Differentially Private Generation of High Fidelity Samples From Diffusion Models

Anonymous Authors[1]

Abstract

Diffusion based generative models achieve unprecedented image quality but are known to leak private information about the training data. Our goal is to provide provable guarantees on privacy leakage of training data while simultaneously enabling generation of high-fidelity samples. Our proposed approach first non-privately trains an ensemble of diffusion models and then aggregates their prediction to provide privacy guarantees for generated samples. We demonstrate the success of our approach on the MNIST and CIFAR-10.

1. Introduction

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019; Song et al., 2021) have emerged as a powerful class of generative models, with publicly available pretrained models (Dhariwal & Nichol, 2021; Rombach et al., 2022) being fine-tuned on potentially sensitive datasets such as chest X-rays (Wang et al., 2017) and brain MRIs (Royer et al., 2022). This is concerning because recent work (Carlini et al., 2023; Somepalli et al., 2022) has shown that diffusion models can directly produce memorized training data during inference, completely violating the privacy of sensitive data. Differential Privacy (DP) is the gold standard for quantifying privacy risks and providing provable guarantees against attacks (Dwork, 2006). When applied to diffusion models, DP guarantees provide a provable upper bound on the privacy leakage from samples.

In this work, we generate high-quality images with DP guarantees from diffusion models by proposing a novel differentially private generation process based on an ensemble of diffusion models. We provide the key aspects of our DP generation process and our particular contributions below.

- We show that sampling from diffusion models via *DP generation* prevents attackers from extracting private data through observed samples, by ensuring that a generated sample cannot depend too much on any single datapoint in the training dataset.
- We find that the stochastic sampling process *is inherently DP* as long as we clip the model predictions at each sampling timestep, albeit with a large privacy cost.
- We propose amplifying the inherent privacy of the sampling process by training a *non-private ensemble* of models on disjoint subsets of the training dataset. During sampling, we clip and average the predictions of all models to generate samples from the ensemble.
- We exploit the few-shot learning abilities of diffusion models to generate high quality samples. Specifically, when each model in the ensemble is trained on as little as $1/100$ of the private training data, we generate high fidelity samples from an ensemble of models.
- We analyze the noise schedule and find that privacy leakage is highest in the final steps of sampling where we add little noise. We find that replacing the ensemble with a publicly available diffusion model for these steps in the sampling process reduces the privacy cost without degrading utility. Our analysis demonstrates that, surprisingly, the steps with the highest privacy cost are actually the least dependent on the data.
- We demonstrate that the end-to-end process generates high fidelity synthetic samples (Fig. 1).

(a) MNIST (b) CIFAR-10

Figure 1. High fidelity images synthesized using our proposed differentially private generation process for diffusion models.

2. Background

Diffusion Models. Diffusion models progressively perturb images by adding an increasing amount of Gaussian noise and generate images by reversing this process through sequential denoising. Specifically, Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) are Markov chains with the following joint distribution:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$

where each marginal $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is parametrized as a Gaussian distribution with learnable mean and a fixed (scaled) variance, $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ that is the target data distribution, and $p(\mathbf{x}_T)$ is a standard Gaussian $\mathcal{N}(0, \mathbf{I})$. The corresponding forward process is a Markov chain given by:

$$q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})$$

where $q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \sqrt{\beta_t}\mathbf{I}))$ for some fixed variance schedule $\beta_1, \ldots, \beta_T \in (0, 1)$.

The sampling process in diffusion model starts with $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, that is iteratively refined to obtain the fully denoised sample $\mathbf{x}_0$.

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

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, f_θ is parameterized as neural network that predicts noise in $\mathbf{x}_t$ in the forward process. We will refer to the predicted noise $\mathbf{n}_t = f_\theta(\mathbf{x}_t, t)$ as predicted perturbation.

Differential Privacy (DP). Differential privacy implies that the outputs of an algorithm do not change much (measured by the privacy budget ε) across two neighboring datasets D and D', where we can go from D to D' by adding, removing, or replacing (remove + add) a single element. We provide the definition below.

Definition 2.1 (Differential Privacy)**.** A randomized mechanism $\mathcal{M}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ preserves (ε, δ)-differential privacy iff for any two neighboring datasets $D, D' \in \mathcal{D}$ and for any subset $S \subseteq \mathcal{R}$ we have $\Pr[\mathcal{M}(D) \in S] \leq e^\varepsilon \Pr[\mathcal{M}(D') \in S] + \delta$.

Thus, differential privacy requires that for all adjacent datasets D, D', the output distribution $\mathcal{M}(D)$ and $\mathcal{M}(D')$ are close, where the closeness is measured by the parameters ε and δ. The parameter ε is a 'privacy budget': as ε increases, our method is able to produce samples that reveal more information about the training data.

3. Methods

3.1. Quantifying the privacy risks of diffusion models

We first define the threat model used in this work. We consider a system that provides black-box query access[1] to the generative model, that was trained on potentially sensitive private data. The attacker's goal is to extract information about the underlying private training data of the model from samples that they generate by querying the model. Carlini et al. (2023) proposes an example of this attacker who queries a diffusion model a number of times with a specific prompt, and with some probability can induce the model to generate a copy of an image from the training dataset. The goal of the defense is to prevent the attacker from reconstructing, extracting, or otherwise learning anything about the private data, while generating high fidelity samples.

Definition 3.1 (Private generation)**.** A generative model G trained on datasets D is a (ε, δ)-DP generator if for all neighboring datasets D and D', all subset S of objects and all conditions c, the sampled images o satisfy:

$$\Pr_{o \sim G(c;D)}[o \in S] \leq e^\varepsilon \Pr_{o \sim G(c;D')}[o \in S] + \delta.$$

Consider a system that satisfies Definition 3.1. In such a system, the attacker will not be able to recover an image X from the private data by using a single instance generated by the model. The generated sample also cannot violate the copyright protections of an image in the model's training set, for the same reasons. Prior work (Ghalebikesabi et al., 2023) satisfies Definition 3.1 by training generative models with DP-SGD (Abadi et al., 2016) so that images generated by the final model are also DP. However, this is not the only way to satisfy Definition 3.1. We now introduce a DP sampling method that satisfies Definition 3.1 without DP-SGD.

3.2. Differentially private sample generation from diffusion models

At a high level, we implement the sample and aggregate framework (Nissim et al., 2007) to ensure DP by bounding the sensitivity of the predictions of the diffusion model.

Definition 3.2 (Sample and Aggregate (Nissim et al. (2007)))**.** Let the sample and aggregate framework be defined as the following. Given a function f and database D, we randomly partition the database into k disjoint subsets, where k is a user-chosen parameter. Let these small databases of size $|D|/k$ be $\{D_0, \ldots, D_{k-1}\}$. We evaluate f on $\{D_0, \ldots, D_{k-1}\}$ to obtain predictions $z_0, \ldots, z_k \in S$

[1] Black-box query access is often provided through APIs to provide users with access to generative models (Ramesh et al., 2022) for multiple reasons, e.g., the model is too large for users to run or the developer does not want to release the model weights.

where S is the support of f. We explicitly bound the ℓ_2 norm of these predictions to a scalar $C/2$. Therefore the sensitivity for one prediction z_i is C because in the worst case when we add, remove, or replace a datapoint in D_i the prediction will change from $C/2$ to $-C/2$. We aggregate the bounded predictions by averaging $z = 1/k \sum_k z_i$, reducing the sensitivity to C/k. We finally add noise calibrated to the sensitivity using some mechanism M. If M is (ε, δ)-DP, then sample and aggregate is (ε, δ)-DP.

We also consider following two formulations of sampling process which are equivalent for image generation but differ significantly in their privacy analysis across steps.

Formulation-A:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}} f_\theta(\mathbf{x}_t, t) + \sigma_t \mathbf{z}$$

Formulation-B:

$$\mathbf{x}_{t-1} = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\tilde{\mathbf{x}}_0 + \sigma_t \mathbf{z}$$

While the underlying diffusion model predicts $f_\theta(\mathbf{x}_t, t)$, $\tilde{\mathbf{x}}_0$ in formulation-B can be easily derived with a simple algebric operation, i.e., $\tilde{\mathbf{x}}_0 = \left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t} f_\theta(\mathbf{x}_t, t)\right) / \sqrt{\bar{\alpha}_t}$

Our proposed approach makes the following crucial changes to the training and sampling process of diffusion models.

Training a non-private ensemble. Given a dataset D, we partition it into k disjoint subsets $\{D_0, \ldots, D_{k-1}\}$ of size $|D|/k$, and non-privately train k distinct diffusion models $f_\theta^{(i)}, i \in [k]$ on each of these subsets D_i. These k distinct non-private DDPM models form an ensemble of diffusion models that is used for DP-enabled sampling.

Sampling privately from the non-privately trained ensemble. At the start of the sampling process, we sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and iteratively refine it to generate $\mathbf{x}_0$. We modify the individual step of the sampling process (Equation 1) by clipping and averaging the model prediction, which is $f_\theta(\mathbf{x}_t, t)$ in formulation-A and $\tilde{\mathbf{x}}_0$ in formulation-B. For example, in formulation-A, we modify the update step as following.

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \bar{f}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z} \quad (2)$$

$$\bar{f}_\theta(\mathbf{x}_t, t) = \frac{1}{k}\sum_{i=0}^{k-1} \texttt{clip}_{C/2}\left(f_\theta^{(i)}(\mathbf{x}_t, t)\right) \quad (3)$$

where $\texttt{clip}_{C/2}(\mathbf{v}) :\mapsto \min\left\{1, \frac{C/2}{\|\mathbf{v}\|_2}\right\} \cdot \mathbf{v} \in \mathbb{R}^d$. Our proposed modification simply clips and averages the predicted perturbation from each non-private model and keeps all other components intact. The hyperparameter C bounds the sensitivity of $\mathbf{x}_t$ to predictions from individual models. Our formulation is motivated by the sample and aggregate framework (Nissim et al., 2007).

Theorem 3.3 (Privacy). *Our method is $(\varepsilon, \delta(\varepsilon))$-DP, where*

$$\delta(\varepsilon) = \Phi\left(-\frac{\varepsilon}{\mu} + \frac{\mu}{2}\right) - e^{\varepsilon}\Phi\left(-\frac{\varepsilon}{\mu} - \frac{\mu}{2}\right)$$

$$\mu_a = \sqrt{\sum_t \left(\frac{C}{k} / \frac{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}\sigma_t}{1-\alpha_t}\right)^2} \quad \mu_b = \sqrt{\sum_t \left(\frac{C}{k} / \frac{1-\bar{\alpha}_t\sigma_t}{\sqrt{\bar{\alpha}_{t-1}}\beta_t}\right)^2}$$

For formulations A and B respectively

Proof. See Appendix 6. □

Better privacy guarantees with ensembling of sampling mechanisms. In each sampling formulation, the relative standard deviation of noise ($\mathbf{z}$) varies with timesteps. Higher noise standard deviation implies lower privacy expenditure (ε) per sampling step. We observe a trade-off between both mechanisms over timesteps and opt to switch from sampling-B to sampling-A when the latter leads to lower privacy expenditure (fig. 2).

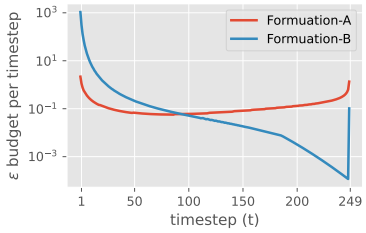
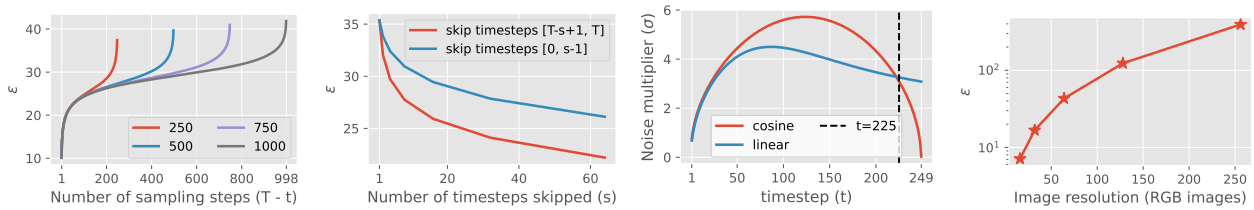


Figure 2. We observe that the privacy expenditure at each sampling step in both diffusion process formulations provides a trade-off, where switching from formulation-B to formulation-A near the end of sampling process reduces overall privacy expenditure.

4. Experimental results

Setup. We train convolutional UNet architecture based diffusion models on MNIST (Deng, 2012) and CIFAR-10 (Krizhevsky et al., 2009). We provide all training hyperparameters in Table 1 in the appendix. We consider 250 sampling steps to preserve image fidelity and enable faster sampling over 1000 sampling steps. By default, we use the linear noise schedule. We set the ensemble size to 100 models on both datasets. We set the clipping norm to 35 and 55, respectively, as a loose upper bound that avoids clipping any predictions during sampling to avoid introducing bias. To achieve better generalization on CIFAR-10, we finetune diffusion models pretrained on 32×32 resolution ImageNet-blurred (Chrabaszcz et al., 2017; Yang et al., 2022) dataset.

We find that both $t = T - 1$ and $t = 0$ sampling steps add noise with a very small noise multiplier (σ) in formulation-A, while near $t = 0$ sampling steps add very small noise in formulation-B. Since we use formulation-B at the start of sampling process, we only skip steps near end of sampling, i.e., denoise with a diffusion model trained on public data. We investigate the effect of skipping later in this section.



(a) (b) (c) (d)

Figure 3. **Disentangling the effect of different factors on our privacy bounds in formulation-A.** (a) *Length of sampling process.* Shorter schedules provides faster sampling and better ε (b) *Skipping timesteps.* Leads to large reduction in ε (c) *Effect of noise schedule.* When skipping last 25 timesteps, cosine schedule provides lower ε (d) *Effect of image resolution.* Modelling higher resolution datasets leads to higher ε. We conduct this analysis with Formulation-A of diffusion process and cosine noise schedule, since cosine noise schedule performs better with this formulation (subfig. c).

Using pretrained models. We use ImageNet-blurred[2] (Yang et al., 2022) and Fashion-mnist (Xiao et al., 2017) as public datasets for CIFAR-10 and MNIST datasets, respectively. We make use of public datasets and models pretrained on these datasets in two scenarios: 1) we finetune the pretrained models on disjoint subsets of the private dataset (only for CIFAR-10), 2) we use pretrained models as public models to denoise a fraction of steps near the end of sampling process (for both CIFAR-10 and MNIST).

Disentangling the effect of diffusion process parameters on privacy guarantees.

Length of diffusion process (Figure 3a). Short diffusion process during sampling enables faster sampling. We ablate the number of sampling steps and measure the cumulative epsilon as the sampling process progresses. With shorter schedules, the noise multiplier magnitude becomes smaller leading to higher ε for individual steps. The cumulative ε for all steps tends to be better for the shorter 250 steps schedule than the longer 1000 steps schedule.

Skipping timesteps to improve ε (Figure 3b). Figure 3a clearly shows that both early ($t \to 0$) and later timesteps ($t \to T$) have the highest privacy cost for formulation-A. We recommend replacing the denoising model for these steps with a public model to significantly reduce the privacy cost.

Effect of noise schedules (Figure 3c). While the cosine noise schedule adds more noise than the linear schedule for most timesteps, it adds significantly less noise at the start of sampling in formulation-A. In contrast, linear noise schedule adds more noise in formulation-B (Figure 6).

Effect of Image resolution (Figure 3d). An increase in image resolution significantly increases ε, e.g., ε increases from 16.7 to 48.4 when increasing image resolution from 32×32 to 64×64 pixels in formulation-A (similar trend in formulation-B). This is because the ℓ_2 norm of predicted perturbation scales linearly with image resolution, and as per Theorem 3.3 ε increases as the sensitivity increases.

Analyzing synthetic images generated by our approach. We first experimentally validate the effect of skipping timesteps with a public model. Based on our analysis in (Figure 3b, 3c), we skip 25 initial and 25 final timesteps when only sampling with formulation-A, and 50 final steps when using formulation-A in combination with formulation-B. We find that doing so doesn't degrade the fidelity of generated synthetic images (Figure 7). This demonstrates that, surprisingly, the steps with the highest privacy cost are actually the least dependent on the data. Next, we visualize the images generated by state-of-the-art DP-training (Ghalebikesabi et al., 2023) and our proposed DP-generation approach.

5. Discussion and Limitations

We demonstrate differentially private generation of high fidelity samples from diffusion based generative models. Our approach provides privacy guarantees by sampling from an ensemble of diffusion models, where individual models are trained on small subsets of the training data. Critical factors in the success of our approach in generating high fidelity samples are transfer learning from pretrained models and the few-shot learning capability of diffusion models.

While our approach enjoys multiple benefits over DP training, such as low training compute cost and the ability to leverage non-private training techniques, it provides a weaker threat model. In particular, DP training enables generating synthetic datasets for downstream analysis without additional privacy cost. However, DP generation requires composing privacy costs for each generation; the total privacy cost quickly becomes unreasonable for large datasets. Our sampling cost is also higher than DP training, though easily parallelized, as we require sampling from an ensemble of models. The threat model of private generation is analogous to private prediction (Dwork & Feldman, 2018; Papernot et al., 2017; 2018) in the context of classification. Specifically, we assume that downstream users cannot view the model, do not combine information across samples, and do not share their results with each other or collude.

[2]ImageNet-blurred dataset blurs faces of individuals in the popular ImageNet-1K dataset.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016. 2, 7

Bu, Z., Wang, Y.-X., Zha, S., and Karypis, G. Differentially private optimization on large model at small cost, 2022. 9

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023. 1, 2

Chrabaszcz, P., Loshchilov, I., and Hutter, F. A downsampled variant of imagenet as an alternative to the cifar datasets, 2017. 3

Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 3

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Neural Information Processing Systems (NeurIPS)*, 2021. 1

Dong, J., Roth, A., and Su, W. J. Gaussian differential privacy, 2019. URL https://arxiv.org/abs/1905.02383. 7

Dwork, C. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052 of *Lecture Notes in Computer Science*, pp. 1–12. Springer Verlag, July 2006. ISBN 3-540-35907-9. URL https://www.microsoft.com/en-us/research/publication/differential-privacy/. 1

Dwork, C. and Feldman, V. Privacy-preserving prediction, 2018. 4, 9

Gaboardi, M., Honaker, J., King, G., Murtagh, J., Nissim, K., Ullman, J., and Vadhan, S. Psi ({\Psi}): a private data sharing interface. *arXiv preprint arXiv:1609.04340*, 2016. 9

Ghalebikesabi, S., Berrada, L., Gowal, S., Ktena, I., Stanforth, R., Hayes, J., De, S., Smith, S. L., Wiles, O., and Balle, B. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*, 2023. 2, 4, 8, 9

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

Krizhevsky, A. et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 3

Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, pp. 75–84, 06 2007. doi: 10.1145/1250790.1250803. 2, 3

Papernot, N., Abadi, M., Úlfar Erlingsson, Goodfellow, I., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data, 2017. 4, 9

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Úlfar Erlingsson. Scalable private learning with pate, 2018. 4, 9

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. 1

Royer, J., Rodríguez-Cruces, R., Tavakol, S., Larivière, S., Herholz, P., Li, Q., Vos de Wael, R., Paquola, C., Benkarim, O., Park, B.-y., et al. An open mri dataset for multiscale neuroscience. *Scientific Data*, 9(1):569, 2022. 1

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015. 1

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models, 2022. 1

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 245–248, 2013. doi: 10.1109/GlobalSIP.2013.6736861. 7

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Neural Information Processing Systems (NeurIPS)*, 2019. 1

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 1

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE CVPR*, volume 7, pp. 46. sn, 2017. 1

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 4

Yang, K., Yau, J. H., Fei-Fei, L., Deng, J., and Russakovsky, O. A study of face obfuscation in imagenet. In *International Conference on Machine Learning*, pp. 25313–25330. PMLR, 2022. 3, 4

6. Theory

Proof of privacy. We prove the privacy guarantee of our method, and introduce any necessary notation along the way. We prove the privacy guarantee for Formulation A, where the only modification for Formulation B is simply in the ξ_t term.

Theorem 6.1 (GDP of Gaussian Mechanism (Dong et al., 2019))**.** *Let the Gaussian mechanism that operates on function f by defined as $M(D) = f(S) + N(0, \Delta^2/\mu^2)$ where Δ is the sensitivity of f. Then, M is μ-GDP.*

Proposition 6.2. *Equation (2) is $\frac{C}{k}/\frac{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}\sigma_t}{1-\alpha_t}$-GDP for Formulation A.*

Proof. The sensitivity according to Definition 3.2 is $\frac{C}{k}$, because we clip each model's prediction norm to $C/2$ (the total l_2 norm for change of the prediction for a different datapoint is C) and take the average prediction across k models. To analyze the noise added to each model, we first rewrite Equation (2) as

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\bar{f}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z} \quad (4)$$

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{f}_\theta(\mathbf{x}_t, t)\right) \quad (5)$$

$$\hat{f}_\theta(\mathbf{x}_t, t) = \bar{f}_\theta(\mathbf{x}_t, t) + \frac{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}\sigma_t}{1-\alpha_t}\mathbf{z} \quad (6)$$

$$\xi_t = \frac{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}\sigma_t}{1-\alpha_t} \text{Formulation A} \quad (7)$$

$$\xi_t = \frac{1-\bar{\alpha}_t\sigma_t}{\sqrt{\bar{\alpha}_{t-1}}\beta_t} \text{Formulation B} \quad (8)$$

$$\mathbf{z_t} \sim N(0, \xi_t^2) \quad (9)$$

$$\hat{f}_\theta(\mathbf{x}_t, t) = \bar{f}_\theta(\mathbf{x}_t, t) + \mathbf{z_t} \quad (10)$$

Where in Equation (6) we applied the post-processing property of DP. Informally, this means that if a single iteration of our sampling method, that clips the score generated by the diffusion model and adds noise to it, is DP, then any post-processing applied to the score is also DP. This means that we can use our newly DP score to generate $\mathbf{x}_{t-1}$ without paying any additional privacy cost. Therefore we only need to analyze the privacy cost of Equation (6), that we finally write as Equation (10). We can see that this is a single instantiation of the Gaussian mechanism. Now we can write the equivalent value of $\mu_t = \frac{C}{k}/\xi_t$, and apply Theorem 6 to finish the proof.

$$\left(\frac{C}{k}\right)^2/\mu_t^2 = \xi_t^2$$

$$\mu_t^2 = \left(\frac{C}{k}\right)^2/\xi_t^2$$

$$\mu_t = \frac{C}{k}/\xi_t$$

□

Proposition 6.3. DP-DDPM *is* $\sqrt{\sum_t \left(\frac{C}{k}/\xi_t\right)^2}$*-GDP.*

Proof. First recall the composition of Gaussians under GDP $\mu = \sqrt{\sum_t \mu_t^2}$ (Dong et al., 2019). We can plug in the expression for μ_t.

$$\mu_t = \frac{C}{k}/\xi_t$$

$$\mu = \sqrt{\sum_t \left(\frac{C}{k}/\xi_t\right)^2}$$

□

Corollary 6.4 (Corollary 2.13 (Dong et al., 2019))**.** *A mechanism is μ-GDP if and only if it is $(\varepsilon, \delta(\varepsilon))$-DP for all $\varepsilon > 0$, where*

$$\delta(\varepsilon) = \Phi\left(-\frac{\varepsilon}{\mu} + \frac{\mu}{2}\right) - e^\varepsilon \Phi\left(-\frac{\varepsilon}{\mu} - \frac{\mu}{2}\right). \quad (11)$$

Now we can apply the GDP to DP conversion to obtain the final (ε, δ)-DP guarantee for our full method.

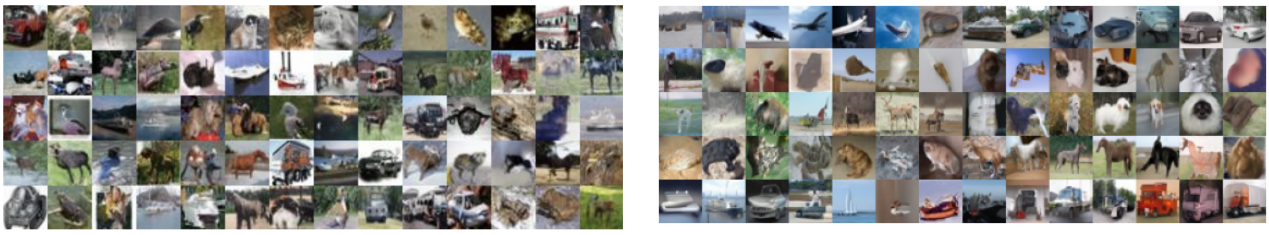
Theorem 6.5 ((Reproduction for clarity))**.** DP-DDPM *is $(\varepsilon, \delta(\varepsilon))$-DP, where*

$$\delta(\varepsilon) = \Phi\left(-\frac{\varepsilon}{\mu} + \frac{\mu}{2}\right) - e^\varepsilon \Phi\left(-\frac{\varepsilon}{\mu} - \frac{\mu}{2}\right)$$

$$\mu = \sqrt{\sum_t \left(\frac{C}{k}/\frac{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}\sigma_t}{1-\alpha_t}\right)^2}$$

7. Limitations

Differentially Private Stochastic Gradient Descent (DP-SGD) (Song et al., 2013; Abadi et al., 2016) is the standard privacy-preserving training algorithm for training neural networks on private data. DP-SGD clips per-sample gradients and adds Gaussian noise to them, introducing bias and variance into SGD and therefore degrading utility. For



(a) Differentially private training ($\varepsilon = 10.0$) (b) Differentially private sampling ($\varepsilon = 7.3$)

Figure 4. Comparing samples from differentially private *training* (Ghalebikesabi et al., 2023) and our proposed differentially private *generation* approach.

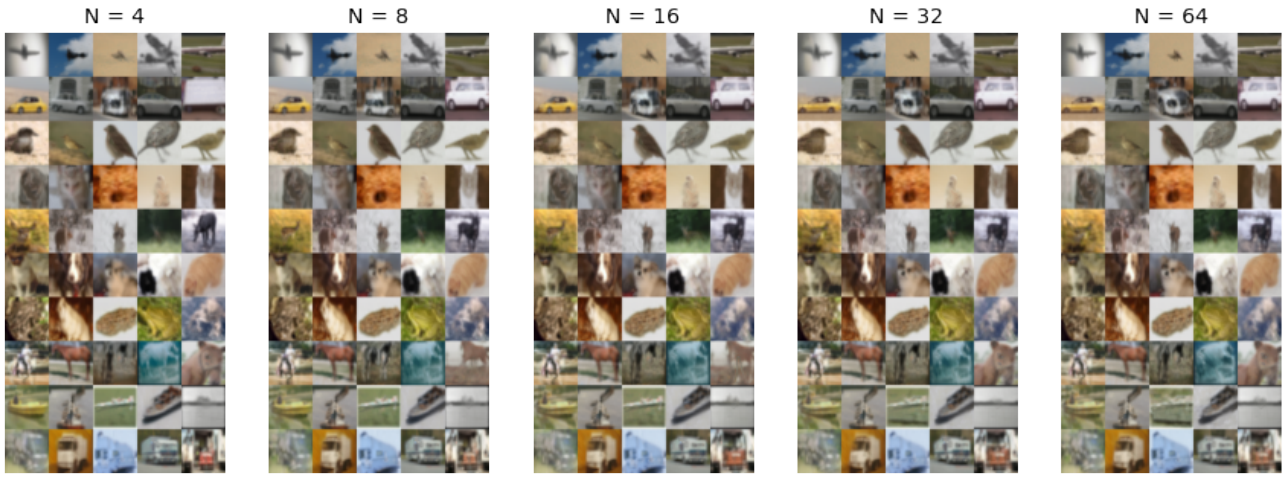
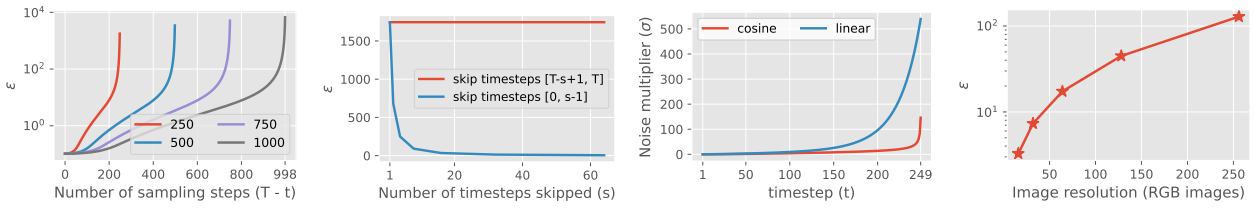


Figure 5. We measure the effect of increasing the number of models in the ensemble on image quality. Each model is the ensemble is trained on 500 images, i.e., $1/100$ of the CIFAR-10 dataset. Even after eight models, the sample quality of the ensemble starts saturating.



(a) (b) (c) (d)

Figure 6. **Disentangling the effect of different factors on our privacy bounds in formulation-B.** (a) *Length of sampling process.* Shorter schedules provides faster sampling and better ε (b) *Skipping timesteps.* Leads to large reduction in ε (c) *Effect of noise schedule.* When skipping last 25 timesteps, cosine schedule provides lower ε (d) *Effect of image resolution.* Modelling higher resolution datasets leads to higher ε. We conduct this analysis with Formulation-B of diffusion process and linear noise schedule, since linear noise schedule performs better with this formulation (subfig. c).

example, the state-of-the-art in differentially private diffusion modeling on CIFAR10 with ImageNet pretraining data achieves an FID of 7.9 at $\varepsilon = 32$.

DP-SGD creates such large utility degradation because it provides an extremely strong guarantee. Even if an unlimited number of adversaries collude together to deanonymize the model, with knowledge of all model gradients, and infinite computational power, DP-SGD still provides an upper bound on how much information can be gained by observing the model. This threat model is far too strong in practice because it defends against an attack that is many orders of magnitude more powerful than current attacks.

Our threat model for private generation follows the threat model of private prediction Gaboardi et al. (2016); Dwork & Feldman (2018); Papernot et al. (2017; 2018) that makes realistic assumptions about adversaries' information and resources. Specifically, we assume that users cannot view the model parameters, do not share their results with each other and do not collude in coordinated attacks on individual training samples. This allows us to independently spend privacy budget for each generated sample, leading to improved utility without compromising data privacy.

For comparison, we also provide the definition for DP training of generative model.

Definition 7.1 (Private training of generative models)**.** A learning algorithm L that operates on dataset D and outputs a generative model G is said to be (ε, δ)-DP if for all neighboring datasets D and D', all subset S of objects we have:

$$\Pr_{G\sim L(D)}[G \in S] \leq e^{\varepsilon} \Pr_{G\sim L(D')}[G \in S] + \delta.$$

Note that any set of images generated by a DP-trained generative model is also DP with the same (ε, δ) by postprocessing.

Challenges of private learning. DP training of a diffusion model can be exceedingly computationally intensive. We estimate the compute required by Ghalebikesabi et al. (2023) to be roughly $1000\times$ that of training a non-private diffusion model. One of the reasons why DP training is so computationally intensive is because models are so hard to train with DP. The best practices for training with DP (Ghalebikesabi et al., 2023) are to take many small steps, that is, we need to modify the learning rate and training time from non-private learning. And each step needs to use a large batch (indeed, optimally we should use the full batch) so these small steps are very slow because we will have to accumulate gradients. And we have to compute the per-sample gradient of each datapoint, that can by itself increase training time by $10\times$ (Bu et al., 2022). Furthermore DP training requires an immense amount of hyperparameter tuning because adding noise to the gradients at each step is so challenging. In order to get an accurate estimate of the noisy gradient, Ghalebikesabi et al. (2023) compute the gradients on as many as 256 augmentations for each datapoint.

By contrast, we are able to harness the few-shot learning capabilities of diffusion models to train k models on subsets of the original dataset of size $|D|/k$, and each model is trained for the same number of epochs on the $k\times$-smaller subset as we would train on the original dataset. Therefore the compute requirement during training of our method is the same as non-private training.

However, sampling from a DP-trained diffusion model is identical to non-private sampling, whereas sampling from an ensemble of k models requires running each sampling process in parallel to run in the same time as non-private sampling.

What does differentially private generation protect? As is clear from the definition, Differentially private generation protects the privacy of each individual generated sample. Note that this notion is strictly weaker than differentially private training of generative models. This is simply because the generation of any generative model that is trained by a differentially private learning algorithm is also differentially private by post-processing.

Of course, differentially private training could also achieve these goals, but the current state of research on differential private training of generative models suggests that training a good generative model with differential privacy guarantees is a really hard task. Private generation could be a significantly easier task and could provide more meaningful guarantees against the threat models of interest. As we have seen in this work, we can achieve private generation with very small modifications to the existing diffusion processes. By contrast, training a private diffusion model from scratch might require changing the architecture and optimization algorithm all together.

We finally note that the notion of private generation is analogous to the notion of private prediction (Dwork & Feldman, 2018; Papernot et al., 2017; 2018) in the context of classification. Private prediction requires that the inference procedure on any given input is differentially private. In comparison, differentially private training of a classification model would guarantee that the weights of a classifier are differentially private.

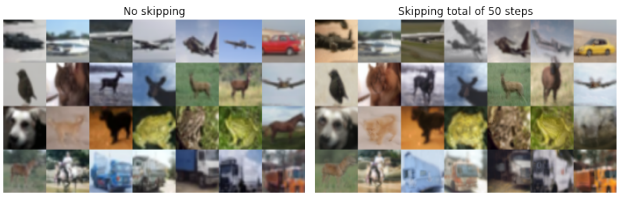


Figure 7. **Minimal effect of skipping timesteps on image quality.** We skip 25 steps at the start and 25 steps at the end of a 250 timestep sampling process. We use a diffusion model pretrained on ImageNet, that we treat as a publicly available dataset, to denoise images in skipped steps. We find that skipping a total of 50 timesteps does not significantly degrade image quality, but does incur semantic changes (such as change in color) compared to the baseline where no steps are skipped.

Table 1. Details of hyperaparameters used in training diffusion models on both MNIST and CIFAR-10 datasets. We use ImageNet-blurred and Fashion-mnist dataset in pretraining for CIFAR-10 and MNIST dataset, respectively.

	MNIST	CIFAR-10
Diffusion steps	1000	1000
Noise Schedule	linear	linear
Channels	64	64
Depth	3	3
Channels multiple	1,2,2,2	1,1,2,3,4
Heads Channels	64	64
Attention resolution	32,16,8	32,16,8
BigGAN up/downsample	✓	✓
Attention pooling	✓	✓
Weight decay	0.01	0.01
Dropout	0.3	0.3
Batch size	128	128
Epochs	500	500
Learning rate	1e-4	5e-4
Public dataset	FMNIST	ImageNet-Blurred
Pretrained model	None	ImageNet-Blurred