Consistent Diffusion Meets Tweedie: Training Exact Ambient Diffusion Models with Noisy Data

Giannis Daras [1 2] Alexandros G. Dimakis [3] Constantinos Daskalakis [4 2]

Abstract

Ambient diffusion is a recently proposed framework for training diffusion models using corrupted data. Both Ambient Diffusion and alternative SURE-based approaches for learning diffusion models from corrupted data resort to approximations which deteriorate performance. We present the first framework for training diffusion models that provably sample from the uncorrupted distribution given only noisy training data, solving an open problem in Ambient diffusion. Our key technical contribution is a method that uses a double application of Tweedie's formula and a consistency loss function that allows us to extend sampling at noise levels below the observed data noise. We also provide further evidence that diffusion models memorize from their training sets by identifying extremely corrupted images that are almost perfectly reconstructed, raising copyright and privacy concerns. Our method for training using corrupted samples can be used to mitigate this problem. We demonstrate this by fine-tuning Stable Diffusion XL to generate samples from a distribution using only noisy samples. Our framework reduces the amount of memorization of the fine-tuning dataset, while maintaining competitive performance.

[*]Equal contribution [1]Department of Computer Science, University of Texas at Austin [2]Archimedes AI [3]Department of Electrical and Computer Engineering, University of Texas at Austin [4]Department of Electrical Engineering and Computer Science, MIT. Correspondence to: Giannis Daras <giannisdaras@utexas.edu>, Alexandros G. Dimakis <dimakis@austin.utexas.edu>, Constantinos Daskalakis <costis@csail.mit.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

In recent years, we have witnessed remarkable progress in image generation as exemplified by state-of-the-art models such as Stable Diffusion (-XL) (Rombach et al., 2022; Podell et al., 2023) and DALL-E (2, 3) (Betker et al., 2023). This progress has been driven by two major enablers: i) the diffusion modeling framework (Ho et al., 2020; Song & Ermon, 2019; Song et al., 2020b); and ii) the existence of massive datasets of image-text pairs (Schuhmann et al., 2022; Gadre et al., 2023).

The need for high-quality, web-scale data and the intricacies involved in curating datasets at that scale often result in the inclusion of copyrighted content. Making things worse, diffusion models memorize training examples more than previous generative modeling approaches (Carlini et al., 2023; Somepalli et al., 2023), such as Generative Adversarial Networks (Goodfellow et al., 2020), often replicating parts or whole images from their training set.

A recently proposed strategy for mitigating the memorization issue is to train (or fine-tune) diffusion models using corrupted data (Daras et al., 2023b; Somepalli et al., 2023; Daras & Dimakis, 2023). Indeed, developing a capability for training diffusion models using corrupted data can also find applications in domains where access to uncorrupted data is expensive or impossible, e.g. in MRI (Aali et al., 2023) or black-hole imaging (Lin et al.; The Event Horizon Telescope Collaboration, 2019). Unfortunately, existing methods for learning diffusion models from corrupted data (Daras et al., 2023b; Aali et al., 2023; Kawar et al., 2023; Xiang et al., 2023) resort to approximations (during training or sampling) that significantly hurt performance. Our contributions are as follows:

i We propose the first *exact* framework for learning diffusion models using only corrupted samples. Our key technical contributions are: i) a computationally efficient method for learning optimal denoisers for all levels of noise $\sigma \geq \sigma_n$, where σ_n is the standard deviation of the noise in the training data, obtained by applying Tweedie's formula twice; and ii) a consistency loss function (Daras et al., 2023a) for learning the optimal denoisers for noise levels $\sigma \leq \sigma_n$. Note that given sam-

Figure 1. Top row: images from LAION (Schuhmann et al., 2022), middle row: masked images, bottom row: reconstructed images with the SDXL (Podell et al., 2023) inpainting model. The accuracy of the reconstructions presents strong evidence that the images on the top-row were in the training set of SDXL (or SDXL Inpainting) and have been memorized. To the best of our knowledge, SDXL does not disclose its training set.

ples at level of noise σ_n it is possible to obtain samples at levels of noise $\sigma > \sigma_n$ (by adding further noise) but prior to our work it was not known how to train diffusion models to obtain samples at levels of noise $\sigma < \sigma_n$.

ii We provide further evidence that foundation diffusion models memorize their training sets by showing that extremely corrupted training images can be almost perfectly reconstructed. Moreover, we show that memorization occurs at a higher rate than previously anticipated.

iii We use our framework to fine-tune diffusion foundation models using corrupted data and show that the performance of our trained model declines (as the corruption in the training data increases) at a much slower rate compared to previously proposed approaches.

iv We evaluate trained models against our as well as a baseline method for testing data replication and we show that models trained under data corruption memorize significantly less.

v We open-source our code to facilitate further research in this area: https://github.com/giannisdaras/ambient-tweedie.

2. Background and Related Work

Consider a distribution of interest admitting a density function p_0. Our goal is to train a diffusion model that generates samples from p_0. However, we only have access to noisy samples from p_0. In particular, we have samples of the form $\boldsymbol{X}_{t_n} = \boldsymbol{X}_0 + \sigma_{t_n}\boldsymbol{Z}$, where $\boldsymbol{X}_0 \sim p_0$ and $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, I_d)$. We denote by p_{t_n} the distribution density of these samples. Throughout the paper we fix an increasing non-negative function $\sigma(t)$, where $t \in [0, T]$, $T > 0$, and $\sigma(0) = 0$, and denote $\sigma(t)$ by σ_t. We take $t_n \in (0, T)$. The subscript 'n' in t_n refers to "nature" and, as stated above, we assume that nature is giving us access to samples at noise level σ_{t_n}. We denote by p_t the distribution of random variable $\boldsymbol{X}_t = \boldsymbol{X}_0 + \sigma_t\boldsymbol{Z}$, where $\boldsymbol{X}_0 \sim p_0$ and $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, I_d)$.

2.1. Background on denoising diffusion models

Diffusion models can equivalently be viewed as denoisers at many different noise levels σ_t, $t \in [0, T]$. They are typically trained with the Denoising Score Matching loss:

$$J_{\mathrm{DSM}}(\theta) = \mathbb{E}_{\boldsymbol{x}_0 \sim p_0(\boldsymbol{x}_0)} \mathbb{E}_{t \sim \mathcal{U}[0,T]} \mathbb{E}_{\boldsymbol{x}_t \sim p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[||\boldsymbol{h}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{x}_0||^2 \right].$$

If the function class $\{\boldsymbol{h}_\theta\}$ is sufficiently rich, the minimizer of this loss satisfies $\boldsymbol{h}_{\theta^*}(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{X}_0 | \boldsymbol{X}_t = \boldsymbol{x}_t]$ for all $t, \boldsymbol{x}_t$. Tweedie's formula connects the conditional expectation, i.e. the best denoiser in the ℓ_2^2 sense, with the score function $\nabla \log p_t(\boldsymbol{x}_t)$,

$$\nabla \log p_t(\boldsymbol{x}_t) = \frac{\mathbb{E}[\boldsymbol{X}_0 | \boldsymbol{X}_t = \boldsymbol{x}_t] - \boldsymbol{x}_t}{\sigma_t^2}. \tag{2.1}$$

stop diffusion sampling at time $t : \sigma_t = \sigma_{t_n}$. This type of approach is adopted by Kawar et al. (2023) and Xiang et al. (2023) and it only guarantees samples from the distribution $\mathbb{E}[\boldsymbol{X}_0|\boldsymbol{X}_{t_n}]$.

Notably, similar problems arise in the setting of training diffusion models from linearly corrupted data, i.e. when the available samples are $\boldsymbol{Y}_0 = A\boldsymbol{X}_0$, for a known matrix A, as considered in the Ambient Diffusion paper (Daras et al., 2023b). In this setting, the authors manage to learn $\mathbb{E}[\boldsymbol{X}_0|A\boldsymbol{X}_t]$ for all t, but not $\mathbb{E}[\boldsymbol{X}_0|\boldsymbol{X}_t]$, where $\boldsymbol{X}_t = \boldsymbol{X}_0 + \sigma_t Z$, as always in this paper. Similar challenges are encountered in the G-SURE paper (Kawar et al., 2023).

In sum, training exact diffusion models, i.e. diffusion models sampling the target distribution p_0, given corrupted data remains unsolved. In this paper, we resolve this open-problem with two key technical contributions: i) an efficiently computable objective for learning the optimal denoisers for all levels of noise $t : \sigma_t \geq \sigma_{t_n}$, obtained by applying Tweedie's formula twice; and ii) a consistency loss for learning the optimal denoisers for levels of noise $t : \sigma_t \leq \sigma_{t_n}$. We describe these contributions in the next section.

3. Method

3.1. Learning the Optimal Denoiser for $\sigma_t > \sigma_{t_n}$

We first present an efficiently computable objective that resembles Denoising Score Matching and enables learning the optimal denoisers for all noise levels $t : \sigma_t > \sigma_{t_n}$.

Theorem 3.1 (Ambient Denoising Score Matching). *Define $\boldsymbol{X}_t$ as in the beginning of Section 2. Suppose we are given samples $\boldsymbol{X}_{t_n} = \boldsymbol{X}_0 + \sigma_{t_n}\boldsymbol{Z}$, where $\boldsymbol{X}_0 \sim p_0$ and $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, I)$. Consider the following objective:*

$$\mathbb{E}_{\boldsymbol{x}_{t_n}}\mathbb{E}_{t\sim\mathcal{U}(t_n,T]}\mathbb{E}_{\boldsymbol{x}_t=\boldsymbol{x}_{t_n}+\sqrt{\sigma_t^2-\sigma_{t_n}^2}\boldsymbol{\eta}}\left[\left|\left|\frac{\sigma_t^2-\sigma_{t_n}^2}{\sigma_t^2}\boldsymbol{h}_\theta(\boldsymbol{x}_t,t)+\frac{\sigma_{t_n}^2}{\sigma_t^2}\boldsymbol{x}_t-\boldsymbol{x}_{t_n}\right|\right|^2\right],$$

where $\boldsymbol{\eta}$ in the above is a standard Gaussian vector. Suppose that the family of functions $\{\boldsymbol{h}_\theta\}$ is rich enough to contain the minimizer of the above objective overall functions $\boldsymbol{h}(\boldsymbol{x}, t)$. Then the minimizer θ^* of J satisfies:

$$\boldsymbol{h}_{\theta^*}(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{X}_0|\boldsymbol{X}_t = \boldsymbol{x}_t], \quad \forall \boldsymbol{x}_t, t > t_n. \tag{3.1}$$

The theorem above states that we can estimate the best l_2^2 denoisers for all noise levels $t : \sigma_t > \sigma_{t_n}$ without ever seeing clean data from p_0 and using an efficiently computable objective that contains no divergence term.

Proof Overview. The central idea for this proof is to apply Tweedie's Formula twice, on appropriate random variables. We start by stating (a generalized version of) Tweedie's formula, the proof of which is given in the Appendix.

Lemma 3.2 (Generalized Tweedie's Formula). *Let:*

$$\boldsymbol{X}_t = \alpha_t\boldsymbol{X}_0 + \sigma_t\boldsymbol{Z}, \tag{3.2}$$

for $\boldsymbol{X}_0 \sim p_0$, $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, I)$, and some positive function α_t of t. Then,

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t) = \frac{\alpha_t\mathbb{E}[\boldsymbol{X}_0|\boldsymbol{X}_t = \boldsymbol{x}_t] - \boldsymbol{x}_t}{\sigma_t^2}. \tag{3.3}$$

For $t : \sigma_t > \sigma_{t_n}$, the R.V. $\boldsymbol{X}_t$ can be written in the following two equivalent ways:

$$\begin{cases} \boldsymbol{X}_t = \boldsymbol{X}_0 + \sigma_t\boldsymbol{Z} \\ \boldsymbol{X}_t = \boldsymbol{X}_{t_n} + \sqrt{\sigma_t^2 - \sigma_{t_n}^2}\boldsymbol{Z} \end{cases}. \tag{3.4}$$

By applying Tweedie's formula twice, we get two alternative expressions for the same score-function since the distribution remains the same, irrespectively of how we choose to express $\boldsymbol{X}_t$. By equating the two expressions for the score, we arrive at the following result:

$$\mathbb{E}[\boldsymbol{X}_{t_n}|\boldsymbol{X}_t = \boldsymbol{x}_t] = \frac{\sigma_t^2 - \sigma_{t_n}^2}{\sigma_t^2}\left(\mathbb{E}[\boldsymbol{X}_0|\boldsymbol{X}_t = \boldsymbol{x}_t] - \boldsymbol{x}_t\right) + \boldsymbol{x}_t.$$

We can train a network with denoising score matching to estimate $\mathbb{E}[\boldsymbol{X}_{t_n}|\boldsymbol{X}_t = \boldsymbol{x}_t]$ and hence we can use the above equation to obtain $\mathbb{E}[\boldsymbol{X}_0|\boldsymbol{X}_t = \boldsymbol{x}_t]$, as desired.

The method we propose is conceptually similar to Noisier2Noise (Moran et al., 2020) but instead of adding noise with a fixed magnitude to create further corrupted iterates, we consider a continuum of noise scales and we train the model jointly in a Denoising Score Matching fashion.

We underline that our method can be easily extended to the Variance Preserving (VP) (Song et al., 2020b) case, i.e. when the available data are $\boldsymbol{X}_{t_n} = \sqrt{1 - \sigma_{t_n}^2}\boldsymbol{X}_0 + \sigma_{t_n}\boldsymbol{Z}$. This is the setting for our Stable Diffusion finetuning experiments (see Section 4). For the sake of simplicity, we avoid these calculations in the main paper and we point the interested reader to the Appendix (see Theorem A.5).

Developing Intuition. A nice interpretation of our method is that it trains the network to predict the denoised image $\mathbb{E}[\boldsymbol{X}_0|\boldsymbol{X}_t = \boldsymbol{x}_t]$ by removing *additional noise* that we introduced to the given samples $\boldsymbol{x}_{t_n}$. This is similar to the idea of *further corruption* developed in Ambient Diffusion (Daras et al., 2023b). The way we create further noisy samples $\boldsymbol{x}_t$ given samples $\boldsymbol{x}_{t_n}$ has some high-level connections to DDRM (Kawar et al.) that reuses noise in the measurements to solve inverse problems with diffusion models.

3.2. Learning the Optimal Denoiser for $\sigma_t \leq \sigma_{t_n}$

Theorem 3.1 allows us to learn the optimal denoisers for $t : \sigma_t > \sigma_{t_n}$. However, to perform exact sampling we need to also learn $\mathbb{E}[\boldsymbol{X}_0|\boldsymbol{X}_t = \boldsymbol{x}_t]$ for $t : \sigma_t \leq \sigma_{t_n}$. We achieve this by *training the network to be consistent*.

Definition 3.3 (Consistent Denoiser (Daras et al., 2023a))**.** Let $p_\theta(\boldsymbol{x}_{t'}, t'|\boldsymbol{x}_t, t)$ be the density of the sample $\boldsymbol{X}_{t'}$ of the stochastic diffusion process of Equation 2.3 at time t' when initialized with $\boldsymbol{x}_t$ at time $t > t'$. The network $\boldsymbol{h}_\theta(\cdot, t)$ that drives the process is a *consistent denoiser* if:

$$\boldsymbol{h}_\theta(\boldsymbol{x}_t, t) = \mathbb{E}_{\boldsymbol{X}_{t'} \sim p_\theta(\boldsymbol{x}_{t'}, t'|\boldsymbol{x}_t, t)} \left[\boldsymbol{h}_\theta(\boldsymbol{X}_{t'}, t')\right]. \tag{3.5}$$

The concept of consistency was introduced by Daras et al. (2023a) as a way to reduce error propagation in diffusion sampling and improve performance. Here, we find a completely different use case: we use consistency to learn the optimal denoisers for levels below the noise level of the available data. We are now ready to state our main theorem.

Theorem 3.4 (Main Theorem (informal))**.** *Define $\boldsymbol{X}_t$ as in the beginning of Section 2. Suppose we are given samples $\boldsymbol{X}_{t_n} = \boldsymbol{X}_0 + \sigma_{t_n}\boldsymbol{Z}$, where $\boldsymbol{X}_0 \sim p_0$ and $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, I)$. Consider the following objective:*

$$\overbrace{\mathbb{E}_{\boldsymbol{x}_{t_n}} \mathbb{E}_{t \sim \mathcal{U}[t_n, T]} \mathbb{E}_{\boldsymbol{x}_t = \boldsymbol{x}_{t_n} + \sqrt{\sigma_t^2 - \sigma_{t_n}^2}\boldsymbol{\eta}} \left[\left|\left|\frac{\sigma_t^2 - \sigma_{t_n}^2}{\sigma_t^2}\boldsymbol{h}_\theta(\boldsymbol{x}_t, t) + \frac{\sigma_{t_n}^2}{\sigma_t^2}\boldsymbol{x}_t - \boldsymbol{x}_{t_n}\right|\right|^2\right]}^{\text{Ambient Score Matching}}$$
$$\underbrace{+ \mathbb{E}_{t \sim \mathcal{U}(t_n, T], t' \sim \mathcal{U}(\epsilon, t), t'' \sim \mathcal{U}(t' - \epsilon, t')} \mathbb{E}_{\boldsymbol{x}_t} \mathbb{E}_{\boldsymbol{x}_{t'}|\boldsymbol{x}_t} \left[\left|\left|\boldsymbol{h}_\theta(\boldsymbol{x}_{t'}, t') - \mathbb{E}_{\boldsymbol{x}_{t''} \sim p_\theta(\boldsymbol{x}_{t''}, t''|\boldsymbol{x}_{t'}, t')}\left[\boldsymbol{h}_\theta(\boldsymbol{x}_{t''}, t'')\right]\right|\right|^2\right]}_{\text{Consistency Loss}}, \tag{3.6}$$

where $\boldsymbol{\eta}$ in the above is a standard Gaussian vector. Suppose that the family of functions $\{\boldsymbol{h}_\theta\}$ is rich enough to contain the minimizer of the above objective overall functions $\boldsymbol{h}(\boldsymbol{x}, t)$. Then the minimizer θ^* satisfies:

$$\boldsymbol{h}_{\theta^*}(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{X}_0|\boldsymbol{X}_t = \boldsymbol{x}_t], \quad \forall \boldsymbol{x}_t, t. \tag{3.7}$$

The formal statement and the proof of this Theorem is given in the Appendix (see Theorem A.6).

Intuition and Proof Overview. It is useful to build some intuition about how this objective works. There are two terms in the loss: i) the Ambient Score Matching term and ii) the Consistency Loss. The Ambient Score Matching term regards only noise levels $t : \sigma_t > \sigma_{t_n}$. Per Theorem 3.1, this term has a unique minimizer that is the optimal denoiser for all levels $t : \sigma_t > \sigma_{t_n}$. The consistency term in the loss, penalizes for violations of the Consistency Property (see Definition 3.3) for all pairs of times t, t'. The desired solution, $\boldsymbol{h}(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{X}_0|\boldsymbol{X}_t = \boldsymbol{x}_t]$, $\forall t, \boldsymbol{x}_t$, minimizes the first term and makes the second term 0, since it corresponds to a consistent denoiser. Hence, the desired solution is an optimal solution for the objective we wrote and the question becomes whether this solution is unique. The uniqueness of the solution arises from the Fokker-Planck PDE that describes the evolution of density: there is unique extension to a function that is $\mathbb{E}[\boldsymbol{X}_0|\boldsymbol{X}_t = \boldsymbol{x}_t]$, $t : \sigma_t > \sigma_{t_n}$ and is consistent for all t. The latter result comes from Theorem 3.2 in Consistent Diffusion Models (Daras et al., 2023a).

Implementation Trade-offs and Design Choices. When it comes to implementing the Consistency Loss there are trade-offs that need to be considered. First, we need to run partially the sampling chain. Doing so at every training step can lead to important slow-downs, as explained in Daras et al. (2023a). To mitigate this, we choose the times t', t'' to be very close to one another, as in Consistent Diffusion, using a uniform distribution with support of width ϵ. This helps us run only 1 step of the sampling chain (without introducing big discretization errors) and it works because local consistency implies global consistency. Second, for the inner-term in the consistency loss we need to compute an expectation over samples of p_θ. To avoid running the sampling chain many times during training, we opt for an unbiased estimator of this term that uses only two samples, following the implementation of Daras et al. (2023a). Specifically, we use the approximation:

$$\left|\left|\boldsymbol{h}_\theta(\boldsymbol{x}_{t'}, t') - \mathbb{E}_{\boldsymbol{x}_{t''} \sim p_\theta(\boldsymbol{x}_{t''}, t''|\boldsymbol{x}_{t'}, t')}\left[\boldsymbol{h}_\theta(\boldsymbol{x}_{t''}, t')\right]\right|\right|^2$$
$$\approx (\boldsymbol{h}_\theta(\boldsymbol{x}_{t''}^1, t'') - \boldsymbol{h}_\theta(\boldsymbol{x}_{t'}, t'))^T(\boldsymbol{h}_\theta(\boldsymbol{x}_{t''}^2, t'') - \boldsymbol{h}_\theta(\boldsymbol{x}_{t'}, t'')),$$

where $\boldsymbol{x}_{t''}^1, \boldsymbol{x}_{t''}^2$ are samples from $p_\theta(\cdot|\boldsymbol{x}_t', t')$. We finally note that our Consistency Loss defined in Eq. 3.6 involves three expectations, instead of two, as in the original definition of Daras et al. (2023a). This is because the consistency property needs to hold for all pairs of times (t', t'') and for $t' > t_n$ we don't have direct access to samples, i.e. we have to use the model to sample them given $\boldsymbol{x}_t$.

3.3. Testing Training Data Replication

Learning from corrupted data is a potential mitigation strategy for the problem of training data replication. Thus, we need effective ways to evaluate the degree to which our models (and baselines trained on clean data) memorize.

A standard approach is to generate a few thousand samples with the trained models (potentially using the dataset prompts) and then measure the similarities of the generated samples with their nearest neighbors in the dataset (Somepalli et al., 2022; Daras et al., 2023b). This approach is known to "systematically underestimate the amount of replication in Stable Diffusion and other models", as noted by Somepalli et al. (2022).

We propose a novel attack that shows that diffusion models memorize their training sets at a higher rate than previously known. We use the trained diffusion priors to solve inverse problems at extremely high corruption levels and we show that the reconstructions are often almost perfect as long as the uncorrupted images belong to the training set.

4. Experimental Evaluation

4.1. Experiments with pre-trained models

In this section, we measure how much pre-trained foundation diffusion models memorize data from their training set. We perform our experiments with Stable Diffusion XL (Podell et al., 2023) (SDXL), as it is the state-of-the-art open-source image generation diffusion model.

We take a random $10,000$ image subset of LAION and we corrupt it severely. We consider two models of corruption. In the first model, we take the LAION images and mask significant portions of them, as shown in Figure 1. The masked regions are selected automatically, using a YOLO object detection network (Redmon et al., 2016), to contain whole faces or large objects that are impossible to perfectly predict by only observing the non-masked content of the image. Yet, as seen in the last row of Figure 1, some posterior samples are almost pixel-perfect matches of the original images. This strongly indicates that the images in the top row of Figure 1 were in the training set of SDXL and have been memorized. Its important to note that the captions (from the LAION dataset) are entered as input in the inpainting model and this attack did not seem to work with null captions.

In the second corruption model, we encode LAION images with the SDXL encoder and we add a significant amount of noise to them. In Figure 2, we show images from LAION dataset, their encodings (visualizing them as 3-channel RGB images), the noisy latents, the MMSE reconstruction (using the model's one-step prediction at the noise level of the corruption) and posterior samples from the model. Again, even if the corruption is severe and the MMSE denoised images are very blurry, the posterior samples from the model are very close to the original images from the dataset, indicating potential memorization. In this corruption model the near-duplicate reconstructed images were obtained with null captions, so no text guidance was needed.

To quantify the degree of memorization and detect replication automatically, we adapt the methodology of Somepalli et al. (2022). In this work, the authors embed both the generated images and the dataset images to the DINOv2 (Oquab et al., 2023) latent space, and for each generated image compute its maximum inner product (similarity score) with its nearest neighbor in the dataset. We repeat this experiment, previously done for Stable Diffusion v1.4, for the latest SDXL model. We empirically find that similarities above 0.95 correspond to almost identical samples to the ones in the training set and similarities above 0.9 correspond to close matches. We compare the distribution obtained using the Somepalli et al. (2022) method with the distribution obtained using our noising approach (for two different noise levels) in Figure 3. As shown, our approach finds significantly more examples that have similarity values close to 1. Also as mentioned, our attack did not need the prompts in this case. This is not necessarily surprising since our approach uses more information (the noisy latents) compared to the previously proposed method that only uses the prompts. Still, our results present evidence that diffusion models memorize significantly more training data compared to what was previously known. For the inpainting case, we only compute embeddings for the infilled regions and hence the similarity numbers are not directly comparable. We present these results in Figure 9 in the Appendix.

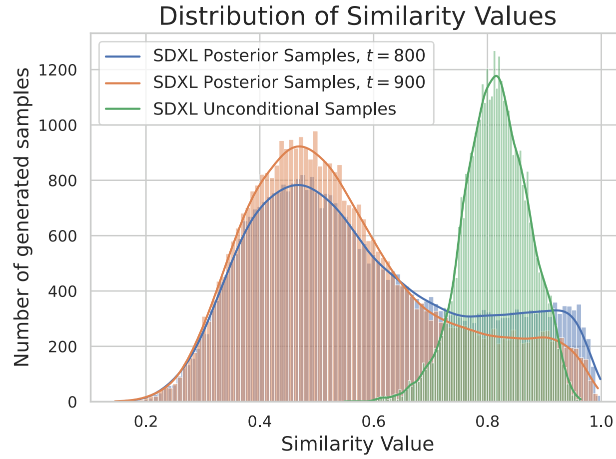


Figure 3. Distribution of image similarities of generated images with their nearest neighbors in the dataset for: i) the Somepalli et al. (2022) method, and ii) for our noising method for two different noise levels. As shown, the fraction of images with similarities above 0.95 (near-identical to training set) is much higher for our method compared to the Somepalli et al. (2022) baseline.

4.2. Finetuning Stable Diffusion XL

The next step is to use our framework, detailed in Sections 3.1, 3.2, to finetune SDXL on corrupted data. We finetune our models on FFHQ, at 1024×1024 resolution, since it is a standard benchmark for image generation. Given that SDXL is a latent model, we first encode the clean images using the SDXL encoder and then we add noise to the latents. We consider four noise levels which we will be referring to as: i) noiseless, $t_n = 0$, $\sigma_{t_n} = 0$, ii) low-noise, $t_n = 100$, $\sigma_{t_n} = 0.325$, iii) medium-noise, $t_n = 500$, $\sigma_{t_n} = 0.850$, and, iv) high-noise, $t_n = 800$, $\sigma_{t_n} = 0.981$. For reference, we fix an image from the training set and we visualize posterior samples for each one of the noise levels in Figure 5. We train models with our Ambient Denoising Score Matching loss, with and without consistency. We provide the training details in the Appendix, Section B.

We first evaluate the denoising performance of our models. To do so, we take 32 evaluation samples from FFHQ, we

Noise Level Eval	Noise Level Training	Latent MSE	Pixel MSE
900	100	**0.3369** (0.0521)	0.0802 (0.0257)
	500	0.3377 (0.0514)	0.0804 (0.0253)
	800	0.3375 (0.0514)	**0.0796** (0.0254)
800	100	**0.2974** (0.0463)	**0.0566** (0.0219)
	500	0.2978 (0.0464)	0.0566 (0.0222)
	800	0.3001 (0.0466)	0.0570 (0.0220)
500	100	**0.2153** (0.0283)	**0.0219** (0.0092)
	500	0.2159 (0.0283)	0.0221 (0.0092)
	800	0.2182 (0.0284)	0.0226 (0.0094)
100	100	**0.0405** (0.0029)	**0.0068** (0.0027)
	500	0.0409 (0.0029)	0.0069 (0.0028)
	800	0.0411 (0.0029)	0.0070 (0.0028)

Table 1. Restoration performance of models trained with noisy data at different noise levels. All the models have comparable performance, irrespective of the noise level of the dataset they were trained with.

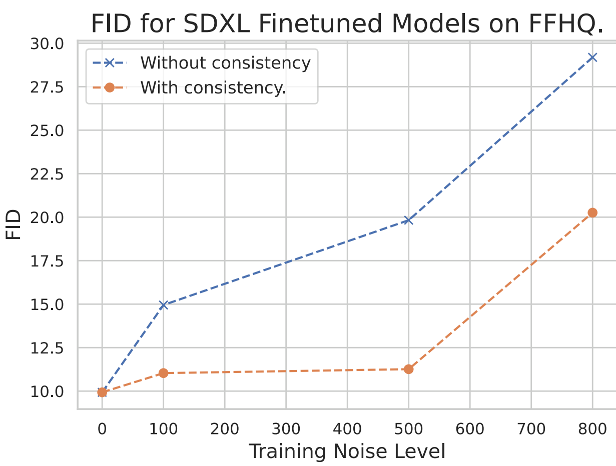
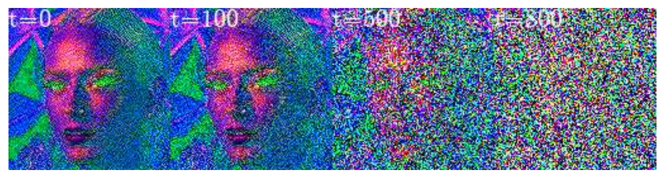


Figure 4. FID results for SDXL finetuned models, with and without consistency, on FFHQ, as we change the corruption level. The performance of models trained without consistency deteriorates significantly as we increase the corruption. Models trained with consistency maintain comparable performance to the baseline model (trained on clean data) for noise levels up to $t_n = 500$.



(a) Noisy Latents for $t_n \in \{0, 100, 500, 800\}$.

(b) Posterior samples for $t_n \in \{0, 100, 500, 800\}$.

Figure 5. Visualization of the noise levels considered in the paper. The top row shows noisy latents, visualized as RGB images. The bottom row shows posterior samples obtained by the SDXL (Podell et al., 2023) model given these noise latents.

add noise to levels $t_{\text{eval}} \in \{900, 800, 500, 100\}$, we use our trained models to denoise and we measure the reconstruction error. Since SDXL is a latent diffusion model, the noise (and the denoising) happens in the latent space. Hence, the MSE reconstruction error can be measured directly in the latent space or pixel space (by decoding the reconstructed latents). We present our results in Table 1. As shown, all the models have comparable performance across all noise levels, irrespective of the noise level of the data they saw during training. This is in line with our theory: all the models are trained to estimate $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ for all levels t.

To understand better the role of consistency, we visualize unconditional samples from our models trained with and without consistency in Figure 6. As shown in the left column of Figure 6, models trained without consistency lead to increasingly blurry generations as the level of noise encountered during training increases. This is not surprising: as explained in Subsection 2.3, models trained without consistency sample from the distribution of MMSE denoised images, $\mathbb{E}[\boldsymbol{X}_0|\boldsymbol{X}_{t_n}]$. As the noise level t_n increases, these images become averaged and high-frequency detail is lost. As shown in the right column on Figure 6, training with consistency recovers high-frequency details and leads to significantly improved images, especially for models trained with highly noisy data ($t_n \in \{500, 800\}$).

We proceed to evaluate unconditional generation performance. For each of our models, we generate $50,000$ images and we compute the FID score. We visualize the performance of our models trained with and without consistency in Figure 4. As shown, the performance of models trained

To further show that our method can be used for data that follow a distribution significantly different to the training one, we finetune SDXL on a dataset of chest x-rays. In Figure 11, we provide same samples of the training dataset (Row 1), generated samples without fine-tuning (Row 2), noisy samples that were used to fine-tune the model (Row 3), generated samples after fine-tuning without consistency (Row 4) and finally generated samples after fine-tuning with consistency. For all our generations, we use prompts from the dataset of interest. The generations of the model without fine-tuning are very different compared to the dataset samples, hinting that the model initially models a different distribution conditioned on the given prompt. After fine-tuning with noisy data, the generated samples are more closely related to the samples from the dataset. As we also observed in the rest of the experiments in this paper, consistency decreases the blurriness of the generated samples.

4.4. Measuring Memorization of Finetuned Models

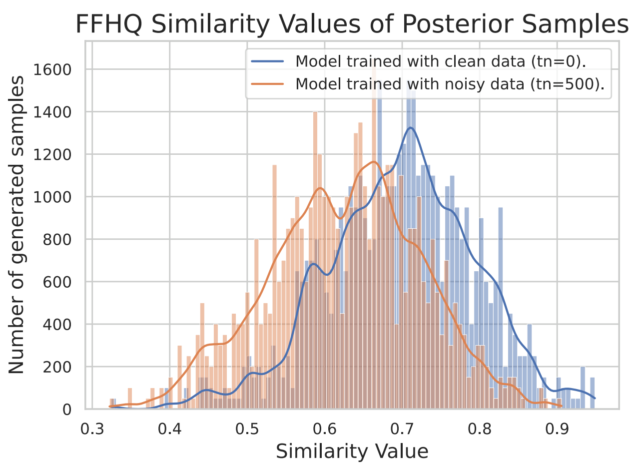


Figure 8. Distribution of similarities of posterior samples to their nearest neighbor in the dataset, given noisy latents (at $t = 900$) for two models. The model trained with clean data (blue curve), has a distribution of similarity values that is more shifted to the right, indicating higher dataset memorization compared to the model trained with corrupted data (orange curve).

The final step in our experimental evaluation is to investigate to what extent training with noisy data reduced the rate of training data replication. To do so, we use the method we proposed in Section 4.1. Specifically, we get the FFHQ training images, we encode them to the latent space of the SDXL Encoder and we add noise to them that corresponds to $t_n = 900$. We then use the model trained with clean images and the model trained with data at $t_n = 500$ noise level to perform posterior sampling, given the noisy latents. For each generated sample, we measure its DINO similarity to the nearest neighbor in the dataset. We plot the resulting distributions for the model trained with clean data and the $t_n = 500$ model in Figure 8. As shown, the model trained with clean data (blue curve), has a distribution of similarity values that is more shifted to the right, indicating higher dataset memorization compared to the model trained with corrupted data (orange curve). Finally, we once again compare with the method of Somepalli et al. (2022) for identifying training data replications. We use the model trained with clean data, we take the $50,000$ images that we used for FID generation and we compute their similarity to their nearest neighbor in the dataset. We compare with our approach in Figure 10 in the Appendix.

5. Discussion and Other Related Work

The concurrent work of Lu et al. (2024) shows that an adversary can "disguise" copyrighted images in the training set. The implication is that training dataset inspection is not enough to detect whether copyrighted images have been used. This is finding conveys a similar message to our work since the training set might contain (severely corrupted) copyrighted images and pure inspection of the noisy images is not enough to determine if that's the case. Finally, we underline that the use of consistency enables sampling images below the observed data noise level, solving an open problem in the space of learning from corrupted data. For a more detailed exposition of diffusion models and consistency, we refer the interested reader to the relevant works of Daras et al. (2023a); De Bortoli et al. (2024); Boffi & Vanden-Eijnden (2023); Shen et al. (2022); Albergo et al. (2023); Lai et al. (2023b;a).

6. Conclusions, Limitations and Future Work

We presented the first exact framework for training diffusion models to sample from an uncorrupted distribution using access to noisy data. We used our framework to finetune SDXL and we showed that training with corrupted data reduces memorization of the training set, while maintaining competitive performance. Our method has several limitations. First, it does not solve the problem of training diffusion models with *linearly* corrupted data that provably sample from the uncorrupted distribution. Second, training with consistency increases the training time (Daras et al., 2023a). Finally, in some preliminary experiments on very limited datasets (< 100 samples), the proposed Ambient Denoising Score Matching objective did not work. We plan to explore all these exciting open directions in future work.

Acknowledgements

This research has been supported by NSF Grants AF 1901292, CNS 2148141, Tripods CCF 1934932, IFML CCF 2019844 and research gifts by Western Digital, Amazon, WNCG IAP, UT Austin Machine Learning Lab (MLL), Cisco and the Stanly P. Finch Centennial Professorship in Engineering. Constantinos Daskalakis has been supported by NSF Awards CCF-1901292, DMS-2022448 and DMS-2134108, a Simons Investigator Award, and the Simons Collaboration on the Theory of Algorithmic Fairness. Giannis Daras has been supported by the Onassis Fellowship (Scholarship ID: F ZS 012-1/2022-2023), the Bodossaki Fellowship and the Leventis Fellowship. Giannis Daras would also like to thank Yuval Dagan and Valentin De Bortoli for useful discussions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Aali, A., Arvinte, M., Kumar, S., and Tamir, J. I. Solving inverse problems with score-based generative priors learned from noisy data. *arXiv preprint arXiv:2305.01166*, 2023.

Aggarwal, H. K., Pramanik, A., John, M., and Jacob, M. Ensure: A general approach for unsupervised training of deep image reconstruction algorithms. *IEEE Transactions on Medical Imaging*, 42(4):1133–1144, 2022.

Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Batson, J. and Royer, L. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pp. 524–533. PMLR, 2019.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2:3, 2023.

Boffi, N. M. and Vanden-Eijnden, E. Probability flow solution of the fokker–planck equation. *Machine Learning: Science and Technology*, 4(3):035012, 2023.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Daras, G. and Dimakis, A. Solving inverse problems with ambient diffusion. In *NeurIPS 2023 Workshop on Deep Learning and Inverse Problems*, 2023.

Daras, G., Dagan, Y., Dimakis, A. G., and Daskalakis, C. Consistent diffusion models: Mitigating sampling drift by learning to be consistent. *arXiv preprint arXiv:2302.09057*, 2023a.

Daras, G., Shah, K., Dagan, Y., Gollakota, A., Dimakis, A., and Klivans, A. Ambient diffusion: Learning clean distributions from corrupted data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=wBJBLy9kBY.

De Bortoli, V., Hutchinson, M., Wirnsberger, P., and Doucet, A. Target score matching. *arXiv preprint arXiv:2402.08667*, 2024.

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*.

Kawar, B., Elata, N., Michaeli, T., and Elad, M. Gsure-based diffusion model training with corrupted data. *arXiv preprint arXiv:2305.13128*, 2023.

Krull, A., Buchholz, T.-O., and Jug, F. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2129–2137, 2019.

Lai, C.-H., Takida, Y., Murata, N., Uesaka, T., Mitsufuji, Y., and Ermon, S. Fp-diffusion: Improving score-based diffusion models by enforcing the underlying score fokker-planck equation. In *International Conference on Machine Learning*, pp. 18365–18398. PMLR, 2023a.

Lai, C.-H., Takida, Y., Uesaka, T., Murata, N., Mitsufuji, Y., and Ermon, S. On the equivalence of consistency-type models: Consistency models, consistent diffusion models, and fokker-planck regularization. *arXiv preprint arXiv:2306.00367*, 2023b.

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., and Aila, T. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018.

Lin, Y. Y., Gao, A. F., and Bouman, K. L. Imaging an evolving black hole by leveraging shared structure.

Lu, Y., Yang, M. Y., Liu, Z., Kamath, G., and Yu, Y. Disguised copyright infringement of latent diffusion model. *arXiv preprint arXiv:2404.06737*, 2024.

Metzler, C. A., Mousavi, A., Heckel, R., and Baraniuk, R. G. Unsupervised learning with stein's unbiased risk estimator. *arXiv preprint arXiv:1805.10531*, 2018.

Moran, N., Schmidt, D., Zhong, Y., and Coady, P. Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12064–12072, 2020.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Pang, T., Zheng, H., Quan, Y., and Ji, H. Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2043–2052, 2021.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

Shen, Z., Wang, Z., Kale, S., Ribeiro, A., Karbasi, A., and Hassani, H. Self-consistency of the fokker planck equation. In *Conference on Learning Theory*, pp. 817–841. PMLR, 2022.

Soltanayev, S. and Chun, S. Y. Training deep learning based denoisers without ground truth data. *Advances in neural information processing systems*, 31, 2018.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. *arXiv preprint arXiv:2212.03860*, 2022.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. *arXiv preprint arXiv:2305.20086*, 2023.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Stein, C. M. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pp. 1135–1151, 1981.

The Event Horizon Telescope Collaboration. First m87 event horizon telescope results. iv. imaging the central supermassive black hole. *The Astrophysical Journal Letters*, 875(1):L4, apr 2019. doi: 10.3847/2041-8213/ab0e85. URL https://dx.doi.org/10.3847/2041-8213/ab0e85.

Xiang, T., Yurt, M., Syed, A. B., Setsompop, K., and Chaudhari, A. Ddm^2: Self-supervised diffusion mri denoising with generative diffusion models. *arXiv preprint arXiv:2302.03018*, 2023.

Xu, J., Huang, Y., Cheng, M.-M., Liu, L., Zhu, F., Xu, Z., and Shao, L. Noisy-as-clean: Learning self-supervised denoising from corrupted image. *IEEE Transactions on Image Processing*, 29:9316–9329, 2020.