PrivImage: Differentially Private Synthetic Image Generation using Diffusion Models with Semantic-Aware Pretraining

Kecen Li, *Institute of Automation, Chinese Academy of Sciences and University of Chinese Academy of Sciences;* Chen Gong, *University of Virginia;* Zhixiang Li, *University of Bristol;* Yuzhong Zhao, *University of Chinese Academy of Sciences;* Xinwen Hou, *Institute of Automation, Chinese Academy of Sciences;* Tianhao Wang, *University of Virginia*

https://www.usenix.org/conference/usenixsecurity24/presentation/li-kecen

This paper is included in the Proceedings of the 33rd USENIX Security Symposium.

August 14–16, 2024 • Philadelphia, PA, USA

978-1-939133-44-1

Open access to the Proceedings of the 33rd USENIX Security Symposium is sponsored by USENIX.

PRIVIMAGE: Differentially Private Synthetic Image Generation using Diffusion Models with Semantic-Aware Pretraining

Kecen Li[1,4,†,‡], Chen Gong[2,†], Zhixiang Li[3], Yuzhong Zhao[4], Xinwen Hou[1], and Tianhao Wang[2]

[1]Institute of Automation, Chinese Academy of Sciences
[2]University of Virginia, [3]University of Bristol
[4]University of Chinese Academy of Sciences
{likecen2023, xinwen.hou}@ia.ac.cn, {chengong, tianhao}@virginia.edu,
ek22436@bristol.ac.uk, zhaoyuzhong20@mails.ucas.ac.cn

Abstract

Differential Privacy (DP) image data synthesis, which leverages the DP technique to generate synthetic data to replace the sensitive data, allowing organizations to share and utilize synthetic images without privacy concerns. Previous methods incorporate the advanced techniques of generative models and pre-training on a public dataset to produce exceptional DP image data, but suffer from problems of unstable training and massive computational resource demands. This paper proposes a novel DP image synthesis method, termed PRIVIMAGE, which meticulously selects pre-training data, promoting the efficient creation of DP datasets with high fidelity and utility. PRIVIMAGE first establishes a semantic query function using a public dataset. Then, this function assists in querying the semantic distribution of the sensitive dataset, facilitating the selection of data from the public dataset with analogous semantics for pre-training. Finally, we pre-train an image generative model using the selected data and then fine-tune this model on the sensitive dataset using Differentially Private Stochastic Gradient Descent (DP-SGD). PRIVIMAGE allows us to train a lightly parameterized generative model, reducing the noise in the gradient during DP-SGD training and enhancing training stability. Extensive experiments demonstrate that PRIVIMAGE uses only **1%** of the public dataset for pre-training and 7.6% of the parameters in the generative model compared to the state-of-the-art method, whereas achieves superior synthetic performance and conserves more computational resources. On average, PRIVIMAGE achieves 6.8% lower FID and 13.2% higher Classification Accuracy than the state-of-the-art method. The replication package and datasets can be accessed online[1].

1 Introduction

Currently, various deep learning models increasingly rely on sensitive personal data during their training process across some crucial fields, involving healthcare [13, 21, 22], finance [27], and social networks [35]. With the rising awareness of personal privacy and data protection [20, 46, 61, 81], it is imperative to adopt specialized methods to enhance the security of training data, ensuring the safeguarding of users' personal information against misuse and breaches of privacy.

Synthetic image generation with Differential Privacy (DP) [16] provides a potential answer to this question. DP image synthesis [6, 10, 15, 18, 80] aims to generate synthetic images that resemble real data while ensuring the original dataset remains private. With DP image synthesis, organizations can share and utilize synthetic images, facilitating various downstream tasks without privacy concerns.

Existing methods. Diffusion models have demonstrated potential in DP image synthesis [15, 18, 31, 54]. Dockhorn et al. [15] advocated for training diffusion models using DP-SGD [1], a widely adopted method for training models satisfying DP. Drawing inspiration from the success of pre-training and fine-tuning across many challenging tasks in computer vision [5, 45, 73, 84], Sabra et al. [18] proposed to first pre-train the diffusion models on a public dataset, and then fine-tune them on the sensitive dataset. They attained state-of-the-art (SOTA) outcomes on datasets more intricate than those used by prior methods.

Our Proposal. We highlight that the dataset with a semantic distribution similar to the sensitive dataset is more suitable for pre-training. Building on this observation, we present PRIVIMAGE, an end-to-end solution to meticulously and privately select a small subset of the public dataset whose semantic distribution aligns with the sensitive one, and train a DP generative model that significantly outperforms SOTA solutions.

PRIVIMAGE consists of three steps. Firstly, we derive a foundational semantic query function from the public dataset. This function could be an image caption method [74, 76] or a straightforward image classifier. Secondly, PRIVIMAGE uses the semantic query function to extract the semantics of each sensitive image. The frequencies of these extracted semantics then shape a semantic distribution, which can be used to select data from the public dataset for pre-training. To

[†]Equal contribution.
[‡]Work done as a remote intern at UVA.
[1]https://dp-image-syn.github.io/privimage/

make this query satisfy DP, we introduce Gaussian noise to the queried semantic distribution. Finally, we pre-train image generative models on the selected dataset and fine-tune pre-trained models on the sensitive dataset with DP-SGD [1].

Compared to previous studies [18, 49], PRIVIMAGE employs a more compact public dataset for pre-training, which conserves not only computational resources and time but also achieves competitive synthesis performance in terms of both fidelity and utility.

Evaluations. We conduct comprehensive experiments to verify the effectiveness of PRIVIMAGE. By utilizing just **1%** of the `ImageNet` [14] dataset for pre-training, we can achieve superior synthesis performance compared to existing solutions that use the full dataset for pre-training. Specifically, under three privacy budgets $\varepsilon = \{1, 5, 10\}$, PRIVIMAGE outperforms all baselines in terms of utility and fidelity of synthetic images. On average, the FID and Classification Accuracy of the downstream classification task of synthetic images from PRIVIMAGE is 30.1% lower and 12.6% higher than the SOTA method, i.e., PDP-Diffusion [18]. Besides, PRIVIMAGE achieves competitive high-quality image synthesis, while just costing 50% lower GPU memory and 48% lower running time compared to the SOTA method.

Our experiments further explore the factors that make the pre-training dataset selected by PRIVIMAGE effective. The results reveal that the success of PRIVIMAGE can be attributed to two aspects: (1) *Win at the starting.* PRIVIMAGE selects a pre-training dataset that more closely resembles the sensitive data than the entire public data. Before fine-tuning on the sensitive dataset, PRIVIMAGE inherently produces synthetic images with a data distribution more aligned with the sensitive data than those generated by existing methods. (2) *Reduce the size of the model.* Previous studies have demonstrated that as the dataset size decreases, the performance gap between large models and lightweight models narrows [29, 37]. Therefore, with the reduction in the size of the pre-training dataset, PRIVIMAGE enables us to well train generative models with fewer parameters, enhancing training efficiency. Additionally, we observe that the privacy budget used in selecting pre-training data is negligible, exerting minimal impact on the fine-tuning of generative models.

We also analyze the effects of hyper-parameters, specifically the selection ratio and model size, on our method. We observed that PRIVIMAGE, with a selection ratio of less than 50%, typically outperforms all baselines that use the entire public dataset. However, the performance of PRIVIMAGE declines as the selection ratio rises. When using over-parameterized models, PRIVIMAGE also experiences inconsistent performance. Our study highlights that *a larger pre-training dataset does not inherently lead to superior outcomes in DP image data synthesis*, emphasizing the need for constructing a more tailored pre-training dataset for DP image dataset synthesis.

In summary, our contributions are three-fold:

- We analyze the pre-training and fine-tuning paradigm for DP image synthesis and made the important observation that the distribution of the public data utilized should be similar to that of the sensitive data.
- We introduce PRIVIMAGE, which uses the semantic distribution of the dataset requiring protection to meticulously select pre-training data. The selected dataset more closely aligns with the sensitive data compared to the entire public dataset, thereby enabling efficient synthesis of DP image datasets with high fidelity and utility.
- Extensive experiments present that by utilizing only 1% of the public dataset for pre-training, we can significantly save computational resources and time. Specifically, the diffusion model in PRIVIMAGE involves a mere 7.6% of parameters in PDP-Diffusion, but achieves superior synthesis performance. PRIVIMAGE achieves SOTA results on `CIFAR-10` [40] and `CelebA` [52].

2 Background

This section introduces the concepts of Differential Privacy, delves into image generative models, and provides an overview of the DP image dataset synthesis techniques.

2.1 Differential Privacy

Differential privacy [17] is a privacy-preserving concept quantifying sensitive data disclosure.

Definition 2.1. *(Differential Privacy [17]) A randomized algorithm M satisfies (ε, δ)-differential privacy, where $\varepsilon > 0$ and $\delta > 0$, if and only if, for any two adjacent datasets D and D', it holds that,*

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

where O denotes the set of all possible outputs of the algorithm M. The privacy budget ε is a non-negative parameter that measures the privacy loss in the data. A smaller ε indicates better privacy. In this study, two datasets D, D' are deemed adjacent, denoted $D \simeq D'$, if $D = D' + x$ or $D' = D + x$, where $D' + x$ is the dataset derived by appending a data entry x to dataset D.

A popular mechanism is the Sub-sampled Gaussian Mechanism (SGM) [58]. Let $f : D_s \subseteq D \to \mathbb{R}^d$ be query function with sensitivity $\Delta_f = \max_{D \simeq D'} \|f(D) - f(D')\|_2$. SGM is parameterized with a sampling rate $q \in (0, 1]$ and noise standard deviation $\sigma > 0$, and is defined as,

$$SGM_{f,q,\sigma}(D) \triangleq f(S) + \mathcal{N}\left(0, \sigma^2 \Delta_f^2 \mathrm{I}\right)$$

where $S = \{x | x \in D$ selected independently with probability $q\}$ and we define $f(\emptyset) = 0$. We suppose that $\Delta_f^2 = 1$ for

simplicity, which can be easily controlled by changing the σ^2. Rényi Differential Privacy (Rényi DP) [58] is usually used to track the privacy cost of SGM.

Definition 2.2. *(Rényi DP [58]) Let* $D_\alpha(Y\|N) = \frac{1}{\alpha-1}\ln \mathbb{E}_{x\sim N}\left[\frac{Y(x)}{N(x)}\right]^\alpha$ *be Rényi divergence with* $\alpha > 1$*, a randomized mechanism* M *is said to be* (α,γ)*-RDP, if* $D_\alpha(M(D)\|M(D')) < \gamma$ *holds for any adjacent dataset* D*,* D'*.*

Theorem 2.1. *(RDP for SGM [58]) Let* p_0 *and* p_1 *denote the PDF of* $\mathcal{N}(0,\sigma^2)$ *and* $\mathcal{N}(1,\sigma^2)$ *respectively. A* $SGM_{M,q,\sigma}(D)$ *satisfies* (α,γ)*-RDP for any* γ *such that,*

$$\gamma \geq D_\alpha([(1-q)\,p_0 + q p_1 \,\|\, p_0]) \tag{1}$$

The above theorem shows that the privacy bound γ can be computed using the term $D_\alpha([(1-q)\,p_0+qp_1\,\|\,p_0])$.

In the domain of machine learning, the most popular approach to satisfy DP is DP-SGD [1], which modifies the standard Stochastic Gradient Descent (SGD) and adding Gaussian noise to the gradient of the model's parameters throughout training, and then updated model via,

$$\theta \leftarrow \theta - \eta\left(\frac{1}{|b|}\sum_{i\in b}\text{Clip}\left(\nabla\mathcal{L}(\theta,x_i),C\right) + C\mathcal{N}(0,\sigma^2 I)\right),$$

where η is the learning rate, $\nabla\mathcal{L}(\theta,x_i)$ is the gradient of the loss function $\mathcal{L}$ with respect to model parameters θ for the data point x_i in a randomly sampled batch b. $\text{Clip}(\nabla\mathcal{L},C)$ refers to a function that clips the gradient vector $\nabla\mathcal{L}$ such that its ℓ_2 norm under the constraint of C, and $\mathcal{N}(0,\sigma^2 I)$ is the Gaussian noise with the variance σ. DP-SGD ensures the model does not overly adapt to specific data points or retain uncommon details that might jeopardize privacy. We provide more details in Algorithm 1 of Section 3.4.

2.2 Image Generative Models

This subsection presents two of the most effective image generative models: diffusion models [69] and Generative Adversarial Networks (GANs) [23].

Diffusion Models: Diffusion Models [31, 59, 69] are a class of likelihood-based generative models that consist of two processes: (1) The *forward diffusion process* that progressively corrupts a clean image x_0 by adding Gaussian noise, which generates gradually noisier images $\{x_1,\ldots,x_T\}$, and T is the number of noising steps. (2) The *reverse process* that progressively denoise a noise to a clean image via a trainable neural network. The forward diffusion process between adjacent noisier images, i.e., x_{t-1} and x_t, follows a multi-dimensional Gaussian distribution, which is formulated as,

$$h(x_t\,|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathrm{I}\right)$$

where β_t regulates the variation of the noise distribution at each step and can be either fixed or learnable. We note $\bar{\alpha}_t := \prod_{s=1}^{t}(1-\beta_s)$. The likelihood between the clean image x_0 and noisier images in step t is formulated as,

$$h(x_t\,|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathrm{I}\right)$$

Therefore, we can sample x_t from $h(x_t\,|x_0)$ directly from x_0 in closed form instead of adding noise t times as,

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + e\sqrt{1-\bar{\alpha}_t}, e\sim\mathcal{N}(0,\mathrm{I})$$

The final objective of diffusion models is defined as by [31],

$$\mathcal{L}_{DM} := \mathbb{E}_{t\sim U(1,T),x_t\sim h(x_t|x_0),e\sim\mathcal{N}(0,\mathrm{I})}\|e - e_\theta(x_t,t)\|^2 \tag{2}$$

where e_θ is a denoising network parameterized with θ. Through minimizing Eq. (2), e_θ learns to predict the noise e of any noisy images x_t. Thus, we can use the noise $e_\theta(x_t,t)$ predicted by e_θ to denoise the noisy images [31, 69]. After being trained well, e_θ can be used to gradually denoise a random Gaussian noise to a clean image.

Generative Adversarial Nets (GAN): GAN [2, 23] is a classical generative model and has achieved great performance on image synthesis task [7, 38, 39]. GAN is composed of two networks, an image generator *Gen* and an image discriminator *Dis*. The image generator *Gen* receives a random noise vector and output an image. The image discriminator *Dis* receives an image and outputs a score, which indicates how real the input image is. The *Gen* is trained to generate more real images and the *Dis* is trained to distinguish whether its input image comes from the true dataset or is generated by the *Gen*. Mathematically, the objective function of GAN is,

$$\begin{aligned}\min_{Gen}\max_{Dis} V(Gen,Dis) &= \mathbb{E}_{x\sim q(x)}[\log Dis(x)] \\ &+\mathbb{E}_{z\sim p(z)}[\log(1-Dis(Gen(z)))]\end{aligned} \tag{3}$$

Where $q(x)$ and $p(z)$ are the distribution of the real image and noise vector respectively. After solving the minimax two-player game between *Gen* and *Dis* as Eq. (3), the well-trained *Gen* generates images of high fidelity with the input of noise.

2.3 DP Image Dataset Synthesis

With the rapid advancements in generative models and deep learning, a range of generative models have demonstrated remarkable image synthesis capabilities [23, 69]. In the DP dataset synthesis, a direct approach involves acquiring noisy gradient via DP-SGD [1] to update model parameters at each training iteration to ensure that the well-trained generative models satisfy DP [15, 18, 72].

DP-CGAN [72] is a representative method which applies DP-SGD to GANs. Since only *Dis* accesses the sensitive data, they sanitize the gradient of *Dis* using DP-SGD. DP-CGAN achieves good synthesis on `MNIST`, but not as good on larger datasets. In recent years, diffusion models have been a promising image generative model which beats GANs [68]. Therefore, DPDM [15] proposes to replace GANs with diffusion

Table 1: A summary of existing methods and PRIVIMAGE across various modules. 'GM' and 'DM' refer to the 'Generative Model' and 'Diffusion Models.' We record the trainable parameters of used models on CIFAR-10 or the number of used images in the pre-training dataset in parentheses. We detail the parameters of DP-CGAN when performing experiments on MNIST. However, DP-CGAN did not run experiments on CIFAR-10, as they failed to synthesize.

Method \ Module	Pre-training	GM
DP-CGAN [72]	-	GAN (6.6M)
DPDM [15]	-	DM (1.8M)
PDP-Diffusion [18]	ImageNet (1.3M)	DM (80M)
PRIVIMAGE+G	Selected Dataset (64K)	GAN (1.8M)
PRIVIMAGE+D	Selected Dataset (64K)	DM (1.8M)

models in DP-CGAN. In particular, the authors use a large number of images for training at one iteration (i.e., large batch size) and use lightly parameterized diffusion models, getting rid of the "curse of dimensionality" [15, 41, 67]. DPDM achieves state-of-the-art (SOTA) synthesis on MNIST [42] and FashionMNIST [75]. While DP image synthesis methods have successfully protected the privacy of the above naive image datasets, many researchers are still working to design more effective DP image synthesis techniques for complex datasets, such as CIFAR-10 [40] and CelebA [52].

As models and datasets grow increasingly complex, the pre-training and fine-tuning paradigm has become prevalent. PDP-Diffusion [18] proposes to exclusively pre-train the diffusion model on a public dataset and subsequently fine-tune it on the sensitive dataset, achieving SOTA image synthesis results on CIFAR-10 [40]. Whereas, public datasets, like ImageNet [14], are typically extensive and include over 1 million high-resolution images.

Table 1 gives a comparison of the existing methods, including DP-CGAN [72], DPDM [15], PDP-Diffusion [18], and two versions of PRIVIMAGE, namely PRIVIMAGE+G and PRIVIMAGE+D. Our method PRIVIMAGE, instantiated with GAN and DM, distinguishes itself from existing methods with a smaller dataset for pre-training and a lower-parameterized model while achieving better performance. The key idea is to align the semantic distribution of public and sensitive datasets, which will be described later.

3 Methodology

This paper focuses on exploring how to leverage the public dataset for more effective pre-training and proposes a novel DP image synthesis method, PRIVIMAGE. We divide PRIVIMAGE into three steps. Initially, we utilize the semantic labels from the public dataset to derive a semantic query function. Then, this function aids in querying the semantic distribution of the sensitive dataset, facilitating data selection for pre-training. Finally, we initiate pre-training of an image generative model using the selected data and then fine-tune this model on the sensitive dataset using DP-SGD.

Figure 1: Two images with their captions. The same and different semantics two images own are in green and red respectively. Although the two images differ a lot in pixel, they have the same semantics, ***man*** and ***dog***.

3.1 Motivation

Studies indicate that if there is minimal overlap between the distributions of pre-training and target dataset (which is the sensitive dataset in our studied problem), the effectiveness of pre-training diminishes [28, 64, 73]. Therefore, drawing inspiration from this understanding, we select the data that is "similar" to sensitive data from the public dataset for the pre-training of generative models to develop the DP image synthesis method. We select semantic distribution as a metric to gauge the similarity between the two datasets and provide our rationale for this choice as follows.

Semantics provide a high-level representation of images and have played a prominent role in various computer vision studies [50, 53, 78]. As illustrated in Figure 1, image semantics refers to the meaning and interpretation derived from an image. Unlike low-level features, such as color, texture, and edges that can be directly extracted from the image, semantics capture the "meaning" or "content" of an image, which often requires higher-level processing and understanding [70, 71]. Most public image datasets have their semantic labels, like caption [12], object [48], and category [14]. Since our focus is on generating low-resolution images, which typically have simple content (e.g., each CIFAR-10 image contains only a single object), we propose utilizing the category labels of the public dataset to query the semantic distribution of the sensitive dataset. This information can then be used to select more relevant data for pre-training.

3.2 Query Image Semantics

We aim to derive a semantic query function that takes an image as input and returns its semantics. In this study, we use the category labels from the public dataset to represent the semantics of sensitive images. The objective while training

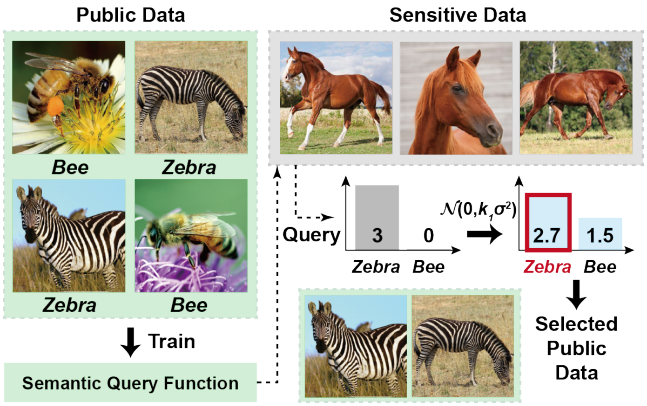

Figure 2: An example of using the semantic query function to retrieve the semantic distribution from the sensitive dataset. We first train the semantic query function using the public dataset. This function is used to obtain the semantic distribution of the sensitive dataset. To ensure privacy, we then incorporate the Gaussian noise into our query results.

a semantic query function Q is to minimize the difference between the predicted semantic labels and the true labels. This is achieved by minimizing a loss function $\mathcal{L}_Q$, e.g., the Cross-Entropy loss [56], which is defined as,

$$\mathcal{L}_Q := \mathbb{E}_{(x,\, s)\sim D_p}\left[\sum_{i=1}^{NS} -s_i \log Q_i(x)\right] \tag{4}$$

where D_p is the public dataset. Each pair in D_p consists of an image x and its semantics labels s. NS indicates the number of possible semantics in D_p. $s_i = 1$ indicates that the image has the i-th semantic label, whereas $s_i = 0$ denotes the absence of this semantic. $Q_i(x)$ is a scalar value between 0 and 1, representing the probability that image x is associated with the semantic s_i, as determined by the semantic query function Q. Through minimizing Eq. (4), we train the semantic query function Q to extract semantics from images.

After training, for each sensitive image, the semantic query function Q can predict the probability that the image has the i-th semantic s_i. Given that each image potentially can have multiple semantic labels, we select the top-k_1 probable semantics as predicted by Q to represent the semantics of the input-sensitive image, where k_1 is a hyper-parameter. A higher k_1 suggests considering a greater number of semantics in images.

3.3 Semantic Distribution Query

PRIVIMAGE aims to select data from the public dataset for pre-training, ensuring its semantic distribution is similar to that of the sensitive dataset, without revealing any sensitive information. Towards this goal, after querying the semantics of each piece of sensitive data, this section delves into constructing a semantic distribution *SD* for the sensitive dataset. Given that the number of semantics from the public dataset is constant, it intuitively follows that we can represent the semantic distribution as the frequency distribution of semantics derived from the sensitive dataset. Specifically, *SD* can be conceptualized as a dictionary where the keys represent the semantics of the public dataset and the values denote the frequency of each semantic within the sensitive dataset. To safeguard the privacy of the sensitive dataset, as presented in Theorem 3.1, we introduce Gaussian noise to the retrieved semantic distribution, in line with the RDP (which is described in Theorem 2.1), ensuring the query results adhere to differential privacy.

Theorem 3.1. *The semantic distribution SD has global sensitivity* $\Delta_{SD} = \sqrt{k_1}$*, where* k_1 *is the number of semantics we query from each sensitive image. For any* $\alpha > 1$*, incorporating noise* $\mathcal{N}\left(0, k_1\sigma_2^2 \mathrm{I}\right)$ *into the semantic distribution SD makes the query results satisfies* $(\alpha, \alpha/\left(2\sigma_2^2\right)$*-RDP.*

We put the proof of Theorem 3.1 in Appendix **??**. After obtaining the sensitive semantic distribution, a straightforward way is to select data from the public dataset to obtain a pre-training dataset whose semantic distribution is close to the sensitive one. However, since the sensitive semantic distribution is noisy for incorporating the Gaussian noise, it is ineffective to focus on the semantics with low frequency, which is more easily affected by the added noise. Therefore, we focus on the semantics with high frequency. Specifically, we select the top-k_2 semantics based on their probabilities in the semantic distribution to represent the semantic description of the sensitive dataset. Finally, the public data, whose semantics are included in the semantic description queried from the sensitive dataset, are selected for pre-training. Our method ensures that all semantics of the selected pre-training data fall within the high-probability regions of the sensitive semantic distribution. We put detailed analysis in Appendix B.

We provide an example to help readers better understand the processes of semantic query and the selection of a pre-training dataset. As shown in Figure 2, each image in the public dataset is labeled with a semantic tag, either *bee* or *zebra*. We first train the semantic query function using the public dataset, then determine the semantics of sensitive images using this trained function. We set $k_1 = 1$ and $k_2 = 1$, implying that each sensitive image considers only one semantic. The initial query results yield $\{zebra : 3, bee : 0\}$ (although the sensitive data contains horses, zebra is the closest label from Q). To ensure privacy, we introduce Gaussian noise to these results, resulting in a modified output of $\{zebra : 2.7, bee : 1.5\}$. Given that $k_2 = 1$, we select the semantic *zebra*, which has the highest frequency among all semantic candidates, as the representative descriptor of the sensitive dataset. Ultimately, we select only those public images labeled as *zebra* for our pre-training. In a word, we select pre-training data from the public dataset guided by the perturbed semantic distribution of the sensitive dataset.

Algorithm 1: Fine-tuning with Gradient Sanitizing

Input: Sensitive dataset D_s; Image generative model G parameterized by θ; Objective of generative models: $\mathcal{L}$; Batch size b; Learning rate η; Clip coefficient C; Gaussian variance σ_1^2; Max iterations T_m
$T \leftarrow 0$
while $T < T_m$ **do**
  Randomly sample the the image training batch $x_{1:b}$ from D_s
  `// Calculate gradient`
  $g_{1:b} \leftarrow \nabla_\theta \mathcal{L}\left(G_\theta\left(x_{1:b}\right)\right)$
  `// Scale gradient`
  $g_{1:b}' \leftarrow \min\left\{1, \frac{C}{\|g_{1:b}\|_2}\right\} g_{1:b}$
  `// Add Gaussian noise`
  $\hat{g} \leftarrow \frac{1}{b}\sum_{i=1}^{b} g_i' + \frac{\sigma_1 C}{b} e, e \sim \mathcal{N}(0, \mathrm{I})$
  `// Update parameter`
  $\theta \leftarrow \theta - \eta \hat{g}$
  $T \leftarrow T + 1$
end
Output: The well-trained image generative model: G^*

Considering applying our method for conditional generation, where the sensitive dataset has been divided into several subsets with corresponding categories, we query a semantic distribution for each subset. This query still satisfies Theorem 3.1 because each sensitive image only provides an independent prediction. Therefore, we can label the selected image with the category of the sensitive subset it belongs to. For example, the selected samples can be zebras or, in unfortunate cases, birds from `ImageNet`, but we label them as horses and pre-train our model accordingly.

3.4 Pre-training and Fine-tuning

We pre-train the image generative models on the selected dataset and subsequently fine-tune these pre-trained models on the sensitive dataset. In particular, in the fine-tuning phase, we use DP-SGD [1] to update the model in adherence to the DP framework, ensuring the protection of sensitive dataset's privacy. Next, we explain how to apply standard DP-SGD works in our fine-tuning in Algorithm 1.

As described in Algorithm 1, let $\mathcal{L}$ be the training objective of generative models (e.g., Eq. (2) and (3)) and we want to minimize $\mathcal{L}$ through updating the parameters θ of our neural network G_θ. At each training iteration, we first compute the gradient $g = \nabla_\theta \mathcal{L}\left(G_\theta\left(x\right)\right)$, where x is an input image. Then, we clip the gradient for controlling the maximal l_2-norm of every gradient, like $g' = \min\left\{1, \frac{C}{\|g\|_2}\right\} g$, where C is a hyper-parameter. Lastly, we introduce Gaussian noise with variance σ_1^2 to the scaled gradient. In a word, we update the G_θ with

Algorithm 2: PRIVIMAGE Workflow

Input: Public dataset D_p with its semantics label S_p; Sensitive dataset D_s; Semantic query function Q parameterized by ϕ; Image generative model G parameterized by θ; Semantics query parameters k_1, k_2; Gaussian variance σ_2^2 for Semantic Distribution (SD); Privacy budget ε and δ.
`// Obtain semantic query function`
Train Q_ϕ on D_p and S_p with Eq. (4)
`// Query semantic distribution`
Initiate the semantic distribution $SD \leftarrow \mathrm{O}_{NS\times 1}$
for $x \in D_s$ **do**
  Query k_1 semantics $[s_1, \cdots s_{k_1}]$ from $Q(x)$
  for $s \in [s_1, \cdots s_{k_1}]$ **do**
    Obtain the index j of s
    $SD[j] += 1$
  end
end
$SD = SD + \mathcal{N}\left(0, k_1\sigma_2^2 \mathrm{I}\right)$
Semantic description $S^* \leftarrow$ top-k_2 semantics in SD
`// Choose data from the public dataset`
Selected dataset $D_{ps} \leftarrow \varnothing$
for $(x, s_p) \in D_p$ **do**
  if $s_p \in S^*$ **then**
    $D_{ps} = x \cup D_{ps}$
  end
end
`// Pre-train and fine-tune` G_θ
Pre-train G_θ on D_{ps}
Fine-tune G_θ on D_s following Algorithm 1
Generate D_s^* by leveraging G_θ
Output: Synthetic dataset D_s^*

the noisy gradient, which is defined as,

$$\hat{g} = \frac{1}{b}\sum_{i=1}^{b} \min\left\{1, \frac{C}{\|g_i\|_2}\right\} g_i + \frac{\sigma_1 C}{b} e, \;\; e \sim \mathcal{N}(0, \mathrm{I}),$$

where b is the batch size. Given privacy budget ε and δ, the maximal training iterations are constrained by the predetermined privacy budget ε and δ.

A previous study has proved that model size scales sublinearly with data size [30]. To achieve good pre-training on such large datasets, we tend to use models with larger parameters. In contrast, these models can often be over-parameterized when applied to the sensitive dataset. Whereas, we recommend to train models with fewer parameters. This is because more parameters indicate higher dimension of gradient. When the l_2-norm of gradient is constrained by clipping, the value of each gradient dimension decreases with the increase of the dimension. Consequently, the gradient becomes more susceptible to perturbations from consistent Gaussian noise, resulting

Table 2: Data split of three datasets used in our experiments.

Dataset \ Subset	Training	Validation	Test
CIFAR-10	45,000	5,000	10,000
CelebA	162,770	19,867	19,962
ImageNet	1,281,167	50,000	100,000

in unstable training of generative models. Thus, it is preferable to train lightly parameterized generative models rather than over-parameterized ones [15, 41, 67]. Moreover, with the reduction in the size of the pre-training dataset, PRIVIMAGE enables us to well train generative models with fewer parameters during the pre-training and fine-tuning stages, thereby benefiting the efficiency of training.

Algorithm 2 elaborates the process for PRIVIMAGE, which produces a synthetic dataset of fidelity and utility similar to the sensitive dataset without privacy leakage. First, we train a semantic query function on a public dataset with Eq. (4) (Line 2). Then, we use the trained semantic query function to query the semantic distribution of the sensitive dataset (Lines 3-10). After adding Gaussian noise to the query results (Line 11), we use it to select data from the public dataset for pre-training a generative model (Line 12-19). Then, we fine-tune the generative model on the sensitive dataset with DP-SGD (Line 20). Finally, the fine-tuned model generates a synthetic image dataset (Line 21).

The privacy analysis of PRIVIMAGE is put into Appendix F in our full version [43]. In particular, in PRIVIMAGE, two processes consume the privacy budget: (1) querying the semantic distribution of the sensitive dataset and (2) querying the gradient during each training iteration while fine-tuning on the sensitive dataset. Both these processes can be viewed as compositions of multiple Sub-sampled Gaussian Mechanism (SGM) instances [58]. Therefore, we use RDP described in Theorem 2.1 to track the privacy budget cost of PRIVIMAGE.

4 Experiment Setup

This section details the experimental settings to evaluate the proposed PRIVIMAGE, including the investigated datasets, baselines, evaluation metrics, and implementation.

4.1 Investigated Datasets

We utilize ImageNet, a widely used dataset for pre-training in computer vision, along with CelebA and CIFAR-10 for DP image synthesis research. These sensitive datasets, featuring over 200K celebrity images and 60,000 natural images across 10 classes respectively, offer a greater synthesis challenge compared to the commonly used MNIST and FashionMNIST. All datasets are divided into a training set, a validation set, and a test set. Please refer to Table 2 for details. All images in CelebA undergo center-cropping and are subsequently resized to dimensions of 64×64 and 32×32, which are named CelebA32 and CelebA64 respectively. Figure 10 presents examples of ImageNet, CIFAR-10, and CelebA, and We present more details in Appendix D.1 in our full version [43].

4.2 Baselines

This paper selects DPDM [15], PDP-Diffusion [18], DP-LDM [54], DPSDA [49], DPGAN [11, 72, 82] and DPGAN with pre-training (DPGAN-p) as our baselines, which follow the same framework of DP image synthesis proposed in Section 2.3. All these baselines are trained with state-of-the-art image generative models, GANs and diffusion models, and incorporate DP-SGD [1] to ensure the protection of sensitive datasets. Additionally, PDP-Diffusion, DP-LDM, DPSDA and DPGAN-p leverage the public dataset for pre-training.

- **DPDM:** DPDM [15] trains diffusion models with lightweight parameters and a substantial number of images (e.g., 'batch sizes') at each training iteration. To protect the sensitive dataset, Gaussian noise is introduced to the gradient of the diffusion model, aligning with the DP-SGD methodology [1].
- **PDP-Diffusion:** PDP-Diffusion [18] adopts a training approach that capitalizes on larger batch sizes to enhance the stability and convergence speed of the model training. By leveraging the public dataset, the model benefits from a broader knowledge base, potentially improving its subsequent fine-tuning on the sensitive dataset.
- **DP-LDM:** Given a pre-trained diffusion model, DP-LDM [54] proposes to fine-tune only its label embedding module and attention module. This approach reduces the number of trainable parameters and requires less noise to achieve the same privacy budget.
- **DPSDA:** DPSDA [49] proposes a Private Evolution algorithm that progressively guides the pre-trained models to generate a synthetic image dataset similar to the sensitive one without the need for fine-tuning.
- **DPGAN:** DPGAN [11, 72, 82] represents a typical method that trains GAN with DP-SGD for sensitive data synthesis. We refer to the implementation of [72], which is the sole work offering executable code along with privacy analysis.
- **DPGAN-p:** To validate the effectiveness of dataset as selected by PRIVIMAGE, we introduce an additional baseline, DPGAN-p. Building upon DPGAN, DPGAN-p first pre-trains the GAN on the public dataset, and subsequently fine-tunes it on the sensitive dataset adopting the pre-training method proposed by [18].

Recently, PDP-Diffusion [18] with 80M parameters has achieved SOTA results on CIFAR-10 [40] under $\varepsilon = 10$. For

the dataset CIFAR-10, we set δ to 10^{-5}, and for CelebA, we set it to 10^{-6}. We adopt three privacy budgets, $\varepsilon = \{10, 5, 1\}$, for both datasets. These specific privacy budgets have been commonly used in prior DP image synthesis research [15, 18]. PRIVIMAGE prioritizes efficient image synthesis using low-parameterized neural networks even under smaller privacy budgets 5 and 1, saving substantial computation resources and achieving superior results.

4.3 Evaluation Metrics

We evaluate the fidelity and utility of the synthetic dataset using two widely accepted metrics: Fréchet Inception Distance (FID) and Classification Accuracy, as commonly employed in prior research [15, 18, 72].

Fréchet Inception Distance (FID): FID is a metric widely used to assess the fidelity of images generated by Generative models [7, 31, 68, 69]. A lower FID suggests that the generated images are higher quality and more akin to the real ones. We generate 5,000 synthetic images to calculate FID.

Classification Accuracy (CA): We assess the utility of synthetic images with specific attributes or categories in the accuracy of downstream classification. Specifically, we select three classification models: Logistic Regression (LR), Multi-Layer Perceptron (MLP), and Convolutional Neural Network (CNN). These models are trained on synthetic datasets using different methods and their classification accuracy is tested on the sensitive test dataset. We generate 50,000 synthetic images to train classifiers.

Semantic Distribution Similarity (SDS): This paper proposes a metric to access the semantic distribution similarity between two image datasets. Let D_1 and D_2 be the two image datasets with semantics of $\left\{s_1^j\right\}_{j=1}^{NS_1}$ and $\left\{s_2^j\right\}_{j=1}^{NS_2}$ respectively, where NS_1 and NS_2 represent the number of semantics in their respective images. The semantic distribution similarity of D_1 and D_2 is calculated as,

$$SDS(D_1, D_2) = \sum_{i=1}^{NS_1} \sum_{k=1}^{NS_2} w_1^i w_2^k \frac{V(s_1^i) \cdot V(s_2^k)}{\left\|V(s_1^i)\right\| \left\|V(s_2^k)\right\|} \tag{5}$$

where V indicates a function that converts a semantic or a word into its embedding vector [62]. Besides, $\frac{V(s_1^i) \cdot V(s_2^k)}{\left\|V(s_1^i)\right\| \left\|V(s_2^k)\right\|}$ calculates the cosine similarity of the embedding vector of s_1^i and s_2^k. The w_1^i represents the frequency of semantic s_1^i in D_1, and we consider semantics with higher frequency can better represent the dataset. The higher value of $SDS(D_1, D_2)$ suggests that D_1 and D_2 have closer semantic distribution.

4.4 Implementation

All image generative methods are realized with Python 3.8 on a server with 4 NVIDIA GeForce A100 and 512GB memory.

We replicate DPDM [15] using open-source implementation repositories, and building upon that, we also reproduce PDP-Diffusion [18]. In particular, we use a larger batch size and pre-train the diffusion models on ImageNet [14]. For DP-LDM [54], we follow their proposed approach and fine-tune only the label embedding and attention modules in the diffusion models, keeping the other parameters frozen. We replicate DPSDA [49] using their open-source code and replace their RANDOM-API with the pre-trained diffusion model used in PDP-Diffusion for a fair comparison. To ensure fairness, we discuss the comparison with their 270M diffusion models in Section 6.2. For DPGAN [72], we refer to the practice adopted for GAN training with gradient sanitizing released in opacus[2]. Our implementation of DPGAN-p is based on DPGAN, incorporating similar training techniques as suggested in PDP-Diffusion [18].

For PRIVIMAGE, we choose diffusion models and GANs as the image generative models within our PRIVIMAGE framework. We refer to these two variations as PRIVIMAGE+D and PRIVIMAGE+G, respectively. For the fair comparison, all Diffusion-based and GAN-based methods use lightly parameterized models with the same parameters. Although PDP-Diffusion proposes to use large-scale diffusion models (e.g. with 80M parameters), we believe this way is ineffective for both the synthesis performance and computational resource saving, which are discussed in Section 6.2 in more details. We recommend readers refer to Appendix D in our full version [43] for more implementation details.

5 Evaluation Results

This section addresses three research questions (RQs) to evaluate the effectiveness of PRIVIMAGE. Specifically, we investigate (1) whether PRIVIMAGE generates higher-quality images than the baselines, (2) reasons why the queried semantic distribution improves fine-tuning, and (3) impact of hyper-parameters on PRIVIMAGE.

RQ1. How effective is PRIVIMAGE for synthesizing useful images?

Experiment Design. We explore whether PRIVIMAGE can generate synthetic images with higher fidelity and utility than baselines. We compare our PRIVIMAGE+G and PRIVIMAGE+D (which is described in Section 4.4) with six baselines introduced in Section 4.2 on CIFAR-10 [40] and CelebA [52] under the privacy budget $\varepsilon = \{10, 5, 1\}$.

Result Analysis. Table 3 shows that PRIVIMAGE outperforms all baselines in terms of the FID and CA of downstream classification tasks using synthetic images on CIFAR-10, CelebA32 and CelebA64 across three distinct privacy budgets.

[2] https://github.com/pytorch/opacus/tree/main/examples

Table 3: FID and CA of PRIVIMAGE and four baselines on CIFAR-10 [40], CelebA32 and CelebA64 [52] with $\varepsilon = 10, 5, 1$. For space limitation, CeA32 and CeA64 refer to CelebA32 and CelebA64 respectively. The best performance in each column is highlighted using the bold font.

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	9.17	8.4	10.5	258	202	121	14.2	14.6	13.0	210	227	190	16.2	17.4	14.8	225	232	162
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	113	115	21.1	24.7	22.0	311	122	127	19.6	22.3	14.7	340	223	243
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

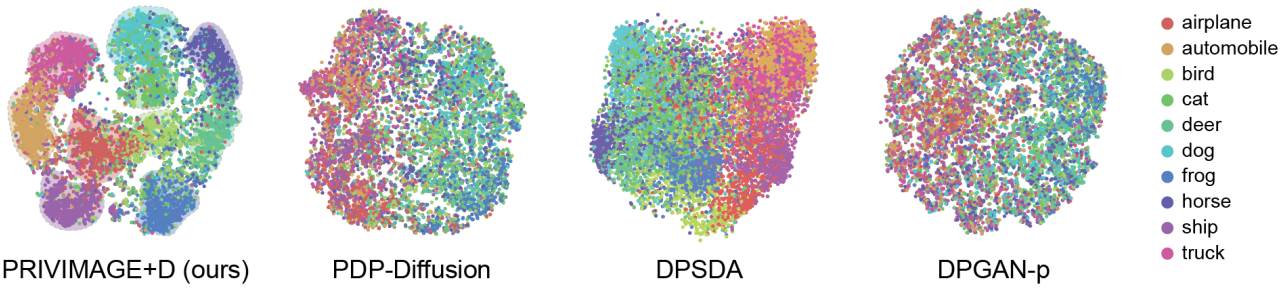

Figure 3: The t-SNE visualizations depict the embedding distribution of the synthetic images using four DP dataset synthetic methods within the CIFAR-10 dataset with $\varepsilon = 10$. We obtain embeddings by using CNN classifiers trained on synthetic images.

When $\varepsilon = 10$, for CIFAR-10, PRIVIMAGE+G and PRIVIMAGE+D outperform all the baselines in terms of FID with specifically 7.7% lower than the SOTA method DPSDA [49], suggesting that the fidelity of images generated by PRIVIMAGE more closely aligns with the sensitive dataset. For the utility of synthetic images, PRIVIMAGE+D achieves superior downstream classification accuracy on all three classification models with especially **20.9%** higher than the SOTA method [49] on the CNN classifier. PRIVIMAGE+G outperforms all GAN-based baselines and DPDM [15], achieving competitive results with the SOTA method. PRIVIMAGE+G outperforms all the baselines in terms of FID with specifically 13.7% and 22.0% lower than the SOTA method on CelebA32 and CelebA64, which implies that our PRIVIMAGE is also effective when there is a substantial difference between the public and sensitive dataset. Although PRIVIMAGE+D does not achieve a lower FID than DPSDA on CelebA64, we consider a practical advantage of PRIVIMAGE+D to be its ability to directly sanitize the diffusion model. Once the model is well-trained, PRIVIMAGE can respond to any number of synthesis queries with fast speed while still protecting data privacy due to the post-processing property [17]. In contrast, DPSDA incurs the same significant time cost for each synthesis query, as discussed in Section 6.2. Moreover, their privacy analysis cannot ensure the security of multiple synthesis queries.

For privacy budget 5 and 1, PRIVIMAGE consistently surpasses all baselines. Particularly, as ε shifts from 10 to 1, the FID of synthetic CIFAR-10 images produced by PRIVIMAGE+D and PRIVIMAGE+G drops only by 7.8% and 7.2%. In contrast, the FID for the state-of-the-art method diminishes by 23.7%, and for DPGAN-p, it decreases by a substantial 41.0%. For CelebA32 and CelebA64, PRIVIMAGE+D and PRIVIMAGE+G only decreases by 36.1% and 18.1% respectively, while the SOTA method decreases by 49.1% and 50.6%. However, our PRIVIMAGE+G decreases by 68.2% on CelebA32. It is well known that GAN suffers from unstable training [2]. Increased gradient noise due to a restricted privacy budget can further intensify this instability. Hence, we believe diffusion models might be more appropriate as image generative models in scenarios with limited privacy budgets.

Additionally, we investigate the distribution characteristics of the synthetic dataset to validate the high utility of our generated data. Specifically, we first use trained CNN classifiers from the dataset generated by PRIVIMAGE+D, PDP-Diffusion, DPSDA, and DPGAN-p, to undertake the classification task within the CIFAR-10. With the aid of the well-trained CNN, each image is transformed into an embedding vector. This vector encapsulates the most pertinent feature information utilized by the CNN for classification. We use t-SNE to visualize them in a two-dimension space.

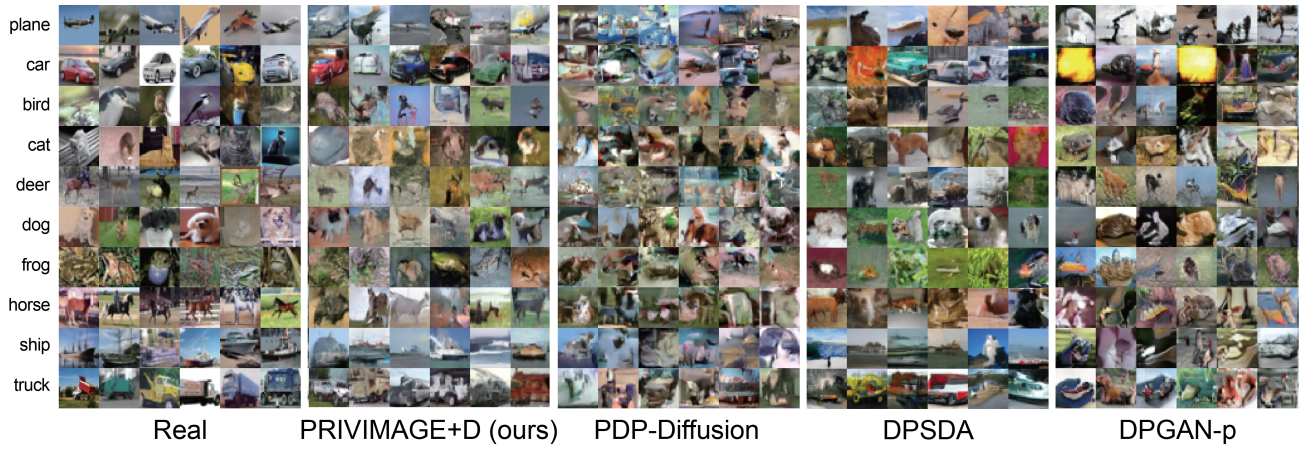

Figure 4: Examples of Synthetic `CIFAR-10` [40] images with $\varepsilon = 10$. These generative models are trained using our tools PRIVIMAGE+D, as well as PDP-Diffusion [18] and DPGAN-p. Each row corresponds to a category from the `CIFAR-10` dataset.

Figure 3 illustrates that embeddings of image data from PRIVIMAGE+D are projected into 10 clusters aligning with 10 categories. This suggests that the classifier adeptly discerns the unique features of images, thereby achieving superior performance compared to the CNNs trained on the dataset provided by other methods. In contrast, dataset embeddings from others do not form distinct clusters, leading their classifiers to struggle with classification. Examples of synthetic images for `CIFAR-10` from various methods are presented in Figure 4. Additionally, we present more synthetic images for `CelebA` and `Camelyon17` [4] in Appendix G.1 in our full version [43].

Answers to RQ1: Synthetic images produced by PRIVIMAGE exhibit greater fidelity and utility compared to all baseline methods with three distinct privacy budgets. On average, the FID of the synthetic dataset is **6.8%** lower, and the CA of the downstream classification task is **13.2%** higher, compared to the state-of-the-art method.

RQ2. How do the semantic distribution queries of PRIVIMAGE improve the fine-tuning?

Experiment Design. We conducted experiments to explore whether the fine-tuning of generative models is enhanced by leveraging our proposed semantic distribution query to select images from the entire public dataset for pre-training. As a result, PRIVIMAGE can synthesize images with greater fidelity and utility. We approach this from two perspectives: data distribution similarity and semantic distribution similarity. To evaluate data distribution similarity, we compute the FID of synthetic images generated by different methods, where generative models have only been pre-trained and not fine-tuned on the sensitive dataset. To evaluate semantic distribution similarity, we assess the SDS (as given in Eq. (5)) between the sensitive dataset and the pre-training dataset across various selection ratios. Additionally, we present a showcase from

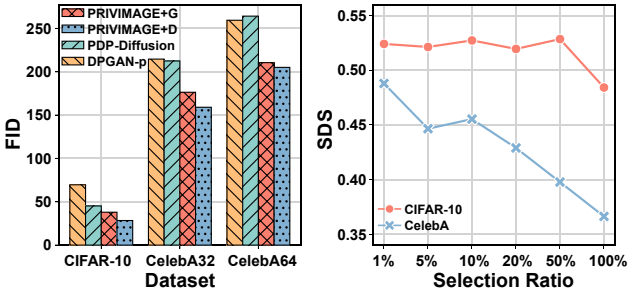

Figure 5: (Left) The FID of baselines pre-trained on the entire public dataset and ours PRIVIMAGE pre-trained on the selected public dataset. (Right) The SDS between the sensitive dataset and different pre-training datasets.

the pre-training dataset selected by our PRIVIMAGE, to further verify the semantic distribution similarity from a visual perspective in Appendix G.2 in our full version [43].

Result Analysis. We analyze our results from two similarity between the synthesized and sensitive dataset as follows.

Data Distribution Similarity. The left panel of Figure 5 presents that when generative models are pre-trained on the selectively curated public dataset, PRIVIMAGE produces datasets with superior FID results compared to baselines pre-trained on the entire public dataset. Before fine-tuning on the sensitive dataset, PRIVIMAGE can intuitively generate synthetic images with a data distribution closer to the sensitive data than images produced by existing methods. PRIVIMAGE *wins at the beginning.* With an equal number of fine-tuning iterations, PRIVIMAGE still achieves superior performance.

Semantic Distribution Similarity. For the `CIFAR-10` dataset [40], we adopt its 10 category labels as its semantic representation. Similarly, for the `CelebA` dataset [52], we use its 40 face attribute labels to represent its semantics. The public dataset `ImageNet` [14] is characterized by its 1000 category labels. We experiment with various selection ratios

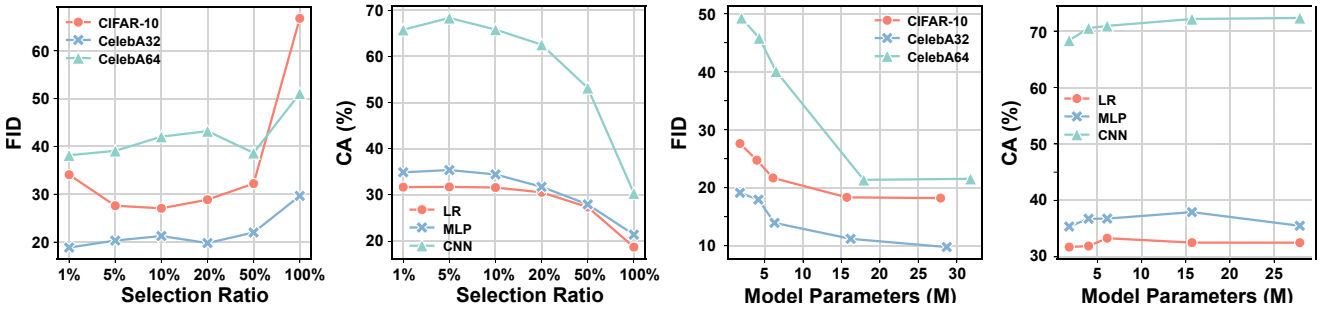

Figure 6: The two left figures present the effect of the selection ratio on the FID of the synthesized image dataset and the CA for downstream classification tasks. The models are trained on datasets selected from the public dataset for pre-training, using various selection ratios. "100%" represents that pre-training generative models on the entire public dataset. The right two figures display the effect of model size, where the selection ratios are set as 5% and 1% for CIFAR-10 and CelebA respectively.

from this public dataset by adjusting the number of queried semantics, k_2, as described in Section 3.3. The right panel of Figure 5 displays that the semantic distribution similarity between ImageNet [14] and CIFAR-10 [40] surpasses that of CelebA [52], aligning with our human visual perceptions as presented in Figure 10 in our full version [43], indicating the utility of our proposed SDS metric. The datasets selected by our PRIVIMAGE exhibit a closer semantic distribution to the two sensitive datasets than the entire public dataset across different selection ratios. Leveraging the semantic distribution query, PRIVIMAGE selects a pre-training dataset whose semantic distribution closely mirrors that of the sensitive dataset, all while operating within a minimal privacy budget.

Answers to RQ2: PRIVIMAGE selects a pre-training dataset that resembles the sensitive data more closely than the entire public data. Before fine-tuning the sensitive dataset, PRIVIMAGE produces synthetic images with a data distribution maligned with the sensitive data. As a result, PRIVIMAGE delivers enhanced DP image synthesis.

RQ3. How do hyper-parameters affect the performance of PRIVIMAGE?

Experiment Design. This experiment investigates how the selection ratio of selecting pre-training data from the entire public data and the generative model size impact the performance of PRIVIMAGE+D. The selection ratio is determined by k_2. For example, ImageNet has 1,000 subsets, each containing nearly the same number of images. When $k_2 = 10$, it means that 10 subsets will be selected for pre-training, resulting in a selection ratio of 1%. We explore selection ratios of 1%, 5%, 10%, 20%, 50%. In terms of model size, we consider 1.8M, 4.0M, 6.1M, 15.7M, 27.9M for CIFAR-10, 1.8M, 4.2M, 6.3M, 16.2M, 28.7M for CelebA32, and 2.0M, 4.3M, 6.5M, 17.9M, 31.8M for CelebA64.

Result Analysis. Figure 6 shows that across all selection ratios, PRIVIMAGE outperforms the full public dataset (where the selection ratio is 100%), revealing the effectiveness of our semantic distribution query. As the selection ratio increases, implying more data is used for pre-training, PRIVIMAGE generally exhibits poorer FID and CA on both CIFAR-10 and CelebA. We believe the reason for this is that only a small portion of ImageNet is similar to our sensitive datasets. Pre-training models on a larger number of dissimilar images may distract the models from learning the distribution of similar images, making it more challenging for the models to learn the distribution of the sensitive datasets. However, this trend is more pronounced for CIFAR-10 than for CelebA. This can be attributed to the fact that CelebA diverges more from the public dataset ImageNet, resulting in fewer opportunities for PRIVIMAGE to select beneficial data for pre-training. For CIFAR-10, the performance of PRIVIMAGE is notably impacted. Specifically, when the selection ratio is at 1%, the FID of synthetic images rises compared to that at 5%. Hence, when there is a significant discrepancy between sensitive and public data, we have the flexibility to select a ratio, which typically outperforms using the entire public dataset. To further investigate this phenomenon, we conduct experiments on another dataset with a larger domain shift from ImageNet. The results of these experiments are presented in Appendix G.1 in our full version [43].

In terms of model size, as it increases, PRIVIMAGE generally achieves superior FID and CA results. However, as the model size becomes significantly larger, the performance gets unstable. For instance, the CA of MLP classifiers trained on

Table 4: The FID and CA of PRIVIMAGE and three non-private baselines on CIFAR-10 [40], CelebA32 and CelebA64 [52] with $\varepsilon = 10$.

Method	CIFAR-10				CelebA32	CelebA64
	CA (%)			FID	FID	FID
	LR	MLP	CNN			
NonPrivG	18.2	24.7	49.8	22.5	13.6	34.7
NonPrivD	35.8	42.2	77.1	19.8	9.01	18.0
NonPriv	37.4	45.7	86.1	-	-	-
PRIVIMAGE+G	19.9	24.5	44.3	28.1	18.9	38.2
PRIVIMAGE+D	31.7	35.4	68.3	27.6	19.1	49.3

synthetic images from PRIVIMAGE at 27.9M is lower than at 15.7M. Similarly, the CA of LR classifiers from PRIVIMAGE at 15.7M is below that of 6.1M. These observations suggest that lighter generative models might be better suited for the DP image synthesis task.

Answers to RQ3: Compared to using the entire public dataset, PRIVIMAGE achieves superior performance with a selection ratio less than 50%. As the divergence between public and sensitive data widens, the influence of the selection ratio diminishes gradually. When the generative model comprises fewer than 15M parameters, PRIVIMAGE produces higher-quality synthesized images with larger models. Whereas, as the parameters increase further, PRIVIMAGE experiences unstable training.

6 Discussion

This section discusses how PRIVIMAGE performs without privacy-protective settings, the computational resources required by PRIVIMAGE, as well as its potential applications and inherent limitations. We also discuss about how to choose the privacy parameter in Appendix A.

6.1 PRIVIMAGE without Privacy Protection

This experiment studies how the synthetic performance of PRIVIMAGE is harmed by adhering to the DP framework. We compared PRIVIMAGE with three methods: (1) "NonPrivD" and "NonPrivG" train diffusion models and GANs respectively on the pre-training dataset chosen by PRIVIMAGE with selection radio 5%, whereas the gradient in fine-tuning stage and semantic distribution in the pre-training stage are not introduced Gaussian noise. (2) "NonPriv" directly uses the sensitive dataset to train the classifier for the downstream classification task. It is noticed that the FID measures the quality of images created by the generative models, so that sensitive datasets do not have an FID value. We conduct these experiments with a privacy budget $(10, 1 \times 10^{-5})$.

Table 4 shows that, on average, PRIVIMAGE+D and PRIVIMAGE+G only decrease by 6.5% and 1.4% in terms of CA on three classification models. Therefore, with sensitive data protected, PRIVIMAGE still generates useful synthetic images. In terms of CA, PRIVIMAGE shows an 11.2% decrease compared to NonPriv. However, when we increase the generative model size of PRIVIMAGE+D by 2.2M parameters, the decrease is just 7.9%. This suggests that more efforts are needed to develop our PRIVIMAGE further.

6.2 Consumption of Computational Resource

This section presents the high efficiency of PRIVIMAGE and the significant trade-off between the computation resource cost and synthesis performance. We trained the state-of-the-art method PDP-Diffusion using an 80M diffusion models [18] and PRIVIMAGE+D using a 6.1M diffusion model on `CIFAR-10`. We also implement DPSDA [49] using a 270M diffusion model pre-trained on `ImageNet`. We compare their GPU memory usage and time cost during pre-training, fine-tuning and synthesis on identical computational configurations. PRIVIMAGE needs additional time to train the SQF.

Table 5: GPU memory cost, Running time, CA and FID of synthetic images of PDP-Diffusion, DPSDA and PRIVIMAGE. The time cost of SQF includes the time taken to train SQF and to use the trained SQF for querying the semantic distribution.

Evaluation Metrics		PDP-Diffusion	DPSDA	PRIVIMAGE
Memory	Pre-train	107GB	0GB	55GB
	Fine-tune	158GB	0GB	89GB
	SQF	0GB	0GB	22GB
	Synthesis	177GB	219GB	83GB
Time	Pre-train	87h	0h	46h
	Fine-tune	24h	0h	3h
	SQF	0h	0h	9h
	Synthesis	0.37h	12h	0.15h
CA (%)	LR	15.07	14.3	33.11
	MLP	14.17	13.4	36.78
	CNN	24.27	15.1	70.91
FID		22.90	16.2	21.69

Table 5 shows that PRIVIMAGE uses only 50% and 59% of the GPU memory used by PDP-Diffusion and DPSDA, respectively. In terms of runtime, PRIVIMAGE is 48% faster than PDP-Diffusion. These results can be attributed to the compact dataset selected by PRIVIMAGE for pre-training and lighter generative models that PRIVIMAGE uses. Although PRIVIMAGE requires more time compared to DPSDA, the majority of PRIVIMAGE's time is spent on training, while the synthesis process is 98% faster than DPSDA. When our synthesis method needs to respond to multiple queries, DPSDA could be significantly slower than PRIVIMAGE. Despite the significant savings in computational resources, our PRIVIMAGE still achieves an FID and average CA that are 5% lower and 29% higher than PDP-Diffusion.

As presented in Table 3, compared to an 1.8M model, PDP-Diffusion with an 80M model does achieve better FID. However, the average CA on three classifiers drops by 5.7%, suggesting that the synthetic images are less useful. This is attributed to the fact that the l_2-norm of the noise added during DP-SGD scales linearly to the dimension of parameters, leading to a "curse of dimensionality" and making the training of generative models unstable [15, 41, 67]. We also conduct another experiment to prove this "curse of dimensionality" and put the result in Appendix C. Besides, we may need to upload our sensitive data to a remote server using such over-parameterized models, presenting potential security concerns [73]. Therefore, we should be more concerned about achieving great synthesis with lightly parameterized

models, which are less affected by DP-SGD and can be used on end-user devices more easily.

6.3 Applications and Limitations

This subsection discusses the application scope and limitations of PRIVIMAGE.

Potential Applicability to Other Fields. Though this paper primarily focuses on image data synthesis, we posit that PRIVIMAGE has high potential applicability to other data types, such as text data. Numerous studies have highlighted the effectiveness of diffusion models in tasks like text generation [44] and audio synthesis [19]. Analogous to the practical adopted in PRIVIMAGE, by judiciously selecting a suitable public dataset as well as integrating pre-training with fine-tuning methodologies, it seems feasible to achieve DP text, audio synthesis, and even other fields.

Dependence on Public Data. The primary advantage of PRIVIMAGE lies in its ability to query the semantic distribution of sensitive data and subsequently select data from a public dataset that closely aligns with the sensitive dataset for pre-training. However, when the distribution of the public data substantially deviates from that of the sensitive data, PRIVIMAGE faces limited opportunities to pick out such pertinent data. In the absence of a suitable public dataset, the effectiveness of PRIVIMAGE in generating complex image datasets while adhering to the DP guarantee diminishes.

7 Related Work

This section explicitly discusses two main types of DP image data synthesis works, including sanitizing generative models through DP-SGD [1] and querying DP information of sensitive datasets for synthesis.

Sanitizing Generative Models via DP-SGD. Currently, most efforts have focused on applying DP-SGD [1] on popular generative models, like GANs [3, 10, 36, 51], diffusion models [15, 18, 54] or variational autoencoders (VAE) [24, 34, 63]. They introduce noise into gradient calculations during training, ensuring that individual data points do not disproportionately influence the model's learning, thereby offering a degree of privacy protection. Diffusion models have shown promise in generating DP image data. For example, PDP-Diffusion [18] and DPDM [15] respectively achieve state-of-the-art performance in terms of fidelity and utility of synthetic images, and demonstrate top-tier classification accuracy, both under the setting of with and without the use of a public dataset for pre-training. However, DPDM only achieves synthesis on some naive datasets (e.g. `MNIST`). PDP-Diffusion demands significant computational resources and struggles to produce images with great fidelity and utility on end-user devices for real-world applications [73].

PRIVIMAGE presented in this paper also relies on sanitizing generative models via DP-SGD for DP image synthesis. Different from existing methods, PRIVIMAGE queries the semantic distribution of the sensitive dataset to select a more compact dataset from the public dataset for pre-training. PRIVIMAGE enables us to use a lightly parameterized model, while achieving superior synthesis performance.

Querying DP Information for Synthesis. Another type of method emphasizes designing a query function to extract valuable features from a sensitive dataset [25,26,33,47,49,77]. These features exhibit low sensitivity and aid in obtaining pertinent information for synthesis. We will next introduce the various feature query functions without delving into the generation process.

Harder et al. presented DP-MERF [25], a method that represents the sensitive dataset using random features like Fourier features. Meanwhile, Seng et al. [47] suggested substituting the random feature with a characteristic function, which offers improved generalization capability. Harder et al. introduced DP-MEPF [26], a method that leverages public data to extract perceptual features. Specifically, they represent each image as a feature vector using an encoder pre-trained on public datasets, such as `ImageNet` [14]. These methods, however, fall short in comparison to DP-SGD [1] in term of generated images qualification, especially for more colorful dataset (e.g., `CIFAR-10` [40]). Recently, DPSDA [49] proposed to utilize foundation model APIs of powerful generative models trained on public datasets, such as DALLE-2 [65], Stable Diffusion [66], and GPT3/4 [8, 60]. Although this method achieves results comparable to the SOTA [18], DPSDA heavily depends on two open-source APIs, RANDOM-API and VARIATION-API, which may not be available in some scenarios, especially for the VARIATION-API.

8 Conclusions and Further Works

This paper explores leveraging the public dataset more effectively to pre-train lightly parameterized models for DP image synthesis. We propose PRIVIMAGE for generating a synthetic image dataset under differential privacy. Compared to existing methods, PRIVIMAGE queries the semantic distribution of sensitive data to select more useful data for pre-training. With lightly parameterized generative models and a small pre-training dataset, PRIVIMAGE still generates synthetic images with excellent fidelity and utility. Besides, PRIVIMAGE saves much computational resource compared to the state-of-the-art method DP-Diffusion and can be employed on end-user devices more easily. This paper calls for attention to construct a more tailored pre-training dataset to advance the practical implementations of DP image dataset synthesis.

Future work plans to address the challenge of generating high-quality images when the sensitive dataset greatly diverges from the public dataset. We aspire to extend the PRIVIMAGE to synthesize other types of data and develop PRIVIMAGE into a more practical DP image synthesis tool.

Acknowledgement

We thank all the anonymous reviewers and our shepherd for their valuable comments. Authors from CAS in this research/project are supported by the National Science and Technology Project (2022ZD0116406). Kecen Li's work was done as a remote intern at UVA.

References

[1] Martín Abadi, Andy Chu, Ian J. Goodfellow, and et al. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[2] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, 2017.

[3] Sean Augenstein, H. Brendan McMahan, Daniel Ramage, and et al. Generative models for effective ML on private, decentralized datasets. *CoRR*, abs/1911.06679, 2019.

[4] Péter Bándi, Oscar Geessink, Quirine Manson, and et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Trans. Medical Imaging*, 38(2):550–560, 2019.

[5] Hangbo Bao, Li Dong, Songhao Piao, and et al. Beit: BERT pre-training of image transformers. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[6] Franziska Boenisch, Adam Dziedzic, Roei Schuster, and et al. When the curious abandon honesty: Federated learning is not private. *CoRR*, abs/2112.02918, 2021.

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR*, 2019.

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[9] Dingfan Chen, Raouf Kerkouche, and Mario Fritz. A unified view of differentially private deep generative modeling. *CoRR*, abs/2309.15696, 2023.

[10] Dingfan Chen, Tribhuvanesh Orekondy, and Mario Fritz. GS-WGAN: A gradient-sanitized approach for learning differentially private generators. In *Advances in Neural Information Processing Systems*, 2020.

[11] Dongjie Chen, Sen-ching Samson Cheung, Chen-Nee Chuah, and et al. Differentially private generative adversarial networks with model inversion. In *IEEE International Workshop on Information Forensics and Security, WIFS*, pages 1–6, 2021.

[12] Xinlei Chen, Hao Fang, Tsung-Yi Lin, and et al. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

[13] Fida Kamal Dankar and Khaled El Emam. Practicing differential privacy in health care: A review. *Trans. Data Priv.*, 6(1):35–67, 2013.

[14] Jia Deng, Wei Dong, Richard Socher, and et al. Imagenet: A large-scale hierarchical image database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, pages 248–255, 2009.

[15] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and et al. Differentially private diffusion models. *CoRR*, 2022.

[16] Jinshuo Dong, Aaron Roth, and Weijie J. Su. Rejoinder: Gaussian differential privacy. *CoRR*, abs/2104.01987, 2021.

[17] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

[18] Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, and et al. Differentially private diffusion models generate useful synthetic images. *CoRR*, 2023.

[19] Karan Goel, Albert Gu, Chris Donahue, and et al. It's raw! audio generation with state-space models. In *International Conference on Machine Learning, ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 7616–7633, 2022.

[20] Chen Gong, Zhou Yang, Yunpeng Bai, and et al. Curiosity-driven and victim-aware adversarial policies. In *Proceedings of the 38th Annual Computer Security Applications Conference*, ACSAC '22, page 186–200, 2022.

[21] Chen Gong, Xiaoxiong Zhang, and Yunyun Niu. Identification of epilepsy from intracranial eeg signals by using different neural network models. *Computational Biology and Chemistry*, 87:107310, 2020.

[22] Chen Gong, Xingchen Zhou, and Yunyun Niu. Pattern recognition of epilepsy using parallel probabilistic neural network. *Applied Intelligence*, 52(2):2001–2012, 2022.

[23] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, and et al. Generative adversarial nets. In *Advances*

in Neural Information Processing Systems, pages 2672–2680, 2014.

[24] Benedikt Groß and Gerhard Wunder. Differentially private synthetic data generation via lipschitz-regularised variational autoencoders. *CoRR*, abs/2304.11336, 2023.

[25] Frederik Harder, Kamil Adamczewski, and Mijung Park. DP-MERF: differentially private mean embeddings with randomfeatures for practical privacy-preserving data generation. In *AISTATS*, volume 130, pages 1819–1827, 2021.

[26] Frederik Harder, Milad Jalali, Danica J. Sutherland, and et al. Pre-trained perceptual features improve differentially private image generation. *Trans. Mach. Learn. Res.*, 2023, 2023.

[27] Muneeb Ul Hassan, Mubashir Husain Rehmani, and Jinjun Chen. Differential privacy in blockchain technology: A futuristic approach. *J. Parallel Distributed Comput.*, 145:50–74, 2020.

[28] Dan Hendrycks, Collin Burns, Anya Chen, and et al. CUAD: an expert-annotated NLP dataset for legal contract review. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*, 2021.

[29] Tom Henighan, Jared Kaplan, and et al. Scaling laws for autoregressive generative modeling. *CoRR*, abs/2010.14701, 2020.

[30] Joel Hestness, Sharan Narang, Newsha Ardalani, and et al. Deep learning scaling is predictable, empirically. *CoRR*, abs/1712.00409, 2017.

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[32] Sanghyun Hong, Varun Chandrasekaran, and et al. On the effectiveness of mitigating data poisoning attacks with gradient shaping. *arXiv preprint arXiv:2002.11497*, 2020.

[33] Dihong Jiang, Sun Sun, and Yaoliang Yu. Functional renyi differential privacy for generative modeling. In *Advances in Neural Information Processing Systems*, 2023.

[34] Dihong Jiang, Guojun Zhang, Mahdi Karami, and et al. Dp^2-vae: Differentially private pre-trained variational autoencoders. *CoRR*, abs/2208.03409, 2022.

[35] Honglu Jiang, Jian Pei, Dongxiao Yu, and et al. Differential privacy and its applications in social network analysis: A survey. *CoRR*, abs/2010.02973, 2020.

[36] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. PATE-GAN: generating synthetic data with differential privacy guarantees. In *7th International Conference on Learning Representations, ICLR*, 2019.

[37] Jared Kaplan, Sam McCandlish, and et al. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020.

[38] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 4401–4410, 2019.

[39] Tero Karras, Samuli Laine, Miika Aittala, and et al. Analyzing and improving the image quality of stylegan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 8107–8116, 2020.

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

[41] Alexey Kurakin, Steve Chien, Shuang Song, and et al. Toward training at imagenet scale with differential privacy. *CoRR*, abs/2201.12328, 2022.

[42] Yann LeCun, Léon Bottou, Yoshua Bengio, and et al. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

[43] Kecen Li, Chen Gong, Zhixiang Li, and et al. Meticulously selecting 1% of the dataset for pre-training! generating differentially private images data with semantics query. *CoRR*, abs/2311.12850, 2023.

[44] Xiang Li, John Thickstun, Ishaan Gulrajani, and et al. Diffusion-lm improves controllable text generation. In *NeurIPS*, 2022.

[45] Xiujun Li, Xi Yin, Chunyuan Li, and et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137. Springer, 2020.

[46] Xuechen Li, Florian Tramèr, Percy Liang, and et al. Large language models can be strong differentially private learners. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[47] Seng Pei Liew, Tsubasa Takahashi, and Michihiko Ueno. PEARL: data synthesis via private embeddings and adversarial reconstruction learning. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[48] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, and et al. Microsoft COCO: common objects in context. In *ECCV*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755, 2014.

[49] Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and et al. Differentially private synthetic data via foundation model apis 1: Images. *CoRR*, abs/2305.15560, 2023.

[50] Huan Ling, Karsten Kreis, Daiqing Li, and et al. Editgan: High-precision semantic image editing. In *Advances in Neural Information Processing Systems*, pages 16331–16345, 2021.

[51] Yi Liu, Jialiang Peng, James Jian Qiao Yu, and et al. PPGAN: privacy-preserving generative adversarial network. In *25th IEEE International Conference on Parallel and Distributed Systems, ICPADS*, pages 985–989, 2019.

[52] Ziwei Liu, Ping Luo, Xiaogang Wang, and et al. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015,*, pages 3730–3738, 2015.

[53] Jia Long and Hongtao Lu. Multi-level gate feature aggregation with spatially adaptive batch-instance normalization for semantic image synthesis. In *MultiMedia Modeling - 27th International Conference, MMM*, volume 12572, pages 378–390, 2021.

[54] Saiyue Lyu, Margarita Vinaroz, Michael F. Liu, and et al. Differentially private latent diffusion models. *CoRR*, 2023.

[55] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*, pages 4732–4738, 2019.

[56] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International Conference on Machine Learning, ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 23803–23828, 2023.

[57] Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. In *2023 IEEE Security and Privacy Workshops (SPW)*, pages 77–83.

[58] Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. *CoRR*, abs/1908.10530, 2019.

[59] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, pages 8162–8171, 2021.

[60] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[61] Yan Pang, Tianhao Wang, Xuhui Kang, Mengdi Huai, and Yang Zhang. White-box membership inference attacks against diffusion models, 2023.

[62] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the EMNLP*, pages 1532–1543, 2014.

[63] Bjarne Pfitzner and Bert Arnrich. Dpd-fvae: Synthetic data generation using federated variational autoencoders with differentially-private decoder. *CoRR*, abs/2211.11591, 2022.

[64] Hieu Pham, Zihang Dai, Golnaz Ghiasi, and et al. Combined scaling for zero-shot transfer learning. *CoRR*, abs/2111.10050, 2021.

[65] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, and et al. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022.

[66] Robin Rombach, Andreas Blattmann, Dominik Lorenz, and et al. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 10674–10685, 2022.

[67] Yinchen Shen, Zhiguo Wang, Ruoyu Sun, and v. Towards understanding the impact of model size on differential private classification. *CoRR*, abs/2111.13895, 2021.

[68] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR*, 2021.

[69] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pages 11895–11907, 2019.

[70] Robin Strudel, Ricardo Garcia Pinel, Ivan Laptev, and et al. Segmenter: Transformer for semantic segmentation. In *International Conference on Computer Vision*, pages 7242–7252, 2021.

[71] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, and et al. Deep semantic segmentation of natural and medical images: a review. *Artif. Intell. Rev.*, 54(1):137–178, 2021.

[72] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. DP-CGAN: differentially private synthetic data and label generation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, pages 98–104, 2019.

[73] Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Considerations for differentially private learning with large-scale public pretraining. *CoRR*, abs/2212.06470, 2022.

[74] Oriol Vinyals, Alexander Toshev, Samy Bengio, and et al. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[75] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, 2017.

[76] Kelvin Xu, Jimmy Ba, Ryan Kiros, and et al. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057, 2015.

[77] Yilin Yang, Kamil Adamczewski, and et al. Differentially private neural tangent kernels for privacy-preserving data generation. *CoRR*, abs/2303.01687, 2023.

[78] Raymond A. Yeh, Chen Chen, Teck-Yian Lim, and et al. Semantic image inpainting with deep generative models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6882–6890, 2017.

[79] Samuel Yeom, Irene Giacomelli, and et al. Privacy risk in machine learning: Analyzing the connection to overfitting. In *31st IEEE Computer Security Foundations Symposium, CSF*, pages 268–282, 2018.

[80] Hongxu Yin, Arun Mallya, Arash Vahdat, and et al. See through gradients: Image batch recovery via gradinversion. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 16337–16346, 2021.

[81] Da Yu, Saurabh Naik, Arturs Backurs, and et al. Differentially private fine-tuning of language models. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[82] Xinyang Zhang, Shouling Ji, and Ting Wang. Differentially private releasing via deep generative model. *CoRR*, abs/1801.01594, 2018.

[83] Zhikun Zhang, Tianhao Wang, Ninghui Li, and et al. Privsyn: Differentially private data synthesis. In *30th USENIX Security Symposium*, pages 929–946, 2021.

[84] Luowei Zhou, Hamid Palangi, Lei Zhang, and et al. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 13041–13049, 2020.

Appendix[3]

A Selection of Privacy Budgets

DP provides a theoretical guarantee for data protection and has been adopted in many data synthesis tasks [9]. The privacy parameter ε simultaneously affects the privacy strength and utility of the synthetic data. As ε decreases, more noise needs to be added to the clean gradient, which leads to a drop in the performance of our algorithm. As our experiments show, the classification accuracy (CA) of PRIVIMAGE with $\varepsilon = 1$ is lower than that with $\varepsilon = 10$. Therefore, it could be useful to choose a proper ε value for some known attacks. In fact, this has been a rarely explored problem in the DP scenario, both theoretically and experimentally. As the definition of DP (Definition 2.1) shows, ε bounds the privacy loss in a probabilistic manner. Different attack approaches require different amounts of privacy loss for a successful attack, which is usually difficult to define.

For membership inference attacks (MIAs), Yeom et al. [79] bound the membership advantage for an ε-DP algorithm to $e^{\varepsilon} - 1$, where a lower membership advantage indicates a lower attack success rate. In the case of data poisoning attacks, Ma et al. [55] demonstrate that 0.1-DP learners are resistant to data poisoning attacks when the adversary is only capable of poisoning a small number of items. Hong et al. [32] discover that DP-SGD, even in configurations that do not provide meaningful privacy guarantees, enhances the model's robustness against data poisoning attacks.

To elaborate on what levels of DP are needed for PRIVIMAGE to be resistant to known attacks and how this affects the utility of synthetic datasets, we choose a white-box MIA [57] for attacking diffusion models. We use the TPR@10%FPR to evaluate the performance of the attacker. Specifically, TPR@10%FPR refers to the True Positive Rate when the False Positive Rate is fixed at 10%, and a higher metric means a higher attack success rate [57]. We study the vulnerability of diffusion models to MIA when trained with six different privacy budgets: $\varepsilon \in \{1, 5, 10, 100, 1000, \infty\}$, where "$\infty$" represents training PRIVIMAGE without DP protection. Figure 7 illustrates that as ε increases, PRIVIMAGE presents a reduction in FID scores, signifying enhanced utility of the synthetic dataset. However, it is observed that an increased ε enables the attacker to attain a higher TPR@10%FPR, suggesting that the data becomes more vulnerable to attacks. When ε is set to 100, the MIA achieves merely an 11% TPR@10%FPR, which approximates the effectiveness of random guessing. Consequently, within the context of the examined MIA, PRIVIMAGE offers an effective defense mechanism. For practical applications, it is recommended to adopt smaller ε values, such as 10 or 1, as suggested by prior studies [15, 18, 83], to defend against some unknown attacks.

[3] Please refer to our full version [43] for Appendix D-G.

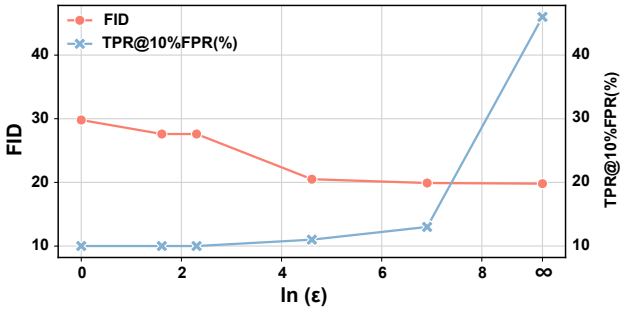

Figure 7: The FID of synthetic images from PRIVIMAGE and TPR@10%FPR of MIA [57] under different ε. "∞" represents training PRIVIMAGE without DP protection.

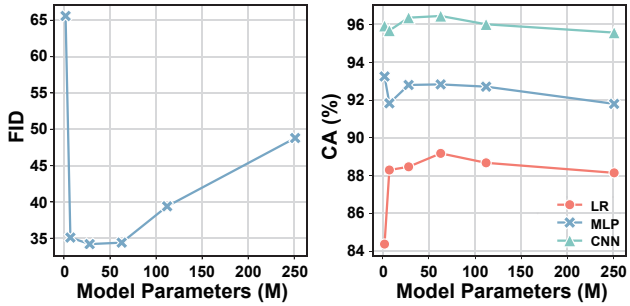

Figure 8: The FID and CA of the synthetic images from diffusion models of different sizes, which are fine-tuned on `MNIST` with DP-SGD.

B Semantics Query and Selection

In Section 3.2, we use a trained semantic query function to query k_1 semantics of each sensitive image. In Section 3.3, we select the top-k_2 semantics based on their probabilities in the queried semantic distribution as the semantics description of the sensitive dataset. In our experiments, we set $k_1 = k_2 = k$. We take the example in Figure 2 to tell the reason for this setting.

(1) $k_1 < k_2$**.** Since the number of semantics in the public dataset is 2, we have $k_1 = 1$ and $k_2 = 2$. The initial query results yield $\{zebra : 3, bee : 0\}$. However, we select both *zebra* and *bee* as the semantics description, meaning the entire public dataset is selected.

(2) $k_1 > k_2$**.** Since the number of semantics in the public dataset is 2, we have $k_1 = 2$ and $k_2 = 1$. The initial query results yield $\{zebra : 3, bee : 3\}$. Since the mean of added Gaussian noise is zero, it is random for us to select *zebra* or *bee* from the noisy semantic distribution as the semantics description of sensitive dataset.

Therefore, we set $k_1 = k_2 = k$ for our PRIVIMAGE. Although the example from Figure 2 is not very general. We find that this setting works well in our experiments and reduces the number of hyper-parameters of PRIVIMAGE.

C Over-parameterized Models are Ineffective

We conduct experiments to verify that for DP image synthesis, where the gradient is noisy, fine-tuning an over-parameterized generative model on the sensitive dataset is not a proper way. Although previous studies have stated a similar view [15, 41, 67], none of them has conducted experiments to show that this phenomenon does exist in diffusion models.

Specifically, we train diffusion models with different parameters 1.75M, 6.99M, 27.9M, 62.7M, 112M, 251M on `MNIST` [42]. `MNIST` contains 60,000 handwritten digits gray images, from 0 to 9, and has been used in more previous DP image synthesis studies [3, 25]. The gradient used to update the models' parameters is added with Gaussian noise for satisfying $(10, 1 \times 10^{-5})$-DP. We use the FID and CA of synthetic images from trained models to assess their synthesis quality.

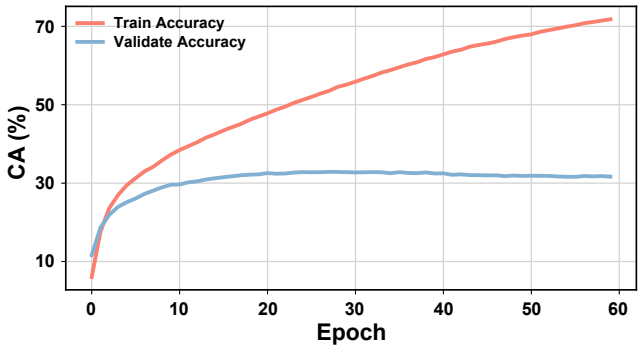

Figure 9: The semantic classification accuracy of the semantic query function on the training and validation set of `ImageNet` [14] during training. The best accuracy on the validation set is 32.9%.

Figure 8 shows the results. When the parameters of diffusion models are very few (e.g. 1.75M), both FID and CA of the synthetic images are very poor. Because such few parameters are not enough for diffusion models to learn the distribution of the sensitive data, thus cannot generate images with high fidelity and utility. When the parameters are between 27.9M and 62.7M, the diffusion models seem to achieve optimal synthesis performance for the lowest FID and highest CA. However, when the parameters get more than 62.7M, both FID and CA get worse with the parameters increase. It is notable that the parameters of SOTA method are 80M. With such a number of parameters, the dimension of gradient gets large, which is called 'curse of dimensionality'. As shown in Algorithm 1, when we scale the gradient into a given maximal l_2-norm C, the scaled gradient added by Gaussian noise with the same variance typically becomes more noisy, leading more unstable training of diffusion models. Therefore, light models are more suitable for DP image synthesis.