Synthesize Privacy-Preserving High-Resolution Images via Private Textual Intermediaries

Haoxiang Wang
Peking University
wanghaoxiang@stu.pku.edu.cn

Zinan Lin
Microsoft Research
zinanlin@microsoft.com

Da Yu
Google Research
dayuwork@google.com

Huishuai Zhang[†]
Wangxuan Institute of Computer Technology, Peking University
State Key Laboratory of General Artificial Intelligence
zhanghuishuai@pku.edu.cn

Abstract

Generating high-fidelity, differentially private (DP) synthetic images offers a promising route to share and analyze sensitive visual data without compromising individual privacy. However, existing DP image synthesis methods struggle to produce high-resolution outputs that faithfully capture the structure of the original data. In this paper, we introduce a novel method, referred to as *Synthesis via Private Textual Intermediaries* (*SPTI*), that can generate high-resolution DP images with easy adoptions. The key idea is to shift the challenge of DP image synthesis from the image domain to the text domain by leveraging state-of-the-art DP text generation methods. *SPTI* first summarizes each private image into a concise textual description using image-to-text models, then applies a modified Private Evolution algorithm to generate DP text, and finally reconstructs images using text-to-image models. Notably, *SPTI* requires no model training, only inferences with off-the-shelf models. Given a private dataset, *SPTI* produces synthetic images of substantially higher quality than prior DP approaches. On the LSUN Bedroom dataset, *SPTI* attains an FID = 26.71 under $\epsilon = 1.0$, improving over Private Evolution's FID of 40.36. Similarly, on MM-CelebA-HQ, *SPTI* achieves an FID = 33.27 at $\epsilon = 1.0$, compared to 57.01 from DP fine-tuning baselines. Overall, our results demonstrate that *Synthesis via Private Textual Intermediaries* provides a resource-efficient and proprietary-model-compatible framework for generating high-resolution DP synthetic images, greatly expanding access to private visual datasets. Our code release: https://github.com/MarkGodrick/SPTI

1 Introduction

The past few years have seen an explosion in the capabilities of state-of-the-art generative models. Large diffusion and autoregressive models can produce photorealistic images, coherent long-form text, and convincing multi-modal outputs with little more than a prompt [36, 17, 30, 43, 29, 56, 48, 47]. However, when these powerful public models are applied to domain- or user-specific tasks, some form of adaptation is often required, whether it be expensive fine-tuning on private data or in-context learning with user examples. At the same time, the use of private or sensitive data raises serious privacy concerns: direct fine-tuning of a model on proprietary images or user uploads risks memorization and potential leakage of personal information [7, 34, 49, 55].


[†]Corresponding author.
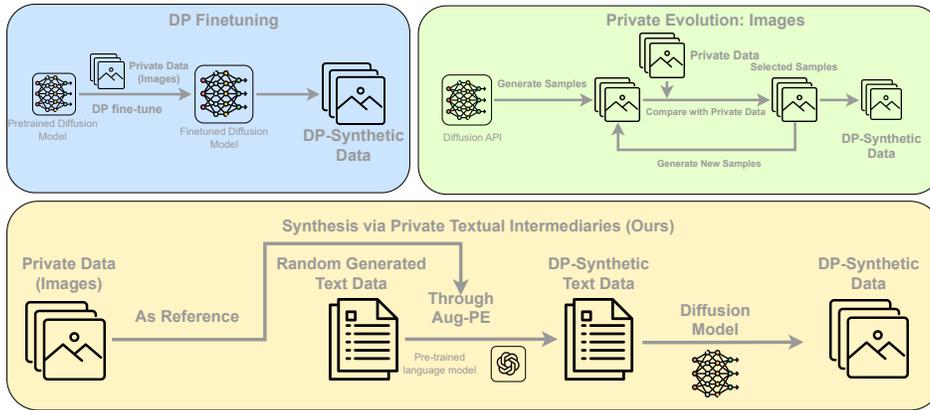
Figure 1: Overview of the *Synthesis via Private Textual Intermediaries* (*SPTI*) framework for differentially private (DP) synthetic data generation. **Top left**: *DP fine-tuning framework.* A pretrained model is fine-tuned on private data under DP constraints, and the resulting model is used to generate DP synthetic data. **Top right**: *Private Evolution (PE) on images.* This method begins by randomly generating candidate samples, which are then compared to private data. Samples are selected based on a voting mechanism and further perturbed to produce a new generation of samples. **Bottom**: Synthesis via Private Textual Intermediaries *(*SPTI*) framework.* Private image data is served as reference. A modified Augmented Private Evolution (Aug-PE) method [53] is then applied to generate DP synthetic text data, which is subsequently transformed into DP synthetic images using a diffusion model API.

Figure 2: DP-synthetic images from dataset **European Art** ($\epsilon = 1.0$).
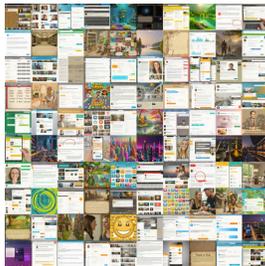
Figure 3: DP-synthetic images from dataset **Wave-ui-25k** ($\epsilon = 1.0$).

Figure 4: DP-synthetic images from dataset **LSUN Bedroom** ($\epsilon = 1.0$).

Broadly speaking, there are two strategies to harness private data under formal differential privacy guarantees [10]. The more direct approach is *differentially private (DP) fine-tuning* [57, 25, 6, 23, 35, 31, 8], in which gradient updates are clipped and noised (e.g., via DP-SGD [1]). While conceptually straightforward, this method demands substantial expertise and computational resources. For instance, fine-tuning a state-of-the-art model such as DeepSeek V3 [29] requires hundreds of GPUs, not to mention the engineering complexity of per-sample gradient clipping and noise calibration of DP-SGD [15]. Moreover, many proprietary foundation models do not even permit user fine-tuning, effectively precluding this route.

An alternative is to generate a *DP synthetic dataset* that mimics the distribution of the private data [27, 53, 28, 12] and then use that synthetic data for downstream adaptation—either by fine-tuning or by providing examples in-context. This "generate-then-adapt" paradigm sidesteps the need to directly privatize a large model but poses its own challenges: how can one synthesize high-fidelity images (or other modalities) that both faithfully capture the structure of the original data and satisfy strong DP guarantees, all without incurring substantial compute overhead?

In this work, we introduce a *Synthesis via Private Textual Intermediaries* framework that leverages the remarkable text understanding of modern models as a bridge between private data and high-quality generation. Our key insight is twofold:

1. **Text generation is one of the most successful modalities for AI, so is the private text generation.** Private text generation can achieve similar performance to directly differentially private fine-tuning LLMs [53, 59], and text models can be privately adapted far more easily than their image counterparts.

2. **Text unifies modalities.** In modern models, natural language often serves as a universal interface, with a wide range of off-the-shelf models available for both generating images, video, audio, and other modalities from text [43, 40, 30], and for generating text from these modalities [47, 18].

Building on these observations, we propose *Synthesis via Private Textual Intermediaries* (*SPTI*) (Figure 1). Given a private image collection, *SPTI* first converts the images into concise textual summaries using a standard captioning model. It then adapts *Private Evolution* [27, 53, 28, 50], a DP synthetic data algorithm that only requires model inference, to generate DP-sanitized captions that reflect the private caption distribution. Finally, it uses an off-the-shelf text-to-image models to sample candidate images conditioned on these DP captions. Notably, all steps in *SPTI* rely solely on the inference APIs of existing models without any additional training, allowing *SPTI* to leverage state-of-the-art models behind proprietary APIs—something that DP fine-tuning approaches cannot achieve.

We evaluate *SPTI* on standard benchmarks. On LSUN Bedroom, *SPTI* achieves FID = 26.71, markedly improving over Private Evolution's prior best of 40.36 for $\epsilon = 1$. On MM-CelebA-HQ, it attains FID = 33.27 versus 57.01 for DP fine-tuning baselines for $\epsilon = 1$. In both cases, *SPTI* generates high-resolution synthetic images with significantly better quality than either cost-intensive DP-SGD approaches [43] or previous API-based approaches [27]. Figures 2, 3, 4 show generated images from *SPTI*.

Our contributions are as follows:

- Conceptually, we introduce *Synthesis via Private Textual Intermediaries*, a novel framework that uses text as an intermediate representation to bridge off-the-shelf multimodal LLMs and achieve privacy-preserving, high-resolution image generation by leveraging state-of-the-art DP text generation methods.

- Technically, we develop a novel Private Evolution voting mechanism that privately select text descriptions by voting in image representation space. It achieves strong privacy while ensuring strong resemblance to the private image data.

- Empirically, we demonstrate that *SPTI* generates high-quality, high-resolution images across diverse datasets, and can readily incorporate future advances in multimodal generation. Extensive ablation studies validate the effect of each components of the algorithm in achieving state-of-the-art performance.

The rest of the paper is organized as follows. In Section 2 we review background and preliminaries; Section 3 details the *SPTI* method; Section 4 presents experimental results and analyses; Section 5 discusses related work; Section 6 concludes with a discussion of future directions; and Section 7 faithfully describes the potential limitations of the proposed method.

2   Preliminaries: Differential Privacy and DP Synthetic Data via APIs

Differential Privacy.   One mechanism M is said to be (ϵ, δ)-DP if for any two neighboring datasets D and D' which differ in a single entry and for any observation set S of outputs of M, one has

$$\Pr\big(M(D) \in S\big) \ \leq \ e^\epsilon \Pr\big(M(D') \in S\big) \ + \ \delta.$$

The intuition of (ϵ, δ)-DP requires that any single sample cannot influence the mechanism's output too much.

DP Synthetic Data via APIs.   The DP synthetic data problem is formulated as DP Wasserstein Approximation in Lin et al. [27]. Given a private dataset $\mathcal{D}_{priv} = \{x_i : i \in [N_{priv}]\}$ with N_{priv} samples (e.g., images), the goal is to design an (ϵ, δ)-DP algorithm M that outputs a synthetic dataset

$\mathcal{D}_{syn} = \{x'_i : i \in [N_{syn}]\}$ with N_{syn} samples, where the Wasserstein p-distance $W_p(\mathcal{D}_{priv}, \mathcal{D}_{syn})$ is minimized, w.r.t. the distance function $d(\cdot, \cdot)$ and some $p \geq 1$.

Private Evolution (PE). [27] is a recently proposed differentially private synthetic data generation that relies solely on API access to off-the-shelf models. The PE algorithm consists of following steps:

- **Random Initialization.** Call RANDOM_API to draw an initial set of synthetic candidates from the foundation model.

- **Repeat the following routine several times**

  1. **Private Voting.** Using the private dataset, each real sample "votes" for its nearest synthetic candidate under the embedding network, producing a (noisy) histogram via the DP voting function.

  2. **Resampling.** Draw a new batch of synthetic points by sampling from this private histogram.

  3. **Variation.** For each drawn point, invoke VARIATION_API to produce novel samples that are semantically similar to the selected candidate (e.g., variations of an object in an image).

This procedure can generate synthetic data whose distribution closely matches that of the private dataset while controlling privacy leakage using a private histogram algorithm.

3 Method: High Quality Image Synthesis via Private Textual Intermediaries

Before describing our method, we first outline the high-level motivation. The goal of differentially private (DP) synthetic data generation is to mimic the distribution of private data, no matter how rare or complex the private data are, e.g., high-resolution or UI-like images, while preserving privacy. However, directly synthesizing such images by using Private Evolution [27] poses two key challenges. First, rare cases are difficult to generate if they are far from typical data distribution that the model has learned on. Second, high-resolution images require the model to correctly render all fine-grained details across a large output space, which is particularly challenging for generation models.

To address these issues, we shift the problem from the image domain to the text domain by introducing textual intermediaries. Text representations are semantically rich yet lower-dimensional, making them significantly more amenable to efficient DP synthesis. Recent advances have shown that privately trained language models [59] or DP synthetic text [53] can achieve strong utility while satisfying DP. Moreover, text serves as a universal interface for multimodal generation: a single caption can condition powerful off-the-shelf models to generate images, videos, audio, and more. This makes our approach have the potential to be broadly applicable across modalities, though we focus on image synthesis in this work.

We propose *Synthesis via Private Textual Intermediaries* (*SPTI*), a novel framework that leverages this insight. *SPTI* first converts private images into captions, applies our newly-designed Private Evolution (PE) algorithm to these descriptions under a DP guarantee, and then synthesizes images using a text-to-image generator. This design ensures strong privacy while maintaining high utility. An overview is shown in Figure 1.

3.1 The Design of *Synthesis via Private Textual Intermediaries*

Our *SPTI* consists of three stages: (1) using image-to-text model to caption private images to obtain text descriptions, (2) applying Private Evolution to the text data under DP constraints, and (3) using text-to-image model to generate synthetic images using the evolved text. The process is outlined in Algorithm 1.

Images to Texts. Given a private image dataset $\mathcal{D}$, we begin by captioning each image into a textual description using an off-the-shelf image captioning model. This results in a corpus of private text data $\mathcal{T}$, which serves as a proxy for the original image data during subsequent privacy-preserving operations.

Private Evolution in Text Space. Augmented Private Evolution (Aug-PE) [53] is an iterative, population-based method for generating synthetic text data under DP constraints. We modify it by replacing its voting mechanism with a new voting strategy. At each iteration, a pool of candidate texts is generated, evaluated, and selected. For selection, we assign each candidate a probability using Image Voting process, which is described with detail in Section 3.2. We then select the candidates by this probability to produce the next generation. This process preserves privacy while exploring the text space efficiently. We provide a detailed version of this modified algorithm 2 in Appendix A.

Evolved Texts to Images. Once we obtain the evolved synthetic text set $\mathcal{T}'$, we use a high-quality text-to-image diffusion model (e.g., state-of-the-art open-source models [44] or commercial APIs) to convert each text back into an image. Since the evolution occurred entirely in the text domain, the resulting images are inherently DP-compliant due to post processing properties of DP.

RANDOM_API and VARIATION_API. Our method follows **Aug-PE** [53] method, also applying `RANDOM_API` and `VARIATION_API` in our algorithm `Aug_PE_Image_Voting`. For `RANDOM_API`, we prompt the LLMs to generate random image captions from scratch, denoted as `RANDOM_API`(N). This means `RANDOM_API` will generate N samples. For `VARIATION_API`, we also prompt the LLMs to generate variations of the given image captions, denoted as `VARIATION_API`$(\mathcal{C})$. This means `VARIATION_API` will generate variants of equal quantity from text data $\mathcal{C}$.

Algorithm 1 *SPTI*: Privately Synthesize High-Resolution Images via *Synthesis via Private Textual Intermediaries*

Require: Private image dataset $\mathcal{D}$, `Aug_PE_Image_Voting`
Ensure: Synthetic images $\mathcal{D}'$
 1: Convert $\mathcal{D}$ to text descriptions $\mathcal{T}$ using a captioning model
 2: Apply Private Evolution: $\mathcal{T}' = $ `Aug_PE_Image_Voting`$(\mathcal{T})$
 3: Generate images $\mathcal{D}'$ from $\mathcal{T}'$ using a text-to-image diffusion model
 4: **return** $\mathcal{D}'$

3.2 Private Evolution in Textual Space with Voting by Image Representations

Running Aug-PE on image captions can generate synthetic captions that are similar to those of the original private images in the text domain. However, the final text-to-image models work in a random and complex way. As a result, even accurate captions in the text domain may not necessarily produce images that closely match the original private ones in the image domain.

Based on this insight, we propose **Image Voting** (Algorithm 2), a cross-modal selection mechanism integrated into the PE process. Rather than selecting candidate texts solely based on textual similarity, we first generate images from the candidate texts using the text-to-image model. These synthesized images are then embedded using an image encoder (e.g., Inception-v3 [46, 39]), along with the private images.

For each private image embedding, we identify its nearest neighbor among the generated image embeddings (e.g., via Euclidean distance). The corresponding synthetic text associated with the nearest image is assigned one vote. After accumulating votes across all private images and adding noise, we normalize the vote counts into a probability distribution over candidate texts, which is then used to sample the next generation. This reverse-nearest-neighbor approach prioritizes synthetic texts that produce images broadly aligned with the private dataset. Additionally, the **Image Voting** process notably **requires no textual reference**, as it directly leverages the private images themselves as guidance during generation.

3.3 Privacy Analysis

We analyze the privacy guarantees of our proposed *SPTI* framework, which synthesizes image data under differential privacy (DP) by operating in the text space. Our goal is to ensure that the final synthetic images $\mathcal{D}'$ are generated through a process that satisfies (ε, δ)-DP with respect to the private dataset $\mathcal{D}$.

Privacy-Critical Step. The only step in the *SPTI* pipeline that accesses the private data is the Private Evolution module, specifically through the similarity-based voting mechanism in algorithm `Aug_PE_Image_Voting` (Algorithm 2). This module operates on embeddings of the private image dataset $\mathcal{D}$ and is used to guide the sampling of synthetic text candidates. Thus, our privacy analysis focuses on this step.

Voting Mechanism. For each private image embedding, we identify its nearest neighbor among the generated image embeddings , and increment a vote count histogram H indexed by the candidate text that produced the nearest image. This step corresponds to a histogram query over the private data.

To ensure differential privacy, we add Gaussian noise $\mathcal{N}(0, \sigma^2\mathbf{I})$ to the vote histogram H , which is equivalent to applying the *Gaussian mechanism* to a function with bounded sensitivity.

Sensitivity Analysis. The histogram query has L_2 sensitivity at most 1, since each private image contributes a vote to only one candidate (i.e., changing a single image can affect the histogram by at most 1 in one coordinate). This satisfies the precondition for the Gaussian mechanism.

According to the Gaussian mechanism [4], adding noise of variance σ^2 per coordinate to a function with L_2 sensitivity 1 ensures (ε, δ)-DP, provided:

$$\Phi\left(\frac{1}{2\sigma} - \varepsilon\sigma\right) - e^\varepsilon \Phi\left(-\frac{1}{2\sigma} - \varepsilon\sigma\right) \le \delta.$$

Privacy Composition. According to the adaptive composition theorem of Gaussian mechanisms [9], applying the above Gaussian mechanism across G iterations satisfy (ε, δ)-DP, provided:

$$\Phi\left(\frac{\sqrt{G}}{2\sigma} - \frac{\varepsilon\sigma}{\sqrt{G}}\right) - e^\varepsilon \Phi\left(-\frac{\sqrt{G}}{2\sigma} - \frac{\varepsilon\sigma}{\sqrt{G}}\right) \le \delta. \tag{1}$$

Post-processing Immunity. All downstream steps in *SPTI*—including text mutation and image generation via a fixed diffusion model—depend solely on the privatized output of the voting mechanism. By the *post-processing property* of DP, these steps incur no additional privacy cost.

Overall Guarantee. Hence, *SPTI* satisfies (ε, δ)-differential privacy with respect to the private image dataset $\mathcal{D}$, provided that the noise scale σ is properly chosen to satisfy Eq. 1.

4 Experiments

In this section, we evaluate the effectiveness of our proposed method through comprehensive experiments. Section 4.1 describes our experimental setup. Section 4.2 and 4.3 present the main results, where we compare our approach against state-of-the-art methods across multiple benchmarks to demonstrate its superior performance. Section 4.4 includes ablation studies analyzing the impact of the image voting mechanism and different model APIs. Our findings show that the image voting technique consistently improves performance, while the quality of the synthetic data remains robust across different language and diffusion model APIs. We also represent extended ablation study in Appendix C.

4.1 Experiment Setup

Datasets. To mitigate concerns that existing benchmarks may have been included in the pretraining data of diffusion models, we carefully select only essential benchmarks with known publication dates, prioritizing those that were introduced recently. Additionally, we create a new dataset by extracting data from the open-source Blender movie **Sprite Fright**. We evaluate our *SPTI* method on six datasets: **LSUN Bedroom** [58], **Cat** [27], **European Art** [41], **Wave-ui-25k** [61], **MM-CelebA-HQ** [51, 52, 33, 21, 24], and our newly constructed **Sprite Fright**. These datasets are selected because each represents a realistic, privacy-sensitive scenario. For instance, **LSUN Bedroom** captures aspects of an individual's private living space, while **Wave-ui-25k** consists of screenshots that may reveal personal information from electronic devices.

Baselines. We compare our method with two state-of-the-art DP baselines:

- **Private Evolution** [26]: A recently proposed DP image synthesis method that operates in the image embedding space via private evolutionary strategies.
- **DP Fine-tuning** [32]: This method fine-tunes a diffusion model on private data under DP constraints using DP-SGD, allowing direct sample generation while preserving privacy.

Models. Our pipeline requires a captioning model, a large language model (LLM), and a diffusion model. For the caption model, we use **GPT-4o-mini** [38] and **Qwen-VL-Max** [3]. Our experiments show these two models produce similar performances, and detailed results are presented in Section 4.4. For the LLM and diffusion model, we use **Meta-Llama-3-8B-Instruct** [2] for text generation and **SDXL-Turbo** [44] for image synthesis. Additional comparisons with newer and more advanced models are discussed in our ablation study in Section 4.4.

Evaluation Metrics. To evaluate the quality of our differentially private synthetic samples, we use the **Fréchet Inception Distance** [16] (FID) between the original private dataset and the synthetic one. Lower FID scores indicate better alignment between the two distributions in terms of visual features.

4.2 Evaluation

FID score. We compare *SPTI* with two state-of-the-art approaches: **Private Evolution** and **DP fine-tuning**, across multiple datasets. The results are presented in Table 1 and Table 2. For a fair comparison, we use a Latent Diffusion Model (LDM) [42, 5] pretrained on LAION-400M [45] as the backbone for all methods. The experimental results demonstrate the superior performance of *SPTI*. For more details on the experimental settings, please refer to Appendix E.

Voted samples and variants. In each iteration, our results generate two categories of samples. One can be categorized as **voted** samples selected from previous generation, while the others can be viewed as variants generated from these samples, which we call **variants**. During experiment evaluation, we often find voted samples perform better on **FID** evaluation, which can be seen in Figures 5, 6. For consistency, we present all our experiment results using FID calculated from voted samples. For more figures on experiment results, please refer to Appendix B.1.
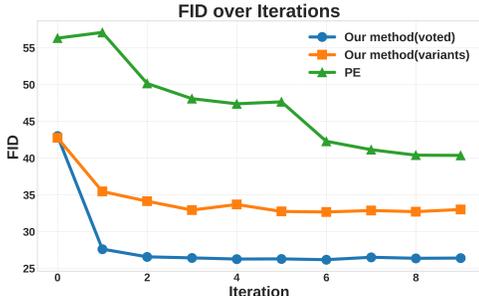

Figure 5: Experiment results on **LSUN Bedroom** dataset
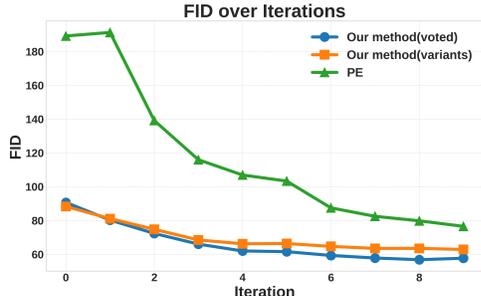
Figure 6: Experiment results on **European Art** dataset

4.3 Downstream Task

To comprehensively evaluate the generalization and transferability of the representations learned by our proposed *SPTI*, we conduct a binary classification task on the CelebA dataset. We select the attribute `Wearing_Lipstick` as the target label, and approximately 53% of the samples are without lipstick and 47% are with lipstick. We generate 2,000, 4,000, 6,000, 8,000, and 10,000 samples using *SPTI*, and PE, respectively, and compare their accuracy to the model trained by the corresponding number of ground-truth training samples. A **WRN-40-4** model is then trained on each generated data samples under identical training settings, and the test accuracy is evaluated using the ground-truth labels.

The test accuracy in Figure 7 shows that *SPTI* is more effective than PE method. This could be due to the higher diversity of data generated by *SPTI* compared to that of PE (see Figures 36, 37). We also

		$\varepsilon = 10$	$\varepsilon = 5$	$\varepsilon = 1$
LSUN Bedroom	*SPTI* (ours)	**25.88**	**25.87**	**26.39**
	PE image	41.72	41.08	40.36
	DP-finetune	31.28	31.34	31.76
Cat	*SPTI* (ours)	101.57	102.31	106.15
	PE image	**51.00**	**51.29**	**64.62**
	DP-finetune	148.05	148.46	148.75
European Art	*SPTI* (ours)	**41.42**	**42.71**	**57.64**
	PE image	76.25	74.41	76.50
	DP-finetune	61.10	61.82	63.97
Wave-ui-25k	*SPTI* (ours)	**20.16**	**22.53**	**35.18**
	PE image	39.28	48.95	50.45
	DP-finetune	49.84	52.09	62.08
Sprite Fright	*SPTI* (ours)	**142.48**	**141.49**	**157.31**
	PE image	195.58	181.77	197.13
	DP-finetune	148.19	151.94	161.25

Table 1: **FID** values (lower is better) across multiple datasets to compare three different DP methods: *SPTI*, PE Image, and DP-finetune.

Method	model	$\varepsilon = 10$	$\varepsilon = 1$
SPTI (our method)	sdxl-turbo	26.02	34.17
	LDM pretrained model	**16.59**	**33.41**
PE	sdxl-turbo	53.32	50.45
DP-finetune	LDM pretrained model	50.13	57.01

Table 2: **FID** values (lower is better) on the MM-Celeba-HQ dataset to compare different DP generation methods: *SPTI*, PE Image, and DP-finetune.

give results from models trained with ground-truth training samples to show that our selected number of samples is sufficient to train a good binary classifier. Additionally, we provide model performance during training and examples of generated samples in Appendix B.3.

4.4 Ablation Study

We conduct a series of ablation studies to investigate the contribution of each component of our proposed method. Specifically, we analyze the impact of **image voting** in the main text, while additional studies on hyperparameters for the LLM and diffusion APIs are provided in appendix due to space constraints. All experiments are conducted under the same conditions described in Section 4.1.

4.4.1 SPTI Without Image Voting

To evaluate the effect of **image voting**, we substitute it with a baseline approach and present the comparative results in Table 3. Specifically, we compare our image voting strategy with the original Aug-PE method, where synthetic text is voted using generated private text data rather than the original private image data. We observe a significant improvement in FID, suggesting that image-based voting
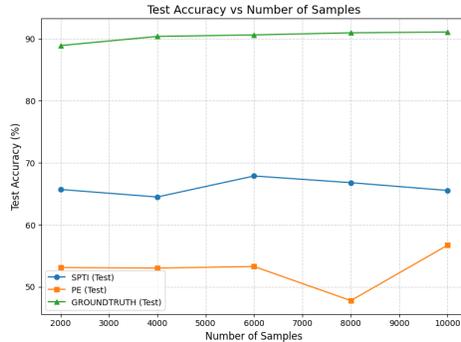
Figure 7: Test Accuracy on **CelebA** dataset with different training samples. *SPTI* achieves better accuracy under the same configuration because SPTI generates samples with more diversity than PE method. The experiment results demonstrate that our method produces data with sufficient fidelity, enabling the model to effectively learn from the generated samples.

plays an important role in enhancing the quality of the generated data. For more figures on experiment results, please refer to Appendix B.2.

		$\varepsilon = 10$	$\varepsilon = 5$	$\varepsilon = 1$
LSUN bedroom	**Image Voting**	**27.92**	**29.50**	**38.00**
	Text Voting	39.82	37.83	41.17
Cat	**Image Voting**	**101.57**	**102.31**	**106.15**
	Text Voting	110.51	106.07	106.02
Europeart	**Image Voting**	**46.69**	**48.45**	**64.04**
	Text Voting	74.37	74.04	74.91
Wave-ui-25k	**Image Voting**	**34.39**	**42.02**	**70.87**
	Text Voting	92.53	92.11	95.77

Table 3: **FID** values (lower is better) across multiple datasets to compare the *SPTI* method with Image Voting and that with Text Voting.

5   Related Work

Differentially private (DP) synthetic data generation has been studied extensively. Early work focused on query-based and statistical approaches, such as the Multiplicative Weights Exponential Mechanism (MWEM) algorithm, which iteratively measures selected queries under DP and synthesizes data to match those noisy answers [14]. PrivBayes learns a Bayesian network with DP noise added to its parameters, then samples synthetic records from the privatized network [60]. While these methods provide provable guarantees, they often struggle with high-dimensional data due to the curse of dimensionality.

The advent of deep generative models enabled DP-GANs and VAEs. Xie et al. [54] apply DP-SGD to GAN training to bound privacy leakage, and PATE-GAN uses the Private Aggregation of Teacher Ensembles framework to train a generator with DP guarantees [20]. Although these models can capture complex distributions, they incur utility loss from noise and training instability.

Recently, diffusion (score-based) models have shown promise under DP. Ghalebikesabi et al. [11] privately fine-tune a pre-trained diffusion model on sensitive images, achieving state-of-the-art FID scores on CIFAR-10 and medical imaging benchmarks. This demonstrates that modern diffusion architectures can yield high-quality private synthetic data. Liu et al. [31] proposed DP-LDM bench-

mark to train latent diffusion models for DP image generation. This provided an outstanding baseline to compare with our method.

A parallel line of work investigates inference-only, API-based DP synthesis. Lin et al. [27] introduce Private Evolution, which injects noise during generation from a frozen model, matching or surpassing retrained approaches on image benchmarks. Xie et al. [53] extend this idea to text, querying a large language model's API with augmented DP sampling to produce synthetic text with strong utility. Lin et al. [28] further extend Private Evolution by incorporating non-neural-network data synthesizers, such as computer graphics tools. This approach delivers better results in domains where suitable pre-trained models are unavailable and unlocks the potential of powerful non-neural-network data synthesizers for DP data synthesis. Google has also demonstrated an inference-only LLM mechanism for DP data generation [13].

A concurrent study also explores generating DP synthetic images by first producing text captions of private images [22]. Their hierarchical DP fine-tuning approach trains a model to generate privatized album descriptions and then photo descriptions conditioned on those. In contrast, we adopt the PE method [53], which requires only API access to foundation models, and further specialize it by using embeddings of the generated images, rather than text captions, during the voting process.

6   Conclusion

We have presented *Synthesis via Private Textual Intermediaries* (*SPTI*), a novel framework for differentially private (DP) image synthesis by leveraging existing powerful image-to-text and text-to-image generation models. By first converting each private image into a concise textual summary via off-the-shelf image-to-text models, then applying a modified Private Evolution algorithm to guarantee (ϵ, δ)-DP on the text, and finally reconstructing high-resolution images with state-of-the-art text-to-image systems, *SPTI* sidesteps the challenges of training DP image generators. Importantly, our approach requires no additional model training, making it both resource-efficient and compatible with proprietary, API-access-only models. Empirical results demonstrate significant improvement over existing methods. These findings validate that shifting the DP burden to the text domain enables the generation of high-fidelity high-resolution synthetic images under strict privacy guarantees, offering a practical path toward sharing and analyzing sensitive visual datasets.

7   Limitations

Domain Generalization   *Synthesis via Private Textual Intermediaries*' performance is intrinsically tied to the coverage and robustness of the underlying image-to-text and text-to-image models. Although multimodal large language models are continually improving, they still underrepresent certain data domains (e.g., specialized medical imagery, remote-sensing data, niche artistic styles). In such underrepresented or out-of-distribution settings, *SPTI* may fail to capture critical visual features or produce coherent synthesis. Moreover, biases and domain shifts in the pre-trained models can further degrade image fidelity and diversity.

Computational Overhead   Our full data generation pipeline involves two major components: an 8B-parameter large language model (LLM) and a 3.5B-parameter diffusion model. Under our current implementation, a complete data generation run takes approximately 19 hours on **a single NVIDIA A800**, or 15.5 hours on **2×NVIDIA A100 GPUs**, with a peak memory requirement of about 70 GB.

In comparison, the baseline PE method [27]uses only a single 3.5B model, requires no LLM, and completes in about 7 hours on **a single NVIDIA A100 GPU**, with a peak memory requirement of about 35 GB. While our method has higher compute demands, it introduces a semantically enriched generation process through language-guided synthesis, which we show leads to superior data diversity and downstream performance.

We believe the additional cost is justified by the significant gains in quality and generalization, and we note that our pipeline can be modularly optimized (e.g., via model distillation or prompt caching) in future work to reduce the runtime.

As for API usage, caption process will take 250M tokens for 10,000 images. During data generation, LLM will process 7.3M tokens during each iteration, and diffusion model will process 5M tokens in each iteration. A total of 10 iterations will count for 123M tokens.

Acknowledgement

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

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

[5] Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. Retrieval-augmented diffusion models, 2022. URL https://arxiv.org/abs/2204.11824.

[6] Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Differentially private bias-term only fine-tuning of foundation models. *arXiv preprint arXiv:2210.00036*, 2022.

[7] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. *USENIX Security Symposium*, 2021.

[8] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *Transactions on Machine Learning Research*.

[9] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

[10] 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.

[11] 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, 2023.

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

[13] Google Research. An inference-only approach to differentially private llm data generation. Technical report, Google, 2024.

[14] Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[15] 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.

[16] 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. *arXiv e-prints*, art. arXiv:1706.08500, June 2017. doi: 10.48550/arXiv.1706.08500.

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[19] Imagen-Team-Google, :, Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, Hongliang Fei, Nando de Freitas, Yilin Gao, Evgeny Gladchenko, Sergio Gómez Colmenarejo, Mandy Guo, Alex Haig, Will Hawkins, Hexiang Hu, Huilian Huang, Tobenna Peter Igwe, Christos Kaplanis, Siavash Khodadadeh, Yelin Kim, Ksenia Konyushkova, Karol Langner, Eric Lau, Rory Lawton, Shixin Luo, Soňa Mokrá, Henna Nandwani, Yasumasa Onoe, Aäron van den Oord, Zarana Parekh, Jordi Pont-Tuset, Hang Qi, Rui Qian, Deepak Ramachandran, Poorva Rane, Abdullah Rashwan, Ali Razavi, Robert Riachi, Hansa Srinivasan, Srivatsan Srinivasan, Robin Strudel, Benigno Uria, Oliver Wang, Su Wang, Austin Waters, Chris Wolff, Auriel Wright, Zhisheng Xiao, Hao Xiong, Keyang Xu, Marc van Zee, Junlin Zhang, Katie Zhang, Wenlei Zhou, Konrad Zolna, Ola Aboubakar, Canfer Akbulut, Oscar Akerlund, Isabela Albuquerque, Nina Anderson, Marco Andreetto, Lora Aroyo, Ben Bariach, David Barker, Sherry Ben, Dana Berman, Courtney Biles, Irina Blok, Pankil Botadra, Jenny Brennan, Karla Brown, John Buckley, Rudy Bunel, Elie Bursztein, Christina Butterfield, Ben Caine, Viral Carpenter, Norman Casagrande, Ming-Wei Chang, Solomon Chang, Shamik Chaudhuri, Tony Chen, John Choi, Dmitry Churbanau, Nathan Clement, Matan Cohen, Forrester Cole, Mikhail Dektiarev, Vincent Du, Praneet Dutta, Tom Eccles, Ndidi Elue, Ashley Feden, Shlomi Fruchter, Frankie Garcia, Roopal Garg, Weina Ge, Ahmed Ghazy, Bryant Gipson, Andrew Goodman, Dawid Górny, Sven Gowal, Khyatti Gupta, Yoni Halpern, Yena Han, Susan Hao, Jamie Hayes, Jonathan Heek, Amir Hertz, Ed Hirst, Emiel Hoogeboom, Tingbo Hou, Heidi Howard, Mohamed Ibrahim, Dirichi Ike-Njoku, Joana Iljazi, Vlad Ionescu, William Isaac, Reena Jana, Gemma Jennings, Donovon Jenson, Xuhui Jia, Kerry Jones, Xiaoen Ju, Ivana Kajic, Christos Kaplanis, Burcu Karagol Ayan, Jacob Kelly, Suraj Kothawade, Christina Kouridi, Ira Ktena, Jolanda Kumakaw, Dana Kurniawan, Dmitry Lagun, Lily Lavitas, Jason Lee, Tao Li, Marco Liang, Maggie Li-Calis, Yuchi Liu, Javier Lopez Alberca, Matthieu Kim Lorrain, Peggy Lu, Kristian Lum, Yukun Ma, Chase Malik, John Mellor, Thomas Mensink, Inbar Mosseri, Tom Murray, Aida Nematzadeh, Paul Nicholas, Signe Nørly, João Gabriel Oliveira, Guillermo Ortiz-Jimenez, Michela Paganini, Tom Le Paine, Roni Paiss, Alicia Parrish, Anne Peckham, Vikas Peswani, Igor Petrovski, Tobias Pfaff, Alex Pirozhenko, Ryan Poplin, Utsav Prabhu, Yuan Qi, Matthew Rahtz, Cyrus Rashtchian, Charvi Rastogi, Amit Raul, Ali Razavi, Sylvestre-Alvise Rebuffi, Susanna Ricco, Felix Riedel, Dirk Robinson, Pankaj Rohatgi, Bill Rosgen, Sarah Rumbley, Moonkyung Ryu, Anthony Salgado, Tim Salimans, Sahil Singla, Florian Schroff, Candice Schumann, Tanmay Shah, Eleni Shaw, Gregory Shaw, Brendan Shillingford, Kaushik Shivakumar, Dennis Shtatnov, Zach Singer, and Evgeny Sluzhaev. Imagen 3. *arXiv e-prints*, art. arXiv:2408.07009, August 2024. doi: 10.48550/arXiv.2408.07009.

[20] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. 2018.

[22] Weiwei Kong and Umar Syed. A picture's worth a thousand (private) words: Hierarchical generation of coherent synthetic photo albums, 2025. URL https://research.google/blog/a-pictures-worth-a-thousand-private-words-hierarchical-generation-of-coherent-synthetic-photo-albums/.

[23] Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*, 2023.

[24] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[25] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *International Conference on Learning Representations*, 2022.

[26] Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. Differentially Private Synthetic Data via Foundation Model APIs 1: Images. *arXiv e-prints*, art. arXiv:2305.15560, May 2023. doi: 10.48550/arXiv.2305.15560.

[27] Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. Differentially private synthetic data via foundation model apis 1: Images. *International Conference on Learning Representations (ICLR)*, 2024.

[28] Zinan Lin, Tadas Baltrusaitis, and Sergey Yekhanin. Differentially private synthetic data via apis 3: Using simulators instead of foundation model. *arXiv preprint arXiv:2502.05505*, 2025.

[29] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[30] 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*, 2024.

[31] Michael F Liu, Saiyue Lyu, Margarita Vinaroz, and Mijung Park. Differentially private latent diffusion models. *Transactions on Machine Learning Research*.

[32] Michael F. Liu, Saiyue Lyu, Margarita Vinaroz, and Mijung Park. DP-LDMs: Differentially Private Latent Diffusion Models. *arXiv e-prints*, art. arXiv:2305.15759, May 2023. doi: 10.48550/arXiv.2305.15759.

[33] 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)*, 2015.

[34] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE Computer Society, 2023.

[35] Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. Differentially private language models for secure data sharing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4873, 2022.

[36] OpenAI. ChatGPT. https://chat.openai.com, 2022.

[37] OpenAI. Dall·e 3. https://openai.com/dall-e, 2023. Accessed: 2025-06-09.

[38] OpenAI. Gpt-4o-mini model. https://openai.com, 2024. Accessed 2025-05-16.

[39] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022.

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[41] Artem Reshetnikov, Maria-Cristina Marinescu, and Joaquim More Lopez. DEArt: Dataset of European Art. *arXiv e-prints*, art. arXiv:2211.01226, November 2022. doi: 10.48550/arXiv.2211.01226.

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[43] 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*, pages 10684–10695, 2022.

[44] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial Diffusion Distillation. *arXiv e-prints*, art. arXiv:2311.17042, November 2023. doi: 10.48550/arXiv.2311.17042.

[45] 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 e-prints*, art. arXiv:2111.02114, November 2021. doi: 10.48550/arXiv.2111.02114.

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *arXiv e-prints*, art. arXiv:1512.00567, December 2015. doi: 10.48550/arXiv.1512.00567.

[47] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[48] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[49] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[50] 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*, 2025.

[51] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[52] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Towards open-world text-guided face image generation and manipulation. *arxiv preprint arxiv: 2104.08910*, 2021.

[53] 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. In *International Conference on Machine Learning*, pages 54531–54560. PMLR, 2024.

[54] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[55] Chejian Xu, Jiawei Zhang, Zhaorun Chen, Chulin Xie, Mintong Kang, Yujin Potter, Zhun Wang, Zhuowen Yuan, Alexander Xiong, Zidi Xiong, et al. Mmdt: Decoding the trustworthiness and safety of multimodal foundation models. *arXiv preprint arXiv:2503.14827*, 2025.

[56] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[57] 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. *International Conference on Learning Representations*, 2022.

[58] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[59] Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, 2023.

[60] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 42(4), October 2017. ISSN 0362-5915.

[61] Longtao Zheng, Zhiyuan Huang, Zhenghai Xue, Xinrun Wang, Bo An, and Shuicheng Yan. AgentStudio: A Toolkit for Building General Virtual Agents. *arXiv e-prints*, art. arXiv:2403.17918, March 2024. doi: 10.48550/arXiv.2403.17918.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract in our paper clearly states that our method can improve the quality of image generation, and the results are listed in the experiment part.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our paper mentioned our the limitations in the last section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Our work has to prove our method follows DP constraints, this theoretical result is provided with a complete and detailed proof in method section.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We represented our idea in method section, and specified our experiment setup so that anyone who read our paper can reproduce our results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provided our code in the supplemental material, and we'll release a more formal version on Github soon.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The appendices in our paper specified most of our non-default hyper-parameters in our experiment.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our main contributions and experiment results don't need information about statistical significance to illustrate.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Our paper provided a part about compute resources in experiment section for reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We're sure that our work conform with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Our paper talked about potential positive societal impacts of our work. Our work provides a method for high quality private image generation, which better preserves privacy for private data source.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper releases no data or models that have such risks for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All creators or original owners of assets used in our paper are properly credited. And licenses and terms of use are explicitly mentioned and properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets are introduced in our paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper described a lot about the usage of LLMs since LLMs take a important role in our methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A   Details on our `Aug_PE_Image_Voting` algorithm

We present the pseudo-code for the `Aug_PE_Image_Voting` algorithm. While it may appear that the generated private text $\mathcal{T}$ is unused in Algorithm 2, this is by design. Image voting serves solely as a quality enhancement technique within our pipeline. In contrast, the original text voting mechanism explicitly requires the generated private text. To maintain the completeness and integrity of the overall pipeline, we retain the captioning process.

Algorithm 2 Aug-PE with Image Voting (`Aug_PE_Image_Voting`)

Require: Private image dataset $\mathcal{D}$, generated private text $\mathcal{T}$, population size N, number of iterations G, `RANDOM_API(·)`, `VARIATION_API(·)`, diffusion model API `Diffusion(·)`, image encoder `EncodeImage(·)`, DP noise multiplier σ, K, L

Ensure: Differentially private synthetic text data $\mathcal{T}'$

1: $\mathcal{C}_0 = \texttt{RANDOM\_API}(N)$
2: $\mathcal{E}_{\text{priv}} = \texttt{EncodeImage}(\mathcal{D})$
3: **for** $g = 0$ to $G - 1$ **do**
4:      **if** $K == 0$ **then**
5:          $\mathcal{I}_g = \texttt{Diffusion}(\mathcal{C}_g)$
6:          $\mathcal{E}_{\text{gen}} = \texttt{EncodeImage}(\mathcal{I}_g)$
7:      **else if** $K > 0$ **then**
8:          $\mathcal{C}_g^k = \texttt{VARIATION\_API}(\mathcal{C}_g), k = 1, 2, ..., K$
9:          $\mathcal{I}_g^k = \texttt{Diffusion}(\mathcal{C}_g^k), k = 1, 2, ..., K$
10:         $\mathcal{E}_{\text{gen}}^k = \texttt{EncodeImage}(\mathcal{I}_g^k), k = 1, 2, ..., K$
11:         $\mathcal{E}_{\text{gen}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{E}_{gen}^k$
12:     **end if**
13:     $H = \mathbf{0}^N = [0, 0, ..., 0]$
14:     **for each** $e_{priv} \in \mathcal{E}_{\text{priv}}$ **do**
15:         $i \leftarrow \arg \min_j d(e_{priv}, \mathcal{E}_{\text{gen}}[j])$
16:         $H[i] \leftarrow H[i] + 1$
17:     **end for**
18:     $H \leftarrow H + \mathcal{N}(0, \sigma^2 \mathbf{I})$
19:     $P \leftarrow H / \sum H$
20:     **if** $L == 1$ **then**
21:         $\mathcal{C}_g' \leftarrow$ draw N samples with replacement from $\mathcal{C}_g$
22:         $\mathcal{C}_{g+1} \leftarrow \texttt{VARIATION\_API}(\mathcal{C}_g')$
23:         save $\mathcal{C}_{g+1}^{syn} \leftarrow \mathcal{C}_g'$
24:     **else if** $L > 1$ **then**
25:         $\mathcal{C}_g' \leftarrow$ **rank samples** by probabilities P and draw top N samples
26:         $\mathcal{C}_{g+1}^l \leftarrow \texttt{VARIATION\_API}(\mathcal{C}_g'), l = 1, 2, ..., L - 1$
27:         $\mathcal{C}_{g+1} = [C_g', C_{g+1}^1, ..., C_{g+1}^{L-1}]$
28:         save $\mathcal{C}_{g+1}^{syn} \leftarrow \mathcal{C}_{g+1}$
29:     **end if**
30: **end for**
31: **return** Final text candidates $\mathcal{C}_G$ as synthetic data $\mathcal{T}'$

B   More on our experiment results

Here we post more results of our experiments, and provide more experiment details.

B.1   FID comparison between PE and SPTI

Here we present details of our experiment results. We use FID [16] to compare between *SPTI* and PE on a variaty of datasets. We give specific dataset name and DP constraint in the caption of each image. We find *SPTI* generally better than PE, and voted samples is generally better than variants in most iterations. Interestingly, voted samples is often slightly worse than variants in the inital iteration.
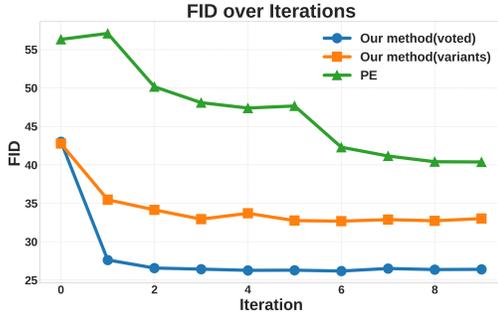
Figure 8: Experiment results on **LSUN Bedroom** dataset. ($\epsilon = 1.0$)
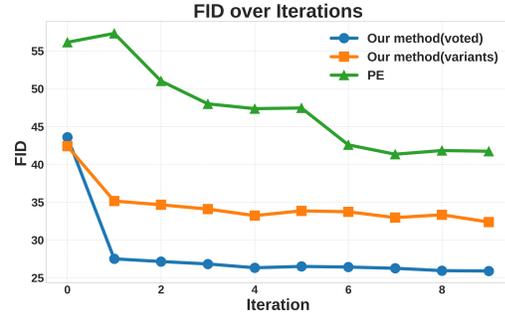
Figure 9: Experiment results on **LSUN Bedroom** dataset. ($\epsilon = 10.0$)
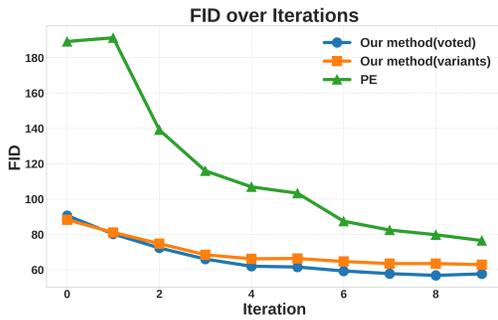

Figure 10: Experiment results on **European Art** dataset. ($\epsilon = 1.0$)
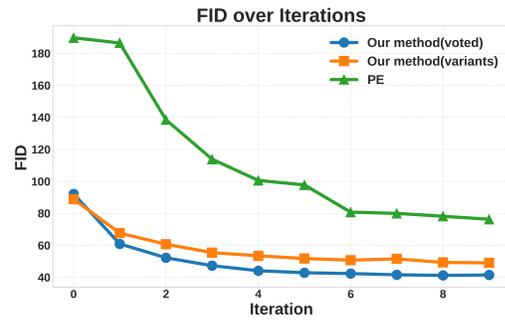
Figure 11: Experiment results on **European Art** dataset. ($\epsilon = 10.0$)
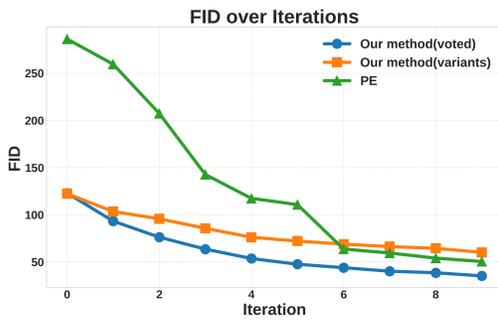
Figure 12: Experiment results on **Wave-ui-25k** dataset. ($\epsilon = 1.0$)
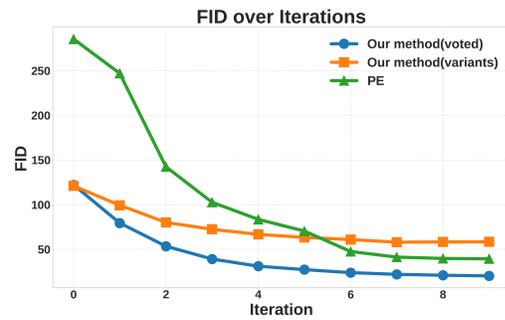
Figure 13: Experiment results on **Wave-ui-25k** dataset. ($\epsilon = 10.0$)
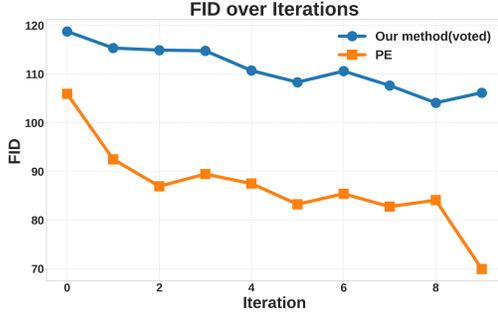
Figure 14: Experiment results on **Cat** dataset. ($\epsilon = 1.0$)
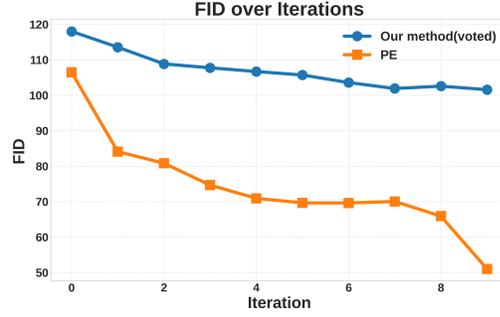
Figure 15: Experiment results on **Cat** dataset. ($\epsilon = 10.0$)
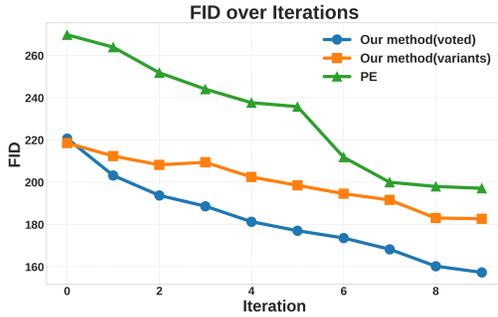
Figure 16: Experiment results on **Sprite Fright** dataset. ($\epsilon = 1.0$)
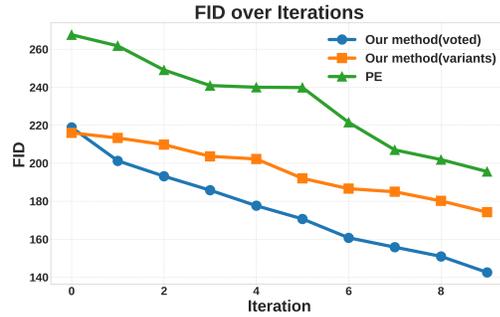
Figure 17: Experiment results on **Sprite Fright** dataset. ($\epsilon = 10.0$)

B.2 FID comparison between Image Voting and Text Voting

Here we present details of our experiment results. We use FID [16] to compare between image voting and text voting on a variaty of datasets. We give specific dataset name and DP constraint in the caption of each image. We find image voting is generally better than text voting under most conditions. Interestingly, image voting and text voting show the same behavior in the inital iteration, i.e. voted samples is slightly worse than variants in the first iteration, no matter image voting or text voting is used.
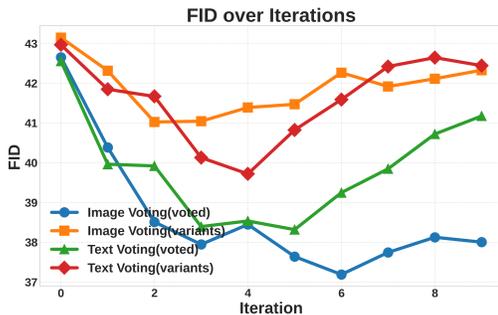
Figure 18: Experiment results on **LSUN Bedroom** dataset. ($\epsilon = 1.0$)
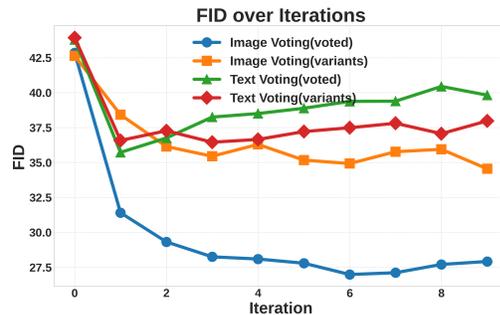
Figure 19: Experiment results on **LSUN Bedroom** dataset. ($\epsilon = 10.0$)

25
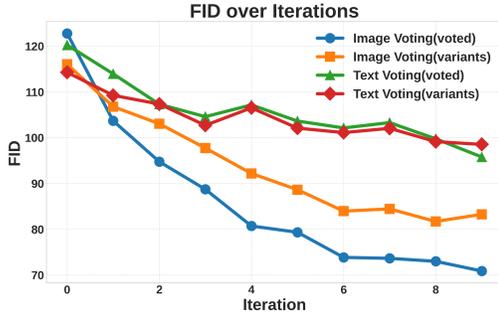
Figure 20: Experiment results on **Wave-ui-25k** dataset. ($\epsilon = 1.0$)
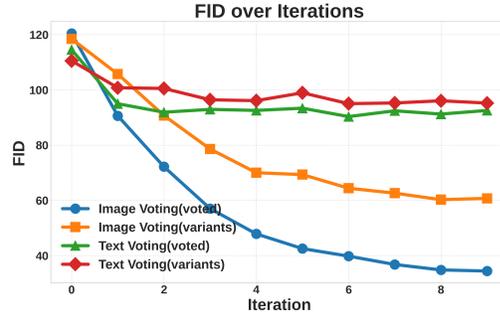
Figure 21: Experiment results on **Wave-ui-25k** dataset. ($\epsilon = 10.0$)
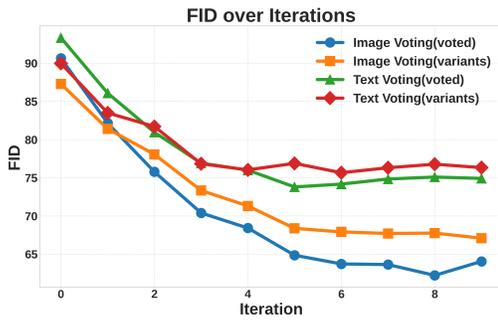
Figure 22: Experiment results on **European Art** dataset. ($\epsilon = 1.0$)
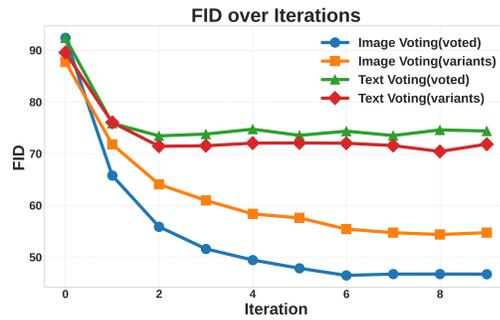
Figure 23: Experiment results on **European Art** dataset. ($\epsilon = 10.0$)
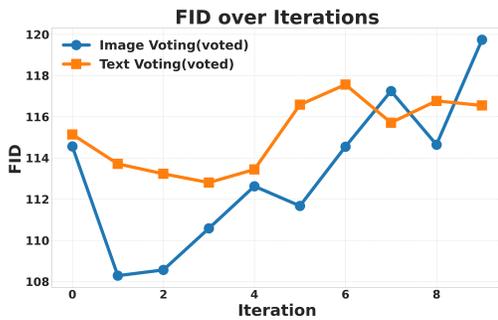
Figure 24: Experiment results on **Cat** dataset. ($\epsilon = 1.0$)
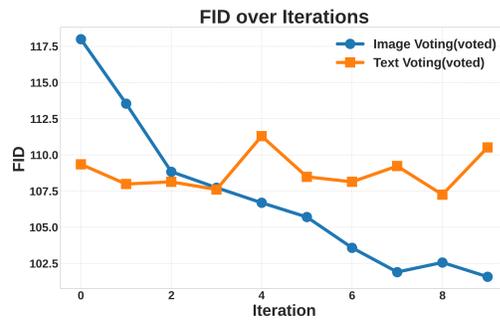
Figure 25: Experiment results on **Cat** dataset. ($\epsilon = 10.0$)

B.3 Downstream Task

Despite the high quality of generated images in our method, we also consolidate their value in usage for downstream tasks. To be more specific, we test the classification accuracy on **CelebA** dataset using **WRN-40-4** model. Here we provide figures of model performance during training.
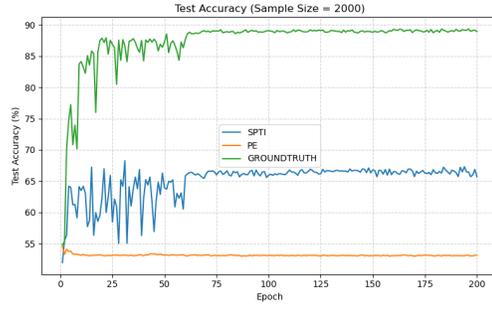
Figure 26: Training performance (test accuracy) on **CelebA** dataset with the same number of generated samples. ($\epsilon = 10.0$, `sample_size=2000`)
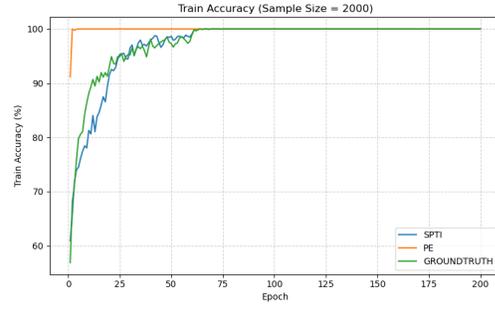
Figure 27: Training performance (train accuracy) on **CelebA** dataset with the same number of generated samples. ($\epsilon = 10.0$, `sample_size=2000`)
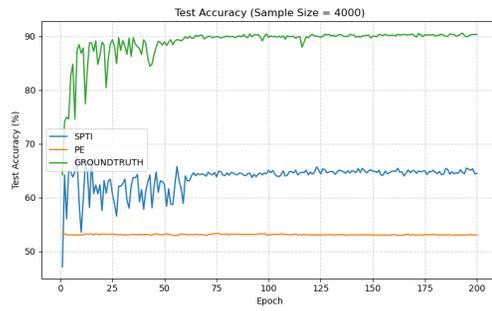
Figure 28: Training performance (test accuracy) on **CelebA** dataset with the same number of generated samples. ($\epsilon = 10.0$, `sample_size=4000`)
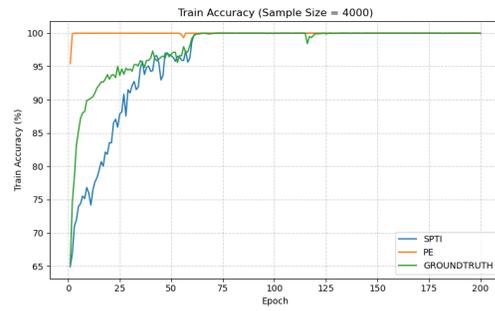
Figure 29: Training performance (train accuracy) on **CelebA** dataset with the same number of generated samples. ($\epsilon = 10.0$, `sample_size=4000`)
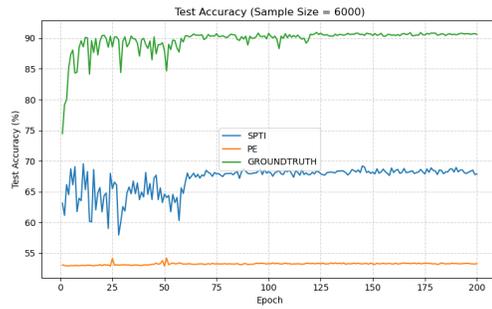
Figure 30: Training performance (test accuracy) on **CelebA** dataset with the same number of generated samples. ($\epsilon = 10.0$, `sample_size=6000`)
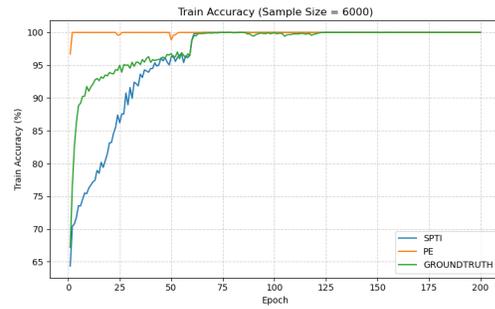
Figure 31: Training performance (train accuracy) on **CelebA** dataset with the same number of generated samples. ($\epsilon = 10.0$, `sample_size=6000`)
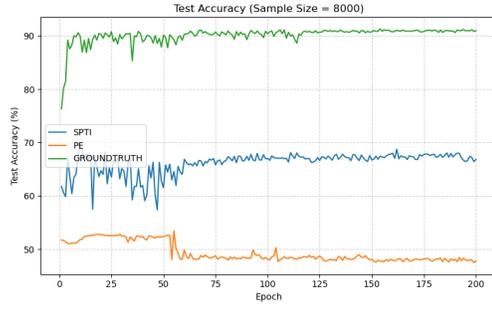
Figure 32: Training performance (test accuracy) on **CelebA** dataset with the same number of generated samples. ($\epsilon = 10.0$, `sample_size=8000`)
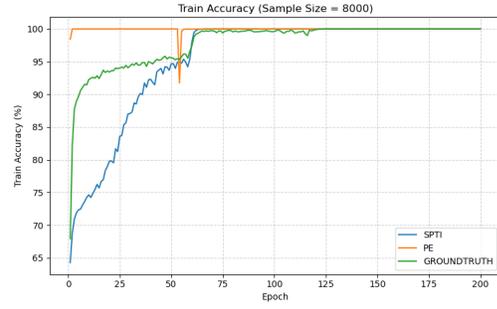
Figure 33: Training performance (train accuracy) on **CelebA** dataset with the same number of generated samples. ($\epsilon = 10.0$, `sample_size=8000`)
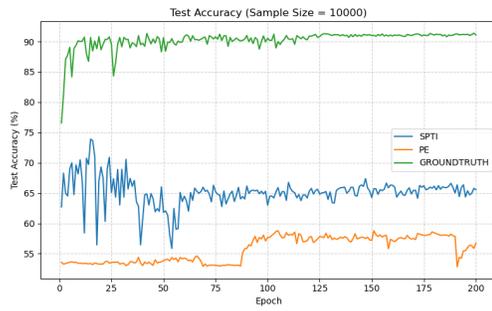
Figure 34: Training performance (test accuracy) on **CelebA** dataset with the same number of generated samples. ($\epsilon = 10.0$, `sample_size=10000`)
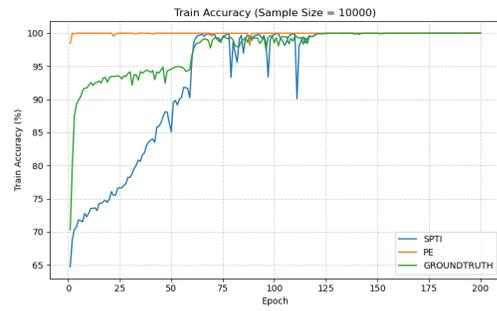
Figure 35: Training performance (train accuracy) on **CelebA** dataset with the same number of generated samples. ($\epsilon = 10.0$, `sample_size=10000`)

Figure 36: Generated samples by *SPTI* ($\epsilon = 10.0$, `sample_size=10000`)

Figure 37: Generated samples by PE image ($\epsilon = 10.0$, `sample_size=10000`)

C Extended Ablation Study

C.1 Large Language Model and Diffusion Model APIs

Along with the effect of our image voting trick, we also tested our method with different LLM and diffusion model APIs along with other hyperparameters. For large language models, we choose **qwen-plus** [3] as comparison to our original **Meta-Llama-3-8B-Instruct** [2] model. For diffusion model APIs, we choose **stable-diffusion-xl-base-1.0** [44] to compare with our original model. We also present Figure 38, 39, 40, 41 generated from DP synthetic data by Imagen3 [19] and DALLE3 [37], i.e. we acquire DP synthetic images by running *SPTI* using **Meta-Llama-3-8B-Instruct** [2] and **stable-diffusion-xl-base-1.0** [44], but generate final DP synthetic images using different image-generation API.

	SDXL-Turbo	SDXL-base-1.0	Infinity
Meta-Llama-3-8B-Instruct	26.71	25.42	30.66
qwen-plus	26.65	24.44	31.28

Table 4: **FID** (lower is better) on **LSUN Bedroom** with $\epsilon = 1.0$, tested using different LLM and diffusion APIs. All other settings are kept identical to the default configuration. The results indicate that while different API backends (e.g., SDXL-Turbo, SDXL-base-1.0, Infinity) and language models (e.g., Meta-Llama-3-8B-Instruct, Qwen-Plus) introduce minor variations in FID, their overall influence on the final image quality is limited. This demonstrates the robustness of our method to the choice of model and API implementation.

Figure 38: Images generated by Imagen3 [19] from DP-synthetic text of dataset **LSUN Bedroom** ($\epsilon = 1.0$). **FID** (lower is better) is 26.58, which is slightly better than using SDXL-Turbo 26.71.
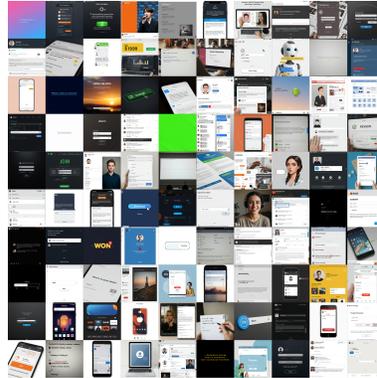
Figure 39: Images generated by Imagen3 [19] from DP-synthetic text of dataset **Wave-ui-25k** ($\epsilon = 1.0$). **FID** (lower is better) is 72.13, which is worse than using SDXL-Turbo 36.52.

Figure 40: Images generated by DALLE3 [37] from DP-synthetic text of dataset **LSUN Bedroom** ($\epsilon = 1.0$). FID (lower is better) is 29.96, which is slightly worse than using SDXL-Turbo 26.71

Figure 41: Images generated by DALLE3 [37] from DP-synthetic text of dataset **Wave-ui-25k** ($\epsilon = 1.0$). FID (lower is better) is 150.68, which is worse than SDXL-Turbo 36.52

D   Explanation of Special Conditions

In the previous experiment on **Cat** dataset, we find our method to be less effective compared to baseline method (**PE**). So we conducted a contrast experiment to analyze the cause behind the phenomenon.

Firstly, we build a small sub-dataset from **ImageNet100** dataset called **bird** dataset. This dataset consists of 100 images labeled *goldfinch* and 100 images labeled *indigo bird*, which corresponds to label id 70 and 2 in the **ImageNet100** dataset. We found similar results to **Cat** dataset, so we take **bird** dataset as an alternate baseline.

Secondly, we proposed two possible reasons for the phenomenon that *SPTI* performs poorer than **PE** on **Cat** and **bird** dataset. One is that the composition of dataset is not diverse, while *SPTI* uses Aug-PE to generate DP data, which brings more diversity than PE image method, causing worse results. The other is that the *SPTI* performs naturally poorer than PE when the total number of dataset is small. We build another 3 datasets 5 according to these two reasons.

Here we reference our newly-built dataset using their composition in the following table. We only use images from **ImageNet100** when building all the four dataset. The experiment settings follows the experiments on **Cat** dataset in Table 6. The experiment results in Table 5 shows that the reason why *SPTI* performs poorer than PE on **Cat** and **bird** dataset is because diversity. *SPTI* performs naturally poorer on dataset that is less diverse. This provides us with insight that text modality may bring more diversity on DP data generation than using image.

E   Details of Experimental Configurations

To facilitate reproducibility, we detail the hyperparameter settings used for all experiments in Section 4. For hyperparameters not mentioned, the default values are used. For experiment settings of ablation studies between image voting and text voting, we use the same hyperparameters in Table 6 below. For comparison experiments on *SPTI*, PE and DP fine-tuning, we present our configuration details in Table 7. For specific configuration of DP fine-tuning, we present our settings in Table 8.

F   Datasets

In this section, we summarize datasets used in our experiments with their stats, i.e. total number of images, resolution, etc. We represent all information in Table 9. Here we record the number of images and resolution in our experiments.

Table 5: **FID** (lower is better) comparison of SPTI and PE on Different Datasets

Dataset Composition	Total Number of Images	SPTI (ours)	PE
goldfinch*100 + indigo bird*100	200	144.18	**110.96**
20 kinds of different birds, each kind 10 images	200	**111.84**	168.39
goldfinch*1000 + indigo bird*1000	2000	106.26	**78.66**
20 kinds of different birds, each kind 100 images	2000	**84.25**	172.93

Dataset	Configurations		SPTI(ours)	PE image
LSUN bedroom	PE.run()	num_samples_schedule	2000	2000
		iterations	10	10
	ImageVotingNN()	lookahead_degree	0	-
	NNhistogram()	lookahead_degree	-	4
	PEPopulation()	initial_variation_api_fold	6	0
		next_variation_api_fold	6	1
Cat	PE.run()	num_samples_schedule	200	200
		iterations	10	10
	ImageVotingNN()	lookahead_degree	4	-
	NearestNeighbors()	lookahead_degree	-	4
	PEPopulation()	initial_variation_api_fold	0	0
		next_variation_api_fold	1	1
Wave-ui-25k	PE.run()	num_samples_schedule	2000	2000
		iterations	10	10
	ImageVotingNN()	lookahead_degree	0	-
	NNhistogram()	lookahead_degree	-	4
	PEPopulation()	initial_variation_api_fold	6	0
		next_variation_api_fold	6	1
Europeart	PE.run()	num_samples_schedule	2000	2000
		iterations	10	10
	ImageVotingNN()	lookahead_degree	0	-
	NNhistogram()	lookahead_degree	-	4
	PEPopulation()	initial_variation_api_fold	6	0
		next_variation_api_fold	6	1
Sprite Fright	PE.run()	num_samples_schedule	1000	1000
		iterations	10	10
	ImageVotingNN()	lookahead_degree	0	-
	NNhistogram()	lookahead_degree	-	4
	PEPopulation()	initial_variation_api_fold	6	0
		next_variation_api_fold	6	1

Table 6: Experiment Settings for **FID** evaluation.

Configurations		SPTI(ours)	PE image
PE.run()	num_samples_schedule	2000	2000
	iterations	10	10
ImageVotingNN()	lookahead_degree	0	-
NNhistogram()	lookahead_degree	-	4
PEPopulation()	initial_variation_api_fold	6	0
	next_variation_api_fold	6	1

Table 7: Experiment Settings of *SPTI* and PE for **FID** evaluation of **MM-Celeba-HQ** dataset

Table 8: Hyperparameters for fine-tuning diffusion models with DP constraints $\epsilon = 10, 1$ and $\delta = 10^{-5}$ on text-conditioned CelebAHQ.

	$\epsilon = 10$	$\epsilon = 1$
batch size	256	256
base learning rate	1×10^{-7}	1×10^{-7}
learning rate	2.6×10^{-5}	2.6×10^{-5}
epochs	10	10
clipping norm	0.01	0.01
noise scale	0.55	1.46
ablation	-1	-1
num of params	280M	280M
use_spatial_transformer	True	True
cond_stage_key	caption	caption
context_dim	1280	1280
conditioning_key	crossattn	crossattn
transformer depth	1	1

Dataset	Total number of images	Resolution	Source
LSUN	300,000	256×256	https://github.com/fyu/lsun
Cat	200	512×512	https://www.kaggle.com/datasets/fjxmlzn/cat-cookie-doudou/
Wave-ui-25k	24,798	256×256	https://huggingface.co/datasets/agentsea/wave-ui-25k
European Art	15,154	256×256	https://huggingface.co/datasets/biglam/european_art
Sprite Fright	13,077	256×256	https://studio.blender.org/projects/sprite-fright/
MM-Celeba-HQ	30,000	256×256	https://github.com/IIGROUP/MM-CelebA-HQ-Dataset
CelebA	202,599	178×178	https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

Table 9: Statistics of datasets used in our experiments.