RPGen: Robust and Differentially Private Synthetic Image Generation

Zihao Wang[1], Hao Peng[2*], Wei Dong[1], Yuecen Wei[2], Li Sun[3], Zhengtao Yu[4]
[1]Nanyang Technological University
[2]Beihang University
[3]North China Electric Power University
[4]Kunming University of Science and Technology
{zihao.wang, wei_dong}@ntu.edu.sg, {penghao, weiyc}@buaa.edu.cn, ccesunli@ncepu.edu.cn, ztyu@hotmail.com

Abstract

Differentially private (DP) image synthesis enables the generation of realistic images while bounding privacy leakage, facilitating secure data sharing across organizations. However, the Gaussian noise injected during DP training, such as via DP-SGD, often severely degrades synthesis quality by disrupting model convergence. To address this, we introduce RPGen, a novel framework that enhances diffusion models' parameter robustness to mitigate DP noise effects without compromising privacy guarantees. At its core, RPGen employs adversarial model perturbation (AMP) during public pre-training to build resilience against perturbations, but we identify and tackle the critical issue of robustness transferability across domains. RPGen achieves this through a three-step process: (1) A pre-trained classifier infers labels for private images, aggregated into a class distribution noised with Gaussian mechanism for DP, and public samples are selected to match this privatized distribution for domain alignment; (2) The diffusion model is pre-trained on this curated subset with adversarial model perturbation to foster robustness; (3) The model undergoes fine-tuning on private data using DP-SGD. This synergy of robustness augmentation and transferability optimization yields high-fidelity synthesis. Extensive evaluations on ImageNet for pre-training, with CelebA and CIFAR-10 for synthesis, show RPGen outperforming state-of-the-art baselines across $\varepsilon \in \{1, 5, 10\}$. On average, it achieves 20.18% lower FID and 5.45% higher classification accuracy. Ablations confirm the efficacy of domain curation and modest perturbations, establishing RPGen as a new benchmark for privacy-utility trade-offs in image generation.

Introduction

Privacy-preserving synthetic image generation seeks to create synthetic images that capture the essential properties of real data, facilitating secure data sharing within and across organizations while mitigating privacy risks (Hu et al. 2024; Dankar and Emam 2013). Differentially private (DP) image synthesis (Lin et al. 2024; Li et al. 2024) provides rigorous theoretical guarantees to quantify and bound privacy leakage from real data through synthetic outputs. Leveraging DP image synthesis enables organizations, such as those in the medical domain handling sensitive patient scans like X-rays or MRIs, to share and exploit synthetic images for diverse downstream tasks, such as classification and detection, without exposing sensitive information.

Diffusion models have emerged as a promising foundation for DP image synthesis (Dockhorn et al. 2022; Ghalebikesabi et al. 2023). For example, Dockhorn et al. (Dockhorn et al. 2022) advocated training diffusion models with DP-SGD (Abadi et al. 2016), a cornerstone technique for enforcing differential privacy during optimization. Building on this, Li et al. (Li et al. 2024) proposed pre-training on public datasets followed by DP fine-tuning on sensitive data, yielding state-of-the-art (SOTA) utility compared to earlier approaches. Nevertheless, the Gaussian noise inherent to DP-SGD often induces substantial degradation in synthesis quality, as it disrupts model convergence and feature learning. This noise sensitivity remains a critical barrier to achieving high-fidelity DP image generation.

To address this challenge, we propose enhancing the diffusion model's robustness to perturbations during pre-training, thereby preserving performance under DP noise without violating privacy guarantees. Specifically, we augment parameter robustness against adversarial perturbations in the public pre-training phase, enabling the model to better tolerate DP noise during subsequent fine-tuning. While effective in principle, this approach can yield suboptimal results due to poor transferability of robustness from public (upstream) to private (downstream) domains. We mitigate this by curating public data samples that closely mimic the private domain, using them for targeted pre-training and robustness enhancement. This domain-aligned strategy promotes transferable robustness, unlocking significant performance gains. We introduce RPGen, a framework for elevating DP image synthesis by bolstering model parameter robustness. RPGen unfolds in three steps. First, a pre-trained classifier (e.g., ImageNet-trained) infers labels for private dataset images, which are aggregated into a class frequency distribution. To uphold DP, we inject Gaussian noise into this distribution and select public images whose frequencies align with the noisy private counterpart. Second, we pre-train the diffusion model on this curated subset, incorporating adversarial model perturbation (AMP) (Zheng, Zhang, and Mao 2021) to simulate worst-case parameter disruptions. This adversarial training fosters resilience to impending DP noise, reducing its detrimental effects on conver-

*Corresponding author.

gence, while the domain alignment ensures robust transfer to the private task. Finally, we apply DP-SGD (Abadi et al. 2016) for fine-tuning on the private data.

We evaluate RPGen using ImageNet (Deng et al. 2009) for pre-training—a standard in computer vision—paired with CelebA and CIFAR-10 for DP synthesis. Across privacy budgets $\varepsilon \in \{1, 5, 10\}$, RPGen outperforms baselines in synthesis fidelity and downstream utility. Averaged over settings, RPGen delivers a 20.18% lower Fréchet Inception Distance (FID) and 5.45% higher classification accuracy compared to the SOTA baseline, with qualitative improvements evident in generated samples (see Table 1).

Our ablation studies examine key hyperparameters, including AMP perturbation magnitude and data selection ratio. Modest perturbations consistently boost performance, affirming the utility of parameter robustness and RPGen's insensitivity to precise tuning. However, excessive magnitudes (e.g., > 0.5) impair convergence. Optimal results emerge at a 5% selection ratio, yielding nearly twofold FID reductions versus full-dataset pre-training, underscoring the value of domain curation for robustness transfer.

Our contributions are:

- To the best of our knowledge, we are the first to systematically enhance parameter robustness for DP image synthesis via diffusion models, while addressing the transferability of such robustness across domains.
- We present RPGen, a pre-training paradigm that integrates AMP for robustness augmentation with privacy-aware public data selection to maximize transferability.
- Through extensive experiments across datasets and privacy levels, we demonstrate RPGen's superior privacy-utility trade-off, establishing new SOTA on CIFAR-10 (Krizhevsky 2009) and CelebA (Liu et al. 2015).

Preliminaries

Differential Privacy. *Differential privacy* (DP) (Dwork et al. 2006) provides a formal framework for protecting individual-level information in a dataset during statistical analyses or model releases. It ensures that the outputs of a randomized mechanism are nearly indistinguishable regardless of whether any single record is included in the input, thereby safeguarding privacy while enabling useful insights into the data distribution. Formally, DP is defined as follows:

Definition 1 ((ε, δ)-Differential Privacy)**.** *Given two neighboring datasets D and D' that differ by one record, a randomized mechanism $\mathcal{M}$ satisfies (ε, δ)-differential privacy if, for all measurable sets S,*

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

where ε is the privacy budget and δ is a failure probability.

A smaller ε implies stronger privacy guarantees, as it limits the distinguishability between D and D'.

Sub-sampled Gaussian Mechanism (SGM). The Sub-sampled Gaussian Mechanism (SGM) (Mironov, Talwar, and Zhang 2019) is a fundamental tool for sanitizing data while controlling privacy loss. For a query function $f : D \subseteq \mathbb{R}^d \to \mathbb{R}^q$ with sensitivity Δ_f (the maximum ℓ_2-norm change in f when altering one record), SGM is defined as $\text{SGM}_{f,q,\sigma}(D) \triangleq f(S) + \mathcal{N}(0, \sigma^2\Delta_f^2 I)$. Here, S is a random subset of D where each element is included independently with probability $q \in (0, 1]$, $\mathcal{N}(0, \sigma^2\Delta_f^2 I)$ adds Gaussian noise with noise scale $\sigma > 0$, and I is the identity matrix. By combining subsampling and noise addition, SGM balances data utility for queries with privacy protection.

Rényi Differential Privacy (RDP). Rényi Differential Privacy (RDP) (Mironov, Talwar, and Zhang 2019) measures privacy loss using Rényi divergence. For a randomized mechanism M, the Rényi divergence of order $\alpha > 1$ between distributions Y and N is $D_\alpha(Y \parallel N) = \frac{1}{\alpha-1} \ln \mathbb{E}_{x\sim N}\left[\left(\frac{Y(x)}{N(x)}\right)^\alpha\right]$.

A mechanism M satisfies (α, γ)-RDP if $D_\alpha(M(D) \parallel M(D')) \leq \gamma$ for any adjacent datasets D and D'. RDP offers a flexible framework for analyzing privacy, providing nuanced bounds compared to traditional metrics, and is particularly useful for tracking cumulative privacy loss in composed mechanisms like SGM.

RDP for SGM. RDP provides tight bounds on the privacy loss of SGM (Mironov, Talwar, and Zhang 2019). Let p_0 and p_1 be the probability density functions of $\mathcal{N}(0, \sigma^2)$ and $\mathcal{N}(1, \sigma^2)$, respectively. Then, $\text{SGM}_{f,q,\sigma}(D)$ satisfies (α, γ)-RDP, where $\gamma \geq D_\alpha\left((1-q)p_0 + qp_1 \parallel p_0\right)$. This bound connects SGM parameters to RDP guarantees, enabling systematic control of the privacy-utility trade-off in applications like machine learning.

DP-SGD. In privacy-preserving deep learning, *differentially private stochastic gradient descent* (DP-SGD) (Abadi et al. 2016) is the standard approach. DP-SGD adapts traditional SGD by clipping each per-sample gradient to a fixed norm and adding Gaussian noise proportional to this norm before aggregation. This masks the contribution of any single example. The total privacy budget is computed by accounting for the per-iteration privacy cost under (ε, δ)-DP and applying composition and subsampling amplification (Bun and Steinke 2016; Dwork, Rothblum, and Vadhan 2010; Dwork, Roth et al. 2014) across iterations.

Diffusion Models. Diffusion models (Song and Ermon 2019) are likelihood-based generative models that learn to reverse a data degradation process, comprising forward noising and reverse denoising phases.

Forward Noising Process. Starting from clean data $x_0 \sim p(x_0)$, Gaussian noise is added iteratively via a Markov chain to produce a sequence of noisy samples $\{x_1, \ldots, x_T\}$:

$$p(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I),$$

where T is the number of steps, $\beta_t \in [0, 1)$ controls the noise level at step t, and I is the identity matrix. As t increases, x_t approaches pure Gaussian noise.

Reverse Denoising Process. The reverse process denoises from $x_T \sim \mathcal{N}(0, I)$ back to x_0 using a network parameterized by θ to predict the noise. The training objective is:

$$\mathcal{L}_{DM} = \mathbb{E}_t \mathbb{E}_{x_0} \mathbb{E}_\epsilon \left\|\epsilon - e_\theta(x_t, t)\right\|^2,$$

where $\epsilon \sim \mathcal{N}(0, I)$. For generation, sample $x_T \sim \mathcal{N}(0, I)$ and iteratively denoise using the predicted noise to obtain synthetic x_0.

Methodology

Our objective is to develop a differentially private (DP) image synthesis framework that generates high-fidelity synthetic images from sensitive datasets while minimizing performance degradation due to DP noise. By enhancing the diffusion model's parameter robustness during public pre-training and ensuring its transferability to the private domain, we aim to preserve generative utility without compromising privacy guarantees.

Formally, let $\mathcal{D}_{\text{pub}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_{\text{pub}}}$ be a labeled public dataset with N_{pub} samples, and $\mathcal{D}_{\text{priv}} = \{\mathbf{z}_j\}_{j=1}^{N_{\text{priv}}}$ an unlabeled private dataset with N_{priv} sensitive images. We seek to train a diffusion model e_θ on $\mathcal{D}_{\text{priv}}$ under (ε, δ)-DP, such that the generated images $\hat{\mathbf{z}} \sim p_\theta(\mathbf{z})$ closely match the distribution of $\mathcal{D}_{\text{priv}}$ in terms of fidelity (e.g., FID) and utility (e.g., classification accuracy), while bounding privacy leakage.

RPGen Overview

Compared to image classification tasks, where models often exhibit some inherent robustness to small perturbations due to simpler architectures and lower-dimensional outputs, diffusion models for image synthesis are particularly vulnerable to noise. Their iterative denoising process and high-dimensional parameter space amplify the disruptive effects of DP-induced Gaussian noise, leading to blurred or semantically incoherent generations and hindering convergence in transfer learning scenarios with domain shifts.

To address these challenges, we design *RPGen*, a three-stage paradigm that first curates domain-aligned public data under privacy constraints, then pre-trains a diffusion model with adversarial perturbations to build transferable parameter robustness, and finally fine-tunes it on private data using DP-SGD. This approach mitigates noise sensitivity by fostering resilience in pre-training, ensuring that robustness transfers effectively to the private task without additional privacy costs.

Formally, RPGen consists of the following three stages:

1. *Privacy-Preserving Data Selection stage:* Train a classifier on $\mathcal{D}_{\text{pub}}$ to infer a noisy class distribution from $\mathcal{D}_{\text{priv}}$ via SGM, then select top-k aligned classes to form $\mathcal{D}_{\text{sel}}$, promoting domain similarity for robust transfer.
2. *Adversarial Pre-training stage:* Pre-train the diffusion model on $\mathcal{D}_{\text{sel}}$ with AMP to enhance parameter robustness against worst-case perturbations, preparing it for DP fine-tuning.
3. *DP Fine-tuning stage:* Fine-tune the robust model on $\mathcal{D}_{\text{priv}}$ using DP-SGD with gradient clipping and noise addition, ensuring (ε, δ)-DP while leveraging pre-built resilience for superior synthesis quality.

Stage 1: Privacy-Preserving Data Selection

To enhance the transferability of parameter robustness from public pre-training to private fine-tuning, we curate a subset of public data that aligns closely with the private domain at the class level. This involves training a classifier on the public dataset to infer domain similarities, followed by privacy-preserving selection via noise injection.

Let $\mathcal{D}_{\text{pub}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_{\text{pub}}}$ be the labeled public dataset with N_{pub} samples across C classes, and $\mathcal{D}_{\text{priv}} = \{\mathbf{z}_j\}_{j=1}^{N_{\text{priv}}}$ the unlabeled private dataset. We train a classifier $f_\theta : \mathcal{X} \to [C]$ on $\mathcal{D}_{\text{pub}}$ using cross-entropy loss, capturing the public feature distribution.

We then infer labels on $\mathcal{D}_{\text{priv}}$: $\hat{y}_j = f_\theta(\mathbf{z}_j)$ for each $\mathbf{z}_j$, aggregating into a count vector $\mathbf{h} \in \mathbb{R}^C$ where $h_c = \sum_{j=1}^{N_{\text{priv}}} \mathbb{I}[\hat{y}_j = c]$. The normalized distribution is $\mathbf{p} = \mathbf{h}/N_{\text{priv}}$.

To protect privacy, we privatize $\mathbf{h}$ using the Sub-sampled Gaussian Mechanism (SGM) applied to the count query $g(\mathcal{D}_{\text{priv}}) = \mathbf{h}$, with sensitivity $\Delta_g = 1$:

$$\tilde{\mathbf{h}} = \text{SGM}_{g,q,\sigma}(\mathcal{D}_{\text{priv}}) = g(S) + \mathcal{N}(0, \sigma^2 \Delta_g^2 I), \quad (1)$$

where S is subsampled with probability q, yielding noisy distribution $\tilde{\mathbf{p}} = \tilde{\mathbf{h}}/|S|$.

This satisfies (α, γ)-RDP for $\alpha > 1$ with $\gamma \geq D_\alpha\left((1-q)p_0 + qp_1 \parallel p_0\right)$, where p_0 and p_1 are densities of $\mathcal{N}(0, \sigma^2)$ and $\mathcal{N}(1, \sigma^2)$. This bound controls leakage and converts to (ε, δ)-DP (Mironov, Talwar, and Zhang 2019).

Finally, we select top-k classes from $\tilde{\mathbf{p}}$: $\mathcal{C}_{\text{top}} = \arg\max_{|\mathcal{C}|=k} \sum_{c \in \mathcal{C}} \tilde{p}_c$. The curated subset is $\mathcal{D}_{\text{sel}} = \{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{pub}} \mid y_i \in \mathcal{C}_{\text{top}}\}$, forming the basis for Stage 2.

Stage 2: Adversarial Pre-training

In Stage 2, we pre-train a diffusion model on the curated public subset $\mathcal{D}_{\text{sel}}$ obtained from Stage 1, incorporating Adversarial Model Perturbation (AMP) to enhance the model's parameter robustness against random perturbations. This step is crucial for mitigating the adverse effects of DP noise in the subsequent fine-tuning stage, as the augmented robustness promotes better convergence under noisy gradients. We build upon the diffusion model framework and adversarial model perturbation (Zheng, Zhang, and Mao 2021), adapting them to foster resilience in the generative setting.

Recall that diffusion models learn to reverse a forward noising process by minimizing a noise prediction loss. For our pre-training, we parameterize the noise predictor as $e_\theta(\cdot, t)$, where θ denotes the model parameters. The standard objective on $\mathcal{D}_{\text{sel}}$ is:

$$\mathcal{L}_{\text{DM}}(\theta) = \mathbb{E}_{t \sim [1,T]} \mathbb{E}_{x_0 \sim \mathcal{D}_{\text{sel}}} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \left\| \epsilon - e_\theta(x_t, t) \right\|^2, \quad (2)$$

where x_t is derived from x_0 via the forward process $p(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I)$.

To integrate AMP, we introduce worst-case perturbations within a norm ball around the current parameters, encouraging the model to perform well even under parameter distortions. Following the AMP formulation, we define the adversarial loss over a mini-batch $\mathcal{B} \subseteq \mathcal{D}_{\text{sel}}$ as:

$$\mathcal{J}_{\mathcal{B}}(\theta) = \max_{\Delta \in \mathbf{B}(\mathbf{0}; \gamma)} \frac{1}{|\mathcal{B}|} \sum_{(x_0, t, \epsilon) \in \mathcal{B}} \left\| \epsilon - e_{\theta + \Delta}(x_t, t) \right\|^2, \quad (3)$$

where $\mathbf{B}(\mathbf{0}; \gamma) = \{\Delta \in \Theta : \|\Delta\| \leq \gamma\}$ is the norm ball with radius $\gamma \geq 0$, controlling the perturbation strength. A larger γ induces greater robustness by training against more severe parameter noise.

The worst-case perturbation $\Delta_{\mathcal{B}}$ is computed via:

$$\Delta_{\mathcal{B}} = \arg \max_{\Delta \in \mathbf{B}(\mathbf{0};\gamma)} \frac{1}{|\mathcal{B}|} \sum_{(x_0,t,\epsilon)\in\mathcal{B}} \|\epsilon - e_{\theta+\Delta}(x_t,t)\|^2 . \tag{4}$$

In practice, this is optimized using gradient ascent on Δ with learning rate η_1, projecting back into the ball if necessary, to simulate perturbations that maximally increase the loss.

The model parameters are then updated based on the perturbed objective:

$$\alpha = \eta_2 \nabla_\theta \mathcal{J}_{\mathcal{B}}(\theta), \tag{5}$$

where η_2 is the update learning rate. To avoid retaining the perturbation (which could degrade clean performance), the final update is applied to the unperturbed parameters: $\theta \leftarrow \theta + \alpha$. This process iterates over mini-batches from $\mathcal{D}_{\text{sel}}$, yielding a pre-trained model e_θ with enhanced parameter robustness, primed for DP fine-tuning in Stage 3.

Stage 3: DP Fine-tuning

In Stage 3, we fine-tune the robust diffusion model obtained from Stage 2 on the private dataset $\mathcal{D}_{\text{priv}}$ using DP-SGD, ensuring that the training process satisfies differential privacy guarantees while preserving the model's generative utility. This step leverages the parameter robustness built during pre-training to mitigate the performance degradation typically caused by DP noise, as the model is now more resilient to perturbations in its gradients. We adapt the standard diffusion objective to the DP setting, drawing on the SGM and RDP for privacy accounting.

The fine-tuning objective remains the noise prediction loss, but now optimized over $\mathcal{D}_{\text{priv}}$:

$$\mathcal{L}_{\text{DM}}(\theta) = \mathbb{E}_{t\sim[1,T]} \mathbb{E}_{z_0\sim\mathcal{D}_{\text{priv}}} \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)} \|\epsilon - e_\theta(z_t,t)\|^2 , \tag{6}$$

where z_t is generated from private sample z_0 via the forward noising process, and e_θ is the pre-trained noise predictor initialized from Stage 2.

To enforce DP, we employ DP-SGD, which modifies the gradient computation to bound sensitivity and add noise. For a mini-batch $\mathcal{B} \subseteq \mathcal{D}_{\text{priv}}$ of size B, we first compute per-sample gradients:

$$\mathbf{g}_i = \nabla_\theta \ell(e_\theta(z_t^{(i)},t),\epsilon^{(i)}), \quad \forall (z_0^{(i)},t,\epsilon^{(i)}) \in \mathcal{B}, \tag{7}$$

where $\ell(\cdot,\cdot) = \|\epsilon - e_\theta(z_t,t)\|^2$ is the sample-wise loss. Each gradient is clipped to bound its L2-norm:

$$\tilde{\mathbf{g}}_i = \mathbf{g}_i / \max\left(1, \frac{\|\mathbf{g}_i\|_2}{C}\right), \tag{8}$$

with clipping threshold $C > 0$ controlling sensitivity.

The aggregated noisy gradient is then formed by averaging the clipped gradients and adding Gaussian noise:

$$\bar{\mathbf{g}} = \frac{1}{B} \sum_{i=1}^{B} \tilde{\mathbf{g}}_i + \mathcal{N}\left(0, \frac{\sigma^2 C^2}{B^2} I\right), \tag{9}$$

where $\sigma > 0$ is the noise multiplier. The model parameters are updated as $\theta \leftarrow \theta - \eta \bar{\mathbf{g}}$, with learning rate η.

This process corresponds to applying SGM to the gradient query function, with sampling probability $q = B/N_{\text{priv}}$. The privacy loss is tracked via RDP: each DP-SGD step satisfies (α,γ)-RDP with $\gamma \geq D_\alpha((1-q)p_0 + qp_1 \parallel p_0)$, where $p_0 = \mathcal{N}(0,\sigma^2)$ and $p_1 = \mathcal{N}(C,\sigma^2)$, convertible to (ε,δ)-DP over multiple steps using composition theorems (Mironov, Talwar, and Zhang 2019). Upon convergence, the fine-tuned model e_θ enables image synthesis by sampling from the reverse denoising process.

In particular, RPGen incurs privacy budget expenditure in two key phases: (1) estimating the privatized class distribution from the private dataset during data selection, and (2) computing noisy gradients in each iteration of the fine-tuning process on the private data. Each of these phases can be modeled as a sequence of composed SGM applications. As such, we employ RDP to accurately account for and compose the overall privacy costs of RPGen, enabling precise control over the (ε,δ)-DP guarantees.

Experiments

Experimental Setup

We utilize ImageNet (Deng et al. 2009), a cornerstone dataset for pre-training in computer vision with 1,281,167 training images, 50,000 validation images, and 100,000 test images across 1,000 classes, alongside CelebA (Liu et al. 2015) and CIFAR-10 (Krizhevsky 2009) for DP image synthesis. CelebA comprises 202,525 celebrity face images with 40 attributes (162,770 training, 19,867 validation, 19,962 test), center-cropped and resized to 32×32 (CelebA32) or 64×64 (CelebA64). CIFAR-10 includes 60,000 32×32 natural images across 10 classes (45,000 training, 5,000 validation, 10,000 test). These datasets pose greater synthesis challenges than simpler ones like MNIST or Fashion-MNIST due to their complexity and scale.

In this work, we compare the proposed RPGen against eight baselines: DPDM (Dockhorn et al. 2022), PDP-Diffusion (Ghalebikesabi et al. 2023), DP-LDM (Lyu et al. 2023), DPSDA (Lin et al. 2024), DPGAN (Torkzadehmahani, Kairouz, and Paten 2019), DPGAN with pre-training (DPGAN-p) (Ghalebikesabi et al. 2023), PRIVIMAGE with GAN (PRIVIMAGE+G) (Li et al. 2024), and PRIVIMAGE with Diffusion (PRIVIMAGE+D) (Li et al. 2024).

- **DPDM** (Dockhorn et al. 2022) trains lightweight diffusion models with large batch sizes, injecting Gaussian noise into gradients via DP-SGD (Abadi et al. 2016) for privacy-preserving generation.
- **PDP-Diffusion** (Ghalebikesabi et al. 2023) uses large batches for stability and pre-trains on public data before DP fine-tuning on sensitive datasets.
- **DP-LDM** (Lyu et al. 2023) fine-tunes only the label embedding and attention modules of a pre-trained diffusion model, reducing parameters and noise requirements.
- **DPSDA** (Lin et al. 2024) employs a Private Evolution algorithm to adapt pre-trained models for synthetic dataset generation without fine-tuning.
- **DPGAN** (Torkzadehmahani, Kairouz, and Paten 2019) trains GANs directly on sensitive data using DP-SGD.

- **DPGAN-p** extends DPGAN by pre-training on public data (Ghalebikesabi et al. 2023) before DP fine-tuning.
- **PRIVIMAGE+G** (Li et al. 2024) queries sensitive data semantics to select public subsets for GAN pre-training.
- **PRIVIMAGE+D** (Li et al. 2024) applies the same selection strategy but for diffusion models.

For CIFAR-10, we set $\delta = 10^{-5}$, and for CelebA, $\delta = 10^{-6}$. Across both datasets, we evaluate under three common privacy budgets: $\varepsilon \in \{1, 5, 10\}$ (Dockhorn et al. 2022; Ghalebikesabi et al. 2023).

We implement RPGen in PyTorch, employing the Denoising Diffusion Probabilistic Model (DDPM) architecture (Song and Ermon 2019) for the diffusion backbone. In the adversarial pre-training stage (Stage 2), we train on the curated subset $\mathcal{D}_{\text{sel}}$ for 4,000 epochs, with a learning rate of $\eta_1 = 0.4$ for worst-case perturbation computation and $\eta_2 = 10^{-4}$ for model updates. Weight decay is disabled (set to 0.0), and we use a batch size of 512. The perturbation magnitude defaults to $\gamma = 0.2$. For the data selection stage (Stage 1), we configure the SGM noise scale as $\sigma = 484$ for CIFAR-10 and $\sigma = 5300$ for CelebA, with a default selection ratio of 5% (corresponding to top-$k = 50$ classes) from ImageNet. During the DP fine-tuning stage (Stage 3), we apply a gradient clipping norm of $C = 0.001$, adjust the noise multiplier to achieve the target ε, and fine-tune for 50 epochs with a batch size of 19,384. All experiments are conducted on NVIDIA A100 GPUs, with results averaged across six random seeds.

We evaluate the fidelity and utility of synthetic datasets using two established metrics: Fréchet Inception Distance (FID) for image quality and Classification Accuracy (CA) for downstream task performance, consistent with prior DP image synthesis studies (Dockhorn et al. 2022; Ghalebikesabi et al. 2023).

- **Fréchet Inception Distance (FID)** (Ho, Jain, and Abbeel 2020; Brock, Donahue, and Simonyan 2019): This metric quantifies the similarity between generated and real image distributions by comparing their feature statistics from an Inception network. Lower FID values indicate higher fidelity and realism. We compute FID using 5,000 synthetic images against the real test set.
- **Classification Accuracy (CA)**: To measure utility, we train classifiers on synthetic images and evaluate their accuracy on the real test set, assessing how well the synthetics capture discriminative features. We employ three models—Logistic Regression (LR), Multi-Layer Perceptron (MLP), and Convolutional Neural Network (CNN)—trained on 50,000 synthetic images per method.

Main Results

Table 1 presents the CA and FID scores across various settings. Overall, RPGen consistently improves CA and reduces FID compared to baselines in all configurations. For example, on CIFAR-10 with $\varepsilon = 10$, RPGen boosts CNN-based CA from the previous best of 68.8% (PRIVIMAGE+D) to 73.4%, a 4.6% gain, while lowering FID from 27.6 (PRIVIMAGE+D) to 26.5, a 3.99% reduction.

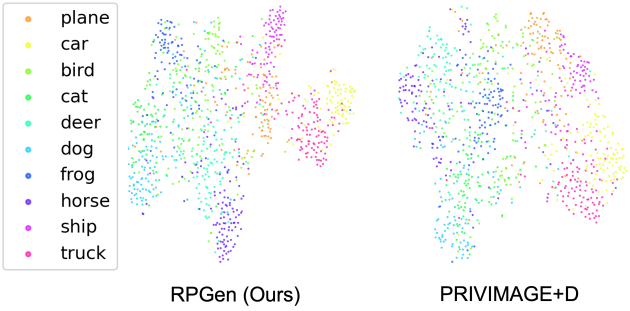

Figure 1: t-SNE visualization of feature embeddings from synthetic CIFAR-10 images generated by RPGen (left) and PRIVIMAGE+D (right).

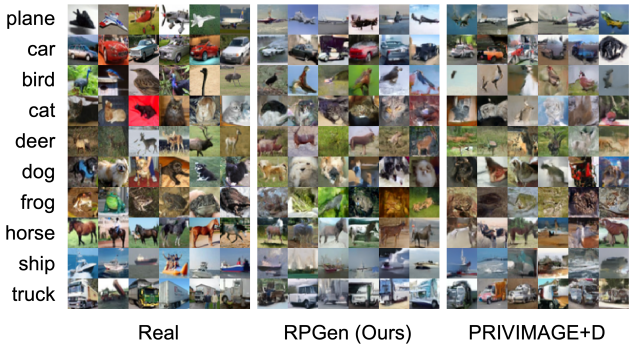

Figure 2: Qualitative comparison of synthetic CIFAR-10 images generated by RPGen (middle) and PRIVIMAGE+D (right), alongside real images (left).

Furthermore, RPGen demonstrates robustness across diverse private datasets. While ImageNet and CIFAR-10 encompass general objects, CelebA focuses on faces, introducing a significant domain shift. Nonetheless, RPGen excels on CelebA; for instance, on CelebA32 with $\varepsilon = 10$, it reduces FID from the prior best of 18.9 (PRIVIMAGE+G) to 16.3, a 13.76% improvement. This resilience highlights the efficacy of our privacy-preserving data selection in bridging domain gaps, ensuring transferable robustness even when public and private distributions differ markedly.

RPGen also handles higher-resolution image generation effectively. On CelebA64 with $\varepsilon = 5$, it decreases FID from 45.2 (PRIVIMAGE+G) to 39.3, yielding a 13.05% reduction. These gains are particularly noteworthy as higher resolutions amplify the challenges of DP noise, yet RPGen's pre-trained robustness mitigates artifacts like blurring or loss of fine details, producing more coherent and realistic outputs.

Although the PRIVIMAGE variants often achieve state-of-the-art results, their performance varies across datasets: PRIVIMAGE+D achieves the best performance on CIFAR-10 under $\varepsilon = 10$, whereas PRIVIMAGE+G performs better on CelebA32 and CelebA64. Notably, on CelebA64, PRIVIMAGE+D even underperforms compared to DPSDA. Similar inconsistencies appear for $\varepsilon = 5$ and $\varepsilon = 1$. In contrast, RPGen delivers the best results across all settings, under-

Method	$\varepsilon = 10$						$\varepsilon = 5$						$\varepsilon = 1$					
	CIFAR-10				CeA32	CeA64	CIFAR-10				CeA32	CeA64	CIFAR-10				CeA32	CeA64
	CA (%)			FID	FID	FID	CA (%)			FID	FID	FID	CA (%)			FID	FID	FID
	LR	MLP	CNN				LR	MLP	CNN				LR	MLP	CNN			
DPGAN	9.2	8.4	10.5	258.0	202.0	121.0	14.2	14.6	13.0	210.0	227.0	190.0	16.2	17.4	14.8	225.0	232.0	162.0
DPGAN-p	13.6	14.9	24.1	49.7	29.7	51.1	13.9	14.3	19.2	48.5	23.9	52.2	11.3	13.8	12.8	70.1	37.9	54.5
DPDM	20.7	24.6	21.3	304.0	113.0	115.0	21.1	24.7	22.0	311.0	122.0	127.0	19.6	22.3	14.7	340.0	223.0	243.0
DP-LDM	15.2	14.1	26.0	48.6	21.9	58.0	15.3	14.6	24.8	48.9	22.2	63.9	12.8	11.8	18.8	50.1	45.5	131.9
PDP-Diffusion	18.7	21.4	30.4	66.8	22.6	51.6	19.3	22.2	28.7	70.0	23.6	55.9	17.7	19.4	22.9	87.5	33.7	77.7
DPSDA	24.1	25.0	47.9	29.9	23.8	49.0	23.5	24.4	46.1	30.1	33.8	49.4	24.2	23.6	47.1	31.2	37.9	54.9
PRIVIMAGE+G	19.9	24.5	44.3	28.1	18.9	38.2	19.6	24.6	39.2	29.9	19.8	45.2	15.8	18.0	25.5	47.5	31.8	45.1
PRIVIMAGE+D	32.6	36.5	68.8	27.6	19.1	49.3	32.4	35.9	69.4	27.6	20.1	52.9	30.2	33.2	66.2	29.8	26.0	71.4
RPGen (Ours)	**33.4**	**37.7**	**73.4**	**26.5**	**16.3**	**37.1**	**32.9**	**37.0**	**72.5**	**27.0**	**17.5**	**39.3**	**31.2**	**35.0**	**70.3**	**28.5**	**22.7**	**43.5**

Table 1: FID and CA of RPGen and eight baselines on CIFAR-10, CelebA32, and CelebA64 under $\varepsilon = \{10, 5, 1\}$. Due to space constraints, CeA32 and CeA64 denote CelebA32 and CelebA64, respectively. Best results in each column are shown in bold.

scoring its stability and versatility. This consistency stems from the synergistic integration of AMP and domain-aligned pre-training, which provides a more reliable defense against DP noise compared to baselines that either lack robustness enhancement or domain curation.

Figure 1 visualizes t-SNE embeddings of RPGen-generated images, forming 10 distinct clusters aligned with CIFAR-10 categories. This indicates that classifiers trained on RPGen data better capture discriminative features, outperforming those on PRIVIMAGE+D (the SOTA on CIFAR-10), whose embeddings exhibit less clear separation and yield inferior classification. Qualitative examples of synthetic CIFAR-10 images from RPGen, PRIVIMAGE+D, and real data are shown in Figure 2, where RPGen's outputs exhibit sharper details and better semantic fidelity.

Method	CIFAR-10				CeA32	CeA64
	CA (%)			FID	FID	FID
	LR	MLP	CNN			
Real	37.4	45.7	86.1	-	-	-
NonPriv	35.8	42.2	77.1	19.8	9.0	18.0
RPGen (Ours)	33.4	37.7	73.4	26.5	16.3	37.1

Table 2: FID and Classification Accuracy (CA) for RPGen, the non-private baseline, and real data on CIFAR-10, CelebA32, and CelebA64 under $\varepsilon = 10$. Lower FID and higher CA indicate better performance.

Furthermore, as detailed in Table 2, RPGen exhibits minimal utility degradation relative to non-private counterparts. On average, it incurs only a 3.5% drop in CA across the three classification models compared to the non-private baseline. When benchmarked against classifiers trained directly on real data, RPGen's synthetic images result in an average CA reduction of 8.2%, a modest gap that underscores the method's effectiveness in preserving downstream utility while enforcing strong DP guarantees.

Ablation Studies

To dissect RPGen's key components, we perform ablations on CIFAR-10 under $\varepsilon = 10$. Omitting AMP during pre-training ($\gamma = 0$ in Figures 3 and 4) results in a 4.3% FID increase and a 3.6% CA drop, highlighting its critical role in building noise tolerance. Removing the data selection stage (Selection Ratio=100% in Figure 5) leads to a 48.1% FID rise and a 40.0% CA decline, affirming the importance of domain alignment for effective robustness transfer.

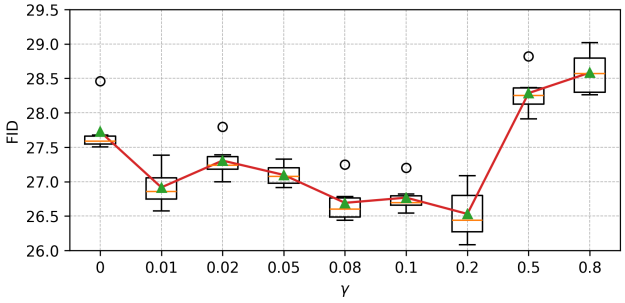

Figure 3: FID on CIFAR-10 under $\varepsilon = 10$ versus perturbation magnitude γ.

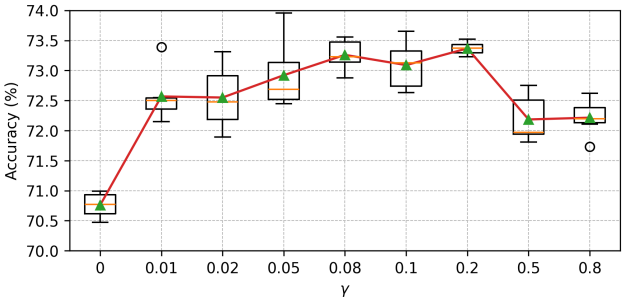

Figure 4: CA on CIFAR-10 under $\varepsilon = 10$ versus perturbation magnitude γ.

Impact of Perturbation. In this section, we analyze how the perturbation in AMP influences performance, evaluating $\gamma \in 0, 0.01, 0.02, 0.05, 0.08, 0.1, 0.2, 0.5, 0.8$ on CIFAR-10 under $\varepsilon = 10$. As shown in Figures 3 and 4, both FID and CA reach their best values at $\gamma = 0.2$. When the perturbation becomes too large ($\gamma > 0.5$), performance drops be-

cause excessive robustness begins to impede effective convergence. Interestingly, even relatively small perturbations ($0 < \gamma < 0.2$) consistently improve results across metrics. This trend not only highlights the value of incorporating parameter robustness, but also shows that RPGen remains stable across a wide range of hyperparameter settings.

Impact of Selection Ratio. Figure 5 further examines the effect of the selection ratio. Ratios between 5% and 20% offer the strongest performance, with 20% yielding the best FID and 5% achieving the highest CA. We choose 5% as the default because it maintains strong performance while keeping computational cost low. Extremely small ratios (below 5%) lack sufficient data diversity, whereas very large ratios (above 20%) introduce less relevant samples, reducing robustness transfer. These observations confirm the importance of the selection stage in ensuring that robustness learned during pre-training carries over effectively to downstream DP training.

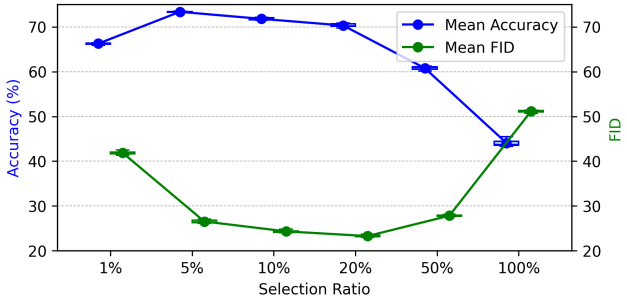

Figure 5: Impact of selection ratio on FID and CA for CIFAR-10 under $\varepsilon = 10$.

Related Work

Differential Privacy

Differential privacy (DP) (Dwork 2008; Dwork and Roth 2014) provides a rigorous mathematical framework for quantifying privacy guarantees, typically parameterized by a privacy budget that balances protection and utility. It has been widely applied to diverse data analysis tasks, including synthetic dataset generation (Yuan et al. 2023; Zhang et al. 2021; Wang et al. 2023; Du et al. 2023), marginal release (Zhang et al. 2018), range queries (Du et al. 2021), and streaming data processing (Wang et al. 2021). A seminal advancement is DP-SGD (Abadi et al. 2016), introduced by Abadi et al. as a foundational algorithm for DP learning.

Recent studies (Park et al. 2023; Shi et al. 2023; Wang et al. 2025) have incorporated AMP-like techniques into DP algorithms to improve test accuracy, although these efforts have focused mainly on image classification tasks. It is still unclear whether such methods generalize to synthetic image generation, which requires modeling complex distributions rather than learning discriminative features. In addition, prior methods strengthen robustness during fine-tuning on private data, while RPGen applies AMP solely in the public pre-training stage. This design choice introduces both unique advantages and specific challenges (Wang et al. 2024). One clear benefit is that robustness accumulation occurs before DP noise is introduced, allowing the model to build parameter resilience without interference from privacy-preserving perturbations. The corresponding challenge arises from domain shift, since public and private data distributions may differ significantly, unlike fine-tuning approaches that operate on the same domain. This issue has not been addressed in previous work (Wang et al. 2024). RPGen tackles this difficulty by explicitly optimizing for robustness transferability through privacy-aware data selection, which forms the central motivation for our framework.

Data Selection for Fine-Tuning

Data selection techniques (Schaul et al. 2016; Loshchilov and Hutter 2015; Katharopoulos and Fleuret 2018) often aim to identify subsets that approximate training on the full dataset, typically involving dynamic updates to sample importance scores throughout the process. They have proven effective across supervised and semi-supervised vision tasks (Sener and Savarese 2018; Coleman et al. 2020; Killamsetty et al. 2021a,b; Mirzasoleiman, Bilmes, and Leskovec 2020; Paul, Ganguli, and Dziugaite 2021; Wei, Iyer, and Bilmes 2015), and have also been applied to enhance DP fine-tuning using public data (Li et al. 2024). Our work departs from prior efforts by using data selection to support not only utility transfer, but also robustness transfer.

Conclusion

In this paper, we addressed the persistent challenge of performance degradation in differentially private (DP) image synthesis caused by noise injection during training, proposing RPGen as a novel framework to enhance diffusion models' parameter robustness while ensuring transferable benefits across domains. By integrating privacy-aware data selection to curate domain-aligned public samples, adversarial model perturbation (AMP) for robustness augmentation during pre-training, and DP-SGD for secure fine-tuning, RPGen effectively mitigates the adverse effects of DP noise, yielding high-fidelity synthetic images without compromising privacy guarantees. Our extensive experiments across datasets like CIFAR-10 and CelebA, under varying privacy budgets, demonstrate RPGen's superiority over state-of-the-art methods. To the best of our knowledge, this is the first work to systematically tackle robustness transfer in DP generative modeling, paving the way for more practical privacy-preserving data sharing in sensitive applications.

While RPGen advances DP image synthesis, it assumes labeled public data, which may not always be available. Future work could extend RPGen to unlabeled public sources via self-supervised selection, incorporate advanced diffusion variants like latent diffusion models, or extend RPGen to other generative architectures, such as GANs or transformers, and explore its efficacy in federated learning scenarios, further advancing the privacy-utility frontier in AI.

Acknowledgments

This work is supported by NSFC through grants 62322202, 62441612, 62432006, and U24A20334, Beijing Natural Science Foundation through grant L253021, Local Science and Technology Development Fund of Hebei Province Guided by the Central Government of China through grants 246Z0102G and 254Z9902G, Major Science and Technology Special Projects of Yunnan Province through grants 202502AD080012, 202502AD080006 and 202402AG050007, and the Fundamental Research Funds for the Central Universities.

References

Abadi, M.; Chu, A.; Goodfellow, I. J.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, 308–318.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Bun, M.; and Steinke, T. 2016. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, 635–658. Springer.

Coleman, C.; Yeh, C.; Mussmann, S.; Mirzasoleiman, B.; Bailis, P.; Liang, P.; Leskovec, J.; and Zaharia, M. 2020. Selection via Proxy: Efficient Data Selection for Deep Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Dankar, F. K.; and Emam, K. E. 2013. Practicing Differential Privacy in Health Care: A Review. *Trans. Data Priv.*, 6(1): 35–67.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 248–255. IEEE Computer Society.

Dockhorn, T.; Cao, T.; Vahdat, A.; and Kreis, K. 2022. Differentially Private Diffusion Models. *CoRR*, abs/2210.09929.

Du, L.; Zhang, Z.; Bai, S.; Liu, C.; Ji, S.; Cheng, P.; and Chen, J. 2021. AHEAD: Adaptive Hierarchical Decomposition for Range Query under Local Differential Privacy. In *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, 1266–1288.

Du, Y.; Hu, Y.; Zhang, Z.; Fang, Z.; Chen, L.; Zheng, B.; and Gao, Y. 2023. LDPTrace: Locally Differentially Private Trajectory Synthesis. *Proc. VLDB Endow.*, 16(8): 1897–1909.

Dwork, C. 2008. Differential Privacy: A Survey of Results. In *Theory and Applications of Models of Computation, 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings*, 1–19.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. D. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, 265–284.

Dwork, C.; and Roth, A. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4): 211–407.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407.

Dwork, C.; Rothblum, G. N.; and Vadhan, S. 2010. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, 51–60. IEEE.

Ghalebikesabi, S.; Berrada, L.; Gowal, S.; Ktena, I.; Stanforth, R.; Hayes, J.; De, S.; Smith, S. L.; Wiles, O.; and Balle, B. 2023. Differentially Private Diffusion Models Generate Useful Synthetic Images. *CoRR*, abs/2302.13861.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Hu, Y.; Wu, F.; Li, Q.; Long, Y.; Garrido, G. M.; Ge, C.; Ding, B.; Forsyth, D. A.; Li, B.; and Song, D. 2024. SoK: Privacy-Preserving Data Synthesis. In *IEEE Symposium on Security and Privacy, SP 2024, San Francisco, CA, USA, May 19-23, 2024*, 4696–4713. IEEE.

Katharopoulos, A.; and Fleuret, F. 2018. Not All Samples Are Created Equal: Deep Learning with Importance Sampling. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, 2530–2539.

Killamsetty, K.; Sivasubramanian, D.; Ramakrishnan, G.; De, A.; and Iyer, R. K. 2021a. GRAD-MATCH: Gradient Matching based Data Subset Selection for Efficient Deep Model Training. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 5464–5474.

Killamsetty, K.; Sivasubramanian, D.; Ramakrishnan, G.; and Iyer, R. K. 2021b. GLISTER: Generalization based Data Subset Selection for Efficient and Robust Learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 8110–8118.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. Tech. report, University of Toronto.

Li, K.; Gong, C.; Li, Z.; Zhao, Y.; Hou, X.; and Wang, T. 2024. PrivImage: Differentially Private Synthetic Image Generation using Diffusion Models with Semantic-Aware Pretraining. In Balzarotti, D.; and Xu, W., eds., *33rd USENIX Security Symposium, USENIX Security 2024, Philadelphia, PA, USA, August 14-16, 2024*. USENIX Association.

Lin, Z.; Gopi, S.; Kulkarni, J.; Nori, H.; and Yekhanin, S. 2024. Differentially Private Synthetic Data via Foundation Model APIs 1: Images. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 3730–3738. IEEE Computer Society.

Loshchilov, I.; and Hutter, F. 2015. Online Batch Selection for Faster Training of Neural Networks. *CoRR*, abs/1511.06343.

Lyu, S.; Vinaroz, M.; Liu, M. F.; and Park, M. 2023. Differentially Private Latent Diffusion Models. *CoRR*, abs/2305.15759.

Mironov, I.; Talwar, K.; and Zhang, L. 2019. Rényi Differential Privacy of the Sampled Gaussian Mechanism. *CoRR*, abs/1908.10530.

Mirzasoleiman, B.; Bilmes, J. A.; and Leskovec, J. 2020. Coresets for Data-efficient Training of Machine Learning Models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, 6950–6960.

Park, J.; Kim, H.; Choi, Y.; and Lee, J. 2023. Differentially Private Sharpness-Aware Training. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 27204–27224. PMLR.

Paul, M.; Ganguli, S.; and Dziugaite, G. K. 2021. Deep Learning on a Data Diet: Finding Important Examples Early in Training. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 20596–20607.

Schaul, T.; Quan, J.; Antonoglou, I.; and Silver, D. 2016. Prioritized Experience Replay. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Shi, Y.; Liu, Y.; Wei, K.; Shen, L.; Wang, X.; and Tao, D. 2023. Make Landscape Flatter in Differentially Private Federated Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 24552–24562. IEEE.

Song, Y.; and Ermon, S. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 11895–11907.

Torkzadehmahani, R.; Kairouz, P.; and Paten, B. 2019. DP-CGAN: Differentially Private Synthetic Data and Label Generation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, 98–104. Computer Vision Foundation / IEEE.

Wang, H.; Zhang, Z.; Wang, T.; He, S.; Backes, M.; Chen, J.; and Zhang, Y. 2023. PrivTrace: Differentially Private Trajectory Synthesis by Adaptive Markov Models. In *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*.

Wang, T.; Chen, J. Q.; Zhang, Z.; Su, D.; Cheng, Y.; Li, Z.; Li, N.; and Jha, S. 2021. Continuous Release of Data Streams under both Centralized and Local Differential Privacy. In *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, 1237–1253.

Wang, Z.; Zhu, R.; Zhou, D.; Zhang, Z.; Mitchell, J.; Tang, H.; and Wang, X. 2024. {DPAdapter}: Improving Differentially Private Deep Learning through Noise Tolerance Pre-training. In *33rd USENIX Security Symposium (USENIX Security 24)*, 991–1008.

Wang, Z.; Zhu, R.; Zhou, D.; Zhang, Z.; Wang, X.; and Tang, H. 2025. {Sharpness-Aware} Initialization: Improving Differentially Private Machine Learning from First Principles. In *34th USENIX Security Symposium (USENIX Security 25)*, 3103–3122.

Wei, K.; Iyer, R. K.; and Bilmes, J. A. 2015. Submodularity in Data Subset Selection and Active Learning. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 1954–1963.

Yuan, Q.; Zhang, Z.; Du, L.; Chen, M.; Cheng, P.; and Sun, M. 2023. PrivGraph: Differentially Private Graph Data Publication by Exploiting Community Information. In *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*.

Zhang, Z.; Wang, T.; Li, N.; He, S.; and Chen, J. 2018. CALM: Consistent Adaptive Local Marginal for Marginal Release under Local Differential Privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018*, 212–229.

Zhang, Z.; Wang, T.; Li, N.; Honorio, J.; Backes, M.; He, S.; Chen, J.; and Zhang, Y. 2021. PrivSyn: Differentially Private Data Synthesis. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, 929–946.

Zheng, Y.; Zhang, R.; and Mao, Y. 2021. Regularizing Neural Networks via Adversarial Model Perturbation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 8156–8165. Computer Vision Foundation / IEEE.