ADAN: ADAPTIVE DIRECTION-AWARE NOISE FOR DEFENDING ITERATIVE GRADIENT INVERSION ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Iterative training data reconstruction is a class of gradient inversion attacks (GIAs) in which an adversary repeatedly updates a dummy input to align its gradient with the gradients of a target model, progressively recovering the ground-truth training data. Existing GIAs operate under a curious-but-honest setting. This setting provides the attackers with a significant advantage because they can utilize the model parameters to calculate the exact gradients during iterative optimization. By contrast, in this work, we consider a more challenging black-box setting in which the attacker has no access to model parameters. Based on this setting, we propose **Adaptive Direction-Aware Noise (ADAN)**, a mechanism that injects input-dependent, non-isotropic Gaussian noise into each queried gradient. ADAN employs a neural network to learn per-coordinate noise magnitudes adapted to the input data distribution, effectively steering the optimization geometry of iterative GIAs. Empirical results show that ADAN substantially outperforms isotropic gradient perturbation methods across multiple datasets and attack variants.

1 INTRODUCTION

In distributed learning (Jochems et al., 2016; McMahan et al., 2017; Yang et al., 2019), each client updates its model using its local data and transmits gradients to a central server, which aggregates them to update a global model. Gradient sharing is also common in federated learning (FL) (Huang et al., 2021), but unlike distributed learning which involves a centrally coordinated distribution of data across clients, FL focuses on preserving data privacy by ensuring that client data remains localized. Although FL avoids directly sharing client data, recent studies (Phong et al., 2017; Zhu et al., 2019; Zhao et al., 2020) demonstrate that leaked gradients can still reveal sensitive information, enabling attackers to reconstruct the original training samples and posing serious privacy risks in practical learning systems. In this work, we use $\mathcal{L}(\boldsymbol{x}, \theta)$ to denote the loss function of a federated learning model parameterized by θ for input $\boldsymbol{x}$.

Gradient inversion attacks (GIAs) aim to reconstruct the training data with the leaked gradients and have been extensively studied (Huang et al., 2021; Phong et al., 2017; Zhu et al., 2019; Zhao et al., 2020; Wei et al., 2020; Geiping et al., 2020; Zhu & Blaschko, 2020; Wang et al., 2020; Yin et al., 2021; Wang et al., 2023; Jeon et al., 2021; Li et al., 2022b; Fang et al., 2023; Wu et al., 2023; Chen & Vikalo, 2024; Wu et al., 2025). These methods generally assume a *curious-but-honest* setting, where the central server is treated as a white-box adversary that responds faithfully to all queries while attempting to infer sensitive information from the client. In this scenario, the adversary's goal is to reconstruct clients' local training data from the gradients shared during federated learning.

However, the assumption of white-box access is often too stringent. In practical scenarios, adversaries typically operate in a black-box setting, possessing query access to the model while remaining agnostic to its internal parameters. Concretely, we consider a threat model illustrated in Figure 1, where an adversary intercepts the communication channel of a client. Specifically, We consider an adversary who has (1) **compromised the input interface** (the data provider feeding the client), allowing them to inject dummy data; (2) **compromised the output interface** (the client-to-server link), allowing them to observe the resulting gradients. Together, these realistic compromises effectively create a functional black-box query system. Similar black-box threat models are well investigated in membership inference attacks (MIAs) and model stealing Shokri et al. (2017); Tramèr

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

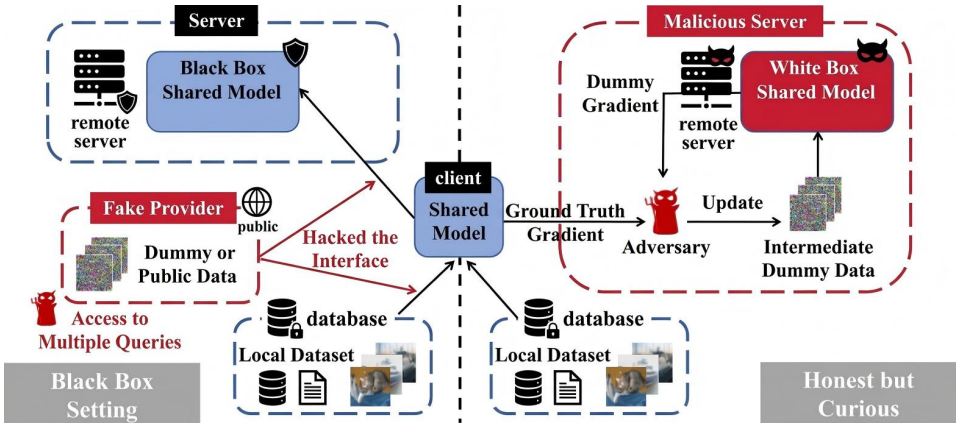


Figure 1: Illustration of the comparison between our proposed threat model and the conventional honest-but-curious assumption. On the left is our proposed black box threat model, while on the right is the conventional honest-but-curious setting.

et al. (2016), where model parameters are unknown and adversaries only have access to outputs. *All* attacks and defenses discussed in this work are evaluated under this proposed black-box threat model.

The defense mechanisms against GIAs are significantly different between white-box and black-box threat models. In the curious-but-honest setting, defense mechanisms are applied to the leaked gradients only, because the adversary has access to model parameters and can calculate the precise gradient of inputs designed by the adversary. By contrast, the defense mechanisms are applied to both the leaked gradients and any gradient query by the adversary in the black-box setting. That is to say, the defense mechanism is integrated into the model, and any gradient obtained via querying, including those during adversary's iterative optimization process, is subject to the defense.

Most existing defense mechanisms against GIAs rely on perturbing gradients. These approaches typically employ data-agnostic rules, such as injecting noise from a fixed distribution or applying gradient sparsification at a predefined ratio. However, the inherent vulnerability to GIAs varies significantly across different data samples. To address this limitation, we propose **Adaptive Direction-Aware Noise (ADAN)** to enhance defense mechanisms under black-box settings. ADAN dynamically generates data-dependent noise tailored to the specific gradients of each input. By strategically sampling noise from data-dependent directions, ADAN protects data privacy against GIAs while better preserving model utility.

We summarize our contributions as following points:

- We investigate defense mechanisms under black-box threat models.
- We propose Adaptive Direction-Aware Noise (ADAN), a defense tailored to iterative gradient inversion attacks. ADAN injects input-dependent, non-isotropic Gaussian noise derived from the attacker's optimization objective.
- Intensive experiments demonstrate that ADAN can achieve a better trade-off between model utillity and data privacy protection against GIAs.

2 Related Works

Gradient Inversion Attacks. Gradient inversion attacks (GIA) commonly arise in the Federated Learning (FL) scenario. In this setting, gradients transmitted from client nodes to a central server can be intercepted and exploited by an adversary to reconstruct private training data. The reconstructed data can be derived from analytical methods, optimization process or generative models.

Analytic-based methods recover private data through direct derivation. One strand of research focuses on gradient sparsification Fowl et al. (2021); Wen et al. (2022); Zhao et al. (2024), where specific gradient components are selectively retained while others are masked to facilitate reconstruction.

Another strand involves closed-form approaches Zhu & Blaschko (2020); Lu et al. (2022), which exploit the algebraic structure of the network to mathematically solve for the input data directly, thereby bypassing the need for iterative optimization loops. However, these analytical methods fundamentally rely on strong prerequisites, such as possessing full white-box access to the federated model parameters or the ability to maliciously manipulate the weights and biases of the model's final layer. As these assumptions contradict the black-box threat model established in this work, where the adversary cannot access or modify model parameters, we exclude these methods from our evaluation.

Deep leakage from gradients (DLG) was one of the early attempts Phong et al. (2017); Zhu et al. (2019) to utilise optimization for gradient inversion attacks. DLG optimizes a randomly initialized "dummy" input to estimate the training data by minimizing the distance between the dummy gradients and the leaked ground-truth gradients. Numerous subsequent studies have enhanced DLG, focusing on either improved optimization strategies or adaptation to more realistic, complex scenarios Wei et al. (2020); Geiping et al. (2020); Zhu & Blaschko (2020); Wang et al. (2020; 2023); Chen & Vikalo (2024).

Generative model-based methods optimize in a latent space instead of the pixel space, they leverage pre-trained generative models to synthesize estimates of the user data. Early attempts Yin et al. (2021) utilize pre-trained generative models to provide image priors for reconstruction. Building on this, GIAS Huang et al. (2021); Jeon et al. (2021) employs a Generative Adversarial Network (GAN) Goodfellow et al. (2014) and alternately searches both the latent space and the parameter space of the generator. To mitigate prohibitive computational complexity of GIAS, GGL Li et al. (2022b) optimizes only the latent code while using a fixed pre-trained GAN as the generator. GIFD Fang et al. (2023) modifies GGL by progressively updating the optimized layers of the GAN and systematically analyzing reconstruction performance under out-of-distribution (OOD) settings. DGGI Wu et al. (2025) further improve the performance by adopting a novel attack pipeline Yin et al. (2021) and replacing the GAN with a pre-trained diffusion model to generate image priors.

Defense Method against GIA. Perturbation-based defenses Abadi et al. (2016); Geyer et al.; Wei et al. (2021a); Kaissis et al. (2023) are common strategies to mitigate leakage involve modifying gradients before transmission. Noisy gradients, often grounded in Differential Privacy (DP), inject random noise (e.g., Gaussian) to provide theoretical privacy guarantees. However, achieving strict worst-case privacy often necessitates excessive noise, which significantly compromises model utility. Gradient Clipping Wei et al. (2021b) restricts gradient norms to a fixed threshold to limit influence. Yet, empirical studies Li et al. (2022b); Fang et al. (2023) indicate that attackers can estimate these clipping bounds from the received gradient norms to guide reconstruction effectively. Gradient Compression (or sparsification) Aji & Heafield (2017) transmits only the most significant components (e.g., Top-k) while zeroing out others. Nevertheless, since the implementation exposes the sparsity mask, attackers can exploit the locations of non-zero entries to recover the private training data Zhu et al. (2019); Li et al. (2022b); Fang et al. (2023).

3 PRELIMINARY

Iterative Attacks. Our method is motivated by the analysis on iterative optimization in gradient inversion attacks Zhu et al. (2019); Wei et al. (2020); Geiping et al. (2020); Wang et al. (2020); Zhu & Blaschko (2020); Yin et al. (2021); Jeon et al. (2021); Huang et al. (2021); Li et al. (2022b); Fang et al. (2023); Wang et al. (2023); Chen & Vikalo (2024); Wu et al. (2025). In this context, an adversary iteratively updates a randomly initialized dummy vector $\boldsymbol{v}_t$ to minimize the discrepancy between its induced gradient and the leaked gradient. Let G denote the function mapping $\boldsymbol{v}_t$ to the input space, G is an identity function in DLG Zhu et al. (2019) since it directly optimizes in the input space, other methods like GIAS Jeon et al. (2021) employ a generative model as G. At each iteration t, the dummy gradient is computed as $\boldsymbol{g}_t = \nabla_\theta \mathcal{L}(G(\boldsymbol{v}_t), \theta)$ and compared against the leaked gradient $\boldsymbol{g} = \nabla_\theta \mathcal{L}(\boldsymbol{x}, \theta)$. The vector $\boldsymbol{v}_t$ is then refined to minimize the mismatch $\mathcal{L}_{\text{GM}}(\boldsymbol{g}_t, \boldsymbol{g})$ via gradient descent with step size η:

$$\boldsymbol{v}_{t+1} = \boldsymbol{v}_t - \eta \nabla_{\boldsymbol{v}_t} \mathcal{L}_{\text{GM}}(\boldsymbol{g}_t, \boldsymbol{g}), \tag{1}$$

By calling (1) iteratively, the attacker exploits the landscape of the loss objective function and the reconstructed input $G(\boldsymbol{x}_t)$ gradually approaches the ground-truth input. In the white-box setting, the adversary has access to model parameters and can precisely calculate $\mathcal{L}_{\text{GM}}(\boldsymbol{g}_t, \boldsymbol{g})$. However, in

the black-box setting, the adversary has no access to model parameters, so it estimates $\mathcal{L}_{\text{GM}}(\boldsymbol{g}_t, \boldsymbol{g})$ through queries and zero-order methods.

The most popular choice for $\mathcal{L}_{\text{GM}}$ is an ℓ_2 distance between gradients, i.e., $\mathcal{L}_{\text{GM}}(\boldsymbol{g}_t, \boldsymbol{g}) = \|\boldsymbol{g}_t - \boldsymbol{g}\|_2^2$. In practice, many attacks adopt composite gradient matching objectives, where the ℓ_1 distance or cosine similarity between gradients Geiping et al. (2020); Li et al. (2022a) is combined with the ℓ_2 term to improve convergence and reconstruction fidelity. Additionally, image-space regularizers such as total variation Yin et al. (2021) are often introduced to enforce natural image priors during optimization. Despite these extensions, the ℓ_2 gradient matching loss serves as the backbone of most existing formulations. Accordingly, we conduct our theoretical analysis based on the ℓ_2 gradient matching loss for clarity and tractability. Importantly, our empirical results demonstrate that the insights derived from the ℓ_2-based analysis generalize well to alternative gradient matching losses.

Noisy Gradient and Threat Model. While injecting random noise to gradients is a standard defense against privacy leakage against gradient inversion attacks Kaissis et al. (2023), many existing methods Abadi et al. (2016); Geyer et al.; Wei et al. (2021a); Kaissis et al. (2023) apply perturbations solely to the leaked gradient $\boldsymbol{g}$ during transmission. This setting provides the attacker with a significant advantage during iterative gradient inversion attacks, and is applicable to scenarios where the attacker has access to model parameters θ. This is because the random noise injection is restricted to the leaked gradients $\boldsymbol{g}$, the attacker can bypass the defense mechanism and utilise the model parameters θ to calculate the *exact* gradients $\boldsymbol{g}_t$ for their dummy inputs $\boldsymbol{v}_t$ to drive the optimization process.

By contrast, we assume a black-box setting in this work. That is, the attacker has no access to the model parameters θ, and the defense mechanism is directly embedded in the gradient generation. We define the defended gradient after random noise injection as $D(\boldsymbol{x}, \theta)$ for the input $\boldsymbol{x}$. The iterative gradient inversion attackers optimize not only on a random noisy leaked gradient $D(\boldsymbol{x}, \theta)$ but also on random noisy gradients $D(G(\boldsymbol{v}_t), \theta)$ during optimization process. Furthermore, because the random noise added to the gradients is generated using a pseudo random number generator with a fixed seed, the gradients cannot be denoised by simply averaging the results of repeated queries. Our setting establishes an interaction paradigm similar to *Machine Learning as a Service (MLaaS)*: the attacker interacts with the model solely through an interface enforcing noise injection on *every query*.

Our threat model and defended shared model are illustrated in Figure 1 and Figure 4 in Appendix A.3. We consider a federated learning scenario where the attacker compromises the interfaces of at least one client node. These interfaces allow the attacker to feed inputs and observe released gradients. Our objective is to design a randomized mechanism within this interface to robustly protect against iterative gradient inversion attacks.

4 METHODOLOGY

A popular defense mechanism against the threat model above is DP-SGD Abadi et al. (2016), which controls the magnitude of the gradient and injects an isotropic Gaussian noise. Injecting isotropic Gaussian noise as in DP-SGD enjoys differential privacy guarantees, which indicates the worst-case guarantees against gradient inversion attacks Kaissis et al. (2023). However, isotropic Gaussian noise consider the worse-case for arbitrary input data and may not be the optimal choice considering the underlying input data distribution for some specific tasks.

In this section, we propose **A**daptive **D**irection-**A**ware **N**oise (ADAN), a random noise mechanism in which the magnitude of each coordinate is learning-based and adapted to the input data distribution of the specific task considered.

4.1 ADAPTIVE NON-ISOTROPIC PERTURBATION

We propose injecting input-dependent, non-isotropic Gaussian noise with a diagonal covariance matrix into the original gradient. Formally, we define the defended gradient as $D(\boldsymbol{x}, \theta) = \nabla_\theta \mathcal{L}(\boldsymbol{x}, \theta) + \delta(\boldsymbol{x})$ where the noise perturbation $\delta(\boldsymbol{x}) \sim \mathcal{N}\left(\mathbf{0}, \text{diag}\left(\sigma^2(\boldsymbol{x})\right)\right)$ where $\sigma(\boldsymbol{x}) = [\sigma_1(\boldsymbol{x}), \sigma_2(\boldsymbol{x}), ..., \sigma_d(\boldsymbol{x})] \in \mathbb{R}^d$ represents the vector of standard deviations corresponding to d model parameters.

To model the noise distribution, we employ a neural network h_ϕ, parameterized by ϕ, which maps the input $\boldsymbol{x}$ to the standard deviation vector, i.e., $\sigma(\boldsymbol{x}) = h_\phi(\boldsymbol{x})$. In this regard, our objective is to optimize ϕ to robustly defend against gradient inversion attacks, subject to constraints on the noise magnitude $\|h_\phi(\boldsymbol{x})\|_2$. Therefore, our method will optimize the direction of the non-isotropic noise, making it adaptive direction-aware.

4.2 Optimization on Coordinate-Wise Variance

The motivation to optimize the direction is to ensure that the expected update direction in iterative gradient inversion attacks is negatively correlated to the direction towards the ground-truth input. The illustration of noise direction is illustrated in Figure 5 in Appendix A.4. Formally, iterative gradient inversion attacks update the dummy vector $\boldsymbol{v}_t$ by gradient descent based on the noisy gradients. In the t-th iteration, the update Δ_t of the reconstructed inputs by iterative gradient inversion attack based on (1) is:

$$\begin{aligned}\Delta_t &= G(\boldsymbol{v}_{t+1}) - G(\boldsymbol{v}_t) = G(\boldsymbol{v}_t - \eta \nabla_{\boldsymbol{v}_t} \mathcal{L}_{\text{dist}}) - G(\boldsymbol{v}_t) \\ &\simeq -\eta \left[\nabla^T_{\boldsymbol{v}_t} G(\boldsymbol{v}_t) \cdot \nabla_{\boldsymbol{v}_t} G(\boldsymbol{v}_t)\right] \nabla_{G(\boldsymbol{v}_t)} \mathcal{L}_{\text{dist}}\end{aligned} \tag{2}$$

where $\mathcal{L}_{\text{dist}} = \|D(\boldsymbol{x}, \theta) - D(G(\boldsymbol{v}_t), \theta)\|_2^2$ is the squared distance between the noisy leaked gradient and the noisy gradients obtained by $G(\boldsymbol{v}_t)$. To defend against gradient inversion attacks, we encourage the update Δ_t in (2) can drive the reconstructed input $G(\boldsymbol{v}_t)$ away from the ground-truth input $\boldsymbol{x}$. Therefore, we define the loss objective as in the following equation. We encourage Δ_t to have small expected signed projections in the direction of $\boldsymbol{x} - G(\boldsymbol{v}_t)$:

$$\mathcal{L}_{\text{proj}}(\phi) = \mathbb{E}_\delta \left[\Delta_t^T \widehat{\boldsymbol{d}}(\boldsymbol{v}_t)\right], \ \widehat{\boldsymbol{d}}(\boldsymbol{v}_t) = \frac{\boldsymbol{x} - G(\boldsymbol{v}_t)}{\|\boldsymbol{x} - G(\boldsymbol{v}_t)\|_2 + \epsilon} \tag{3}$$

where $\widehat{\boldsymbol{d}}(\boldsymbol{v}_t)$ is the normalized direction from the reconstructed input $G(\boldsymbol{v}_t)$ to the ground-truth input $\boldsymbol{x}$ and ϵ is a predefined small constant to avoid zero division. We focus on analyzing Δ_t as defined in (2), since it is the only term in (3) that depends on δ and thus ϕ. First, we use a symmetric semi-definite $H(\boldsymbol{v}_t) = \nabla^T_{\boldsymbol{v}_t} G(\boldsymbol{v}_t) \cdot \nabla_{\boldsymbol{v}_t} G(\boldsymbol{v}_t)$ to simplify the notation. $H(\boldsymbol{v}_t)$ can also be consider the kernel matrix when calculating the inner product between Δ_t and $\boldsymbol{x} - G(\boldsymbol{v}_t)$. Then, we simplify $\nabla_{G(\boldsymbol{v}_t)} \mathcal{L}_{\text{dist}}$ as follows:

$$\begin{aligned}\mathbb{E}_\delta[\nabla_{G(\boldsymbol{v}_t)} \mathcal{L}_{\text{dist}}] &= \nabla_{G(\boldsymbol{v}_t)} \mathbb{E}_\delta \|D(\boldsymbol{x}, \theta) - D(G(\boldsymbol{v}_t), \theta)\|_2^2 \\ &= \nabla_{G(\boldsymbol{v}_t)} \mathbb{E}_\delta \|\boldsymbol{g} + \delta(\boldsymbol{x}) - \boldsymbol{g}_t - \delta(G(\boldsymbol{v}_t))\|_2^2 \\ &= \nabla_{G(\boldsymbol{v}_t)} \|\boldsymbol{g} - \boldsymbol{g}_t\|_2^2 + \nabla_{G(\boldsymbol{v}_t)} \|h_\phi(G(\boldsymbol{v}_t))\|_2^2\end{aligned} \tag{4}$$

The last equation above is based on two facts: (1) the noise $\delta(\boldsymbol{x})$, $\delta(G(\boldsymbol{v}_t))$ have zero mean and diagonal variance as defined in Section 4.1; (2) $\boldsymbol{g}$, $\delta(\boldsymbol{x})$ do not depend on $G(\boldsymbol{v}_t)$ and thus have zero gradients. Moreover, we can see that the first term on the right hand side of (4) does not depend on ϕ, only the second term is a function on ϕ.

Plugging (2), (4) and the definition of $\widehat{\boldsymbol{d}}(\boldsymbol{v}_t)$, $H(\boldsymbol{v}_t)$ into (3), we have the following equation:

$$\mathcal{L}_{\text{proj}}(\phi) = -\eta \nabla^T_{G(\boldsymbol{v}_t)} \|h_\phi(G(\boldsymbol{v}_t))\|_2^2 H(\boldsymbol{v}_t) \widehat{\boldsymbol{d}}(\boldsymbol{v}_t) + C \tag{5}$$

where C is a constant term independent of ϕ. In addition, the step size η is a predefined constant. Therefore, we can discard η and C when optimizing ϕ.

Now, we are ready to optimize the noise direction which is reflected by the parameter ϕ. Specifically, we need to constrain the magnitude of the injected random noise, i.e., $\|h_\phi(G(\boldsymbol{v}_t))\|_2^2$, not only to control the impact of noisy gradient in model training but also to enable fair comparison with other defensive methods. Therefore, we formulate the constrained optimization problem on ϕ as follows:

$$\min_\phi -\nabla^T_{G(\boldsymbol{v}_t)} \|h_\phi(G(\boldsymbol{v}_t))\|_2^2 H(\boldsymbol{v}_t) \widehat{\boldsymbol{d}}(\boldsymbol{v}_t), \quad s.t. \ \|h_\phi(G(\boldsymbol{v}_t))\|_2^2 \leq \alpha^2 d \tag{6}$$

where d is the number of parameters and α indicates the average magnitude per coordinate. We consider the augmented Lagrangian function $\mathcal{L}_{\text{aug}}$ of the optimization problem (6):

$$\begin{aligned}\mathcal{L}_{\text{aug}}(\phi, \lambda) = &-\nabla^T_{G(\boldsymbol{v}_t)} \|h_\phi(G(\boldsymbol{v}_t))\|_2^2 H(\boldsymbol{v}_t) \widehat{\boldsymbol{d}}(\boldsymbol{v}_t) + \\ &\lambda \left(\|h_\phi(G(\boldsymbol{v}_t))\|_2^2 - \alpha^2 d\right) + \frac{\beta}{2} \left(\|h_\phi(G(\boldsymbol{v}_t))\|_2^2 - \alpha^2 d\right)^2\end{aligned} \tag{7}$$

where λ is the dual variable and β is the hyperparameter for the quadratic term. We adopt Alternating Direction Method of Multipliers (ADMM) to alternatively update the primal variable ϕ and the dual variable λ. For primal variable ϕ, we apply gradient-based method to update it since $\mathcal{L}_{\text{aug}}$ is fully differentiable. For dual variable λ, we update it by:

$$\lambda \leftarrow \lambda + \beta \left(\|h_\phi(G(\boldsymbol{v}_t))\|_2^2 - \alpha^2 d \right) \tag{8}$$

We use different data collected in the next subsection as $\boldsymbol{v}_t$ to optimize ϕ until convergence. We provide the pseudo-code of our method in Algorithm 1.

Algorithm 1 Offline Data Collection and ADAN Training

Require: Auxiliary dataset $\mathcal{D}_{\text{aux}}$; defense model h_ϕ; attack method $\mathcal{A}$; number of references M; samples per trajectory N; noise scale α; penalty β; learning rate η_{def}.

Initialize dataset $\mathcal{S} \leftarrow \emptyset$
for $m = 1$ to M **do**
  Sample $\boldsymbol{x}^{(m)} \sim \mathcal{D}_{\text{aux}}$, run attack $\mathcal{A}$ to get $\{\boldsymbol{v}_0, \ldots, \boldsymbol{v}_T\}$
  Uniformly sample $\{\boldsymbol{v}_{t_j}\}_{j=1}^N$, then $\mathcal{S} \leftarrow \mathcal{S} \cup \{(\boldsymbol{v}_{t_j}, \boldsymbol{x}^{(m)})\}_{j=1}^N$
end for
Initialize ϕ: final layer weights $\mathbf{0}$, bias softplus$^{-1}(\alpha)$; dual variable $\lambda \leftarrow 0$
repeat
  Sample $(\boldsymbol{v}_t, \boldsymbol{x}) \sim \mathcal{S}$
  Compute reconstruction $G(\boldsymbol{v}_t)$, direction $\widehat{\boldsymbol{d}}(\boldsymbol{v}_t) \leftarrow \frac{\boldsymbol{x} - G(\boldsymbol{v}_t)}{\|\boldsymbol{x} - G(\boldsymbol{v}_t)\|_2 + \epsilon}$
  Compute $\mathcal{L}_{\text{aug}}(\phi, \lambda)$, update $\phi \leftarrow \phi - \eta_{\text{def}} \nabla_\phi \mathcal{L}_{\text{aug}}$
  Update $\lambda \leftarrow \lambda + \beta\big(\mathbb{E}[\|h_\phi(G(\boldsymbol{v}_t))\|_2^2] - \alpha^2 d\big)$
until convergence
return Trained defense model h_ϕ

4.3 Model Design and Data Collection

Compared with injecting isotropic random noise, our proposed adaptive non-isotropic perturbation requires a model to predict the noise variance for each coordinate. We apply the softplus activation function (Zheng et al., 2015) to the expanded output to ensure that the model's predictions are strictly positive. In addition, to prevent gradient explosion when calculating the gradient in (7), we initialize the weights of the final layer to zero and set the bias term to a specific value $b = \text{softplus}^{-1}(\alpha)$, on which applying softplus function obtains α at the start of training. Consequently, the optimization is initially dominated by the directional objective (the first term in Equation 7), allowing the model to learn the optimal tangential direction for defense while maintaining training stability.

More details about ADAN model architecture and the method to reduce memory consumption are deferred to Appendix A.3.

The optimization of ϕ relies on a set of intermediate dummy inputs $\{\boldsymbol{v}_t\}$ representative of real attack trajectories. To construct the training dataset, we assume the defender has access to a public or auxiliary dataset that shares a similar distribution with the clients' private data, a standard assumption in defensive literature. We sample M reference inputs $\{\boldsymbol{x}^{(m)}\}_{m=1}^M$ from this auxiliary data domain. For each reference input $\boldsymbol{x}^{(m)}$, we simulate the iterative gradient inversion attack process offline. During each simulation, we collect N intermediate dummy latent codes $\{\boldsymbol{v}_{t_j}^{(m)}\}_{j=1}^N$ from the optimization trajectory. Consequently, the training set consists of $M \times N$ pairs of intermediate reconstructions and their corresponding reference targets: $\mathcal{D}_{\text{train}} = \bigcup_{m=1}^M \{(G(\boldsymbol{v}_{t_j}^{(m)}), \boldsymbol{x}^{(m)})\}_{j=1}^N$. This data collection strategy is conducted offline; thus, it does not incur computational overhead during the online federated learning phase. Moreover, by generalizing across M reference samples, the objective encourages the variance generation model to capture the general geometric relationship between the intermediate dummy inputs and the data manifold, rather than overfitting to a single sample. This ensures that the generated noise consistently drives the optimization of $\boldsymbol{v}_t$ away from the correct reconstruction direction across the data domain. ADAN *does not* require retraining for each attack, model, or dataset. Its generalization ability is deferred to Appendix B.3.

The implementation of ADAN defense is deferred to Appendix A.2.

5 EXPERIMENTAL RESULTS

5.1 EXPERIMENTAL SETUP

Datasets and Models. We evaluate our proposed defense on two standard image classification benchmarks: CIFAR-10 (Krizhevsky et al., 2009) and ImageNet (Deng et al., 2009). For CIFAR-10, we employ LeNet (LeCun et al., 1998) as the federated learning (FL) model. For the large-scale ImageNet dataset, we utilize ResNet-50 (He et al., 2016). More results of different dataset-model pairs are deferred to Appendix B.4.

Attacks and Metrics. To verify the robustness of our defense, we select two representative iterative gradient inversion attacks: Deep Leakage from Gradients (DLG) (Zhu et al., 2019) and GGL (Li et al., 2022b). DLG represents *optimization-based methods*, which directly optimize dummy inputs to match gradients. We employ DLG to attack the CIFAR-10 dataset. GGL represents generative model-based methods, which optimize the latent code of a pre-trained generator. We utilize GGL with BigGAN (Brock et al., 2018) to attack the ImageNet dataset, as optimization-based methods struggle to reconstruct high-resolution images. To accommodate black-box threat model, we use estimate gradients of the optimization objective $\mathcal{L}_{\text{GM}}$ in Equation (1) by zero-order methods.

To quantitatively and comprehensively evaluate the quality of reconstructed training data (and thus the privacy leakage), we employ four metrics: Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). While MSE and PSNR measures pixel-level differences, SSIM and LPIPS provide better assessments of perceptual quality and structural similarity, which are critical for determining if private semantic information has been leaked.

The configurations of FL training and GIAs are deferred to Appendix B.1.

5.2 PERFORMANCE COMPARISON WITH BASELINES

5.2.1 COMPARISON WITH ISOTROPIC NOISE

We first compare our proposed adaptive non-isotropic perturbation against the standard isotropic Gaussian noise defense. To ensure a fair comparison, we enforce the total energy of our adaptive noise to equal that of the isotropic noise. Specifically, we scale the output of our noise generation model such that the expected ℓ_2-norm of the perturbation matches the isotropic baseline (i.e., $\sum \sigma_i^2 = \sigma_{\text{iso}}^2 \cdot d$).

Table 1: Quantitative comparison of reconstruction quality under different defense mechanisms. We evaluate the defense performance against DLG attack on CIFAR-10 and GGL attack on ImageNet with varying noise levels (σ).

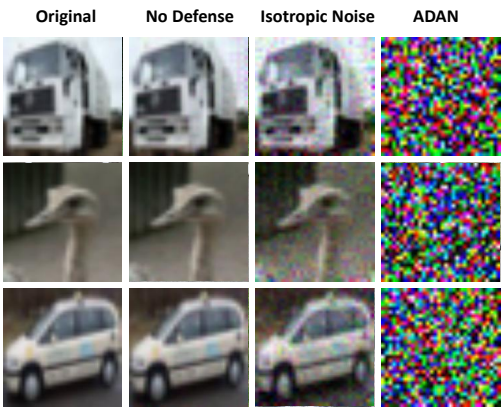
σ	Defense	MSE ↑	PSNR ↓	LPIPS ↑	SSIM ↓
Attack: DLG (on CIFAR-10)					
0	None	0.0126	18.9963	0.0290	0.8797
0.02	Isotropic	0.0242	16.1618	0.0373	0.8127
	ADAN	0.1588	7.9915	0.3114	0.0990
0.04	Isotropic	0.0339	14.6980	0.0572	0.6005
	ADAN	**0.1869**	**7.2839**	0.3498	0.0375
0.06	Isotropic	0.0401	13.9686	0.0715	0.5235
	ADAN	0.1844	7.3424	**0.3579**	**0.0311**
0.08	Isotropic	0.0458	13.3913	0.0866	0.4674
	ADAN	0.1789	7.4739	0.3439	0.0345

σ	Defense	MSE ↑	PSNR ↓	LPIPS ↑	SSIM ↓
Attack: GGL (on ImageNet)					
0	None	0.0698	11.2090	0.6123	0.1850
0.05	Isotropic	0.0757	12.5643	0.6170	0.2033
	ADAN	0.0724	11.4026	0.6450	0.1918
0.10	Isotropic	0.0809	11.4836	0.6491	0.1814
	ADAN	0.0813	11.0128	0.6715	0.1703
0.15	Isotropic	0.0857	10.5129	0.6796	0.1598
	ADAN	0.0849	10.9843	0.7132	0.1397
0.20	Isotropic	**0.0899**	9.3237	0.6925	0.1340
	ADAN	0.0870	**9.3065**	**0.7538**	**0.1105**

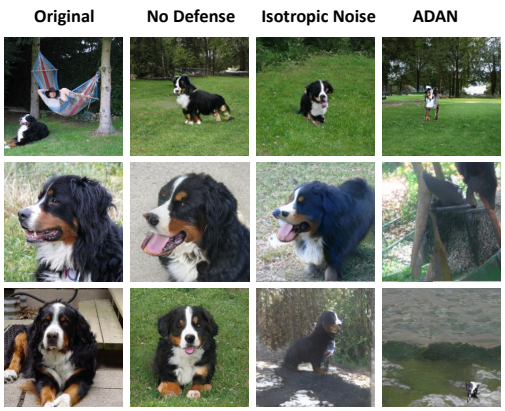
Table 1 summarizes the reconstruction metrics under different defense strategies. Our method significantly outperforms the isotropic baseline across all metrics. On CIFAR-10 (against DLG), the defense is remarkably effective: the reconstructed images are completely corrupted even at a low noise level of $\sigma = 0.02$. Since DLG relies solely on the ℓ_2 norm for gradient matching, our method—which explicitly optimizes the noise direction to maximize the reconstruction error based on the geometry of the loss landscape—successfully renders the attack ineffective. As shown in

Figure 2(a), while isotropic noise allows the recovery of rough shapes, our proposed ADAN results in completely unrecognizable noise patterns.

On ImageNet (against GGL), Defending against GGL is inherently more challenging than DLG for two reasons: (1) GGL employs gradient clipping and matches sparse gradients, which naturally filters out certain noise components; (2) GGL adopts a more complex gradient matching objective, which integrates the ℓ_1 norm, ℓ_2 norm, and cosine similarity, and further applies layer-wise weighting to gradients when computing the overall matching loss. Despite these challenges, our adaptive defense demonstrates superior performance compared to isotropic noise and no-defense baselines. As shown in Figure 2(b), while isotropic noise may alter the background, angle, or texture, the semantic category of the object often remains identifiable. In contrast, our adaptive perturbation achieves semantic destruction, preventing the adversary from inferring the *object class*, let alone recovering fine-grained details.



(a) Comparison of isotropic noise and ADAN under equal noise scale ($\sigma = 0.02$) against DLG attacks. From left to right: ground truth, reconstruction without defense, reconstruction with isotropic noise, and reconstruction with ADAN. The ground-truth image corresponds to CIFAR-10 dataset.



(b) Comparison of isotropic noise and ADAN under equal noise scale ($\sigma = 0.05$) against GGL attacks. From left to right: ground truth, reconstruction without defense, reconstruction with isotropic noise, and reconstruction with ADAN. The ground-truth image corresponds to the Bernese Mountain Dog class from the ImageNet dataset.

Figure 2: Comparison of defense methods on CIFAR-10 and ImageNet

We observe a notable phenomenon: although ADAN consistently provides superior protection in terms of semantic privacy, the MSE metric occasionally yields lower values (indicating "worse" defense) compared to isotropic noise. We attribute this discrepancy to the fundamental difference in attack mechanisms. DLG performs reconstruction at the pixel level, where defense manifests as visible high-frequency noise; thus, MSE correlates well with defense strength. Conversely, GGL reconstructs images via a latent space mapping, leveraging a strong generative prior to ensure high perceptual fidelity. Consequently, defense against GGL manifests as semantic drifts—altering attributes such as identity, age, orientation, or position—rather than pixel-level degradation. Since GGL leverages a strong generative prior to ensure high perceptual fidelity, standard metrics such as MSE is insufficient for comprehensively evaluating defense efficacy against GGL. Therefore, visual inspection remains essential to assess the privacy preservation capabilities against generative model-based attacks.

5.2.2 Comparison with other defense methods (Utility-Privacy Trade-off)

Comparing noise-based defenses with other defense mechanisms, such as Gradient Clipping Wei et al. (2021b) and Gradient Compression Aji & Heafield (2017), is non-trivial because it is difficult to control confounding factors and to fairly compare privacy protection at a fixed level of utility. Gradient clipping limits the ℓ_2 norm of shared gradients by thresholding their magnitude, while gradient compression reduces information leakage by sparsifying gradients, typically by pruning

a fixed proportion of coordinates with the smallest magnitudes. To provide a holistic evaluation, we analyze the trade-off between privacy protection (measured by reconstruction MSE) and model utility (measured by the accuracy of the FL model). We sweep the hyperparameters for each defense (noise standard deviation σ, clipping threshold, and compression pruning rate) to generate a set of operating points. Figure 3 illustrates the Privacy-Utility trade-off curves obtained on the CIFAR-10 dataset. To quantitatively evaluate the overall performance, we calculate the Area Under the Curve (AUC) for each defense mechanism. To ensure a fair comparison, the AUC is computed specifically over the common high-utility interval (the intersection of valid accuracy ranges across all methods), focusing on the regime where the federated model remains functional with minimal utility loss. A larger AUC signifies a superior trade-off, indicating that the method maintains higher reconstruction error (privacy) for a given cost in accuracy.

The results demonstrate that our adaptive defense (ADAN) achieves the highest AUC, consistently providing stronger privacy guarantees than gradient clipping and pruning at comparable utility levels. We observe distinct behaviors among the defense strategies: First, the gradient clipping curve suggests that clipping offers a favorable trade-off initially; a substantial boost in privacy (MSE) can be achieved with only a negligible drop in utility. However, it struggles to reach the high-privacy ceiling maintained by our method without drastically sacrificing accuracy. In addition, while ADAN operates within a similar utility envelope to isotropic noise, it sustains a significantly higher baseline of privacy protection. As shown in the figure 3, ADAN maintains a high MSE even at high accuracy levels. This corroborates our earlier findings in Section 5.2.1 that even minute noise variance (e.g., $\sigma = 0.01$) in our adaptive approach is sufficient to render reconstructed images completely unrecognizable (high MSE), whereas isotropic noise requires significantly larger variance to achieve a comparable defensive effect.

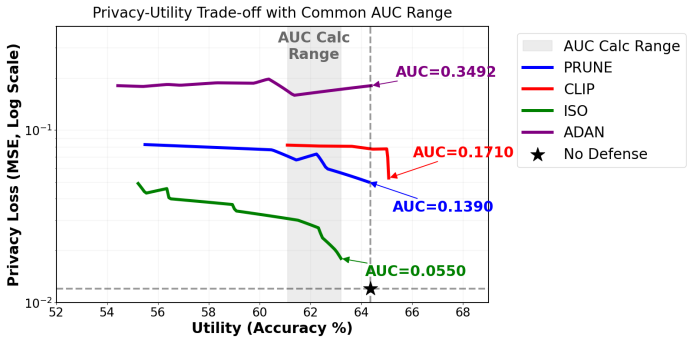


Figure 3: Illustration of Utility-Privacy curve of different defense methods.

5.3 More Analyses and Ablation Studies

Ablation studies, generalization evaluation and performance under defense-aware attacks are deferred to Appendices B.2, B.3, and B.5, respectively.

6 Conclusion

In this work, we revisit defending against iterative training data reconstruction attacks under a more realistic black-box threat model, where the server only observes compromised client communication channels rather than assuming a curious-but-honest white-box server. This reveals new opportunities to act directly on the attacker's optimization process. We propose **Adaptive Direction-Aware Noise (ADAN)**, a learning-based, non-isotropic perturbation mechanism that injects input-dependent noise into every queried gradient. By shaping the optimization geometry, ADAN effectively disrupts iterative gradient inversion attacks across different datasets and attack paradigms. Experiments show that ADAN provides substantially stronger privacy than isotropic noise and other common defenses, while maintaining favorable utility–privacy trade-offs. Ablations further show that ADAN is data-efficient and scalable, requiring only limited trajectory coverage and lightweight architectures to generalize across large-scale FL models. This highlights the importance of aligning defense design with realistic threat models and opens new directions for optimization-aware privacy defenses in federated learning.

7 AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript to improve readability and to assist with coding, including code debugging and implementation support. We have not used generative AI tools for tasks with required disclosure, and the required disclosure tasks are not applicable to this work. All AI-assisted text and code were reviewed, verified, and tested by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021*, 2017.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Huancheng Chen and Haris Vikalo. Recovering labels from local updates in federated learning. *arXiv preprint arXiv:2405.00955*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Hao Fang, Bin Chen, Xuan Wang, Zhi Wang, and Shu-Tao Xia. Gifd: A generative gradient inversion method with feature domain optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4967–4976, 2023.

Liam Fowl, Jonas Geiping, Wojtek Czaja, Micah Goldblum, and Tom Goldstein. Robbing the fed: Directly obtaining private data in federated learning with modified models. *arXiv preprint arXiv:2110.13057*, 2021.

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.

RC Geyer, T Klein, and M Nabi. Differentially private federated learning: A client level perspective. arxiv 2017. *arXiv preprint arXiv:1712.07557*.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7232–7241. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/3b3fff6463464959dcd1b68d0320f781-Paper.pdf.

Jinwoo Jeon, Kangwook Lee, Sewoong Oh, Jungseul Ok, et al. Gradient inversion with generative image prior. *Advances in neural information processing systems*, 34:29898–29908, 2021.

Arthur Jochems, Timo M Deist, Johan Van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and Andre Dekker. Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital–a real life proof of concept. *Radiotherapy and Oncology*, 121(3):459–467, 2016.

Georgios Kaissis, Jamie Hayes, Alexander Ziller, and Daniel Rueckert. Bounding data reconstruction attacks with the hypothesis testing interpretation of differential privacy. *arXiv preprint arXiv:2307.03928*, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Weiyu Li, Zhaoxian Wu, Tianyi Chen, Liping Li, and Qing Ling. Communication-censored distributed stochastic gradient descent. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11): 6831–6843, 2021.

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 (CVPR)*, pp. 10132–10142, June 2022a.

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*, pp. 10132–10142, 2022b.

Jiahao Lu, Xi Sheryl Zhang, Tianli Zhao, Xiangyu He, and Jian Cheng. April: Finding the achilles' heel on privacy for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10051–10060, 2022.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Le Trieu Phong, Yoshinori Aono, Takuya Hayashi, Lihua Wang, and Shiho Moriai. Privacy-preserving deep learning: Revisited and enhanced. In *Applications and Techniques in Information Security: 8th International Conference, ATIS 2017, Auckland, New Zealand, July 6–7, 2017, Proceedings*, pp. 100–110. Springer, 2017.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pp. 601–618, 2016.

Yijue Wang, Jieren Deng, Dan Guo, Chenghong Wang, Xianrui Meng, Hang Liu, Caiwen Ding, and Sanguthevar Rajasekaran. Sapag: A self-adaptive privacy attack from gradients. *arXiv preprint arXiv:2009.06228*, 2020.

Zihan Wang, Jason Lee, and Qi Lei. Reconstructing training data from model gradient, provably. In *International Conference on Artificial Intelligence and Statistics*, pp. 6595–6612. PMLR, 2023.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Hang Su, Bo Zhang, and H Vincent Poor. User-level privacy-preserving federated learning: Analysis and performance optimization. *IEEE Transactions on Mobile Computing*, 21(9):3388–3401, 2021a.

Wenqi Wei, Ling Liu, Margaret Loper, Ka-Ho Chow, Mehmet Emre Gursoy, Stacey Truex, and Yanzhao Wu. A framework for evaluating gradient leakage attacks in federated learning. *arXiv preprint arXiv:2004.10397*, 2020.

Wenqi Wei, Ling Liu, Yanzhao Wu, Gong Su, and Arun Iyengar. Gradient-leakage resilient federated learning. In *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*, pp. 797–807. IEEE, 2021b.

Yuxin Wen, Jonas Geiping, Liam Fowl, Micah Goldblum, and Tom Goldstein. Fishing for user data in large-batch federated learning via gradient magnification. *arXiv preprint arXiv:2202.00580*, 2022.

Liwen Wu, Zhizhi Liu, Bin Pu, Kang Wei, Hangcheng Cao, and Shaowen Yao. Dggi: Deep generative gradient inversion with diffusion model. *Information Fusion*, 113:102620, 2025.

Ruihan Wu, Xiangyu Chen, Chuan Guo, and Kilian Q Weinberger. Learning to invert: Simple adaptive attacks for gradient inversion in federated learning. In *Uncertainty in Artificial Intelligence*, pp. 2293–2303. PMLR, 2023.

Wensi Yang, Yuhang Zhang, Kejiang Ye, Li Li, and Cheng-Zhong 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*, pp. 18–32. Springer, 2019.

Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16337–16346, 2021.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

Joshua C Zhao, Atul Sharma, Ahmed Roushdy Elkordy, Yahya H Ezzeldin, Salman Avestimehr, and Saurabh Bagchi. Loki: Large-scale data reconstruction attack against federated learning through model manipulation. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 1287–1305. IEEE, 2024.

Hao Zheng, Zhanlei Yang, Wenju Liu, Jizhong Liang, and Yanpeng Li. Improving deep neural networks using softplus units. In *2015 International joint conference on neural networks (IJCNN)*, pp. 1–4. IEEE, 2015.

Junyi Zhu and Matthew Blaschko. R-gap: Recursive gradient attack on privacy. *arXiv preprint arXiv:2010.07733*, 2020.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/60a6c4002cc7b29142def8871531281a-Paper.pdf.

A Method Supplementary

A.1 Fairness and Strength of Black-Box Evaluation

The shift to a black-box setting does not "weaken" the attack logic; instead, it allows for a more realistic evaluation where protection is active throughout the entire attack process.

Attack Algorithm Consistency. Existing Gradient Inversion Attacks (GIAs) assume white-box setting because they require access to the model's computation graph to compute the "gradient of the gradient" regarding the input ($\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}, \theta)$). We adapted these into a black-box setting by employing the finite difference method (first-order estimation). This ensures the attack algorithm remains unchanged.

Why Black Box. In White-Box settings, defenses are restricted to protect the ground truth gradient. Attackers can bypass this by using "clean" local parameters during their attack iterations. In our Black-Box setting, the defense is embedded within the interface. Noise is injected into every single

query during the attacker's optimization trajectory. This is a more realistic and challenging scenario for the attacker.

Methodological Fairness. To ensure a rigorous comparison, all baseline defenses were integrated into the black-box interface in the exact same manner as ADAN. Every query made by the attacker was subjected to the respective defense, ensuring that ADAN's superior performance stems from its "direction-aware" geometry rather than a more favorable setup.

A.2 ADAN IMPLEMENTATION

During the implementation stage, the perturbation is implemented via the reparameterization trick, so that the defended gradient is computed as $D(\boldsymbol{x}, \theta) = \nabla_\theta \mathcal{L}(\boldsymbol{x}, \theta) + h_\phi(\boldsymbol{x}) \odot \epsilon,\ \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. As a result, during iterative reconstruction attacks, the adversary can backpropagate through the perturbed gradients with respect to the dummy input, which is consistent with the attacker's optimization perspective formalized in Equation 4. In addition, we explicitly normalize the generated noise to ensure that its overall magnitude *exactly* matches the target scale, since the soft constraint in the objective does not strictly enforce equality. The online ADAN implementation pseudo-code is provided in Algorithm 2. Instead of the whole FL system, Algorithm 2 demonstrates one client which needs to share gradients with other nodes via communication. That is to say, Algorithm 2 focuses on the internal gradient processing, which is the key of the defense mechanism in this paper. By contrast, our framework is compatible with generic inter-node FL protocols, including client sampling and server aggregation, which can adopt, for example, the popular FedAvg framework.

Algorithm 2 Local Gradient Computation with ADAN Defense

Require: Private dataset $\mathcal{D}_{\text{train}}$; FL model parameters θ; trained defense model h_ϕ; learning rate η_{FL}.

for each FL round **do**
  Sample batch $(\boldsymbol{x}, y) \sim \mathcal{D}_{\text{train}}$
  Compute gradient $\boldsymbol{g} \leftarrow \nabla_\theta \mathcal{L}_{\text{FL}}(\boldsymbol{x}, y; \theta)$
  Predict noise scale $\boldsymbol{\sigma} \leftarrow h_\phi(\boldsymbol{x})$
  Sample Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Generate perturbation $\delta(\boldsymbol{x}) \leftarrow \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$
  Defend gradient $\tilde{\boldsymbol{g}} \leftarrow \boldsymbol{g} + \delta(\boldsymbol{x})$
  Update model $\theta \leftarrow \theta - \eta_{\text{FL}} \tilde{\boldsymbol{g}}$
end for
return Trained FL model θ

A.3 ILLUSTRATION OF OUR DEFENDED SHARED MODEL

To minimize memory overhead, we downsample the reconstructed dummy data $G(\boldsymbol{v}_t)$ to a fixed resolution of $3 \times 32 \times 32$ before feeding it into the MLP. Furthermore, since target models like ResNet-50 contain over 25 million parameters, predicting a unique variance for each coordinate is computationally prohibitive. To address this, we restrict the MLP output to a lower dimensionality and map these values to the full gradient dimension via a fixed random permutation, implicitly assuming that subsets of gradient coordinates share the same predicted variance.

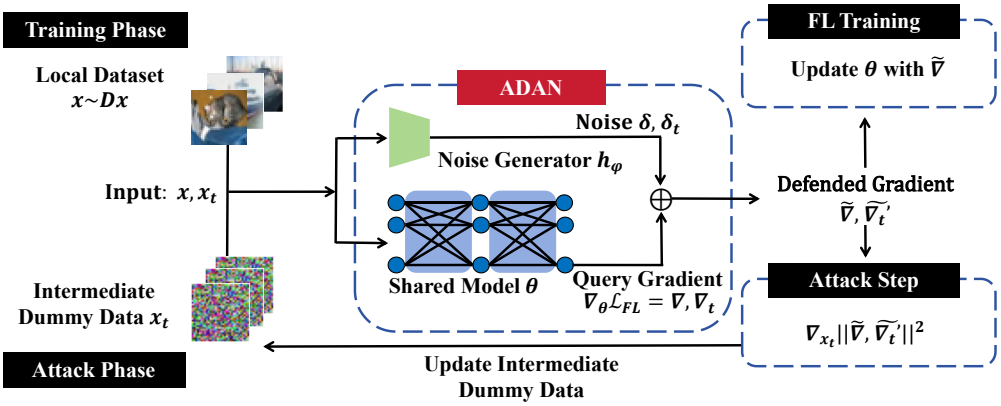


Figure 4: Illustration of details of our defended shared model (inside the black box shared model).

A.4 Illustration of Defense Mechanism

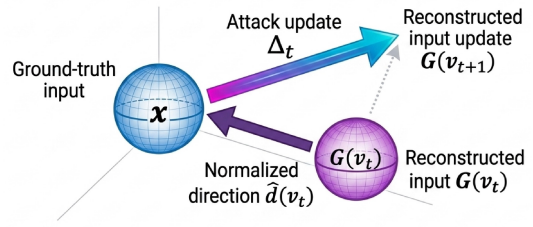


Figure 5: Illustration of Noise Direction Learning

B Experimental Supplementary

B.1 Implementation Details of FL Training and GIAs

Unless otherwise specified, for training the ADAN model, we randomly select $M = 10$ reference data batches from the auxiliary dataset to minimize training costs, with a batch size of 1. During the offline data collection phase, We simulate the attack trajectory and collect intermediate dummy inputs at uniform intervals. For DLG, the total attack iterations are set to 20000; for GGL, they are set to 1000. For each reference image, we simulate the attack trajectory and collect $N = 1000$ intermediate samples. To evaluate defense performance, we execute these same attack protocols against the defended model and report the metrics averaged over 100 batches.

Our experiments utilize a simple FedAvg (Huang et al., 2021) protocol involving 10 clients by default. When calculating the model utility under defense, the training process consists of 800 rounds. In each round, we randomly select 1 client to participate and use standard coordinate-wise averaging for aggregation. For defense performance, privacy metrics (e.g., MSE) are independent of data partitioning or node count, as GIAs analyze the leakage of a specific leaked gradient, regardless of its origin. Therefore, we focus on a single local step at Epoch 1, the stage where model weights are untrained and most vulnerable. The above experimental setup follows the configuration in CENSOR (Li et al., 2021).

B.2 Ablation Studies

B.2.1 Trajectory-Based Optimization

The optimization of our defense parameter ϕ relies on a set of intermediate dummy inputs $\{\boldsymbol{v}_t\}$, which are representative of real attack trajectories. As detailed in Section 4.3, we construct the training dataset $\mathcal{D}_{\text{train}}$ by sampling M reference images and simulating the attack process. Crucially, to capture the temporal evolution of the gradient inversion, we collect N intermediates sampled at equal intervals spanning the entire optimization path. We investigate the impact of dataset construction parameters on the CIFAR-10 dataset. As shown in Table 2, trajectory depth (i.e., the number of trajectory iterations N) proves to be the decisive factor for robust defense. This is evidenced by the observation that under equal $M \times N$ settings, a larger N yields stronger defensive performance. This finding aligns with the principle that a sufficiently large N ensures robustness across various stages of the attack, thereby enhancing overall defense effectiveness. In addition, increasing the trajectory coverage N from $1,000$ to $5,000$ (while keeping $M = 10$ fixed) leads to a significant improvement in reconstruction error. Finally, performance saturates at $N = 5,000$, indicating that this level of coverage is sufficient to capture the essential characteristics of the gradient landscape.

Table 2: Ablation study on Trajectory Coverage. We report average metrics across varying noise levels ($\sigma \in [0.01, 0.1]$). Comparing configurations reveals that trajectory depth (N) is more critical than breadth (M).

Configuration	**Coverage**	**Avg Performance**		
	($M \times N$)	MSE ↑	LPIPS ↑	SSIM ↓
Isotropic Baseline	-	0.035	0.061	0.603
Ours (Varying M & N)				
M=10, N=1k	1.0×10^4	0.186	0.349	0.037
M=50, N=1k	5.0×10^4	0.199	0.351	0.031
M=10, N=5k	5.0×10^4	**0.206**	**0.357**	**0.027**
M=10, N=10k	10.0×10^4	**0.206**	0.353	0.035

B.2.2 Model Design and Efficiency

Finally, we analyze the design choices of the lightweight MLP used to parameterize h_ϕ. To balance efficiency and performance, we impose a constraint on the output dimensionality. For ResNet-50, predicting a distinct variance for all approximately 25 million parameters is computationally prohibitive. Instead, we cap the MLP output at a lower dimensionality (150,000) and map these values to the full gradient dimension via a fixed random permutation. We compare reconstruction performance as we vary the MLP output dimensionality from 15,000 to 750,000; larger dimensions quickly lead to excessive memory consumption. Evaluated on ImageNet with ResNet-50 under the GGL attack (one example visual result shown in the Figure 6), we observe that while $15,000$ yields some degradation, $150,000$ is already sufficient to provide an effective defense. This design enables our method to scale to large architectures with minimal overhead.



Figure 6: Comparison of defense performance across different output dimensions of the ADAN model.

B.3 Generalization under Distribution Shifts

ADAN *does not* require retraining for each attack, model, or dataset. It learns a repulsive force field around the ground-truth manifold: for *any* dummy data, the generated noise pushes its update direction opposite to the vector pointing toward the ground truth, as shown in Figure 5. This objective is attack-agnostic—trajectory-based simulation is only an empirical choice for faster convergence. To handle black-box shifts, we unify input/output dimensions via down-sampling (input) and dimension expansion (output), as detailed in Section 4.3 (Model Architecture). Thus, ADAN can be applied to different models or datasets without modification. We evaluate three shifts on CIFAR-10 under the DLG attack ($\sigma = 0.04$). The results are shown in Table 3.

Table 3: Generalization of ADAN under distribution shifts (CIFAR-10, DLG, $\sigma = 0.04$).

Shift Type	Training Condition	Testing Condition	MSE ↑	PSNR ↓	LPIPS ↑	SSIM ↓
No Defense	–	DLG on CIFAR-10	0.0126	18.9963	0.0290	0.8797
No Shift	DLG, LeNet, CIFAR-10	DLG, LeNet, CIFAR-10	**0.1869**	**7.2839**	**0.3498**	**0.0375**
Model Shift	DLG, LeNet, CIFAR-10	DLG, ResNet18, CIFAR-10	0.1723	7.6412	0.3256	0.0521
Attack Shift	DLG, LeNet, CIFAR-10	IG, LeNet, CIFAR-10	0.1405	8.5217	0.2689	0.0983
Dataset Shift	DLG, LeNet, CIFAR-10	DLG, ResNet18, Tiny-ImageNet	0.0987	9.9541	0.2012	0.2456

Table 3 demonstrates that ADAN remains highly effective under all shifts (far better than no defense). Dataset shift causes the largest degradation, yet defense does not collapse, because different natural image manifolds are still much closer to each other than to randomly initialized dummy data. Attack shift also weakens performance due to different gradient matching objectives. Model shift has the smallest impact, as ADAN is independent of the FL model architecture (see Appendix A.3, Figure 4). Without any shift, ADAN naturally achieves the best results (overfitting to a specific attack yields highest defense). Overall, retraining is only recommended under severe dataset mismatch; otherwise, the pre-trained defense can be deployed directly.

B.4 Complementary Experiments on More Dataset-Model Pairs

Table 4 shows complementary results for larger models and different datasets. As seen in the table, ADAN's defense performance actually improves with model size, while isotropic defense's performance decreases. This is because, although the expansion of the dimensionality of "information" to exploit, it simultaneously provides ADAN with a higher degree of freedom to manipulate the noise distribution.

B.5 Improved ADAN against Defense-Aware Attacks

To counter adaptive attackers attempting to mitigate noise via repeated-query averaging, a straightforward **Improved ADAN** can address this by leveraging a stateful, *deterministic* and input-dependent noise injection strategy within the black-box interface: By mapping the specific state of a dummy input $\boldsymbol{x}$ to a deterministic (yet non-isotropic) noise, the interface ensures that repeated queries for the same dummy data return the same noisy gradient, preventing "averaging out" the noise, as the error remains rather than vanishing. Furthermore, since iterative GIAs already require hundreds of iterations, any attempt at multi-query averaging would significantly increase the attacker's computational cost and query volume, making the attack easily detectable by standard rate-limiting and anomaly detection mechanisms in a real-world MLaaS or TEE-based deployment.

Table 4: Comparison of defense performance under different dataset-model pairs.

Dataset	Attack	Model	Defense	σ	MSE ↑	PSNR ↓	LPIPS ↑	SSIM ↓
CIFAR-10	DLG	ResNet-18	None	0	0.0115	19.0862	0.0272	0.8921
			Isotropic	0.02	0.0215	16.6754	0.0410	0.7845
			Isotropic	0.04	0.0298	15.2576	0.0621	0.5832
			ADAN	0.02	0.1685	7.7341	0.3245	0.0821
			ADAN	0.04	**0.1942**	**7.1172**	**0.3681**	**0.0342**
CIFAR-10	DLG	ViT	None	0	0.0112	19.5076	0.0254	0.8943
			Isotropic	0.02	0.0182	17.4001	0.0354	0.8210
			Isotropic	0.04	0.0254	15.9521	0.0512	0.6541
			ADAN	0.02	0.1856	7.3142	0.3521	0.0452
			ADAN	0.04	**0.2231**	**6.5154**	**0.4124**	**0.0215**
ImageNet	GGL	VGG-16	None	0	0.0652	11.8564	0.5982	0.1985
			Isotropic	0.05	0.0712	12.8241	0.6012	0.2145
			Isotropic	0.10	0.0784	11.6521	0.6321	0.1952
			ADAN	0.05	0.0754	11.2261	0.6582	0.1824
			ADAN	0.10	**0.0842**	**10.7451**	**0.6954**	**0.1524**
ImageNet	GGL	ViT	None	0	0.0582	12.3512	0.5621	0.2214
			Isotropic	0.05	0.0685	13.0142	0.5842	0.2312
			Isotropic	0.10	0.0741	11.9542	0.6154	0.2014
			ADAN	0.05	0.0892	10.4952	0.7241	0.1254
			ADAN	0.10	**0.1054**	**9.7681**	**0.7854**	**0.0984**