batch size of 8 for CIFAR-10 and 1 for ImageNet, with 4,000 attack iterations. For GIAS, 800 iterations are allocated for latent vector updates and 3,200 for generator updates. The defense parameters are configured as follows: DP-SGD: Implemented with a Gaussian mechanism, where noise has a mean of zero and standard deviations ranging from 10^{-4} to 10^{-2}, without clipping norms. Soteria: Perturbation percentages are set to 0.04, 0.06, 0.08, 0.1, 0.2, 0.4, 0.6,

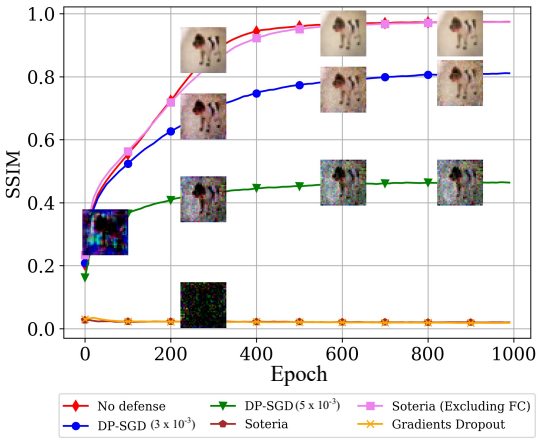


Fig. 3. The mean SSIM from CI-Net against the defense methods. In "Soteria (excluding FC)", the fully connected layer gradients are ignored in attack.

and 0.8. In our defense, we set the noise standard deviation $\sigma = 5 \times 10^{-3}$, while the p values are varied as 0.96, 0.94, 0.92, 0.9, 0.8, 0.7, and 0.6.

A. Comparison with Peer Defense Methods

Figure 2 shows the trade-off between model accuracy and privacy protection for three defenses under four reconstruction attacks. Defenses positioned in the top-left region exhibit the most favorable balance, achieving high model accuracy while maintaining low SSIM scores. Our proposed defense method is positioned near this optimal region, achieving an accuracy of 88% compared to 89% for the baseline (without defense), while maintaining an SSIM score below 0.2. This result indicates that Gradient Dropout effectively prevents attackers from reconstructing images while incurring only a minimal accuracy reduction.

The strength of our method lies in its gradient perturbation strategy, which combines scaling a portion of the gradients and replacing the rest with Gaussian noise. This technique generates diverse gradients that effectively prevent the reconstruction of the original training data. In contrast, achieving a similar SSIM score below 0.2 with DP-SGD requires injecting substantial noise, which reduces model accuracy by at least 12%. This highlights the advantage of Gradient Dropout, in achieving a superior balance between privacy protection and model performance through precise noise calibration. While Soteria achieves comparable accuracy to ours and can reach SSIM < 0.2 when $p \geq 0.4$, it still allows attackers to reconstruct images with moderate accuracy. Moreover, Soteria only perturbs gradients in the fully connected layer, leaving convolutional layers exposed. Consequently, attackers can exploit these unprotected gradients to successfully reconstruct training data. In the figure, the pink marker represents the scenario under the Soteria defense where the attacker excludes gradients from the fully connected layer during optimization, referred to as "Soteria (excluding FC)". As expected, the SSIM score in this case closely aligns with that of an unprotected model, confirming a major limitation of Soteria: when only the fully connected layer is perturbed, attackers can leverage unprotected convolutional gradients to reconstruct training data effectively.

To further evaluate the effectiveness of our method and other defense mechanisms, we examine small batch-size scenarios, which pose a higher risk of reconstruction than larger batch sizes, as attackers can more easily reconstruct individual images. Following the previous experimental protocol, we reduce the batch size to one and conduct 1,000 attack iterations using the CI-Net attack. For each defense method, we select hyperparameter configurations from the top-left region of Figure 2 (a), as these settings offer the best balance between data protection and model performance.

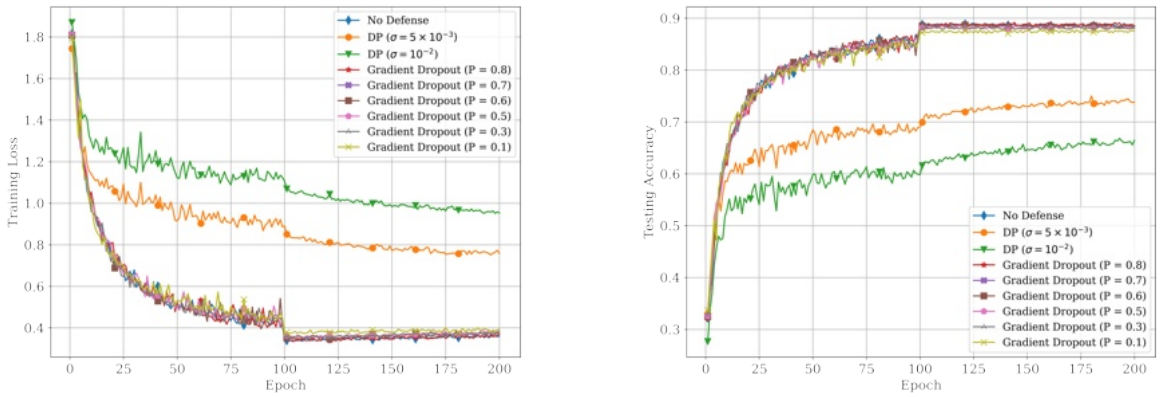
For DP-SGD, we set $\sigma = 3 \times 10^{-3}$ and 5×10^{-3}, while for Soteria, we set $p = 0.8$. In this experiment, we also assess a scenario where the attacker ignores gradients from the fully connected layer, denoted as "Soteria (excluding FC)". For our proposed method, we set $p = 0.6$ and $\sigma = 5 \times 10^{-3}$. Figure 3 illustrates the data leakage process for the three defense methods under the CI-Net attack, along with the corresponding SSIM values. The results indicate that recalibrating the noise level is critical for maintaining privacy when the batch size is set to 1. DP-SGD fails to adequately protect training data in this scenario, when applying small noise levels, rendering it ineffective against reconstruction attacks. In contrast, both our method and Soteria exhibit robust privacy protection. However, in the "Soteria (excluding FC)" configuration, where attackers ignore the fully connected layer gradients, the leakage process closely resembles that of a model without defense. This finding underscores that while prior analyses identify the fully connected layer as a key source of data representation, other layers retain sufficient information for attackers to reconstruct images.

These findings highlight the importance of achieving a balance between privacy and utility in FL. Our proposed method provides a practical solution, offering strong privacy protection while maintaining high model accuracy. This balance makes our approach particularly well-suited for privacy-sensitive applications, such as healthcare and finance, where safeguarding data confidentiality is essential.

B. Convergence and Prediction accuracy

To demonstrate that our proposed method effectively controls weight differences without causing uncontrolled growth—thus ensuring stable training—we compare the convergence speed and model accuracy of each defense method. In the FL setup, the dataset is randomly partitioned in an IID manner among three participants. Training is conducted over 200 iterations, starting with an initial learning rate of 0.01, which decays to 0.001 after 100 iterations. A batch size of 16 is used for all experiments. For DP-SGD, Gaussian noise is applied with a mean of 0 and standard deviations of 5×10^{-3} and 10^{-2}. In our proposed Gradient Dropout defense, the probability parameter p is varied from 0.1 to 0.8, while the noise standard deviation is set to 5×10^{-3}.

Figure 4 illustrates the training stability and model accuracy of ResNet-18 on the CIFAR-10 dataset under different defense



(a) Training Loss

(b) Accuracy on Test Set

Fig. 4. Comparison of Gradient Dropout with other methods in terms of training stability and model accuracy using ResNet-18 on the CIFAR-10 dataset. The figure consists of two subplots: Training Accuracy vs. Test Accuracy, showing performance trends. Training is conducted with a batch size of 16.

methods over 200 training iterations. The training and testing loss curves indicate that all models exhibit stable convergence. From the accuracy curves, we observe that the baseline (No Defense) achieves the highest accuracy, converging at approximately 89%, which serves as the upper bound for model performance. In contrast, DP-SGD demonstrates a significant decline in accuracy as the noise scale increases. With a standard deviation of $\sigma = 5 \times 10^{-3}$, DP-SGD converges to approximately 72%. However, when the noise increases to $\sigma = 10^{-2}$, accuracy drops further and stabilizes at around 60%, highlighting the performance degradation caused by excessive noise.

Our proposed Gradient Dropout method demonstrates superior performance compared to DP-SGD while preserving privacy. Specifically, for $p = 0.8$, 0.7, and 0.6, testing accuracy converges near the baseline, reaching approximately 88%. This corresponds only a minor accuracy reduction of less than 2% compared to the baseline, despite gradient perturbation applied. In contrast, when the replacement probability is largely reduced, *e.g.*, $p = 0.1$, the final accuracy decreases by approximately 3%, indicating that excessive gradient replacement can negatively impact model performance. These results highlight the trade-off between privacy and utility governed by p: smaller values provide stronger privacy protection but reduce model utility, while larger values preserve accuracy at the cost of weaker privacy guarantees.

To further demonstrate that the proposed defense maintains model accuracy in federated learning settings with a larger number of participants, we evaluate the classification performance of ResNet-18 on CIFAR-10 with 16 participating clients. We additionally vary the probability parameter p from 0.6 to 0.96, with a noise standard deviation of 5×10^{-3}. As shown in Figure 5, the proposed defense preserves model accuracy across a wide range of p values and maintains performance close to the no-defense baseline. These results demonstrate the effectiveness of the proposed approach in practical federated learning scenarios.

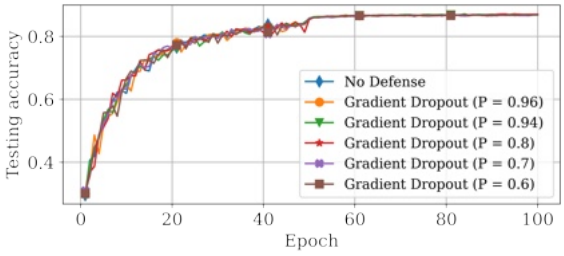


Fig. 5. Comparison of Gradient Dropout with varied probability p in terms of training stability and model accuracy using ResNet-18 on the CIFAR-10 dataset with 16 clients. Training is conducted with a batch size of 16.

Overall, these results confirm that Gradient Dropout effectively balances privacy protection and model performance. Unlike DP-SGD, which experiences substantial accuracy degradation due to excessive noise injection, Gradient Dropout successfully reduces information leakage while maintaining accuracy close to that of the baseline.

C. Privacy Leakage on Adaptive attacker

To evaluate the robustness of our proposed method, we further conduct the experiments with adaptive attackers, including Mjölnir [32] and LTI [38]. Mjölnir employs a gradient diffusion model to denoise protected gradients prior to inversion, whereas LTI directly learns a mapping from gradients to training samples using auxiliary data with distributions similar to those of local clients. Experiments are conducted on FashionMNIST for Mjölnir and CIFAR-10 for LTI, both using a LeNet-5 model with batch sizes of 1 and 4 respectively, following the original experimental settings [32, 38]. We compare our proposed defence ($p = 0.6$, noise standard deviation 5×10^{-3}) against gradient noise baselines with standard deviations of 0.008 for Mjölnir and LTI, as specified in the original papers. From the results in Figure 6, we can

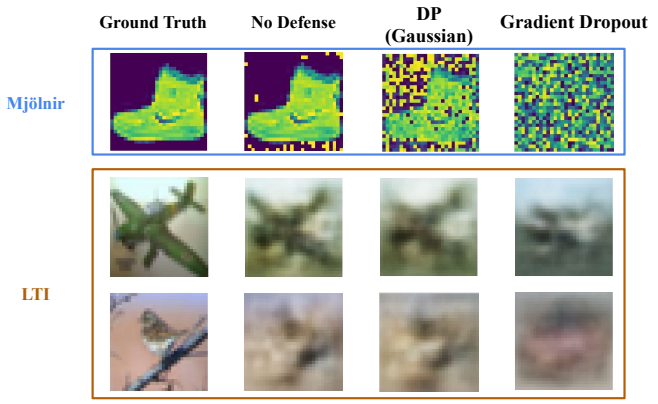


Fig. 6. Comparison of reconstructed images from differential privacy and our proposed method against gradient inversion attacks from Mjölnir and LTI.

see that both Mjölnir and LTI are able to reconstruct the training data when no defense is applied, and can roughly recover the objects in the images under DP-SGD defense. However, they are unable to reconstruct the training images when our proposed method is applied. Notably, our defense remains effective even when Mjölnir attempts gradient denoising. Unlike DP-SGD, where the cumulative noise increases with the number of training iterations, our method dynamically changes the gradient positions that are scaled and replaced at each iteration. This forces the attack model to identify and compensate for varying noise locations, which is substantially more challenging. For LTI, the attack assumes access to an auxiliary dataset with a distribution similar to that of the local clients. As a result, the reconstructed images roughly reflect the global data distribution; however, they fail to recover fine-grained details and accurate structures, capturing only coarse color information. These results demonstrate that the proposed defense remains effective against adaptive attackers such as Mjölnir and LTI.

D. Privacy Leakage on High Resolution Image Dataset

To evaluate the effectiveness of our defense approach on high-resolution image datasets, we conduct experiments on the ImageNet and NIH chest x-ray datasets with the CI-Net attack, comparing our method to other defense mechanisms. The experiments are conducted with a batch size of 1, and the number of attack iterations is set to 30,000. For DP-SGD, we apply two noise configurations, both with a mean of 0 and standard deviations of 10^{-2} and 10^{-3}. For Soteria, the probability parameter is set to $p = 0.8$, and for our proposed method, we set probability $p = 0.8$ and a noise standard deviation of $\sigma = 5 \times 10^{-3}$.

The mean and standard deviation of SSIM, PSNR, and LPIPS are calculated over 10 experimental runs, with sample reconstructed images presented in Figure 7. The results show that our proposed method performs comparably to DP-SGD with a large noise scale (10^{-2}), effectively generating diverse gradients that prevent data leakage from high-resolution images while maintaining model performance. In contrast, DP-SGD introduces excessive noise at larger scales, significantly degrading model performance, which highlights the challenge of balancing privacy and utility.

Although Soteria offers partial protection, it proves less robust than our proposed method. Soteria perturbs gradients only in the fully connected layer, leaving convolutional layers unprotected. As illustrated in the *Soteria (excluding FC)* column, attackers can accurately reconstruct training images by disregarding the fully connected layer gradients, underscoring the vulnerability of Soteria when gradient perturbation is limited to a single layer.

To evaluate model performance when applying our defense to a high-resolution dataset, we conduct training on the NIH dataset in a federated learning setting with three clients. The dataset is randomly partitioned across the three clients. We use a batch size of 16 and train the model for a total of 50 iterations. The learning rate is initialized at 2×10^{-2}, reduced to 2×10^{-3} at iteration 10, and further reduced to 2×10^{-5} at iteration 20. Figure 8 shows that when the noise scale varies from 5×10^{-3} to 10^{-1}, the testing accuracy remains comparable to the no-defense baseline. In contrast, when the noise magnitude becomes excessively large, even with a replacement ratio of 20%, model performance degrades, and the accuracy decreases from 0.67 to 0.61. Notably, when $p = 0.8$ and the noise scale is 5×10^{-3}, the proposed defense preserves classification accuracy while protecting the training data. Overall, the results indicate that appropriate choices of the replacement ratio p and noise scale allow the proposed defense to preserve model performance while limiting reconstruction effectiveness.

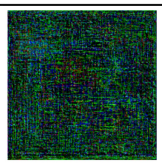
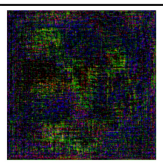
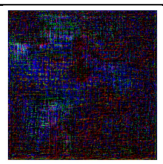
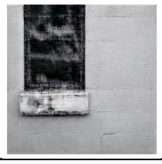
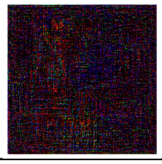
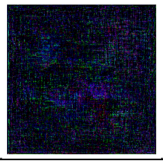
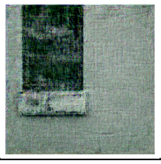
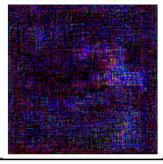
Overall, our proposed Gradient Dropout method effectively mitigates reconstruction attacks on high-resolution images, achieving a superior trade-off between privacy and accuracy compared to both DP-SGD and Soteria.

E. Distribution Comparison of Original Gradients and Noise Replacements

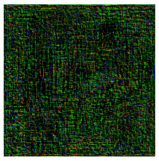
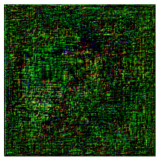
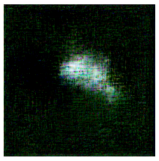
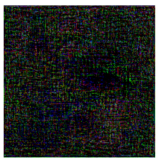
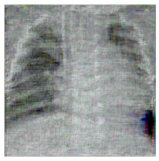
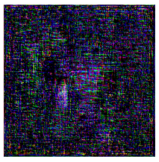
To demonstrate that the distributions of the original gradients and the noise replacements are statistically similar, we empirically analyse the gradient distributions in the first layer before and after applying the defense. The experimental settings follow those described in Section IV-D. As shown in Figure 9, the gradient distributions exhibit substantial overlap, making it difficult in practice for an attacker to separate injected noise from genuine gradient information. While formal indistinguishability is not claimed, this observation supports the effectiveness of the proposed defense against gradient reconstruction attacks.

V. Conclusion

In this paper, we presented Gradient Dropout, a novel defense framework designed to safeguard training data against reconstruction attacks in FL systems. By integrating gradient scaling with Gaussian noise, our method constructs a distinct gradient space that effectively mitigates reconstruction attacks while maintaining high model accuracy. A key contribution of this work is the provision of safety guarantees, ensuring

Ground Truth	No Defense	Gradient Dropout	DP-SGD(10^{-2})	DP-SGD(10^{-3})	Soteria	Soteria (Exclude FC)
						
						
SSIM	0.986 ± 0.004	0.024 ± 0.005	0.021 ± 0.005	0.186 ± 0.069	0.034 ± 0.002	0.704 ± 0.150
PSNR	24.917 ± 0.102	5.514 ± 0.594	6.873 ± 2.279	12.085 ± 3.715	5.197 ± 0.236	19.667 ± 4.310
LPIPS	0.2976 ± 0.0150	0.848 ± 0.058	0.807 ± 0.038	0.694 ± 0.041	0.806 ± 0.009	0.379 ± 0.078

(a) Results on ImageNet

Ground Truth	No Defense	Gradient Dropout	DP-SGD(10^{-2})	DP-SGD (10^{-3})	Soteria	Soteria (Exclude FC)
						
						
SSIM	0.885 ± 0.022	0.016 ± 0.003	0.021 ± 0.003	0.3619 ± 0.120	0.021 ± 0.006	0.338 ± 0.104
PSNR	23.279 ± 5.373	5.514 ± 0.594	6.481 ± 1.418	17.143 ± 3.021	5.657 ± 0.941	16.007 ± 3.109
LPIPS	0.308 ± 0.045	0.807 ± 0.035	0.842 ± 0.057	0.655 ± 0.054	0.821 ± 0.041	0.648 ± 0.055

(b) Results on the NIH Chest X-ray dataset

Fig. 7. Sample reconstructed images from the ImageNet and NIH Chest X-ray datasets using CI-Net under various defense methods, including No Defense, DP-SGD, Soteria, and Soteria (excluding FC), compared to the proposed Gradient Dropout. The mean and standard deviation of the Structural Similarity Index (SSIM) are reported for each method, demonstrating their effectiveness in mitigating reconstruction attacks.

that the protected gradients remain sufficiently distinct from the original gradients, thereby preventing data leakage. Additionally, we establish convergence guarantees for convex problems, enabling participants to balance privacy protection and model performance effectively. Comprehensive experiments on the CIFAR-10 and ImageNet datasets demonstrate that Gradient Dropout achieves an excellent trade-off between privacy and model performance. Specifically, it outperforms existing defense mechanisms like DP-SGD and Soteria, offering high levels of data protection with minimal impact on accuracy. Moreover, our method does not require parameter optimization during each training iteration, resulting in reduced computational cost. These findings underscore the practical applicability of Gradient Dropout in privacy-sensitive domains such as healthcare and finance, where both data confidentiality and model performance are essential. While Gradient Dropout offers significant advantages, its effectiveness depends on the choice of noise scale and probability parameters, which influence the balance between privacy and utility. Future research could explore adaptive parameter tuning strategies to further optimize this trade-off across diverse distributed learning scenarios. Additionally, while our current evaluation focuses on convolutional neural networks, investigating the effectiveness of Gradient Dropout on Transformer-based architectures represents an important direction for future research, given the distinct characteristics of attention mechanisms.

APPENDIX

A. *Proof for Theorem 1*

Proof: Let us consider the gradient leakage case, where the attacker use some images $(\tilde{x}, \tilde{y})$ to generate $\nabla\tilde{w}_t$ and minimize its gap with the true gradient ∇w_t:

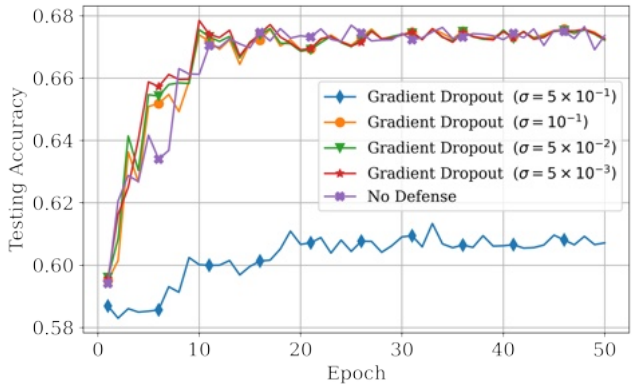


Fig. 8. Comparison of Gradient Dropout with varying noise scales in terms of training stability and model accuracy for ResNet-18 on the NIH dataset with three clients. All experiments are conducted with a batch size of 16. For the proposed method, the probability parameter p is fixed at 0.8, while the noise standard deviation varies from 5×10^{-3} to 5×10^{-1}.

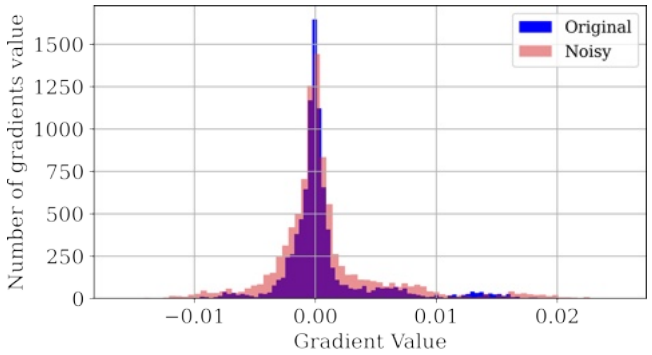


Fig. 9. Comparison of gradient distributions in the first layer of ResNet-18 during training on the NIH dataset.

$$\ell := \frac{1}{N}\sum_{i}^{N} \|\nabla w_t^i - \nabla \tilde{w}_t^i\|^2, \tag{7}$$

where i denotes the i-th component of the model.

Note in the ideal case where $(\tilde{x}, \tilde{y}) = (x, y)$, we have an zero total loss and the attacker fully reconstructs the hidden images.

But with gradient dropout, each component of ℓ can be obtained by:

$$\ell^i = \begin{cases} \frac{1}{N}\|s\nabla w_t - \nabla w_t\|^2, & \text{w.p. } p, \\ \frac{1}{N}\|\sigma B_t - \nabla w_t\|^2, & \text{w.p. } 1-p, \end{cases} \tag{8}$$

where $B_t \sim N(0,1)$.

$$\begin{aligned}\mathbb{E}_w[\ell^i] &= \frac{1}{N}p(s-1)^2\|\nabla w_t\|^2 + \frac{1}{N}(1-p)\|\sigma B_t - \nabla w_t\|^2 \\ &= \frac{1}{N}\frac{(1-p)^2}{p}\|\nabla w_t\|^2 + \frac{1}{N}(1-p)\sigma^2\|B_t\|^2 \\ &\quad - 2(1-p)\sigma B_t \cdot \nabla w_t + \frac{1}{N}(1-p)^2\|\nabla w_t\|^2,\end{aligned}$$

Consequently,

$$\begin{aligned}\mathbb{E}_{w,B}[\ell^i] &= \frac{1}{N}\frac{(1-p)^2}{p}\|\nabla w_t\|^2 + \frac{1}{N}(1-p)\sigma^2\mathbb{E}\|B_t\|^2 \\ &\quad + \frac{1}{N}(1-p)^2\|\nabla w_t\|^2, \\ &\geq \frac{1}{N}(1-p)\sigma^2\end{aligned}$$

To guarantee $\mathbb{E}_{w,B}[\ell] > \ell_0$, it is sufficient to have

$$p < 1 - \frac{\ell_0}{\sigma^2}.$$

■

B. *Proof of Theorem 2*

Proof: A fundamental tool to analyze the dynamics of stochastic differential equations is Itô's formula [60]. Consider a stochastic process $X(t)$ governed by the Itô process

$$dX(t) = \mu(t)\,dt + \sigma(t)\,dB(t),$$

where $\mu(t)$ is the drift term, $\sigma(t)$ is the volatility term, and $B(t)$ is a Brownian motion. For a twice differentiable function $f(X(t))$, Itô's formula states

$$df(X(t)) = \frac{\partial f}{\partial X(t)}\,dX(t) + \frac{1}{2}\frac{\partial^2 f}{\partial X(t)^2}(dX(t))^2.$$

Using $(dB(t))^2 = dt$, this becomes

$$df(X(t)) = \frac{\partial f}{\partial X(t)}\,(\mu(t)\,dt + \sigma(t)\,dB(t)) + \frac{1}{2}\frac{\partial^2 f}{\partial X(t)^2}\sigma^2(t)\,dt.$$

We now apply this to the weight dynamics. Let $\delta(t) = \mathbb{E}[\tilde{w}(t)] - w(t)$ denote the difference between the stochastic and deterministic trajectories at time t. From the continuous-time approximations in Eqs. (4) and (5), we obtain

$$d\delta(t) = -\left(\nabla f(\tilde{w}(t)) - \nabla f(w(t))\right)dt - (1-p)\sigma\,dB(t).$$

Consider the squared norm of the weight difference $\delta^2(t) := \|\delta(t)\|^2$. Applying Itô's formula to $f(\delta) = \|\delta\|^2$ yields

$$\begin{aligned}d\delta^2(t) &= 2\delta(t)^\top d\delta(t) + \big(d\delta(t)\big)^\top d\delta(t) \\ &= -2\left[\mathbb{E}[\tilde{w}(t)] - w(t)\right]^\top \left[\nabla f(\tilde{w}(t)) - \nabla f(w(t))\right]dt \\ &\quad - 2(1-p)\sigma\left[\mathbb{E}[\tilde{w}(t)] - w(t)\right]^\top dB(t) + (1-p)^2\sigma^2\,dt \\ &= \Big[-2\delta(t)^\top\big(\nabla f(\tilde{w}(t)) - \nabla f(w(t))\big) + (1-p)^2\sigma^2\Big]dt \\ &\quad - 2(1-p)\sigma^2\delta(t)^\top dB(t).\end{aligned}$$

We have used the standard properties of Brownian motion:

$$dt \cdot dt = 0, \quad dB(t) \cdot dt = 0, \quad dB(t) \cdot dB(t) = dt.$$

Taking expectations on both sides and using the fact that the stochastic integral has zero mean, we obtain

$$\frac{d}{dt}\mathbb{E}\big[\delta^2(t)\big] = -2\mathbb{E}\Big[\delta(t)^\top\big(\nabla f(\tilde{w}(t)) - \nabla f(w(t))\big)\Big] + (1-p)^2\sigma^2.$$

Now assume that ∇f is Lipschitz continuous with constant $L > 0$, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\| \quad \text{for all } x, y.$$

Then we have

$$\begin{aligned} &-2\delta(t)^\top \big(\nabla f(\tilde{w}(t)) - \nabla f(w(t))\big) \\ &\le 2\|\delta(t)\| \,\big\|\nabla f(\tilde{w}(t)) - \nabla f(w(t))\big\| \\ &\le 2L\|\delta(t)\|^2 = 2L\delta^2(t), \end{aligned}$$

and therefore

$$\frac{d}{dt}\mathbb{E}\big[\delta^2(t)\big] \le 2L\,\mathbb{E}\big[\delta^2(t)\big] + (1-p)^2\sigma^2.$$

This is a linear differential inequality of the form $y'(t) \le 2Ly(t) + (1-p)^2\sigma^2$ with $y(t) = \mathbb{E}[\delta^2(t)]$. By Grönwall's inequality, we obtain

$$\mathbb{E}\big[\delta^2(t)\big] \le \frac{(1-p)^2\sigma^2}{2L}\left(e^{2Lt} - 1\right).$$

This proves Theorem 2. ∎

C. Proof of Theorem 3

Proof: We start from the same expression for the dynamics of $\mathbb{E}[\delta^2(t)]$ obtained in the proof of Theorem 2:

$$\frac{d}{dt}\mathbb{E}\big[\delta^2(t)\big] = -2\mathbb{E}\Big[\delta(t)^\top\big(\nabla f(\tilde{w}(t)) - \nabla f(w(t))\big)\Big] + (1-p)^2\sigma^2.$$

Now assume that f is μ-strongly convex with parameter $\mu > 0$, i.e.,

$$\big(\nabla f(x) - \nabla f(y)\big)^\top (x - y) \ge \mu\|x - y\|^2 \quad \text{for all } x, y.$$

By choosing $x = \tilde{w}(t)$ and $y = w(t)$, we have

$$\big(\nabla f(\tilde{w}(t)) - \nabla f(w(t))\big)^\top \big(\tilde{w}(t) - w(t)\big) \ge \mu\|\tilde{w}(t) - w(t)\|^2.$$

Recalling that $\delta(t) = \mathbb{E}[\tilde{w}(t)] - w(t)$ and treating $\delta(t)$ as the difference variable in the bound, we obtain

$$-2\delta(t)^\top\big(\nabla f(\tilde{w}(t)) - \nabla f(w(t))\big) \le -2\mu\|\delta(t)\|^2 = -2\mu\delta^2(t).$$

Taking expectations yields

$$\frac{d}{dt}\mathbb{E}\big[\delta^2(t)\big] \le -2\mu\,\mathbb{E}\big[\delta^2(t)\big] + (1-p)^2\sigma^2.$$

Let $y(t) = \mathbb{E}[\delta^2(t)]$. The inequality above has the form

$$y'(t) \le -2\mu y(t) + (1-p)^2\sigma^2.$$

The corresponding equality ODE $y'(t) = -2\mu y(t) + (1-p)^2\sigma^2$ has solution

$$y(t) = \left(y(0) - \frac{(1-p)^2\sigma^2}{2\mu}\right)e^{-2\mu t} + \frac{(1-p)^2\sigma^2}{2\mu}.$$

Assuming $y(0) = 0$ (i.e., the two trajectories start from the same initialization), we obtain

$$y(t) \le \frac{(1-p)^2\sigma^2}{2\mu}\left(1 - e^{-2\mu t}\right),$$

and thus

$$\mathbb{E}\big[\delta^2(t)\big] \le \frac{(1-p)^2\sigma^2}{2\mu}\left(1 - e^{-2\mu t}\right),$$

which proves Theorem 3. ∎

Theorem 4 (High-Dimensional Cosine Similarity Bound under Gradient Dropout)**.** *Let $g \in \mathbb{R}^d$ be a deterministic gradient vector with $\|g\|_2 > 0$, and suppose it satisfies the regularity conditions*

$$\alpha d \;\le\; \|g\|_2^2 \;\le\; \beta d, \qquad \|g\|_\infty \;\le\; G, \tag{9}$$

for some fixed constants $0 < \alpha \le \beta < \infty$ and $G > 0$ independent of d. Let $\tilde{g} \in \mathbb{R}^d$ be the perturbed gradient produced by the gradient dropout mechanism with retention probability $p \in (0,1)$ and noise variance $\sigma^2 > 0$, i.e., for each coordinate $j = 1, \dots, d$,

$$\tilde{g}_j = \begin{cases} \dfrac{g_j}{p}, & \text{with probability } p, \\ \varepsilon_j, & \text{with probability } 1-p, \end{cases}$$

where $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$, and all $\{(m_j, \varepsilon_j)\}_{j=1}^d$ are independent. Define the cosine similarity between g and $\tilde{g}$ as

$$\cos(g, \tilde{g}) := \frac{\langle g, \tilde{g}\rangle}{\|g\|_2\,\|\tilde{g}\|_2}.$$

Then, for any $\varepsilon > 0$, there exist constants $c > 0$ and $d_0 \in \mathbb{N}$ (depending only on $\alpha, \beta, G, p, \sigma, \varepsilon$ but not on d) such that, for all $d \ge d_0$,

$$\mathbb{P}(\cos(g, \tilde{g}) \;\ge\; \sqrt{p} + \varepsilon) \;\le\; 4e^{-cd}. \tag{10}$$

In particular, since $p < 1$, the cosine similarity between the original and perturbed gradients is strictly bounded away from 1 with probability tending to 1 as $d \to \infty$.

Proof: Throughout the proof we condition on the fixed vector g satisfying (9), and all expectations and probabilities are with respect to the randomness in the masks and noise.

a) Step 1: Expectations of numerator and denominator.: Define the per-coordinate random variables

$$X_j := g_j\tilde{g}_j, \qquad Y_j := \tilde{g}_j^2, \qquad j = 1, \dots, d.$$

Then

$$\langle g, \tilde{g}\rangle = \sum_{j=1}^d X_j, \qquad \|\tilde{g}\|_2^2 = \sum_{j=1}^d Y_j.$$

By the definition of the mechanism,

$$\tilde{g}_j = \begin{cases} g_j/p, & \text{with prob. } p, \\ \varepsilon_j, & \text{with prob. } 1-p, \end{cases}$$

with $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$ and independent of the Bernoulli mask. Hence

$$\mathbb{E}[\tilde{g}_j \mid g] = p \cdot \frac{g_j}{p} + (1-p)\cdot 0 = g_j,$$

and

$$\mathbb{E}[\tilde{g}_j^2 \mid g] = p\left(\frac{g_j}{p}\right)^2 + (1-p)\mathbb{E}[\varepsilon_j^2] = \frac{1}{p}g_j^2 + (1-p)\sigma^2.$$

Therefore,

$$\mathbb{E}[\langle g,\tilde{g}\rangle \mid g] = \sum_{j=1}^{d} \mathbb{E}[X_j \mid g] = \sum_{j=1}^{d} g_j\, \mathbb{E}[\tilde{g}_j \mid g] = \|g\|_2^2,$$
$$\mathbb{E}[\|\tilde{g}\|_2^2 \mid g] = \sum_{j=1}^{d} \mathbb{E}[Y_j \mid g] = \sum_{j=1}^{d} \Big(\frac{1}{p} g_j^2 + (1-p)\sigma^2\Big)$$
$$= \frac{1}{p}\|g\|_2^2 + d(1-p)\sigma^2.$$

Define the empirical averages and their expectations:

$$\bar{X}_d := \frac{1}{d}\sum_{j=1}^{d} X_j, \quad \mu_1 := \mathbb{E}[\bar{X}_d \mid g] = \frac{\|g\|_2^2}{d},$$

$$\bar{Y}_d := \frac{1}{d}\sum_{j=1}^{d} Y_j, \quad \mu_2 := \mathbb{E}[\bar{Y}_d \mid g] = \frac{1}{p}\frac{\|g\|_2^2}{d} + (1-p)\sigma^2.$$

By (9), we have

$$\mu_1 \in [\alpha,\beta], \quad \mu_2 \in \Big[\frac{\alpha}{p} + (1-p)\sigma^2,\ \frac{\beta}{p} + (1-p)\sigma^2\Big].$$

b) Step 2: Sub-exponential tails and concentration.: Using the boundedness of g in ℓ_∞, we can bound the tails of X_j and Y_j. Specifically:

- $X_j = g_j\tilde{g}_j$ equals g_j^2/p with prob. p and $g_j\varepsilon_j$ with prob. $(1-p)$. Since $|g_j| \le G$, the Gaussian part $g_j\varepsilon_j$ is sub-Gaussian, and thus X_j is sub-exponential with a ψ_1-norm bounded by a constant $K_X < \infty$ depending only on G, p, σ.
- $Y_j = \tilde{g}_j^2$ equals $(g_j/p)^2$ with prob. p and ε_j^2 with prob. $(1-p)$. Gaussian squares are sub-exponential, so Y_j is also sub-exponential with ψ_1-norm bounded by a constant $K_Y < \infty$ depending only on G, p, σ.

By Bernstein's inequality for sub-exponential random variables, there exist absolute constants $c_1, c_2 > 0$ (depending only on K_X, K_Y) such that, for all $t > 0$,

$$\mathbb{P}\Big(|\bar{X}_d - \mu_1| \ge t \,\Big|\, g\Big) \le 2\exp\Big(-c_1 d \min\big(t^2/K_X^2,\ t/K_X\big)\Big),$$
$$\mathbb{P}\Big(|\bar{Y}_d - \mu_2| \ge t \,\Big|\, g\Big) \le 2\exp\Big(-c_2 d \min\big(t^2/K_Y^2,\ t/K_Y\big)\Big).$$

Fix $\delta > 0$ small (to be specified later). For sufficiently large d, the quadratic term dominates, and there exists $c > 0$ such that

$$\mathbb{P}\Big(|\bar{X}_d - \mu_1| \ge \delta \,\Big|\, g\Big) \le 2e^{-cd},$$
$$\mathbb{P}\Big(|\bar{Y}_d - \mu_2| \ge \delta \,\Big|\, g\Big) \le 2e^{-cd}.$$

Define the high-probability event

$$\mathcal{E}_d := \Big\{|\bar{X}_d - \mu_1| \le \delta \text{ and } |\bar{Y}_d - \mu_2| \le \delta\Big\}.$$

By union bound,

$$\mathbb{P}(\mathcal{E}_d^c \mid g) \le 4e^{-cd}. \tag{11}$$

c) Step 3: Bounding the cosine on the event $\mathcal{E}_d$.: On $\mathcal{E}_d$, we have

$$\langle g,\tilde{g}\rangle = d\bar{X}_d \le d(\mu_1+\delta), \qquad \|\tilde{g}\|_2^2 = d\bar{Y}_d \ge d(\mu_2-\delta).$$

Therefore,

$$\cos(g,\tilde{g}) = \frac{\langle g,\tilde{g}\rangle}{\|g\|_2\|\tilde{g}\|_2} \le \frac{d(\mu_1+\delta)}{\|g\|_2\sqrt{d(\mu_2-\delta)}}$$
$$= \frac{\mu_1+\delta}{(\|g\|_2/\sqrt{d})\,\sqrt{\mu_2-\delta}}.$$

Using (9), $\|g\|_2/\sqrt{d} \ge \sqrt{\alpha}$ and $\mu_1 \le \beta$, hence

$$\cos(g,\tilde{g}) \le \frac{\beta+\delta}{\sqrt{\alpha}\,\sqrt{\mu_2-\delta}} \quad \text{on } \mathcal{E}_d. \tag{12}$$

From the bounds on μ_2, we have

$$\mu_2 \ge \underline{\mu}_2 := \frac{\alpha}{p} + (1-p)\sigma^2.$$

Choose $\delta > 0$ small enough such that $\mu_2 - \delta \ge \underline{\mu}_2/2 > 0$. Then

$$\sqrt{\mu_2-\delta} \ge \sqrt{\underline{\mu}_2/2},$$

and (12) implies

$$\cos(g,\tilde{g}) \le \frac{\beta+\delta}{\sqrt{\alpha}\,\sqrt{\underline{\mu}_2/2}} =: C_\delta, \quad \text{on } \mathcal{E}_d. \tag{13}$$

The constant C_δ depends only on $\alpha, \beta, p, \sigma, \delta$.

Next, consider the function

$$\phi(u) := \frac{u}{\sqrt{\frac{1}{p}u + (1-p)\sigma^2}}, \qquad u > 0.$$

For any $u > 0$,

$$\phi(u) \le \frac{u}{\sqrt{\frac{1}{p}u}} = \sqrt{p},$$

so $\sup_{u>0}\phi(u) \le \sqrt{p}$. Moreover, for $u \in [\alpha,\beta]$, $\phi(u)$ is continuous and thus attains a maximum on that compact interval. Denote

$$\phi^\star := \max_{u\in[\alpha,\beta]} \phi(u) \le \sqrt{p}.$$

By the definitions of μ_1 and μ_2, $\phi(\mu_1)$ corresponds to the "ideal" (no-fluctuation) cosine constructed from the expectations μ_1 and μ_2. As $\delta \to 0$, C_δ in (13) converges to a constant no larger than $\phi^\star \le \sqrt{p}$. Hence, for any fixed $\varepsilon > 0$, we can choose $\delta > 0$ sufficiently small such that

$$C_\delta \le \sqrt{p} + \frac{\varepsilon}{2}.$$

d) Step 4: High-probability bound and conclusion.: With this choice of δ, and for d large enough so that (11) holds, we obtain

$$\mathbb{P}\Big(\cos(g, \tilde{g}) \leq \sqrt{p} + \varepsilon \,\Big|\, g \Big) \geq \mathbb{P}(\mathcal{E}_d \mid g) \geq 1 - 4e^{-cd}.$$

Equivalently,

$$\mathbb{P}\Big(\cos(g, \tilde{g}) \geq \sqrt{p} + \varepsilon \,\Big|\, g \Big) \leq 4e^{-cd},$$

for all $d \geq d_0$, where d_0 is large enough (depending on $\alpha, \beta, G, p, \sigma, \varepsilon$). This proves (10) and completes the proof. ■

REFERENCES

[1] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *CoRR*, vol. abs/1610.05492, 2016. [Online]. Available: https://arxiv.org/pdf/1610.05492

[2] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records," *International journal of medical informatics*, vol. 112, pp. 59–67, 2018.

[3] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[4] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific reports*, vol. 10, no. 1, p. 12598, 2020.

[5] W. Yang, Y. Zhang, K. Ye, L. Li, and C.-Z. Xu, "FFD: A federated learning based method for credit card fraud detection," in *Big Data–BigData 2019: 8th International Congress, Held as Part of the Services Conference Federation, SCF 2019, San Diego, CA, USA, June 25–30, 2019, Proceedings 8*. Springer, 2019, pp. 18–32.

[6] D. Kawa, S. Punyani, P. Nayak, A. Karkera, and V. Jyotinagar, "Credit risk assessment from combined bank records using federated learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 6, no. 4, pp. 1355–1358, 2019.

[7] R. Shokri, M. Stronati, C. S ong, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[8] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 739–753.

[9] H. Hu, Z. Salcic, L. Sun, G. Dobbie, and X. Zhang, "Source inference attacks in federated learning," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 1102–1107.

[10] O. Zari, C. Xu, and G. Neglia, "Efficient passive membership inference attack in federated learning," *CoRR*, vol. abs/2111.00430, 2021. [Online]. Available: https://arxiv.org/abs/2111.00430

[11] Y. Gu, Y. Bai, and S. Xu, "CS-MIA: Membership inference attack based on prediction confidence series in federated learning," *Journal of Information Security and Applications*, vol. 67, p. 103201, 2022.

[12] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 691–706.

[13] F. Mo, A. Borovykh, M. Malekzadeh, H. Haddadi, and S. Demetriou, "Layer-wise characterization of latent information leakage in federated learning," *CoRR*, vol. abs/2010.08762, 2021. [Online]. Available: https://arxiv.org/abs/2010.08762

[14] X. Luo, Y. Wu, X. Xiao, and B. C. Ooi, "Feature inference attack on model predictions in vertical federated learning," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021, pp. 181–192.

[15] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 603–618.

[16] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 774–14 784.

[17] B. Zhao, K. R. Mopuri, and H. Bilen, "iDLG: Improved deep leakage from gradients," *CoRR*, vol. abs/2001.02610, 2020. [Online]. Available: https://arxiv.org/pdf/2001.02610

[18] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients - how easy is it to break privacy in federated learning?" in *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2020.

[19] H. Yin, A. Mallya, A. Vahdat, J. M. Alvarez, J. Kautz, and P. Molchanov, "See through gradients: Image batch recovery via gradinversion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 337–16 346.

[20] J. Jeon, K. Lee, S. Oh, J. Ok *et al.*, "Gradient inversion with generative image prior," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 898–29 908, 2021.

[21] E. Sotthiwat, L. Zhen, C. Zhang, Z. Li, and R. S. M. Goh, "Generative image reconstruction from gradients," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[22] C. Zhang, Z. Xiaoman, E. Sotthiwat, Y. Xu, P. Liu, L. Zhen, and Y. Liu, "Generative gradient inversion via over-parameterized networks in federated learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5126–5135.

[23] B. Li, H. Gu, R. Chen, J. Li, C. Wu, N. Ruan, X. Si, and L. Fan, "Temporal gradient inversion attacks with robust optimization," *IEEE Transactions on Dependable and Secure Computing*, 2025.

[24] W. Yu, H. Fang, B. Chen, X. Sui, C. Chen, H. Wu, S.-T. Xia, and K. Xu, "GI-NAS: Boosting gradient inversion attacks through adaptive neural architecture search," *IEEE Transactions on Information Forensics and Security*, 2025.

[25] M. Balunovic, D. I. Dimitrov, R. Staab, and M. Vechev, "Bayesian framework for gradient leakage," in *International Conference on Learning Representations*, 2022.

[26] A. El Ouadrhiri and A. Abdelhadi, "Differential privacy for deep and federated learning: A survey," *IEEE access*, vol. 10, pp. 22 359–22 380, 2022.

[27] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE transactions on information forensics and security*, vol. 15, pp. 3454–3469, 2020.

[28] R. Xue, K. Xue, B. Zhu, X. Luo, T. Zhang, Q. Sun, and J. Lu, "Differentially private federated learning with an adaptive noise

mechanism," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 74–87, 2023.
[29] H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, H. Möllering, T. D. Nguyen, P. Rieger, A.-R. Sadeghi, T. Schneider, H. Yalame *et al.*, "Safelearn: Secure aggregation for private federated learning," in *2021 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2021, pp. 56–62.
[30] C. Zhang, S. Ekanut, L. Zhen, and Z. Li, "Augmented multi-party computation against gradient leakage in federated learning," *IEEE Transactions on Big Data*, 2022.
[31] Z. You, X. Dong, S. Li, X. Liu, S. Ma, and Y. Shen, "Local differential privacy is not enough: A sample reconstruction attack against federated learning with local differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 1519–1534, 2024.
[32] X. Liu, S. Cai, Q. Zhou, S. Guo, R. Li, and K. Lin, "Mjölnir: Breaking the shield of perturbation-protected gradients via adaptive diffusion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 25, 2025, pp. 26 308–26 316.
[33] A. C.-C. Yao, "How to generate and exchange secrets," in *27th annual symposium on foundations of computer science (Sfcs 1986)*. IEEE, 1986, pp. 162–167.
[34] E. Sotthiwat, L. Zhen, Z. Li, and C. Zhang, "Partially encrypted multi-party computation for federated learning," in *2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*. IEEE, 2021, pp. 828–835.
[35] Y. Lin, S. Han, H. Mao, Y. Wang, and B. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *International Conference on Learning Representations*, 2018.
[36] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
[37] Z. Zhang, Z. Tianqing, W. Ren, P. Xiong, and K.-K. R. Choo, "Preserving data privacy in federated learning through large gradient pruning," *Computers & Security*, vol. 125, p. 103039, 2023.
[38] R. Wu, X. Chen, C. Guo, and K. Q. Weinberger, "Learning to invert: Simple adaptive attacks for gradient inversion in federated learning," in *Uncertainty in Artificial Intelligence*. PMLR, 2023, pp. 2293–2303.
[39] X. Liu, S. Cai, L. Li, R. Zhang, and S. Guo, "MGIA: Mutual gradient inversion attack in multi-modal federated learning (student abstract)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 16 270–16 271.
[40] X. Liu, S. Cai, R. He, and J. Yuan, "Mutual gradient inversion: Unveiling privacy risks of federated learning on multi-modal signals," *IEEE Signal Processing Letters*, 2024.
[41] R. Zhang, S. Guo, J. Wang, X. Xie, and D. Tao, "A survey on gradient inversion: Attacks, defenses and future directions," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2023, pp. 5678–685.
[42] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.
[43] C. Zhao, S. Zhao, M. Zhao, Z. Chen, C.-Z. Gao, H. Li, and Y.-a. Tan, "Secure multi-party computation: theory, practice and applications," *Information Sciences*, vol. 476, pp. 357–372, 2019.
[44] J. Sun, A. Li, B. Wang, H. Yang, H. Li, and Y. Chen, "Provable defense against privacy leakage in federated learning from representation perspective," *CoRR*, vol. abs/2012.06043, 2020. [Online]. Available: https://arxiv.org/abs/2012.06043
[45] Z. He, T. Zhang, and R. B. Lee, "Attacking and protecting data privacy in edge–cloud collaborative inference systems," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9706–9716, 2020.
[46] D. Scheliga, P. Mäder, and M. Seeland, "Dropout is not all you need to prevent gradient leakage," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, 2023, pp. 9733–9741.
[47] K. Gao, T. Zhu, D. Ye, and W. Zhou, "Defending against gradient inversion attacks in federated learning via statistical machine unlearning," *Knowledge-Based Systems*, vol. 299, p. 111983, 2024.
[48] H. Zhou, Y. Chen, Z. Qin, X. Deng, and Z. Peng, "Thwarting gradient inversion in federated learning via generative shadow mapping defense," *Journal of Systems Architecture*, p. 103671, 2025.
[49] Z. Ye, W. Luo, Q. Zhou, Z. Zhu, Y. Shi, and Y. Jia, "Gradient inversion attacks: Impact factors analyses and privacy enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
[50] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th computer security foundations symposium (CSF)*. IEEE, 2017, pp. 263–275.
[51] B. Chang, L. Meng, E. Haber, L. Ruthotto, D. Begert, and E. Holtham, "Reversible architectures for arbitrarily deep residual neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
[52] E. Haber and L. Ruthotto, "Stable architectures for deep neural networks," *Inverse problems*, vol. 34, no. 1, p. 014004, 2017.
[53] C. Zhang, C. Jingpu, Y. Xu, and Q. Li, "Parameter-efficient fine-tuning with controls," in *Forty-first International Conference on Machine Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=C4nalr0DoE
[54] A. Krizhevsky, V. Nair, and G. Hinton, "CIFAR-10 (Canadian Institute for Advanced Research)." [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html
[55] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," *CoRR*, vol. abs/1708.07747, 2017. [Online]. Available: https://arxiv.org/abs/1708.07747
[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
[57] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.
[58] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[59] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
[60] K. Itô, "109. stochastic integral," *Proceedings of the Imperial Academy*, vol. 20, no. 8, pp. 519–524, 1944.