Differentially Private Adaptation of Diffusion Models via Noisy Aggregated Embeddings

Anonymous Author(s)
Affiliation
Address
email

Abstract

Personalizing large-scale diffusion models poses serious privacy risks, especially when adapting to small, sensitive datasets. A common approach is to fine-tune the model using differentially private stochastic gradient descent (DP-SGD), but this suffers from severe utility degradation due to the high noise needed for privacy, particularly in the small data regime. We propose an alternative that leverages Textual Inversion (TI), which learns an embedding vector for an image or set of images, to enable adaptation under differential privacy (DP) constraints. Our approach, Differentially Private Aggregation via Textual Inversion (DPAgg-TI), adds calibrated noise to the aggregation of per-image embeddings to ensure formal DP guarantees while preserving high output fidelity. We show that DPAgg-TI outperforms DP-SGD finetuning in both utility and robustness under the same privacy budget, achieving results closely matching the non-private baseline on style adaptation tasks using private artwork from a single artist and Paris 2024 Olympic pictograms. In contrast, DP-SGD fails to generate meaningful outputs in this setting.

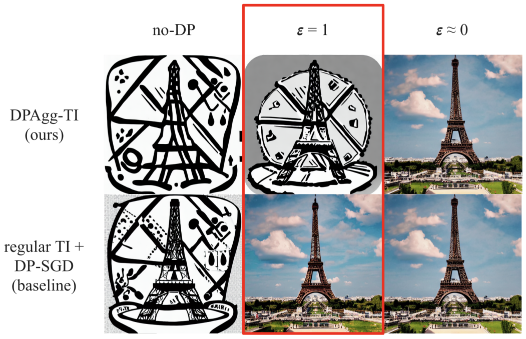

Figure 1: We compare our method (DPAgg-TI, top) to a baseline applying DP-SGD to Textual Inversion (bottom), using the prompt "an icon of the Eiffel Tower in the style of the Paris 2024 Olympic Pictograms." While the baseline learns a single embedding over the dataset, our method privately aggregates per-image embeddings. At privacy budget $\varepsilon = 1$, DPAgg-TI preserves visual fidelity much better than the baseline, and closely matches the non-private output (left), demonstrating a superior privacy-utility tradeoff.

1 Introduction

The rapid adoption of diffusion models Ho et al. [2020], Song et al. [2021b], Rombach et al. [2022] has raised significant privacy and legal concerns. These models are vulnerable to privacy attacks, such as membership inference Duan et al. [2023], where attackers determine if a specific data point was used for training, and data extraction Carlini et al. [2023], which enables reconstruction of training data. This risk is amplified during fine-tuning on smaller, domain-specific datasets, where each record has a greater impact. Additionally, reliance on large datasets scraped without consent raises copyright concerns Vyas et al. [2023], as diffusion models can reproduce original artworks without credit or compensation. These issues highlight the urgent need for privacy-preserving technologies and clearer ethical and legal guidelines for generative models.

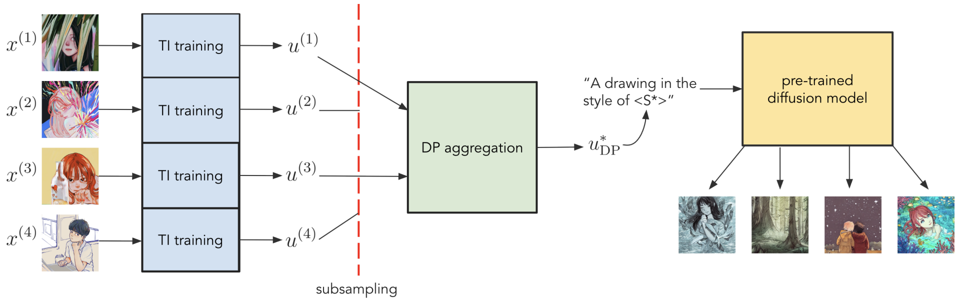

Figure 2: Overview of DPAgg-TI. We first apply Textual Inversion to extract embeddings for each image in the private dataset. These embeddings are then aggregated with differentially private mechansim, incorporating subsampling to produce a private embedding u^*_{DP}. Finally, images are generated using the corresponding token $<S^*>$.

Differential privacy (DP) Dwork [2006] is a widely adopted framework for addressing these challenges. One standard approach for ensuring DP in deep learning is Differentially Private Stochastic Gradient Descent (DP-SGD) Abadi et al. [2016], which modifies traditional SGD by adding noise to clipped gradients. However, applying DP-SGD to train diffusion models poses several challenges. It introduces significant computational and memory overhead due to per-sample gradient clipping Hoory et al. [2021], which is essential for bounding gradient sensitivity Dwork et al. [2006], Abadi et al. [2016]. DP-SGD is also incompatible with batch-wise operations like batch normalization, as these link samples and hinder sensitivity analysis. Furthermore, training large models with DP-SGD often leads to substantial performance degradation, particularly under realistic privacy budgets since the required noise scales with the gradient norm. Consequently, existing diffusion models trained with DP-SGD are limited to relatively small-scale images.

Independent of privacy concerns, Textual Inversion (TI) Gal et al. [2023] provides an effective method for adapting diffusion models to specific styles or content without modifying the model. Instead, TI learns an external embedding vector that captures the style or content of a target image set, which is then incorporated into text prompts to guide the model's outputs. A key advantage of TI is its ability to compress a style into a compact vector, reducing computational and memory demands while simplifying the application of privacy-preserving mechanisms, as privacy constraints can be applied directly to embeddings rather than the full model. Additionally, since TI avoids direct model optimization, it remains efficient and compatible with DP constraints on smaller datasets.

In this work, we propose a novel privacy-preserving adaptation method for smaller datasets, leveraging TI to avoid the extensive model updates required by DP-SGD. Standard TI does not offer formal privacy guarantees, so to address this limitation, we introduce a private variant of TI, called Differentially Private Aggregation via Textual Inversion (DPAgg-TI) and summarize it in Figure 2. Our method decouples interactions among samples by learning a separate embedding for each target image, which are then aggregated into a noisy centroid. This approach ensures efficient and secure adaptation to private datasets.

Our experiments demonstrate the effectiveness of DPAgg-TI, showing that TI remains robust in preserving stylistic fidelity even under privacy constraints. Applying our method to a private artwork collection by `@eveismyname` and Paris 2024 Olympics pictograms Paris 2024, we show that DPAgg-TI captures nuanced stylistic elements while ensuring privacy. We observe a trade-off between privacy (controlled by DP parameter ε) and image quality: lower ε reduces fidelity but maintains the target style under moderate noise. Subsampling further amplifies privacy by reducing sensitivity to individual data points, mitigating noise impact on image quality. This framework enables privacy-preserving adaptation of diffusion models to new styles and domains while protecting sensitive data.

Our contributions can be summarized as follows:

(1) We propose DPAgg-TI that ensures privacy by learning separate embeddings for individual images and aggregating them into a noisy centroid.

(2) Our approach enables style adaptation without extensive model updates, reducing computational overhead while preserving privacy.

(3) We analyze the trade-off between privacy and image quality, showing that moderate noise maintains stylistic fidelity while protecting sensitive data.

(4) We validate our method on diverse datasets, demonstrating its effectiveness in capturing stylistic elements under privacy constraints.

2 Background and Related Work

2.1 Diffusion Models

Diffusion models Ho et al. [2020], Song et al. [2021b,a], Rombach et al. [2022] leverage an iterative denoising process to generate high-quality images that align 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 broken down 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 Radford et al. [2021], outputting a single text-conditional vector y that serves as the conditioning input. This vector y is then incorporated at each denoising step, guiding the model to align 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.

The objective function for a text-conditioned diffusion model, given both the noisy image x_t and the text conditioning y, 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:

$$\theta^* = \arg\min_\theta \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,I),t\sim[T]}[\|\epsilon - \epsilon_\theta(x_t, y, t)\|^2]. \tag{1}$$

2.2 Textual Inversion.

Textual Inversion (TI) Gal et al. [2023] is an adaptation technique that enables personalization using a small dataset of typically 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, representing a placeholder token, denoted as S. During training, images are conditioned on 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 given in (1) 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:

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

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.

2.3 Differential Privacy.

In this work, we adopt differential privacy (DP) Dwork et al. [2006], Dwork [2006] 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 with respect to 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 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 Dwork et al. [2014], Balle and Wang [2018], 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] Balle and Wang [2018], 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 Abadi et al. [2016], which applies Gaussian noise during model updates to achieve DP.

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 $m \ll n$) before the privatization mechanism (such as the Gaussian mechanism) is 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

$$\varepsilon' = \log\left(1 + \frac{m}{n}\left(e^{\varepsilon} - 1\right)\right) = O\left(\frac{m}{n}\varepsilon\right). \tag{3}$$

This result is well-known (Steinke [2022, Theorem 29]), with tighter amplification bounds available for Gaussian mechanisms Mironov [2017].

2.4 Private Adaptation of Diffusion Models

Recent advancements in applying DP to diffusion models have aimed to balance privacy preservation with the high utility of generative outputs. Dockhorn et al. Dockhorn et al. [2023] proposed a Differentially Private Diffusion Model (DPDM) that enables privacy-preserving generation of realistic samples, setting a foundational approach for adapting diffusion processes using DP-SGD. Another common strategy involves training a model on a large public dataset, followed by differentially private fine-tuning on a private dataset, as explored by Ghalebikesabi et al. [2023]. While effective in certain contexts, this approach raises privacy concerns, particularly around risks of information leakage during the fine-tuning phase Tramèr et al. [2024].

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 fine-tuning, as in the VIP approach Yu et al. [2024], which demonstrates the feasibility of applying DP in later adaptation stages. Other approaches explore differentially private learning of feature representations Sander et al. [2024], 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 Textual Inversion

Let $x^{(1)}, \ldots, x^{(n)}$ represent a target dataset of images whose characteristics we wish to privately adapt our image generation towards. Instead of training a single token embedding on the entire

[1] In practice, we use numerical privacy accountant such as Balle and Wang [2018], Mironov [2017] to calibrate the noise.

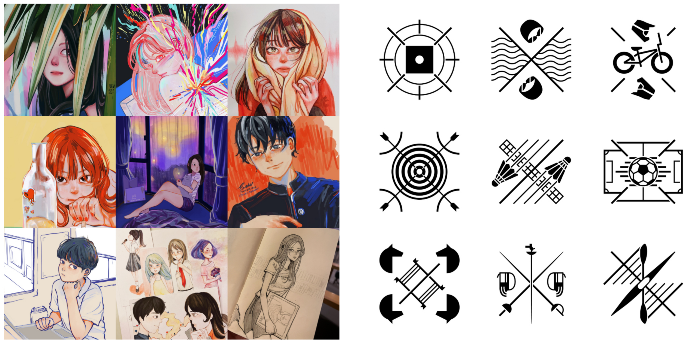

Figure 3: Samples of images used in our style adaptation experiments. **Left:** artwork by `@eveismyname` ($n = 158$). **Right:** Paris 2024 Olympic pictograms ($n = 47$), © *International Olympic Committee, 2023*.

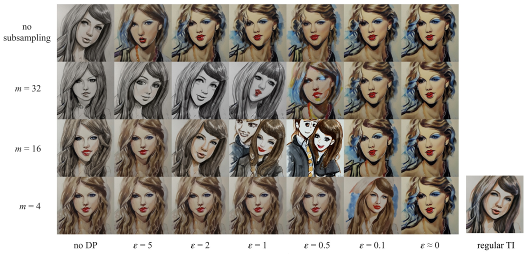

Figure 4: Images generated by Stable Diffusion v1.5 using the prompt "A painting of Taylor Swift in the style of <`@eveismyname`>", with the embedding <`@eveismyname`> trained using different values of m and ε.

dataset as in regular TI, we train a separate embedding $u^{(i)}$ on each $x^{(i)}$ to obtain a set of embeddings $u^{(1)}, \dots, u^{(n)}$, as illustrated in Figure 2. We can formalize the encoding process as follows:

$$u^{(i)} = \arg\min_{u} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I),t}[\|\epsilon - \epsilon_\theta(x_t^{(i)}, y_u, t)\|^2]. \tag{4}$$

Then, we can aggregate the embeddings $u^{(1)}, \dots, u^{(n)}$ 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 guarantees, we also add isotropic Gaussian noise to the centroid. We can therefore define the resulting embedding vector u^*_{DP} as follows:

$$u^*_{\text{DP}} = \frac{1}{n}\sum_{i=1}^{n} u^{(i)} + \mathcal{N}(0, \sigma^2 I), \tag{5}$$

where the minimum σ required to provide (ε, δ)-DP is given by the following expression based on Balle and Wang [2018, Theorem 1]:

$$\sigma = \frac{\Delta}{n} \cdot \frac{\sqrt{2\ln(1.25/\delta)}}{\varepsilon}. \tag{6}$$

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^*_{DP} can then be used to adapt the downstream image generation process. Similar to regular TI's u^*, we can use u^*_{DP} to represent a new placeholder token S^* that can be incorporated into prompts for personalized image generation. While u^*_{DP} may not fully solve the TI optimization problem presented in (2), it provides provable privacy guarantees, with only a minimal trade-off in accurately representing the style of the target dataset.

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 (3)) can be applied. Formally,

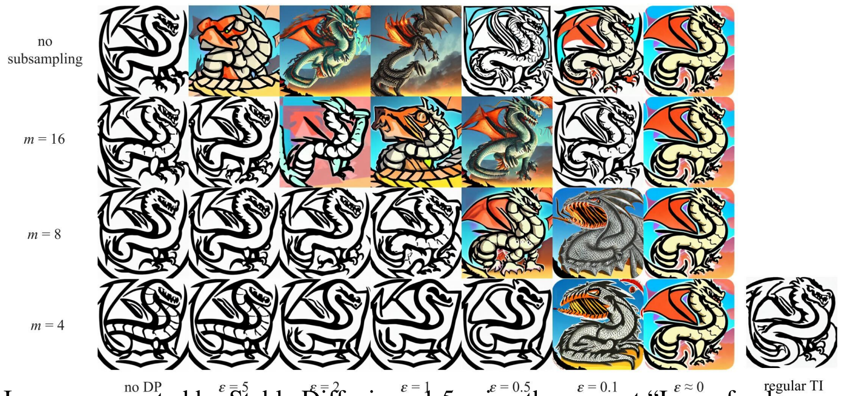

Figure 5: Images generated by Stable Diffusion v1.5 using the prompt "Icon of a dragon in the style of <Paris 2024 Pictograms>", with the embedding <Paris 2024 Pictograms> trained using different values of m and ε.

we sample $D_{\text{sub}} \subseteq \{u^{(1)}, \ldots, u^{(n)}\}$ where $|D_{\text{sub}}| = m$, and compute the output embedding as follows:

$$u^*_{\text{DP}} = \frac{1}{m} \sum_{u^{(i)} \in D_{\text{sub}}} u^{(i)} + \mathcal{N}(0, \sigma^2 I), \tag{7}$$

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

4 Experimental Results

4.1 Datasets

We compiled two datasets to evaluate our style adaptation method, specifically selecting content unlikely to be recognized by Stable Diffusion v1.5, our base model.

The first dataset consists of 158 artworks by the artist `@eveismyname`, who has granted consent for non-commercial use. This dataset allows us to assess whether models can capture artistic styles without memorizing individual works. While some of these artworks may have been publicly accessible on social media, making incidental inclusion in Stable Diffusion's pretraining possible, the artist's limited recognition and relatively small portfolio reduce the likelihood that the model has internalized her unique style. This dataset serves as a controlled test for privacy-preserving style transfer on individual artistic collections.

The second dataset contains 47 pictograms from the Paris 2024 Olympics Paris 2024, permitted strictly for non-commercial editorial use International Olympic Committee. These pictograms were officially released in February 2023, several months after the release of Stable Diffusion v1.5, ensuring they were absent from the model's pretraining data. This dataset allows us to assess how well our approach adapts to newly introduced visual styles that the base model has never encountered.

Both datasets are used to test the ability of our method to extract and transfer stylistic elements while preserving privacy. Representative samples are shown in Figure 3.

4.2 Style Transfer Results

Using both the `@eveismyname` and Paris 2024 pictograms dataset, we trained TI Gal et al. [2023] embeddings on Stable Diffusion v1.5 Rombach et al. [2022] using DPAgg-TI. Our primary goal is to investigate how DP configurations, specifically the privacy budget ε and subsampling size m, affect the generated images quality and privacy resilience. For regular TI, we utilize the default process to embed the private dataset without any additional noise. For the DPAgg-TI, we test multiple configurations of m and ε to analyze the trade-off between image fidelity and privacy.

Figures 4 and 5 present generated images across two key configurations: (1) regular TI without DP, (2) DPAgg-TI with DP at different values of m and ε. We used the same random seed to generate embeddings, subsample images, and sample DP noise for ease of visual comparison 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 $\varepsilon = 10^{-5}$. The purpose of this parameter value is to demonstrate the image generated when u^*_{DP} contains zero information about the target dataset. 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 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 diminished stylistic fidelity.

As $\varepsilon \to 0$, the resulting token embedding u^*_{DP} gradually loses its semantic meaning, leading to a loss of stylistic fidelity. In particular, $y_{u^*_{\text{DP}}}$ 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 a dragon (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^*_{DP} towards infinity, gradually steering the generated image away from the target style in exchange for stronger privacy guarantees. We also observe instances where there is a temporary drop in prompt fidelity (e.g., $m = 16, \varepsilon \in [0.5, 1]$ in Figure 4 and intermediate ε values in Figure 5) which restores as u^*_{DP} drifts even further from its optimal value. We hypothesize that this is due to drifted u^*_{DP} capturing a different meaning unrelated to the prompt, before losing any meaning that could be interpreted by Stable Diffusion's text encoder, causing u^*_{DP} to be disregarded from $y_{u^*_{\text{DP}}}$ and the prompt fidelity to be restored. Another possible explanation is that the temporary drop in prompt fidelity is due to the drift path of u^*_{DP} passing through non-linear regions within embedding space. We leave further investigations into this observation for future work.

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 4 and 5, 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 biggest outlier in the dataset.

4.3 Quantitative Evaluation

User Study To evaluate the utility of our approach under different DP and subsampling configurations, we conducted a user study with 25 participants. Each participant was shown reference images from the target dataset and asked to compare pairs of generated images, selecting the one that better captured the style of the reference images. Images were generated using 10 prompts and adapted TI embeddings for the `@eveismyname` and Paris 2024 Pictogram datasets, resulting in 20 groups of images. Each participant evaluated two groups, one randomly selected from each dataset, with comparisons focusing on model configurations differing by DP noise and subsampling size.

Survey results, summarized in Table 1 in Appendix A, align with our design goals. Participants showed no clear preference between regular TI and DPAgg-TI, suggesting that our privacy-preserving approach maintains perceptual quality. As expected, both DP noise and reduced subsampling size degraded style fidelity, consistent with the trade-offs inherent in differential privacy. Preferences at $\varepsilon = 1$ were split, but subsampling was generally favored, reinforcing its role in reducing noise impact while preserving style.

Kernel Inception Distance The Kernel Inception Distance (KID) Bińkowski et al. [2018] is a metric for evaluating generative models by measuring the difference between the distributions of generated and training images in an embedding space. To compute KID, images generated by the model and real training images are passed through an Inception network Szegedy et al. [2015], and their distributional differences are estimated. Unlike the more commonly used Fréchet Inception Distance (FID) Heusel et al. [2017], KID is an unbiased estimator of the true divergence between the learned and target distributions Jayasumana et al. [2024], making it more suitable for smaller datasets, as in our case.

We report KID scores for different parameters in Tables 2 and 3 (see Appendix B), showing that DPAgg-TI maintains the style transfer fidelity of TI while ensuring differential privacy. Further discussion of these results is also provided in Appendix B.

Ablation Study: Textual Inversion with DP-SGD A natural question that arises is how well our approach compares to the naive method of applying DP-SGD to regular TI training. We therefore integrated DP-SGD into the TI codebase using the Opacus library and trained similar embeddings on the `@eveismyname` and Paris 2024 datasets. We found that in most cases, notably the `@eveismyname` dataset, the amount of noise required for DP-SGD to achieve a reasonable value of ε for DP is so high that the resulting embedding contains negligible information about the training dataset. In particular, the results for $\varepsilon = 1$ are almost indistinguishable to $\varepsilon \approx 0$, as shown in Figure 6. We believe that this is simply because DP-SGD is not designed to handle such small datasets in the order of 100 images. Additional results can be found in Appendix F.

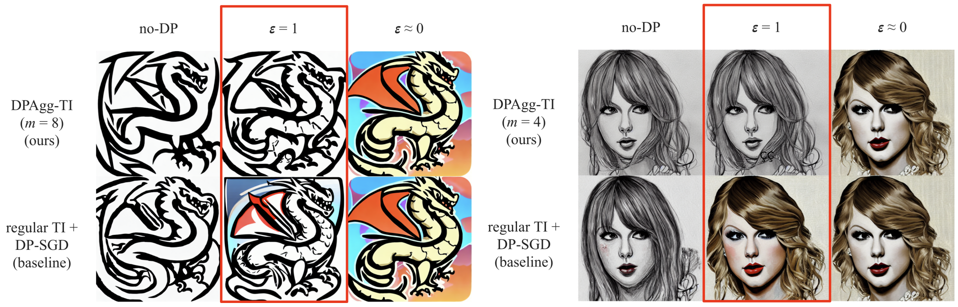

Figure 6: Comparing our approach to applying DP-SGD to regular TI using prompts “an icon of a dragon in the style of the Paris 2024 Olympic Pictograms” and “a painting of Taylor Swift in the style of `@eveismyname`” respectively. Note that our method aggregates individual TI embeddings for each training image, whereas the baseline trains a single TI embedding over the entire dataset.

5 Copyright Protection Implications

Our proposed mechanism can also be interpreted through the lens of *copyright protection*. This connection is grounded in the framework of *Near Access-Freeness (NAF)* [Vyas et al., 2023], which evaluates whether a model’s outputs reveal undue influence from specific data points by comparing them to those from a safe model trained without access to the same data. Since DPAgg-TI satisfies ε-DP, it also satisfies ε-NAF, which means the adapted model behaves similarly to one that never saw the private images, under the NAF criterion. However, we emphasize that this guarantee holds only within the NAF framework; it does not constitute a general claim about content similarity or legal compliance. Crucially, DPAgg-TI is designed to adapt to the *style* of private images, not their specific content. Prior work and legal precedent suggest that style imitation is generally considered *fair use* and does not constitute infringement [Vyas et al., 2023]. Thus, our mechanism aligns with the intended protections of NAF: it avoids memorization while still enabling meaningful personalization and stylistic adaptation. We defer the details of copyright protection to Appendix D.

6 Conclusion

We presented a differentially private adaptation method for diffusion models based on Textual Inversion, enabling privacy-preserving style transfer without the need for full model fine-tuning. By learning per-image embeddings and aggregating them with calibrated noise, our method, DPAgg-TI, achieves strong formal privacy guarantees while maintaining high output fidelity. Experiments on private artwork and Paris 2024 pictograms show that DPAgg-TI consistently outperforms DP-SGD, which fails to produce meaningful results under comparable privacy budgets. These results highlight the effectiveness of embedding-level adaptation as an efficient and scalable alternative to traditional gradient-based approaches, especially in low-data regimes. Unlike DP-SGD, which introduces significant computational overhead and utility degradation, DPAgg-TI is lightweight, modular, and compatible with existing diffusion backbones. Our findings suggest that embedding-centric approaches offer a promising direction for privacy-aware personalization, and motivate further research into cross-modal extensions, improved aggregation techniques, and integration with broader privacy-preserving frameworks.

References

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *ACM SIGSAC*, 2016.

B. Balle and Y.-X. Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *ICML*, 2018.

A. Bansal, H.-M. Chu, A. Schwarzschild, S. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein. Universal guidance for diffusion models. In *ICLR*, 2024.

M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans, 2018.

N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. In *USENIX Security*, 2023.

T. Dockhorn, T. Cao, A. Vahdat, and K. Kreis. Differentially private diffusion models. *TMLR*, 2023.

J. Duan, F. Kong, S. Wang, X. Shi, and K. Xu. Are diffusion models vulnerable to membership inference attacks? In *ICML*, 2023.

C. Dwork. Differential privacy. In *ICALP*, 2006.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

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

N. Elkin-Koren, U. Hacohen, R. Livni, and S. Moran. Can copyright be reduced to privacy? *arXiv preprint arXiv:2305.14822*, 2023.

R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023.

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

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

S. Hoory, A. Feder, A. Tendler, S. Erell, A. Peled-Cohen, I. Laish, H. Nakhost, U. Stemmer, A. Benjamini, A. Hassidim, et al. Learning and evaluating a differentially private pre-trained language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1178–1189, 2021.

Innat. Van gogh paintings. https://www.kaggle.com/datasets/ipythonx/van-gogh-paintings.

International Olympic Committee. Olympic properties. https://olympics.com/ioc/olympic-properties.

S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 9307–9315. IEEE, June 2024. doi: 10.1109/cvpr52733.2024.00889. URL http://dx.doi.org/10.1109/CVPR52733.2024.00889.

I. Mironov. Rényi differential privacy. In *CSF*, 2017.

Paris 2024. Paris 2024 - pictograms. https://olympics.com/en/paris-2024/the-games/the-brand/pictograms.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

T. Sander, Y. Yu, M. Sanjabi, A. O. Durmus, Y. Ma, K. Chaudhuri, and C. Guo. Differentially private representation learning via image captioning. In *ICML*, 2024.

J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021b.

T. Steinke. Composition of differential privacy & privacy amplification by subsampling. *arXiv preprint arXiv:2210.00597*, 2022.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision, 2015. URL https://arxiv.org/abs/1512.00567.

F. Tramèr, G. Kamath, and N. Carlini. Position: Considerations for differentially private learning with large-scale public pretraining. In *ICML*, 2024.

N. Vyas, S. M. Kakade, and B. Barak. On provable copyright protection for generative models. In *ICML*, 2023.

Y. Yu, M. Sanjabi, Y. Ma, K. Chaudhuri, and C. Guo. Vip: A differentially private foundation model for computer vision. In *ICML*, 2024.

A User Study

	regular TI	No Adaptation	Unsure
@eveismyname	19	4	2
Paris 2024	16	6	3

	DPAgg-TI (no DP, no subsampling)	No Adaptation	Unsure
@eveismyname	16	9	0
Paris 2024	15	4	6

	regular TI	DPAgg-TI (no DP, no subsamp.)	Unsure
@eveismyname	12	13	0
Paris 2024	9	10	6

	regular TI	DPAgg-TI (no DP, subsamp. $m = 8$)	Unsure
@eveismyname	16	6	3
Paris 2024	7	13	5

	DPAgg-TI (no DP, no subsampling)	DPAgg-TI (no DP, subsamp. $m = 8$)	Unsure
@eveismyname	18	4	3
Paris 2024	10	8	7

	DPAgg-TI ($\varepsilon = 1$) no subsampling	DPAgg-TI ($\varepsilon = 1$, subsamp. $m = 8$)	Unsure
@eveismyname	14	10	1
Paris 2024	3	16	6

	DPAgg-TI (no DP, no subsampling)	Style Guidance	Unsure
@eveismyname	16	8	1
Paris 2024	20	2	3

	DPAgg-TI ($\varepsilon = 1$, subsamp. $m = 8$)	Style Guidance	Unsure
@eveismyname	16	8	1
Paris 2024	19	2	4

	DPAgg-TI (no DP, subsamp. $m = 8$)	DPAgg-TI ($\varepsilon = 1$, subsamp. $m = 8$)	Unsure
@eveismyname	8	5	12
Paris 2024	15	4	6

Table 1: Survey Results.

A.1 Study Design and Objective

The user study aimed to assess the utility of our approach under different DP and subsampling configurations by evaluating the models' ability to adapt to novel styles. The study involved 25 participants, each of whom was tasked with comparing images generated using various configurations and selecting the one that better captured the style of reference images.

A.2 Experimental Setup

Participants were shown reference images from two datasets:

- The @eveismyname dataset of private artwork.

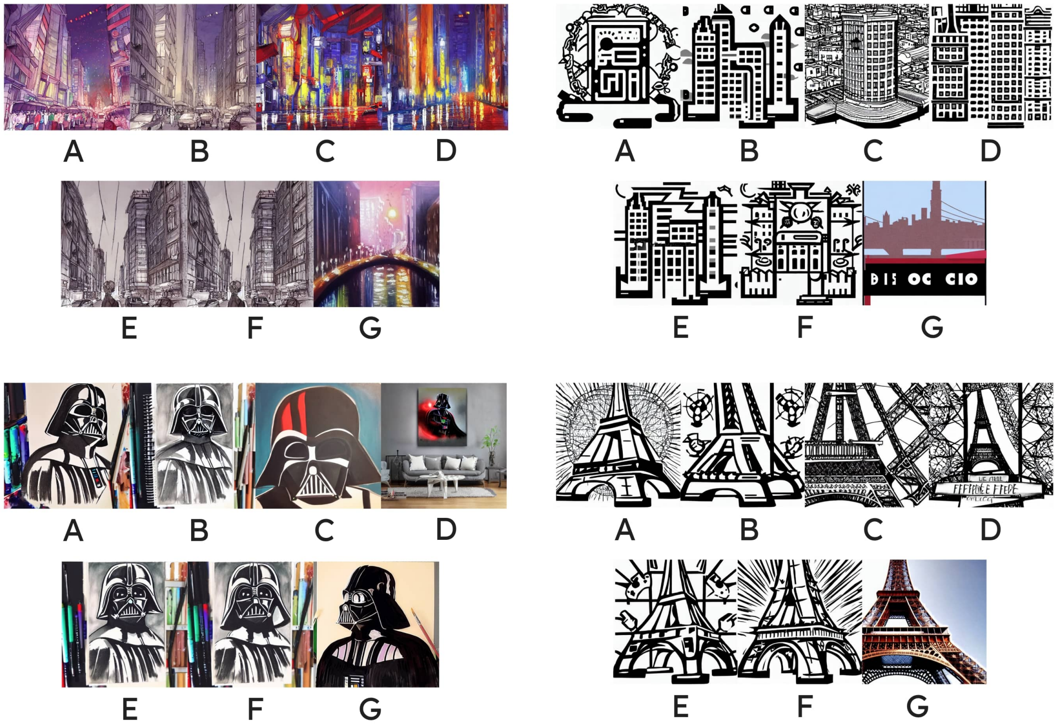

Figure 7: Samples of image sets used in our user study. Participants are asked to compare 2 images at a time.

- The Paris 2024 Pictogram dataset.

For each dataset, 10 prompts were used to generate images, resulting in 20 groups of images (10 prompts per dataset). Each group included images generated using the same prompt and dataset but with different model configurations. Configurations varied in the addition of DP noise and the size of subsampling.

- Original Textual Inversion (TI)
- DPAgg-TI ($\varepsilon = \infty$, no DP) w/o subsampling
- DPAgg-TI ($\varepsilon = 1$) without subsampling
- No Adaptation
- DPAgg-TI ($\varepsilon = \infty$, no DP) with subsampling $(m = 8)$
- DPAgg-TI ($\varepsilon = 1$) with subsampling $(m = 8)$
- Style Guidance (SG)

A.3 Survey Procedure

Participants were asked to evaluate two groups of images: one randomly selected from the `@eveismyname` dataset and one from the Paris 2024 Pictogram dataset. For each group:

1. Participants were shown reference images from the target dataset.
2. They were presented with pairs of images generated using different model configurations for the same prompt.
3. Participants selected the image they felt better captured the style of the reference images.

A.4 Evaluation Metrics

The study focused on assessing:

- Participants' preference between regular TI and DPAgg-TI for style adaptation.
- The impact of DP noise and subsampling size on the perceived utility of style transfer.

A.5 Results and Analysis

The results are summarized in Table 1. Key observations include:

- Participants showed no clear preference between regular TI and DPAgg-TI in capturing styles for either dataset.
- Both DP noise and reduced subsampling size decreased the perceived quality of style transfer.
- Preferences were split between configurations with $\varepsilon = 1$ with and without subsampling, though subsampling generally had favorable outcomes.

These findings highlight the trade-off between increased DP robustness and reduced utility, suggesting that the optimal configuration may depend on subjective preferences and specific application requirements.

B Kernel Inception Distance

m	No DP	$\varepsilon = 5.0$	$\varepsilon = 1.0$	$\varepsilon = 0.5$	$\varepsilon = 0.1$	$\varepsilon \approx 0$
–	0.0441 ± 0.0027	0.0798 ± 0.0032	0.0526 ± 0.0022	0.0688 ± 0.0020	0.1114 ± 0.0032	0.0654 ± 0.0027
32	0.0753 ± 0.0047	0.0836 ± 0.0042	0.1166 ± 0.0037	0.0295 ± 0.0019	0.0644 ± 0.0021	0.0650 ± 0.0025
16	0.0350 ± 0.0020	0.0381 ± 0.0018	0.0663 ± 0.0025	0.1303 ± 0.0033	0.0438 ± 0.0030	0.0660 ± 0.0029
8	0.0359 ± 0.0018	0.0364 ± 0.0017	0.0366 ± 0.0019	0.0394 ± 0.0025	0.0527 ± 0.0033	0.0654 ± 0.0024
4	0.0246 ± 0.0013	0.0251 ± 0.0016	0.0249 ± 0.0014	0.0256 ± 0.0012	0.0313 ± 0.0017	0.0653 ± 0.0023
`ctrl`	0.0314 ± 0.0010	–	–	–	–	–

Table 2: KID scores of DPAgg-TI on `@eveismyname` dataset for various ε values ranging from $\varepsilon = 10^{-5}, 0.1, 0.5, 1.0, 5.0$ (including no DP) under different subsampling levels ($m = 4, 8, 16, 32$) as well as regular TI (`ctrl`). Reported values are the mean ± standard deviation over 100 random subsamples.

m	No DP	$\varepsilon = 5.0$	$\varepsilon = 1.0$	$\varepsilon = 0.5$	$\varepsilon = 0.1$	$\varepsilon \approx 0$
–	0.1153 ± 0.0055	0.1194 ± 0.0054	0.1306 ± 0.0046	0.1395 ± 0.0057	0.1201 ± 0.0053	0.1274 ± 0.0055
32	0.1222 ± 0.0066	0.1036 ± 0.0065	0.1375 ± 0.0047	0.1311 ± 0.0048	0.1248 ± 0.0060	0.1258 ± 0.0054
16	0.1321 ± 0.0057	0.1411 ± 0.0077	0.1309 ± 0.0061	0.1380 ± 0.0047	0.1359 ± 0.0060	0.1273 ± 0.0057
8	0.1303 ± 0.0084	0.1303 ± 0.0074	0.1112 ± 0.0062	0.1311 ± 0.0064	0.1318 ± 0.0052	0.1267 ± 0.0056
4	0.1158 ± 0.0057	0.1085 ± 0.0056	0.1184 ± 0.0068	0.1194 ± 0.0065	0.1592 ± 0.0065	0.1268 ± 0.0055
`ctrl`	0.1383 ± 0.0066	–	–	–	–	–

Table 3: KID scores of DPAgg-TI on Paris dataset for various ε values ranging from $\varepsilon = 1e - 5, 0.1, 0.5, 1.0, 5.0$ (including no DP) under different subsampling levels ($m = 4, 8, 16, 32$) as well as regular TI (`ctrl`). Reported values are the mean ± standard deviation over 100 random subsamples.

Our results indicate that DPAgg-TI preserves the style transfer fidelity of TI while also ensuring differential privacy. Notably, for `@eveismyname` ($m = 4$) at low privacy budgets, we observe even lower KID values than standard TI, suggesting enhanced style alignment. Similarly, results for the Paris 2024 dataset follow a comparable trend, with DPAgg-TI achieving KID scores similar to TI at low privacy budgets. However, the overall KID scores for this dataset remain high within the context of diffusion model style transfer.

Upon inspecting the generated images (Figure 8), we hypothesize that the abstract and out-of-distribution nature of the Paris 2024 images poses a challenge for the Inception network, leading to less meaningful feature embeddings. This likely inflates the measured embedding distances between generated and reference images, resulting in higher-than-expected KID values.

For KID evaluations, we used prompts similar to those employed during TI training: “A painting/icon in the style of S^*”. Consistent with the training image captions, these prompts do not specify a subject. For each parameter configuration, we generate 100 images and compute KID by repeatedly subsampling the larger of the real and generated sets to match the size of the smaller set, 100 times, then averaging the resulting KID scores.

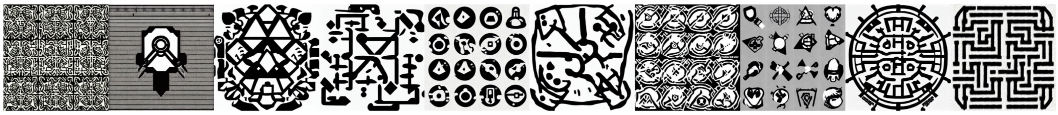

Figure 8: Sample of generated images for KID evaluations with respect to the Paris 2024 dataset.

C Differentially Private Adaptation via Style Guidance

C.1 Background: Denoising Diffusion Implicit Models

Denoising Diffusion Implicit Models (DDIM) sampling Song et al. [2021a] 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:

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, y, t)}{\sqrt{\alpha_t}} \tag{8}$$

$$x_{t-1} = \sqrt{\alpha_{t-1}}\hat{x}_0 + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_t, y, t). \tag{9}$$

C.2 Implementation

We extend our approach to style guidance (SG) by leveraging the framework of Universal Guidance Bansal et al. [2024]. Specifically, we focus on CLIP-based style guidance, which optimizes the similarity between the CLIP embeddings of a target image and the generated image.

We encode each target image $x^{(i)}$ as $u^{(i)}$ via a CLIP image encoder, then aggregate the embeddings $u^{(1)}, \ldots, u^{(n)}$ into u^*_{DP} using (5) or (7), depending on whether subsampling is applied. The aggregated embedding u^*_{DP} is then incorporated into the reverse diffusion process as a style guide.

Let x_c denote the target style image, x_t the noisy image at step t, and $\mathcal{E}(\cdot)$ the CLIP image encoder. The forward guidance process is defined as follows:

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

where w is a guidance weight and $\ell_{\cos}$ is the negative cosine similarity loss. For a detailed description of Universal Guidance, including the backward guidance process and per-step self-recurrence, we refer the reader to the original paper. The reverse diffusion step replaces $\epsilon_\theta(x_t, y, t)$ with $\hat{\epsilon}_\theta(x_t, y, t)$, generating an image x_0 that aligns with the text conditioning y while incorporating the stylistic characteristics of x_c.

To integrate differential privacy, we encode each target image $x^{(i)}$ into $u^{(i)} = \mathcal{E}(x^{(i)})$ and aggregate these embeddings into u^*_{DP} using the centroid method. The aggregated u^*_{DP} guides the reverse diffusion process:

$$\hat{\epsilon}_\theta(x_t, y, t) = \epsilon_\theta(x_t, y, t) + w\sqrt{1-\alpha_t}\nabla_{x_t}\ell_{\cos}(u^*_{\text{DP}}, \mathcal{E}(\hat{x}_0)). \tag{11}$$

This ensures privacy-preserving style transfer while maintaining high stylistic fidelity.

C.3 Style Transfer Results

We apply our SG-based approach to both datasets. While it provides privacy protection by obfuscating embedding details, the resulting images captured only generalized stylistic elements and lack the detailed fidelity and coherence achieved with the TI-based method. As shown in Figure 9, this highlights the superiority of TI in balancing privacy and high-quality image generation.

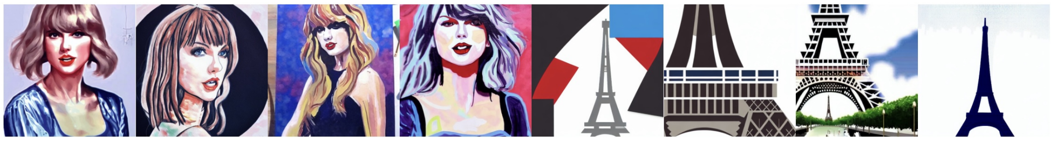

Figure 9: Attempts of using universal guidance to generate drawings of Taylor Swift and icons of the Eiffel Tower in the styles of `@eveismyname` and Paris 2024 Pictograms respectively. Here, we apply no subsampling or DP-noise.

The reduced effectiveness of SG for style transfer may stem from its sensitivity to hyperparameters such as the guidance weight w, leading to instability. Although Bansal et al. [2024] proposed remedies, namely backward guidance and per-step self-recurrence, these proved insufficient for our application. Additionally, the CLIP embeddings may not retain enough stylistic detail after the aggregation.

C.4 Ablation

To better understand the limited effectiveness of style guidance in our experiments, despite its success in Bansal et al. [2024], we applied our approach to a dataset of 143 paintings from Van Gogh's Saint-Paul Asylum, Saint-Rémy collectionInnat (Figure 10). Unlike the `@eveismyname` and Paris 2024 datasets, it is highly likely that Stable Diffusion has been trained on these images. Additionally, Bansal et al. [2024] demonstrated successful adaptation towards the style of Van Gogh's Starry Night as a single reference image, making this dataset a reasonable interpolation between their successful results and our more limited findings.

Without DP noise or subsampling, we obtained reasonable style transfer results, as shown in Figure 11. This suggests that style guidance struggles when applied to previously unseen target styles, and that its effectiveness may depend on prior exposure within the pre-training data.

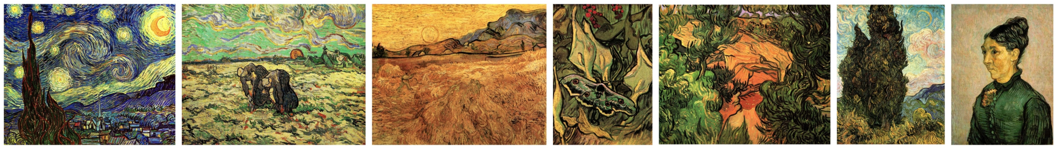

Figure 10: Sample of paintings by Van Gogh used to generate style guidance embeddings.

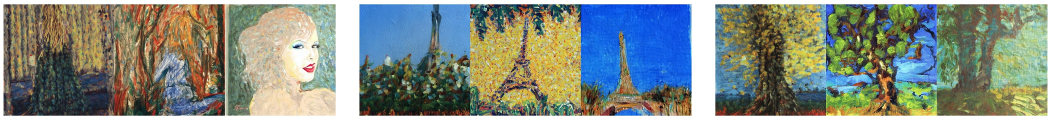

Figure 11: Images generated by Stable Diffusion v1.5 with style guidance towards Van Gogh's *Saint-Paul Asylum, Saint-Rémy* collection using prompts "A painting of Taylor Swift (left) / the Eiffel Tower (center) / a tree (right)".

D Copyright Protection

Modern generative models typically produce outputs via randomized sampling. Leveraging this inherent randomness, Vyas et al. [2023] introduced *Near Access-Freeness* (NAF) as a metric to quantify the similarity between a model's output and copyrighted content. The key idea is to compare the output distribution of a potentially infringing model to that of a *safe* model – one trained without access to the target content.

Formally, let safe be a mapping from a data point $x \in \mathcal{C}$ (where $\mathcal{C}$ is the collection of copyrighted samples) to a generative model $\mathsf{safe}(x) \in \mathcal{W}$ that is trained without using x. A canonical example is the *leave-one-out-safe* model, trained on the full dataset excluding x. Since $\mathsf{safe}(x)$ does not have access to x, the probability that it generates content resembling x is exponentially small. Any such resemblance is considered fortuitous. Formally, the NAF criterion is defined as follows:

Definition 2 (Near Access-Freeness [Vyas et al., 2023])**.** *Let $\mathcal{C}$ be a set of copyrighted samples and $\mathcal{W}$ a set of generative models. Given a mapping* $\mathsf{safe} : \mathcal{C} \to \mathcal{W}$ *and a divergence measure* Δ*, we say a model w is k_y-*near access-free *(or k_y-NAF) on prompt $y \in \mathsf{Y}$ if for every $x \in \mathcal{C}$,*

$$\Delta\left(p(\cdot|y) \,\|\, \mathsf{safe}_x(\cdot|y)\right) \leq k_y.$$

If $k_y = 0$, the model is indistinguishable from a safe model, meaning any resemblance to copyrighted material is by random chance. More generally, a small k_y suggests the model is unlikely to generate outputs resembling x with higher probability than a model that has never seen x.

D.1 Connection to Differential Privacy

NAF is closely related to concepts in *Differential Privacy (DP)* [Elkin-Koren et al., 2023]. Depending on the divergence measure Δ, NAF resembles different DP variants – for example, ε-DP when $\Delta = \Delta_{\max}$ [Dwork et al., 2006], and $(1, \varepsilon)$-Rényi DP when $\Delta = \Delta_{\mathrm{KL}}$.

Translating DP to generative models yields the following definition:

Definition 3 (Differentially Private Generation (DPG))**.** *Let S and S' be neighboring datasets. Denote by $P_S(\cdot|y)$ the distribution over outputs generated by a model trained on, or adapted from, S with algorithm $\mathcal{A}$, where randomness includes both training and generation stages. The generation is said to satisfy ε-*Differentially Private Generation *(ε-DPG) if for every $y \in \mathcal{Y}$,*

$$\Delta\left(P_S(\cdot|y) \,\|\, P_{S'}(\cdot|y)\right) \leq \varepsilon.$$

Here, *neighboring datasets* differ by a single data point (or privacy unit). If the training process is ε-DP, then the outputs naturally satisfy ε-DPG via the data processing inequality. One benefit of DPG is the flexibility to add noise during generation rather than training, potentially improving the utility-privacy tradeoff.

However, there are notable distinctions. ε-DP offers protection under arbitrary post-processing and multiple outputs, whereas ε-DPG only guarantees privacy for single outputs. Also, under DP, the trained model can be released, but under DPG, only the outputs are safe to share.

Elkin-Koren et al. [2023] highlight further differences: NAF is *one-sided*—comparing a model to a fixed safe reference—whereas DPG is *symmetric*. This asymmetry in NAF can enable better utility. Additionally, NAF allows more flexibility in choosing the safe model, which can be exploited in algorithm design.

Given these conceptual overlaps, both DP-SGD based training and our proposed private adaptation method DPAgg-TI satisfies ε-DP, so they naturally satisfy ε-NAF with the leave-one-out safe model.

We emphasize that this guarantee is meaningful only *within the formal framework of NAF*. It does not imply broader legal immunity or empirical indistinguishability from the original content. However, within this framework, satisfying ε-NAF allows us to argue that any close resemblance between outputs and private training data is no more likely than would be expected from a model that never had access to that data. This theoretical grounding supports the privacy and safety claims of our adaptation method.

Importantly, the goal of DPAgg-TI is to adapt to the *style* of a private image set—not its precise content. This distinction matters: style transfer is widely considered to fall under the doctrine of *fair use*, particularly in artistic and creative contexts. As discussed in Elkin-Koren et al. [2023] and further elaborated in legal analysis such as Carlini et al. [2023], generating new content in the style of a work, without reproducing its substantive elements, is generally not considered copyright infringement. Therefore, the use of DPAgg-TI to learn and reproduce stylistic attributes does not contradict the spirit or intent of the NAF framework. Instead, it offers a promising direction for responsibly fine-tuning generative models on private or copyrighted sources while respecting both privacy and intellectual property boundaries.

E Computational Cost Comparisons

Direct comparisons of computational cost across methods are inherently challenging due to differing training paradigms, optimization procedures, and parameter settings. Nonetheless, to provide a concrete sense of scale, we report representative computational costs for each method based on experiments conducted using a Stable Diffusion v1.5 model on a single NVIDIA A100 GPU. Below we summarize both training and inference overheads (the number of steps are optimized for each setup):

Method	Steps	Batch Size	Time	Memory Usage
TI (no DP)	10,000 (for 150 images)	1	25 min	7 GB
		8	2.5 hours	20 GB
TI (DP-SGD)	30,000 (for 150 images)	1	80 min	7 GB
		8	7 hours	20 GB
DPAgg-TI	2,000 per image	N/A	$\sim$5 min/image	7 GB
SG	N/A	N/A	N/A	N/A

Table 4: Training cost comparison across methods. Overhead from DP-SGD is relatively modest due to the low-dimensional embedding being optimized. N/A for SG means nothing is trained aside from the base model.

Method	Steps	Batch Size	Time	Memory Usage
TI (no DP, DP-SGD, DPAgg-TI)	50	1	1–2 sec	4 GB
	100	1	1–2 min	58 GB
SG (no DP, DPAgg-SG)	500	1	$\sim$30 min	17 GB

Table 5: Inference cost comparison across methods.

F Additional Style Transfer and Ablation Results

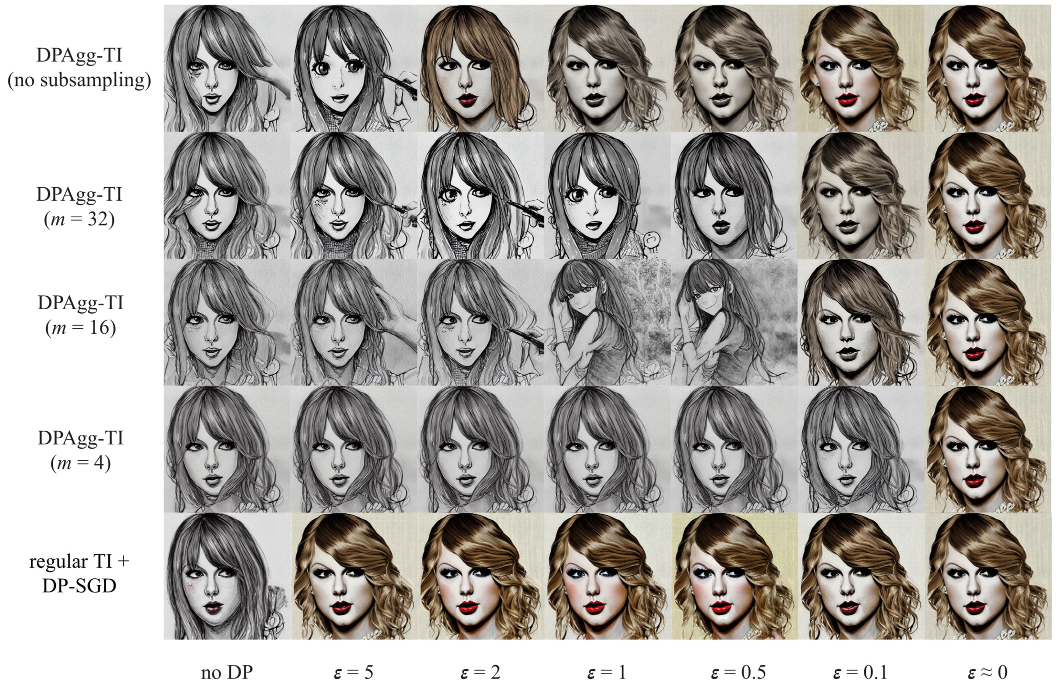

Figure 12: Images generated by Stable Diffusion v1.5 using the prompt "A painting of Taylor Swift in the style of <`@eveismyname`>", with the embedding <`@eveismyname`> trained using DPAgg-TI (with different subsample sizes m) and TI with DP-SGD using different values of ε.

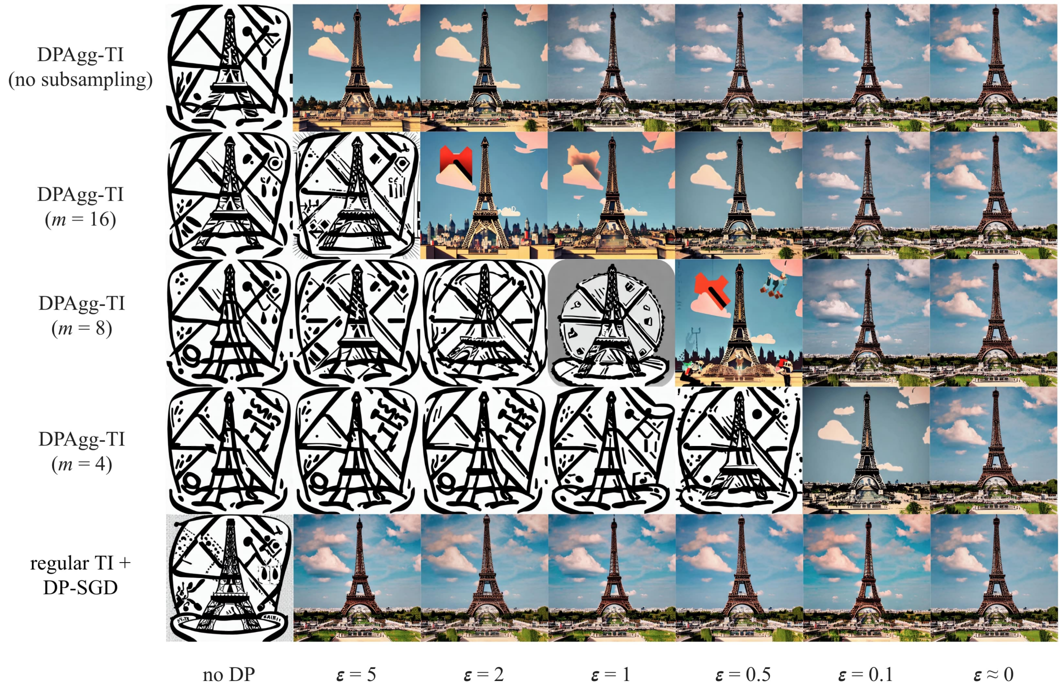

Figure 13: Images generated by Stable Diffusion v1.5 using the prompt "An icon of the Eiffel Tower in the style of <Paris 2024 Pictograms>", with the embedding <Paris 2024 Pictograms> trained using DPAgg-TI (with different subsample sizes m) and TI with DP-SGD using different values of ε.

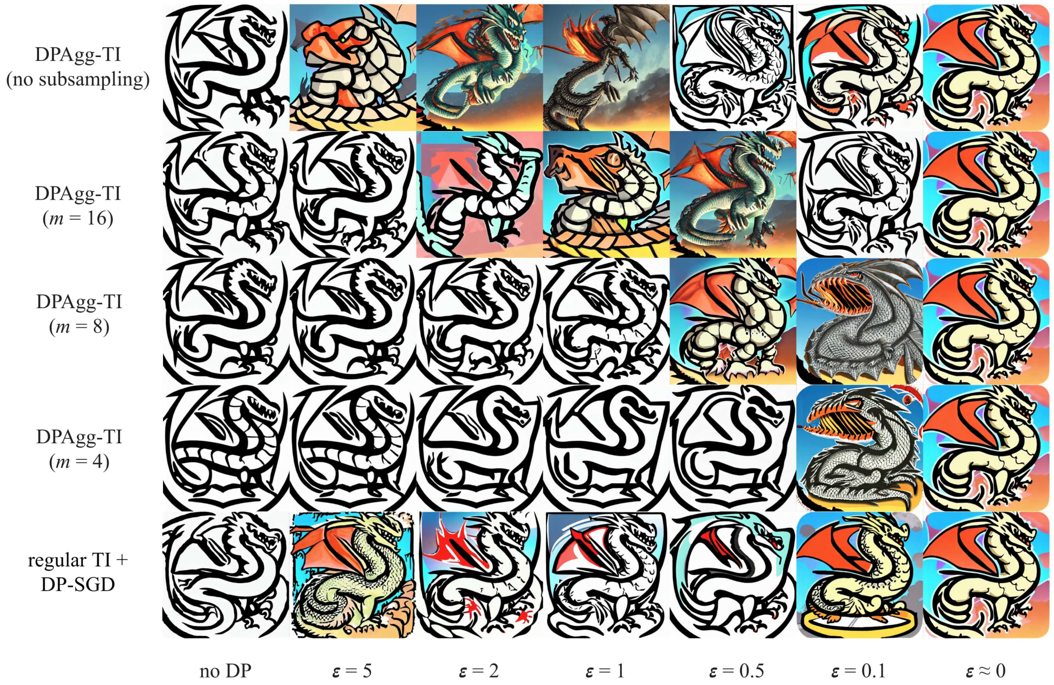

Figure 14: Images generated by Stable Diffusion v1.5 using the prompt "An icon of a dragon in the style of <Paris 2024 Pictograms>", with the embedding <Paris 2024 Pictograms> trained using DPAgg-TI (with different subsample sizes m) and TI with DP-SGD using different values of ε.