PROTECTING SENSITIVE DATA IN IMAGE SYNTHESIS VIA PAC-PRIVATE ADAPTATION FOR DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic data are increasingly used as an alternative to sharing sensitive records. However, synthetic data generation does not guarantee privacy, as diffusion models trained or adapted on sensitive data remain susceptible to reconstruction attacks. Moreover, while approaches that use differential privacy (DP), such as DP-SGD, achieve provably private diffusion model training, the repeated gradient clipping and noise injection they require result in significant utility loss. An important limitation of DP-based privacy is that, although it has a provable relationship to reconstruction privacy (RP), that relationship is indirect. RP is defined in terms of limiting how much an adversary's posterior distribution over sensitive data differs from the prior, whereas DP provides guarantees by bounding the sensitivity of outputs to changes in individual records. This indirection is an important source of the utility loss. To address this, we propose a PAC-private diffusion model adaptation to achieve reconstruction privacy. Since PAC-privacy is defined directly with respect to posterior advantage over the prior, it directly implicates RP. To obtain scalable PAC privatization in high dimensions, we first learn a compact data-dependent diffusion model component using LoRA or Textual Inversion, and then calibrate anisotropic Gaussian noise from the covariance of repeated mechanism outputs. Unlike DP-SGD, our method perturbs the learned component only once after optimization, thereby avoiding privacy composition across gradient updates. We evaluate the framework on few-shot concept personalization and full-dataset image synthesis, and show that the proposed approach better preserves subject identity, generation quality, and downstream classification accuracy than DP while achieving the same reconstruction privacy.

1 INTRODUCTION

Synthetic data provide an appealing alternative to the direct use of raw sensitive data in downstream applications, such as clinical data analytics, allowing useful data to be generated and shared without directly exposing the original records (Hu et al., 2024). For image synthesis, diffusion models have emerged as a leading generative paradigm for synthetic data, demonstrating strong performance and supporting a wide range of modern image-generation applications (Nichol et al., 2021; Rombach et al., 2022; Podell et al., 2024).

However, diffusion-based data synthesis does not necessarily prevent the reconstruction of sensitive training data (Carlini et al., 2023; Stadler et al., 2022; Webster, 2023). Indeed, diffusion models may even memorize individual training examples (Matsumoto et al., 2023). Accordingly, for synthetic data to serve as a privacy-preserving alternative to the direct use of raw data, the synthesis process itself must provide an explicit privacy guarantee with respect to the sensitive training data. A widely used framework for providing such formal privacy guarantees for diffusion-based synthesis is differential privacy (DP) (Dwork et al., 2006), with DP-SGD (Abadi et al., 2016; Li et al., 2024; Liu et al., 2024a;b) commonly employed to train or fine-tune diffusion models on sensitive data. DP-SGD bounds per-example gradient contributions through clipping and adds Gaussian noise during private optimization. However, differentially private training also introduces significant utility degradation. Several approaches aim to reduce this degradation, for example, DP-Promise (Wang et al., 2024), which exploits the noise inherent in the diffusion forward process, and DP-FETA (Li

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2025), which adds a private preprocessing step to warm start diffusion model training. Given the widespread use of latent diffusion models for scalable, high-resolution image synthesis (Rombach et al., 2022), we focus on privacy-preserving synthesis in this practically important setting. However, these approaches work by exploiting the properties of images, and cannot be directly extended to latent diffusion, which requires additional sensitivity control because the VAE latent representations need not have a bounded input domain.

Stepping back, there are two principal sources of utility loss in DP-based diffusion model training. First, privacy must be enforced at every optimization step, using up privacy budget. Second, DP is a worst-case and somewhat abstract notion of privacy which considers output sensitivity to any single-record changes. However, the main motivation for generating synthetic data is to avoid revealing real sensitive data, which naturally translates to the goal of preventing successful reconstruction of original data or sensitive attributes within it, or *reconstruction privacy (RP)*. RP, in turn, can be formally captured as ensuring that the attacker's posterior distribution over the training data conditional on observing the diffusion model, or the synthetic data generated by it, does not significantly differ from the prior data distribution—as measured, for example, using KL-divergence. While bounds on this difference can be derived from DP guarantees (as we show), the indirectness which results (since no bounds are entirely tight, and worst-case adversarial assumptions of DP are stronger than what we need for RP) entails significant utility loss.

Together, these considerations motivate a different route to privacy-preserving latent diffusion synthesis. First, rather than enforcing privacy at every optimization step, we calibrate it directly to the learned adaptation that is ultimately used for privacy-preserving synthetic image generation. Because the synthetic images are generated from the final learned adaptation, the data-processing inequality allows us to enforce this objective by privatizing the adaptation itself. Specifically, instead of clipping and perturbing gradients throughout optimization, we apply a single calibrated perturbation to the learned adaptation parameters after training, thereby avoiding explicit privacy composition across gradient updates. Second, since applying standard DP through output perturbation requires calibration to the worst-case sensitivity of the learned parameters, it tends to be overly conservative and result in significant utility loss. To address this, we instead use an information-theoretic privacy notion, PAC privacy (Xiao & Devadas, 2023), as the building block. PAC privacy is defined directly in terms of the adversarial posterior reconstruction advantage over the prior, thereby providing a natural formalization of RP. Additionally, we enable a principled comparison with DP by showing that (ε, δ)-DP bounds the reconstruction advantage.

The key remaining technical challenge in applying the privatization procedure to achieve PAC privacy is that it scales poorly with dimension as a result of a Monte Carlo calibration step it requires (Xiao & Devadas, 2023; Sridhar et al., 2025; Zhang & Vorobeychik, 2026). To make it practical in the high-dimensional image synthesis setting, we employ parameter-efficient diffusion adaptation. For tasks that require parameter fine-tuning, we learn compact LoRA adapters instead of optimizing the full parameter space. Because functionally equivalent LoRA factors can exhibit different scalings, we canonicalize the factors before covariance estimation to remove this ambiguity and improve calibration stability. For simpler adaptation tasks that can be represented through text conditioning, we instead privatize the special-token embedding learned by Textual Inversion (Gal et al., 2022). In both cases, PAC calibration is performed on a substantially lower-dimensional learned component rather than on the full diffusion model.

We evaluate the proposed framework in two complementary settings. For few-shot personalized generation, we compare PAC output perturbation with DP-SGD using both Textual Inversion and LoRA, and evaluate concept or identity fidelity, text alignment, and perceptual image quality. For general-purpose synthetic data generation, we fine-tune diffusion models on private datasets using LoRA and evaluate the downstream classification accuracy of models trained on generated data. Across these settings, our results show that the proposed PAC-based final-output privatization typically preserves substantially greater utility than DP-SGD under matched privacy budgets.

In summary, we make the following contributions:

- We develop end-to-end privacy-preserving image synthesis based on PAC-private diffusion-model adaptation that avoids per-example perturbation over individual gradient updates.
- We develop a scalable PAC calibration framework for parameter-efficient diffusion adaptation using LoRA and Textual Inversion. For LoRA, we canonicalize the learned factors before calibration

to remove scale and sign ambiguities, yielding more stable covariance estimation and improved downstream utility.

- Experimental results demonstrate that our PAC-private diffusion-adaptation framework preserves substantially greater utility than DP-SGD in challenging high-resolution and small-data regimes.

2 Related Work

Differentially Private Diffusion Models. Private diffusion training is commonly implemented using DP-SGD. DPDM (Dockhorn et al., 2022) introduces noise multiplicity, which averages gradients over multiple diffusion noise realizations of each training example without additional privacy cost. Ghalebikesabi et al. (2023) improve private diffusion fine-tuning through public pretraining, augmentation multiplicity, and biased timestep sampling. Other approaches exploit the structure of diffusion training more directly, e.g., DP-Promise (Wang et al., 2024), which incorporates the forward diffusion process into its privacy analysis, and DP-FETA (Li et al., 2025), which adopts an easy-to-hard curriculum that first learns coarse structure from privately aggregated central images and subsequently performs DP-SGD training. As image resolution and model scale increase, private diffusion training becomes expensive. Another line of work therefore improves efficiency by leveraging pretrained diffusion models and restricting private optimization to a small subset of parameters. Jiang et al. (2025) use public data to construct a library of denoising trajectories, retrieve a similar public trajectory for each private training example, and apply DP training only to the remaining refinement stage. DP-LDM (Liu et al., 2024b) performs private training in latent space and fine-tunes only the attention modules. Liu et al. (2024a) propose a low-dimensional adaptation strategy for computationally efficient private fine-tuning, and DP-LoRA (Tsai et al., 2025) applies DP-SGD to LoRA modules in pretrained latent diffusion models and demonstrates that parameter-efficient private fine-tuning can achieve a better privacy-utility trade-off than full-parameter fine-tuning.

PAC Privacy and Its Applications. Our approach uses tools from PAC Privacy (Xiao & Devadas, 2023), which provides a framework that limits the improvement in an adversary's ability to recover private data after observing the output of a mechanism. It also enables automatic privacy analysis of arbitrary black-box data-processing mechanisms. Sridhar et al. (2025) further develop a practical framework for automatically privatizing such mechanisms through output canonicalization and calibrated perturbation. Subsequent work extends PAC Privacy to database systems (Sridhar et al., 2026; Battiston et al., 2026), while Zhu (2026) establishes linear mutual information composition for adaptively and adversarially selected PAC-private model responses. Zhang & Vorobeychik (2026) show that the Gaussian log-determinant bound used by existing PAC mechanisms is in general conservative for non-Gaussian outputs and introduce Residual-PAC Privacy with an automatic privatization algorithm termed Stackelberg PAC privatization to tightly implement the target privacy budget with less noise. PAC Privacy guarantees have also been studied for learnable data encodings (Xiao et al., 2024), constrained one-sided perturbations (Xiao et al., 2025), verifiable privatization using zero-knowledge proofs (Repetto et al., 2026), and autoregressive generation (Mirzadehsarcheshmeh & Khandani, 2026). More recently, Ertan et al. (2026) apply PAC Privacy to sign-quantized language-model fine-tuning.

3 Preliminaries

A major privacy concern in diffusion model adaptation for generating synthetic data is that someone (an "adversary") with access to the adapted diffusion model or the data it generates should not be able to reconstruct the original sensitive training data *beyond what an adversary armed solely with the prior could*. For example, if the training data includes sensitive faces as well as common non-sensitive background elements, such as trees and cars, we are only concerned that the faces should not be effectively reconstructed. In this section, we formalize this notion of *reconstruction privacy (RP)* in information-theoretic terms; it then serves as our underlying privacy goal for diffusion model adaptation. Subsequently, we overview the DP tools for achieving RP, as well as diffusion models.

Reconstruction Privacy. We begin by formalizing reconstruction privacy. The key idea RP captures is that an adversary with access to a model trained on sensitive data should learn little more about these data than if they only knew the prior data distribution. Let $\mathcal{X}$ denote the sample space and $\mathcal{D}$ the distribution over $\mathcal{X}$. Define $X = (x_1, \ldots, x_n)$ as the private dataset with each $x_i \in \mathcal{X}$

drawn i.i.d. from $\mathcal{D}$. We capture reconstruction attack success in terms of the attacker successfully identifying an estimate $\hat{x}$ which is close to some data point in X. To formalize, let $d(\hat{x}, x)$ be the distance (e.g., ℓ_p norm) between a pair $\widehat{x}, x \in \mathcal{X}$ and r a reconstruction threshold, such that the estimate $\hat{x}$ is successful if $\min_{x \in X} d(\widehat{x}, x) \leq r$.

Let $M(X)$ be a function that maps a dataset X to some output space $\mathcal{O}$. In our setting, M is the algorithm for diffusion model adaptation outputting an adapter. The privacy goal is that an attacker who observes the output of M should not thereby obtain much more *information* about X than through knowing $\mathcal{D}$ alone. More precisely, let $\delta(M)$ be the attacker-optimal posterior reconstruction error defined as $\delta(M) = 1 - \sup_{\mathcal{A}} \Pr_{X \sim \mathcal{D}^n}[\min_{x \in X} d(\mathcal{A}(M(X)), x) \leq r]$, where $\mathcal{A}$ is a reconstruction function, whereas the optimal prior reconstruction error is $\delta^o = 1 - \sup_{\widehat{x} \in \mathcal{X}} \Pr_{X \sim \mathcal{D}^n}[\min_{x \in X} d(\widehat{x}, x) \leq r]$. Treating prior and posterior reconstruction errors in distributional terms, we wish that the difference between these two distributions is small, naturally formalized in terms of KL-divergence as in PAC privacy (Xiao & Devadas, 2023) of which our notion of β-reconstruction privacy is a special case.

Definition 1 (β-reconstruction privacy)**.** *Let $\Delta_{\mathrm{KL}}(\delta^o, \delta(M)) \equiv D_{\mathrm{KL}}(\mathrm{Bern}(1 - \delta(M)) \| \mathrm{Bern}(1 - \delta^o))$ be the* reconstruction advantage *of M, where D_{KL} is the KL-divergence and $\mathrm{Bern}(\cdot)$ the Bernoulli distribution. Then M satisfies β-reconstruction privacy (β-RP) if $\Delta_{\mathrm{KL}}(\delta^o, \delta(M)) \leq \beta$.*

Moreover, $\Delta_{\mathrm{KL}}(\delta^o, \delta(M)) \leq I(X; M(X))$ (Xiao & Devadas, 2023), where $I(X; M(X))$ is the mutual information between X and $M(X)$ (Shannon, 1948). Thus, it suffices to bound $I(X; M(X))$ to achieve β-RP.

Differential Privacy. Unlike RP, differential privacy (DP) enforces indistinguishability between the output distributions induced by neighboring datasets that differ in one sample.

Definition 2 ((ε, δ)-differential privacy (Dwork et al., 2006))**.** *A randomized mechanism M satisfies (ε, δ)-differential privacy if for any neighboring datasets $X \simeq X'$ differing in the addition or removal of one sample, and any measurable set $S \subseteq \mathcal{O}$, $\Pr[M(X) \in S] \leq e^{\varepsilon} \Pr[M(X') \in S] + \delta$.*

For private neural network training, DP-SGD (Abadi et al., 2016) is the most widely adopted approach that applies the Gaussian mechanism to clipped stochastic gradients. In learning step t, let $g_t^{(i)} = \nabla_{\theta_t} \ell(\theta_t; x_i)$ denote the gradient of sample x_i. Each gradient is clipped as $\mathrm{clip}_C(g_t^{(i)}) = g_t^{(i)} \min\{1, C/\|g_t^{(i)}\|_2\}$, and for a minibatch $\mathcal{B}_t$, DP-SGD computes $\widetilde{g}_t = |\mathcal{B}_t|^{-1}\{\sum_{i \in \mathcal{B}_t} \mathrm{clip}_C(g_t^{(i)}) + Z_t\}$, where $Z_t \sim \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 C^2 I)$ and σ_{DP} is obtained from privacy accounting, and updates the model using $\widetilde{g}_t$.

Diffusion Models. Diffusion models learn to generate data by reversing a progressive Gaussian noising process (Ho et al., 2020; Song et al., 2020), often implemented in a lower-dimensional latent space to improve efficiency (Rombach et al., 2022). Specifically, given an image x, a pretrained encoder $\mathcal{E}$ produces its latent representation $z_0 = \mathcal{E}(x)$. For a variance schedule $\{\beta_t\}_{t=1}^T$, let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. A noisy latent at timestep t is sampled as $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. For text-conditioned generation, a text encoder maps a prompt y to a conditioning representation $c(y)$. The denoising network ϵ_θ is trained to predict the injected noise by minimizing $\mathcal{L}(\theta) = \mathbb{E}_{x \sim \mathcal{D},\, t \sim \mathrm{Unif}([T]),\, \epsilon \sim \mathcal{N}(0,I)}[\|\epsilon - \epsilon_\theta(z_t, t, c(y))\|_2^2]$. During generation, the reverse process starts from $z_T \sim \mathcal{N}(0, I)$. We use classifier-free guidance (Ho & Salimans, 2022), which combines conditional and unconditional predictions as $\widehat{\epsilon}_\theta = \epsilon_\theta(z_t, t, c(\varnothing)) + w[\epsilon_\theta(z_t, t, c(y)) - \epsilon_\theta(z_t, t, c(\varnothing))]$, where w is the guidance scale. A diffusion sampler iteratively estimates the noisy latent at the previous timestep, and the final latent is decoded as $\tilde{x} = \mathcal{G}(z_0)$, where $\mathcal{G}$ is the decoder jointly trained with $\mathcal{E}$.

4 Reconstruction-Private Diffusion Adaptation

4.1 Overall Framework

Training a diffusion model from scratch on private data requires many optimization steps to learn basic denoising behavior and general visual structure, a process difficult to make private without considerable sacrifice in utility. Instead, we initialize the model from publicly pretrained parameters θ_0 and use the private dataset only to learn a compact adaptation component ϕ. Our approach for

training ϕ proceeds in two phases: the noise calibration phase, in which we estimate the covariance matrix of an adaptation algorithm for ϕ, and the privatization phase, in which the trained adaptation is privatized by adding Gaussian noise to obtain a provable β-RP guarantee.

Noise calibration. Our first step involves estimating the covariance of an adapter ϕ trained on bootstrapped subsamples of the dataset X. Specifically, given the public initialization θ_0, let Train denote the standard gradient-based training procedure for ϕ, and let $\mathcal{Q}$ denote the bootstrap distribution induced by the private training set. We repeatedly sample $X^{(k)} \sim \mathcal{Q}(X_{\mathrm{train}})$ and run the training algorithm to obtain $\phi^{(k)} = \mathrm{Train}(\theta_0, X^{(k)})$. Let $\bar{\phi} = \sum_{k=1}^{K} \phi^{(k)}/K$. We estimate the covariance of the trained adapter ϕ (viewed as a random vector) as $\widehat{\Sigma} = \sum_{k=1}^{K}(\phi^{(k)} - \bar{\phi})(\phi^{(k)} - \bar{\phi})^\top/(K-1)$. We progressively increase K until the relative change in $\mathrm{tr}(\widehat{\Sigma})$ falls below the convergence threshold τ, and then perform the eigendecomposition $\widehat{\Sigma} = V\,\mathrm{diag}(\lambda_1, \ldots, \lambda_d)V^\top$. The noise variance along direction V_i is set to $e_i = \sqrt{\lambda_i}\sum_{j=1}^{d}\sqrt{\lambda_j}/(2\beta)$, giving $\widehat{\Sigma}_B = V\,\mathrm{diag}(e_1, \ldots, e_d)V^\top$. Algorithm 1 summarizes the complete procedure.

Algorithm 1 Noise Calibration for Diffusion Adapters

Require: Private dataset X_{train}; frozen public model θ_0 and pretrained encoder $\mathcal{E}$; trainable component $\phi \in \mathbb{R}^d$ with fixed initialization ϕ_{init}; adaptation steps S; learning rate η; bootstrap distribution $\mathcal{Q}(X_{\mathrm{train}})$ that samples $|X_{\mathrm{train}}|$ records with replacement; mutual-information budget β; tolerance τ

Ensure: Noise covariance $\widehat{\Sigma}_B$

$\Phi \leftarrow \varnothing$, $K \leftarrow 0$, $c \leftarrow \infty$, $\nu_{\mathrm{old}} \leftarrow \perp$
while $c \geq \tau \ \vee \ K \leq 2$ **do**
  $K \leftarrow K + 1$
  $X^{(K)} \sim \mathcal{Q}(X_{\mathrm{train}})$; $\phi_0 \leftarrow \phi_{\mathrm{init}}$
  for $s = 0, \ldots, S-1$ **do**
    Sample a minibatch $\mathcal{B}_s = \{(x_{s,j}, y_{s,j})\}_{j=1}^{B}$ from $X^{(K)}$
    Independently sample $t_{s,j} \sim \mathrm{Unif}([T])$ and $\epsilon_{s,j} \sim \mathcal{N}(0, I)$ for each $j = 1, \ldots, B$
    $z^0_{s,j} \leftarrow \mathcal{E}(x_{s,j})$; $z^t_{s,j} \leftarrow \sqrt{\bar{\alpha}_{t_{s,j}}}z^0_{s,j} + \sqrt{1-\bar{\alpha}_{t_{s,j}}}\epsilon_{s,j}$
    $\mathcal{L}_s \leftarrow \sum_{j=1}^{B}\|\epsilon_{s,j} - \epsilon_{\theta_0,\phi_s}(z^t_{s,j}, t_{s,j}, c(y_{s,j}))\|_2^2/B$
    $\phi_{s+1} \leftarrow \phi_s - \eta\nabla_{\phi_s}\mathcal{L}_s$ *# Update only the adaptation component*
  end for
  $\phi^{(K)} \leftarrow \mathrm{vec}(\phi_S)$; $\Phi \leftarrow \Phi \cup \{\phi^{(K)}\}$
  if $K \geq 2$ **then**
    $\bar{\phi} \leftarrow \sum_{k=1}^{K}\phi^{(k)}/K$; $\widehat{\Sigma} \leftarrow \sum_{k=1}^{K}(\phi^{(k)} - \bar{\phi})(\phi^{(k)} - \bar{\phi})^\top/(K-1)$; $\nu \leftarrow \mathrm{tr}(\widehat{\Sigma})$
    $c \leftarrow \infty$ if $\nu_{\mathrm{old}} = \perp$; otherwise $c \leftarrow |\nu - \nu_{\mathrm{old}}|/\max\{\nu_{\mathrm{old}}, 10^{-12}\}$
    $\nu_{\mathrm{old}} \leftarrow \nu$
  end if
end while
Apply the eigendecomposition $\widehat{\Sigma} = V\,\mathrm{diag}(\lambda_1, \ldots, \lambda_d)V^\top$
$e_i \leftarrow \sqrt{\lambda_i}\sum_{j=1}^{d}\sqrt{\lambda_j}/(2\beta)$ for $i = 1, \ldots, d$
$\widehat{\Sigma}_B \leftarrow V\,\mathrm{diag}(e_1, \ldots, e_d)V^\top$
return $\widehat{\Sigma}_B$

Private release and generation. Once the noise covariance $\widehat{\Sigma}_B$ has been determined, let $X^\star \sim \mathcal{D}^n$ denote the private dataset used for the final adaptation. We run the training procedure once more to obtain $\phi^\star = \mathrm{Train}(\theta_0, X^\star)$. Finally, we release the privatized adaptation $\widetilde{\phi} = \phi^\star + B$, where $B \sim \mathcal{N}(0, \widehat{\Sigma}_B)$. Next, we show that our full procedure yields β-RP.

Theorem 1 (Privacy Guarantee). *Let $\Sigma = \mathrm{Cov}(\phi^\star)$ denote the population covariance. Let Σ_B be obtained by applying the noise allocation step of Algorithm 1 to Σ at budget β. For $B \sim \mathcal{N}(0, \Sigma_B)$ independent of $X^\star$ and $\phi^\star$, the release $\widetilde{\phi} = \phi^\star + B$ satisfies $I(X^\star; \widetilde{\phi}) \leq \beta$. For any collection of images Y generated using $\widetilde{\phi}$, $I(X^\star; Y) \leq I(X^\star; \widetilde{\phi}) \leq \beta$.*

Proof. The mutual-information argument follows the PAC privacy calibration analysis of Sridhar et al. (2025). We have the Markov chain $X^\star \to \phi^\star \to \widetilde{\phi}$, and therefore $I(X^\star;\widetilde{\phi}) \le I(\phi^\star;\widetilde{\phi})$. The Gaussian maximum-entropy bound gives

$$I(\phi^\star;\widetilde{\phi}) \le \frac{1}{2}\log\det\big(I+\Sigma\Sigma_B^{-1}\big) = \frac{1}{2}\sum_{i=1}^{d}\log\left(1+\frac{\lambda_i}{e_i}\right) \le \frac{1}{2}\sum_{i=1}^{d}\frac{\lambda_i}{e_i} = \beta.$$

Finally, generation forms the Markov chain $X^\star \to \widetilde{\phi} \to Y$. The data-processing inequality therefore gives $I(X^\star;Y) \le I(X^\star;\widetilde{\phi}) \le \beta$. □

Finally, β-RP follows immediately by $\Delta_{\mathrm{KL}}(\delta^o, \delta(M)) \le I(X; M(X))$ where M can be either the full training and privatization procedure for $\widetilde{\phi}$ or this procedure composed with image generation.

4.2 Parameter-Efficient Diffusion Adaptation

The approach above relies on training an adapter ϕ. We instantiate it using two concrete adaptation methods: low-rank adaptation (LoRA) (Hu et al., 2021) and textual inversion.

Low-Rank Adaptation. For tasks requiring greater adaptation capacity, we apply Low-Rank Adaptation (LoRA) only to the key and value projections of the cross-attention modules in the denoising network. For a pretrained weight $W_0 \in \mathbb{R}^{d_{\mathrm{out}}\times d_{\mathrm{in}}}$, LoRA parameterizes the adapted weight as $W = W_0 + \frac{\alpha}{r}BA$, where $A \in \mathbb{R}^{r\times d_{\mathrm{in}}}$, $B \in \mathbb{R}^{d_{\mathrm{out}}\times r}$, and r is the LoRA rank. We keep the remaining model parameters frozen and optimize only A and B on the private dataset, with A randomly initialized and B initialized to zero.

Directly calibrating the raw factors is unstable because different factorizations can represent the same update, since $BA = (c^{-1}B)(cA)$ for any $c \neq 0$. We therefore canonicalize each learned pair before covariance estimation. Given $BA = U\,\mathrm{diag}(s)V^\top$, we set $B^{\mathrm{can}} = U\,\mathrm{diag}(\sqrt{s})$ and $A^{\mathrm{can}} = \mathrm{diag}(\sqrt{s})V^\top$, with singular values ordered and signs fixed deterministically. This transformation preserves the effective update while removing functionally irrelevant scaling and sign ambiguities. We estimate the covariance and apply privatization in this canonicalized parameter space before restoring the privatized factors to their corresponding attention layers for generation.

Textual Inversion. When a private concept can be learned without modifying the diffusion model, we adopt Textual Inversion (TI) (Gal et al., 2022), which introduces a pseudo-token S_* with a learnable embedding v. Given private reference images X and a prompt y containing S_*, TI freezes the diffusion model, text encoder, and all other token embeddings and optimizes only v as $v^\star(X) = \arg\min_v \mathbb{E}_{x\sim X,t,\epsilon}[\|\epsilon - \epsilon_{\theta_0}(z_t, t, c(y;v))\|_2^2]$, where $t \sim \mathrm{Unif}(\{1,\dots,T\})$ and $\epsilon \sim \mathcal{N}(0,I)$.

4.3 Privacy Budget Matching

We can also leverage (ε,δ)-DP to achieve β-RP guarantees. As shown by Dong et al. (2022), (ε,δ)-DP is equivalent to $f_{\varepsilon,\delta}$-DP where $f_{\varepsilon,\delta}(\alpha) = \max\{0, 1-\delta-e^\varepsilon\alpha, e^{-\varepsilon}(1-\delta-\alpha)\}$. If the prior reconstruction success probability is α, $f_{\varepsilon,\delta}$-DP bounds the posterior success probability by $1 - f_{\varepsilon,\delta}(\alpha)$ (Kulynych et al., 2025). Thus, for any $f_{\varepsilon,\delta}(\alpha) > 0$, setting $\beta(\alpha) = D_{\mathrm{KL}}(\mathrm{Bern}(1 - f_{\varepsilon,\delta}(\alpha))||\mathrm{Bern}(\alpha))$ ensures $\beta(\alpha)$-RP, allowing us to directly compare the bounds implied by DP and our approach. In our experiments, we use $\alpha^\star = (1-\delta)/(1+e^\varepsilon)$, with $1-f_{\varepsilon,\delta}(\alpha^\star) = (e^\varepsilon+\delta)/(1+e^\varepsilon)$ for (ε,δ)-DP, which gives $\beta_{\varepsilon,\delta} = D_{\mathrm{KL}}(\mathrm{Bern}((e^\varepsilon+\delta)/(1+e^\varepsilon)) \,\|\, \mathrm{Bern}((1-\delta)/(1+e^\varepsilon)))$ as the mutual-information budget to use for our approach.

5 Experiments

5.1 Experimental Setup

We evaluate the proposed approach, which we refer to as PAC as it builds on PAC privacy tools, in comparison with (ε,δ)-DP for diffusion model adaptation under reconstruction risk. In our evaluation, we fix ε and δ for DP, use these to recover $\beta(\varepsilon,\delta)$ for our approach (PAC), thereby leveling the corresponding privacy guarantees, and present our results by comparing utility between DP and PAC across tasks, datasets, and efficacy measures.

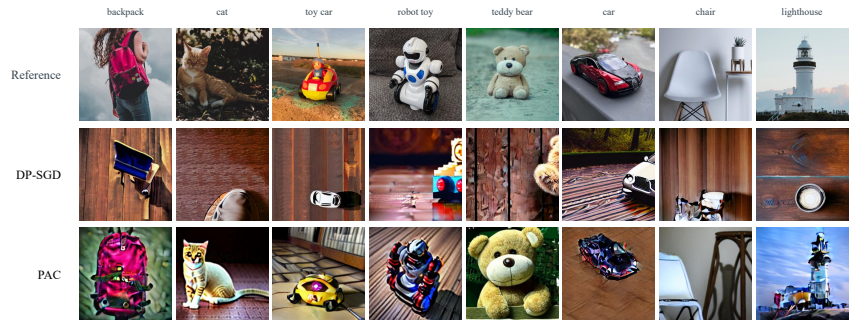

Figure 1: Qualitative comparison of synthetic images generated by PAC-privatized and DP-SGD-trained LoRA adapters at the matched privacy level $\varepsilon = 10$ and $\delta = 10^{-5}$.

Tasks and datasets. Our evaluation uses DreamBooth (Ruiz et al., 2023), CustomConcept101 (Kumari et al., 2023), and CelebA-HQ (Liu et al., 2015; Karras et al., 2017) for few-shot personalized generation. For general data synthesis, we use FashionMNIST (Xiao et al., 2017), EuroSAT (Helber et al., 2019), and BloodMNIST from MedMNIST v2 (Yang et al., 2023).

Backbones and adaptation mechanisms. We use Stable Diffusion V-1.5 (Rombach et al., 2022) for few-shot personalized generation with both LoRA and Textual Inversion, and BK-SDM-Small (Kim et al., 2024) for LoRA-based general data synthesis. Textual Inversion optimizes only the embedding of a special token. LoRA adapts the key and value projections in the cross-attention layers using rank-1 factors ($r = 1$), while keeping the pretrained backbone frozen. Detailed hyperparameters are provided in the Appendix B.

Evaluation. For object personalization on DreamBooth and CustomConcept101, we evaluate subject fidelity using DINO-I (Caron et al., 2021; Oquab et al., 2023) and CLIP-I (Radford et al., 2021), text alignment using CLIP-T (Radford et al., 2021), and perceptual quality using BRISQUE (Mittal et al., 2012). For face personalization, following Anti-DreamBooth (Van Le et al., 2023), we select 50 CelebA-HQ identities with 12 reference images each. We compute ISM as the cosine similarity between the normalized ArcFace (Deng et al., 2019) embedding of the largest detected generated face and the normalized mean embedding of its reference images. Because ISM excludes images without a detected face, we jointly report FDFR, the fraction of generated images in which no face is detected. We also report DINO-I, CLIP-I, and BRISQUE, but omit CLIP-T because all methods use the same fixed face prompt. Metrics are averaged over generation seeds and across concepts or identities. For general data synthesis, we follow the train-on-synthetic-test-on-real (TSTR) protocol: a classifier is trained exclusively on class-balanced synthetic images and evaluated on the official real test set. We report test accuracy as the primary metric and provide qualitative samples in the appendix.

5.2 Few-Shot Personalized Generation

LoRA. We evaluate LoRA adaptation on DreamBooth and CustomConcept101, comparing DP-SGD and PAC. Figure 1 presents qualitative examples of concepts drawn from these datasets at a matched privacy level. We can see that DP-SGD often fails to acquire conceptual information from the reference images. For example, the generated images for *bag* and *lighthouse* do not capture the corresponding concepts. For *teddy bear*, DP-SGD reproduces a fluffy texture but fails to preserve the overall shape. In contrast, PAC better preserves the distinctive attributes learned by the clean adapter. For example, for *robot*, PAC retains the overall appearance, and generated examples for *teddy bear* and *bag* also preserve visual characteristics that clearly resemble their reference images.

Table 1 presents the quantitative comparison. On both datasets, PAC achieves substantially higher concept fidelity across all evaluated privacy budgets. On DreamBooth, PAC improves DINO-I from 0.140 to 0.232 at $\varepsilon = 3$, from 0.149 to 0.352 at $\varepsilon = 10$, and from 0.179 to 0.379 at $\varepsilon = 50$. The corresponding CLIP-I scores increase from 0.577 to 0.596, from 0.579 to 0.645, and from 0.589 to 0.657, respectively. A similar trend appears on CustomConcept101. PAC improves DINO-I from 0.131 to 0.199 at $\varepsilon = 3$, from 0.141 to 0.353 at $\varepsilon = 10$, and from 0.174 to 0.390 at $\varepsilon = 50$. It also

Table 1: Comparison of PAC and DP-SGD for few-shot personalized generation. Results are averaged over 15 releases and then macro-averaged across subjects.

Dataset	ε	DINO-I ↑		CLIP-I ↑		CLIP-T ↑		BRISQUE ↓	
		PAC	DP	PAC	DP	PAC	DP	PAC	DP
DreamBooth	3	**.232**	.140	**.596**	.577	.212	**.232**	30.2	**29.4**
	5	**.306**	.141	**.627**	.577	.232	**.233**	**27.7**	28.0
	10	**.352**	.149	**.645**	.579	**.245**	.235	26.2	**25.6**
	20	**.373**	.158	**.655**	.581	**.252**	.240	25.4	**25.0**
	50	**.379**	.179	**.657**	.589	**.254**	.247	25.1	**21.8**
CustomConcept101	3	**.199**	.131	**.561**	.542	.200	**.233**	31.7	**27.5**
	5	**.286**	.129	**.596**	.543	.221	**.232**	**27.8**	**27.8**
	10	**.353**	.141	**.625**	.546	**.236**	**.236**	25.3	**24.5**
	20	**.383**	.152	**.639**	.550	**.244**	.240	**24.2**	24.4
	50	**.390**	.174	**.642**	.557	**.247**	**.247**	23.9	**22.0**

improves CLIP-I from 0.542 to 0.561, from 0.546 to 0.625, and from 0.557 to 0.642 at the same budgets. PAC sacrifices some text alignment under tight privacy constraints, but this gap closes as the budget increases: on DreamBooth, PAC surpasses DP-SGD in CLIP-T from $\varepsilon = 10$ onward, while on CustomConcept101 the two methods are comparable at larger budgets. BRISQUE does not exhibit a consistent winner, indicating that PAC's advantage primarily comes from preserving concept-specific information rather than uniformly improving perceptual quality. Overall, PAC retains substantially more of the learned concept than DP-SGD under matched privacy budgets.



Figure 2: Qualitative comparison of private personalized generation via TI on CelebA-HQ at $\varepsilon = 10$ and $\delta = 10^{-5}$.

Table 2: Private Textual Inversion results on CelebA-HQ. Results are averaged over 50 identities. All private mechanisms use $\delta = 10^{-5}$.

ε	ISM ↑		DINO-I ↑		CLIP-I ↑		BRISQUE ↓		FDFR ↓	
	PAC	DP	PAC	DP	PAC	DP	PAC	DP	PAC	DP
3	**.037**	.002	**.301**	.197	**.522**	.471	**7.1**	**7.1**	.24	**.22**
5	**.057**	.002	**.347**	.198	**.554**	.471	**6.5**	7.0	.23	**.22**
10	**.103**	.002	**.409**	.200	**.603**	.471	**6.0**	6.9	.23	**.22**
20	**.165**	.003	**.458**	.205	**.650**	.472	**5.4**	6.8	**.21**	.22
50	**.195**	.004	**.478**	.214	**.670**	.474	**5.3**	6.9	**.21**	.22

Textual Inversion. We compare the performance of TI privatized using DP-SGD and PAC on CelebA-HQ. Figure 2 presents qualitative results. DP-SGD largely fails to capture identity-specific facial characteristics, whereas PAC preserves substantially more facial structure and identity information. The quantitative results in Table 2 support this observation. Even at $\varepsilon = 3$, PAC improves ISM from 0.002 to 0.037, DINO-I from 0.197 to 0.301, and CLIP-I from 0.471 to 0.522 compared with DP-SGD. As the privacy budget increases, PAC improves steadily, reaching 0.195 ISM, 0.478 DINO-I, and 0.670 CLIP-I at $\varepsilon = 50$, whereas DP-SGD remains nearly unchanged across all budgets. PAC also achieves better perceptual quality for $\varepsilon \geq 5$, while the two methods exhibit comparable face-detection failure rates. Overall, final-output PAC perturbation preserves personalized identity considerably better than iterative DP-SGD perturbation for Textual Inversion.

Table 3: TSTR test accuracy (%) across privacy budgets. Each entry reports the mean ± the half-width of the 95% confidence interval.

Dataset	Method	$\varepsilon = 3$	$\varepsilon = 5$	$\varepsilon = 10$	$\varepsilon = 20$	$\varepsilon = 50$
FMNIST	DP	35.7 ± 4.9	36.1 ± 5.9	37.2 ± 6.5	39.4 ± 4.4	39.9 ± 6.9
	PAC	$\mathbf{36.4 \pm 4.8}$	$\mathbf{37.7 \pm 4.5}$	$\mathbf{43.1 \pm 4.6}$	$\mathbf{50.9 \pm 5.6}$	$\mathbf{52.6 \pm 4.2}$
EuroSAT	DP	33.1 ± 2.9	34.3 ± 2.8	35.5 ± 2.0	35.8 ± 2.2	37.6 ± 3.2
	PAC	$\mathbf{33.4 \pm 3.0}$	$\mathbf{36.6 \pm 3.9}$	$\mathbf{39.6 \pm 3.8}$	$\mathbf{44.1 \pm 3.4}$	$\mathbf{46.4 \pm 3.8}$
BloodMNIST	DP	12.3 ± 3.2	13.3 ± 3.5	14.1 ± 3.8	14.5 ± 3.7	14.8 ± 2.8
	PAC	$\mathbf{12.7 \pm 1.6}$	$\mathbf{18.5 \pm 2.5}$	$\mathbf{26.4 \pm 5.8}$	$\mathbf{28.6 \pm 6.8}$	$\mathbf{30.2 \pm 5.0}$

5.3 General Data Synthesis

We further evaluate PAC-private LoRA adaptation for general data synthesis. Table 3 reports the resulting test accuracy with 95% confidence intervals. Both PAC and DP-SGD reduce utility because they introduce noise into the learned model. However, compared with DP-SGD, PAC preserves substantially more of the utility provided by the clean adapter. At $\varepsilon = 50$, which corresponds to a very loose privacy budget, the synthetic data generated after PAC perturbation achieve test accuracies of 52.6% on FashionMNIST, 46.4% on EuroSAT, and 30.2% on BloodMNIST, approaching the corresponding clean LoRA ceilings. In contrast, DP-SGD yields average downstream accuracies below 40% on FashionMNIST and EuroSAT and only 14.8% on BloodMNIST. When the budget decreases to $\varepsilon = 10$, PAC still exceeds DP-SGD by 5.9%, 4.1%, and 12.3% on the three datasets, respectively. Under tighter privacy budgets, both PAC and DP-SGD struggle to preserve the utility of the clean adapter, although PAC continues to achieve slightly higher average accuracy.

Table 4: Effect of LoRA canonicalization on FashionMNIST. Each entry reports TSTR test accuracy (%) as the mean ± the half-width of the 95% confidence interval.

Mechanism	$\varepsilon = 3$	$\varepsilon = 5$	$\varepsilon = 10$	$\varepsilon = 20$	$\varepsilon = 50$
Raw factors	34.9 ± 4.4	35.6 ± 5.8	40.6 ± 4.3	48.3 ± 4.4	50.6 ± 4.6
Canonicalized factors	$\mathbf{36.4 \pm 4.8}$	$\mathbf{37.7 \pm 4.5}$	$\mathbf{43.1 \pm 4.6}$	$\mathbf{50.9 \pm 5.6}$	$\mathbf{52.6 \pm 4.2}$

5.4 Effect of LoRA Canonicalization

Finally, we evaluate the effect of the LoRA canonicalization step on PAC privatization. The raw-factor and canonicalized variants differ only in whether canonicalization is applied before covariance estimation and noise perturbation. Table 4 shows that canonicalization improves TSTR accuracy by 1.5, 2.1, 2.5, 2.6, and 2.0 percentage points at $\varepsilon \in \{3, 5, 10, 20, 50\}$, respectively. Although the confidence intervals overlap at individual privacy budgets, the canonicalized variant achieves a higher mean accuracy at every operating point. This consistent trend indicates that eliminating the arbitrary scale of equivalent LoRA factorizations improves the stability and utility of PAC noise calibration.

6 Conclusion

In this work, we propose an alternative to DP-SGD for private diffusion generation by perturbing the learned output once at the end of training, thereby avoiding privacy composition across optimization steps. We consider a distribution reconstruction privacy notion and derive a reconstruction-based conversion that maps (ε, δ)-DP parameters to a comparable mutual-information budget. Given this budget, we repeatedly execute the training mechanism, estimate the covariance of its outputs, and calibrate an anisotropic noise distribution for the final release. To make this procedure computationally feasible for diffusion models, we instantiate it with two low-dimensional adaptation mechanisms: LoRA and Textual Inversion. For LoRA, we further introduce a function-preserving canonicalization that removes factor-scale ambiguity and improves calibration stability. We evaluate the framework on both general-purpose data synthesis and few-shot personalized generation. The results show that the proposed approach preserves substantially more downstream and concept-specific utility than DP-SGD across a broad range of matched privacy budgets.

AI USE STATEMENT

We used generative AI tools to improve the clarity and grammar of the manuscript and to identify typographical errors. We also used these tools to assist with code debugging, LaTeX formatting, and the presentation of figures and tables. The synthetic image datasets evaluated in this work were generated by the diffusion models studied in the paper as part of our experimental methodology, rather than by an external AI assistant. We did not use generative AI tools for research problem formulation or ideation. The authors reviewed all AI-assisted content and verified the assisted code through executed experiments. All reported numerical results were obtained from these experiments and were not generated by an AI assistant. We take full responsibility for the final content of this work, including any content produced with the assistance of generative AI tools.

ETHICS STATEMENT

This work studies privacy-preserving image synthesis and is intended to reduce the risk of reconstructing sensitive training data from released generative models or synthetic images. Our experiments use publicly available benchmark datasets, and we do not collect new human-subject data. CelebA-HQ contains identifiable facial images and is used only for evaluating privacy-preserving personalization under the established benchmark protocol.

REPRODUCIBILITY STATEMENT

We provide the complete PAC calibration procedure in Algorithm 1. The appendix documents the model configurations, optimization settings, generation settings, and evaluation metrics. The supplementary material includes the core implementation of our method.

REFERENCES

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*, pp. 308–318, 2016.

Ilaria Battiston, Dandan Yuan, Xiaochen Zhu, and Peter Boncz. Simd-pac-db: Pretty performant pac privacy. *arXiv preprint arXiv:2603.15023*, 2026.

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)*, pp. 5253–5270, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

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

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

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pp. 486–503. Springer, 2006.

Murat Bilgehan Ertan, Xiaochen Zhu, Phuong Ha Nguyen, Marten van Dijk, and Srinivas Devadas. Paczero: Pac-private fine-tuning of language models via sign quantization. *arXiv preprint arXiv:2605.06505*, 2026.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

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.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, 34:11631–11642, 2021.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

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. *arXiv preprint arXiv:2106.09685*, 2021.

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

Tanqiu Jiang, Changjiang Li, Fenglong Ma, and Ting Wang. Rapid: Retrieval augmented training of differentially private diffusion models. *arXiv preprint arXiv:2502.12794*, 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. Bk-sdm: A lightweight, fast, and cheap version of stable diffusion. In *European Conference on Computer Vision*, pp. 381–399. Springer, 2024.

Bogdan Kulynych, Juan Gomez, Georgios Kaissis, Jamie Hayes, Borja Balle, Flavio Calmon, and Jean Raisaro. Unifying re-identification, attribute inference, and data reconstruction risks in differential privacy. *Advances in Neural Information Processing Systems*, 38:111967–112003, 2025.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1941. IEEE, 2023.

Kecen Li, Chen Gong, Zhixiang Li, Yuzhong Zhao, Xinwen Hou, and Tianhao Wang. PrivImage: Differentially private synthetic image generation using diffusion models with Semantic-Aware pretraining. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 4837–4854, 2024.

Kecen Li, Chen Gong, Xiaochen Li, Yuzhong Zhao, Xinwen Hou, and Tianhao Wang. From easy to hard: Building a shortcut for differentially private image synthesis. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 3988–4006. IEEE, 2025.

Jing Liu, Andrew Lowy, Toshiaki Koike-Akino, Kieran Parsons, and Ye Wang. Efficient differentially private fine-tuning of diffusion models. *arXiv preprint arXiv:2406.05257*, 2024a.

Michael F Liu, Saiyue Lyu, Margarita Vinaroz, and Mijung Park. Differentially private latent diffusion models. *Transactions on Machine Learning Research*, 2024b.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

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

Mina Mirzadehsarcheshmeh and Amir Keyvan Khandani. Pac-private autoregressive generation: Calibrating noise to ensemble disagreement. *arXiv preprint arXiv:2609.05676*, 2026.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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. In *International Conference on Learning Representations*, pp. 1862–1874, 2024.

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*, pp. 8748–8763. PmLR, 2021.

Guilhem Repetto, Nojan Sheybani, Gabrielle De Micheli, and Farinaz Koushanfar. Pac to the future: Zero-knowledge proofs of pac private systems. In *Companion Proceedings of the ACM Web Conference 2026*, pp. 1532–1540, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510. IEEE, 2023.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

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

Mayuri Sridhar, Hanshen Xiao, and Srinivas Devadas. Pac-private algorithms. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 3839–3857. IEEE, 2025.

Mayuri Sridhar, Michael A Noguera, Chaitanyasuma Jain, Kevin Kristensen, Srinivas Devadas, Hanshen Xiao, and Xiangyao Yu. Pac-private databases. *Cryptology ePrint Archive*, 2026.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1451–1468, 2022.

Yu-Lin Tsai, Yizhe Li, Chia-Mu Yu, Xuebin Ren, Po-Yu Chen, Zekai Chen, and Francois Buet-Golfouse. Differentially private fine-tuning of diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4561–4571, 2025.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2116–2127. IEEE, 2023.

Haichen Wang, Shuchao Pang, Zhigang Lu, Yihang Rao, Yongbin Zhou, and Minhui Xue. dp-promise: Differentially private diffusion probabilistic models for image synthesis. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1063–1080, 2024.

Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Hanshen Xiao and Srinivas Devadas. Pac privacy: Automatic privacy measurement and control of data processing. In *Annual International Cryptology Conference*, pp. 611–644. Springer, 2023.

Hanshen Xiao, G Edward Suh, and Srinivas Devadas. Formal privacy proof of data encoding: The possibility and impossibility of learnable encryption. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1834–1848, 2024.

Hanshen Xiao, Jun Wan, Elaine Shi, and Srinivas Devadas. One-sided bounded noise: Theory, optimization algorithms and applications. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 4214–4228, 2025.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific data*, 10(1):41, 2023.

Tao Zhang and Yevgeniy Vorobeychik. Residual-PAC privacy: Automatic privacy control beyond the gaussian barrier. In *35th USENIX Security Symposium (USENIX Security 26)*, 2026.

Xiaochen Zhu. *PAC-private responses with adversarial composition*. PhD thesis, Massachusetts Institute of Technology, 2026.

A Additional Theoretical Background

A.1 f-Differential Privacy

We briefly review f-differential privacy (f-DP), which characterizes privacy through binary hypothesis testing between the outputs of a mechanism on neighboring datasets Dong et al. (2022).

Definition 3 (Trade-off Function)**.** *Let P and Q be two distributions over the same output space. For a hypothesis test ψ, let $\alpha_\psi = \mathbb{E}_P[\psi]$ and $\beta_\psi = 1 - \mathbb{E}_Q[\psi]$ denote its type-I and type-II errors. The trade-off function between P and Q is*

$$T(P, Q)(\alpha) := \inf_{\psi:\, \alpha_\psi \le \alpha} \beta_\psi, \qquad \alpha \in [0, 1]. \tag{1}$$

Definition 4 (f-Differential Privacy Dong et al. (2022))**.** *Let $f : [0, 1] \to [0, 1]$ be a trade-off function. A randomized mechanism M satisfies f-DP if, for every pair of neighboring datasets $D \simeq D'$,*

$$T(M(D), M(D'))(\alpha) \ge f(\alpha), \qquad \forall \alpha \in [0, 1]. \tag{2}$$

The standard (ε, δ)-DP guarantee is equivalent to $f_{\varepsilon,\delta}$-DP, where

$$f_{\varepsilon,\delta}(\alpha) = \max\{0,\, 1 - \delta - e^{\varepsilon}\alpha,\, e^{-\varepsilon}(1 - \delta - \alpha)\}. \tag{3}$$

A.2 Reconstruction Advantage Provided by DP

Kulynych et al. (2025) use strong reconstruction robustness (SRR) to interpret differential privacy in terms of reconstruction risk. SRR considers a strong adversary that knows all but one record of the input dataset and attempts to reconstruct the remaining record from the mechanism output. We show that the resulting reconstruction-advantage bound also applies to the weaker adversary considered in our setting, who observes only the mechanism output.

Definition 5 (Datapoint Reconstruction Advantage (Kulynych et al., 2025))**.** *Let $\mathcal{X}$ be the data domain, $\mathcal{D}$ a distribution over $\mathcal{X}$, and $M : \mathcal{X}^n \to \Theta$ a randomized mechanism. Let $\bar{D} \in \mathcal{X}^{n-1}$ be a fixed partial dataset known to the adversary, and let $Z \sim \mathcal{D}$ denote the remaining target record.*

Given a distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a reconstruction radius r, define the optimal prior reconstruction error as

$$\delta_p^o := 1 - \sup_{\widehat{z} \in \mathcal{X}} \Pr_{Z \sim \mathcal{D}}\big[d(Z, \widehat{z}) \le r\big]. \tag{4}$$

Let $O \sim M(\bar{D} \cup \{Z\})$ be the released mechanism output. After observing both $\bar{D}$ and O, the optimal posterior reconstruction error is

$$\delta_p(M \mid \bar{D}) := 1 - \sup_{\mathcal{A}} \Pr_{\substack{Z \sim \mathcal{D} \\ O \sim M(\bar{D} \cup \{Z\})}} \big[d\big(Z, \mathcal{A}(\bar{D}, O)\big) \le r\big], \tag{5}$$

where $\mathcal{A} : \mathcal{X}^{n-1} \times \Theta \to \mathcal{X}$ is a reconstruction attack. The SRR reconstruction advantage is

$$\mathrm{Adv}(M, \bar{D}, \mathcal{D}; d, r) := \delta_p^o - \delta_p(M \mid \bar{D}). \tag{6}$$

The datapoint reconstruction advantage can be bounded using f-DP.

Theorem 2 (Reconstruction Advantage under (ε, δ)-DP (Kulynych et al., 2025))**.** *If M satisfies (ε, δ)-DP, then, for every partial dataset $\bar{D}$,*

$$\mathrm{Adv}(M, \bar{D}, \mathcal{D}; d, r) \le \delta_p^o - f_{\varepsilon,\delta}(1 - \delta_p^o). \tag{7}$$

Equivalently,

$$\delta_p(M \mid \bar{D}) \ge f_{\varepsilon,\delta}(1 - \delta_p^o). \tag{8}$$

Corollary 1 (Output-Only Reconstruction Advantage)**.** *Let $\bar{D} \sim \mathcal{D}^{n-1}$, and let $\delta_p(M)$ denote the optimal reconstruction error of an adversary that observes only the mechanism output. If M satisfies (ε, δ)-DP, then*

$$\delta_p^o - \delta_p(M) \le \delta_p^o - f_{\varepsilon,\delta}(1 - \delta_p^o). \tag{9}$$

Proof. An adversary observing only the mechanism output is no stronger than one observing both the output and the partial dataset. More precisely, for any output-only attack $\mathcal{A}_0$, define the corresponding informed attack by $\mathcal{A}(\bar{D}, O) = \mathcal{A}_0(O)$. Therefore,

$$\begin{aligned} 1 - \delta_p(M) &\le \int \sup_{\mathcal{A}} \Pr\big[d(Z, \mathcal{A}(\bar{d}, O)) \le r \mid \bar{D} = \bar{d}\big] p(\bar{d})\, d\bar{d} \\ &= 1 - \int \delta_p(M \mid \bar{D} = \bar{d}) p(\bar{d})\, d\bar{d}. \end{aligned} \tag{10}$$

Equivalently,

$$\delta_p(M) \ge \int \delta_p(M \mid \bar{D} = \bar{d}) p(\bar{d})\, d\bar{d}. \tag{11}$$

By Theorem 2, $\delta_p(M \mid \bar{D} = \bar{d}) \ge f_{\varepsilon,\delta}(1 - \delta_p^o)$ for every $\bar{d}$. Hence, $\delta_p(M) \ge f_{\varepsilon,\delta}(1 - \delta_p^o)$. Subtracting both sides from δ_p^o proves the result. □

Let $\alpha := 1 - \delta_p^o$ denote the optimal prior reconstruction success rate. Corollary 1 implies $1 - \delta_p(M) \le 1 - f_{\varepsilon,\delta}(\alpha)$. For any fixed α satisfying $f_{\varepsilon,\delta}(\alpha) > 0$, we define $\beta(\alpha) := D_{\mathrm{KL}}(\mathrm{Bern}(1 - f_{\varepsilon,\delta}(\alpha)) \,\|\, \mathrm{Bern}(\alpha))$. Because the KL reconstruction advantage is bounded by mutual information, $I(X; M(X)) \le \beta(\alpha)$ gives the same upper bound on posterior reconstruction success as (ε, δ)-DP at the fixed prior success rate α.

B Experimental Details

B.1 Training Configurations

Table 5 summarizes the main training configurations. All experiments use the AdamW optimizer. PAC and DP-SGD use identical pretrained models, initialization, adaptation architecture, optimization schedule, preprocessing, and training prompts.

For DreamBooth and CustomConcept101, we train rank-one LoRA adapters on the key and value projections of the cross-attention layers. The private reference sets contain 4–6 images per DreamBooth subject and 3–15 images per CustomConcept101 subject. Subject identifiers are used only

Table 5: Training configurations used in our experiments. Generation resolution matches training resolution.

Dataset	Backbone	Resolution	Adaptation	Batch	Updates	Learning rate	Warm-up
DreamBooth	SD V-1.5	256×256	LoRA, $r = 1$	4	250	2×10^{-4}	7
CustomConcept101	SD V-1.5	256×256	LoRA, $r = 1$	4	250	2×10^{-4}	7
CelebA-HQ	SD V-1.5	512×512	Textual Inversion	1	500	5×10^{-4}	0
FashionMNIST	BK-SDM-Small	256×256	LoRA, $r = 1$	8	7,500	10^{-4}	25
EuroSAT	BK-SDM-Small	256×256	LoRA, $r = 1$	8	8,100	10^{-4}	25
BloodMNIST	BK-SDM-Small	256×256	LoRA, $r = 1$	8	7,475	10^{-4}	25

to locate the reference images and organize the outputs. The text condition always uses the associated class word. For example, `dog2` is trained with "a photo of a dog," `pet_cat3` with "a photo of a cat," and `transport_car10` with "a photo of a car." Consequently, subjects belonging to the same class receive identical prompts and are distinguished only through their private reference images and learned LoRA adapters. For CelebA-HQ, we optimize one 768-dimensional Textual Inversion embedding using all 12 reference images of each identity. At every update, the training prompt is sampled uniformly from the eight standard Textual Inversion templates. For FashionMNIST, EuroSAT, and BloodMNIST, we use class-specific prompt templates tailored to the visual domain of each dataset. The details are provided in Table 6.

Privacy mechanisms. DP-SGD clips each per-example gradient to $C = 0.01$. Given a private dataset of size n and a minibatch of size b, the sampling rate is $q = \min\{b/n, 1\}$. We use the PRV accountant Gopi et al. (2021) to determine the noise multiplier for each target ε, with $\delta = 10^{-5}$. For PAC calibration, we repeat the complete training procedure $K = 96$ times to estimate the output covariance. At $K = 96$, the relative change in the trace of the estimated covariance falls below the convergence threshold $\tau = 0.05$.

B.2 Evaluation Settings

Generation protocol. Table 7 presents the evaluation prompts used for each dataset. For DreamBooth and CustomConcept101, each released adapter is evaluated using ten fixed DreamBooth prompt templates instantiated with the class name only and without an identifier token. We generate four images per prompt using 50 denoising steps, classifier-free guidance of 7.5, and a resolution of 256×256. For CelebA-HQ, we use the single fixed prompt "a photo of sks person" and generate 16 images per identity using 50 denoising steps, classifier-free guidance of 7.5, and a resolution of 512×512. For FashionMNIST, EuroSAT, and BloodMNIST, we generate class-balanced synthetic training sets using the same class-specific prompts employed during training. Generation uses 25 denoising steps, classifier-free guidance of 7.5, and a resolution of 256×256. Table 8 summarizes the synthetic training-set sizes and real evaluation splits.

Object-personalization metrics. We evaluate subject fidelity using DINO-I and CLIP-I. DINO-I uses `dinov2_vits14`, while CLIP-I uses the image encoder of CLIP ViT-L/14. For each generated image, we compute its cosine similarity to every reference image. Similarities are first averaged within each prompt and then across prompts. CLIP-T is the cosine similarity between the CLIP ViT-L/14 embeddings of a generated image and its generation prompt. Perceptual image quality is measured using BRISQUE, for which lower values indicate better quality.

Face-personalization metrics. For facial identity evaluation, we specifically adopt identity score matching (ISM) to measure identity preservation. ISM is computed using InsightFace `buffalo_l`. Faces are detected using SCRFD at 640×640, and identity embeddings are extracted using its ArcFace ResNet-50 model trained on WebFace600K. We average and normalize the ArcFace embeddings of the 12 reference images to obtain one reference identity vector. For each generated image, we retain the largest detected face and compute its cosine similarity to the reference vector. ISM is averaged only over images in which a face is detected. We therefore jointly report the face-detection failure rate (FDFR), defined as the fraction of generated images containing no detected face.

Table 6: Training prompts used in our experiments. For LoRA personalization, {class} denotes the semantic class specified by the dataset metadata rather than the subject-directory name.

Dataset	Training prompts
DreamBooth	`a photo of a {class}`
CustomConcept101	`a photo of a {class}`
CelebA-HQ	`a photo of sks person`
	`a rendering of sks person`
	`a cropped photo of sks person`
	`the photo of sks person`
	`a close-up photo of sks person`
	`a bright photo of sks person`
	`a good photo of sks person`
	`a photo of one sks person`
FashionMNIST	**Template:** `a centered grayscale product image of {X} on a plain black background`
	Class descriptions:
	`a short-sleeved T-shirt`
	`a pair of trousers`
	`a pullover sweater`
	`a dress`
	`a coat`
	`a sandal`
	`a long-sleeved shirt`
	`a sneaker`
	`a bag`
	`an ankle boot`
EuroSAT	**Template:** `an overhead satellite image of {X}`
	Class descriptions:
	`annual cropland`
	`a forest`
	`herbaceous vegetation`
	`a highway`
	`an industrial area`
	`pasture land`
	`permanent cropland`
	`a residential area`
	`a river`
	`a sea or lake`
BloodMNIST	**Template:** `a microscopy image of {X}`
	Class descriptions:
	`a basophil blood cell`
	`an eosinophil blood cell`
	`an erythroblast`
	`an immature granulocyte`
	`a lymphocyte blood cell`
	`a monocyte blood cell`
	`a neutrophil blood cell`
	`a platelet`

TSTR evaluation. For general data synthesis, we follow the train-on-synthetic-test-on-real protocol. A ResNet-50 initialized with ImageNet-1K V2 weights is trained exclusively on the generated images after replacing its final classification layer. The classifier is optimized for 30 epochs using SGD with a learning rate of 0.01, momentum 0.9, weight decay 10^{-4}, a cosine learning-rate schedule, and a batch size of 128. Images are resized to 224×224, randomly flipped horizontally during training, and normalized using ImageNet statistics. EuroSAT is split into 21,600 training, 2,700 validation, and 2,700 test images. We select the number of optimization steps for EuroSAT and BloodMNIST using only their validation sets. All results reported in the main paper are evaluated on the corresponding held-out test sets.

C Additional Qualitative Results

Figures 3, 4, and 5 present qualitative examples generated using PAC-private adaptation and DP-SGD on FashionMNIST, EuroSAT, and BloodMNIST, respectively, under the matched privacy bud-

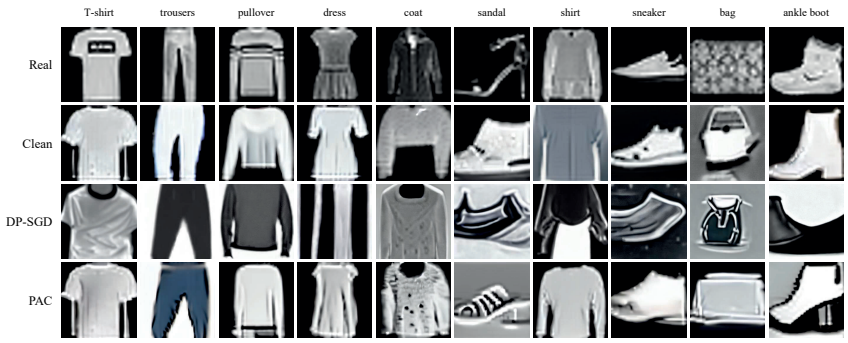

Figure 3: Qualitative comparison of synthetic images generated using PAC privatization and DP-SGD on FashionMNIST at $\varepsilon = 10$ and $\delta = 10^{-5}$.



Figure 4: Qualitative comparison of synthetic images generated using PAC privatization and DP-SGD on EuroSAT at $\varepsilon = 10$ and $\delta = 10^{-5}$.

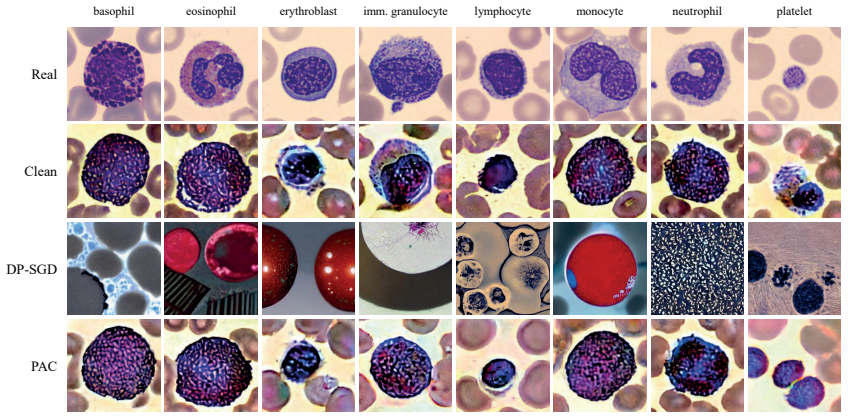

Figure 5: Qualitative comparison of synthetic images generated using PAC privatization and DP-SGD on BloodMNIST at $\varepsilon = 10$ and $\delta = 10^{-5}$.

Table 7: Generation and evaluation prompts used in our experiments. For object personalization, {class} is replaced by the semantic class associated with each subject; no identifier token is used. General-synthesis prompts are identical to their training prompts in Table 6.

Dataset	Generation and evaluation prompts

DreamBooth and CustomConcept101	`a {class} in the jungle`
	`a {class} in the snow`
	`a {class} on the beach`
	`a {class} on a cobblestone street`
	`a {class} on top of pink fabric`
	`a {class} on top of a wooden floor`
	`a {class} with a city in the background`
	`a {class} with a mountain in the background`
	`a {class} with the Eiffel Tower in the background`
	`a {class} floating on top of water`

CelebA-HQ	**Generation:** `a photo of sks person` **Class-only reference:** `a photo of a person`
FashionMNIST	Same as the class-specific FashionMNIST training prompts in Table 6.
EuroSAT	Same as the class-specific EuroSAT training prompts in Table 6.
BloodMNIST	Same as the class-specific BloodMNIST training prompts in Table 6.

Table 8: Generation and downstream evaluation settings for general data synthesis.

Setting	FashionMNIST	EuroSAT	BloodMNIST
Synthetic images per class	500	2,000	1,500
Total synthetic images	5,000	20,000	12,000
Denoising steps	25	25	25
Guidance scale	7.5	7.5	7.5
Generation resolution	256×256	256×256	256×256
Real test examples	10,000	2,700	3,421
Evaluation split	Official test	Held-out test	Official test

get $\varepsilon = 10$ and $\delta = 10^{-5}$. Compared with DP-SGD, PAC-private generation more clearly preserves the class-specific visual characteristics of the training data.

D Additional Quantitative Results

Table 9: TSTR test accuracy (%) across privacy budgets. Each entry reports the mean $\pm$ the half-width of the 95% confidence interval.

Dataset	Method	$\varepsilon = 3$	$\varepsilon = 4$	$\varepsilon = 5$	$\varepsilon = 7$	$\varepsilon = 10$
FMNIST	DP	35.7 ± 4.9	35.7 ± 4.6	36.1 ± 5.9	36.6 ± 6.6	37.2 ± 6.5
	PAC	$\mathbf{36.4 \pm 4.8}$	$\mathbf{37.1 \pm 4.5}$	$\mathbf{37.7 \pm 4.5}$	$\mathbf{39.1 \pm 5.8}$	$\mathbf{43.1 \pm 4.6}$
EuroSAT	DP	33.1 ± 2.9	34.3 ± 2.9	34.3 ± 2.8	34.4 ± 2.6	35.5 ± 2.0
	PAC	$\mathbf{33.4 \pm 3.0}$	$\mathbf{34.8 \pm 3.4}$	$\mathbf{36.6 \pm 3.9}$	$\mathbf{39.4 \pm 3.2}$	$\mathbf{39.6 \pm 3.8}$
BloodMNIST	DP	12.3 ± 3.2	13.2 ± 3.4	13.3 ± 3.5	13.8 ± 3.6	14.1 ± 3.8
	PAC	$\mathbf{12.7 \pm 1.6}$	$\mathbf{15.9 \pm 2.6}$	$\mathbf{18.5 \pm 2.5}$	$\mathbf{19.3 \pm 3.5}$	$\mathbf{26.4 \pm 5.8}$

We additionally evaluate a finer range of privacy budgets with $\varepsilon \leq 10$ to further compare DP-SGD and PAC privatization. Table 9 reports TSTR accuracy at $\varepsilon \in \{3, 4, 5, 7, 10\}$. PAC achieves higher mean accuracy than DP-SGD on all three datasets at every evaluated budget, with the performance gap generally increasing as ε increases. The improvement is particularly pronounced on BloodMNIST, where DP-SGD remains nearly unchanged while PAC progressively recovers downstream utility.

When averaging the results over the 50 CelebA-HQ identities, we observe substantial variation in performance across identities. To reflect this variability rather than reporting only the mean, we present 95% confidence intervals based on the variation across the 50 identities in Table 10. The results remain consistent with those in the main paper: PAC achieves higher ISM, DINO-I, and CLIP-I scores across all privacy budgets, whereas DP-SGD performance changes only marginally as ε increases.

Table 10: Private Textual Inversion results on 50 CelebA-HQ identities. Each entry reports the mean $\pm$ the half-width of the 95% confidence interval. All private mechanisms use $\delta = 10^{-5}$.

ε	ISM ↑		DINO-I ↑		CLIP-I ↑	
	PAC	DP	PAC	DP	PAC	DP
3	**.037±.009**	.002±.002	**.301±.016**	.197±.006	**.522±.012**	.471±.010
5	**.057±.010**	.002±.002	**.347±.018**	.198±.006	**.554±.013**	.471±.010
10	**.103±.013**	.002±.002	**.409±.021**	.200±.006	**.603±.014**	.471±.010
20	**.165±.017**	.003±.002	**.458±.019**	.205±.007	**.650±.014**	.472±.010
50	**.195±.021**	.004±.002	**.478±.018**	.214±.007	**.670±.014**	.474±.010