DIFFERENTIALLY PRIVATE ADAPTATION OF DIFFUSION MODELS VIA NOISY AGGREGATED EMBEDDINGS

Pura Peetathawatchai
Stanford University
pura@stanford.edu

Wei-Ning Chen
Microsoft
weiningchen@microsoft.com

Berivan Isik
Google
berivan@google.com

Sanmi Koyejo
Stanford University
sanmi@cs.stanford.edu

Albert No
Yonsei University
albertno@yonsei.ac.kr

November 21, 2024

ABSTRACT

We introduce novel methods for adapting diffusion models under differential privacy (DP) constraints, enabling privacy-preserving style and content transfer without fine-tuning. Traditional approaches to private adaptation, such as DP-SGD, incur significant computational overhead and degrade model performance when applied to large, complex models. Our approach instead leverages embedding-based techniques: Universal Guidance and Textual Inversion (TI), adapted with differentially private mechanisms. We apply these methods to Stable Diffusion for style adaptation using two private datasets: a collection of artworks by a single artist and pictograms from the Paris 2024 Olympics. Experimental results show that the TI-based adaptation achieves superior fidelity in style transfer, even under strong privacy guarantees, while both methods maintain high privacy resilience by employing calibrated noise and subsampling strategies. Our findings demonstrate a feasible and efficient pathway for privacy-preserving diffusion model adaptation, balancing data protection with the fidelity of generated images, and offer insights into embedding-driven methods for DP in generative AI applications.

1 Introduction

In recent years, diffusion models [1, 2], particularly latent diffusion models [3], have spearheaded high quality text-to-image generation, and have been widely adopted by researchers and the general public alike. Trained on massive datasets like LAION-5B [4], these models have developed a broad understanding of visual concepts, enabling new creative and practical applications. Notably, tools like Stable Diffusion [3, 5] have been made readily accessible for general use. Building on this foundation, efficient adaptation methods such as parameter efficient fine-tuning (PEFT) [6, 7, 8], guidance based approaches [9, 10, 11], and pseudo-word generation [12] enable users to leverage this extensive pretraining for customizing models that can specialize on downstream tasks with smaller datasets.

However, the rapid adoption of diffusion models has also raised significant privacy, ethical and legal concerns. One critical issue is the vulnerability of these models to privacy attacks, from membership inference [13], where an attacker determines whether a specific data point was used to train a particular model, to data extraction [14], which enables an attacker to reconstruct particular images from the training dataset. This issue is even more severe during the fine-tuning phase where the model is fine-tuned on smaller specialized datasets from a possibly different domain and each data record has more impact on the final model. This risk underscores the importance of privacy-preserving technologies, particularly as diffusion models often rely on vast datasets scraped from the internet without explicit consent from content owners. Copyright concerns further complicate the landscape [15]. In particular, diffusion models can closely reproduce original artworks without credit or compensation to the artist, as evident in data extraction attacks [14]. Such

concerns intensify the debate on balancing technological advancement with the rights and privacy of content owners, underscoring the need for clearer ethical and legal guidelines in the training and deployment of generative models.

Differential privacy (DP) [16, 17] is a widely adopted approach to addressing these challenges, where controlled noise is added during training or inference to prevent information leakage from individual data points while still enabling the model to learn effectively from the overall dataset. The standard method for ensuring DP in deep learning is Differentially Private Stochastic Gradient Descent (DP-SGD) [18], which modifies traditional SGD by adding noise to clipped gradients.

While, in theory, DP-SGD can be applied to train diffusion models, it poses several significant practical challenges. First, it is well known that DP-SGD introduces considerable computational and memory overhead due to the need for per-sample gradient clipping, which is crucial for bounding sensitivity [16, 18]. Second, DP-SGD is fundamentally incompatible with batch-wise operations, such as batch normalization, since these operations inherently link samples together, rendering sensitivity analysis infeasible. Moreover, training a large network with DP-SGD typically leads to significant performance degradation (i.e. the curse of dimensionality), even under relatively lenient privacy budgets. To the best of our knowledge, existing diffusion models trained with DP-SGD are constrained to relatively small-scale datasets [19, 20].

As a result, recent research has increasingly turned to alternative strategies for adapting diffusion models in a privacy-preserving manner without the need for training from scratch under DP-SGD. A promising approach leverages large, publicly pre-trained models, adapting them to new domains under differential privacy (DP) constraints, which allows models to capitalize on existing representational strengths while avoiding the computational and memory burdens of full DP training [20]. Similarly, parameter-efficient fine-tuning (PEFT) methods, such as DP-LoRA [21], fine-tune only a subset of parameters, allowing for efficient, targeted adaptation with limited privacy costs. Methods like DP-RDM (Differentially Private Retrieval-Augmented Diffusion Model) [22] illustrate this trend by enabling model adaptation through a retrieval mechanism that conditions image generation on private data retrieved at inference time, avoiding direct model updates. These approaches collectively demonstrate how large models can be adapted privately, circumventing the complexities and high costs associated with full DP training.

Meanwhile, alternative methods such as Universal Guidance [11] and Textual Inversion (TI) [12] enable model adaptation to specific content or style without modifying the diffusion model itself, instead relying on external embedding vectors. Universal Guidance's style guidance (SG) mechanism generates a target CLIP embedding by passing the reference image through a CLIP image encoder. During the reverse diffusion process, the cosine distance between this target embedding and the CLIP embedding of the generated image is minimized, guiding the generation process to match the desired style. TI, in contrast, learns a new token embedding representing a set of target images, which is then incorporated into the text prompt to adapt the model to the style or content associated with those images. Both SG and TI avoid direct model optimization, reducing memory and computation demands while also involving fewer trainable parameters, making them highly compatible with DP constraints on smaller datasets.

In this work, we propose two novel privacy-preserving adaptation methods for smaller datasets, leveraging SG and TI to bypass the need for DP-SGD or DP-PEFT approaches that require extensive model updating. Our first approach adapts Universal Guidance for privacy by using a differentially private CLIP embedding. Although SG is typically applied to a single target image, we extend it to handle multiple images under DP by computing an embedding for each target image and aggregating these embeddings via a centroid calculation. To ensure differential privacy, we add calibrated noise to the centroid, which protects individual data points in the dataset while preserving the model's guidance capabilities.

Our second approach enhances Textual Inversion (TI) with differential privacy. While standard TI compresses training data into a single, low-dimensional vector, offering some intrinsic privacy benefits, it lacks the rigorous guarantees required by DP. We address this by introducing a private variant of TI, in which we decouple interactions among samples by learning a separate embedding for each target image. These embeddings are then aggregated into a noisy centroid, similar to our SG approach, ensuring differential privacy while preserving TI's data compression advantages.

Our experimental results underscore the practicality and effectiveness of the proposed differentially private embedding-based adaptation methods, with Textual Inversion (TI) proving particularly robust for preserving stylistic fidelity even under privacy constraints. In applying our approach to two distinct datasets–a private collection of artworks by `@eveismyname` and pictograms from the Paris 2024 Olympics [23]–we demonstrate that differentially private TI can successfully capture nuanced stylistic elements while maintaining privacy guarantees. Our findings reveal a trade-off between privacy levels (determined by the DP parameter ε) and image quality, where lower ε values yield diminished stylistic fidelity, yet still preserve the target style under moderate noise. Furthermore, subsampling offers an effective means of improving privacy resilience by controlling sensitivity to individual data points, significantly mitigating the impact of added noise on image quality. Ultimately, our approach provides a versatile framework for privacy-preserving

model adaptation, enabling diffusion models to generalize effectively to novel styles and domains without risking exposure of sensitive data points.

2 Background and Related Works

2.1 Diffusion Models

Diffusion models [1, 2, 24, 3] leverage an iterative denoising process to generate high-quality images that aligns with a given conditional input from random noise. In text-to-image generation, this conditional input is based on a textual description (a prompt) that guides the model in shaping the image to reflect the content and style specified by the text. To convert the text prompt into a suitable conditional format, it is first tokenized into discrete tokens, each representing a word or sub-word unit. These tokens are then converted into a sequence of embedding vectors v_i that encapsulate the meaning of each token within the model's semantic space. Next, these embeddings pass through a transformer text encoder, such as CLIP [25], outputting a single text-conditional vector y that serves as the conditioning input. This text-conditioning vector y is then incorporated at each denoising step, guiding the model in aligning the output image with the specific details outlined in the prompt.

The image generation process, also known as the reverse diffusion process, comprises of T discrete timesteps and starts with pure Gaussian noise x_T. At each decreasing timestep t, the denoising model, which often utilizes a U-Net structure with cross-attention layers, takes a noisy image x_t and text conditioning y as inputs and predicts the noise component $\epsilon_\theta(x_t, y, t)$, where θ denotes the denoising model's parameters. The predicted noise is then used to make a reverse diffusion step from x_t to x_{t-1}, iteratively refining the noisy image closer to a coherent output x_0 that aligns with the text conditional y.

In particular, denoising diffusion implicit models (DDIM) sampling [24] uses the predicted noise $\epsilon_\theta(x_t, y, t)$ and a noise schedule represented by an array of scalars $\{\alpha_t\}_{t=1}^T$ to first predict a clean image $\hat{x}_0$, then makes a small step in the direction of $\hat{x}_0$ to obtain x_{t-1}. The reverse diffusion process for DDIM sampling can be formalized as follows:

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, y, t)}{\sqrt{\alpha_t}} \tag{1}$$

$$x_{t-1} = \sqrt{\alpha_{t-1}}\hat{x}_0 + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_t, y, t) \tag{2}$$

The training of a text-conditioned diffusion model involves minimizing a loss function that guides the model to predict the noise added to an image, given both the noisy image x_t and the text conditioning y. The objective function is typically a mean squared error (MSE) between the true noise ϵ and the predicted noise $\epsilon_\theta(x_t, y, t)$. The denoising model is therefore trained over the following optimization problem:

$$\theta^* = \arg\min_\theta \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,I),t\sim[T]}\left[\|\epsilon - \epsilon_\theta(x_t, y, t)\|^2\right] \tag{3}$$

2.1.1 Style Guidance

Bansal et al. [11] propose a universal guidance algorithm that allows the reverse diffusion process to be guided by arbitrary guidance functions, including object detection, image segmentation and facial recognition. In this section, we formally describe CLIP style guidance, one of the many processes explored in the universal guidance study. We also only describe the forward guidance process. For a generalized and complete description of universal guidance process, including the backward guidance process and per-step self-recurrence, we defer the reader to the original paper.

Let x_c denote a target image whose style we wish to replicate and $\mathsf{CLIP}(\cdot)$ denote a CLIP image encoder. To incorporate style guidance, given a noisy image x_t, we first compute $\hat{x}_0$ using the predicted noise $\epsilon_\theta(x_t, y, t)$ and Equation (1). Using $\hat{x}_0$, we then compute the following:

$$\hat{\epsilon}_\theta(x_t, y, t) = \epsilon_\theta(x_t, y, t) + w\sqrt{1-\alpha_t}\nabla_{x_t}\ell_{\cos}(\mathsf{CLIP}(x_c), \mathsf{CLIP}(\hat{x}_0)) \tag{4}$$

where w is a guidance weight parameter and $\ell_{\cos}(\cdot,\cdot)$ is the negative cosine loss.

We then perform the reverse diffusion step using the same procedure given in Equations (1) and (2), but replacing $\epsilon_\theta(x_t, y, t)$ with $\hat{\epsilon}_\theta(x_t, y, t)$. Using this reverse diffusion procedure, we generate a high-quality image x_0 that aligns with the text conditioning y while also optimizing the cosine similarity between the CLIP embeddings of x_c and x_0, resulting in x_0 having the stylistic characteristics of x_c.

2.1.2 Textual Inversion

Textual Inversion (TI) [12] is an adaptation technique that enables personalization using a small dataset of typically only 3-5 images. This approach essentially learns a new token that encapsulates the semantic meaning of the training images, allowing the model to associate specific visual features with a custom token.

To achieve this, TI trains a new token embedding, denoted as u, which represents a placeholder token, often denoted as S. During training, images are conditioned of phrases such as "A photo of S" or "A painting in the style of S". However, unlike the fixed embeddings of typical tokens v_i, u is a learnable parameter. Let y_u denote the text conditioning vector resulting from a prompt containing the token S. Through gradient descent, TI minimizes the diffusion model loss with respect to u, while keeping the diffusion model parameters θ fixed, iteratively refining this embedding to capture the unique characteristics of the training images. The resulting optimal embedding u_* is formalized as follows:

$$u_* = \arg\min_u \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,I),t\sim[T]}[\|\epsilon - \epsilon_\theta(x_t, y_u, t)\|^2]. \tag{5}$$

Hence, u_* represents an optimized placeholder token S_*, which can employed in prompts such as "A photo of S_* floating in space" or "A drawing of a capybara in the style of S_*", enabling the generation of personalized images that reflect the learned visual characteristics.

2.2 Differential Privacy

In this work, we adopt differential privacy (DP) [16, 17] as our privacy framework. Over the past decade, DP has become the gold standard for privacy protection in both research and industry. It measures the stability of a randomized algorithm concerning changes in an input instance, thereby quantifying the extent to which an adversary can infer the existence of a specific input based on the algorithm's output.

Definition 2.1 ((Approximate) Differential Privacy) *For $\varepsilon, \delta \geq 0$, a randomized mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ε, δ)-DP if for all* neighboring *datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^n$ which differ in a single record (i.e., $\|\mathcal{D} - \mathcal{D}'\|_{\mathsf{H}} \leq 1$ where $\|\cdot\|_{\mathsf{H}}$ is the Hamming distance) and all measurable $\mathcal{S}$ in the range of $\mathcal{M}$, we have that*

$$\mathbb{P}\left(\mathcal{M}(\mathcal{D}) \in \mathcal{S}\right) \leq e^{\varepsilon}\mathbb{P}\left(\mathcal{M}(\mathcal{D}') \in \mathcal{S}\right) + \delta.$$

When $\delta = 0$, we say $\mathcal{M}$ satisfies ε-pure DP or (ε-DP).

To achieve DP, the Gaussian mechanism is often applied [26, 27], adding Gaussian noise scaled by the sensitivity of the function f and privacy parameters ε and δ. Specifically, noise with standard deviation $\sigma = \frac{\Delta_f\sqrt{2\ln(1.25/\delta)}}{\varepsilon}$ is added to the output[1] [27], where Δ_f represents ℓ_2-sensitivity of the target function $f(\cdot)$. When the context is clear, we may omit the subscript f. This mechanism enables a smooth privacy-utility tradeoff and is widely used in privacy-preserving machine learning, including in DP-SGD [18], which applies Gaussian noise during model updates to achieve DP.

2.2.1 Privacy Amplification by Subsampling

Subsampling is a standard technique in DP, where a full dataset of size n is first subsampled to m records without replacement (typically with $m \ll n$), and the privatization mechanism (such as the Gaussian mechanism) is then applied. Specifically, if a mechanism provides (ε, δ)-DP on a dataset of size m, it achieves (ε', δ')-DP on the subsampled dataset, where $\delta' = \frac{m}{n}\delta$ and

$$\varepsilon' = \log\left(1 + \frac{m}{n}\left(e^{\varepsilon} - 1\right)\right) = O\left(\frac{m}{n}\varepsilon\right). \tag{6}$$

This result is well-known (see, for example, [29, Theorem 29]). Notably, if the DP mechanism is Gaussian, even tighter amplification bounds can be applied [28].

2.3 Differentially Private Adaptation of Diffusion Models

Recent advancements in applying differential privacy (DP) to diffusion models have aimed to balance privacy preservation with the high utility of generative outputs. Dockhorn et al. [19] proposed a differentially private diffusion model (DP diffusion) that enables privacy-preserving generation of realistic samples, setting a foundational approach for adapting diffusion processes under DP constraints. Another common strategy involves training a model on a large public dataset, followed by differentially private finetuning on a private dataset, as explored by Ghalebikesabi et al. [20].

[1] In practice, we use numerical privacy accountant such as [27, 28] to calibrate the noise.

While effective in certain contexts, this approach raises privacy concerns, particularly around risks of information leakage during the finetuning phase [30].

In response to these limitations, various adaptation techniques have emerged. Although not specific to diffusion models, some methods focus on training models on synthetic data followed by DP-constrained finetuning, as in the VIP approach [31], which demonstrates the feasibility of applying DP in later adaptation stages. Other approaches explore differentially private learning of feature representations [32], aiming to distill private information into a generalized embedding space while maintaining DP guarantees. Although these adaptations are not yet implemented for diffusion models, they lay essential groundwork for developing secure and efficient privacy-preserving generative models.

3 Differentially Private Adaptation via Noisy Aggregated Embeddings

3.1 General Approach

Let $x^{(1)}, \ldots, x^{(n)}$ represent a target dataset of images whose characteristics we wish to privately adapt our image generation towards. Assuming that we can encode each image $x^{(i)}$ as a directional embedding vector $u^{(i)}$, we can aggregate the embeddings $u^{(1)}, \ldots, u^{(n)}$ by calculating the centroid. The purpose of this aggregation is to limit the sensitivity of the final output to each $x^{(i)}$. In order to provide DP guanrantees, we also add isotropic Gaussian noise to the centroid. We can therefore define the resulting embedding vector u^* as follows:

$$u^* = \frac{1}{n} \sum_{i=1}^{n} u^{(i)} + \mathcal{N}(0, \sigma^2 I) \tag{7}$$

where the minimum σ required to provide (ε, δ)-DP is given by the following expression based on [27, Theorem 1]:

$$\sigma = \frac{\Delta}{n} \cdot \frac{\sqrt{2 \log(1.25)/\delta}}{\varepsilon}. \tag{8}$$

In the context of our problem, $\Delta = \sup_{i,j} \|u^{(i)} - u^{(j)}\|$. Since our embedding vectors are directional, we can normalize each $u^{(i)}$, allowing us to set $\Delta = 2$.

The noisy centroid embedding u^* can then be used to adapt the downstream image generation process. In Sections 3.3 and 3.4, we describe how to apply our method to TI and SG respectively, outlining the encoding of $x^{(i)}$ as $u^{(i)}$ and the role of u^* in the image generation process. To reduce the amount of noise needed to provide the same level of DP, we employ subsampling: instead of computing the centroid over all n embedding vectors, we randomly sample $m \leq n$ embedding vectors without replacement and compute the centroid over only the sampled vectors. Then the standard privacy amplification by subsampling bounds (such as (6)) can be applied. Formally, we sample $D_{\text{sub}} \subseteq \{u^{(1)}, \ldots, u^{(n)}\}$ where $|D_{\text{sub}}| = m$, and compute the output embedding as follows:

$$u^* = \frac{1}{m} \sum_{u^{(i)} \in D_{\text{sub}}} u^{(i)} + \mathcal{N}(0, \sigma^2 I), \tag{9}$$

where σ can be computed numerically for any target ε, δ and subsampling rate $\frac{m}{n}$.

3.2 Using Style Guidance

Applying our method to style guidance is straightforward. To obtain the embeddings, we simply use the CLIP image encoder to obtain the embedding for each image. More formally, for each $x^{(i)}$, let $u^{(i)} = \mathsf{CLIP}(x^{(i)})$. We can then aggregate the embeddings $u^{(i)}$ to obtain u^* via Equation (7) or (9), depending on whether we wish to incorporate subsampling. Finally, we use u^* to guide the reverse diffusion process, as outlined in Section 2.1.1, by computing $\hat{\epsilon}_\theta(x_t, y, t)$ via a modification of Equation (2):

$$\hat{\epsilon}_\theta(x_t, y, t) = \epsilon_\theta(x_t, y, t) + w\sqrt{1 - \alpha_t} \nabla_{x_t} \ell_{\cos}(u^*, \mathsf{CLIP}(\hat{x}_0)) \tag{10}$$

3.3 Using Textual Inversion

Textual Inversion (TI) is inherently parameter-efficient and offers certain privacy benefits, as information from an entire dataset of images is compressed into a single token embedding vector. This compression limits the model's capacity to memorize specific images, making data extraction attacks difficult. However, this privacy is merely heuristic and

yet to be proven, so TI may still be vulnerable to privacy attacks such as membership inference. A similar adapation technique with privacy guarantees may therefore be desirable.

To incorporate our approach into TI, instead of training a single token embedding on the entire dataset, we instead train a separate embedding $u^{(i)}$ on each $x^{(i)}$ to obtain a set of embeddings $u^{(1)}, \ldots, u^{(n)}$. We can formalize the encoding process as follows:

$$u^{(i)} = \arg\min_{u} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I),t}[\|\epsilon - \epsilon_\theta(x_t^{(i)}, y_u, t)\|^2] \tag{11}$$

We can then aggregate the embeddings $u^{(i)}$ to obtain u^* via Equation (7) or (9), depending on whether we wish to utilize subsampling. Similar to TI's u_*, we can use u^* to represent a new placeholder token S^* that can be incorporated into prompts for personalized image generation. While u^* may not fully solve the TI optimization problem presented in Equation (5), it provides provable privacy guarantees, with only a minimal trade-off in accurately representing the content or style of the training data.

4 Experimental Results

4.1 Datasets

We compiled two datasets for our style adaptation experiments, designed to proxy for private datasets unrecognized by the base Stable Diffusion model. The first dataset consists of 158 artworks created by the artist `@eveismyname`, who has provided consent for non-commercial use. This dataset serves to examine how models can capture and replicate artistic styles without directly copying individual works. Although we acknowledge the possibility that a few of these artworks may have been inadvertently included in the pre-training of Stable Diffusion due their public presence on social media platforms, the artist's limited recognition and modest portfolio size make it unlikely that their unique style could be effectively extracted from the pre-trained model. The second dataset consists of 47 pictograms from the Paris 2024 Olympics [23], approved strictly for non-commercial editorial use [33]. These pictograms were unveiled in February 2023, several months after the release of Stable Diffusion v1.5, which serves as the base model for our experiments. Through this dataset, we hope to evaluate how effectively our approach adapts to novel and unfamiliar visual patterns. Representative samples from both datasets are presented in Figure 1.

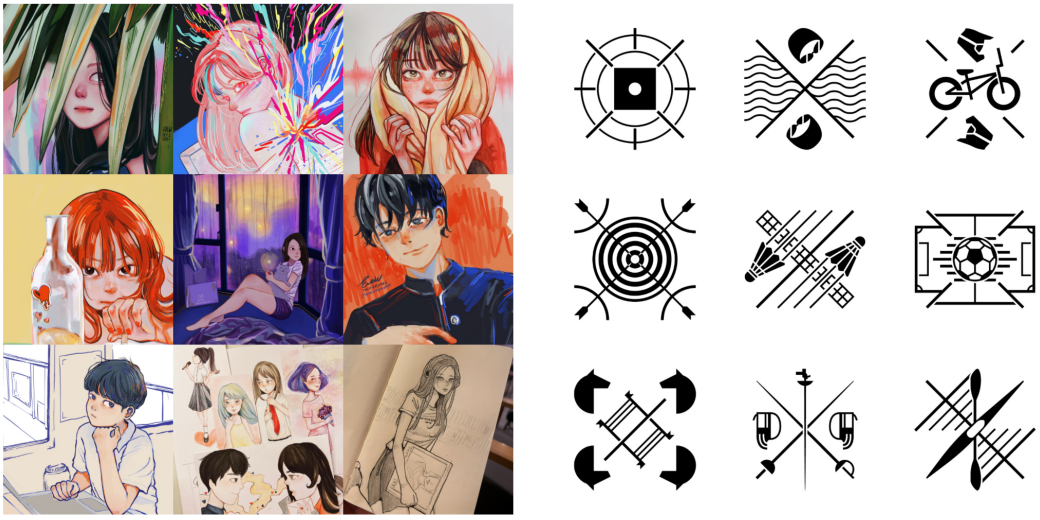

Figure 1: Samples of images used in our style adaptation experiments. **Left:** artwork by `@eveismyname` ($n = 158$). **Right:** Paris 2024 Olympic pictograms ($n = 47$), © *International Olympic Committee, 2023*.

4.2 Style Guidance Results

We applied our SG-based [11] approach, detailed in Section 3.2 to both datasets. While this method successfully provided privacy protection by obfuscating embedding details, the resulting image quality and style capture was notably less compared to the TI-based approach. The CLIP embeddings, even without noise or subsampling, led to images that

retained only generalized stylistic elements, lacking the detailed fidelity and coherence achieved with TI. Results of our attempt can be seen in Figure 2. This result underscores the superiority of our TI method for maintaining both privacy and high-quality image generation.

A possible explanation for the difference in style transfer effectiveness is that SG, as with other guidance methods, suffers from instability and sensitive dependency to hyperparameters such as guidance weight w. While Bansal et al. proposed using backward guidance and per-step self-recurrence to remedy these issues, we found that they were insufficient for our application. Another possible explanation is that the stylistic signal of the image embedding vectors $\mathsf{CLIP}(x^{(i)})$ are simply not resilient to our aggregation process.

Figure 2: Attempts of using universal guidance to generate drawings of Taylor Swift and icons of the Eiffel Tower in the styles of `@eveismyname` and Paris 2024 Pictograms respectively. Here, we apply no subsampling or DP-noise.

4.3 Textual Inversion Results

Figure 3: Images generated by Stable Diffusion v1.5 using the prompt "A painting of Taylor Swift in the style of `<@eveismyname>`", with the embedding `<@eveismyname>` trained using different values of m and ε.

Using both the `@eveismyname` and Paris 2024 pictograms dataset, we trained TI [12] embeddings on Stable Diffusion v1.5 [3] using our proposed method detailed in Section 3.3. Our primary goal is to investigate how DP configurations, specifically the ε and subsampling sizes (m), affect the generated images' quality and privacy resilience. For regular TI, we utilize Stable Diffusion's default process to embed the private dataset without any additional noise. For the DP-enhanced experiments using our method, we test multiple configurations of m and ε to analyze the trade-off between image fidelity and privacy resilience.

Figures 3 and 4 present generated images across three key configurations: (1) regular TI without DP, (2) centroid TI with DP at various epsilon values, and (3) centroid TI with DP at different subsampling sizes within the private dataset. We used the same random seed to generate embeddings and sample the images for ease of visual comparision between different configurations. As with common practice, we set $\delta = 1/n$. Since σ is undefined for $\varepsilon = 0$, we demonstrate the results of $\varepsilon \approx 0$, in other words, infinite noise, by setting $\varepsilon = 10^{-5}$. Images generated without DP closely resemble the unique stylistic elements of the target dataset. In particular, images adapted using `@eveismyname` images displayed crisp details and nuanced color gradients characteristic of the artist's work, while those of Paris 2024

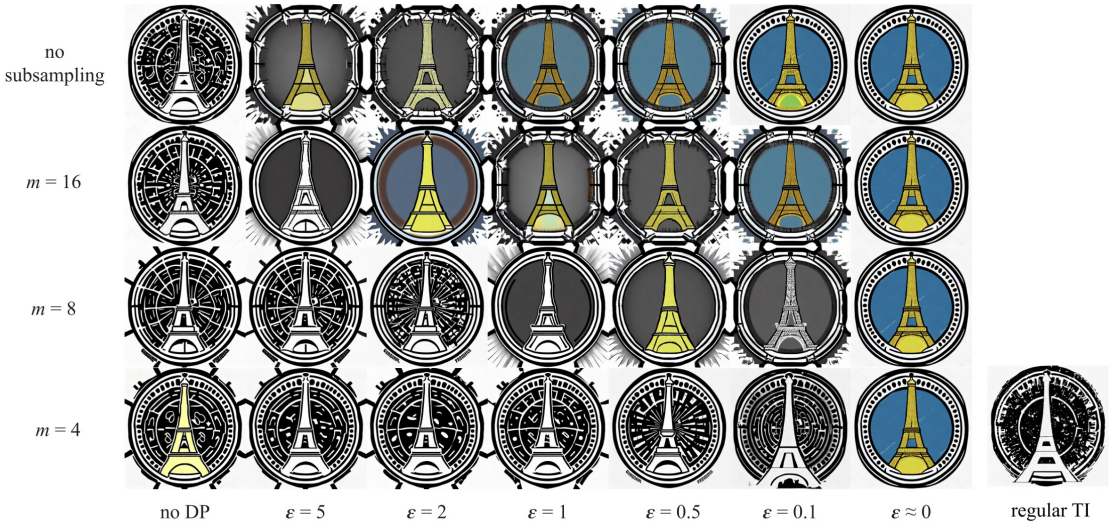

Figure 4: Images generated by Stable Diffusion v1.5 using the prompt "Icon of the Eiffel Tower in the style of <Paris 2024 Pictograms>", with the embedding <Paris 2024 Pictograms> trained using different values of m and ε.

pictograms captured the logo's original structure. In contrast, DP configurations introduce a discernible degradation in image quality, with lower epsilon values and smaller subsampling sizes resulting in more noticeable noise and diminished stylistic fidelity.

As $\varepsilon \to 0$, the resulting token embedding u^* gradually loses its semantic meaning, leading to a loss of stylistic fidelity. In particular, y_{u^*} tends towards y (a conditioning vector independent of the learnable embedding). In our results, this manifests as a painting of Taylor Swift devoid of the artist-specific stylistic elements, or a generic icon of the Eiffel Tower (with color, as opposed to the black and white design of the pictograms). With this in mind, ε can be interpreted as a drift parameter, representing the progression from the optimal u^* towards infinity, gradually steering the generated image away from the target style in exchange for stronger privacy guarantees.

Meanwhile, reducing m also reduces the sensitivity of the generated image to ε, as evident by the observation that, on both datasets at $m = 4$, (subsampling rate below 0.1) image generation can tolerate ε as low as 0.5 without significant changes in visual characteristics, and retaining stylistic elements of the target dataset at ε as low as 0.1. This strong boost in robustness comes at a small price of base style capture fidelity. As observed in figures 3 and 4, we can also treat subsampling as an introduction of noise. Mathematically, the subsample centroid is an unbiased estimate of the true centroid, and so the subsampling process itself defines a distribution centered at the true centroid. However, the amount of noise introduced by the subsampling process is limited by the individual image embeddings, as a subsample centroid can only stray from the true centroid as much as the most anomalous point in the dataset.

5 Conclusion

We presented a differentially private adaptation approach for diffusion models using embedding-based methods–Universal Guidance and Textual Inversion (TI)–to enable privacy-preserving style transfer from small, sensitive datasets. Through experiments on a private artwork dataset and the Paris 2024 Olympics pictograms, we demonstrated that our approach maintains high stylistic fidelity while providing strong privacy guarantees. Notably, TI proved more effective in capturing detailed stylistic elements compared to SG, especially with added noise and subsampling to protect privacy.

Our findings suggest that embedding-driven methods can serve as efficient and scalable alternatives to DP-SGD for privacy-preserving model adaptation, minimizing computational demands and offering resilience against privacy attacks. This approach provides a balance between style quality and privacy, enabling generative models to adapt responsibly to new domains.

6 Acknowledgments

We sincerely thank Tatchamon Wongworakul (`@eveismyname`) for graciously providing her artwork for use in this study.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[2] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[3] 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.

[4] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[5] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[6] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[7] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

[8] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[10] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022.

[11] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023.

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

[13] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pages 8717–8730. PMLR, 2023.

[14] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

[15] Nikhil Vyas, Sham M Kakade, and Boaz Barak. On provable copyright protection for generative models. In *International Conference on Machine Learning*, pages 35277–35299. PMLR, 2023.

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

[17] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[18] 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.

[19] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *Transactions on Machine Learning Research*, 2023.

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

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

[22] Jonathan Lebensold, Maziar Sanjabi, Pietro Astolfi, Adriana Romero-Soriano, Kamalika Chaudhuri, Mike Rabbat, and Chuan Guo. Dp-rdm: Adapting diffusion models to private domains without fine-tuning. *arXiv preprint arXiv:2403.14421*, 2024.

[23] Paris 2024. Paris 2024 - pictograms. `https://olympics.com/en/paris-2024/the-games/the-brand/pictograms`.

[24] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[26] 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.

[27] 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.

[28] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[29] Thomas Steinke. Composition of differential privacy & privacy amplification by subsampling. *arXiv preprint arXiv:2210.00597*, 2022.

[30] Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Position: Considerations for differentially private learning with large-scale public pretraining. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 48453–48467. PMLR, 21–27 Jul 2024.

[31] Yaodong Yu, Maziar Sanjabi, Yi Ma, Kamalika Chaudhuri, and Chuan Guo. Vip: A differentially private foundation model for computer vision. In *Forty-first International Conference on Machine Learning*, 2024.

[32] Tom Sander, Yaodong Yu, Maziar Sanjabi, Alain Oliviero Durmus, Yi Ma, Kamalika Chaudhuri, and Chuan Guo. Differentially private representation learning via image captioning. In *Forty-first International Conference on Machine Learning*, 2024.

[33] International Olympic Committee. Olympic properties. `https://olympics.com/ioc/olympic-properties`.