MORE NOISE, LESS DAMAGE: ENHANCED NOISE INJECTION AGAINST GRADIENT-INVERSION IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient inversion attacks expose a key vulnerability of federated learning: even when raw data remain on local devices, shared gradients can still leak sensitive information about training samples. Existing defenses typically mitigate this risk by perturbing gradients before transmission, but stronger perturbations often interfere with optimization and reduce model performance. In this paper, we show that larger noise need not incur a proportional loss in performance when introduced in a structured way. Building on this insight, we propose TIDE, a structured noise injection framework that improves the privacy–utility trade-off by exploiting a telescoping effect across training steps. TIDE injects relatively large perturbations at each iteration while causing their cumulative effect on the optimization trajectory to partially cancel over time. This allows the transmitted gradients to receive stronger protection without imposing equally large distortion on the overall training dynamics. We theoretically show that, for smooth nonconvex objectives with a fixed learning rate η, TIDE's residual stationarity floor scales as $\mathcal{O}(\eta^2)$ whereas standard DPSGD incurs an $\mathcal{O}(\eta)$ floor; equivalently, with optimally tuned step sizes TIDE attains a strictly faster $T^{-2/3}$ convergence rate compared to DPSGD's $T^{-1/2}$. This separation explains why TIDE can remain stable even under substantially larger per-step noise. Empirically, TIDE supports noise levels up to $30\times$ larger than standard approaches while consistently maintaining better model performance. Experiments across multiple datasets and gradient inversion attacks further show that TIDE strengthens gradient privacy while preserving competitive test accuracy.

1 INTRODUCTION

Gradients are not merely numerical updates for optimization; they also encode rich information about the data from which they are computed. Once shared beyond the local device, this information creates a potential channel for privacy leakage. Studies of gradient inversion attacks make this risk explicit by reconstructing private training samples directly from communicated gradients, in some cases with striking fidelity or even exact recovery (Zhu et al., 2019; Geiping et al., 2020; Zhang et al., 2022; Jeon et al., 2021; Zhang et al., 2023). This vulnerability challenges a central premise of federated learning (FL) (Konečnỳ et al., 2016; McMahan et al., 2017): enabling collaborative model training without exposing the underlying raw data.

Various defense mechanisms have therefore been proposed to mitigate the privacy risks before sharing gradients. Among them, differential privacy (DP) (Dwork, 2008; Dwork & Roth, 2014) has become a standard approach, adding calibrated noise during training to provide formal privacy guarantees. DPSGD (Abadi et al., 2016) and its variants (Bu et al., 2023; Chilukoti et al., 2025), for example, combine gradient clipping with isotropic noise to bound the contribution of any individual training sample. Other defenses take similar routes by changing the structure or content of the shared gradient itself. For instance, gradient pruning (Huang et al., 2021; Xue et al., 2024) reduces leakage by removing components strongly correlated with private inputs, while Soteria (Sun et al., 2021) suppresses reconstruction-relevant information through gradient transformation.

Such protection is rarely free. Noise injection and gradient modification can alter the optimization process, often slowing down convergence and reducing the final model accuracy. In DP-based

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.

methods, stronger privacy is typically associated with larger per-step noise, while pruning- and transformation-based defenses may discard gradient components that are also useful for learning. These effects give rise to a persistent privacy–utility trade-off, where stronger protection against gradient inversion is often accompanied by reduced optimization efficiency or model accuracy.

In this paper, we show that stronger noise injection need not lead to a proportional loss in model utility. The key insight is that the effect of noise depends not only on its magnitude, but also on how it accumulates over the course of training. Building on this intuition, we propose TIDE, a differential noise mechanism with error correction for gradient defense in federated learning. TIDE injects the first difference $\xi(t) - \xi(t-1)$ of i.i.d. random vectors and exhibits telescoping properties that prevent unbounded noise accumulation, enabling fixed noise scales without dynamic retuning. This allows the proposed method to introduce larger noise for enhanced privacy protection while maintaining stable optimization dynamics. Consequently, TIDE achieves improved protection against gradient inversion with reduced loss in model utility.

Our validations of this new scheme are two-fold. Theoretically, for smooth nonconvex objectives with a fixed learning rate η, TIDE's residual stationarity floor (measured by the time-averaged squared gradient norm) scales as $\mathcal{O}(\eta^2)$, whereas standard noise injection methods such as DPSGD incur an $\mathcal{O}(\eta)$ floor; equivalently, with optimally tuned step sizes, TIDE attains a strictly faster $T^{-2/3}$ rate compared to DPSGD's $T^{-1/2}$. Importantly, this improved stability is achieved while injecting larger noise at each iteration, thereby providing stronger protection against gradient inversion. Empirically, TIDE tolerates noise levels up to $30\times$ larger than conventional settings (e.g., 10^{-1} versus 3×10^{-3} on CIFAR-10), while still delivering superior model performance. Extensive experiments on multiple datasets further show that TIDE preserves competitive test accuracy while substantially reducing the success of diverse gradient inversion attacks.

Our main contributions are summarized as follows:

- **Structured Noise Defense Mechanism:** We propose TIDE, a differential noise mechanism with exponential moving average error-correction. TIDE exploits a telescoping cancellation effect that bounds cumulative noise accumulation without dynamic noise scaling.
- **Theoretical Convergence & Safety Analysis:** Under smooth nonconvex objectives, we prove that TIDE achieves an $O(\eta^2)$ residual stationarity floor (vs. $O(\eta)$ for DPSGD) and a strictly faster $O(T^{-2/3})$ convergence rate (vs. $O(T^{-1/2})$ for DPSGD), while guaranteeing an enhanced single-round inversion resistance margin (ℓ_1 lower bound).
- **Extensive Empirical Validation:** Across multiple benchmarks (CIFAR-10, CIFAR-100, Tiny-ImageNet, ViT-B/16, and scaled cross-device FL), TIDE tolerates up to $30\times$ larger per-step noise multiplier σ while preserving model utility and reducing attack SSIM across 5 representative inversion attacks.

2 Related Work

Gradient Inversion Attacks Federated Learning (FL) (Konečnỳ et al., 2016; McMahan et al., 2017) enables collaborative model training without sharing raw data, but gradients exchanged during FL can still leak training samples. In particular, gradient inversion attacks (Zhu et al., 2019; Wu et al., 2023; Carletti et al., 2025; Guo et al., 2025; Sami et al., 2025; Li et al., 2025) exploit this fact to reconstruct sensitive data from the shared gradients. Early work applied reverse-engineering ideas to shadow networks (Phong et al., 2017) or jointly optimizes pseudo-inputs and labels (Zhu et al., 2019; Zhao et al., 2020). To increase visual fidelity, many works add regularizers or priors (Geiping et al., 2020; Yin et al., 2021; Wang et al., 2019; Yu et al., 2025), and analytical approaches have also been proposed to directly invert gradients under certain conditions (Fowl et al., 2021; Boenisch et al., 2023). Recent efforts leverage generative models to produce high-quality reconstructions even for large batches or high-resolution images (Jeon et al., 2021; Zhang et al., 2023; Sotthiwat et al., 2025), with GANs (Jeon et al., 2021) or diffusion models (Wu et al., 2025). Extensions to modern architectures (e.g., Vision Transformers and diffusion models) further broaden the attack surface for image-processing systems (Hatamizadeh et al., 2022; Lu et al., 2022; Huang et al., 2024; Wu et al., 2025). Beyond single-sample attacks, methods have been developed for batch reconstruction (Wen et al., 2022; Balunović et al., 2022), recursive attacks across multiple training rounds (Lam et al., 2021; Fowl et al., 2022), and attacks on language models (Gupta et al., 2022; Deng et al., 2021).

Defenses Against Inversion To mitigate this risk, various defense mechanisms have been proposed in the past studies. Cryptographic solutions, such as multi-party computation (MPC) and homomorphic encryption (HE), offer strong theoretical protection by avoiding clear-text gradient exposure, but they incur substantial communication and computation overhead (Yao, 1982; 1986; Rivest et al., 1978; Zhao et al., 2019; Fereidooni et al., 2021; Ma et al., 2022). Secure aggregation protocols provide a practical middle ground by enabling aggregation without revealing individual gradients (Bonawitz et al., 2017; Bell et al., 2020). Differential privacy (DP) and DPSGD introduce calibrated noise to obscure individual contributions and are more scalable for large FL deployments (Dwork, 2008; Dwork & Roth, 2014; Abadi et al., 2016; Mironov, 2017; Wei et al., 2020; Noble et al., 2022; Kurakin et al., 2022; Xue et al., 2024). Complementary approaches include pruning or sparsification (Huang et al., 2021; Xue et al., 2024), gradient compression techniques (Scheliga et al., 2022), and adaptive masking or representation transformations (Sun et al., 2021; Sotthiwat et al., 2026) that aim to hide reconstructible features while retaining utility (Sun et al., 2021; Huang et al., 2020). Model perturbation methods such as mixup-based defenses (Qu et al., 2023) have also been proposed to complicate reconstruction. Despite diverse mechanisms, these defenses share the core goal of transforming gradient information to impede inversion while preserving model training for image or video tasks (Takahashi et al., 2023; Tramèr et al., 2022).

3 Gradient Inversion and Desired Defense Properties

In this section, we first review gradient inversion attacks in federated learning and then identify two key properties that an effective defense should satisfy.

3.1 Gradient Inversion Attack

In federated learning (FL), clients communicate gradients rather than raw training data. Yet gradient inversion attacks (Zhu et al., 2019; Geiping et al., 2020) have shown that these updates can still reveal sensitive information about the underlying samples, and in some cases enable their reconstruction.

Primary Threat Model: Single-Step Inversion. We first consider the standard single-round gradient inversion setting (Zhu et al., 2019). An honest-but-curious server or eavesdropper observes the released gradient g'_t and is assumed to know the model architecture, global parameters w_t, and defense hyperparameters (β, σ, a), but not the client data, the realized noise ξ_t, or the internal state e_{t-1}. The adversary attempts to reconstruct a private sample x by solving

$$\hat{x} = \arg\min_{x'} \|\nabla_w L(w, x') - g\|^2 + \mathcal{R}(x'),$$

where $\mathcal{R}(x')$ denotes an optional image prior or regularizer (Geiping et al., 2020; Zhang et al., 2023). This single-step setting underlies our main evaluations and theoretical safety analysis.

Extended Threat Model: Multi-Round Attackers. We further consider more recent adversary (Li et al., 2025) and its variants that observe T consecutive releases $\{g'_t\}_{t=1}^T$ and know TIDE's noise-mixing matrix M. Such adversaries may exploit correlations across rounds through joint inversion or aggregation attacks. Details of this extended threat model and the corresponding evaluations are provided in Appendix F.

3.2 Desired Properties for Defense Algorithms

Motivated by the above attacks, we consider a general defense operation $C(g, x)$ applied to client gradient g and sample x, and evaluate it under two key criteria.

3.2.1 Convergence Condition

Let $\{w_t\}_{t\geq 0}$ denote the defended iterates. For a smooth, lower-bounded loss L, optimization performance is measured by the time-averaged squared gradient norm $\bar{G}_T := \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla L(w_t)\|^2]$. A defense is *convergent* if $\bar{G}_T \to 0$ as $T \to \infty$. Its residual stationarity floor at fixed learning rate η quantifies optimization degradation relative to standard training without requiring loss-gap or curvature assumptions.

3.2.2 SAFETY CONDITION

Meanwhile, a defense should make the released gradient difficult to invert back to the original training sample. Let $G_w(x) := \nabla_w L(w, x)$ and define $d_w(x; g') := \inf\{\|x' - x\|_2 : x' \in D,\ G_w(x') = g'\} \in [0, +\infty]$, with $\inf \emptyset = +\infty$. For all $x \in D$ and $w \notin W_\epsilon := \{w \mid \|\nabla L(w, x)\|_2 \le \epsilon,\ \forall x\}$, we require

$$d_w(x; C(G_w(x), x)) \ge \ell_1.$$

A larger ℓ_1 margin guarantees stronger input obfuscation.

Theorem 3.1. *If G_w is l_g-Lipschitz for a finite $l_g > 0$, then $d_w(x; C(G_w(x), x)) \ge \|C(G_w(x), x) - G_w(x)\|_2 / l_g$.*

Taking the infimum over x gives the corresponding uniform margin. For a fixed-context batch mean of size B, the target map is l_g/B-Lipschitz, so an additive mean perturbation Δ gives target-distance scale $B\|\Delta\|_2/l_g$. An effective defense must simultaneously achieve a small stationarity floor $\bar{G}_T$ and a large ℓ_1 margin.

4 TIDE: TELESCOPING INTEGRATED DIFFERENTIAL ERROR-CORRECTION

Standard DPSGD injects independent noise at every iteration. Although effective for privacy protection, these perturbations accumulate over time: the cumulative perturbation $\sum_t \xi_t$ grows like $O(\sigma\sqrt{T})$, causing the defended trajectory drifts arbitrarily far from the undefended one as $T \to \infty$. To control this accumulation, we introduce ***TIDE*** (Telescoping Integrated Differential Error-correction) in this part, which replaces the i.i.d. perturbations with the first differences of an i.i.d. Gaussian sequence. The key property is telescoping: over T iterations, the accumulated perturbation stays bounded uniformly in T. We further generalize this construction with an exponential moving average (EMA) filter parameterized by $\beta \in [0, 1)$, which smooths the injected perturbations while retaining the same bounded-accumulation structure. We next present the TIDE update rule and its key structural properties, with the convergence and safety analysis deferred to Section 4.2.

4.1 TELESCOPING INTEGRATED DIFFERENTIAL ERROR-CORRECTION

TIDE is built on top of the same per-sample ℓ_2-norm clipping used by DPSGD, with one structural twist: the Gaussian noise $\xi_t^{(i)}$ is drawn *independently per sample* (so each released per-sample gradient inherits the standard Gaussian-mechanism DP guarantee), while the EMA error-correction state e_{t-1} is maintained at the *batch (client) level* and shared across the B per-sample releases of the same batch. Concretely, at iteration t with mini-batch $\{x_i\}_{i=1}^B$ and per-sample gradients $g_t^{(i)} := \nabla_w L(w_t, x_i)$, we clip each per-sample gradient via $\Pi_a(g) := g \cdot \min\{1, a/\|g\|\}$ and release

$$g_t'^{(i)} = \Pi_a(g_t^{(i)}) + \xi_t^{(i)} + e_{t-1}, \qquad \xi_t^{(i)} \overset{\text{iid}}{\sim} \mathcal{N}(0, a^2\sigma_p^2 I_d), \quad i = 1, \dots, B, \tag{1}$$

$$e_t = \beta\, e_{t-1} - (1-\beta)\,\bar{\xi}_t, \qquad \bar{\xi}_t := \tfrac{1}{B}\sum_{i=1}^{B} \xi_t^{(i)} \sim \mathcal{N}\big(0, a^2\sigma^2 I_d\big), \qquad \sigma := \sigma_p/\sqrt{B}, \tag{2}$$

where $\beta \in [0, 1)$ is the smoothing parameter, σ_p is the *per-sample* Gaussian noise multiplier (used by Algorithm 1 and the DP guarantee), and σ is the corresponding *iterate-level* per-coordinate scale of the batch-aggregated noise $\bar{\xi}_t$ (used by the convergence/safety analysis). The model update aggregates the B per-sample releases, yielding $w_{t+1} = w_t - \eta\bar{g}_t' = w_t - \eta(\bar{g}_t + \bar{\xi}_t + e_{t-1})$ with $\bar{g}_t = \frac{1}{B}\sum_i \Pi_a(g_t^{(i)})$. To initialize the recursion, we draw $e_{-1} \sim \mathcal{N}(0, \sigma_e^2 I_d)$ once with $\sigma_e^2 := \frac{1-\beta}{1+\beta}\, a^2\sigma^2$, the stationary variance of the EMA driven by $\bar{\xi}_t$; with this choice, $e_t \sim \mathcal{N}(0, \sigma_e^2 I_d)$ for every $t \ge -1$, eliminating any transient term in the convergence and safety bounds. The clipping step ensures bounded per-sample sensitivity (so the same DP-style accounting applies on each release), while the additive structure $\xi_t^{(i)} + e_{t-1}$ is independent of $g_t^{(i)}$ and enables the telescoping behavior analyzed below. When clipping is inactive on the trajectory (i.e. $\|g_t^{(i)}\| \le a$ for every i), $\Pi_a(g_t^{(i)}) = g_t^{(i)}$; the convergence theory below is scoped to this regime. Clipping remains part of the algorithm and is enabled in the experiments.

Such a design endows TIDE with several important properties:

Algorithm 1 TIDE: Telescoping Integrated Differential Error-correction (one client, round t)

Require: Per-sample raw gradients $\{g_t^{(i)}\}_{i=1}^B \subset \mathbb{R}^d$, clipping threshold $a > 0$, per-sample noise multiplier $\sigma_p > 0$, smoothing parameter $\beta \in [0, 1)$, batch-level EMA state e_{t-1}

Ensure: per-sample defended gradients $\{g_t'^{(i)}\}$ and updated state e_t

1: **Initialize once** (t=0): sample $e_{-1} \sim \mathcal{N}\big(0,\ \frac{1-\beta}{1+\beta} \cdot \frac{a^2\sigma_p^2}{B} I_d\big)$ *(stationary warm-start; client-side state)*
2: **for** $i = 1, \ldots, B$ **do**
3: Clip per-sample gradient: $\bar{g}_t^{(i)} \leftarrow \Pi_a(g_t^{(i)}) = g_t^{(i)} \cdot \min\{1,\ a/\|g_t^{(i)}\|\}$
4: Sample per-sample noise: $\xi_t^{(i)} \sim \mathcal{N}(0, a^2\sigma^2 I_d)$
5: Release per-sample defended gradient: $g_t'^{(i)} \leftarrow \bar{g}_t^{(i)} + \xi_t^{(i)} + e_{t-1}$
6: **end for**
7: Form batch-aggregated noise: $\bar{\xi}_t \leftarrow \frac{1}{B}\sum_{i=1}^B \xi_t^{(i)}$
8: Update batch-level EMA state: $e_t \leftarrow \beta\, e_{t-1} - (1-\beta)\,\bar{\xi}_t$
9: **return** $\{g_t'^{(i)}\}_{i=1}^B$ and e_t

1. Telescoping structure. Rearranging the EMA error-correction recursion $e_t = \beta e_{t-1} - (1-\beta)\bar{\xi}_t$ yields $(1-\beta)\bar{\xi}_t = \beta e_{t-1} - e_t$. Substituting into the released additive perturbation gives $\bar{\xi}_t + e_{t-1} = \frac{e_{t-1} - e_t}{1-\beta}$. Summing over T iterations yields:

$$\sum_{t=0}^{T-1}(\bar{\xi}_t + e_{t-1}) = \frac{1}{1-\beta}\sum_{t=0}^{T-1}(e_{t-1} - e_t) = \frac{e_{-1} - e_{T-1}}{1-\beta}.$$

By stationarity, e_{-1} and e_{T-1} are Gaussian vectors with per-coordinate variance $\sigma_e^2 = \frac{1-\beta}{1+\beta}a^2\sigma^2$. Hence, holding clipping fixed, cumulative additive perturbation depends only on boundary correction terms rather than T. Gaussian concentration (Lemma A.4) guarantees $\|\sum_{t=0}^{T-1}(\bar{\xi}_t + e_{t-1})\| \le \frac{\|e_{-1}\| + \|e_{T-1}\|}{1-\beta} = O(a\sigma\sqrt{d}/(1-\beta))$, which is *bounded uniformly in* T, contrasting with DPSGD's $O(a\sigma\sqrt{dT})$ noise accumulation.

2. Interpretation of β. When $\beta = 0$, $e_t = -\bar{\xi}_t$, recovering pure differential noise $g_t' = \Pi_a(g_t) + \bar{\xi}_t - \bar{\xi}_{t-1}$. For $\beta \in (0, 1)$, e_t acts as an EMA filter, smoothing high-frequency noise while preserving telescoping properties.

3. Bounded perturbation by design. Preventing long-term noise accumulation avoids dynamic noise tuning and stabilizes optimization under large per-step noise. We also analyze TIDE's conditional sensitivity under a Matrix Gaussian model and give conservative full-participation accounting (Appendix A.6).

4.2 Theoretical Analysis

Notational convention. We write w_t for the iterate at round t (with x reserved for a data sample as in Sec. 3), and use two related noise multipliers: the *per-sample* σ_p in Algorithm 1, and the *iterate-level* $\sigma := \sigma_p/\sqrt{B}$ of the batch-aggregated noise $\bar{\xi}_t$ used in the analysis below (Theorems 4.3–4.7); the same convention applies to DPSGD, so with B fixed the comparison at matched (a, σ_p, η, T) coincides with that at matched (a, σ, η, T).

To study TIDE with respect to the convergence and safety conditions introduced in Section 3, we begin by imposing several standard assumptions.

Assumption 4.1 (l-Smoothness and Lower-Boundedness)**.** *The loss function L is l-smooth with lower bound: $\|\nabla L(w') - \nabla L(w)\| \le l\|w' - w\|$ for all $w, w' \in W$, and $L^* := \inf_w L(w) > -\infty$.*

Assumption 4.2 (Gaussian Noise)**.** *Noise vectors satisfy $\xi_t \sim \mathcal{N}(0, a^2\sigma^2 I_d)$, i.i.d. across t, with $\mathbb{E}[\xi_t] = 0$ and $\mathbb{E}\|\xi_t\|^2 = d\,a^2\sigma^2$. Here $a > 0$ is the clip threshold (set $a = 1$ when no clipping is used) and $\sigma > 0$ is the per-coordinate noise multiplier.*

We work throughout in the standard nonconvex setting: no curvature condition (Polyak–Łojasiewicz, strong convexity, etc.) is imposed on L, and convergence is measured by the time-averaged squared gradient norm $\bar{G}_T$ defined in Section 3.

4.2.1 Convergence Condition ($\bar{G}_T$)

We instantiate the time-averaged squared gradient norm on the original iterates $\{w_t\}$ for both methods:

$$\bar{G}_T^{\text{TIDE}} := \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla L(w_t^{\text{TIDE}})\|^2, \qquad \bar{G}_T^{\text{DPSGD}} := \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla L(w_t^{\text{DPSGD}})\|^2, \tag{3}$$

so that the bounds below are directly comparable across the two methods at matched (a, σ, η, T). The TIDE bound is obtained by analyzing an auxiliary trajectory $z_t := w_t - \frac{\eta}{1-\beta}e_{t-1}$ on which TIDE reduces to gradient descent with a small bias controlled by the EMA term, and then converting back to $\{w_t\}$ via l-smoothness; the full derivation is in Appendix A.2.

Theorem 4.3 (Stationarity Bounds for TIDE and DPSGD, Fixed Step Size)**.** *Under Assumptions 4.1 and 4.2, and assuming $\|g_t\| \le a$ holds along the trajectory (so that $\Pi_a(g_t) = g_t$, matching the standard analysis assumption for DPSGD (Chen et al., 2020)), for any $\beta \in [0, 1)$ and any fixed step size $0 < \eta \le 1/(4l)$,*

$$\textit{TIDE:}\quad \bar{G}_T^{\textit{TIDE}} \le \frac{4(L(w_0) - L^*)}{\eta T} + \frac{2l\eta a^2 d\sigma^2}{T(1-\beta^2)} + \frac{10l^2\eta^2 a^2 d\sigma^2}{1-\beta^2}, \tag{4}$$

$$\textit{DPSGD:}\quad \bar{G}_T^{\textit{DPSGD}} \le \frac{2(L(w_0) - L^*)}{\eta T} + l\eta\, a^2 d\sigma^2. \tag{5}$$

The warm-start term decays as $1/T$; hence the fixed-η asymptotic TIDE noise floor is $O(\eta^2)$, compared with the $O(\eta)$ DPSGD floor. The complete proof is provided in Appendix A.2.

Theorem 4.4 (Convergence Rates with Tuned Step Size)**.** *Under Theorem 4.3, set $\Delta = L(w_0) - L^*$. Use the stationary DPSGD minimizer and, for TIDE, the cube-root step balancing the optimization term and asymptotic η^2 floor.*

$$\textit{DPSGD: } \eta = \sqrt{\tfrac{2\Delta}{l\, a^2 d\sigma^2\, T}} \quad\Longrightarrow\quad \bar{G}_T^{\textit{DPSGD}} \le 2\sqrt{\tfrac{2l\Delta\, a^2 d\sigma^2}{T}} = O\big(T^{-1/2}\big), \tag{6}$$

$$\textit{TIDE: } \eta = \Big(\tfrac{\Delta(1-\beta^2)}{5l^2 a^2 d\sigma^2 T}\Big)^{1/3} \quad\Longrightarrow\quad \bar{G}_T^{\textit{TIDE}} \le 3\Big(\tfrac{40l^2\Delta^2 a^2 d\sigma^2}{(1-\beta^2)T^2}\Big)^{1/3} + O(T^{-4/3}) = O\big(T^{-2/3}\big). \tag{7}$$

Here the $O(T^{-4/3})$ term is the explicit warm-start contribution in Theorem 4.3, with l, Δ, a, d, σ, and β fixed. Hence TIDE attains a strictly faster $T^{-2/3}$ convergence rate compared with DPSGD's $T^{-1/2}$ at the same per-sample noise multiplier σ_p. The complete derivation is provided in Appendix A.3. This rate falls outside the i.i.d. stochastic-oracle model of Arjevani et al. (2023) because TIDE's noise is correlated across rounds (see Appendix A.3).

Remark 4.5 (Noise Floor Separation and Selection of β)**.** *In the fixed-η, $T \to \infty$ limit, the asymptotic noise-floor ratio TIDE/DPSGD equals $10l\eta/(1-\beta^2)$; the finite-T TIDE bound also has the warm-start term in* (4)*. Increasing β worsens this asymptotic optimization factor, but reduces the off-diagonal magnitude $1-\beta$ in M^{-1} and therefore reduces the multi-round sensitivity-energy diagnostic, while also reducing the single-round safety multiplier $\sqrt{2/(1+\beta)}$. These quantities define a three-way trade-off. We use $\beta = 0.5$ as an empirical default based on Appendix D.*

4.2.2 Safety Condition (ℓ_1) and Differential Privacy (DP)

By Theorem 3.1,

$$\ell_1 \ge \frac{1}{l_g}\min_{x\in D}\|C(g, x) - g\|_2. \tag{8}$$

For TIDE, $C(g_t, x) - g_t = \big(\Pi_a(g_t) - g_t\big) + \xi_t + e_{t-1}$ for any input sample x. Under the same clipping-inactive regime as in Theorem 4.3 the deterministic clip residual vanishes, so the released perturbation reduces to $\xi_t + e_{t-1}$, which is independent of x. Let

$$\sigma_e^2 := \frac{1-\beta}{1+\beta}\, a^2\sigma^2 \tag{9}$$

(the per-coordinate variance of e_{t-1}, which is constant in t thanks to the stationary warm-start $e_{-1} \sim \mathcal{N}(0, \sigma_e^2 I_d)$). Then we have the following explicit lower bound.

Theorem 4.6 (ℓ_1 Lower Bound for TIDE)**.** *Suppose Assumptions 4.1 and 4.2 hold, $\beta \in [0, 1)$, $l_g > 0$, and clipping is inactive at round t. With the stationary warm-start, for every $t \geq 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\ell_{1,\textit{TIDE}} \geq \frac{\tau}{l_g} \max\left\{\mu_d - \sqrt{2 \log \delta^{-1}}, 0\right\}, \qquad \tau^2 := a^2\sigma^2 + \sigma_e^2 = \frac{2}{1+\beta}\, a^2\sigma^2. \tag{10}$$

Here $\mu_d = \mathbb{E}\|Z\|_2 = \sqrt{2}\,\Gamma((d+1)/2)/\Gamma(d/2)$; the bound is nontrivial when $\mu_d > \sqrt{2 \log \delta^{-1}}$. The complete proof is provided in Appendix A.4.

For DPSGD, the released perturbation is $C(g_t, x) - g_t = \xi_t \sim \mathcal{N}(0, a^2\sigma^2 I_d)$ for any input sample x, so $\min_{x \in D} \|C(g_t, x) - g_t\|_2$ is itself the random quantity $\|\xi_t\|_2$. We therefore lower-bound it with high probability using Gaussian norm concentration:

Theorem 4.7 (ℓ_1 Lower Bound for DPSGD)**.** *Under Assumptions 4.1 and 4.2, let $l_g > 0$ and suppose clipping is inactive at round t. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\ell_{1,\textit{DPSGD}} \geq \frac{a\sigma}{l_g} \max\left\{\mu_d - \sqrt{2 \log \delta^{-1}}, 0\right\}. \tag{11}$$

The complete proof is provided in Appendix A.5.

Multi-Round Sensitivity Accounting. In addition to the single-round reconstruction margin above, we analyze TIDE's multi-round sensitivity under the Matrix Gaussian model (Denisov et al., 2022; Choquette-Choo et al., 2022).

Proposition 4.8 (Conditional Matrix-Gaussian Sensitivity)**.** *Consider the normalized Matrix Gaussian model $\bar{g} = g + \sigma_p M \xi$. Conditional on a fixed participation vector $s \in \{0, 1\}^T$, the order-α divergence for the corresponding neighboring shift is $\alpha \|M^{-1}s\|_2^2/(2\sigma_p^2)$. For i.i.d. $S_j \sim \mathrm{Bern}(q)$ with $q > 0$, the ratio $r(T, q) := \mathbb{E}\|M^{-1}\mathbf{1}_S\|^2 / \mathbb{E}\|\mathbf{1}_S\|^2$ equals $1 + A_{\beta,q}(T-1) + B_{\beta,q}(T-1)^2$. This quantity is the expected conditional sensitivity-energy diagnostic; Rényi accounting for the Poisson-subsampled mixture is determined by its log moment. At full participation, $r(T, 1)$ is the exact deterministic divergence ratio and gives conservative accounting without subsampling amplification (Appendix A.6).*

Remark 4.9 (Comparison with DPSGD and Empirical Compensation)**.** *At matched (a, σ_p), TIDE's single-round reconstruction margin (ℓ_1 bound) exceeds DPSGD's by $\sqrt{2/(1+\beta)} \approx 1.15\times$ (at $\beta = 0.5$). Multi-round noise correlation increases both the conservative full-participation sensitivity and the expected conditional sensitivity energy reported in Table 2. However, TIDE's $O(\eta^2)$ stationarity floor (Theorem 4.3) nevertheless permits substantially larger noise σ_p in our experiments, producing a favorable empirical privacy–utility trade-off. Formal Poisson-subsampled RDP accounting is determined by the mixture log moment and remains a direction for future work.*

5 Experiments

We now evaluate the effectiveness of the proposed method through a series of experiments. Our goal is to assess both its ability to protect against gradient inversion attacks and its impact on model performance. We compare TIDE with standard baselines across multiple datasets and attack settings, considering a wide range of noise levels.

5.1 Experiment Setup

Our experimental setup follows common practices from prior gradient-inversion work (Zhu et al., 2019; Zhang et al., 2023). We use ResNet-18 models with sigmoid activations, which have been shown to increase susceptibility to inversion attacks (Zhu et al., 2019). We compare TIDE with six baseline defenses: DP (Abadi et al., 2016), Outpost (Wang et al., 2024), DP-Auto-S (Bu et al., 2023), DGP (Xue et al., 2024), DP-Step-Decay (Chilukoti et al., 2025) and GradDropout (Sotthiwat et al., 2026). To assess defense effectiveness, we evaluate against six representative gradient inversion attacks: GI (Zhu et al., 2019), IG (Geiping et al., 2020), iDLG (Zhao et al., 2020), CI-Net (Zhang et al., 2023), DGGI (Wu et al., 2025) and TGIAs (Li et al., 2025) (used for multi-round attack in Appendix F). We report both model performance and privacy leakage, where the latter is quantified using SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018). Detailed settings are in Appendix B.

Table 1: Main benchmark performance on CIFAR-10 and CIFAR-100 under defense methods, reporting test accuracy (%) and SSIM values under five gradient inversion attacks. Default noise scale is $\sigma = 3 \times 10^{-3}$ for DP families unless otherwise specified; TIDE uses $\beta = 0.5$.

(a) CIFAR-10 Benchmark

Defense	Test Acc (%) ↑	Attack SSIM ↓				
		GI	IG	iDLG	CI-Net	DGGI
None	$90.55_{\pm 0.55}$	$0.4578_{\pm 0.1211}$	$0.8848_{\pm 0.0500}$	$0.8697_{\pm 0.0484}$	$0.8284_{\pm 0.0961}$	$0.9391_{\pm 0.0642}$
Outpost	$89.40_{\pm 0.77}$	$0.3848_{\pm 0.1252}$	$0.7832_{\pm 0.0491}$	$0.7688_{\pm 0.0601}$	$0.6761_{\pm 0.1128}$	$0.8800_{\pm 0.0664}$
DGP	$90.34_{\pm 0.34}$	$0.4007_{\pm 0.1106}$	$0.8503_{\pm 0.0628}$	$0.8253_{\pm 0.0613}$	$0.7129_{\pm 0.1382}$	$0.9025_{\pm 0.0619}$
GradDropout (p=0.3)	$87.84_{\pm 0.41}$	$0.0671_{\pm 0.0146}$	$0.1705_{\pm 0.0533}$	$0.1588_{\pm 0.0495}$	$0.1645_{\pm 0.0678}$	$0.2204_{\pm 0.0296}$
DP-SGD (no-clip)	$87.45_{\pm 0.37}$	$0.1181_{\pm 0.0282}$	$0.4204_{\pm 0.0816}$	$0.4056_{\pm 0.0874}$	$0.4544_{\pm 0.1050}$	$0.4365_{\pm 0.0879}$
DP-SGD	$87.69_{\pm 0.39}$	$0.0274_{\pm 0.0076}$	$0.0287_{\pm 0.0082}$	$0.0280_{\pm 0.0081}$	$0.0460_{\pm 0.0160}$	$0.0346_{\pm 0.0097}$
DP-Auto-S	$88.26_{\pm 0.28}$	$0.0298_{\pm 0.0089}$	$0.0336_{\pm 0.0070}$	$0.0311_{\pm 0.0066}$	$0.0488_{\pm 0.0152}$	$0.0340_{\pm 0.0056}$
DP-Step-Decay	$91.00_{\pm 0.20}$	$0.0274_{\pm 0.0076}$	$0.0287_{\pm 0.0082}$	$0.0280_{\pm 0.0081}$	$0.0460_{\pm 0.0160}$	$0.0346_{\pm 0.0097}$
TIDE (no-clip)	$90.86_{\pm 0.13}$	$0.1016_{\pm 0.0294}$	$0.3470_{\pm 0.0665}$	$0.3398_{\pm 0.0700}$	$0.3977_{\pm 0.0837}$	$0.3701_{\pm 0.0686}$
TIDE (σ=3×10^{-3})	$91.16_{\pm 0.17}$	$0.0286_{\pm 0.0111}$	$0.0261_{\pm 0.0065}$	$0.0269_{\pm 0.0067}$	$0.0429_{\pm 0.0125}$	$0.0303_{\pm 0.0085}$
TIDE (σ=1×10^{-2})	$\mathbf{91.33}_{\pm 0.10}$	$0.0277_{\pm 0.0077}$	$0.0190_{\pm 0.0055}$	$0.0214_{\pm 0.0046}$	$0.0309_{\pm 0.0089}$	$0.0216_{\pm 0.0035}$
TIDE (σ=3×10^{-2})	$90.60_{\pm 0.45}$	$\mathbf{0.0276}_{\pm 0.0100}$	$\mathbf{0.0171}_{\pm 0.0038}$	$\mathbf{0.0166}_{\pm 0.0038}$	$\mathbf{0.0290}_{\pm 0.0060}$	$\mathbf{0.0205}_{\pm 0.0065}$

(b) CIFAR-100 Benchmark

Defense	Test Acc (%) ↑	Attack SSIM ↓				
		GI	IG	iDLG	CI-Net	DGGI
None	$68.52_{\pm 0.49}$	$0.5946_{\pm 0.0855}$	$0.9384_{\pm 0.0178}$	$0.9257_{\pm 0.0215}$	$0.9202_{\pm 0.0741}$	$0.9321_{\pm 0.0355}$
Outpost	$65.99_{\pm 0.62}$	$0.5426_{\pm 0.1008}$	$0.8636_{\pm 0.0249}$	$0.8430_{\pm 0.0217}$	$0.8148_{\pm 0.1342}$	$0.8862_{\pm 0.0397}$
DGP	$67.50_{\pm 0.11}$	$0.8282_{\pm 0.0780}$	$0.9058_{\pm 0.0320}$	$0.9016_{\pm 0.0314}$	$0.8998_{\pm 0.0895}$	$0.9410_{\pm 0.0348}$
GradDropout (p=0.3)	$64.15_{\pm 0.19}$	$0.0342_{\pm 0.0107}$	$0.1001_{\pm 0.0117}$	$0.0986_{\pm 0.0089}$	$0.0637_{\pm 0.0228}$	$0.1231_{\pm 0.0138}$
DP-SGD (no-clip)	$63.67_{\pm 0.34}$	$0.1302_{\pm 0.0326}$	$0.4545_{\pm 0.0422}$	$0.4441_{\pm 0.0434}$	$0.4934_{\pm 0.0597}$	$0.4782_{\pm 0.0455}$
DP-SGD	$62.86_{\pm 0.41}$	$0.0318_{\pm 0.0094}$	$0.0323_{\pm 0.0076}$	$0.0351_{\pm 0.0068}$	$0.0459_{\pm 0.0129}$	$0.0379_{\pm 0.0081}$
DP-Auto-S	$63.12_{\pm 0.13}$	$0.0293_{\pm 0.0063}$	$0.0296_{\pm 0.0045}$	$0.0310_{\pm 0.0060}$	$0.0413_{\pm 0.0143}$	$0.0340_{\pm 0.0056}$
DP-Step-Decay	$67.89_{\pm 0.45}$	$0.0318_{\pm 0.0094}$	$0.0323_{\pm 0.0076}$	$0.0351_{\pm 0.0068}$	$0.0459_{\pm 0.0129}$	$0.0379_{\pm 0.0081}$
TIDE (no-clip)	$68.50_{\pm 0.21}$	$0.1108_{\pm 0.0198}$	$0.3743_{\pm 0.0390}$	$0.3640_{\pm 0.0317}$	$0.4249_{\pm 0.0541}$	$0.3987_{\pm 0.0405}$
TIDE (σ=3×10^{-3})	$\mathbf{68.57}_{\pm 0.50}$	$0.0324_{\pm 0.0098}$	$0.0276_{\pm 0.0046}$	$0.0287_{\pm 0.0040}$	$0.0458_{\pm 0.0130}$	$0.0313_{\pm 0.0057}$
TIDE (σ=1×10^{-2})	$68.40_{\pm 0.09}$	$\mathbf{0.0277}_{\pm 0.0077}$	$0.0190_{\pm 0.0055}$	$0.0214_{\pm 0.0046}$	$0.0309_{\pm 0.0089}$	$\mathbf{0.0216}_{\pm 0.0035}$
TIDE (σ=3×10^{-2})	$67.46_{\pm 0.30}$	$0.0291_{\pm 0.0077}$	$\mathbf{0.0174}_{\pm 0.0062}$	$\mathbf{0.0184}_{\pm 0.0054}$	$\mathbf{0.0301}_{\pm 0.0087}$	$0.0257_{\pm 0.0068}$

5.2 CIFAR-10 Results

Table 1a reports test accuracy and gradient reconstruction metrics across defense methods on CIFAR-10. In the absence of defense, the model achieves 90.55% accuracy but remains severely vulnerable to gradient inversion, with SSIM reaching up to 0.9391 under DGGI and 0.8848 under IG. Heuristic defenses present clear limitations: Outpost (89.40%) and DGP (90.34%) reduce SSIM under mild optimization-based attacks like GI, but fail to defend against generative inversion models (CI-Net SSIM > 0.67). GradDropout ($p = 0.3$) reduces reconstruction leakage but suffers a substantial utility penalty, dropping test accuracy to 87.84%.

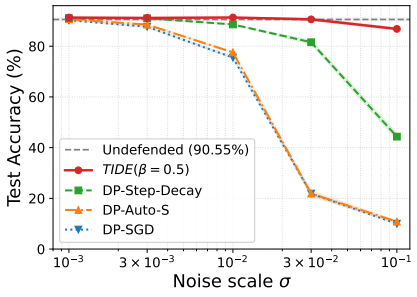

Figure 1: Test accuracy of DP and TIDE on CIFAR-10 across iterate-level noise scales σ.

Standard DP illustrates the classic privacy–utility conflict: applying per-sample clipping and noise ($\sigma = 3 \times 10^{-3}$) suppresses attack SSIM but degrades accuracy to 87.69%. Modern adaptive DP baselines like DP-Auto-S (88.26%) and DP-Step-Decay (91.00%) mitigate utility loss, yet TIDE consistently matches or outperforms them while offering far stronger privacy protection under higher noise.

Specifically, at baseline noise $\sigma = 3 \times 10^{-3}$, TIDE achieves 91.16% accuracy (+3.47% over DP) with comparable SSIM. Crucially, as the noise scale increases to $\sigma = 1 \times 10^{-2}$ and $\sigma = 3 \times 10^{-2}$

($10\times$ higher), TIDE maintains near-optimal accuracy (91.33% and 90.60%) while suppressing attack SSIM across all five attacks to extraordinarily low levels (≤ 0.0290). Figure 1 illustrates this stability across noise multipliers up to $\sigma = 10^{-1}$, where DP families experiences catastrophic accuracy collapse. As such, TIDE **tolerates noise levels up to 30$\times$ larger** than conventional DP families (e.g., 10^{-1} versus 3×10^{-3} on CIFAR-10), while still delivering superior model performance.

5.3 CIFAR-100 Results

We extend our evaluation to CIFAR-100 in Table 1b, testing scalability under a $10\times$ larger classification space. Under 100 classes, the vulnerability of unprotected training remains acute (SSIM > 0.92 for IG, iDLG, and DGGI). Standard DP-SGD ($\sigma = 3 \times 10^{-3}$) incurs a severe 5.66% accuracy drop down to 62.86%, while DP-Auto-S reaches 63.12% and DP-Step-Decay reaches 67.89%. In contrast, TIDE at the same noise scale achieves 68.57% test accuracy, outperforming all DP baselines and matching the undefended baseline (68.52%).

Furthermore, when scaling the noise multiplier up to $\sigma = 1\times10^{-2}$ and $\sigma = 3\times10^{-2}$, TIDE maintains 68.40% and 67.46% accuracy while driving SSIM down to ≤ 0.0309 across all five inversion attacks. This confirms that TIDE's telescoping error-correction property prevents error accumulation even in complex, multi-class optimization landscapes, enabling stronger per-step privacy without sacrificing model performance.

5.4 Generalization and Robustness Evaluation

To verify that TIDE's design is broadly applicable, we evaluate it across larger vision benchmarks, modern Vision Transformers (ViT), scaled federated learning regimes, and multi-round adaptive attacks:

Tiny-ImageNet Benchmark. On 64×64 Tiny-ImageNet, TIDE ($\sigma = 3 \times 10^{-2}$) reaches 53.09% test accuracy (matching undefended 52.84%), outperforming DP-SGD (49.65%) and DP-Step-Decay (52.52%) while maintaining SSIM ≤ 0.0311 under all 5 attacks (Table 3 in Appendix C).

Vision Transformer (ViT-B/16) Fine-Tuning. Fine-tuning ViT-B/16 on CIFAR-100, TIDE ($\sigma = 3 \times 10^{-2}$) retains 90.52% accuracy versus DP's 83.17% ($+7.35\%$ gain), while suppressing CI-Net generative reconstruction SSIM to 0.034 (Appendix J). Moreover, under large noise ($\sigma = 3 \times 10^{-1}$), DP-SGD collapses to 27.45% whereas TIDE maintains 89.13% accuracy.

Scaled Cross-Device FL & Non-IID Partitions. We further extend the setting to: (1) partial client participation and (2) Dirichlet non-IID partitions up to 100 clients. For both settings, TIDE consistently preserves optimization stability (Appendix I).

Ablations & LPIPS Evaluation. TIDE remains robust across the tested $\beta \in \{0, 0.1, 0.5, 0.9\}$ values (Appendix D); $\beta = 0.5$ is selected as an empirical default from this grid. Perceptual LPIPS evaluation (Appendix E) further confirms TIDE severely degrades feature similarity under CI-Net and DGGI attacks.

Multi-Round & Adaptive Attacks. We further evaluate adversaries that jointly exploit $T = 10$ consecutive gradient releases from the same client mini-batch, including TGIAs-RO and a TIDE-aware De-telescope GLS attack that explicitly inverts the temporal transform and accounts for the exact shared-state covariance. TIDE keeps SSIM below 0.065 under TGIAs-RO and achieves only 0.0111 SSIM under De-telescope GLS, indicating that the evaluated multi-round correlation-aware attacks do not recover meaningful reconstructions (Appendix F).

6 Conclusion

TIDE uses telescoping differential noise and error correction to defend shared gradients while limiting disruption to optimization. Its improved stationarity guarantees and consistent results across datasets and attacks demonstrate a stronger privacy–utility trade-off than standard noise injection. These results demonstrate that temporally structured perturbations can provide a stronger privacy–utility trade-off.

AI USE STATEMENT

LLMs are used only for minor language polishing of the manuscript (e.g., improving clarity). They are not used to generate scientific claims, design experimental results, or perform any form of automated analysis. All key ideas, methodological contributions, and experimental findings are developed and verified by the authors.

ETHICS STATEMENT

This work studies defenses against reconstruction of private training data from shared gradients. The experiments use standard research datasets and are intended to improve privacy protection. The proposed techniques could also inform stronger adaptive attacks; we therefore report the threat model, limitations, and defense-aware evaluations to support responsible use.

REPRODUCIBILITY STATEMENT

The algorithm and its hyperparameters are specified in Algorithm 1; assumptions and complete proofs are provided in the appendix. Experimental settings, additional datasets, ablations, and adaptive multi-round evaluations are also documented in the appendix and supplementary material.

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. doi: 10.1145/2976749.2978318.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1): 165–214, 2023. doi: 10.1007/s10107-022-01822-7.

Mislav Balunović, Dimitar I Dimitrov, Nikola Jovanović, and Martin Vechev. Lamp: Extracting text from gradients with language model priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7641–7654, 2022. doi: 10.52202/068431-0555.

James Henry Bell, Kallista A Bonawitz, Adrià Gascón, Tancrède Lepoint, and Mariana Raykova. Secure single-server aggregation with (poly)logarithmic overhead. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2020. doi: 10.1145/3372297.3417885.

Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, and Nicolas Papernot. When the curious abandon honesty: Federated learning is not private. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2023. doi: 10.1109/EuroSP57164.2023.00020.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017. doi: 10.1145/3133956.3133982.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. volume 36, pp. 41727–41764, 2023.

Vincenzo Carletti, Pasquale Foggia, Carlo Mazzocca, Giuseppe Parrella, and Mario Vento. SoK: Gradient inversion attacks in federated learning. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 6439–6459, 2025. URL `https://www.usenix.org/conference/usenixsecurity25/presentation/carletti`.

Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private SGD: A geometric perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 13773–13782, 2020.

Sai Venkatesh Chilukoti, Md Imran Hossen, Liqun Shan, Vijay Srinivas Tida, Mahathir Mohammad Bappy, Wenmeng Tian, and Xiali Hei. Dp-sgd-global-adapt-v2-s: Triad improvements of privacy, accuracy and fairness via step decay noise multiplier and step decay upper clipping threshold. *Electronic Commerce Research and Applications*, 70:101476, 2025.

Christopher A Choquette-Choo, H Brendan McMahan, Keith Rush, and Abhradeep Thakurta. Multi-epoch matrix factorization mechanisms for private machine learning. *arXiv preprint arXiv:2211.06530*, 2022.

Jieren Deng, Yijue Wang, Ji Li, Chenghong Wang, Chao Shang, Hang Liu, Sanguthevar Rajasekaran, and Caiwen Ding. Tag: Gradient attack on transformer-based language models. In *Findings of the association for computational linguistics: EMNLP 2021*, pp. 3600–3610, 2021. doi: 10.18653/v1/2021.findings-emnlp.305.

Sergey Denisov, H Brendan McMahan, John Rush, Adam Smith, and Abhradeep Guha Thakurta. Improved differential privacy for sgd via optimal private linear operators on adaptive streams. *Advances in Neural Information Processing Systems*, 35:5910–5924, 2022.

Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pp. 1–19. Springer, 2008. doi: 10.1007/978-3-540-79228-4_1.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–487, 2014. doi: 10.1561/0400000042.

Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Helen Möllering, Thien Duc Nguyen, Phillip Rieger, Ahmad-Reza Sadeghi, Thomas Schneider, Hossein Yalame, et al. Safelearn: Secure aggregation for private federated learning. In *2021 IEEE Security and Privacy Workshops (SPW)*, pp. 56–62. IEEE, 2021. doi: 10.1109/SPW53761.2021.00017.

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.

Liam Fowl, Jonas Geiping, Steven Reich, Yuxin Wen, Wojtek Czaja, Micah Goldblum, and Tom Goldstein. Decepticons: Corrupted transformers breach privacy in federated learning for language models. *arXiv preprint arXiv:2201.12675*, 2022.

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

Pengxin Guo, Runxi Wang, Shuang Zeng, Jinjing Zhu, Haoning Jiang, Yanran Wang, Yuyin Zhou, Feifei Wang, Hui Xiong, and Liangqiong Qu. Exploring the vulnerabilities of federated learning: A deep dive into gradient inversion attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Samyak Gupta, Yangsibo Huang, Zexuan Zhong, Tianyu Gao, Kai Li, and Danqi Chen. Recovering private text in federated learning of language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. doi: 10.52202/068431-0590.

Ali Hatamizadeh, Hongxu Yin, Holger R Roth, Wenqi Li, Jan Kautz, Daguang Xu, and Pavlo Molchanov. Gradvit: Gradient inversion of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10011–10020, 2022. doi: 10.1109/CVPR52688.2022.00978.

Jiyue Huang, Chi Hong, Lydia Y Chen, and Stefanie Roos. Gradient inversion of federated diffusion models. *arXiv preprint arXiv:2405.20380*, 2024.

Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. InstaHide: Instance-hiding schemes for private distributed learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 4507–4518. PMLR, 2020. URL `https://proceedings.mlr.press/v119/huang20i.html`.

Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in neural information processing systems*, 34:7232–7241, 2021.

Jinwoo Jeon, Jaechang Kim, Kangwook Lee, Sewoong Oh, and Jungseul Ok. Gradient inversion with generative image prior. *Advances in Neural Information Processing Systems*, 34:29898–29908, 2021.

Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning (ICML)*, pp. 5213–5225, 2021.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Alexey Kurakin, Steve Chien, Shuang Song, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022.

Maximilian Lam, Gu-Yeon Wei, David Brooks, Vijay Janapa Reddi, and Michael Mitzenmacher. Gradient disaggregation: Breaking privacy in federated learning by reconstructing the user participant matrix. In *International Conference on Machine Learning (ICML)*, 2021.

Bowen Li, Hanlin Gu, Ruoxin Chen, Jie Li, Chentao Wu, Na Ruan, Xueming Si, and Lixin Fan. Temporal gradient inversion attacks with robust optimization. *IEEE Transactions on Dependable and Secure Computing*, 22(4):3383–3397, 2025. doi: 10.1109/TDSC.2025.3532339.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

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. 10041–10050, 2022. doi: 10.1109/CVPR52688.2022.00981.

Jing Ma, Si-Ahmed Naas, Stephan Sigg, and Xixiang Lyu. Privacy-preserving federated learning based on multi-key homomorphic encryption. *International Journal of Intelligent Systems*, 37(9): 5880–5901, 2022. doi: 10.1002/int.22818.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017. doi: 10.1109/CSF.2017.11.

Maxence Noble, Aurélien Bellet, and Aymeric Dieuleveut. Differentially private federated learning on heterogeneous data. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 10110–10145. PMLR, 2022. URL https://proceedings.mlr.press/v151/noble22a.html.

Le Trieu Phong, Yoshinori Aono, Takuya Hayashi, Lihua Wang, and Shiho Moriai. Privacy-preserving deep learning: Revisited and enhanced. In *International Conference on Applications and Techniques in Information Security*, pp. 100–110. Springer, 2017. doi: 10.1007/978-981-10-5421-1_9.

Zhe Qu, Xingyu Li, Xiao Han, Rui Duan, Chengchao Shen, and Lixing Chen. How to prevent the poor performance clients for personalized federated learning? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12167–12176, 2023. doi: 10.1109/CVPR52729.2023.01171.

Ronald L Rivest, Len Adleman, Michael L Dertouzos, et al. On data banks and privacy homomorphisms. *Foundations of secure computation*, 4(11):169–180, 1978.

Hasin Us Sami, Swapneel Sen, Amit K Roy-Chowdhury, Srikanth V Krishnamurthy, and Basak Guler. Gradient inversion attacks on parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10224–10234, 2025. doi: 10.1109/CVPR52734.2025.00956.

Daniel Scheliga, Patrick Mäder, and Marco Seeland. Precode: A generic model extension to prevent deep gradient leakage. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. doi: 10.1109/WACV51458.2022.00366.

Ekanut Sotthiwat, Liangli Zhen, Chi Zhang, Zengxiang Li, and Rick Siow Mong Goh. Generative image reconstruction from gradients. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):21–31, 2025. doi: 10.1109/TNNLS.2024.3383722.

Ekanut Sotthiwat, Chi Zhang, Xiaokui Xiao, and Liangli Zhen. Safeguarding federated learning from data reconstruction attacks via gradient dropout. *IEEE Transactions on Information Forensics and Security*, 21:1874–1888, 2026. doi: 10.1109/TIFS.2026.3659401.

Jingwei Sun, Ang Li, Binghui Wang, Huanrui Yang, Hai Li, and Yiran Chen. Soteria: Provable defense against privacy leakage in federated learning from representation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9307–9315, 2021. doi: 10.1109/CVPR46437.2021.00919.

Hideaki Takahashi, Jingjing Liu, and Yang Liu. Breaching fedmd: image recovery via paired-logits inversion attack. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12198–12207. IEEE, 2023.

Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2022.

Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Fei Wang, Ethan Hugh, and Baochun Li. More than enough is too much: Adaptive defenses against gradient leakage in production federated learning. *IEEE/ACM Transactions on Networking*, 32(4):3061–3075, 2024.

Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 2512–2520. IEEE, 2019. doi: 10.1109/INFOCOM.2019.8737416.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE transactions on information forensics and security*, 15:3454–3469, 2020. doi: 10.1109/TIFS.2020.2988575.

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. doi: 10.1016/j.inffus.2024.102620.

Ruihan Wu, Xiangyu Chen, Chuan Guo, and Kilian Q Weinberger. Learning to invert: Simple adaptive attacks for gradient inversion in federated learning. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216, pp. 2293–2303. PMLR, 2023. URL `https://proceedings.mlr.press/v216/wu23a.html`.

Lulu Xue, Shengshan Hu, Ruizhi Zhao, Leo Yu Zhang, Shengqing Hu, Lichao Sun, and Dezhong Yao. Revisiting gradient pruning: A dual realization for defending against gradient attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6404–6412, 2024.

Andrew C Yao. Protocols for secure computations. In *23rd annual symposium on foundations of computer science (sfcs 1982)*. IEEE, 1982. doi: 10.1109/SFCS.1982.38.

Andrew Chi-Chih Yao. How to generate and exchange secrets. In *27th annual symposium on foundations of computer science (Sfcs 1986)*, pp. 162–167. IEEE, 1986. doi: 10.1109/SFCS.1986.25.

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. 16332–16341, 2021. doi: 10.1109/CVPR46437.2021.01607.

Wenbo Yu, Hao Fang, Bin Chen, Xiaohang Sui, Chuan Chen, Hao Wu, Shu-Tao Xia, and Ke Xu. Gi-nas: Boosting gradient inversion attacks through adaptive neural architecture search. *IEEE Transactions on Information Forensics and Security*, 2025. doi: 10.1109/TIFS.2025.3589127.

Chi Zhang, Zhang Xiaoman, Ekanut Sotthiwat, Yanyu Xu, Ping Liu, Liangli Zhen, and Yong Liu. Generative gradient inversion via over-parameterized networks in federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5103–5112, 2023. doi: 10.1109/ICCV51070.2023.00473.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Rui Zhang, Song Guo, Junxiao Wang, Xin Xie, and Dacheng Tao. A survey on gradient inversion: Attacks, defenses and future directions. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 5678–5685, 2022. doi: 10.24963/ijcai.2022/791.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. iDLG: Improved deep leakage from gradients. *CoRR*, 2020. URL `https://arxiv.org/abs/2001.02610`.

Chuan Zhao, Shengnan Zhao, Minghao Zhao, Zhenxiang Chen, Chong-Zhi Gao, Hongwei Li, and Yu-an Tan. Secure multi-party computation: theory, practice and applications. *Information Sciences*, 476:357–372, 2019. doi: 10.1016/j.ins.2018.10.024.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Proceedings of the Advances in Neural Information Processing Systems*, 2019.

APPENDIX CATALOGUE

A Complete Proofs for Theorems and Lemmas **16**

A.1 Preliminary lemmas 16

A.2 Proof of Theorem 4.3 (stationarity bounds, fixed step size) 17

A.3 Proof of Theorem 4.4 (convergence rates with tuned step size) 19

A.4 Proof of Theorem 4.6 (ℓ_1 lower bound for TIDE) 19

A.5 Proof of Theorem 4.7 (ℓ_1 lower bound for DPSGD) 20

A.6 Proof of Proposition 4.8 (Conditional Matrix-Gaussian Sensitivity) 20

B Detailed Experiment Settings **22**

C Experiments on Additional Datasets **23**

C.1 Tiny-ImageNet Results 23

C.2 Summary Across Datasets 23

D Ablation Studies **24**

E Perceptual Quality: LPIPS Evaluation on CIFAR-10 **24**

F Evaluation under Multi-Round and Adaptive Attacks **25**

G Visualization of Reconstructed Images **26**

H Runtime Cost **27**

I Cross-Device Federated Learning Experiment **28**

J Vision Transformer (ViT-B/16) Full Fine-Tuning Check **28**

A Complete Proofs for Theorems and Lemmas

Throughout this appendix we adopt the notation from Section 3 and Section 4.1, and work under Assumptions 4.1–4.2 unless stated otherwise. No curvature condition (Polyak–Łojasiewicz, strong convexity, etc.) is used anywhere in this appendix.

A.1 Preliminary lemmas

Lemma A.1 (Standard l-smoothness inequalities)**.** *Let $L : \mathbb{R}^m \to \mathbb{R}$ be differentiable with l-Lipschitz gradient for $l > 0$, and let L^* be a global lower bound: $L^* \le L(v)$ for every v. Then for all u, v,*

$$L(v) \le L(u) + \langle \nabla L(u), v - u \rangle + \tfrac{l}{2}\|v - u\|^2, \tag{12}$$

$$\|\nabla L(u)\|^2 \le 2l\,(L(u) - L^*). \tag{13}$$

Proof. Equation (12) is the standard descent lemma. Since $l > 0$, set $v = u - \nabla L(u)/l$ to obtain $L(v) \le L(u) - \|\nabla L(u)\|^2/(2l)$. The global bound $L^* \le L(v)$ then gives (13). □

Lemma A.2 (Stationary variance of the EMA correction under warm-start)**.** *For TIDE with smoothing parameter $\beta \in [0, 1)$, $\xi_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, a^2\sigma^2 I_d)$, and stationary warm-start $e_{-1} \sim \mathcal{N}(0, \sigma_e^2 I_d)$ (independent of $\{\xi_t\}_{t\ge 0}$) with $\sigma_e^2 := \frac{1-\beta}{1+\beta}\, a^2\sigma^2$, the recursion $e_t = \beta e_{t-1} - (1-\beta)\xi_t$ produces, for every $t \ge -1$, a Gaussian vector $e_t \sim \mathcal{N}(0, \sigma_e^2 I_d)$. In particular, $\mathbb{E}\|e_t\|^2 = d\sigma_e^2 = \frac{1-\beta}{1+\beta} a^2 d\sigma^2$ exactly, and ξ_t is independent of e_{t-1} for all $t \ge 0$.*

Proof. If $e_{t-1} \sim \mathcal{N}(0, \sigma_e^2 I_d) \perp \xi_t \sim \mathcal{N}(0, a^2\sigma^2 I_d)$, then e_t is centered Gaussian with per-coordinate variance $\beta^2\sigma_e^2 + (1-\beta)^2 a^2\sigma^2$. Substituting σ_e^2 gives $\beta^2 \frac{1-\beta}{1+\beta} a^2\sigma^2 + (1-\beta)^2 a^2\sigma^2 = \frac{1-\beta}{1+\beta} a^2\sigma^2 = \sigma_e^2$. Induction propagates the stationary law from the base case $t = -1$. Independence follows because e_{t-1} depends only on $(e_{-1}, \xi_0, \ldots, \xi_{t-1})$. □

Lemma A.3 (The z-transformation)**.** *Let $z_t := w_t - \frac{\eta}{1-\beta}\, e_{t-1}$ for $t \ge 0$. Then the TIDE iterates satisfy*

$$z_{t+1} \;=\; z_t \;-\; \eta\,\nabla L(w_t) \;=\; z_t \;-\; \eta\big(\nabla L(z_t) + b_t\big), \qquad b_t := \nabla L(w_t) - \nabla L(z_t), \tag{14}$$

and the bias term satisfies $\|b_t\| \le l\|w_t - z_t\| = \frac{l\eta}{1-\beta}\|e_{t-1}\|$.

Proof. Step 1 (telescoping identity). The recursion $e_t = \beta e_{t-1} - (1-\beta)\xi_t$ rearranges to $(1-\beta)\xi_t = \beta e_{t-1} - e_t$, so

$$\xi_t \;=\; \frac{\beta e_{t-1} - e_t}{1-\beta} \quad\Longrightarrow\quad \xi_t + e_{t-1} \;=\; \frac{\beta e_{t-1} - e_t}{1-\beta} + e_{t-1} \;=\; \frac{e_{t-1} - e_t}{1-\beta},$$

where the last equality uses $\beta e_{t-1}/(1-\beta) + e_{t-1} = \big(\beta + (1-\beta)\big)e_{t-1}/(1-\beta) = e_{t-1}/(1-\beta)$.

Step 2 (substitute into the TIDE update). The TIDE iterate update under clipping-inactivity is $w_{t+1} = w_t - \eta\big(\nabla L(w_t) + \xi_t + e_{t-1}\big)$. Using Step 1,

$$w_{t+1} \;=\; w_t \;-\; \eta\,\nabla L(w_t) \;-\; \frac{\eta}{1-\beta}\,(e_{t-1} - e_t).$$

Step 3 (rewrite in the z-variable). Subtract $\frac{\eta}{1-\beta} e_t$ from both sides:

$$\underbrace{w_{t+1} - \frac{\eta}{1-\beta}\, e_t}_{=\, z_{t+1}} \;=\; \underbrace{w_t - \frac{\eta}{1-\beta}\, e_{t-1}}_{=\, z_t} \;-\; \eta\,\nabla L(w_t).$$

This is the first equality $z_{t+1} = z_t - \eta\,\nabla L(w_t)$. The second equality follows from inserting $\nabla L(w_t) = \nabla L(z_t) + b_t$.

Step 4 (bias bound). By Assumption 4.1, $\|b_t\| = \|\nabla L(w_t) - \nabla L(z_t)\| \le l\|w_t - z_t\|$. By definition $w_t - z_t = \frac{\eta}{1-\beta} e_{t-1}$, so $\|w_t - z_t\| = \frac{\eta}{1-\beta}\|e_{t-1}\|$. □

Lemma A.4 (Gaussian norm concentration; cf. Vershynin (2018)). *Let $Z \sim \mathcal{N}(0, I_d)$. Its norm has the chi distribution with d degrees of freedom and explicit mean $\mu_d = \mathbb{E}\|Z\|_2 = \sqrt{2}\,\Gamma((d+1)/2)/\Gamma(d/2)$. For integer d,*

$$\mu_d = \begin{cases} \sqrt{\pi/2}\,(2k-1)!!/(2k-2)!!, & d = 2k,\ k \geq 1, \\ \sqrt{2/\pi}\,(2k)!!/(2k-1)!!, & d = 2k+1,\ k \geq 0, \end{cases} \tag{15}$$

with $(-1)!! = 0!! = 1$. For any $\delta \in (0,1)$, $\mathbb{P}\Big(\|Z\|_2 \geq \max\{\mu_d - \sqrt{2\log\delta^{-1}}, 0\}\Big) \geq 1 - \delta$.

A.2 Proof of Theorem 4.3 (stationarity bounds, fixed step size)

Throughout this proof we assume that clipping is inactive on the trajectory, so $\Pi_a(g_t) = g_t = \nabla L(w_t)$ at every iteration. The convergence theorem is scoped to this regime; clipping remains enabled in the empirical evaluation.

TIDE bound (4). **Step 1 (one-step descent on the z-trajectory).** Under clipping-inactivity the TIDE update is $w_{t+1} = w_t - \eta(\nabla L(w_t) + \xi_t + e_{t-1})$, and Lemma A.3 gives the auxiliary recursion $z_{t+1} = z_t - \eta(\nabla L(z_t) + b_t)$ with bias $b_t = \nabla L(w_t) - \nabla L(z_t)$ satisfying $\|b_t\| \leq \frac{l\eta}{1-\beta}\|e_{t-1}\|$. Apply the descent inequality (12) of Lemma A.1 with $u = z_t$ and $v = z_{t+1}$:

$$\begin{aligned} L(z_{t+1}) &\leq L(z_t) + \langle \nabla L(z_t),\, z_{t+1} - z_t \rangle + \tfrac{l}{2}\|z_{t+1} - z_t\|^2 \\ &= L(z_t) - \eta\langle \nabla L(z_t),\, \nabla L(z_t) + b_t \rangle + \tfrac{l\eta^2}{2}\|\nabla L(z_t) + b_t\|^2. \end{aligned}$$

Expand the inner product and apply two elementary inequalities. First, by Young's inequality $|\langle u, v\rangle| \leq \frac{1}{2\alpha}\|u\|^2 + \frac{\alpha}{2}\|v\|^2$ with $\alpha = 2$,

$$-\eta\,\langle \nabla L(z_t), b_t \rangle \;\leq\; \eta\,|\langle \nabla L(z_t), b_t \rangle| \;\leq\; \eta\big(\tfrac{1}{4}\|\nabla L(z_t)\|^2 + \|b_t\|^2\big)\,.$$

Second, by the triangle inequality $\|u + v\|^2 \leq 2\|u\|^2 + 2\|v\|^2$,

$$\|\nabla L(z_t) + b_t\|^2 \;\leq\; 2\|\nabla L(z_t)\|^2 + 2\|b_t\|^2.$$

Substituting both and writing $-\eta\langle \nabla L(z_t), \nabla L(z_t) + b_t\rangle = -\eta\|\nabla L(z_t)\|^2 - \eta\langle \nabla L(z_t), b_t\rangle$,

$$\begin{aligned} L(z_{t+1}) &\leq L(z_t) - \eta\|\nabla L(z_t)\|^2 + \eta\big(\tfrac{1}{4}\|\nabla L(z_t)\|^2 + \|b_t\|^2\big) + \tfrac{l\eta^2}{2}\big(2\|\nabla L(z_t)\|^2 + 2\|b_t\|^2\big) \\ &= L(z_t) - \eta\big(\tfrac{3}{4} - l\eta\big)\,\|\nabla L(z_t)\|^2 + \big(\eta + l\eta^2\big)\|b_t\|^2. \end{aligned}$$

Step 2 (use the step-size restriction $\eta \leq 1/(4l)$). The condition $\eta \leq 1/(4l)$ gives $l\eta \leq 1/4$, hence

$$\tfrac{3}{4} - l\eta \;\geq\; \tfrac{3}{4} - \tfrac{1}{4} = \tfrac{1}{2}, \qquad \eta + l\eta^2 \;=\; \eta(1 + l\eta) \;\leq\; \tfrac{5}{4}\eta \;\leq\; 2\eta.$$

Therefore

$$L(z_{t+1}) \;\leq\; L(z_t) - \tfrac{\eta}{2}\|\nabla L(z_t)\|^2 + 2\eta\,\|b_t\|^2. \tag{16}$$

Step 3 (take expectation and bound the bias term). Take total expectation of (16). The bias bound from Lemma A.3 gives $\|b_t\|^2 \leq \frac{l^2\eta^2}{(1-\beta)^2}\|e_{t-1}\|^2$, and Lemma A.2 gives $\mathbb{E}\|e_{t-1}\|^2 = d\sigma_e^2 = \frac{1-\beta}{1+\beta}a^2 d\sigma^2$ for every $t \geq 0$ (here we use the stationary warm-start). Hence

$$2\eta\,\mathbb{E}\|b_t\|^2 \;\leq\; 2\eta \cdot \frac{l^2\eta^2}{(1-\beta)^2} \cdot \frac{1-\beta}{1+\beta}a^2 d\sigma^2 \;=\; \frac{2l^2\eta^3\,a^2 d\sigma^2}{(1-\beta)(1+\beta)} \;=\; \frac{2l^2\eta^3\,a^2 d\sigma^2}{1-\beta^2}.$$

This yields the per-step expected descent

$$\mathbb{E}L(z_{t+1}) \;\leq\; \mathbb{E}L(z_t) - \tfrac{\eta}{2}\,\mathbb{E}\|\nabla L(z_t)\|^2 + \frac{2l^2\eta^3 a^2 d\sigma^2}{1-\beta^2}. \tag{17}$$

Step 4 (initial-iterate offset $\mathbb{E}L(z_0) \leq \ldots$). Since $z_0 = w_0 - \frac{\eta}{1-\beta}e_{-1}$ with w_0 deterministic, applying (12) at $(u, v) = (w_0, z_0)$ gives

$$\begin{aligned} L(z_0) \;&\leq\; L(w_0) + \langle \nabla L(w_0),\, z_0 - w_0 \rangle + \tfrac{l}{2}\|z_0 - w_0\|^2 \\ &=\; L(w_0) - \frac{\eta}{1-\beta}\langle \nabla L(w_0), e_{-1}\rangle + \frac{l\eta^2}{2(1-\beta)^2}\|e_{-1}\|^2. \end{aligned}$$

Taking expectations: $\mathbb{E}\langle \nabla L(w_0), e_{-1}\rangle = 0$ since $\nabla L(w_0)$ is deterministic and $\mathbb{E}[e_{-1}] = 0$, and $\mathbb{E}\|e_{-1}\|^2 = d\sigma_e^2$ by Lemma A.2. Therefore

$$\mathbb{E}L(z_0) \;\le\; L(w_0) + \frac{l\eta^2}{2(1-\beta)^2} \cdot \frac{1-\beta}{1+\beta} a^2 d\sigma^2 \;=\; L(w_0) + \frac{l\eta^2 a^2 d\sigma^2}{2(1-\beta^2)}. \tag{18}$$

Step 5 (telescope and rearrange). Sum (17) from $t = 0$ to $T-1$. The left-hand side telescopes to $\mathbb{E}L(z_T) - \mathbb{E}L(z_0)$. Rearranging,

$$\tfrac{\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla L(z_t)\|^2 \;\le\; \mathbb{E}L(z_0) - \mathbb{E}L(z_T) + T \cdot \frac{2l^2\eta^3 a^2 d\sigma^2}{1-\beta^2}.$$

Using $\mathbb{E}L(z_T) \ge L^*$ (Assumption 4.1) and (18),

$$\tfrac{\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla L(z_t)\|^2 \;\le\; \big(L(w_0) - L^*\big) + \frac{l\eta^2 a^2 d\sigma^2}{2(1-\beta^2)} + T \cdot \frac{2l^2\eta^3 a^2 d\sigma^2}{1-\beta^2}.$$

Dividing both sides by $\eta T/2$,

$$\bar{G}_T^{\mathrm{TIDE}} \;:=\; \tfrac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla L(z_t)\|^2 \;\le\; \frac{2(L(w_0) - L^*)}{\eta T} + \frac{l\eta a^2 d\sigma^2}{T(1-\beta^2)} + \frac{4l^2\eta^2 a^2 d\sigma^2}{1-\beta^2}. \tag{19}$$

Step 6 (smoothness transfer to the original iterates). Since $z_t - w_t = -\frac{\eta}{1-\beta} e_{t-1}$, by Assumption 4.1 and $\|u+v\|^2 \le 2\|u\|^2 + 2\|v\|^2$,

$$\|\nabla L(w_t)\|^2 \;\le\; 2\|\nabla L(z_t)\|^2 + 2\|\nabla L(w_t) - \nabla L(z_t)\|^2 \;\le\; 2\|\nabla L(z_t)\|^2 + 2l^2\|w_t - z_t\|^2.$$

Taking expectation and using $\mathbb{E}\|e_{t-1}\|^2 = \frac{1-\beta}{1+\beta} a^2 d\sigma^2$,

$$\mathbb{E}\|\nabla L(w_t)\|^2 \;\le\; 2\,\mathbb{E}\|\nabla L(z_t)\|^2 + \frac{2l^2\eta^2}{1-\beta^2}\, a^2 d\sigma^2.$$

Averaging over t and combining with (19),

$$\tfrac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla L(w_t)\|^2 \le \frac{4(L(w_0) - L^*)}{\eta T} + \frac{2l\eta a^2 d\sigma^2}{T(1-\beta^2)} + \frac{10l^2\eta^2 a^2 d\sigma^2}{1-\beta^2}.$$

This is (4). If $T \ge 1/(4l\eta)$ is additionally imposed, the warm term can be absorbed, yielding the alternative two-term bound

$$\bar{G}_T^{\mathrm{TIDE}} \le \frac{4(L(w_0) - L^*)}{\eta T} + \frac{18l^2\eta^2 a^2 d\sigma^2}{1-\beta^2}.$$

DPSGD bound (5). **Step 1 (one-step expected descent).** For DPSGD the update is $w_{t+1} = w_t - \eta(\nabla L(w_t) + \xi_t)$ with $\xi_t \sim \mathcal{N}(0, a^2\sigma^2 I_d)$ independent of w_t. Apply (12) with $u = w_t$, $v = w_{t+1}$:

$$L(w_{t+1}) \le L(w_t) - \eta\langle \nabla L(w_t),\, \nabla L(w_t) + \xi_t\rangle + \tfrac{l\eta^2}{2}\|\nabla L(w_t) + \xi_t\|^2.$$

Take expectation conditional on w_t: by independence and $\mathbb{E}[\xi_t \mid w_t] = 0$,

$$\mathbb{E}\langle \nabla L(w_t), \xi_t \mid w_t\rangle = 0, \qquad \mathbb{E}\|\nabla L(w_t) + \xi_t\|^2 \,\big|\, w_t = \|\nabla L(w_t)\|^2 + \mathbb{E}\|\xi_t\|^2 = \|\nabla L(w_t)\|^2 + a^2\sigma^2 d.$$

Substituting,

$$\mathbb{E}\big[L(w_{t+1}) \,\big|\, w_t\big] \;\le\; L(w_t) - \eta\big(1 - \tfrac{l\eta}{2}\big)\|\nabla L(w_t)\|^2 + \tfrac{l\eta^2}{2}\, a^2\sigma^2 d.$$

Step 2 (use $\eta \le 1/(4l) \le 1/l$). Then $1 - l\eta/2 \ge 1/2$, so taking total expectation,

$$\mathbb{E}L(w_{t+1}) \;\le\; \mathbb{E}L(w_t) - \tfrac{\eta}{2}\,\mathbb{E}\|\nabla L(w_t)\|^2 + \tfrac{l\eta^2}{2}\, a^2\sigma^2 d. \tag{20}$$

Step 3 (telescope). Sum (20) from $t = 0$ to $T-1$ and use $\mathbb{E}L(w_t) \ge L^*$, $\mathbb{E}L(w_0) = L(w_0)$:

$$\tfrac{\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla L(w_t)\|^2 \;\le\; L(w_0) - L^* + T \cdot \tfrac{l\eta^2}{2}\, a^2\sigma^2 d.$$

Dividing by $\eta T/2$ yields exactly (5):

$$\bar{G}_T^{\mathrm{DPSGD}} \;\le\; \frac{2(L(w_0) - L^*)}{\eta T} + l\eta\, a^2 d\sigma^2. \qquad \square$$

A.3 Proof of Theorem 4.4 (convergence rates with tuned step size)

We evaluate the stated horizon-dependent choices under $\eta \in (0, 1/(4l)]$. Throughout, $\Delta := L(w_0) - L^* \geq 0$.

DPSGD. The bound (5) is $f_{\mathrm{DP}}(\eta) := 2\Delta/(\eta T) + l\eta a^2 d\sigma^2$. Setting $f'_{\mathrm{DP}}(\eta) = 0$ gives $\eta^\star = \sqrt{2\Delta/(la^2 d\sigma^2 T)}$, so

$$\bar{G}_T^{\mathrm{DPSGD}} \leq 2\sqrt{\tfrac{2l\Delta a^2 d\sigma^2}{T}} = O(T^{-1/2}).$$

The constraint $\eta^\star \leq 1/(4l)$ holds for all $T \geq 32l\Delta/(a^2 d\sigma^2)$.

TIDE. The complete bound contains the warm term $2l\eta a^2 d\sigma^2/[T(1-\beta^2)]$. Balancing the leading $4\Delta/(\eta T) + 10l^2\eta^2 a^2 d\sigma^2/(1-\beta^2)$ terms gives

$$\eta^\star = \Big(\tfrac{\Delta(1-\beta^2)}{5l^2 a^2 d\sigma^2 T}\Big)^{1/3}.$$

The leading two terms attain their AM–GM minimum, while the warm contribution is retained:

$$\bar{G}_T^{\mathrm{TIDE}} \leq 3\Big(\tfrac{40l^2\Delta^2 a^2 d\sigma^2}{(1-\beta^2)T^2}\Big)^{1/3} + \frac{2la^2 d\sigma^2}{T(1-\beta^2)}\Big(\tfrac{\Delta(1-\beta^2)}{5l^2 a^2 d\sigma^2 T}\Big)^{1/3}.$$

The second term is $O(T^{-4/3})$, so the total is $O(T^{-2/3})$. The constraint $\eta^\star \leq 1/(4l)$ holds for all $T \geq 64l\Delta(1-\beta^2)/(5a^2 d\sigma^2)$.

Comparison. The DPSGD rate $T^{-1/2}$ matches the i.i.d. stochastic-oracle lower bound of Arjevani et al. (2023), while TIDE's $T^{-2/3}$ rate is strictly faster (the gap arises because TIDE's per-round noise is correlated, falling outside the i.i.d. oracle model; see Remark A.5 below). □

Remark A.5 (Tightness and the role of correlated noise)**.** *The information-theoretic lower bound for first-order stochastic optimization with i.i.d. unbiased noise of variance Σ^2 on a smooth nonconvex objective is $\Omega(l\Delta\Sigma^2/T)$ on the time-averaged squared gradient norm $\bar{G}_T$ Arjevani et al. (2023), matched by the DPSGD bound* (5) *(with $\Sigma^2 = a^2\sigma^2 d$ and the calibration $\eta \sim 1/\sqrt{T}$). TIDE's $T^{-2/3}$ rate is* not *in conflict with this lower bound because TIDE's per-step noise is* correlated *across rounds, placing it outside the i.i.d. stochastic-oracle model of Arjevani et al. (2023). The improvement is therefore an explicit consequence of structured noise correlation, not of better usage of i.i.d. noise—placing TIDE in the same algorithmic family as DP-FTRL Kairouz et al. (2021) and matrix-factorization DP Denisov et al. (2022); Choquette-Choo et al. (2022), but with a much simpler online recursion.*

A.4 Proof of Theorem 4.6 (ℓ_1 lower bound for TIDE)

This proof uses the clipping-inactive condition $\Pi_a(g_t) = g_t$. Under active clipping the deterministic residual can cancel part of the random perturbation; a valid extension must subtract a residual-norm bound and take the positive part.

Step 1 (reduction to perturbation norm). By Theorem 3.1 (proved in Section 3),

$$\ell_1 \;\geq\; \frac{1}{l_g}\min_{x\in D}\|C(g_t, x) - g_t\|_2.$$

Under the clipping-inactive regime, $\Pi_a(g_t) = g_t$, so the TIDE-induced perturbation is

$$C(g_t, x) - g_t \;=\; \big(\Pi_a(g_t) - g_t\big) + \xi_t + e_{t-1} \;=\; \xi_t + e_{t-1},$$

which does not depend on the input x. Therefore $\min_{x\in D}\|C(g_t, x) - g_t\|_2 = \|\xi_t + e_{t-1}\|_2$.

Step 2 (joint distribution of $\xi_t + e_{t-1}$). By Lemma A.2, $e_{t-1} \sim \mathcal{N}(0, \sigma_e^2 I_d)$ with $\sigma_e^2 = \frac{1-\beta}{1+\beta}a^2\sigma^2$, and $\xi_t \perp e_{t-1}$. Hence $\xi_t + e_{t-1}$ is the sum of two independent centered Gaussians and is itself centered Gaussian with per-coordinate variance

$$\tau^2 \;:=\; \mathrm{Var}(\xi_t) + \mathrm{Var}(e_{t-1}) \;=\; a^2\sigma^2 + \sigma_e^2 \;=\; a^2\sigma^2\left(1 + \frac{1-\beta}{1+\beta}\right) \;=\; \frac{2}{1+\beta}\,a^2\sigma^2.$$

Write $\xi_t + e_{t-1} = \tau\, Z$ with $Z \sim \mathcal{N}(0, I_d)$.

Step 3 (Gaussian-norm concentration). By Lemma A.4, with probability at least $1-\delta$,

$$\|\xi_t + e_{t-1}\|_2 \geq \tau \max\{\mu_d - \sqrt{2\log\delta^{-1}}, 0\}.$$

Combining with Step 1,

$$\ell_{1,\text{TIDE}} \geq \frac{\tau}{l_g} \max\{\mu_d - \sqrt{2\log\delta^{-1}}, 0\}, \qquad \tau^2 = \frac{2}{1+\beta} a^2\sigma^2. \qquad \square$$

A.5 Proof of Theorem 4.7 (ℓ_1 lower bound for DPSGD)

As in the TIDE case, this proof operates under the clipping-inactive regime explicitly stated in the hypothesis of Theorem 4.7.

Step 1 (reduction). As in the TIDE case, Theorem 3.1 gives $\ell_1 \geq \frac{1}{l_g} \min_{x\in D} \|C(g_t, x) - g_t\|_2$. For DPSGD under clipping-inactivity the perturbation reduces to $C(g_t, x) - g_t = \xi_t \sim \mathcal{N}(0, a^2\sigma^2 I_d)$, independent of x, so $\min_x \|C(g_t, x) - g_t\|_2 = \|\xi_t\|_2$.

Step 2 (Gaussian-norm concentration). Write $\xi_t = a\sigma Z$. By Lemma A.4, with probability at least $1-\delta$,

$$\ell_{1,\text{DPSGD}} \geq \frac{a\sigma}{l_g} \max\{\mu_d - \sqrt{2\log\delta^{-1}}, 0\}. \qquad \square$$

A.6 Proof of Proposition 4.8 (Conditional Matrix-Gaussian Sensitivity)

Proposition A.6 (Formal Restatement of Proposition 4.8: Conditional Sensitivity Analysis)**.** *Consider TIDE (Algorithm 1) with ℓ_2-clip threshold $a > 0$, per-sample noise multiplier $\sigma_p > 0$, smoothing parameter $\beta \in [0, 1)$, and stationary warm-start $e_{-1} \sim \mathcal{N}(0, \sigma_e^2 I_d)$ drawn independently of all gradient and noise variables, with $\sigma_e^2 := \frac{1-\beta}{1+\beta} \cdot \frac{a^2\sigma_p^2}{B}$. Across T rounds with Poisson subsampling rate q:*

(i) *(*Conditional Matrix Gaussian Representation.*) Conditional on the independent warm-start e_{-1}, its contribution is a data-independent shift and the temporally shaped Gaussian component of the normalized transcript is $\sigma_p M\xi$. Here $M[t,t] = 1$ and $M[t,s] = -(1-\beta)\beta^{t-1-s}$ for $s < t$. The inverse M^{-1} has 1s on the diagonal and $(1-\beta)$ everywhere below.*

(ii) *(*Conditional divergence and expected energy ratio.*) For a fixed normalized participation vector $s \in \{0,1\}^T$, the corresponding Gaussian-shift divergence at order $\alpha > 1$ is $\alpha\|M^{-1}s\|_2^2/(2\sigma_p^2)$. If $S_j \sim \text{Bern}(q)$ independently with $q > 0$, the expected conditional sensitivity-energy ratio relative to DPSGD is*

$$r(T,q) \;:=\; \frac{\mathbb{E}[\|M^{-1}\mathbf{1}_S\|^2]}{\mathbb{E}[\|\mathbf{1}_S\|^2]} \;=\; 1 \;+\; A_{\beta,q}\,(T-1) \;+\; B_{\beta,q}\,(T-1)^2,$$

where $A_{\beta,q} := \frac{(1-q)(1-\beta)^2}{2} + q(1-\beta) + \frac{q(1-\beta)^2}{6}$ and $B_{\beta,q} := \frac{q(1-\beta)^2}{3}$. For $q < 1$, this expectation is the conditional energy ratio; Rényi accounting for the latent Poisson-subsampled mixture uses its log moment. At $q = 1$, it is the exact deterministic divergence ratio and also gives conservative accounting without subsampling amplification.

Proof. (i) Unrolling the EMA error-correction recursion $e_t = \beta e_{t-1} - (1-\beta)\bar{\xi}_t$ for $t = 0, 1, \ldots, T-1$ gives

$$e_{t-1} = -(1-\beta)\sum_{s=0}^{t-1} \beta^{t-1-s}\bar{\xi}_s + \beta^t e_{-1}.$$

Under the stationary warm-start $e_{-1} \sim \mathcal{N}\big(0, \frac{1-\beta}{1+\beta}a^2\sigma^2 I_d\big)$ (where $\sigma = \sigma_p/\sqrt{B}$), the total noise added to the released gradient g_t at step t is $\xi_t^{(i)} + e_{t-1}$. Conditional on e_{-1}, the terms $\beta^t e_{-1}$ form a data-independent shift. After sensitivity normalization, the remaining Gaussian component is

shaped by the lower-triangular matrix $M \in \mathbb{R}^{T\times T}$ with diagonal entries $M[t,t] = 1$ and strictly lower-triangular entries $M[t,s] = -(1-\beta)\beta^{t-1-s}$ for $s < t$.

To determine the inverse matrix M^{-1}, consider the linear system $Mv = w$, which corresponds to the lower-triangular equations:

$$v_t - (1-\beta)\sum_{s=0}^{t-1}\beta^{t-1-s}v_s = w_t, \qquad \text{for } t = 0, 1, \ldots, T-1.$$

We solve for v_t explicitly by taking the first difference of adjacent equations:

$$\begin{aligned} w_t - \beta w_{t-1} &= \Big(v_t - (1-\beta)\sum_{s=0}^{t-1}\beta^{t-1-s}v_s\Big) - \beta\Big(v_{t-1} - (1-\beta)\sum_{s=0}^{t-2}\beta^{t-2-s}v_s\Big) \\ &= v_t - \beta v_{t-1} - (1-\beta)v_{t-1} = v_t - v_{t-1}. \end{aligned}$$

Rearranging gives $v_t = v_{t-1} + w_t - \beta w_{t-1}$. Summing this telescoping relation over time yields

$$v_t = w_t + (1-\beta)\sum_{s=0}^{t-1} w_s.$$

Thus, the operator M^{-1} maps $w \mapsto v$ via $v_t = \sum_{s=0}^{t} M^{-1}[t,s]w_s$, proving that M^{-1} is a lower-triangular matrix with 1s on the main diagonal ($M^{-1}[t,t] = 1$) and constant value $(1-\beta)$ on all strictly sub-diagonal positions ($M^{-1}[t,s] = 1-\beta$ for $s < t$).

(ii) Under Poisson sample-level subsampling at rate q, two neighboring datasets D, D' differing in a single sample x produce a Bernoulli presence vector $S = (S_0, S_1, \ldots, S_{T-1})^\top \in \{0,1\}^T$ with $S_j \overset{\text{iid}}{\sim} \mathrm{Bern}(q)$. The effective adjacency sensitivity vector across the T rounds is $\mathbf{1}_S$.

Conditional on a fixed realization $S = s$, the Matrix Gaussian formula Denisov et al. (2022); Choquette-Choo et al. (2022) gives the Gaussian-shift divergence $\alpha\|M^{-1}s\|_2^2/(2\sigma_p^2)$. We compare the expected conditional sensitivity energies of TIDE and DPSGD (whose inverse noise-mixing matrix is I_T):

$$r(T,q) \;:=\; \frac{\mathbb{E}_S\big[\|M^{-1}\mathbf{1}_S\|_2^2\big]}{\mathbb{E}_S\big[\|\mathbf{1}_S\|_2^2\big]}.$$

For a fixed realization of S, the i-th component of the vector $Y = M^{-1}\mathbf{1}_S$ is $Y_i = S_i + (1-\beta)\sum_{j=0}^{i-1} S_j$. The total squared norm is:

$$\|M^{-1}\mathbf{1}_S\|_2^2 = \sum_{i=0}^{T-1} Y_i^2 = \sum_{i=0}^{T-1}\left(S_i + (1-\beta)\sum_{j=0}^{i-1} S_j\right)^2.$$

Expanding each term Y_i^2 yields:

$$Y_i^2 = S_i^2 + 2(1-\beta)S_i\sum_{j=0}^{i-1} S_j + (1-\beta)^2\left(\sum_{j=0}^{i-1} S_j\right)^2.$$

Since $S_j \overset{\text{iid}}{\sim} \mathrm{Bern}(q)$, we have $\mathbb{E}[S_j^2] = \mathbb{E}[S_j] = q$, and for $j \neq k$, $\mathbb{E}[S_jS_k] = q^2$. Thus,

$$\mathbb{E}\left[\sum_{j=0}^{i-1} S_j\right] = iq, \qquad \text{and} \qquad \mathbb{E}\left[\left(\sum_{j=0}^{i-1} S_j\right)^2\right] = iq + i(i-1)q^2.$$

Taking expectations for Y_i^2 gives:

$$\mathbb{E}[Y_i^2] = q + 2(1-\beta)iq^2 + (1-\beta)^2\big(iq + i(i-1)q^2\big).$$

Summing $\mathbb{E}[Y_i^2]$ over $i = 0, 1, \ldots, T-1$ using the summation identities $\sum_{i=0}^{T-1} 1 = T$, $\sum_{i=0}^{T-1} i = \frac{T(T-1)}{2}$, and $\sum_{i=0}^{T-1} i(i-1) = \frac{T(T-1)(T-2)}{3}$:

$$\mathbb{E}_S\big[\|M^{-1}\mathbf{1}_S\|_2^2\big] = qT + 2(1-\beta)q^2\frac{T(T-1)}{2} + (1-\beta)^2 q\frac{T(T-1)}{2} + (1-\beta)^2q^2\frac{T(T-1)(T-2)}{3}.$$

Dividing by $\mathbb{E}_S[\|\mathbf{1}_S\|_2^2] = qT$ yields the exact expected sensitivity ratio:

$$r(T,q) = 1 + (1-\beta)q(T-1) + \frac{(1-\beta)^2}{2}(T-1) + \frac{(1-\beta)^2 q}{3}(T-1)(T-2).$$

Rearranging terms by powers of $(T-1)$ using $(T-2) = (T-1) - 1$:

$$r(T,q) = 1 + \left[\frac{(1-q)(1-\beta)^2}{2} + q(1-\beta) + \frac{q(1-\beta)^2}{6}\right](T-1) + \frac{q(1-\beta)^2}{3}(T-1)^2,$$

which proves the exact quadratic representation $r(T,q) = 1 + A_{\beta,q}(T-1) + B_{\beta,q}(T-1)^2$ with $A_{\beta,q} = \frac{(1-q)(1-\beta)^2}{2} + q(1-\beta) + \frac{q(1-\beta)^2}{6}$ and $B_{\beta,q} = \frac{q(1-\beta)^2}{3}$. This expression is the expected conditional quadratic-energy ratio. Full participation gives $S = \mathbf{1}$ deterministically, so $r(T,1)$ is the exact divergence ratio under the normalized matrix model; since M^{-1} has nonnegative entries, this vector also maximizes the energy over $s \in \{0,1\}^T$. Accounting for latent Poisson subsampling requires the corresponding mixture log moment. □

Sensitivity-Energy Ratio $r(T,q)$. Table 2 reports deterministic divergence ratios at $q = 1$ and expected conditional energy diagnostics at $q = 0.026$.

Table 2: Conditional sensitivity-energy ratio $r(T,q)$ at $\beta = 0.5$. The $q = 1$ rows report deterministic divergence ratios, while the $q = 0.026$ rows report expected conditional-energy diagnostics.

Regime	T	Adjacency	Matched-σ ratio	Matched-utility ratio (TIDE at 10σ)
Attack observation	10	Full participation ($q = 1$)	$12.6\times$	0.126 ($8\times$ better)
Attack observation	10	Poisson sample ($q = 0.026$)	$\approx 2.40\times$	≈ 0.024 ($42\times$ better)
Training-time DP	200	Full participation ($q = 1$)	$3409\times$	$34.1\times$
Training-time DP	200	Poisson sample ($q = 0.026$)	$114\times$	$1.14\times$ (matched)

For nondegenerate q and β, $r(T,q) = 1$ only at $T = 1$. Solving $r(T,q) = 100$ gives a conditional-energy crossover for the $10\sigma_p$ comparison; formal privacy-budget crossover analysis for $q < 1$ uses the mixture log moment.

B Detailed Experiment Settings

Our experimental setup follows standard practices in prior work on gradient inversion Zhu et al. (2019); Zhang et al. (2023). We use ResNet-18 as the backbone model and replace ReLU with sigmoid activations, which have been shown to make models more susceptible to gradient inversion attacks Zhu et al. (2019). Unless otherwise stated, all experiments are conducted in a federated setting with three simulated clients.

We compare TIDE with four representative baseline defenses. The first is DP Abadi et al. (2016), which protects gradients by adding random noise during training. We also include Outpost **?**, DGP **?**, and GradDropout Sotthiwat et al. (2026), which represent recent gradient-level defense methods. Together, these baselines cover both noise-based and gradient-modification-based approaches.

To evaluate privacy protection, we test each defense against five representative gradient inversion attacks: GI Zhu et al. (2019), IG Geiping et al. (2020), iDLG Zhao et al. (2020), CI-Net Zhang et al. (2023), and DGGI Wu et al. (2025). These attacks include both optimization-based and generative reconstruction methods, providing a broad assessment of defense robustness. We report model performance using test accuracy. Privacy leakage is quantified using SSIM Wang et al. (2004) and LPIPS Zhang et al. (2018); lower SSIM and higher LPIPS indicate weaker reconstruction and therefore stronger privacy protection.

Models are trained using SGD with momentum 0.9. For all datasets and defense methods, we use a local batch size of 128 and an initial learning rate of 0.3. The learning rate is decayed to zero following a cosine schedule Loshchilov & Hutter (2016). We keep the optimizer and learning-rate schedule fixed across all methods to ensure that performance differences mainly reflect the effect of the defense mechanism rather than changes in the training protocol. All experiments are run on NVIDIA A100 GPUs with 40GB memory.

C EXPERIMENTS ON ADDITIONAL DATASETS

Our main findings that TIDE preserves model performance while improving privacy protection are not limited to CIFAR-10. In this section, we provide additional experiments on two more challenging datasets: CIFAR-100 and Tiny-ImageNet.

C.1 TINY-IMAGENET RESULTS

We further evaluate the defense methods on Tiny-ImageNet, a more challenging dataset with higher image resolution and more classes. This setting provides a stronger test of whether a defense can scale beyond CIFAR-style benchmarks while maintaining both privacy and utility.

Table 3: Performance on Tiny-ImageNet: Best test accuracy and average SSIM under five gradient-inversion attacks across various defense methods. The best performance is in bold.

Defense	Test Acc (%) ↑	Attack SSIM ↓				
		GI	IG	iDLG	CI-Net	DGGI
None	$52.84_{\pm 0.44}$	$0.2528_{\pm 0.0289}$	$0.9377_{\pm 0.0376}$	$0.9358_{\pm 0.0360}$	$0.8931_{\pm 0.0310}$	$0.8189_{\pm 0.0606}$
Outpost	$47.63_{\pm 0.69}$	$0.1109_{\pm 0.0187}$	$0.4465_{\pm 0.0344}$	$0.3713_{\pm 0.0338}$	$0.4675_{\pm 0.1205}$	$0.5568_{\pm 0.0554}$
DGP	$52.33_{\pm 0.49}$	$0.3267_{\pm 0.0323}$	$0.9235_{\pm 0.0427}$	$0.9210_{\pm 0.0429}$	$0.8686_{\pm 0.0471}$	$0.7862_{\pm 0.0622}$
GradDropout (p=0.3)	$50.47_{\pm 0.17}$	$0.0287_{\pm 0.0085}$	$0.0345_{\pm 0.0064}$	$0.0319_{\pm 0.0040}$	$0.0423_{\pm 0.0144}$	$0.0488_{\pm 0.0079}$
DP-SGD (no-clip)	$49.97_{\pm 0.38}$	$0.0298_{\pm 0.0090}$	$0.0511_{\pm 0.0171}$	$0.0506_{\pm 0.0160}$	$0.0662_{\pm 0.0233}$	$0.0574_{\pm 0.0163}$
DP-SGD	$49.65_{\pm 0.56}$	$\mathbf{0.0165}_{\pm 0.0018}$	$0.0125_{\pm 0.0017}$	$0.0126_{\pm 0.0017}$	$0.0394_{\pm 0.0146}$	$0.0114_{\pm 0.0024}$
DP-Auto-S	$50.11_{\pm 0.36}$	$0.0181_{\pm 0.0030}$	$0.0135_{\pm 0.0027}$	$0.0123_{\pm 0.0015}$	$0.0363_{\pm 0.0135}$	$0.0121_{\pm 0.0026}$
DP-Step-Decay	$52.52_{\pm 0.25}$	$\mathbf{0.0165}_{\pm 0.0018}$	$0.0125_{\pm 0.0017}$	$0.0126_{\pm 0.0017}$	$0.0394_{\pm 0.0146}$	$0.0114_{\pm 0.0024}$
TIDE (no-clip)	$52.61_{\pm 0.59}$	$0.0291_{\pm 0.0099}$	$0.0419_{\pm 0.0139}$	$0.0399_{\pm 0.0135}$	$0.0635_{\pm 0.0234}$	$0.0436_{\pm 0.0123}$
TIDE	$52.06_{\pm 0.90}$	$0.0174_{\pm 0.0035}$	$0.0114_{\pm 0.0021}$	$0.0108_{\pm 0.0019}$	$0.0370_{\pm 0.0151}$	$0.0120_{\pm 0.0027}$
TIDE (σ=1×10^{-2})	$52.24_{\pm 0.31}$	$0.0180_{\pm 0.0032}$	$\mathbf{0.0097}_{\pm 0.0027}$	$0.0116_{\pm 0.0020}$	$0.0352_{\pm 0.0149}$	$0.0111_{\pm 0.0018}$
TIDE (σ=3×10^{-2})	$\mathbf{53.09}_{\pm 0.67}$	$0.0183_{\pm 0.0031}$	$0.0103_{\pm 0.0022}$	$\mathbf{0.0100}_{\pm 0.0016}$	$\mathbf{0.0311}_{\pm 0.0125}$	$\mathbf{0.0108}_{\pm 0.0020}$

Table 3 reports the test accuracy and defense performance against various attacks. The undefended model achieves competitive accuracy (52.84%), but its gradients remain highly exposed to reconstruction attacks, especially under IG, iDLG, CI-Net, and DGGI. Several baseline defenses reduce this vulnerability, but their behavior differs substantially. Outpost lowers SSIM for some attacks, yet it also incurs clear accuracy losses (47.63%). DGP preserves accuracy (52.33%), but offers limited privacy improvement, with SSIM remaining high under most attacks. GradDropout provides strong protection, achieving low SSIM values across attacks, but still reduces test accuracy (50.47%) relative to the undefended model.

The contrast between DP and TIDE is particularly pronounced on Tiny-ImageNet. Standard DP achieves 49.65% test accuracy under moderate noise ($\sigma = 3 \times 10^{-3}$). Modern DP baselines like DP-Auto-S (50.11%) and DP-Step-Decay (52.52%) improve accuracy, but TIDE consistently matches or outperforms them in both utility and privacy protection.

TIDE shows remarkable stability across noise scales. At $\sigma = 3 \times 10^{-3}$, TIDE achieves 52.06% test accuracy, close to the undefended baseline (52.84%). As the noise scale is increased to $\sigma = 1 \times 10^{-2}$ and $\sigma = 3 \times 10^{-2}$, TIDE retains high test accuracy (52.24% and 53.09%, respectively)—with $\sigma = 3 \times 10^{-2}$ reaching the highest accuracy across all methods (53.09%) while achieving the lowest SSIM values across all five attacks (≤ 0.0311). These results confirm that TIDE's structured noise mechanism scales effectively to larger and more complex datasets like Tiny-ImageNet without sacrificing training stability.

C.2 SUMMARY ACROSS DATASETS

Across CIFAR-10, CIFAR-100, and Tiny-ImageNet, the comparison with DP highlights the key advantage of TIDE. Standard DP provides stronger protection as the noise scale increases, but this improvement comes with a sharp drop in test accuracy, especially on CIFAR-100 and Tiny-ImageNet. In contrast, TIDE remains stable even at the larger noise scale σ=1×10^{-2}. It achieves the best or near-best test accuracy while also attaining the lowest SSIM values under most attacks across the

three datasets. These results show that TIDE supports substantially stronger noise injection without the severe optimization degradation observed in standard DP.

Overall, the empirical results demonstrate that the benefits of TIDE are not dataset-specific. Its structured noise mechanism consistently improves the privacy–utility trade-off across datasets of increasing difficulty, from CIFAR-10 to CIFAR-100 and Tiny-ImageNet. This confirms that TIDE provides robust protection against diverse gradient inversion attacks while maintaining competitive model performance.

D Ablation Studies

We further study the effect of the correction parameter β in TIDE. This parameter controls the balance between e_{t-1} and ξ_t and therefore determines how aggressively past perturbations are compensated during training. This ablation helps assess whether TIDE is sensitive to the choice of β and whether its performance depends on careful tuning.

Table 4: Ablation study on CIFAR-10: test accuracy for TIDE at $\sigma \in \{3\times10^{-3}, 1\times10^{-2}\}$ and $\beta \in \{0, 0.1, 0.5, 0.9\}$, under both clip and no-clip settings. Best per column is bold, second-best is underlined.

	Clip		No-clip	
Method	σ=3×10^{-3}	σ=1×10^{-2}	σ=3×10^{-3}	σ=1×10^{-2}
TIDE (β=0)	$91.06_{\pm0.34}$	$\underline{91.12}_{\pm0.22}$	$\underline{90.52}_{\pm0.09}$	$\mathbf{90.84}_{\pm0.35}$
TIDE (β=0.1)	$91.00_{\pm0.04}$	$91.11_{\pm0.16}$	$90.07_{\pm1.55}$	$89.95_{\pm0.59}$
TIDE (β=0.5)	$\mathbf{91.16}_{\pm0.17}$	$\mathbf{91.33}_{\pm0.10}$	$\mathbf{90.86}_{\pm0.13}$	$\underline{90.79}_{\pm0.58}$
TIDE (β=0.9)	$\underline{91.13}_{\pm0.16}$	$90.46_{\pm0.05}$	$89.78_{\pm1.63}$	$90.01_{\pm0.22}$

Table 4 reports the test accuracy of TIDE on CIFAR-10 under different values of β, considering both clipped and unclipped settings and two noise scales. Overall, TIDE remains stable across a wide range of β values. In the clipped setting, $\beta = 0.5$ achieves the best accuracy for both σ=3×10^{-3} and σ=1×10^{-2}, while $\beta = 0$ and $\beta = 0.9$ also remain competitive. This suggests that error correction can improve optimization when combined with clipping, especially at larger noise scales.

In the unclipped setting, performance also remains relatively stable across different β values. The best accuracy is obtained when $\beta = 0$, while $\beta = 0.5$ gives the second-best result at σ=1×10^{-2} and remains competitive at σ=3×10^{-3}. These results show stable performance across the tested β grid.

Overall, the choice of β embodies a two-sided trade-off. A smaller β increases per-release noise variance ($\propto \frac{2}{1+\beta}\sigma^2$), enhancing single-round gradient inversion resistance (Theorem 4.6). It also increases the off-diagonal entries $(1-\beta)$ of M^{-1} and hence both the conservative full-participation divergence ratio and the expected conditional sensitivity-energy diagnostic in Proposition 4.8. Conversely, a larger $\beta \to 1$ reduces error-correction effectiveness and weakens the convergence bound ($\propto \frac{1}{1-\beta^2}$). Together with the ablation results, these quantities motivate $\beta = 0.5$ as the empirical default.

E Perceptual Quality: LPIPS Evaluation on CIFAR-10

In addition to the commonly-used SSIM metric, we evaluate privacy protection using LPIPS on CIFAR-10 under CI-Net (the strongest attack in our suite). Unlike SSIM, where lower values indicate weaker reconstruction, higher LPIPS values correspond to lower perceptual similarity and therefore stronger privacy protection. This evaluation provides a complementary view of reconstruction quality beyond pixel- or structure-based similarity.

Table 5 reports average LPIPS on CIFAR-10 under CI-Net. Without defense, the reconstructed images remain perceptually close to the original samples, leading to a low LPIPS score (0.1238). Traditional gradient-modification methods like Outpost and DGP increase LPIPS to $0.19 \sim 0.23$, while GradDropout ($p = 0.7$) reaches 0.6192. DP and TIDE both reduce perceptual reconstruction quality effectively; notably, TIDE ($\sigma = 1 \times 10^{-2}$) achieves the highest LPIPS (0.6767), delivering stronger perceptual privacy protection than all baseline defenses.

Table 5: CIFAR-10 LPIPS under CI-Net (higher is better). Core DP-family rows use $\sigma = 3 \times 10^{-3}$ unless otherwise specified. Best performance is in **bold**, second-best is underlined.

Defense Method	CIFAR-10 LPIPS under CI-Net ↑
None	$0.1238_{\pm 0.0575}$
Outpost	$0.2260_{\pm 0.0772}$
DGP	$0.1905_{\pm 0.0855}$
GradDropout ($p = 0.3$)	$0.5438_{\pm 0.0550}$
GradDropout ($p = 0.7$)	$0.6192_{\pm 0.0323}$
DP-SGD (no-clip)	$0.3992_{\pm 0.0486}$
DP-SGD (σ=3×10^{-3})	$0.6295_{\pm 0.0347}$
DP-Auto-S (σ=3×10^{-3})	$0.6269_{\pm 0.0492}$
DP-Step-Decay (σ=3×10^{-3})	$0.6295_{\pm 0.0347}$
TIDE (no-clip)	$0.4258_{\pm 0.0506}$
TIDE ($\beta = 0.5, \sigma$=3×10^{-3})	$0.6387_{\pm 0.0379}$
TIDE ($\beta = 0.5, \sigma$=1×10^{-2})	$\mathbf{0.6767_{\pm 0.0283}}$
TIDE ($\beta = 0.5, \sigma$=3×10^{-2})	$\underline{0.6480_{\pm 0.0321}}$

F Evaluation under Multi-Round and Adaptive Attacks

In addition to single-step gradient inversion attacks, we evaluate TIDE against multi-round and adaptive adversaries that observe a sequence of $T = 10$ consecutive gradient releases from a target client on the **same** local mini-batch. Table 6 reports the reconstruction SSIM across three multi-round attack scenarios:

1. **TGIAs-RO (LBFGS)**: Robust median-based multi-round aggregation (Li et al., 2025) using L-BFGS optimization per round.
2. **TGIAs-RO (IG)**: TGIAs-RO using Inverting Gradients (Geiping et al., 2020) as the per-round solver, with 5 outer refinement rounds and 1000 inner steps.
3. **De-telescope GLS**: A TIDE-aware attack that applies the closed-form inverse temporal transform and weights the multi-round residuals by their exact shared-state covariance.

Derivation of the De-telescope GLS Attack. Consider T consecutive rounds where the target client performs local training on a fixed mini-batch x^* (the scenario in which TIDE's temporal noise correlation is strongest). In round $t \in \{1, \dots, T\}$, let $c_t(x^*) := \nabla_w L(w_t, x^*)$ be the true gradient under global model parameters w_t. The released protected gradient is $y_t = c_t(x^*) + n_t + e_{t-1}$, where $n_t \sim \mathcal{N}(0, \sigma^2 I_d)$ is independent per-round Gaussian noise, and e_{t-1} is TIDE's error correction state. The error state evolves according to:

$$e_t = \beta e_{t-1} + (1-\beta)(-n_t) = e_{t-1} - (1-\beta)(y_t - c_t(x^*)). \tag{21}$$

Unrolling this state recurrence from round 1 with a stationary warm-start $e_0 \sim \mathcal{N}\left(0, \frac{1-\beta}{1+\beta}\sigma^2 I_d\right)$ gives:

$$e_{t-1} = e_0 - (1-\beta)\sum_{k=1}^{t-1}\big(y_k - c_k(x^*)\big). \tag{22}$$

Substituting e_{t-1} back into the release equation $y_t = c_t(x^*) + n_t + e_{t-1}$ yields:

$$y_t - c_t(x^*) = n_t + e_0 - (1-\beta)\sum_{k=1}^{t-1}\big(y_k - c_k(x^*)\big). \tag{23}$$

Rearranging terms to collect known public quantities y_k on the left-hand side and parameter-dependent operators $c_k(x^*)$ on the right-hand side gives:

$$\underbrace{y_t + (1-\beta)\sum_{k=1}^{t-1} y_k}_{\tilde{y}_t} \; - \; \underbrace{\left(c_t(x^*) + (1-\beta)\sum_{k=1}^{t-1} c_k(x^*)\right)}_{\tilde{c}_t(x^*)} \; = \; n_t + e_0. \tag{24}$$

Here, $\tilde{y}_t$ is a known linear combination of public releases up to round t, and $\tilde{c}_t(x)$ is the corresponding transformed forward gradient evaluation. The noise term $n_t + e_0$ consists of independent Gaussian noise $n_t \sim \mathcal{N}(0, \sigma^2 I_d)$ plus the shared initial state $e_0 \sim \mathcal{N}\left(0, \frac{1-\beta}{1+\beta}\sigma^2 I_d\right)$. Each transformed residual has variance $\frac{2}{1+\beta}\sigma^2 I_d$, but the residuals are not independent because they share e_0. Writing $r_t(x) := \tilde{y}_t - \tilde{c}_t(x)$ and $\kappa := (1-\beta)/(1+\beta)$, their temporal covariance is

$$K = \sigma^2(I_T + \kappa \mathbf{1}\mathbf{1}^\top).$$

Hence the exact Gaussian negative log-likelihood, up to constants and the common factor $1/(2\sigma^2)$, is

$$\sum_{t=1}^{T} \|r_t(x)\|_2^2 - \frac{\kappa}{1+\kappa T}\left\|\sum_{t=1}^{T} r_t(x)\right\|_2^2.$$

The covariance-aware quadratic is combined with cosine and total-variation regularization. Writing $Z_y = \sum_t \|\tilde{y}_t\|_2^2$, the implemented objective is

$$\mathcal{L}_{\mathrm{GLS}}(x) = \frac{1}{Z_y}\left[\sum_{t=1}^{T} \|r_t(x)\|_2^2 - \frac{\kappa}{1+\kappa T}\left\|\sum_{t=1}^{T} r_t(x)\right\|_2^2\right] + \lambda \sum_{t=1}^{T}\left(1 - \frac{\langle y_t, c_t(x)\rangle}{\|y_t\|_2 \|c_t(x)\|_2}\right) + \gamma\,\mathrm{TV}(x). \tag{25}$$

The bracketed term is the exact Gaussian GLS quadratic up to a common positive scale. The average transformed residual has per-coordinate variance $\sigma^2(1/T + \kappa)$ and approaches the shared-state floor $\kappa\sigma^2$.

Table 6: Multi-round attack evaluation on CIFAR-10 ($T = 10$, $a = 1$, $\sigma_p = 3 \times 10^{-3}$, 3-client IID). Lower SSIM indicates stronger defense; values below 0.5 represent uninterpretable reconstruction.

Defense	TGIAs-RO (LBFGS) ↓	TGIAs-RO (IG) ↓	De-telescope GLS ↓
None	$0.1603_{\pm 0.0421}$	$0.6440_{\pm 0.3530}$	–
DP-SGD	$0.0054_{\pm 0.0045}$	$0.0697_{\pm 0.0278}$	–
DP-Auto-S	$\mathbf{0.0038_{\pm 0.0045}}$	$0.0693_{\pm 0.0291}$	–
DP-Step-Decay	$0.0069_{\pm 0.0035}$	$0.0676_{\pm 0.0290}$	–
TIDE ($\beta = 0.5$)	$0.0058_{\pm 0.0044}$	$\mathbf{0.0646_{\pm 0.0237}}$	$0.0111_{\pm 0.0053}$

While the undefended baseline suffers substantial leakage under TGIAs-RO (IG) (SSIM $= 0.644$), all defended methods maintain SSIM < 0.07. The covariance-aware De-telescope GLS attack obtains SSIM 0.0111 ± 0.0053 on TIDE. Reconstruction remains far below the meaningful threshold, providing empirical robustness evidence against the evaluated correlation-aware attacks.

G Visualization of Reconstructed Images

We further visualize the reconstructed images to provide a qualitative comparison of different defense methods. Figure 2 compares the original CIFAR-10 image with its reconstructions under different combinations of attacks and defenses. Each row corresponds to a gradient inversion attack, while each column represents a defense method.

The visual results show that strong attacks such as CI-Net and DGGI can still recover recognizable information from the original image under several defense methods. Although defenses such as Outpost, DGP, and GradDropout reduce reconstruction quality to varying degrees, the reconstructed images often retain visible semantic structures, indicating that some private information remains exposed.

In contrast, DP and TIDE largely prevent meaningful reconstruction. Under these defenses, the recovered images appear close to random noise, suggesting that the shared gradients no longer contain sufficient information for the attacker to reconstruct the original sample. However, as shown in the quantitative results, DP achieves this protection at the cost of reduced model performance, especially when the noise scale is large. TIDE provides similar qualitative protection while better preserving test accuracy, further supporting its improved privacy–utility trade-off.

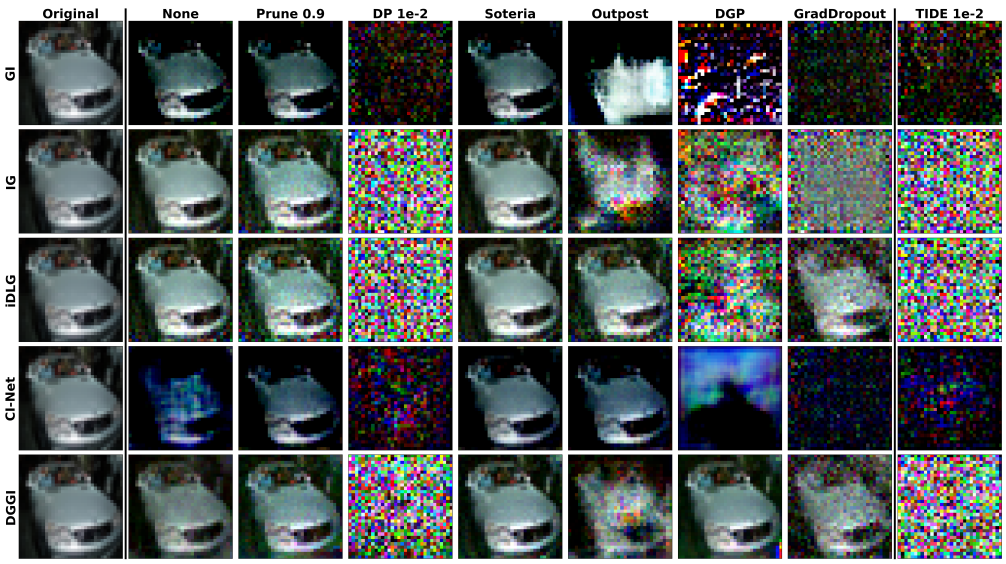

Figure 2: Comparison between the original image and its reconstructions on CIFAR-10 across different combinations of attack and defense methods. Rows denote the attack applied, while columns illustrate the corresponding defense strategies.

H RUNTIME COST

We also compare the runtime cost of different defense methods to assess whether the improved privacy protection of TIDE introduces additional computational overhead. The per-step runtime is measured under the same model, dataset, and training configuration used in the main experiments.

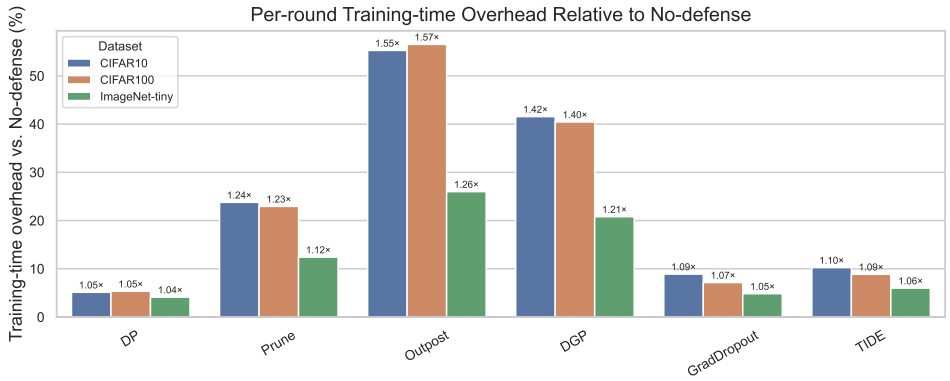

Figure 3: Average running time for one epoch of each defense method.

Figure 3 shows that all methods have similar runtime per training step. In particular, TIDE introduces only lightweight noise generation and correction operations, which add negligible overhead compared with the forward and backward passes of the model. Its runtime is therefore comparable to standard DP and other gradient-level defenses.

These results indicate that the stronger privacy protection achieved by TIDE does not come from increased computational cost. Instead, TIDE improves the privacy–utility trade-off while maintaining a practical training efficiency similar to existing defense methods.

I Cross-Device Federated Learning Experiment

To probe TIDE under a more cross-device-like deployment, we run CIFAR-10 under partial participation in three regimes that differ in client population, data distribution, and sampling rate:

- **Setting A**: 3 simulated clients with 1 client sampled per round ($1/3$ participation, 200 rounds, IID).
- **Setting B**: 10 simulated clients with 1 client sampled per round ($1/10$ participation, 800 rounds, IID).
- **Setting C**: 100 simulated clients with 10 clients sampled per round ($10/100$ participation, 1500 rounds, Dir($\alpha = 0.1$) non-IID balanced partition).

In all regimes, each participating client maintains its own EMA state e_t across its local participation rounds; non-sampled rounds introduce gaps in the local trajectory, mimicking the stateful-but-intermittent regime of cross-device FL. All other settings ($\beta = 0.5$ for TIDE) match Section 5.

Table 7: CIFAR-10 test accuracy under partial-participation and cross-device federated training regimes. Setting A: 3 clients, 1/round ($1/3$), 200 rounds. Setting B: 10 clients, 1/round ($1/10$), 800 rounds. Setting C: 100 clients, 10/round ($10/100$), Dir($\alpha = 0.1$) non-IID, 1500 rounds. Default noise scale is $\sigma = 3 \times 10^{-3}$ unless otherwise specified; TIDE uses $\beta = 0.5$. All results are mean $\pm$ std over three random seeds.

Defense	Setting A (3 clients, $1/3$)	Setting B (10 clients, $1/10$)	Setting C (100 clients, $10/100$, non-IID)
None	$88.97_{\pm 0.23}$	$\mathbf{88.88}_{\pm 0.40}$	$66.70_{\pm 7.60}$
DP-SGD	$82.65_{\pm 0.27}$	$82.45_{\pm 0.74}$	$62.50_{\pm 1.90}$
DP-Auto-S	$87.24_{\pm 0.38}$	$87.29_{\pm 0.01}$	$62.00_{\pm 3.60}$
DP-Step-Decay	$88.99_{\pm 0.16}$	$87.39_{\pm 0.25}$	$64.40_{\pm 2.60}$
TIDE	$\mathbf{89.05}_{\pm 0.44}$	$88.64_{\pm 0.53}$	$\mathbf{75.20}_{\pm 2.40}$

Table 7 shows that the trend from the full-participation main results carries over consistently across all partial-participation regimes. In Setting A ($1/3$ participation, 3 clients), TIDE achieves 89.05% accuracy, matching the undefended baseline (88.97%) and outperforming DP-SGD (82.65%) and DP-Auto-S (87.24%). In Setting B ($1/10$ participation, 10 clients), TIDE reaches 88.64%, surpassing all DP baselines ($82.45\% \sim 87.39\%$). In the large-scale non-IID regime (Setting C, 100 clients, $10/100$ participation), TIDE achieves 75.20%, outperforming DP-SGD (62.50%) by $+12.7\%$, DP-Auto-S (62.00%) by $+13.2\%$, and DP-Step-Decay (64.40%) by $+10.8\%$. Crucially, TIDE even outperforms the undefended baseline (66.70%) by $+8.5\%$ in Setting C, demonstrating that structured error-correction acts as an effective regularizer under heterogeneous data distributions. These results confirm that TIDE's telescoping error-correction structure remains highly effective even when client participation is intermittent and data distribution is non-IID.

J Vision Transformer (ViT-B/16) Full Fine-Tuning Check

To evaluate TIDE's architectural generalization beyond CNNs, we test ViT-B/16 on CIFAR-100 under the same setting as Table 1b with 10 epochs and 1 warmup epoch. The input embedding layer is frozen while all remaining parameters are fully fine-tuned.

Table 8: ViT-B/16 full fine-tuning best validation accuracy (%) on CIFAR-100 (3-client IID, 10 rounds). All entries report mean $\pm$ std across three random seeds.

Defense	$\sigma = 3 \times 10^{-2}$	$\sigma = 1 \times 10^{-1}$	$\sigma = 3 \times 10^{-1}$
None	$90.91_{\pm 0.21}$	–	–
DP-SGD	$86.65_{\pm 0.30}$	$27.45_{\pm 0.45}$	–
TIDE ($\beta = 0.5$)	$89.27_{\pm 0.23}$	$89.21_{\pm 0.22}$	$89.13_{\pm 0.21}$

As shown in Table 8, under large noise ($\sigma = 3 \times 10^{-1}$), **DP-SGD collapses whereas TIDE maintains $\mathbf{89.13}\%$ accuracy**—matching its performance at $\sigma = 3 \times 10^{-2}$ (89.27%) and $\sigma = 1 \times 10^{-1}$ (89.21%),

remaining close to the undefended baseline (90.91%). This confirms that TIDE's noise-correction mechanism generalizes seamlessly to transformer architectures without experiencing noise-induced optimization breakdown.

To further verify reconstruction privacy on ViT-B/16, Table 9 presents the reconstruction metrics under $40,000$-step CI-Net attack. Both DP-SGD and TIDE suppress SSIM below 0.0470 and increase LPIPS above 0.80, effectively destroying visual semantic reconstruction on Vision Transformers.

Table 9: ViT-B/16 CI-Net attack reconstruction metrics ($40,000$ steps, $n = 5$) on CIFAR-100 ($BatchSize = 1$, seed 42).

Defense	SSIM ↓	LPIPS ↑
None	$0.5572_{\pm 0.1533}$	$0.5328_{\pm 0.0727}$
DP-SGD (σ=3×10^{-2})	$0.0447_{\pm 0.0122}$	$\mathbf{0.8273_{\pm 0.0264}}$
TIDE (β=0.5, σ=3×10^{-2})	$\mathbf{0.0470_{\pm 0.0076}}$	$0.8059_{\pm 0.0320}$