Differentially Private Synthetic Data via APIs 3: Using Simulators Instead of Foundation Models

Abstract

Differentially private (DP) synthetic data enables the use of sensitive datasets while providing strong privacy guarantees. Recently, Private Evolution (PE) has emerged as a promising framework for generating DP synthetic data using only inference APIs of foundation models. However, suitable foundation models are not always available for every private data domain. In this paper, we show that PE is more general than previously understood and can incorporate APIs beyond foundation models. In particular, we show that state-of-the-art *non–neural-network data synthesizers*, including computer graphics-based image generators and physics simulation tools (which we refer to as *simulators*), integrate naturally into PE. We call the resulting method Sim-PE. **Sim-PE broadens the applicability of PE while improving performance**. For image synthesis, Sim-PE improves downstream classification accuracy by up to 3×, reduces FID by up to 80%, and is significantly more efficient. We further show that simulators and foundation models can be seamlessly combined within Sim-PE for additional gains. Beyond performance improvements, **Sim-PE enables new DP applications that are difficult or impossible for existing methods**, including generating DP control signals whose execution reproduces the distribution of private observations.

1 Introduction

Leaking sensitive user information is a major concern in data-driven applications. A common solution is to generate differentially private (DP) (Dwork et al., 2006) synthetic data that resembles the original while ensuring strong privacy guarantees. Such data can substitute the original in tasks like model fine-tuning, statistical analysis, and data sharing, while preserving user privacy (Bowen & Snoke, 2019; Lin, 2022; Tao et al., 2021; Hu et al., 2024).

Private Evolution (PE) (Lin et al., 2023; Xie et al., 2024) has recently emerged as a promising method for DP data synthesis, adopted in Microsoft and Apple (Apple, 2025; Afonja et al., 2024). It begins by probing foundation models to produce random samples, then iteratively selects those most similar to private data and uses the model to generate more like them. Unlike prior methods that fine-tune open-source models, PE relies solely on model inference–making it up to 66× faster (Xie et al., 2024). More importantly, this allows PE to easily leverage cutting-edge blackbox models like GPT-5 (Singh et al., 2025), achieving SoTA performance on multiple image and text benchmarks (Lin et al., 2023; Xie et al., 2024; Hou et al., 2024; Zou et al., 2025; Hou et al., 2025; Wang et al., 2025a;b; Chien et al., 2026a).

However, PE relies on foundation models suited to the private data domain, which may not always be available. When the model's distribution significantly differs from the private data, PE's performance lags far behind training-based methods (Gong et al., 2025).

To address this question, we note that in the traditional synthetic data field—where private data is *not* involved—*non-neural-network data synthesizers* remain widely used, especially in domains where foundation models struggle. Examples include computer graphics-based renders for images, videos, and 3D data (e.g., Blender (Community, 2018) and Unreal (Epic Games)) and physics-based simulators for robotics data (e.g., Gym/Gymnasium (Towers et al., 2024; Brockman et al., 2016), Genesis (Authors, 2024)). For brevity, we refer to these tools as *simulators*. While these simulators have been successful, their applications in *DP* data

Figure 1: Unlike prior PE work that relies solely on foundation models, we show that PE is also compatible with non-neural-network data synthesis tools, which we call *simulators*. This greatly broadens PE's applicability, enables SoTA simulators for DP data synthesis, and unlocks new DP applications.

synthesis remain underexplored. This is understandable, as adapting these simulators to fit private data in a DP fashion requires non-trivial, case-by-case modifications. Our key insight is that PE only requires two APIs: RANDOM_API that generates random samples and VARIATION_API that generates samples similar to the given one. These APIs do not have to come from foundation models. Thus, we ask: *Can PE use simulators in place of models?* If viable, this approach could greatly expand PE's capabilities and unlock the potential of a wide range of simulators for DP data synthesis.

In this paper, we propose Sim-PE (Fig. 1) to exploit this potential. While the underlying idea is conceptually simple, its correct realization requires substantial effort. Through systematic exploration, we arrive at the following design, which distinguishes between two types of simulator access: **(1) The simulator is accessible.** We define RANDOM_API as generating a sample with random simulator parameters, and VARIATION_API as slightly perturbing the simulator parameters of the given sample. **(2) The simulator is inaccessible—only its generated data is released.** This scenario is quite common (Wood et al., 2021; Bae et al., 2023), especially when simulator assets are proprietary (Kar et al., 2019; Devaranjan et al., 2020). In this case, we define RANDOM_API as randomly selecting a sample from the dataset, and VARIATION_API as randomly selecting a nearest neighbor of the given sample. We demonstrate that with suitable simulators, Sim-PE can outperform PE with models. Moreover, Sim-PE enables new applications that are difficult or infeasible for prior methods. Our key contributions are:

- **Advancing PE. (1)** We discover that PE can leverage tools beyond foundation models and propose Sim-PE, which integrates simulators into PE. This integration alleviates two key limitations of PE. **(a) Improved performance under large distribution shift.** PE performs poorly when the private data distribution deviates significantly from the pretraining data of foundation models. By incorporating simulators, Sim-PE achieves substantially better performance in such settings. For example, on the MNIST dataset with $\epsilon = 1$, downstream classification accuracy improves to 89.1%, compared to 27.9% with the original PE. **(b) Improved efficiency.** As an additional benefit, Sim-PE reduces the high cost of repeated foundation model API calls in PE by offloading computation to potentially cheaper simulators. For example, on MNIST, Sim-PE running on CPU is 80× faster than PE on GPU, while also benefiting from the lower cost of CPU computation.
  (2) We further introduce a unified framework that allows *models and simulators to be used interchangeably*, enabling PE to leverage the complementary strengths of both and yielding additional performance gains.
- **Unlocking new DP applications.** We show that Sim-PE supports applications in which the generation target and the private observations lie in *different* spaces, such as *generating DP control signals whose execution in a robotics simulator reproduces the distribution of private observations.* **This setting is fundamentally challenging for prior training-based methods** for two reasons: (1) they are designed for scenarios in which the generation target and observations reside in the same space, and (2) they are gradient-based, making them incompatible with simulators through which backpropagation is not possible.

2 Preliminaries and Motivation

2.1 Preliminaries

Synthetic data refers to "fake" data generated by models or software for various applications, including data augmentation, model training, and software testing (Lin, 2022). While neural-network-based generative models such as GANs (Goodfellow et al., 2020), diffusion models (Sohl-Dickstein et al., 2015), and auto-regressive models (OpenAI, 2023; Liu et al., 2024a; 2025) are widely used, non-neural-network tools remain SoTA in many applications. For example, ns (Issariyakul et al., 2009; Riley & Henderson, 2010) can simulate networks and generate network packets based on network configurations. Blender (Community, 2018), given 3D models and lighting configurations, can render images and videos, and is extensively used in movie production. **In this paper, we refer to these tools as *simulators*.** See App. H.1 for a discussion on the continued importance of simulators, even in the era of large foundation models.

DP synthetic data requires the synthetic data to be close to a given private dataset, while having a differential privacy (DP) (Dwork et al., 2006) guarantee. Formally, a mechanism $\mathcal{M}$ is (ϵ, δ)-DP if for any two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ (i.e., $\mathcal{D}'$ has one extra entry compared to $\mathcal{D}$ or vice versa) and for any set S of outputs of $\mathcal{M}$, we have $\mathbb{P}(\mathcal{M}(\mathcal{D}) \in S) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(\mathcal{D}') \in S) + \delta$. Current SoTA DP image and text synthesis methods typically requires neural network training (Lin et al., 2020a; Beaulieu-Jones et al., 2019; Dockhorn et al., 2022; Yin et al., 2022; Yu et al., 2021; He et al., 2022; Li et al., 2021; Ghalebikesabi et al., 2023a; Yue et al., 2022; Jordon et al., 2019; Harder et al., 2023; 2021b; Vinaroz et al., 2022; Cao et al., 2021; Chen et al., 2022).

Private Evolution (PE) (Lin et al., 2023; Xie et al., 2024) is a recent training-free framework for DP data synthesis. Therefore, unlike prior training-based methods, PE can leverage state-of-the-art models even when they are accessible only via APIs (e.g., GPT-5), while also being more computationally efficient (Lin et al., 2023; Xie et al., 2024). In more detail, **(1) Performance.** PE and its follow-up methods have achieved state-of-the-art performance on several *image* and *text* benchmarks (Lin et al., 2023; Xie et al., 2024; Wang et al., 2025a;b; Chien et al., 2026b). *When using comparable open-source pre-trained models*, PE achieves an *image* quality score of FID ≤ 7.9 on CIFAR10 with a privacy cost of $\epsilon = 0.67$, representing a significant improvement over the previous state-of-the-art, which required $\epsilon = 32$ (Lin et al., 2023). *By leveraging state-of-the-art models behind APIs—where training-based methods are not applicable—*PE further improves performance, outperforming all prior approaches in downstream *text* classification accuracy on the OpenReview dataset (Xie et al., 2024). **(2) Efficiency.** Furthermore, since PE only requires model inference, it can be up to $66\times$ more efficient than training-based methods for DP *text* generation (Xie et al., 2024). **(3) Broader applications.** Additionally, PE can be applied to *federated learning* to shift model training from devices to central servers in a differentially private and more efficient manner (Hou et al., 2024; Zou et al., 2025; Hou et al., 2025). While PE itself does not require model training, it can also be extended to fine-tune model weights via reinforcement (Xu et al., 2026; Hou et al., 2025) or supervised (Zou et al., 2026) fine-tuning. PE has been adopted by some of the largest IT companies, including Microsoft and Apple (Apple, 2025; Afonja et al., 2024).

2.2 Motivation

While PE is promising, its performance significantly drops when there is a large distribution shift between the private data and the foundation model's pre-trained data (Gong et al., 2025). For example, using MNIST (LeCun, 1998) as private data, training a downstream digit classifier (10 classes) on DP synthetic data generated by PE with $\epsilon = 1$—via a foundation model pre-trained on ImageNet—achieves only 27.9% accuracy (Gong et al., 2025). Because suitable foundation models may not exist for every domain, this limitation restricts PE's applicability in practice. Extending PE to leverage simulators can substantially broaden its scope.

More broadly, simulators cannot be replaced by foundation models for data synthesis in many domains (§ 2.1 and App. H.1), yet many SoTA DP synthetic data methods rely heavily on (training) models and cannot leverage simulators. By extending PE to simulators, we aim to unlock their potential for DP data synthesis and enable new applications infeasible for prior training-based methods.

3 Sim-PE: Private Evolution with Simulators

While the idea is conceptually simple, its implementation requires careful design. We first present an overview of the Sim-PE algorithm (§ 3.1), and then address three key questions: (1) How can PE be adapted to simulators (§ 3.2)? (2) Can the strengths of models and simulators be leveraged jointly (§ 3.3)? (3) Beyond performance improvements, what unique capabilities does Sim-PE provide (§ 3.4)?

3.1 Overview

Key insight. PE requires a data generation backend that supports (1) RANDOM_API, which generates a random sample (e.g., a random bird image), and (2) VARIATION_API, which produces slight variations of a given sample (e.g., a similar bird image). Prior work on PE (Lin et al., 2023; Xie et al., 2024; Hou et al., 2024; Zou et al., 2025; Hou et al., 2025; Swanberg et al., 2025; Wang et al., 2025a;b) has exclusively used foundation models to implement these APIs. Our key insight is that *these APIs do not have to be powered by foundation models*: traditional data synthesizers that do not rely on neural networks—referred to as *simulators*—can also be used to implement RANDOM_API and VARIATION_API.

Algorithm overview. Here, we describe the full workflow of Sim-PE. For completeness, the pseudocode is attached in App. A. We first use RANDOM_API to generate an initial set of random samples (Line 4). Then, we iteratively refine these samples using the private data. In each iteration:

- Each private sample casts a vote for its closest synthetic sample. This yields a histogram (denoted DP_NN_HISTOGRAM) reflecting how well each synthetic sample aligns with the private data (Line 6). To ensure differential privacy, Gaussian noise is added to it.
- We sample synthetic data according to the noisy histogram, giving higher likelihood to those samples that align more closely with private data (Lines 7 and 8).
- We apply VARIATION_API to the drawn samples to generate additional variants (Line 9). These samples become the initialization of the next iteration.

Final-iteration samples form the DP synthetic dataset.

Theoretical analysis. Sim-PE provides strict DP guarantees with respect to the private dataset (including labels, if available) and guarantees convergence under assumptions in Lin et al. (2023). We defer the privacy and convergence analysis, adapted from PE (Lin et al., 2023), to App. B.

3.2 Simulator Adaptation in Sim-PE

Given the setup in § 3.1, the core challenge is to design RANDOM_API and VARIATION_API for simulators. In practice, simulators expose different levels of access.

Simulator code access. Some simulators are open-sourced. Examples include KUBRIC (Greff et al., 2022), a Blender-based renderer for multi-object images/videos; 3D TEAPOT (Lin et al., 2020b; Eastwood & Williams, 2018), an OpenDR-based renderer for teapot images; and PYTHON-AVATAR (Escartín, 2021), a rule-based generator for avatars.

Data-only access. Since the assets in simulators (e.g., 3D models in Blender) are often proprietary, many works release only generated datasets without the simulator code. Examples include the FACE SYNTHETICS (Wood et al., 2021) and the DIGIFACE-1M (Bae et al., 2023) datasets, both produced with Blender-based renderers for human faces.

In § 3.2.1 and 3.2.2, we present designs under simulators with code access and data-only access, respectively.

3.2.1 Sim-PE with Simulator Access

While different simulators have very different programming interfaces, most of them can be abstracted in the same way. A simulator generates samples based on a set of input *parameters*. These parameters are *not* part of the generated samples themselves; rather, they act as *control signals* that determine certain properties of

the outputs.[1] For example, face image renderers in Wood et al. (2021); Bae et al. (2023) generate images conditioned on parameters such as facial pose, lighting intensity, and the IDs of the underlying 3D face models and hairstyles.

These parameters can be further categorized into two types, depending on whether their values admit a natural ordering. *Numerical parameters* refer to those with meaningful ordering (e.g., 1 is closer to 2 than to 3), whether discrete (e.g., $\in \{0, 1, 2\}$) or continuous (e.g., $\in [0, 2]$). In contrast, *categorical parameters* are those *without* inherent ordering, or where even a small change in the parameter can cause a large, unrelated change in the output. In the face renderer example above, facial pose and lighting intensity are numerical parameters, as their magnitudes are meaningful and variations lead to smooth changes in the generated images. In contrast, the IDs of the 3D face models and hairstyles are categorical parameters, since different IDs correspond to distinct entities without any notion of proximity or ordering.

Under this formulation, a simulator $\mathcal{S}$ generates a sample denoted by $\mathcal{S}(\xi_1, \ldots, \xi_p, \phi_1, \ldots, \phi_q)$, where $\xi_1, \ldots, \xi_p \in \Xi_1 \times \cdots \times \Xi_p$ are p *categorical* parameters and $\phi_1, \ldots, \phi_q \in \Phi_1 \times \cdots \times \Phi_q$ are q *numerical* parameters.

For RANDOM_API, we simply draw each parameter randomly from its corresponding feasible set:

$$\text{RANDOM\_API} = \mathcal{S}(\xi_1, \ldots, \xi_p, \phi_1, \ldots, \phi_q), \qquad \text{where } \xi_i \sim \text{Uniform}(\Xi_i) \text{ and } \phi_i \sim \text{Uniform}(\Phi_i). \tag{1}$$

Here, Uniform (S) denotes drawing a sample uniformly at random from the set S.

For VARIATION_API, we generate variations by perturbing the input simulator parameters. For numerical parameters ϕ_i, we simply add noise. However, for categorical parameters ξ_i, where no natural ordering exists among feasible values in Ξ_i, adding noise is not applicable. Instead, we re-draw the parameter from the entire feasible set Ξ_i with a certain probability. Formally, it is defined as

$$\text{VARIATION\_API}(\mathcal{S}(\xi_1, \ldots, \xi_p, \phi_1, \ldots, \phi_q)) = \mathcal{S}(\xi_1', \ldots, \xi_p', \phi_1', \ldots, \phi_q'), \tag{2}$$

$$\text{where } \phi_i' \sim \text{Uniform}([\phi_i - \alpha, \phi_i + \alpha] \cap \Phi_i) \text{ and } \xi_i' = \begin{cases} \text{Uniform}(\Xi_i), & \text{with probability } \beta \\ \xi_i, & \text{with probability } 1 - \beta \end{cases}.$$

Here, α and β control the degree of variation. At one extreme, when $\alpha = \infty$ and $\beta = 1$, VARIATION_API completely discards the information of the input sample and reduces to RANDOM_API. Conversely, when $\alpha = \beta = 0$, VARIATION_API outputs the input sample unchanged.

3.2.2 Sim-PE with Simulator-generated Data

We assume a simulator-generated dataset of m samples, $S_{\text{sim}} = \{z_1, \ldots, z_m\}$. The goal is to select N_{syn} samples to form the DP synthetic dataset S_{syn}. Although this problem appears simple, it requires careful design. We illustrate this by explaining why two straightforward approaches fail.

Baseline 1: Applying DP_NN_HISTOGRAM on S_{syn}. One immediate solution is to apply DP_NN_HISTOGRAM in PE (Alg. 2) by treating S_{sim} as the generated set S. In other words, each private sample votes for its nearest neighbor in S_{sim}, and the final histogram, aggregating all votes, is privatized with Gaussian noise. We then draw samples from S_{sim} according to the privatized histogram (i.e., Line 8 in Alg. 1) to obtain S_{syn}.

However, the size of the simulator-generated dataset (i.e., m) is typically very large (e.g., 1.2 million in Bae et al. (2023)), and the total amount of added Gaussian noise grows with m. This means that the resulting histogram suffers from a low signal-to-noise ratio, leading to poor fidelity in S_{syn}.

Baseline 2: Applying DP_NN_HISTOGRAM on cluster centers of S_{syn}. To improve the signal-to-noise ratio of the histogram, one solution is to have private samples vote on the cluster centers of S_{sim} instead of the raw samples. Specifically, we first cluster the samples in S_{sim} into N_{cluster} clusters with centers

---

[1] For well-documented simulators, the list of parameters is typically readily available. For example, PYTHON-AVATAR, used in § 5.1.1, explicitly documents its parameters in the README. Alternatively, one can adopt the approach in § 3.2.2, which does not require explicit identification of parameters.

$\{w_1, \ldots, w_{N_{\text{cluster}}}\}$ and have private samples vote on these centers rather than individual samples in S_{sim}.[2] Since the number of bins in the histogram decreases from m to N_{cluster}, the signal-to-noise ratio improves. Following the approach of the previous baseline, we then draw N_{syn} cluster centers (with replacement) based on the histogram and randomly select a sample from each chosen cluster to construct the final S_{syn}.

However, when the total number of samples m is large, each cluster may contain a diverse set of samples, including those both close to and far from the private dataset. While DP voting on clusters improves the accuracy of the DP histogram and helps select better clusters, there remains a risk of drawing unsuitable samples from the chosen clusters.

Our approach. Our key insight is that the unavoidable trade-off between DP histogram accuracy and selection precision (clusters vs. individual samples) stems from forcing private samples to consider all of S_{sim}—either directly (baseline 1) or via cluster centers (baseline 2). But this is not necessary: if a sample is far from the private data, its nearest neighbors in S_{sim} are likely far too (see App. C for experiments). Thus, we can avoid wasting privacy budget on evaluating such samples.

The iterative selection and refinement process in PE naturally aligns with this idea. For each sample z_i, we define its nearest neighbors in S_{sim} as $q_1^i, \ldots, q_m^i$, sorted by closeness, where $q_1^i = z_i$ is the closest. We define RANDOM_API as drawing a random sample from S_{sim}:

$$\text{RANDOM\_API} \sim \text{Uniform}\left(S_{\text{sim}}\right).$$

Since we draw only N_{syn} samples (instead of all m) from RANDOM_API, the DP histogram has a higher signal-to-noise ratio. In the following steps (Lines 6 to 8 in Alg. 1), we discard samples far from the private data indicated by the DP histogram, and apply VARIATION_API only to the remaining as follows:

$$\text{VARIATION\_API}\left(z_i\right) = \text{Uniform}\left(\left\{q_1^i, \ldots, q_\gamma^i\right\}\right),$$

thus avoiding consideration of nearest neighbors of the removed samples (unless they are also nearest neighbors of retained samples). Similar to α and β in § 3.2.1, γ controls the degree of variation. At one extreme, when $\gamma = m$, VARIATION_API disregards the input sample and reduces to RANDOM_API. At the other extreme, when $\gamma = 1$, VARIATION_API returns the input sample unchanged.

We verify the effectiveness of these designs in App. F.1.

Broader applications. While our main experiments (§ 5) focus on simulator-generated data, the proposed algorithm also applies to any public datasets, as shown in App. E.2.

3.3 Sim-PE with both Simulators and Models

As discussed in § 2.1, simulators and foundation models complement each other across different data domains. Moreover, even within a single domain, they excel in different aspects. For example, computer graphics-based face image generators (Bae et al., 2023; Wood et al., 2021) allow controlled diversity in race, lighting, and makeup while mitigating potential biases in foundation models. However, the generated faces may appear less realistic than those from SoTA foundation models. Thus, combining the strengths of both methods for DP data synthesis is highly appealing.

We find that Sim-PE naturally supports this integration, as RANDOM_API and VARIATION_API work the same for both foundation models and simulators. Specifically, we explore a simple strategy: using simulators in the early PE iterations to generate diverse seed samples, then switching to foundation models in later iterations to refine details and enhance realism. As shown in § 5, this approach outperforms using either simulators or foundation models alone. We discuss more alternative approaches in App. H.4.

3.4 Enabling New Applications with Sim-PE

Prior work (§ 2.1) typically assumes that the private observations and the target data to be generated lie in the same domain (e.g., generating face images from private ones). However, in many applications, the

---

[2]Note that voting in PE (Lin et al., 2023) is conducted in the embedding space. Here, w_is represent cluster centers in the embedding space, and each private sample uses its image embedding to find the nearest cluster center.

observed data is *not* the ideal generation target. Instead, the goal is to generate *latent control signals* whose execution in a system reproduces the observed behavior. For example, given recordings of human behavior or gameplay, one may wish to generate robot or player control signals rather than the observations themselves. Such unobserved control signals are more useful for downstream tasks such as training robot models and game-playing agents (Argall et al., 2009; Hester et al., 2018).

Existing training-based methods (§ 2.1), which assume that the private and generated data lie in the same space, are *not* applicable to this setting. A natural extension—feeding generated samples into the system (e.g., a robot or game engine) and defining a loss on the resulting outputs—fails because the system is generally non-differentiable.

In contrast, Sim-PE naturally supports these applications by optimizing over simulator parameters (e.g., control signals) that are decoupled from the private observations in the simulator's output space. By treating the system itself as a simulator, we can apply the algorithm in § 3.2.1 to these settings. We demonstrate this in § 5.4.

4 Evaluation Setup

We consider two settings: (1) the traditional case where the private and target generated data lie in the same domain; and (2) a new setting in which they lie in different domains, where prior methods are not directly applicable (§ 3.4).

4.1 The Traditional Application

We use widely studied *DP image generation* benchmarks.

4.1.1 Datasets and Simulators

Datasets. Following prior work (Gong et al., 2025), we use three private datasets: **(1) MNIST** (LeCun, 1998), where the image class labels are digits '0'-'9', **(2) CelebA** (Liu et al., 2015), where the image class labels are male and female, and **(3) CIFAR10** (Krizhevsky et al., 2009), which contains 10 classes of natural images. We aim at *conditional generation* for these datasets (i.e., each generated image is associated with the class label). Due to space constraints, CIFAR10 experiments are deferred to App. E.2.

Simulators. To showcase the broad applicability of Sim-PE, we use four diverse simulators.

(1) Text rendering program. Generating images with readable text using foundation models is a known challenge (Betker et al., 2023). Simulators can address this gap, as generating text images through computer programs is straightforward. To illustrate this, we use Python PIL library as the text rendering program and treat MNIST as the private dataset. There are two categorical and three numerical parameters in total (details in App. D.1). We set the feasible sets of these parameters to be large enough so that the random samples differ significantly from MNIST (see Fig. 2b).

(2) Computer graphics-based renderer for face images. Computer graphics-based rendering is widely used in real-world applications such as game development, cartoons, and movie production. This experiment aims to assess whether these advanced techniques can be adapted for DP synthetic image generation via Sim-PE. We use CelebA as the private dataset and a Blender-based face image renderer from Bae et al. (2023) as the API. Since the source code for their renderer is not publicly available, we apply our data-based algorithm from § 3.2.2 on their released dataset of 1.2 million face images. Note that this renderer may not necessarily represent the SoTA. As visualized in Fig. 3b, the generated faces exhibit various unnatural artifacts and appear less realistic than images produced by SoTA generative models (e.g., Rombach et al. (2022)). Therefore, this experiment serves as a preliminary study, and *the results could potentially improve with more advanced rendering techniques.*

(3) Rule-based avatar generator. We further investigate whether Sim-PE remains effective *when the simulator's data significantly differs from the private dataset.* We use CelebA as the private dataset and a rule-based avatar generator (Escartín, 2021) as the API. This simulator has 16 categorical parameters

that control attributes of the avatar, including eyes, noses, background colors, skin colors, etc (details in App. D.2). As in Fig. 7b, the generated avatars have a cartoon-like appearance and lack fine-grained details, which contrasts sharply with real human face photos in CelebA.

(4) Public images. To demonstrate that Sim-PE (§ 3.2.2) generalizes beyond simulator-generated data, we evaluate it using ImageNet as S_{sim}. (Results deferred to App. E.2.)

Class label information from simulators. In main experiments, we assume the simulator does not provide class label information (denoted as "**ClassUnavail**"). In App. F.8, we further investigate a setting where the simulator provides class labels alongside generated images and find that this improves performance. See App. F.8 for details and results.

4.1.2 Metrics and Evaluation Pipelines

We follow the evaluation settings of DPImageBench (Gong et al., 2025), a recent benchmark for DP image synthesis. Specifically, we use two metrics: **(1) Fidelity: FID** (Heusel et al., 2017), which computes the distribution distance between real and generated images in the embedding space; **(2) Utility: the accuracy of downstream classifiers** trained on the synthetic data. We carefully design the evaluation process so that the reported accuracy is not inflated due to train-test overlap or DP violations in classifier hyperparameter tuning. See App. H.2 for the detailed procedure and the rationale behind our choice of metrics.

Following Gong et al. (2025), we set DP parameter $\delta = 1/(N_{\text{priv}} \cdot \log N_{\text{priv}})$, where N_{priv} is the number of samples in the private dataset, and $\epsilon = 1$ or 10.

4.1.3 Baselines

We compare Sim-PE with 12 SoTA DP image synthesizers reported in Gong et al. (2025), including DP-MERF (Harder et al., 2021a), DP-NTK (Yang et al., 2023), DP-Kernel (Jiang et al., 2023), GS-WGAN (Chen et al., 2020), DP-GAN (Xie et al., 2018), DPDM (Dockhorn et al., 2023), PDP-Diffusion (Ghalebikesabi et al., 2023b), DP-LDM (Liu et al., 2024b), DP-LoRA (Tsai et al., 2024), PrivImage (Li et al., 2024), and PE with foundation models (Lin et al., 2023). Except for PE, all other baselines require model training. When experimenting with simulator-generated data, we additionally compare Sim-PE against the two baselines introduced in § 3.2.2 (see App. F.1).

Note: Different methods rely on different prior knowledge. For instance, many baselines use pre-trained models or public data (see Tab. 2 of Gong et al. (2025)), whereas Sim-PE does not. Instead, Sim-PE leverages simulators, which are not used by others. **As such, Sim-PE represents a new evaluation setting, and baseline results serve primarily to contextualize this new paradigm and inspire future work, rather than an apple-to-apple comparison.**

4.2 The New Application

Environment. We use the popular Gymnasium library[3] (Towers et al., 2024) for our experiments. Gymnasium provides a wide range of simulated environments for robotics, control, physics, and games, originally designed for developing and benchmarking reinforcement learning algorithms. These environments align well with the motivating scenarios in § 3.4.

Specifically, we use the CartPole environment in the Gymnasium library as the example. In this environment, a robot cart balances a pole placed vertically on top of it. At each timestep, the *observation* consists of four quantities: cart position, cart velocity, pole angle, and pole angular velocity. A *control signal* moves the cart left or right to keep the pole balanced. The episode terminates when the pole angle exceeds $\pm 12°$ or the cart moves outside the display. The task is considered successful if the pole remains balanced for 500 frames. See App. D.4 for more implementation details.

---

[3]https://github.com/Farama-Foundation/Gymnasium, previously known as the Gym library (Brockman et al., 2016) by OpenAI.

(a) Real (private) images (b) Simulator-generated images (c) Sim-PE images ($\epsilon = 10$)

Figure 2: The real and generated images on MNIST under the "**ClassUnavail**" setting. Each row corresponds to one class. The simulator generates images that are very different from the real ones and are from the incorrect classes. Starting from these bad images, Sim-PE can effectively guide the generation of the simulator towards high-quality images with correct classes.

Sim-PE adoption. We follow the approach described in § 3.4. In particular, we use four numerical parameters that map *observations* to *control signals* via a linear function. See App. D.4 for additional implementation details.

Metrics. We propose two metrics: **(1) Utility**: For each generated policy, we run it in the simulator to determine whether it succeeds or fails, compare this outcome to the assigned label, and report the accuracy. This metric directly aligns with the objective of the environment—successfully balancing the pole. **(2) Fidelity**: We compute the Wasserstein distance between the distributions of episode lengths (i.e., observation sequence lengths) in the generated and private datasets. This provides a more fine-grained measure than utility alone, as episode lengths can vary significantly even among failed policies.

5 Results

We focus on six research questions:
(1) Does Sim-PE improve upon PE (§ 5.1)?
(2) Does Sim-PE work with *bad* simulators (§ 5.2)?
(3) Can Sim-PE leverage the strengths of both simulators and foundation models (§ 5.3)?
(4) How does Sim-PE perform in new application settings where prior methods are inapplicable (§ 5.4)?
(5) How efficient is Sim-PE (§ 5.5)?
(6) How sensitive is Sim-PE to hyperparameters, and how does it perform without embedding models (§ 5.5)?

5.1 Sim-PE Improves PE

5.1.1 Sim-PE with Simulator Access

We evaluate Sim-PE with a text rendering program on MNIST. The results (Tab. 1 and Fig. 2) show that:

Sim-PE effectively guides the simulator to generate high-quality samples. As shown in Fig. 2b, without any information from the private data or guidance from Sim-PE, the simulator initially produces poor-quality images with incorrect digit sizes, rotations, and stroke widths. These low-quality samples serve as the starting point for Sim-PE (via RANDOM_API). Through iterative refinement and private data voting, Sim-PE gradually optimizes the simulator parameters, ultimately generating high-quality MNIST samples, as illustrated in Fig. 2c.

Quantitative results in Tab. 1 further support this. Without private data guidance, the simulator often generates digits from incorrect classes, yielding a classifier accuracy of just 11.6%–near random guess. In contrast, Sim-PE boosts accuracy to around 90%. FID scores also confirm that Sim-PE produces images more similar to real data.

Sim-PE can improve the performance of PE by a large margin. The PE baseline (Lin et al., 2023) uses a diffusion model pre-trained on ImageNet, which primarily contains natural object images (e.g., plants,

Table 1: Accuracy and FID. The best between PE methods is in **bold**, and the best between all methods is underlined. "Simulator" refers to samples from the simulator's RANDOM_API. Results other than Sim-PE and Simulator are taken from Gong et al. (2025).

(a) Accuracy (%) of downstream classifiers.

Algorithm	MNIST		CelebA	
	$\epsilon = 1$	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 10$
DP-MERF	80.3	81.3	81.0	81.2
DP-NTK	50.0	91.3	61.2	64.2
DP-Kernel	94.0	93.6	83.0	83.7
GS-WGAN	72.4	75.3	61.4	61.5
DP-GAN	92.4	92.7	77.9	89.2
DPDM	89.2	97.7	74.5	91.8
PDP-Diffusion	94.5	97.4	89.4	94.0
DP-LDM (SD)	78.8	94.4	84.4	89.1
DP-LDM	44.2	95.5	85.8	92.4
DP-LoRA	82.2	97.1	87.0	92.0
PrivImage	94.0	97.8	90.8	92.0
Simulator	11.6 ($\epsilon = 0$)		61.4 ($\epsilon = 0$)	
PE	27.9	32.7	70.5	74.2
Sim-PE (ours)	**89.1**	**93.6**	**80.0**	**82.5**

(b) FID of synthetic images.

Algorithm	MNIST		CelebA	
	$\epsilon = 1$	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 10$
DP-MERF	113.7	106.3	176.3	147.9
DP-NTK	382.1	69.2	350.4	227.8
DP-Kernel	33.7	38.9	140.3	128.8
GS-WGAN	57.0	47.7	611.8	290.0
DP-GAN	82.3	30.3	112.5	31.7
DPDM	36.1	4.4	153.99	28.8
PDP-Diffusion	8.9	3.8	17.1	8.1
DP-LDM (SD)	31.9	18.7	46.2	24.1
DP-LDM	155.2	99.1	124.1	40.4
DP-LoRA	112.8	95.4	53.3	32.2
PrivImage	7.6	2.3	11.4	11.3
Simulator	86.2 ($\epsilon = 0$)		37.2 ($\epsilon = 0$)	
PE	48.8	45.3	**23.4**	22.0
Sim-PE (ours)	**20.7**	**9.4**	24.7	**20.8**

Table 2: Accuracy (%) of classifiers trained on synthetic images and FID of synthetic images on CelebA. The best results are highlighted in bold. Using a combination of both (weak) simulators and foundation models outperforms using either one alone.

Algorithm	FID ↓		Classification Acc. ↑	
	$\epsilon = 1$	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 10$
PE with foundation models	23.4	22.0	70.5	74.2
PE with weak simulators (i.e., Sim-PE)	101.4	99.5	62.6	63.2
PE with both	**15.0**	**11.9**	**72.7**	**78.1**

(a) Real (private) images   (b) Simulator-generated images   (c) Sim-PE images ($\epsilon = 10$)

Figure 3: The real and generated images on CelebA. The top rows correspond to the "female" class, and the bottom rows correspond to the "male" class. The simulator generates images with incorrect classes. However, starting from these misclassified images, Sim-PE effectively selects those that better match the correct class.

animals, cars). Since MNIST differs significantly from such data, PE, as a training-free method, struggles to generate meaningful MNIST-like images. Most PE-generated images lack recognizable digits (see Gong et al. (2025)), resulting in a classification accuracy of only ∼ 30% (Tab. 1a). By leveraging a simulator better suited for this domain, Sim-PE achieves much better results, tripling the classification accuracy and reducing the FID by 80% at $\epsilon = 10$.

Sim-PE achieves competitive results among SoTA methods. When the foundation model or public data differs significantly from the private data, training-based baselines can still adapt the model to the private data distribution by updating its weights, whereas PE cannot. This limitation accounts for the substantial performance gap between PE and other methods. Specifically, PE records the lowest classification accuracy among all 12 methods (Tab. 1a). By leveraging domain-specific simulators, Sim-PE substantially narrows this gap, achieving classification accuracy within 5.4% and 4.2% of the best-performing method for $\epsilon = 1$ and $\epsilon = 10$, respectively.

5.1.2 Sim-PE with Simulator-generated Data

We evaluate Sim-PE on CelebA with a computer graphics-based renderer. The results (Tab. 1 and Fig. 3) show that:

Sim-PE selects samples that better match target classes. Without private data, the simulator naturally generates images with incorrect labels (Fig. 3b). As a result, a gender classifier trained on this data achieves at most 61.4% accuracy–the majority class (female) rate. Sim-PE iteratively refines selection from this noisy pool, ultimately choosing samples more aligned with target classes (Fig. 3c), boosting accuracy by up to 21.1% (Tab. 1a).

Sim-PE maintains the strong data quality of PE. Sim-PE improves PE's classification accuracy by a large margin (Tab. 1a) and maintains the good FID (Tab. 1b). Unlike in MNIST (§ 5.1.1) where Sim-PE brought large gains, the modest FID improvement on CelebA stems from two factors. First, PE with foundation models already ranks 3rd in FID, leaving little room to improve. Second, Sim-PE here only selects from a fixed simulator-generated dataset. As seen in Fig. 3b, these images differ from real CelebA (e.g., having larger faces), and it is impossible for Sim-PE to correct such discrepancies. Having access to simulator code, as in § 5.1.1, could help alleviate such errors by parameter adjustments. A hybrid approach combining foundation models and simulators to be shown in § 5.3 may also help.

5.2 Sim-PE with Bad Simulators

Here, we examine Sim-PE under bad simulators. As in the previous section, we use CelebA as the private dataset, but replace the simulator with a rule-based cartoon avatar generator (Escartín, 2021). As shown in Fig. 7b, the generated avatars differ significantly from real images.

Sim-PE with bad simulators still learns useful features. From Tab. 2, we observe that downstream classifiers trained on Sim-PE with bad simulators achieve poor classification accuracy. However, two interesting results emerge: **(1)** Despite the significant difference between avatars and real face images, Sim-PE still captures certain characteristics of the two classes correctly. Specifically, Sim-PE tends to generate faces with long hair for the female class and short hair for the male class (Fig. 7c). **(2)** Although the FID of Sim-PE is quite poor (Tab. 2), they still outperform many baselines (Tab. 1b). This can be explained by the fact that, as shown in Gong et al. (2025), high DP noise makes the training of many baselines unstable. This results in images with noisy patterns, non-face images, or significant mode collapse, particularly for DP-NTK, DP-Kernel, and GS-WGAN. In contrast, Sim-PE is training-free, and thus it avoids these issues. See App. E.1 for more analysis.

5.3 Sim-PE Can Leverage both Simulators and Models

Next, we explore the feasibility of using PE with both foundation models and the bad avatar simulator (§ 3.3).

Sim-PE benefits from utilizing simulators and foundation models together. We observe (Tab. 2) that using both simulators and foundation models yields the best results. This result is intuitive: the foundation model, pre-trained on the diverse ImageNet dataset, has a low probability of generating faces via RANDOM_API. While avatars differ from CelebA, they retain the correct image layout, such as facial boundaries, eyes, and nose. Using these avatars as seed samples for variation allows the foundation model to focus on images closer to real faces rather than random patterns.

Unlike other SoTA methods tied to a specific data synthesizer, this result suggests that PE is a promising framework that can combine the strengths of multiple data synthesizers.

5.4 New Applications Enabled by Sim-PE

Since prior methods are not directly applicable (§ 3.4), we focus on evaluating the properties of Sim-PE.

Sim-PE's data quality improves across iterations. Fig. 4 show that both utility and fidelity metrics steadily improve over time. More specifically: (1) since random policies fail roughly 99% of the time, the initial accuracy is near 50%. Sim-PE progressively discovers policies that match the success/failure patterns of the private data, ultimately reaching 76% accuracy. (2) The Wasserstein distance also decreases by roughly $4\times$ over the iterations. These results demonstrate that Sim-PE is effective even in this new application.

(a) Accuracy vs. Sim-PE iteration

(b) Wassertein distance vs. Sim-PE iteration

Figure 4: The quality of Sim-PE's generated data improves across iterations in the CartPole environment ($\epsilon = 1$).

(a) Accuracy vs. epsilon

(b) Wassertein distance vs. epsilon

Figure 5: The quality of Sim-PE's generated data improves with more privacy budget in the CartPole environment.

Sim-PE's data quality improves as the privacy budget increases. We vary the privacy budget ϵ from 0.2 to 8. Fig. 5 shows a clear trend: performance improves as the privacy budget increases. With $\epsilon = 8$, the accuracy reaches 96%, a very strong result. This confirms that Sim-PE can leverage larger privacy budgets effectively when they are available.

5.5 Further Analyses and Discussions

Sim-PE converges quickly. Like PE, Sim-PE achieves good results in a few (e.g., 6) iterations (Fig. 10).

Sim-PE is efficient. Since simulators could be cheaper to run than foundation models, Sim-PE could be much more efficient than PE. For instance, on MNIST, each PE's API call takes over *2400 GPU seconds*, whereas Sim-PE takes less than *30 CPU seconds*–an 80x speedup, not to mention the lower cost of CPU than GPU. (More results in App. G.)

Embedding model. Computing the sample distance in Line 3 is performed in an embedding space, following Lin et al. (2023). The VARIATION_API in § 3.2.2 also retrieves nearest samples in this embedding space. We clarify the following points: **(1) Experimental fairness.** For image experiments, we use embeddings from a pre-trained Inception model, assuming its training data (ImageNet) is public and independent of the private dataset. This setup is consistent across methods: PE uses the same ImageNet-pretrained Inception embedding, and all training-based baselines also treat ImageNet as public data and rely on it similarly, typically via pretraining (Gong et al., 2025). **(2) Reasonableness of the assumption.** Sim-PE assumes access to a public embedding model, which aligns with standard assumptions in DP fine-tuning, where large pretrained models are treated as public and non-private. In practice, such pretraining often relies on massive datasets, sometimes including potentially sensitive corpora (e.g., LAION (Schuhmann et al., 2021) in Liu et al. (2024b)). In contrast, Sim-PE only depends on an Inception embedding trained on ImageNet, making this assumption comparatively weaker. **(3) Applicability without pretrained embeddings.** Sim-PE can also operate without pretrained embeddings in domains lacking public data or strong representation models. For example, in our CartPole experiments (§ 5.4), we directly use raw samples as embeddings (App. D.4), demonstrating that Sim-PE could remain effective even without any pretrained embedding.

Hyperparameter sensitivity. We conduct ablation studies in App. F, including stress tests of Sim-PE under unreasonable parameter settings. Overall, Sim-PE is robust to simulator parameter ranges, its hyperparameters (α, β, γ), and the timing of switching from simulators to foundation models, unless these settings are made excessively poor. We also study the effect of sample size and distribution alignment.

6 Limitations and Future Work

In this paper, we demonstrate the potential of the PE framework for utilizing simulators in DP data synthesis. While the results are promising, important questions remain open: **(1)** Since Sim-PE (in § 3.2.2) can be applied to any public dataset (App. E.2), it could support applications such as pre-training data selection for private fine-tuning (Yu et al., 2023; Li et al., 2024). **(2)** In domains like networking and systems, simulators are more common than foundation models, and Sim-PE could offer even greater potential. **(3)** Sim-PE for image synthesis is still outperformed by the best baseline. Exploring better ways to combine simulators and foundation models could further push the limits of the PE framework. **(4)** It remains an open question how to effectively apply PE in domains where suitable foundation models, simulators, and public datasets are all unavailable.

References

Gbola Afonja, Robert Sim, Zinan Lin, Huseyin Atahan Inan, and Sergey Yekhanin. The crossroads of innovation and privacy: Private synthetic data for generative ai. https://www.microsoft.com/en-us/research/blog/the-crossroads-of-innovation-and-privacy-private-synthetic-data-for-generative-ai/, 2024.

Apple. Understanding aggregate trends for apple intelligence using differential privacy. https://machinelearning.apple.com/research/differential-privacy-aggregate-trends, 2025.

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Genesis Authors. Genesis: A universal and generative physics engine for robotics and beyond, December 2024. URL https://github.com/Genesis-Embodied-AI/Genesis.

Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3526–3535, 2023.

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pp. 394–403. PMLR, 2018.

Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Claire McKay Bowen and Joshua Snoke. Comparative study of differentially private synthetic data algorithms from the nist pscr differential privacy synthetic data challenge. *arXiv preprint arXiv:1911.12704*, 2019.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

Tianshi Cao, Alex Bie, Arash Vahdat, Sanja Fidler, and Karsten Kreis. Don't generate me: Training differentially private generative models with sinkhorn divergence. *Advances in Neural Information Processing Systems*, 34:12480–12492, 2021.

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.

Jia-Wei Chen, Chia-Mu Yu, Ching-Chia Kao, Tzai-Wei Pang, and Chun-Shien Lu. Dpgen: Differentially private generative energy-guided network for natural image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8387–8396, 2022.

Eli Chien, Yuzheng Hu, Ryan McKenna, Shanshan Wu, Zheng Xu, and Peter Kairouz. Maple: Metadata augmented private language evolution. *arXiv preprint arXiv:2603.19258*, 2026a.

Eli Chien, Yuzheng Hu, Ryan McKenna, Shanshan Wu, Zheng Xu, and Peter Kairouz. Maple: Metadata augmented private language evolution, 2026b. URL https://arxiv.org/abs/2603.19258.

Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. URL http://www.blender.org.

Jeevan Devaranjan, Amlan Kar, and Sanja Fidler. Meta-sim2: Unsupervised learning of scene structure for synthetic data generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pp. 715–733. Springer, 2020.

Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *arXiv preprint arXiv:2210.09929*, 2022.

Tim Dockhorn, Tianshi Cao, Arash Vahdat, et al. Differentially private diffusion models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=ZPpQk7FJXF.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006.

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

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations*, 2018.

Epic Games. Unreal engine. https://www.unrealengine.com. URL https://www.unrealengine.com.

Ibon Escartín. python avatars. https://github.com/ibonn/python_avatars, 2021.

Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L Smith, Olivia Wiles, and Borja Balle. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*, 2023a.

Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, et al. Differentially private diffusion models generate useful synthetic images. *CoRR*, abs/2302.13861, 2023b.

Chen Gong, Kecen Li, Zinan Lin, and Tianhao Wang. Dpimagebench: A unified benchmark for differentially private image synthesis. *arXiv preprint arXiv:2503.14681*, 2025.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Google. Google fonts. https://github.com/google/fonts, 2022.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. 2022.

Frederik Harder, Kamil Adamczewski, and Mijung Park. DP-MERF: differentially private mean embeddings with random features for practical privacy-preserving data generation. In *AISTATS*, pp. 1819–1827, 2021a.

Frederik Harder, Kamil Adamczewski, and Mijung Park. Dp-merf: Differentially private mean embeddings with randomfeatures for practical privacy-preserving data generation. In *International conference on artificial intelligence and statistics*, pp. 1819–1827. PMLR, 2021b.

Frederik Harder, Milad Jalali, Danica J Sutherland, and Mijung Park. Pre-trained perceptual features improve differentially private image generation. *Transactions on Machine Learning Research*, 2023.

Jiyan He, Xuechen Li, Da Yu, Huishuai Zhang, Janardhan Kulkarni, Yin Tat Lee, Arturs Backurs, Nenghai Yu, and Jiang Bian. Exploring the limits of differentially private deep learning with group-wise clipping. *arXiv preprint arXiv:2212.01539*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Charlie Hou, Akshat Shrivastava, Hongyuan Zhan, Rylan Conway, Trang Le, Adithya Sagar, Giulia Fanti, and Daniel Lazar. Pre-text: Training language models on private federated data in the age of llms. *arXiv preprint arXiv:2406.02958*, 2024.

Charlie Hou, Mei-Yu Wang, Yige Zhu, Daniel Lazar, and Giulia Fanti. Private federated learning using preference-optimized synthetic data. *arXiv preprint arXiv:2504.16438*, 2025.

Yuzheng Hu, Fan Wu, Qinbin Li, Yunhui Long, Gonzalo Munilla Garrido, Chang Ge, Bolin Ding, David Forsyth, Bo Li, and Dawn Song. Sok: Privacy-preserving data synthesis. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 4696–4713. IEEE, 2024.

Teerawat Issariyakul, Ekram Hossain, Teerawat Issariyakul, and Ekram Hossain. *Introduction to network simulator 2 (NS2)*. Springer, 2009.

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

James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2019.

Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4551–4560, 2019.

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

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Kecen Li, Chen Gong, Zhixiang Li, et al. PrivImage: Differentially private synthetic image generation using diffusion models with Semantic-Aware pretraining. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 4837–4854, 2024. ISBN 978-1-939133-44-1.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021.

Zinan Lin. *Data Sharing with Generative Adversarial Networks: From Theory to Practice*. PhD thesis, Carnegie Mellon University, 2022.

Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In *Proceedings of the ACM Internet Measurement Conference*, pp. 464–483, 2020a.

Zinan Lin, Kiran Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *international conference on machine learning*, pp. 6127–6139. PMLR, 2020b.

Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. Differentially private synthetic data via foundation model APIs 1: Images. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*, 2023. URL https://openreview.net/forum?id=7GbfIEvoS8.

Enshu Liu, Xuefei Ning, Yu Wang, and Zinan Lin. Distilled decoding 1: One-step sampling of image auto-regressive models with flow matching. *arXiv preprint arXiv:2412.17153*, 2024a.

Enshu Liu, Qian Chen, Xuefei Ning, Shengen Yan, Guohao Dai, Zinan Lin, and Yu Wang. Distilled decoding 2: One-step sampling of image auto-regressive models with conditional score distillation. *arXiv preprint arXiv:2510.21003*, 2025.

Michael F. Liu, Saiyue Lyu, Margarita Vinaroz, and Mijung Park. Differentially private latent diffusion models. 2024b.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Frank D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pp. 19–30, 2009.

OpenAI. Gpt-4 technical report, 2023.

George F Riley and Thomas R Henderson. The ns-3 network simulator. In *Modeling and tools for network simulation*, pp. 15–34. Springer, 2010.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Danyu Sun, Joann Qiongna Chen, Chen Gong, Tianhao Wang, and Zhou Li. Netdpsyn: synthesizing network traces under differential privacy. In *Proceedings of the 2024 ACM on Internet Measurement Conference*, pp. 545–554, 2024.

Marika Swanberg, Ryan McKenna, Edo Roth, Albert Cheu, and Peter Kairouz. Is api access to llms useful for generating private synthetic tabular data? *arXiv preprint arXiv:2502.06555*, 2025.

Yuchao Tao, Ryan McKenna, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Benchmarking differentially private synthetic data generation algorithms. *arXiv preprint arXiv:2112.09238*, 2021.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Yu-Lin Tsai, Yizhe Li, Zekai Chen, Po-Yu Chen, Chia-Mu Yu, Xuebin Ren, and Francois Buet-Golfouse. Differentially private fine-tuning of diffusion models. *arXiv preprint arXiv:2406.01355*, 2024.

Margarita Vinaroz, Mohammad-Amin Charusaie, Frederik Harder, Kamil Adamczewski, and Mi Jung Park. Hermite polynomial features for private data generation. In *International Conference on Machine Learning*, pp. 22300–22324. PMLR, 2022.

Haoxiang Wang, Zinan Lin, Da Yu, and Huishuai Zhang. Synthesize privacy-preserving high-resolution images via private textual intermediaries. *arXiv preprint arXiv:2506.07555*, 2025a.

Shuaiqi Wang, Vikas Raunak, Arturs Backurs, Victor Reis, Pei Zhou, Sihao Chen, Longqi Yang, Zinan Lin, Sergey Yekhanin, and Giulia Fanti. Struct-bench: A benchmark for differentially private structured text generation. *arXiv preprint arXiv:2509.10696*, 2025b.

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3681–3691, 2021.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, et al. Differentially private synthetic data via foundation model apis 2: Text. *arXiv preprint arXiv:2403.01749*, 2024.

Liyang Xie, Kaixiang Lin, and et al. Differentially private generative adversarial network. *CoRR*, abs/1802.06739, 2018. URL http://arxiv.org/abs/1802.06739.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Fangyuan Xu, Sihao Chen, Zinan Lin, Taiwei Shi, Sydney Graham, Pei Zhou, Mengting Wan, Alex Stein, Virginia Estellers, Charles Chen, et al. Dp-rft: Learning to generate synthetic text via differentially private reinforcement fine-tuning. *arXiv preprint arXiv:2602.18633*, 2026.

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

Yucheng Yin, Zinan Lin, Minhao Jin, Giulia Fanti, and Vyas Sekar. Practical gan-based synthetic ip header trace generation using netshare. In *Proceedings of the ACM SIGCOMM 2022 Conference*, pp. 458–472, 2022.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

Da Yu, Sivakanth Gopi, Janardhan Kulkarni, Zinan Lin, Saurabh Naik, Tomasz Lukasz Religa, Jian Yin, and Huishuai Zhang. Selective pre-training for private fine-tuning. *arXiv preprint arXiv:2305.13865*, 2023.

Xiang Yue, Huseyin A Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Huan Sun, David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe. *arXiv preprint arXiv:2210.14348*, 2022.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Tianyuan Zou, Yang Liu, Peng Li, Yufei Xiong, Jianqing Zhang, Jingjing Liu, Xiaozhou Ye, Ye Ouyang, and Ya-Qin Zhang. Contrastive private data synthesis via weighted multi-plm fusion. *arXiv preprint arXiv:2502.00245*, 2025.

Tianyuan Zou, Zinan Lin, Sivakanth Gopi, Yang Liu, Ya-Qin Zhang, Robert Sim, Xin Deng, and Sergey Yekhanin. PE-SGD: Differentially private deep learning via evolution of gradient subspace for text. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=713ywmTZHv.

Appendix

Contents

A   Private Evolution

[Alg. 1](#) presents the Private Evolution (PE) algorithm, reproduced from [Lin et al. (2023)](#). This algorithm represents the conditional version of PE, where each generated image is associated with a class label. It can be interpreted as running the unconditional version of PE separately for each class ([Line 2](#)).

Algorithm 1: Private Evolution (PE)

Input: The set of private classes: C ($C = \{0\}$ if for unconditional generation)

Private samples: $S_{\text{priv}} = \{(x_i, y_i)\}_{i=1}^{N_{\text{priv}}}$, where x_i is a sample and $y_i \in C$ is its label

Number of iterations: T

Number of generated samples: N_{syn} (assuming $N_{\text{syn}} \bmod |C| = 0$)

Noise multiplier for DP Nearest Neighbors Histogram: σ

Threshold for DP Nearest Neighbors Histogram: H

1 $S_{\text{syn}} \leftarrow \emptyset$

2 **for** $c \in C$ **do**

3 $private\_samples \leftarrow \{x_i | (x_i, y_i) \in S_{\text{priv}} \text{ and } y_i = c\}$

4 $S_0 \leftarrow$ RANDOM_API $(N_{\text{syn}} / |C|)$

5   **for** $t \leftarrow 1, \ldots, T$ **do**

6 $histogram_t \leftarrow$ DP_NN_HISTOGRAM $(private\_samples, S_{t-1}, \sigma, H)$ // See [Alg. 2](#)

7 $\mathcal{P}_t \leftarrow histogram_t / \text{sum}(histogram_t)$ // $\mathcal{P}_t$ is a distribution on S_t

8 $S'_t \leftarrow$ draw $N_{\text{syn}} / |C|$ samples with replacement from $\mathcal{P}_t$ // S'_t is a multiset

9 $S_t \leftarrow$ VARIATION_API (S'_t)

10 $S_{\text{syn}} \leftarrow S_{\text{syn}} \cup \{(x, c) | x \in S_T\}$

11 **return** S_{syn}

Algorithm 2: DP Nearest Neighbors Histogram (DP_NN_HISTOGRAM)

Input : Private samples: S_{priv}

Generated samples: $S = \{z_i\}_{i=1}^n$

Noise multiplier: σ

Threshold: H

Distance function: $d(\cdot, \cdot)$

Output: DP nearest neighbors histogram on S

1 $histogram \leftarrow [0, \ldots, 0]$

2 **for** $x_{\text{priv}} \in S_{\text{priv}}$ **do**

3 $i = \arg\min_{j \in [n]} d(x_{\text{priv}}, z_j)$

4 $histogram[i] \leftarrow histogram[i] + 1$

5 $histogram \leftarrow histogram + \mathcal{N}(0, \sigma I_n)$ // Add noise to ensure DP

6 $histogram \leftarrow \max(histogram - H, 0)$ // 'max', '-' are element-wise

7 **return** $histogram$

B   Theoretical Analysis

Convergence analysis. Since Sim-PE only changes PE's RANDOM_API and VARIATION_API, the original convergence analysis (App. E in [Lin et al. (2023)](#)) remains valid for Sim-PE. Specifically:

- The analysis assumes that RANDOM_API generates samples within a ball of diameter D covering the private samples. Changing RANDOM_API affects D but not the overall analysis procedure.

- The analysis assumes that `VARIATION_API` draws samples from a Gaussian distribution—an essential assumption for any tractable analysis involving complex, unknown foundation models and simulators. Under this assumption, modifying `VARIATION_API` does not impact the analysis procedure.

Privacy analysis. Since Sim-PE does not change the way `PE` utilizes private data, the privacy analysis of `PE` (Lin et al., 2023) applies to Sim-PE. We repeat the privacy analysis in `PE` below for completeness.

The only step that directly accesses private data is `DP_NN_HISTOGRAM` (Line 6 in Alg. 1, detailed in Alg. 2). Let us analyze its privacy cost first:

- Each private sample contributes only one vote (Line 5 in Alg. 2).
- Adding or removing a single sample alters the resulting histogram by at most 1 in the l_2 norm—so the sensitivity is 1.
- Line 5 adds i.i.d. Gaussian noise with standard deviation σ to each histogram bin. This corresponds to the standard Gaussian mechanism (Dwork et al., 2014) with noise multiplier σ (i.e., $\sigma/1$).

Now, consider the entire Sim-PE procedure (Alg. 1).

(1) Let's first consider the procedure for one class c (Line 3-Line 10). All steps other than `DP_NN_HISTOGRAM` do not access private data and simply post-process results from earlier DP steps. Therefore, Alg. 1 can be interpreted as T *adaptive compositions* of Gaussian mechanisms.

According to Corollary 3.3 in Dong et al. (2022),[4] *T compositions of Gaussian mechanisms* with noise multiplier σ can be viewed as a *single Gaussian mechanism* with effective noise multiplier $\sigma/\sqrt{T}$.

Thus, the final privacy cost reduces to computing the ε and δ of a standard Gaussian mechanism with noise $\sigma/\sqrt{T}$. We use the tight results from Balle & Wang (2018) for this computation: the mechanism is (ε, δ)-differentially private if

$$\Phi\left(\frac{\sqrt{T}}{2\sigma} - \frac{\varepsilon\sigma}{\sqrt{T}}\right) - e^\varepsilon\,\Phi\left(-\frac{\sqrt{T}}{2\sigma} - \frac{\varepsilon\sigma}{\sqrt{T}}\right) \leq \delta$$

where Φ denotes the Gaussian cumulative distribution function (CDF). Given δ and σ, we can use the above equation to solve for ε.

(2) Now, let's first consider the full algorithm involving multiple classes. Since the private samples from different classes are disjoint, due to *parallel composition property* (McSherry, 2009), the privacy cost remains the same as above. See Lin et al. (2023) for an alternative analysis that gives the same conclusion.

C   Justifying the Methodological Design of Sim-PE with Simulator-generated Data

In this section, we provide experimental evidence to support the claim in § 3.2.2: "If we already know that a sample z_i is far from the private dataset, then its nearest neighbors in S_{sim} are also likely to be far from the private dataset."

Consider Sim-PE's synthetic dataset S_t from the t-th iteration. For each private sample x_i in the private dataset S_{priv}, we find its nearest sample in S_t as z_i. $S(selected)_t = \{z_i\}_{i=1}^{N_{\text{priv}}}$ contains samples close to S_{priv} (chosen by `DP_NN_HISTOGRAM` (Alg. 2) in PE under non-DP settings). Conversely, the remaining samples, $S(unselected)_t = S_t \setminus S(selected)_t$ are farther from S_{priv}.

Now, we need to confirm that $S(unselected)_t$'s nearest neighbors in S_{sim} are farther away from the private dataset S_{priv} than $S(selected)_t$. We define:

- $z(Y)_i$ as the nearest neighbor of the private sample x_i in $S(Y)_t$, for $Y \in \{selected, unselected\}$
- $q(Y)_j^i$ as the $j-th$ nearest neighbor of $z(Y)_i$ in S_{sim}
- $FID(Y)$ as the FID between S_{priv} and $\{q(Y)_j^i\}_{i=1,j=1}^{N_{\text{priv}},\gamma}$.

---

[4]Please see their arXiv version https://arxiv.org/pdf/1905.02383.

Fig. 6 shows that $FID(selected)$ is smaller (better) than $FID(unselected)$ across most PE iterations, supporting our claim.

Figure 6: FID of the nearest neighbors of the selected and unselected samples.

D Experimental Details

In this section, we provide more experimental details.

D.1 MNIST with Text Rendering Program

The categorical parameters include: (1) Font. We use Google Fonts (Google, 2022), which offers 3589 fonts in total. (2) Text. The text consists of digits '0' - '9'. **The numerical parameters include:** (1) Font size, ranging from 10 to 29. (2) Stroke width, ranging from 0 to 2. (3) Digit rotation degree, ranging from $-30°$ to $30°$.

Tabs. 3 and 4 show their variation degrees. The total number of PE iterations is 4. Following Gong et al. (2025), we set the number of generated samples to be 60,000.

Table 3: The configurations of the categorical parameters in MNIST with Text Rendering Program experiments.

Categorical Parameter (ξ)	Feasible Set (Ξ)	Variation Degrees (β) across PE Iterations
Font	1 - 3589	0.8, 0.4, 0.2, 0.0
Text	'0' - '9'	0, 0, 0, 0

Table 4: The configurations of the numerical parameters in MNIST with Text Rendering Program experiments.

Numerical Parameter (ϕ)	Feasible Set (Φ)	Variation Degrees (α) across PE Iterations
Font size	[10, 30]	5, 4, 3, 2
Font rotation	[-30, 30]	9, 7, 5, 3
Stroke width	[0, 2]	1, 1, 0, 0

D.2 CelebA with Generated Images from Computer Graphics-based Render

The variation degrees γ across PE iterations are [1000, 500, 200, 100, 50, 20]. The total number of PE iterations is 6. Following Gong et al. (2025), we set the number of generated samples to be 60,000.

D.3 CelebA with Rule-based Avatar Generator

The full list of the categorical parameters are

- Style
- Background color
- Top
- Hat color
- Eyebrows
- Eyes
- Nose
- Mouth
- Facial hair
- Skin color
- Hair color
- Facial hair color
- Accessory
- Clothing
- Clothing color
- Shirt graphic

These are taken from the input parameters to the library (Escartín, 2021). There is no numerical parameter.

For the experiments with only the simulator, the variation degrees β across PE iterations are [0.8, 0.6, 0.4, 0.2, 0.1, 0.08, 0.06]. The total number of PE iterations is 7. Following Gong et al. (2025), we set the number of generated samples to be 60,000.

For the experiments with both foundation models and the simulator, we use a total of 5 PE iterations so as to be consistent with the setting in Gong et al. (2025). For the RANDOM_API and the first PE iteration, we use the simulator ($\beta = 0.8$). For the next 4 PE iterations, we use the same foundation model as in Lin et al. (2023) with variation degrees [96, 94, 92, 90]. Following Gong et al. (2025), we set the number of generated samples to be 60,000.

D.4 The CartPole Environment

Policy. To map observations to control signals, we adopt a simple linear policy of the form

$$\mathrm{sign}(\langle \mathrm{observations}, \boldsymbol{\pi} \rangle),$$

where $\boldsymbol{\pi}$ is a 4-dimensional policy vector and $\langle \cdot, \cdot \rangle$ denotes the inner product.

Private Dataset. To create a "private dataset," we enumerate all policies with each coordinate in $\{-5, -4.5, \ldots, 5\}$, resulting in a total of 21^4 policies. We then uniformly sample 2500 policies that succeed (label $= 1$) and 2500 policies that fail (label $= 0$).

Embedding. We use raw observations as the embedding. Since episodes may terminate at different lengths (due to success or failure), we pad all observation sequences to a fixed length of 500 frames by repeating the final observation. To keep the setup simple, we use the raw values directly without any normalization, and we find that this simple embedding already yields strong results. This also highlights that Sim-PE can still function effectively without any pretrained embeddings.

RANDOM and VARIATION APIs. We follow the same method as in § 3.2.1:

- RANDOM_API samples policy parameters uniformly at random.

- VARIATION_API perturbs existing policy parameters.

(a) Real (private) images   (b) Simulator-generated images   (c) Sim-PE images ($\epsilon = 10$)

Figure 7: The real and generated images on CelebA. The simulator is a weak rule-based avatar generator (Escartín, 2021) significantly different from the real dataset. The top rows correspond to the "female" class, and the bottom rows correspond to the "male" class. The simulator generates images with incorrect classes. Sim-PE tends to generate faces with long hair for the female class and short hair for the male class (correctly), but the generated images have mode collapse issues.

E Additional Results

E.1 Sim-PE with Weak Simulators

In Fig. 7c, we see that Sim-PE's generated images have many duplicates. In this section, we conduct more analysis on this phenomenon.

Fig. 8 shows that FID between generated images and the private dataset decreases (improves) across iterations, confirming that generated images better align with the private distribution over iterations. However, the number of unique samples in the generated dataset drops significantly (Fig. 9), leading to duplicates in Fig. 7c. This may be due to the simulator being too weak, generating most images far from the private dataset. As a result, Sim-PE converges to a small set of good images.

Figure 8: Sim-PE (using DIGIFACE-1M)'s FID on CelebA improves over the course of the PE iterations.

Figure 9: The number of unique generated samples in each iteration of Sim-PE (using DIGIFACE-1M) on CelebA.

E.2 Sim-PE with CIFAR10 as the Private Dataset and ImageNet as the Simulator-generated Dataset

In this section, we apply the algorithm in § 3.2.2 on the setting where CIFAR10 is treated as the private dataset and ImageNet is treated as the simulator-generated dataset. We conduct this experiment for the following reasons:

- In § 3.2.2, we discuss that Sim-PE can apply to any public dataset beyond simulator-generated data. This experiment provides evidence for that claim.

- The original PE approach on CIFAR10 (Lin et al., 2023) uses a foundation model pretrained on ImageNet, which aligns with the simulator data in this experiment. Comparing their performance removes the impact of distribution alignment and highlights the differences in the RANDOM_API and VARIATION_API designs of PE and Sim-PE.

Experiment settings. We use $\epsilon = 10$. The downstream task is classifying images into the 10 classes in CIFAR10, following DPImageBench (Gong et al., 2025) and the original PE paper (Lin et al., 2023).

Results. Sim-PE achieves FID=8.76, and acc=70.06%. In comparison, PE with a foundation model achieves similarly with FID=9.2 and acc=75.3%. These results suggest that Sim-PE and PE can be competitive when both suitable simulators and foundation models are available for the private dataset. In addition, the results are comparable to state-of-the-art training-based methods, demonstrating that Sim-PE is a promising framework.

E.3 Sim-PE for Generating Networking Data

In our main paper, we demonstrated that Sim-PE works for images and robotics simulations. In this section, we additionally provide a prelimiality expeirments to show that Sim-PE can also be used for generating networking data.

We follow the experimental setup of Sun et al. (2024), a recent work on DP networking data synthesis.

Data. We aim to generate differentially private *network flow header* data using the TON dataset (Sun et al., 2024) as the private dataset. Each sample contains the following fields:

- Source IP address
- Destination IP address
- Source port
- Destination port
- Protocol (e.g., UDP, TCP)
- Timestamp of the first packet
- Duration of the flow
- Number of packets
- Number of bytes
- Flow label (10 classes, e.g., normal, background, DDoS)

RANDOM and VARIATION APIs. We follow the same method as in § 3.2.1:

- RANDOM_API generates random flows.

- VARIATION_API perturbs individual fields of a flow (e.g., modifying the IP address).

Embedding. To keep the setup minimal, we embed samples directly in their normalized raw format, using one-hot encodings for categorical fields and normalized values for continuous fields. Thus, **the embedding does not rely on any machine learning model**.

Metrics. Following (Sun et al., 2024), we report:

- **1-way and 2-way Total Variation Distance (TVD)** between synthetic and private data (lower is better).
- **Classification accuracy**, obtained by training a classifier on synthetic data and testing on held-out private data (higher is better).

This classification task represents a standard downstream application for network flow data.

Privacy budget. As in (Sun et al., 2024), we set $\epsilon = 2$.

Results. Tab. 5 shows how the metrics evolve across Sim-PE iterations. We highlight two key observations:

Iteration	Accuracy	1-way TVD	2-way TVD
0	56.48	0.29	0.46
1	56.48	0.29	0.45
2	56.48	0.28	0.44
3	56.75	0.27	0.43
4	57.08	0.27	0.42
5	58.28	0.26	0.41
6	60.29	0.25	0.40
7	61.28	0.24	0.38
8	63.47	0.23	0.37
9	65.95	0.22	0.36
10	67.37	0.21	0.35
11	68.47	0.21	0.33
12	72.18	0.20	0.32
13	72.77	0.19	0.31
14	73.24	0.18	0.30

Table 5: Performance of Sim-PE on networking data across iterations.

- **Consistent improvement across iterations.** All metrics steadily improve, with final performance significantly better than the initial random samples (e.g., accuracy improves by **16.76%**). This demonstrates that Sim-PE is effective for networking data even without ML-based embeddings or sophisticated simulators.
- **Strong performance compared to prior work.** Compared to Figure 3 in Sun et al. (2024), this simple approach already outperforms two baselines in accuracy. Notably, NetShare (Yin et al., 2022), a widely used DP fine-tuning–based synthesizer in networking, is surpassed by Sim-PE in classification accuracy, highlighting its potential beyond the image domain.

F   Additional Ablation Studies

F.1   Data Selection Algorithms

In § 3.2.2, we discussed two simple alternatives for simulator data selection. The comparison is in Tab. 6. We see that Sim-PE with iterative data selection outperforms the baselines on most metrics, validating the intuition outlined in § 3.2.2. However, the clustering approach used in the second baseline still has merit, as it results in a better FID for $\epsilon = 10$. This idea is orthogonal to the design of Sim-PE and could potentially be combined for further improvement. We leave this exploration to future work.

Table 6: Accuracy (%) of classifiers trained on synthetic images and FID of synthetic images on CelebA. The best results are highlighted in bold. Sim-PE outperforms the baselines in most metrics.

Algorithm	FID ↓		Classification Acc. ↑	
	$\epsilon = 1$	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 10$
DP_NN_HISTOGRAM on S_{syn}	36.2	29.3	61.5	71.9
DP_NN_HISTOGRAM on cluster centers of S_{syn}	26.4	**18.3**	74.7	77.7
Sim-PE	**24.7**	20.8	**80.0**	**82.5**

F.2   Performance across Iterations

Fig. 10 shows that both the FID and the downstream classifier's accuracy generally improve as PE progresses. This confirms that PE's iterative data refinement process is effective when combined with simulators.

Figure 10: Sim-PE's FID and accuracy generally improve over the course of the iterations.

F.3 Sensitivity Analysis of α, β, γ

These hyperparameters (defined in § 3.2.1 and 3.2.2) control the variation degree in VARIATION_API. Higher values induce larger variations. **These are the only new hyperparameters introduced in Sim-PE.**

Following the PE paper (Lin et al., 2023), we gradually decrease them across iterations. The idea is that larger variations at the beginning help Sim-PE/PE explore and find good seeds, while smaller variations are needed later for convergence. In our experiments, most are set to follow an arithmetic or geometric sequence across iterations (App. D) based on heuristics. In the following, we modify their values: for each hyperparameter, we either fix it to the largest or smallest value in its base sequence. Figs. 11 and 12 show that:

- Setting these hyperparameters to a fixed *large* value is not ideal. For instance, setting a large text variation throughout greatly hurts FID and accuracy. Conversely, using a fixed *small* value has a smaller impact.
- The default parameters do not always yield the best result (e.g., setting γ to a fixed small value improves FID). This suggests that our main results in the paper could be improved with better hyperparameters.

Figure 11: Accuracy and FID on MNIST with different α, β schedules ($\epsilon = 10$). The "base" parameter schedule across iterations is: (a) font size's $\alpha = [5, 4, 3, 2]$, (b) rotation degree's $\alpha = [9, 7, 5, 3]$, (c) stroke width's $\alpha = [1, 1, 0, 0]$, (d) font's $\beta = [0.8, 0.4, 0.2, 0.0]$, (e) text's $\beta = [0.8, 0.4, 0.2, 0.0]$. "Small" ("large") means fixing α/β to be the smallest (largest) value across iterations. For example, the unit under "font variation" and "large" indicates the results when we set font's $\beta = [0.8, 0.8, 0.8, 0.8]$.

Figure 12: Accuracy and FID on CelebA using simulator-generated data with different γ schedules ($\epsilon = 10$). The "base" parameter schedule across iterations is $[1000, 500, 200, 100, 50, 20]$. "Small" ("large") means fixing γ to be the smallest (largest) value across iterations. For example, the unit under "large" indicates the results when we set $\gamma = [1000, 1000, 1000, 1000, 1000, 1000]$.

F.4 When to Switch from the Simulator to the Foundation Model

In § 5.3, the foundation model was introduced after iteration 1. We vary it to be either earlier (after iteration 0, using only simulators for the initial samples) or later (after iteration 2). The result is in Tab. 7. We can see that:

- Our reported main results can be improved with better parameters. For instance, starting the foundation model after iteration 0 improves accuracy to 74.3% (+1.6%) and reduces FID to 14.2 (-0.8) at $\varepsilon = 1$.
- Nevertheless, the results across different schedules are relatively close. In most cases, they outperform using either the foundation model or the simulator alone (see Tab. 2), reinforcing the conclusions in § 5.3.

Diffusion Start	FID ($\varepsilon = 1$)	Acc ($\varepsilon = 1$)	FID ($\varepsilon = 10$)	Acc ($\varepsilon = 10$)
After iter 0	14.2	74.3	13.2	73.6
After iter 1	15.0	72.7	11.9	78.1
After iter 2	16.0	73.6	14.2	76.5

Table 7: Comparison of FID and accuracy with different diffusion start iterations for $\varepsilon = 1$ and $\varepsilon = 10$.

F.5 Range of Simulator Parameters

We vary the range of all numerical parameters in the MNIST experiment $\varepsilon = 1$, including font size, rotation, and stroke width. Results across all Sim-PE iterations are presented.

- **Font size.** In our main experiments, we used the range $[10, 30)$. We evaluate two alternatives: $[15, 25)$ (narrower) and $[5, 35)$ (wider). See Tab. 8 for the results.
- **Rotation degree.** In our main experiments, we used the range $[-30, 30]$. We evaluate two alternatives: $[-15, 15]$ (narrower) and $[-60, 60]$ (wider). See Tab. 9 for the results.
- **Stroke width.** In our main experiments, we used the range $[0, 2]$. We evaluate two alternatives: $[0, 1]$ (narrower) and $[0, 3]$ (wider). See Tab. 10 for the results.

We can see from the above three experiments that:

- For font size and rotation, varying the parameter range has minimal effect on accuracy or FID across iterations. In contrast, a suboptimal stroke width range ($[0, 3]$) leads to poor initial samples (FID 109.1), but Sim-PE gradually improves quality, narrowing the final FID gap to the best setting to 4.3.
- Regardless of the specific parameter range chosen, the overall trend is consistent: Sim-PE consistently improves the synthetic data over iterations and outperforms standard PE significantly (see Table 1). These are the core messages of the paper, and they hold even when the parameter ranges are varied.

Range	Metric	Iter=0	Iter=1	Iter=2	Iter=3	Iter=4
$[15, 25)$	Accuracy	13.6	71.6	81.4	87.4	84.8
	FID	88.1	61.8	45.3	31.1	22.4
$[10, 30)$	Accuracy	16.6	76.7	86.7	89.5	89.1
	FID	85.8	57.0	42.4	28.7	20.7
$[5, 35)$	Accuracy	13.3	66.0	82.8	90.6	86.2
	FID	87.2	58.2	41.4	28.7	20.2

Table 8: Effect of font size range on accuracy and FID.

Range	Metric	Iter=0	Iter=1	Iter=2	Iter=3	Iter=4
$[-15, 15]$	Accuracy	14.5	75.7	86.0	88.3	87.7
	FID	83.1	56.7	42.3	28.6	19.7
$[-30, 30]$	Accuracy	16.6	76.7	86.7	89.5	89.1
	FID	85.8	57.0	42.4	28.7	20.7
$[-60, 60]$	Accuracy	13.3	72.3	83.0	88.5	86.3
	FID	86.6	56.0	40.9	27.5	19.4

Table 9: Effect of rotation degree range on accuracy and FID.

Range	Metric	Iter=0	Iter=1	Iter=2	Iter=3	Iter=4
[0, 1]	Accuracy	12.8	70.0	87.6	91.5	91.3
	FID	69.0	48.7	35.3	25.1	18.1
[0, 2]	Accuracy	16.6	76.7	86.7	89.5	89.1
	FID	85.8	57.0	42.4	28.7	20.7
[0, 3]	Accuracy	20.4	71.6	75.4	84.6	85.2
	FID	109.1	67.6	47.4	31.3	22.4

Table 10: Effect of stroke width range on accuracy and FID.

F.6 Number of Samples Generated from the Simulator.

We fixed the number of images to 60,000 across all experiments and iterations, following the DPImageBench evaluation protocol (Gong et al., 2025) for fair and consistent comparison across methods. Additionally, we run MNIST experiments ($\varepsilon = 1$) with 30,000 and 120,000 samples in Tab. 11.

Consistent with the claims in the original PE paper (Appendix N in Lin et al. (2023)), we see that more samples don't always improve results. On one hand, more samples increase the chance of generating examples close to the private data, potentially improving synthetic data. On the other hand, a larger sample size flattens the DP Nearest Neighbors Histogram, reducing its signal-to-noise ratio and possibly degrading quality. The overall effect is nuanced. Following Lin et al. (2023), we recommend setting the number of generated samples to be close to the number of private samples.

#Samples	FID	Accuracy
30,000	14.4	87.2
60,000	20.7	89.1
120,000	32.9	85.4

Table 11: Performance comparison with different numbers of samples.

F.7 The Impact of Distribution Alignment

On CelebA, we show that Sim-PE with a simulator that aligns well with the private data outperforms standard PE (§ 5.1.2 and Tab. 1). However, Sim-PE with a simulator less aligned with the private data performs worse than PE (§ 5.3 and Tab. 2).

To provide a more controlled experiment, we fix the simulator type and only vary the alignment between the simulator and private data.

Experiment settings. We use the same 1.2M simulator-generated images from a computer graphics renderer as in § 5.1.2 and artificially adjust their alignment to CelebA. Specifically, for each image, we compute its distance to the closest CelebA image in the Inception embedding space, sort all images by this distance, and divide them into five subsets, $D_0, ..., D_4$ (e.g., the closest 0.24M images form D_0). Fig. 13 confirms that the FID between D_i and CelebA increases with i, meaning that D_0 is the most aligned and D_4 the least. We then apply Sim-PE to each D_i independently.

Results. Fig. 14 shows that as alignment decreases, the sample quality generally drops. Specifically, Sim-PE with $D_0, ..., D_3$ yields better classification accuracy than PE, while D_4 yields worse results than PE. This confirms that Sim-PE's performance is influenced by the degree of alignment. Additionally, in all cases, Sim-PE's FID is better than the original simulator data's FID, demonstrating Sim-PE's ability to select useful samples.

Figure 13: FID between the 5 constructed subsets of DIGIFACE-1M and CelebA.

(a) FID

(b) Classification accuracy

Figure 14: Results of Sim-PE-generated data using the subsets in Fig. 13 ($\epsilon = 10$).

F.8 Class Label Information from the Simulators

For simulator 1, the target class label (i.e., the digit) is fully controlled by one parameter. For simulators 2 and 3, the target class label (i.e., the gender) is not directly controlled by any parameter, but could potentially be obtained by an external image gender classifier. One benefit of using domain-specific simulators is that we can potentially use the class label information to enhance data quality (Wood et al., 2021; Bae et al., 2023). To get a more comprehensive understanding of Sim-PE, we consider two settings: **(1) Class label information is unavailable (abbreviated as "ClassUnavail").** We artificially make the problem more challenging by assuming that the class label information is *not* available. Therefore, Sim-PE has to learn to synthesize images with the correct class by itself. Our main experiments are based on this setting. **(2) Class label information is available (abbreviated as "ClassAvail").** On MNIST, we further test how Sim-PE can be improved if the class label information is available. In this case, the RANDOM_API and VARIATION_API (Eqs. (1) and (2)) are restricted to draw parameters from the corresponding class (i.e., the digit is set to the target class).

Results: Class label information from the simulators can be helpful. The results are presented in Tab. 12 and Fig. 15. We observe that with digit information, the simulator-generated data achieves much higher classification accuracy (92.2%), although the FID remains low due to the generated digits exhibiting incorrect characteristics (Fig. 15b). The fact that Sim-PE outperforms the simulator in both FID and classification accuracy across all settings suggests that Sim-PE effectively incorporates private data information to enhance both data fidelity and utility, even when compared to such a strong baseline. As expected, Sim-PE under **ClassAvail** matches or surpasses the results obtained in **ClassUnavail** across all settings, suggesting the usefulness of leveraging class label information.

(a) Real (private) images   (b) Simulator-generated images   (c) Sim-PE images ($\epsilon = 10$)

Figure 15: The real and generated images on MNIST under the "**ClassAvail**" setting. Each row corresponds to one class. The simulator generates images that are very different from the real ones. Starting from these bad images, Sim-PE can effectively guide the generation of the simulator towards high-quality images that are more similar to real data.

Table 12: Accuracy (%) of classifiers trained on synthetic images and FID of synthetic images on MNIST under the "**ClassAvail**" setting. See Tab. 1 for results under the "**ClassUnavail**" setting for reference.

Algorithm	FID ↓		Classification Acc. ↑	
	$\epsilon = 1$	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 10$
Simulator	86.0 ($\epsilon = 0$)		92.2 ($\epsilon = 0$)	
Sim-PE	**20.7**	**8.6**	**93.9**	**95.5**

G Efficiency Evaluation

As simulators could be much cheaper to generate samples than the foundation models, we show in this section that *Sim-PE is much more efficient than PE* in our experiments.

Note that the only difference between Sim-PE and PE is the RANDOM_API and VARIATION_API. Therefore, we focus on comparing the computation time and peak CPU/GPU memory of the APIs. Tabs. 13 to 15 show that Sim-PE APIs require far less time than PE. For instance, on MNIST, each PE's API call takes over *2400 GPU seconds*, whereas Sim-PE takes less than *30 CPU seconds*–an 80x speedup, not to mention the lower cost of CPU than GPU. Consequently, each Sim-PE iteration is much faster than PE (Tab. 16). The only Sim-PE operation requiring GPU is computing the embedding and nearest neighbors of simulator-generated data (§ 3.2.2), which is a one-time cost per dataset. Even this one-time process is significantly faster than one PE API call. Sim-PE with data access requires more CPU memory to store simulator-generated data, but this can be easily optimized by loading only the needed data.

		Time	Peak CPU Memory	Peak GPU Memory
PE	RANDOM_API	3920.30 seconds (GPU)	15292.98 MB	13859.10 MB
	VARIATION_API	2422.44 seconds (GPU)	15880.25 MB	16203.89 MB
Sim-PE	RANDOM_API	27.17 seconds (CPU)	758.90 MB	0 MB
	VARIATION_API	18.51 seconds (CPU)	1083.11 MB	0 MB

Table 13: Efficiency comparison of PE and Sim-PE on the MNIST dataset. Tested on a Linux server with `AMD EPYC 7V12 64-Core Processor` and one `NVIDIA RTX A6000 GPU`.

		Time	Peak CPU Memory	Peak GPU Memory
PE	RANDOM_API	39272.33 seconds (GPU)	15293.29 MB	13859.10 MB
	VARIATION_API	31028.24 seconds (GPU)	15879.47 MB	16203.89 MB
Sim-PE	RANDOM_API	0.04 seconds (CPU)	61416.62 MB	0 MB
	VARIATION_API	0.03 seconds (CPU)	61427.29 MB	0 MB
Sim-PE Setup		1450.94 seconds (GPU)	61420.69 MB	3728.16 MB

Table 14: Efficiency comparison of PE and Sim-PE (with data access) on the CelebA dataset. Tested on a Linux server with `AMD EPYC 7V12 64-Core Processor` and one `NVIDIA RTX A6000 GPU`.

		Time	Peak Memory	Peak GPU Memory
PE	RANDOM_API	39272.33 seconds (GPU)	15293.29 MB	13859.10 MB
	VARIATION_API	31028.24 seconds (GPU)	15879.47 MB	16203.89 MB
Sim-PE	RANDOM_API	46.69 seconds (CPU)	811.72 MB	0 MB
	VARIATION_API	48.78 seconds (CPU)	1067.93 MB	0 MB

Table 15: Efficiency comparison of PE and Sim-PE (with code access) on the CelebA dataset. Tested on a Linux server with `AMD EPYC 7V12 64-Core Processor` and one `NVIDIA RTX A6000 GPU`.

	MNIST	CelebA (Sim-PE with data access)	CelebA (Sim-PE with code access)
PE	22243.32	279695.52	279695.52
Sim-PE	607.95	441.63	880.38

Table 16: The runtime (seconds) of one iteration in PE and Sim-PE. Tested on a Linux server with `AMD EPYC 7V12 64-Core Processor` and one `NVIDIA RTX A6000 GPU`.

H Extended Discussions

H.1 The Prevalence and Importance of Simulators

In this paper, we define *simulators* as any data synthesizers that do not rely on neural networks. These simulators are widely used across various applications due to their unique advantages over neural network-based approaches.

The prevalence of simulators. Despite the widespread adoption of foundation models, simulators remain highly prevalent. Here are a few notable examples:

- **Genesis:**[5] A physics-based simulator used in robotics, embodied AI, and physical AI applications. Since its release in late 2024, it has received over 24k GitHub stars and 84k downloads as of April 2024.
- **Blender:**[6] A rendering framework widely used for image and video production, including in movie-making (see examples at Blender Studio[7]). One simulator used in our face experiment (the DIGIFACE-1M dataset) is built on Blender.
- **Unreal:**[8] A widely adopted game engine with image/video rendering capability. It holds a 14.85% market share in the game development industry.[9] Note that competing engines also qualify as "simulators" under our definition.

The importance of simulators. As demonstrated above, simulators continue to play a crucial role in industry applications. Even synthetic data generation also frequently relies on simulators rather than foundation models (e.g., META-SIM (Kar et al., 2019), DIGIFACE-1M (Bae et al., 2023)). This preference stems from several unique advantages of simulators over foundation models:

- **Rich annotations**: Simulators provide additional structured labels due to explicit control over the data generation process. For instance, besides face images, FACE SYNTHETICS (Wood et al., 2021) offers pixel-level segmentation masks (e.g., identifying eyes, noses), which are highly valuable for downstream tasks such as face parsing.
- **Task-specific strengths**: Certain generation tasks remain challenging for foundation models. For example, despite recent advances, foundation models still struggle with generating images containing text (Rombach et al., 2022), whereas simulators can handle it easily. Our MNIST experiment was specifically designed to highlight this distinction.
- **Domain-specific strengths**: In domains like networking, where robust foundation models are lacking, network simulators such as ns-3[10] remain a more reliable and scalable solution.
- **Greater reliability and control**: Because simulators model the data generation process, they mitigate issues such as generating anatomically incorrect images (e.g., faces without noses[11]).

That said, there are indeed data domains for which no suitable simulators exist. However, the goal and contribution of this work is not to cover every possible data domain–which is inherently infeasible–but rather to significantly expand the range of domains where PE can be applied. For example, we demonstrate its effectiveness on challenging domains for PE such as text rendering and face generation.

Moreover, the proposed approach is not limited to using simulators. The method in § 3.2.2 can also be applied to any public dataset, as demonstrated in App. E.2. This highlights the broader potential of our approach that sidesteps the limitation of requiring a suitable simulator.

H.2 Discussion on the Chosen Metrics

In the main experiments, we use **FID** and **the accuracy of downstream classifiers** as the metrics.

---

[5]https://genesis-embodied-ai.github.io/
[6]https://github.com/blender/blender
[7]https://studio.blender.org/films/
[8]https://www.unrealengine.com/en-US
[9]https://6sense.com/tech/game-development/unreal-engine-market-share
[10]https://www.nsnam.org/
[11]https://www.reddit.com/r/StableDiffusion/comments/1e7dd62/weird_and_distorted_images_with_a1111_sd3/

Details on the accuracy of downstream classifiers. To generate synthetic data, we use the conditional version of PE (App. A), so that each generated images are associated with the class labels (i.e., '0'-'9' digits in MNIST, male vs. female in CelebA). These class labels are the targets for training the classifiers. We employ a strict train-validation-test split and account for the privacy cost of classifier hyperparameter selection. Specifically, we divide the private dataset into disjoint training and validation sets. We then run Sim-PE on the training set to generate synthetic data. Next, we train three classifiers—ResNet (He et al., 2016), WideResNet (Zagoruyko & Komodakis, 2016), and ResNeXt (Xie et al., 2017)—on the synthetic data and evaluate their accuracy on the validation set. Since the validation set is part of the private data, we use the Report Noisy Max algorithm (Dwork et al., 2014) to select the best classifier checkpoint across all epochs of all three classifiers. Finally, we report the accuracy of this classifier on the test set. This procedure ensures that the reported accuracy is not inflated due to train-test overlap or DP violations in classifier hyperparameter tuning.

Why we pick these metrics. We chose FID and classification accuracy because they are the most widely used metrics in DP image synthesis. *Most of the baselines we compared to utilize only these two metrics or just one of them (Gong et al., 2025).* The only exception is GS-WGAN, which also uses Inception Score (IS) (Salimans et al., 2016). However, IS is more relevant for natural images (e.g., ImageNet images) because it evaluates the diversity across ImageNet classes, making it unsuitable for the datasets we considered in the main paper. Additionally, for the downstream task (ten-class classification for MNIST and binary classification for CelebA), we also follow prior works (Gong et al., 2025).

The differences between these metrics. FID measures the "fidelity" of synthetic data by mapping both synthetic and private data to an embedding space, approximating each with a Gaussian distribution, and calculating the Wasserstein distance between the two distributions. Classification accuracy measures the "utility" of synthetic data by evaluating its performance when used to train a downstream classifier, simulating real-world use where users expect the synthetic data to support good classifier performance on real data. *FID and classification accuracy are complementary*: FID focuses on distribution-level closeness, while classification accuracy is sensitive to outliers or samples near the classification boundary. A high score in one does not necessarily correlate with a high score in the other. This is why we often see cases where the best methods for these metrics differ (Gong et al., 2025).

H.3 Bias in Simulators

Because Sim-PE relies on simulators, biases inherent in the simulator can be transferred to the synthetic data produced by Sim-PE.

H.4 The Ordering of Applying the Simulator and the Foundation Model

In § 3.3, we propose to use simulators in early iterations and then switch to foundation models. The reverse ordering–"foundation model → simulator"–is also feasible with appropriate modifications. In this setup, when transitioning from the foundation model to the simulator, one would need to find the optimal simulator parameters corresponding to each sample generated by the foundation model, potentially using gradient-free optimization methods.

However, we only explore "simulator → foundation model" in the paper for two main reasons:

- This ordering keeps the approach simple and avoids the need for special treatments like the one described above.
- It is also more promising in data quality. Simulators can generate diverse samples that span a wide range of variations, making them ideal for initialization. However, their image quality is bad. Foundation models, on the other hand, are better at producing high-fidelity, realistic samples, which is more desirable in the final stages of the pipeline, which creates the final synthetic data.