Venom: Liquid Diffusion-Guided Gradient Inversion for Breaking Differential Privacy in Federated Learning

Bin Hu[1,2], **Jingling Yuan**[2,1*], **Jiawei Jiang**[3], **Chuang Hu**[4,1*]
[1]Hubei Key Laboratory of Transportation Internet of Things, School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China
[2]State Key Laboratory of Silicate Materials for Architectures, Wuhan University of Technology, Wuhan, China
[3]School of Computer Science, Wuhan University, Wuhan, China
[4]State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau SAR
hubin@whut.edu.cn, yjl@whut.edu.cn, jiawei.jiang@whu.edu.cn, handc@whu.edu.cn

Abstract

Gradient perturbation mechanisms, such as differential privacy (DP), aim to defend against gradient inversion attacks (GIA) by injecting noise into the shared gradients. Recent studies have shown that DP-based defenses lack robustness against advanced GIAs. However, existing gradient inversion methods typically rely on iterative refinement and assume static noise, resulting in low efficiency and limited reconstruction fidelity under high-noise conditions. In this paper, we propose Venom, a novel gradient inversion attack method based on a liquid diffusion mechanism. Venom reconstructs private data directly from DP-protected gradients without requiring any prior knowledge of the noise distribution. Specifically, we design a Structural Prior Extraction (SPE) module that analytically extracts deep feature representations from perturbed gradients through energy-based aggregation, enabling stable pre-reconstruction of users' latent data features. We further introduce a Diffusion-driven Liquid Recovery Network (Diff-LRN) for high-fidelity image reconstruction. Unlike traditional diffusion models that rely on iterative sampling with predefined noise schedules, Diff-LRN performs deterministic single-step reconstruction using adaptive liquid neural dynamics to handle spatially heterogeneous noise patterns. Experiments across four benchmarks demonstrate that Venom achieves an speedup of up to 38,315× over state-of-the-art attacks while maintaining high reconstruction fidelity under strong DP settings. These results challenge prevailing assumptions about DP robustness and underscore the need for more resilient privacy-preserving mechanisms in federated learning.

Introduction

Federated learning (FL) is a distributed model training paradigm that enables multiple clients to collaboratively train a shared model while keeping their raw data local (McMahan et al. 2017; Gong et al. 2024; Huang et al. 2024a; Wei and Liu 2025). Under this paradigm, a central server coordinates training by sending the global model to each client; each client then updates the model using its own local data and computes a model update (gradient), which the server aggregates to refine the global model, repeating this process over multiple rounds until convergence.

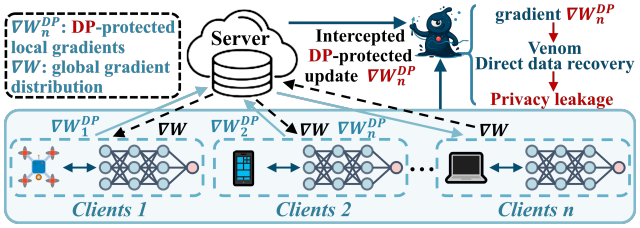

Figure 1: Threat model. In a standard federated learning setup, clients compute local gradients and upload DP-protected updates ∇W_n^{DP} to a central server. We assume a passive adversary who intercepts the transmitted gradients. While existing attacks rely on noise priors or surrogate training pipelines, Venom enables direct data reconstruction from DP-contaminated gradients without requiring knowledge of the DP mechanism or iterative optimization.

However, although no raw data leaves the clients, the gradients shared in FL can still leak sensitive information about the users' data. Indeed, researchers have shown that an adversary can exploit these gradient updates to reconstruct a client's original training examples, known as the **Gradient Inversion Attack (GIA)** (Zhu, Liu, and Han 2019; Liang et al. 2023; Gao et al. 2025). To mitigate such privacy risks, **Differential Privacy (DP)** (Dwork 2006) mechanisms have been widely incorporated into FL systems (Hu et al. 2023; Wang, Hugh, and Li 2024; Zhang et al. 2025). By injecting noise into each client's gradients before upload, DP provides formal privacy guarantees and significantly reduces the information an attacker can infer from the gradients.

Although DP significantly raises the difficulty of gradient inversion attacks, recent work (Liu et al. 2025) demonstrates that it is still possible to circumvent DP by first removing the noise from gradients and then reconstructing the data (as shown in Figure 1). In particular, they propose Mjölnir, which constructs surrogate models to capture the pattern of DP-induced gradient perturbations, then employs a diffusion-based model to denoise the gradients, and finally applies DLG (Deep Leakage from Gradients) techniques to recover the original images. However, Mjölnir suffers from two critical limitations. First, it strongly depends on knowledge of the noise distribution used by DP (e.g., assuming

*Co-corresponding authors.

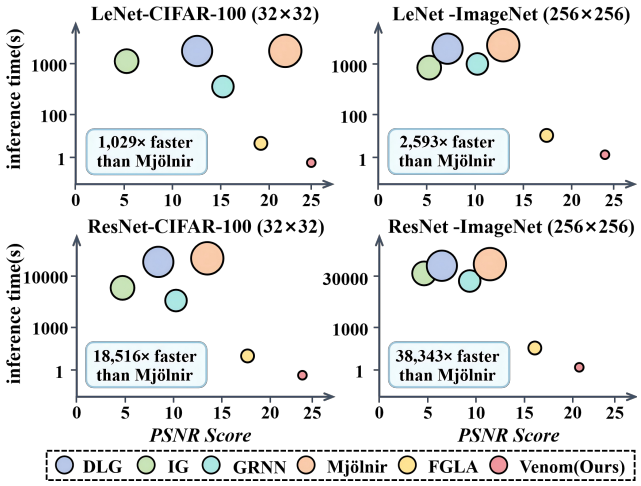

Figure 2: Trade-off analysis between reconstruction quality and runtime of CIFAR-100 (32×32) and ImageNet (256×256) using LeNet and ResNet. Circle size reflects per-image runtime. Venom achieves up to $1{,}010\times$ speedup on LeNet and $38{,}315\times$ on ResNet while maintaining high reconstruction quality.

Gaussian or Laplace noise), which is often unavailable in real federated learning deployments. Second, the approach relies on a computationally intensive iterative optimization process with substantial training overhead, making it prone to convergence difficulties for large models (as shown in Fig. 2). These shortcomings collectively point to two core challenges: (1) *How can we achieve robust and efficient reconstruction without relying on any noise prior?* (2) *How can we design an analytical (non-iterative) attack mechanism to improve reconstruction efficiency and stability?*

Liquid Neural Networks (LNNs) (Hasani et al. 2021) possess a unique ability to adapt their internal states in response to input dynamics without requiring any prior knowledge of the input signal characteristics. Their adaptive state-update mechanism enables efficient processing of time-varying information without resorting to computationally intensive iterative optimization procedures. Inspired by these properties, we propose a novel two-stage gradient inversion framework, dubbed **Venom**, to exploit the synergy of "liquid" neural adaptivity and potent attack capabilities. Venom first employs a **Structural Prior Extraction (SPE)** module that analytically recovers deep semantic features from perturbed gradients. The SPE module leverages an energy-based multi-class feature aggregation strategy to counteract the distortion amplification caused by differential privacy noise and further enhances feature fidelity through wavelet-domain structural refinement. However, due to the complex, heterogeneous patterns of DP noise, the deep features output by SPE may still retain residual perturbations. To address this, we introduce a **Diffusion-driven Liquid Recovery Network (Diff-LRN)** as the second stage, which incorporates adaptive mechanisms to further purify features and reconstruct the input image. Specifically, Diff-LRN first estimates the residual noise strength via structural divergence analysis, then dynamically adjusts its reconstruction strategy by integrating the temporal adaptivity of LNNs with a deterministic diffusion process. Unlike conventional diffusion-based models that require hundreds of iterative denoising steps, Diff-LRN performs a one-pass deterministic reconstruction in feature space, eliminating expensive iterative sampling while still achieving high-fidelity image recovery without any prior knowledge of the noise distribution.

Compared to the recent state-of-the-art Mjölnir attack—which relies on a pre-trained diffusion model and iterative reverse steps to remove gradient noise—Venom introduces a more efficient two-stage pipeline that significantly improves both runtime and reconstruction fidelity. As shown in Figure 2, Venom achieves up to $38{,}315\times$ speedup while preserving high-quality reconstructions. By combining analytical noise mitigation (SPE) with LNN-driven adaptive recovery (Diff-LRN), Venom effectively counteracts privacy noise and enables robust data reconstruction under strong differential privacy constraints, all with substantially reduced computational overhead.

Related Work

Differential Privacy in Federated Learning

Differential Privacy (DP) has emerged as a principled defense mechanism in federated learning, offering formal guarantees against individual data leakage by injecting calibrated noise into model updates (Wei et al. 2023; Ren et al. 2024; Fu et al. 2024). A randomized mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP if, for any adjacent datasets D and D' differing in at most one sample, and for all measurable subsets S of outputs, it holds that:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D') \in S] + \delta \qquad (1)$$

Here, ϵ quantifies the maximum privacy loss, and δ represents a small probability of failure.

In practice, federated learning with DP typically follows four steps: local gradient clipping, sensitivity computation, calibrated noise injection (e.g., Laplace or Gaussian mechanisms), and the upload of privatized model updates. Following the assumptions in Mjölnir (Liu et al. 2025), we consider both the Laplace mechanism (achieving pure ϵ-DP) and the Gaussian mechanism (providing (ϵ, δ)-DP). The noise scale is governed by the gradient sensitivity $\nabla_s = 2C/m$, where C is the clipping threshold and m denotes the minimum client-side data size.

Gradient Reconstruction Attacks with DP Noise

Classical Gradient Inversion. Zhu et al. (Zhu, Liu, and Han 2019) introduced Deep Leakage from Gradients (DLG), reconstructing training data by optimizing dummy inputs to match observed gradients. Follow-up work improved both accuracy and convergence: iDLG (Zhao, Mopuri, and Bilen 2020) analytically extracted ground-truth labels, and Geiping et al. (Geiping et al. 2020) employed cosine similarity to stabilize optimization. These approaches have since been extended to large-batch settings (Wen et al. 2022), vision transformers (Hatamizadeh et al. 2023), and even large language model training (Feng et al. 2024).

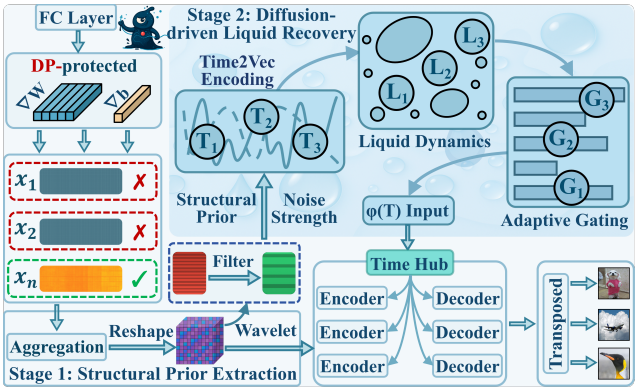

Figure 3: Overview of the proposed Venom framework for gradient inversion under differential privacy. It comprises two stages: (1) Structural Prior Extraction, which reconstructs deep representations from DP-protected gradients via energy-aware class selection and wavelet refinement; and (2) Diffusion-driven Liquid Recovery Network, which denoises noisy features into high-fidelity images using a deterministic, noise-adaptive architecture. Diff-LRN integrates Time2Vec encoding, liquid neural dynamics, and adaptive gating. Venom enables high-precision reconstruction without requiring knowledge of the DP noise distribution or costly iterative processes.

Attacks under DP Protection. Classical gradient inversion methods often fail under differential privacy, as injected noise disrupts gradient structure. However, Carlini et al. (Carlini et al. 2021) showed that membership inference remains feasible even under strong DP guarantees, and Dimitrov et al. (Dimitrov et al. 2022) demonstrated that partial data recovery is possible when the noise distribution is known. Building on these insights, Mjölnir (Liu et al. 2025) introduces a diffusion-based framework for attacking DP-protected gradients, but it relies on noise distribution priors and incurs computational overhead, leading to limited scalability in practice (Yang et al. 2024; Huang et al. 2024b; Ye et al. 2024; Yang et al. 2025). In parallel, Scale-MIA (Shi et al. 2023) investigates model inversion in secure aggregation settings, where the adversary observes only aggregated updates. In contrast, Venom focuses on gradient inversion from individual client updates before aggregation.

Liquid Neural Networks

Liquid Neural Networks (LNNs) (Hasani et al. 2021) adapt to non-stationary inputs through time-varying dynamics, enabling effective modeling of evolving or unknown noise patterns. They have shown utility in domains such as robotics (Hasani et al. 2022), time-series forecasting (Pawlak et al. 2024), and neuromorphic computing (Ahmad et al. 2025), particularly where static architectures struggle under dynamic conditions. However, their potential in privacy-preserving gradient inversion remains unexplored. Motivated by this gap, we integrate LNN dynamics with efficient reconstruction strategies to enhance robustness and computational efficiency under DP.

Methodology

Venom Architecture Overview

We propose Venom (as shown in Figure 3), a novel two-stage gradient inversion framework that circumvents DP defenses through analytical gradient modeling and adaptive noise-aware recovery. Venom reconstructs high-fidelity images from DP-protected gradients by first extracting informative features via Structural Prior Extraction (SPE). This module efficiently performs analytical recovery of deep features from final-layer gradients. To further denoise and refine these features, we introduce a Diffusion-driven Liquid Recovery Network (Diff-LRN), a one-step deterministic diffusion model enhanced with liquid neural dynamics that adapts its denoising behavior to unknown and heterogeneous DP noise. Unlike existing methods that rely on hundreds of diffusion steps or iterative optimization, Venom achieves efficient inference through single-step deterministic reconstruction while maintaining high reconstruction fidelity.

Structural Prior Extraction

Recent work by Dimitrov et al. (Dimitrov et al. 2022) demonstrates that gradient structures can retain informative patterns even under DP perturbations. Building on this insight, we propose an SPE module that analytically recovers noise-injected input representations from final-layer gradients, thereby avoiding the high computational complexity $\mathcal{O}(TKd)$ and convergence instability typically associated with iterative optimization under DP constraints.

Consider a fully connected (FC) classification layer with parameters $\mathbf{W} \in \mathbb{R}^{K\times d}$ and $\mathbf{b} \in \mathbb{R}^K$, where K is the number of classes and d is the feature dimension. The gradient structure induced by the cross-entropy loss satisfies:

$$\nabla \mathbf{W}_j = \delta_j \cdot \mathbf{x}, \quad \nabla b_j = \delta_j, \tag{2}$$

where $\delta_j = \partial \mathcal{L}/\partial z_j$ denotes the loss gradient for logit z_j of class j. Under this structure, the input representation $\mathbf{x}$ can be analytically recovered as $\mathbf{x} = \nabla \mathbf{W}_j / \nabla b_j$ with $\mathcal{O}(Kd)$ complexity. However, this direct division is highly sensitive to DP noise, particularly when $\tilde{\nabla} b_j \approx 0$, leading to numerical instability and amplified error. To address this, we introduce an energy-based multi-class aggregation mechanism that exploits statistical redundancy among class-wise gradients. We define an energy score for each class j as:

$$E_j = \|\tilde{\nabla} \mathbf{W}_j\|_2 \cdot |\tilde{\nabla} b_j|, \tag{3}$$

favoring classes in which both weight and bias gradients exhibit strong signals. This multiplicative design exploits the statistical independence of DP noise across coordinates, making E_j an effective detector for signal-dominated classes where true gradients consistently dominate noise perturbations. We select statistically reliable classes as:

$$\mathcal{S} = \{j : E_j > \mu_E + \gamma \sigma_E\}, \tag{4}$$

where μ_E and σ_E denote the mean and standard deviation of $\{E_j\}_{j=1}^K$, and $\gamma = \min(2.0, \sqrt{\log K})$ balances robustness and diversity. The final estimate is obtained via energy-weighted fusion:

$$\hat{\mathbf{x}} = \sum_{j\in\mathcal{S}} w_j \cdot \frac{\tilde{\nabla} \mathbf{W}_j}{\tilde{\nabla} b_j + \lambda \sigma^2}, \tag{5}$$

where $w_j = E_j / \sum_{k \in \mathcal{S}} E_k$ normalizes the energy weights, and $\lambda = \sigma^2 / \max_j |\delta_j|^2$ ensures numerical stability. The SPE module achieves $O(Kd + d \log d)$ time complexity, providing substantial speedup over iterative methods requiring $O(TKd)$ with $T \approx 1000$ steps.

To further suppress high-frequency DP noise, the aggregated feature $\hat{\mathbf{x}}$ is refined using discrete wavelet transforms, which isolate structured signal components from stochastic noise. The result serves as a structural prior that guides image reconstruction. While SPE efficiently extracts dominant gradient-aligned structures, residual noise persists due to the limitations of purely analytical recovery under strong DP constraints. To overcome this, we introduce Diff-LRN, which refines $\hat{\mathbf{x}}$ into perceptually faithful reconstructions.

Diff-LRN: Adaptive Diffusion Recovery Network

Building upon the structural features extracted by SPE, we propose the Diff-LRN to reconstruct original images from noisy deep features. Unlike Mjölnir, which assumes known noise and requires hundreds of iterative steps, Diff-LRN performs deterministic single-step reconstruction while adapting DP noise patterns.

(A) Standard Diffusion Pitfalls. Conventional diffusion models assume a known Gaussian noise schedule and uniform noise injection across all dimensions. The reverse process is given by:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}, \quad (6)$$

where $\boldsymbol{\epsilon}_\theta$ predicts the noise at step t, and $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. However, under DP, noise is injected in a data-dependent and spatially heterogeneous manner. This breaks core diffusion assumptions, causing fixed schedules to over-smooth clean regions and under-denoise high-noise areas. Furthermore, standard diffusion requires hundreds of steps, making it impractical for real-time gradient inversion.

(B) Deterministic Noise-Adaptive Recovery. To address the computational inefficiency of iterative diffusion sampling, we propose a deterministic one-pass reconstruction process that adapts to the corruption level of DP-contaminated features. Instead of relying on pre-defined noise schedules, we estimate the noise strength T for a given SPE-extracted feature $\hat{\mathbf{x}}$ via structure-aware divergence:

$$T = \gamma \log \left(1 + \frac{\|\hat{\mathbf{x}} - \mathbf{c}\|_2^2}{\|\mathbf{c}\|_2^2} \right), \mathbf{c} = \text{WaveletRefine}(\hat{\mathbf{x}}) \quad (7)$$

where $\mathbf{c}$ serves as a smooth structural baseline and $\gamma = 2.0$ provides logarithmic scaling that ensures numerical stability while maintaining appropriate sensitivity to local corruption levels. This relative divergence measure adapts to feature-dependent noise characteristics without requiring explicit noise distribution knowledge. Critically, Venom inference requires only the observed gradients and does not rely on DP parameters or noise priors. We then train a reconstruction network $\mathcal{R}_\theta(\hat{\mathbf{x}}, T)$ to directly map noisy features to clean images. To generate training data without explicit noise modeling, we apply SPE to gradients corrupted by clipped DP noise (Algorithm 1), yielding triplets $\{(\hat{\mathbf{x}}_i, T_i, \mathbf{z}_i)\}$. The network is trained to minimize:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \|\mathcal{R}_\theta(\hat{\mathbf{x}}_i, T_i) - \mathbf{z}_i\|_2^2 \quad (8)$$

This deterministic approach achieves computational speedups of 2-4 orders of magnitude over iterative methods while maintaining reconstruction quality. However, static network parameters cannot adapt to the spatially heterogeneous noise patterns in DP perturbations, motivating our liquid neural dynamics design for adaptive denoising.

Algorithm 1: Dataset Construction for Diff-LRN Training

Require: Auxiliary images $\mathcal{D}_{\text{aux}} = \{\mathbf{z}_i\}_{i=1}^N$, target model f_θ, DP parameters (ϵ, δ), clipping bound C
Ensure: Training dataset $\{(\hat{\mathbf{x}}_i, T_i, \mathbf{z}_i)\}_{i=1}^N$
for each image $\mathbf{z}_i \in \mathcal{D}_{\text{aux}}$ **do**
  Forward pass: $\mathbf{y}_i = f_\theta(\mathbf{z}_i)$; Label: $\hat{\mathbf{y}}_i$
  Compute and clip gradients: $\nabla \mathbf{W} \leftarrow \text{clip}(\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{\text{fc}}}, C)$, $\nabla \mathbf{b} \leftarrow \text{clip}(\frac{\partial \mathcal{L}}{\partial \mathbf{b}_{\text{fc}}}, C)$
  Sample DP noise: $\sigma^2 = \frac{2C^2 \ln(1.25/\delta)}{\epsilon^2}$
  Add DP noise: $\tilde{\nabla}\mathbf{W} = \nabla \mathbf{W} + \mathcal{N}(0, \sigma^2 \mathbf{I})$, $\tilde{\nabla}\mathbf{b} = \nabla \mathbf{b} + \mathcal{N}(0, \sigma^2)$
  Extract features: $\hat{\mathbf{x}}_i = \text{SPE}(\tilde{\nabla}\mathbf{W}, \tilde{\nabla}\mathbf{b})$
  Estimate noise: $T_i = \gamma \log \left(1 + \frac{\|\hat{\mathbf{x}}_i - \mathbf{c}_i\|_2^2}{\|\mathbf{c}_i\|_2^2}\right)$ where $\mathbf{c}_i = \text{WaveletRefine}(\hat{\mathbf{x}}_i)$
  Handle degenerate cases: if $|\mathcal{S}| = 0$, select $j^* = \arg\max_j E_j$, set $\hat{\mathbf{x}}_i$ from j^*, $T_i = T_{\max}$
  Store triplet: $(\hat{\mathbf{x}}_i, T_i, \mathbf{z}_i)$
end for
return Training dataset $\{(\hat{\mathbf{x}}_i, T_i, \mathbf{z}_i)\}_{i=1}^N$

(C) Adaptive Liquid Neural Integration. While the deterministic diffusion approach offers computational efficiency and noise-aware conditioning, it applies uniform processing to all features, limiting its capacity to complex DP noise. To overcome this, we introduce a Liquid Time Embedding (LTE) module that adaptively modulates the denoising process based on the estimated level of corruption.

Time2Vec Encoding. The noise strength parameter T is first transformed into rich temporal representations using Time2Vec encoding (Kazemi et al. 2019):

$$\mathbf{v}_{\text{linear}} = \mathbf{W}_l T + \mathbf{b}_l, \quad \mathbf{v}_{\text{periodic}} = \sin(\mathbf{W}_p T + \mathbf{b}_p) \quad (9)$$

$$\mathbf{t}_{\text{enc}} = \text{concat}(\mathbf{v}_{\text{linear}}, \mathbf{v}_{\text{periodic}}) \quad (10)$$

This encoding captures both magnitude and phase information, providing richer representations than standard sinusoidal embeddings.

Liquid Neural Dynamics. The encoded temporal features undergo adaptive evolution through a liquid neural unit inspired by liquid time-constant (LTC) neurons:

$$\mathbf{h}_{\text{new}} = \mathbf{h}_{\text{old}} + \alpha \cdot (\tanh(\mathbf{W}_{\text{in}} \mathbf{t}_{\text{enc}} + \mathbf{W}_{\text{rec}} \mathbf{h}_{\text{old}}) - \mathbf{h}_{\text{old}}) \quad (11)$$

This update rule enables dynamic temporal integration, where the representation adapts its evolution rate based on

Algorithm 2: Diff-LRN Inference with Liquid Integration

Require: DP gradients $\tilde{\nabla}\mathbf{W}, \tilde{\nabla}\mathbf{b}$; recovery network $\mathcal{R}_\theta$
Ensure: Reconstructed image $\mathbf{z}$

feature extraction: $\hat{\mathbf{x}} \leftarrow \text{SPE}(\tilde{\nabla}\mathbf{W}, \tilde{\nabla}\mathbf{b})$
noise estimation: $\mathbf{c} \leftarrow \text{WaveletRefine}(\hat{\mathbf{x}})$;
$T \leftarrow \gamma \cdot \log\left(1 + \frac{\|\hat{\mathbf{x}}-\mathbf{c}\|_2^2}{\|\mathbf{c}\|_2^2}\right)$
temporal embedding: $\phi(T) \leftarrow \text{LTE}(T)$
time-conditioned UNet:
for each ResNet block in UNet **do**
  $\mathbf{x} \leftarrow \mathbf{x} \cdot (\text{scale}(\phi(T)) + 1) + \text{shift}(\phi(T))$
end for
image reconstruction: $\mathbf{z} \leftarrow \mathcal{R}_\theta(\hat{\mathbf{x}}, \phi(T))$
return $\mathbf{z}$

noise characteristics. The learnable parameter $\alpha \in (0, 1)$ (initialized to 0.1 and learned jointly with network parameters) controls the adaptation rate. Heavily corrupted regions benefit from slower integration for deeper denoising, while cleaner areas stabilize quickly to preserve fine details.

Adaptive Gating. The liquid dynamics are refined via feature-dependent gating:

$$\mathbf{g} = \sigma(\mathbf{W}_g \mathbf{h}_{\text{proj}} + \mathbf{b}_g), \quad \phi(T) = \mathbf{g} \odot \mathbf{h}^{(t+1)} \tag{12}$$

where $\mathbf{h}_{\text{proj}}$ is a linear projection of $\mathbf{h}^{(t+1)}$, and $\odot$ denotes element-wise multiplication. The resulting temporal embedding $\phi(T)$ replaces conventional sinusoidal embeddings in the deterministic reconstruction network $\mathcal{R}_\theta(\hat{\mathbf{x}}, T)$.

The liquid-enhanced temporal embedding $\phi(T)$ is integrated into the UNet architecture through adaptive normalization layers. Specifically, $\phi(T)$ generates scale and shift parameters for each ResNet block: $\text{scale}(\phi(T))$ and $\text{shift}(\phi(T))$ via linear projections. This allows the network to dynamically adjust its processing intensity based on estimated noise level T and local feature characteristics encoded in $\phi(T)$. Unlike fixed diffusion schedules, this adaptive conditioning enables fine-grained control over denoising strength across different feature regions, making single-step reconstruction both efficient and noise-aware. The complete inference procedure is summarized in Algorithm 2.

Theoretical Analysis

We provide theoretical guarantees for the proposed framework, focusing on the reconstruction error bounds of SPE and the convergence properties of Diff-LRN.

Theorem 1 (SPE Reconstruction Error Bound)**.** *Under (ϵ, δ)-DP with noise variance σ^2, assuming bounded gradients and at least one reliable class in $\mathcal{S}$ with $|\delta_j| \geq \delta_{\min} > 0$, the SPE reconstruction error satisfies:*

$$\mathbb{E}[\|\hat{\mathbf{x}} - \mathbf{x}^*\|_2^2] \leq \frac{C_1 \sigma^2}{|\mathcal{S}|} + C_2 \lambda^2 \sigma^4 + O(K^{-1}) \tag{13}$$

where C_1, C_2 are constants depending on the gradient magnitude distribution and feature dimension d, $\mathcal{S}$ is the selected class set, and K is the number of classes.

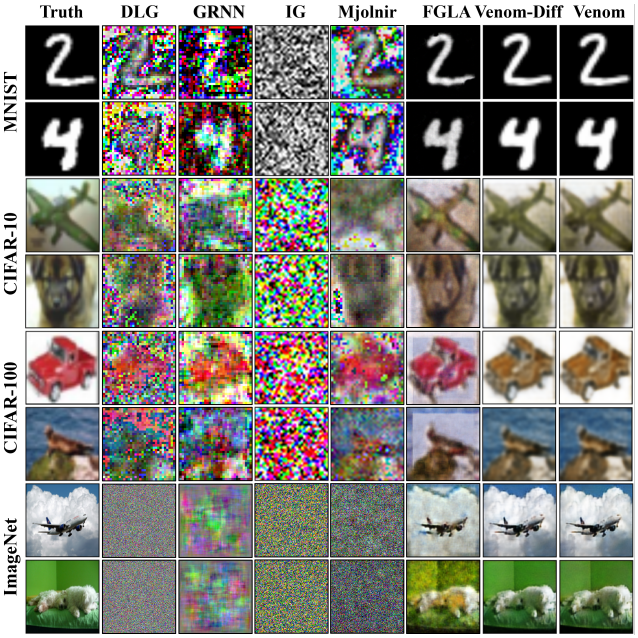

Figure 4: Comparison of ground truth client images and reconstructed results by Venom, traditional, and state-of-the-art gradient inversion attacks ($\delta = 10^{-5}$, $\varepsilon = 10$).

The error bound demonstrates that reconstruction error decreases inversely with the number of selected classes $|\mathcal{S}|$, validating our multi-class aggregation strategy by leveraging statistical redundancy across gradient computations.

Theorem 2 (Diff-LRN Convergence)**.** *Under Lipschitz continuity assumptions on the activation functions and bounded weight matrices, the liquid time embedding $\phi(T)$ converges to the optimal noise-adaptive representation with exponential rate $O((1-\alpha)^t)$, where $\alpha \in (0, 1)$ controls the adaptation speed and t denotes the processing depth.*

This result establishes that the liquid neural dynamics form a contraction mapping, ensuring stable convergence regardless of input noise characteristics. The convergence rate is controlled by α, allowing fine-tuned adaptation between fast convergence (α large) and detailed processing (α small).

Evaluation

Experimental Setup

Datasets. We evaluate on four widely used benchmarks: MNIST (LeCun et al. 1998), CIFAR-10 and CIFAR-100 (Krizhevsky and Hinton 2009), and ImageNet (Deng et al. 2009). These datasets span from simple handwritten digits to high-resolution natural scenes, providing comprehensive evaluation across semantic richness and diversity.

Implementation Details. We simulate different privacy levels using differential privacy (DP) with gradient perturbation under the (ϵ, δ)-DP framework, where $\delta = 10^{-5}$ and $\epsilon \in \{1, 5, 10\}$ correspond to strong, moderate, and weak privacy guarantees, respectively. We use DP-SGD with per-sample gradient clipping ($C = 1.0$). Following standard gradient inversion protocols (Zhu, Liu, and Han 2019;

Dataset	Model	ϵ=1				ϵ=5				ϵ=10			
		PSNR	*SSIM*	*LPIPS*	*ASR*	*PSNR*	*SSIM*	*LPIPS*	*ASR*	*PSNR*	*SSIM*	*LPIPS*	*ASR*
MNIST	GRNN	6.38	0.026	0.782	0%	14.27	0.321	0.747	0%	16.12	0.353	0.663	0%
	IG	4.27	0.018	0.992	0%	4.69	0.028	0.994	0%	4.80	0.029	0.989	0%
	DLG	5.48	0.013	0.965	0%	9.67	0.273	0.739	0%	15.08	0.391	0.719	0%
	FGLA	15.84	0.472	0.523	0%	23.81	0.853	0.161	92%	26.55	0.836	0.092	90%
	Mjölnir	17.87	0.568	0.328	8%	24.17	0.877	0.127	95%	**33.09**	**0.976**	**0.016**	**99%**
	Venom-Diff	**21.04**	**0.832**	**0.105**	**89%**	**27.76**	**0.933**	**0.058**	**96%**	32.48	0.973	0.019	98%
	Venom	**21.31**	**0.835**	**0.099**	**90%**	**28.75**	**0.940**	**0.055**	**97%**	**33.38**	**0.979**	**0.015**	**99%**
CIFAR-10	GRNN	11.58	0.408	0.629	0%	14.98	0.431	0.562	0%	16.67	0.416	0.480	0%
	IG	7.44	0.056	0.897	0%	7.45	0.051	0.895	0%	7.51	0.045	0.900	0%
	DLG	5.67	0.014	0.969	0%	9.60	0.077	0.910	0%	15.19	0.394	0.529	0%
	FGLA	15.57	0.479	0.651	0%	18.36	0.602	0.475	12%	20.11	0.678	0.412	25%
	Mjölnir	18.42	0.664	0.379	30%	19.24	0.691	0.299	35%	21.78	0.752	0.332	62%
	Venom-Diff	**21.63**	**0.804**	**0.267**	**70%**	**23.74**	**0.867**	**0.184**	**84%**	**24.94**	**0.870**	**0.179**	**85%**
	Venom	**21.77**	**0.805**	**0.259**	**72%**	**25.03**	**0.878**	**0.179**	**86%**	**26.06**	**0.895**	**0.155**	**88%**
CIFAR-100	GRNN	10.97	0.398	0.633	0%	13.47	0.325	0.700	0%	15.91	0.481	0.553	0%
	IG	7.10	0.050	0.902	0%	7.17	0.052	0.896	0%	7.23	0.055	0.899	0%
	DLG	5.44	0.018	0.975	0%	9.08	0.074	0.926	0%	14.83	0.405	0.612	0%
	FGLA	15.18	0.581	0.548	3%	16.48	0.563	0.445	8%	18.98	0.694	0.367	45%
	Mjölnir	18.25	0.647	0.401	20%	19.04	0.684	0.318	40%	21.32	0.762	0.241	55%
	Venom-Diff	**20.54**	**0.777**	**0.284**	**62%**	**22.76**	**0.825**	**0.221**	**75%**	**23.64**	**0.878**	**0.170**	**84%**
	Venom	**20.86**	**0.783**	**0.281**	**65%**	**22.89**	**0.839**	**0.198**	**78%**	**24.58**	**0.891**	**0.157**	**88%**
ImageNet	GRNN	8.08	0.061	0.822	0%	10.44	0.178	0.800	0%	10.59	0.199	0.792	0%
	IG	6.01	0.005	0.940	0%	6.09	0.006	0.906	0%	6.70	0.008	0.862	0%
	DLG	7.09	0.009	0.887	0%	7.63	0.009	0.852	0%	7.95	0.010	0.827	0%
	FGLA	13.86	0.379	0.673	0%	15.44	0.443	0.559	3%	17.08	0.483	0.486	5%
	Mjölnir	9.19	0.128	0.791	0%	10.99	0.211	0.740	0%	12.59	0.252	0.632	2%
	Venom-Diff	**19.19**	**0.499**	**0.635**	**8%**	**22.30**	**0.656**	**0.450**	**25%**	**23.61**	**0.700**	**0.368**	**40%**
	Venom	**19.78**	**0.539**	**0.605**	**10%**	**22.67**	**0.659**	**0.442**	**28%**	**24.32**	**0.726**	**0.340**	**45%**

Table 1: Performance comparison on image reconstruction across datasets under varying ϵ ($\delta = 1 \times 10^{-5}$). Metrics include PSNR (↑), SSIM (↑), LPIPS (↓), and Attack Success Rate (ASR@0.5).

Geiping et al. 2020; Liu et al. 2025), we focus on single-gradient privacy leakage scenarios. All experiments are repeated 10 times with different random seeds, and we report mean ± standard deviation. Experiments use PyTorch 2.4 on an NVIDIA RTX 4090 GPU (24GB) running Ubuntu 22.04, under identical DP settings.

Evaluation Metrics. We evaluate reconstruction quality using PSNR, SSIM, and LPIPS. Attack Success Rate (ASR@0.5) is defined as the percentage of reconstructions with SSIM above 0.5. Following standard gradient inversion protocols, reconstruction time is measured as the end-to-end processing time for a single image. Results are averaged over 50 images per dataset, and statistical significance is assessed using paired t-tests with Bonferroni correction.

Experimental Results and Analysis

We compare Venom against five state-of-the-art gradient inversion attacks: DLG (Zhu, Liu, and Han 2019), IG (Geiping et al. 2020), GRNN (Ren, Deng, and Xie 2022), FGLA (Yang et al. 2024), and Mjölnir (Liu et al. 2025), along with two ablated variants of Venom: (1) Venom-Diff (Diff), combining the SPE module with deterministic reconstruction, and (2) Venom, our complete framework.

Reconstruction Quality. Our quantitative evaluation demonstrates Venom's superiority across all datasets and privacy regimes (as shown in Table 1). Venom achieves improvements over the baseline Mjölnir, with PSNR gains ranging from 3.4 dB (MNIST, $\epsilon = 1$) to 11.7 dB (ImageNet, $\epsilon = 10$). Under strong privacy constraints ($\epsilon = 1$), traditional iterative methods collapse with PSNR below 6 dB, while Venom maintains high-quality reconstruction. The performance gap widens with dataset complexity: on ImageNet, Venom preserves semantic structure (LPIPS = 0.340) while Mjölnir produces distorted outputs (LPIPS = 0.632). Unlike Mjölnir's fixed diffusion schedule that applies uniform processing, our liquid time embedding enables fine-grained adaptation to feature-dependent DP noise patterns, making single-step reconstruction efficient and noise-aware.

Attack Success Rate. Venom consistently achieves high attack success rates across all benchmarks, illustrating its

Method	LeNet		ResNet	
	CIFAR-100	ImageNet	CIFAR-100	ImageNet
DLG	808±12	2,160±35	14,977±180	33,717±420
IG	781±9	884±15	7,030±95	30,633±380
GRNN	251±6	1,357±22	1,880±25	26,512±310
Mjölnir	823±11	2,178±28	15,183±195	33,742±445
FGLA	1.10±0.03	1.54±0.04	1.16±0.02	1.66±0.05
Diff	0.90±0.02	1.25±0.03	0.93±0.02	1.32±0.04
Venom	**0.80±0.02**	**0.84±0.02**	**0.82±0.02**	**0.88±0.03**

Table 2: Runtime comparison across different gradient inversion methods. Values: mean ± std in seconds across 10 runs (Device: NVIDIA GeForce RTX 4090 GPU).

practical threat to privacy. Under weak DP guarantees ($\varepsilon = 10$), it reaches 99% ASR@0.5 on MNIST, and 88% on both CIFAR-10 and CIFAR-100—indicating reliable recovery of nearly 9 out of 10 images. On ImageNet, Venom attains 45% ASR@0.5, a 43-point increase over Mjölnir. This robustness persists under strong privacy regimes. At $\varepsilon = 1$, Venom retains 90% ASR@0.5 on MNIST and 72% on CIFAR-10, while all baselines drop below 10%. The full Venom framework with Diff-LRN provides an additional 2-5% ASR improvement over Venom-Diff through adaptive noise handling, demonstrating the practical value of liquid neural dynamics for robust privacy attacks.

Computational Efficiency. Runtime analysis highlights Venom's efficiency gains (Table 2). Iterative approaches like DLG require over 9 hours to reconstruct a single ImageNet sample using ResNet (33,717s), with Mjölnir providing no meaningful speedup. In contrast, Venom achieves up to four orders of magnitude acceleration through a two-stage analytical approach. First, SPE extracts structural features from gradients with $\mathcal{O}(Kd + d \log d)$ complexity. Second, Diff-LRN performs deterministic single-step reconstruction from these noisy deep features to original images with $\mathcal{O}(d)$ complexity, eliminating the need for iterative denoising. On LeNet, this delivers 1,029× and 2,593× speedups on CIFAR-100 and ImageNet, respectively. On ResNet, Venom reduces inference time from 9.4 hours to 0.88 seconds. These improvements render gradient inversion tractable for real-time vulnerability assessment across federated networks, compared to $\mathcal{O}(TKd)$ in iterative methods with $T \approx 1000$ diffusion steps.

Visual Quality Assessment. Qualitative results validate Venom's high-fidelity reconstruction under differential privacy, as shown in Figure 4. While DLG and IG produce heavily distorted outputs, and Mjölnir often exhibits blur and structural artifacts, Venom preserves structural coherence and semantic content across all datasets. MNIST reconstructions display crisp digit contours, CIFAR examples retain clear object boundaries and plausible colors, and ImageNet scenes maintain identifiable shapes and textures. On ImageNet, Venom attains a lower LPIPS score (0.340 vs. 0.632 for Mjölnir) and a higher ASR (45% vs. 2%), indicating that our single-step approach provides efficient inference together with improved perceptual reconstruction quality.

Discussion

Key Findings and Limitations. Our evaluation reveals that gradient structure exploitation significantly outperforms iterative optimization under differential privacy noise, achieving speedups up to 38,315×. Venom adopts a two-stage analytical framework that combines energy-based feature extraction with liquid neural dynamics, enabling robust reconstruction without noise priors or iterative sampling. These results challenge prevailing assumptions about DP robustness and highlight the need for more resilient privacy-preserving mechanisms in federated learning.

However, Venom has several limitations: (1) *Architectural Dependency*: The SPE module relies on gradient structures of fully connected classification layers, limiting applicability to architectures without traditional FC heads. Future work will investigate exploiting gradients from attention layers; (2) *Privacy Boundary*: Performance degrades significantly when $\varepsilon < 0.5$, though such extreme privacy settings are rarely used in practice due to severe utility loss; (3) *Gradient Sensitivity*: Performance decreases under aggressive gradient clipping, label smoothing, or low-bit quantization.

Defense Strategies and Implications. To counter Venom, we outline several defense strategies. Since Venom exploits gradient structures from fully connected layers, architectural defenses such as replacing FC layers with prototype-based classifiers or adopting split learning can reduce the attack surface. Algorithmic approaches include injecting correlated noise between weight and bias gradients to disrupt the linear relationships Venom leverages, as well as applying layer-wise privacy budgeting with gradient compression. Additional protections via homomorphic encryption and secure multi-party computation can mitigate structured attacks. This work shows that gradient-based attacks remain challenging for differential privacy in federated learning, underscoring the need for more robust privacy mechanisms. Venom provides both analytical insights and practical tools for privacy vulnerability assessment, highlighting the importance of security evaluation when deploying privacy-preserving techniques in decentralized systems.

Conclusion

This paper investigates gradient inversion vulnerabilities under differential privacy and introduces Venom, a framework that combines structural gradient analysis with adaptive liquid neural recovery. Venom employs a structural prior extraction module for robust feature modeling and a diffusion-driven liquid recovery network for noise-resilient reconstruction. By reformulating gradient inversion from iterative optimization into efficient single-pass analytical processing, Venom achieves substantial improvements in both reconstruction quality and computational efficiency. Overall, our results reveal critical privacy risks in federated learning and underscore the need for stronger, analytically resilient defense mechanisms for important future systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62472332) and the Hubei Provincial International Science and Technology Cooperation Project (No. 2024EHA031).

References

Ahmad, S.; Bano, S.; Verma, S.; Rawat, Y. S.; Chanda, S.; Vipparthi, S. K.; and Murala, S. 2025. PULSE: Physiological Understanding with Liquid Signal Extraction. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4574–4584. IEEE.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium*, 2633–2650.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. IEEE.

Dimitrov, D. I.; Balunović, M.; Konstantinov, N.; and Vechev, M. 2022. Data leakage in federated averaging. *Transactions on Machine Learning Research*.

Dwork, C. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, 1–12. Springer.

Feng, X.; Ma, Z.; Wang, Z.; Chegne, E. J.; Ma, M.; Abuadbba, A.; and Bai, G. 2024. Uncovering Gradient Inversion Risks in Practical Language Model Training. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 3525–3539.

Fu, J.; Hong, Y.; Ling, X.; Wang, L.; Ran, X.; Sun, Z.; Wang, W. H.; Chen, Z.; and Cao, Y. 2024. Differentially private federated learning: A systematic review. *arXiv preprint arXiv:2405.08299*.

Gao, Y.; Xie, Y.; Deng, H.; and Zhu, Z. 2025. Gradient Inversion Attack in Federated Learning: Exposing Text Data through Discrete Optimization. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2582–2591.

Geiping, J.; Bauermeister, H.; Dröge, H.; and Moeller, M. 2020. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33: 16937–16947.

Gong, X.; Li, S.; Bao, Y.; Yao, B.; Huang, Y.; Wu, Z.; Zhang, B.; Zheng, Y.; and Doermann, D. 2024. Federated learning via input-output collaborative distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22058–22066.

Hasani, R.; Lechner, M.; Amini, A.; Liebenwein, L.; Ray, A.; Tschaikowski, M.; Teschl, G.; and Rus, D. 2022. Closed-form continuous-time neural networks. *Nature Machine Intelligence*, 4(11): 992–1003.

Hasani, R.; Lechner, M.; Amini, A.; Rus, D.; and Grosu, R. 2021. Liquid time-constant networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7657–7666.

Hatamizadeh, A.; Yin, H.; Molchanov, P.; Myronenko, A.; Li, W.; Dogra, P.; Feng, A.; Flores, M. G.; Kautz, J.; Xu, D.; et al. 2023. Do gradient inversion attacks make federated learning unsafe? *IEEE Transactions on Medical Imaging*, 42(7): 2044–2056.

Hu, J.; Wang, Z.; Shen, Y.; Lin, B.; Sun, P.; Pang, X.; Liu, J.; and Ren, K. 2023. Shield against gradient leakage attacks: Adaptive privacy-preserving federated learning. *IEEE/ACM Transactions on Networking*, 32(2): 1407–1422.

Huang, W.; Ye, M.; Shi, Z.; Wan, G.; Li, H.; Du, B.; and Yang, Q. 2024a. Federated learning for generalization, robustness, fairness: A survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Huang, Y.; Gupta, S.; Song, Z.; Arora, S.; and Li, K. 2024b. Evaluating gradient inversion attacks and defenses. In *Federated Learning*, 105–122. Elsevier.

Kazemi, S. M.; Goel, R.; Eghbali, S.; Ramanan, J.; Sahota, J.; Thakur, S.; Wu, S.; Smyth, C.; Poupart, P.; and Brubaker, M. 2019. Time2vec: Learning a vector representation of time. In *International Conference on Machine Learning*, 3426–3436. PMLR.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Liang, H.; Li, Y.; Zhang, C.; Liu, X.; and Zhu, L. 2023. Egia: An external gradient inversion attack in federated learning. *IEEE Transactions on Information Forensics and Security*, 18: 4984–4995.

Liu, X.; Cai, S.; Zhou, Q.; Guo, S.; Li, R.; and Lin, K. 2025. Mjölnir: Breaking the Shield of Perturbation-Protected Gradients via Adaptive Diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 26308–26316.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and Arcas, B. A. y. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Pawlak, W. A.; Isik, M.; Le, D.; and Dikmen, I. C. 2024. Exploring liquid neural networks on loihi-2. *arXiv preprint arXiv:2407.20590*.

Ren, H.; Deng, J.; and Xie, X. 2022. Grnn: generative regression neural network—a data leakage attack for federated learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4): 1–24.

Ren, X.; Yang, S.; Zhao, C.; McCann, J.; and Xu, Z. 2024. Belt and Braces: When Federated Learning Meets Differential Privacy. *Commun. ACM*, 67(12): 66–77.

Shi, S.; Wang, N.; Xiao, Y.; Zhang, C.; Shi, Y.; Hou, Y. T.; and Lou, W. 2023. Scale-MIA: A Scalable Model Inversion Attack against Secure Federated Learning via Latent Space Reconstruction. arXiv:2311.05808.

Wang, F.; Hugh, E.; and Li, B. 2024. More than enough is too much: Adaptive defenses against gradient leakage in production federated learning. *IEEE/ACM Transactions on Networking*.

Wei, K.; Li, J.; Ma, C.; Ding, M.; Chen, W.; Wu, J.; Tao, M.; and Poor, H. V. 2023. Personalized Federated Learning With Differential Privacy and Convergence Guarantee. *IEEE Transactions on Information Forensics and Security*, 18: 4488–4503.

Wei, W.; and Liu, L. 2025. Trustworthy distributed ai systems: Robustness, privacy, and governance. *ACM Computing Surveys*, 57(6): 1–42.

Wen, Y.; Geiping, J. A.; Fowl, L.; Goldblum, M.; and Goldstein, T. 2022. Fishing for User Data in Large-Batch Federated Learning via Gradient Magnification. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 23668–23684. PMLR.

Yang, H.; Xue, D.; Ge, M.; Li, J.; Xu, G.; Li, H.; and Lu, R. 2024. Fast generation-based gradient leakage attacks: An approach to generate training data directly from the gradient. *IEEE Transactions on Dependable and Secure Computing*.

Yang, W.; Wang, S.; Wu, D.; Cai, T.; Zhu, Y.; Wei, S.; Zhang, Y.; Yang, X.; Tang, Z.; and Li, Y. 2025. Deep learning model inversion attacks and defenses: a comprehensive survey. *Artificial Intelligence Review*, 58(8): 1–52.

Ye, Z.; Luo, W.; Zhou, Q.; Zhu, Z.; Shi, Y.; and Jia, Y. 2024. Gradient inversion attacks: Impact factors analyses and privacy enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, B.; Mao, Y.; He, X.; Ping, P.; Huang, H.; and Wu, J. 2025. Exploring the Privacy-Accuracy Trade-off Using Adaptive Gradient Clipping in Federated Learning. *IEEE Transactions on Network Science and Engineering*.

Zhao, B.; Mopuri, K. R.; and Bilen, H. 2020. iDLG: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, 14774–14784.