Ambient Diffusion: Learning Clean Distributions from Corrupted Data

Giannis Daras
UT Austin
giannisdaras@utexas.edu

Kulin Shah
UT Austin
kulinshah@utexas.edu

Yuval Dagan
UC Berkeley
yuvald@berkeley.edu

Aravind Gollakota
UT Austin
aravindg@cs.utexas.edu

Alexandros G. Dimakis
UT Austin
dimakis@austin.utexas.edu

Adam Klivans
UT Austin
klivans@utexas.edu

Abstract

We present the first diffusion-based framework that can learn an unknown distribution using only highly-corrupted samples. This problem arises in scientific applications where access to uncorrupted samples is impossible or expensive to acquire. Another benefit of our approach is the ability to train generative models that are less likely to memorize individual training samples since they never observe clean training data. Our main idea is to introduce *additional measurement distortion* during the diffusion process and require the model to predict the original corrupted image from the further corrupted image. We prove that our method leads to models that learn the conditional expectation of the full uncorrupted image given this additional measurement corruption. This holds for any corruption process that satisfies some technical conditions (and in particular includes inpainting and compressed sensing). We train models on standard benchmarks (CelebA, CIFAR-10 and AFHQ) and show that we can learn the distribution even when all the training samples have 90% of their pixels missing. We also show that we can finetune foundation models on small corrupted datasets (e.g. MRI scans with block corruptions) and learn the clean distribution without memorizing the training set.

1 Introduction

Diffusion generative models [48, 24, 51] are emerging as versatile and powerful frameworks for learning high-dimensional distributions and solving inverse problems [34, 11, 35, 28]. Numerous recent developments [52, 30] have led to text conditional foundation models like Dalle-2 [41], Latent Diffusion [45] and Imagen [47] with incredible performance in general image domains. Training these models requires access to high-quality datasets which may be expensive or impossible to obtain. For example, direct images of black holes cannot be observed [12, 19] and high-quality MRI images require long scanning times, causing patient discomfort and motion artifacts [28].

Recently, Carlini et al. [8], Somepalli et al. [49], and Jagielski et al. [27] showed that diffusion models can memorize examples from their training set. Further, an adversary can extract dataset samples given only query access to the model, leading to privacy, security and copyright concerns. For many applications, we may want to learn the distribution but not individual training images e.g. we might want to learn the distribution of X-ray scans but not memorize images of specific patient scans from the dataset. Hence, we may want to introduce corruption as a design choice. We show that it is possible to train diffusions that learn a distribution of clean data by only observing highly corrupted samples.

Preprint. Work in progress.

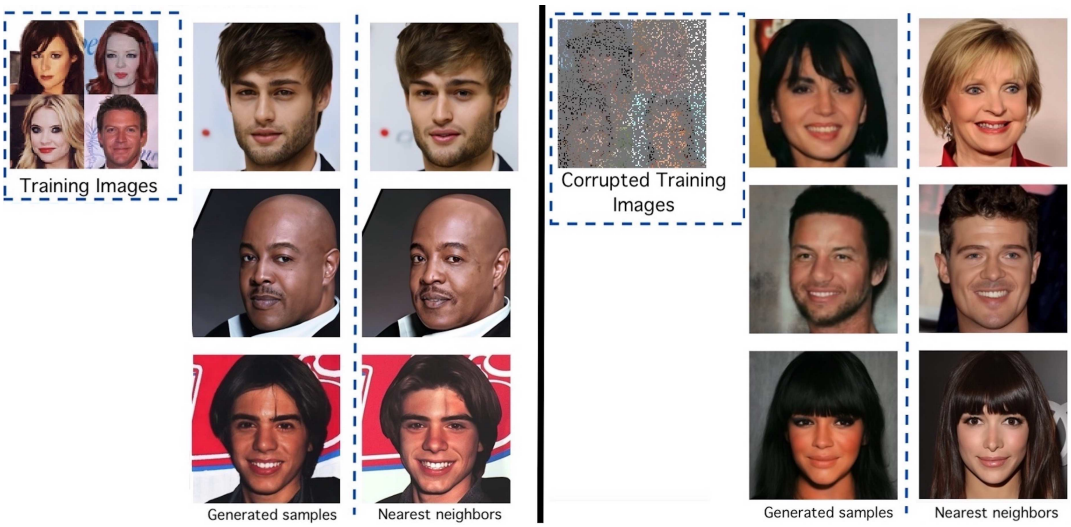

Figure 1: **Left panel:** Baseline method of vanilla finetuning Deepfloyd IF using 3000 images from CelebA-HQ. We show generated sample images and nearest neighbors from the finetuning set. As shown, the generated samples are often near-identical copies from training data. This verifies related work Carlini et al. [8], Somepalli et al. [49], and Jagielski et al. [27] that pointed out that diffusions often generate training samples. **Right panel:** We finetune the same foundation model (Deepfloyd IF) using our method and 3000 highly corrupted training images. The corruption adds noise and removes 80 percent random pixels. We show generated samples and nearest neighbors from the training set. Our method still learns the clean distribution of faces (with some quality deterioration, as shown) but does not memorize training data. We emphasize that our training is performed without ever accessing clean training data.

Prior work in supervised learning from corrupted data. The traditional approach to solving such problems involves training a restoration model using supervised learning to predict the clean image based on the measurements [43, 44, 57, 39]. The seminal Noise2Noise [38] work introduced a practical algorithm for learning how to denoise in the absence of any non-noisy images. This framework and its generalizations [5, 37, 53] have found applications in electron microscopy [16], tomographic image reconstruction [56], fluorescence image reconstruction [59], blind inverse problems [20, 5], monocular depth estimation and proteomics [6]. Another related line of work uses Stein's Unbiased Risk Estimate (SURE) to optimize an unbiased estimator of the denoising objective without access to non-noisy data [18]. We stress that the aforementioned research works study the problem of *restoration*, whereas are interested in the problem of *sampling* from the clean distribution. Restoration algorithms based on supervised learning are only effective when the corruption level is relatively low [15]. However, it might be either not possible or not desirable to reconstruct individual samples. Instead, the desired goal may be to learn to *generate* fresh and completely unseen samples from the distribution of the uncorrupted data but *without reconstructing individual training samples*.

Indeed, for certain corruption processes, it is theoretically possible to perfectly learn a distribution only from highly corrupted samples (such as just random one-dimensional projections), even though individual sample denoising is usually impossible in such settings. Specifically, AmbientGAN [7] showed that general d dimensional distributions can be learned from *scalar* observations, by observing only projections on one-dimensional random Gaussian vectors, in the infinite training data limit. The theory requires an infinitely powerful discriminator and hence does not apply to diffusion models.

Our contributions. We present the first diffusion-based framework to learn an unknown distribution $\mathcal{D}$ when the training set only contains highly-corrupted examples drawn from $\mathcal{D}$. Specifically, we consider the problem of learning to sample from the target distribution $p_0(\boldsymbol{x}_0)$ given corrupted samples $A\boldsymbol{x}_0$ where $A \sim p(A)$ is a random corruption matrix (with known realizations and prior distribution) and $\boldsymbol{x}_0 \sim p_0(\boldsymbol{x}_0)$. Our main idea is to introduce *additional measurement distortion* during the diffusion process and require the model to predict the original corrupted image from the further corrupted image.

Vincent [55] showed that we can learn the score function at level t by optimizing for the score-matching objective:

$$J(\theta) = \frac{1}{2}\mathbb{E}_{(\boldsymbol{x}_0,\boldsymbol{x}_t)} \left|\left|\boldsymbol{h}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{x}_0\right|\right|^2. \tag{2.1}$$

Specifically, the score function can be written in terms of the minimizer of this objective as:

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) = \frac{\boldsymbol{h}_{\theta^*}(\boldsymbol{x}_t, t) - \boldsymbol{x}_t}{\sigma_t}. \tag{2.2}$$

This result reveals a fundamental connection between the score-function and the best restoration model of $\boldsymbol{x}_0$ given $\boldsymbol{x}_t$, known as Tweedie's Formula [17]. Specifically, the optimal $\boldsymbol{h}_{\theta^*}(\boldsymbol{x}_t, t)$ is given by $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$, which means that

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) = \frac{\overbrace{\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]}^{\text{best restoration}} - \boldsymbol{x}_t}{\sigma_t}. \tag{2.3}$$

Inspired by this restoration interpretation of diffusion models, the Soft/Cold Diffusion works [14, 4] generalized diffusion models to look at non-Markovian corruption processes: $\boldsymbol{x}_t = C_t\boldsymbol{x}_0 + \sigma_t\boldsymbol{\eta}$. Specifically, Soft Diffusion proposes the Soft Score Matching objective:

$$J_{\text{soft}}(\theta) = \frac{1}{2}\mathbb{E}_{(\boldsymbol{x}_0,\boldsymbol{x}_t)} \left|\left|C_t\left(\boldsymbol{h}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{x}_0\right)\right|\right|^2, \tag{2.4}$$

and shows that it is sufficient to recover the score function via a generalized Tweedie's Formula:

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) = \frac{C_t\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] - \boldsymbol{x}_t}{\sigma_t}. \tag{2.5}$$

For these generalized models, the matrix C_t is a design choice (similar to how we could choose the functions $\boldsymbol{f}, g$). Most importantly, for $t = 0$, the matrix C_t becomes the identity matrix and the noise σ_t becomes zero, i.e. we observe samples from the true distribution.

3 Method

As explained in the introduction, in many cases we do not observe uncorrupted images $\boldsymbol{x}_0$, either by design (to avoid memorization and leaking of sensitive data) or because it is impossible to obtain clean data. Here we study the case where a learner only has access to linear measurements of the clean data, i.e. $\boldsymbol{y}_0 = A\boldsymbol{x}_0$, and the corruption matrices $A : \mathbb{R}^{m\times n}$. We note that we are interested in non-invertible corruption matrices. We ask two questions:

1. Is it possible to learn $\mathbb{E}[\boldsymbol{x}_0|A(\boldsymbol{x}_0 + \sigma_t\boldsymbol{\eta}), A]$ for all noise levels t, given only access to corrupted samples $(\boldsymbol{y}_0 = A\boldsymbol{x}_0, A)$?
2. If so, is it possible to use this restoration model $\mathbb{E}[\boldsymbol{x}_0|A(\boldsymbol{x}_0 + \sigma_t\boldsymbol{\eta}), A]$ to recover $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ for any noise level t, and thus sample from the true distribution through the score function as given by Tweedie's formula (Eq. 2.3)?

We investigate these questions in the rest of the paper. For the first, the answer is affirmative but only after introducing additional corruptions, as we explain below. For the second, at every time step t, we approximate $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ directly using $\mathbb{E}[\boldsymbol{x}_0|A\boldsymbol{x}_t, A]$ (for a chosen A) and substitute it into Eq. 2.3. Empirically, we observe that the resulting approximate sampler yields good results.

3.1 Training

For the sake of clarity, we first consider the case of random inpainting. If the image $\boldsymbol{x}_0$ is viewed as a vector, we can think of the matrix A as a diagonal matrix with ones in the entries that correspond to the preserved pixels and zeros in the erased pixels. We assume that $p(A)$ samples a matrix where each entry in the diagonal is sampled i.i.d. with a probability $1 - p$ to be 1 and p to be zero.

We would like to train a function $\boldsymbol{h}_\theta$ which receives a corruption matrix A and a noisy version of a corrupted image, $\boldsymbol{y}_t = A\underbrace{(\boldsymbol{x}_0 + \sigma_t\boldsymbol{\eta})}_{\boldsymbol{x}_t}$ where $\boldsymbol{\eta} \sim \mathcal{N}(0, I)$, and produces an estimate for the

conditional expectation. The simplest idea would be to simply ignore the missing pixels and optimize for:

$$J^{\text{corr}}_{\text{naive}}(\theta) = \frac{1}{2}\mathbb{E}_{(\boldsymbol{x}_0,\boldsymbol{x}_t,A)} \left|\left|A\left(\boldsymbol{h}_\theta(A, A\boldsymbol{x}_t, t) - \boldsymbol{x}_0\right)\right|\right|^2, \tag{3.1}$$

Despite the similarities with Soft Score Matching (Eq 2.4), this objective will not learn the conditional expectation. The reason is that the learner is never penalized for performing arbitrarily poorly in the missing pixels. Formally, any function $\boldsymbol{h}_{\theta'}$ satisfying $A\boldsymbol{h}_{\theta'}(A, \boldsymbol{y}_t, t) = A\mathbb{E}[\boldsymbol{x}_0|A\boldsymbol{x}_t, A]$ is a minimizer.

Instead, we propose to *further corrupt* the samples before feeding them to the model, and ask the model to predict the original corrupted sample from the further corrupted image.

Concretely, we randomly corrupt A to obtain $\tilde{A} = BA$ for some matrix B that is selected randomly given A. In our example of missing pixels, $\tilde{A}$ is obtained from A by randomly erasing an additional fraction δ of the pixels that survive after the corruption A. Here, B will be diagonal where each element is 1 with probability $1-\delta$ and 0 w.p. δ. We will penalize the model on recovering all the pixels that are visible in the sample $A\boldsymbol{x}_0$: this includes both the pixels that survive in $\tilde{A}\boldsymbol{x}_0$ and those that are erased by $\tilde{A}$. The formal training objective is given by minimizing the following loss:

$$J^{\text{corr}}(\theta) = \frac{1}{2}\mathbb{E}_{(\boldsymbol{x}_0,\boldsymbol{x}_t,A,\tilde{A})} \left|\left|A\left(\boldsymbol{h}_\theta(\tilde{A}, \tilde{A}\boldsymbol{x}_t, t) - \boldsymbol{x}_0\right)\right|\right|^2, \tag{3.2}$$

The key idea behind our algorithm is as follows: the learner does not know if a missing pixel is missing because we never had it (and hence do not know the ground truth) or because it was deliberately erased as part of the further corruption (in which case we do know the ground truth). Thus, the best learner cannot be inaccurate in the unobserved pixels because with non-zero probability it might be evaluated on some of them. Notice that the trained model behaves as a denoiser in the observed pixels and as an inpainter in the missing pixels. We also want to emphasize that the probability δ of further corruption can be arbitrarily small as long as it stays positive.

The idea of further corruption can be generalized from the case of random inpainting to a much broader family of corruption processes. For example, if A is a random Gaussian matrix with m rows, we can form $\tilde{A}$ by deleting one row from A at random. If A is a block inpainting matrix (i.e. a random block of fixed size is missing from all of the training images), we can create $\tilde{A}$ by corrupting further with one more non-overlapping missing block. Examples of our further corruption are shown in Figure 2. In our Theory Section, we prove conditions under which it is possible to recover $\mathbb{E}[\boldsymbol{x}_0|\tilde{A}\boldsymbol{x}_t, \tilde{A}]$ using our algorithm and samples $(\boldsymbol{y}_0 = A\boldsymbol{x}_0, A)$. Our goal is to satisfy this condition while adding minimal further corruption, i.e. while keeping $\tilde{A}$ close to A.

3.2 Sampling

Fixed mask sampling. To sample from $p_0(\boldsymbol{x}_0)$ using the standard diffusion formulation, we need access to $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$, which is equivalent to having access to $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ (see Eq. 2.3). Instead, our model is trained to predict $\mathbb{E}[\boldsymbol{x}_0|\tilde{A}\boldsymbol{x}_t, \tilde{A}]$ for all matrices A in the support of $p(A)$.

We note that for random inpainting, the identity matrix is technically in the support of $p(A)$. However, if the corruption probability p is at least a constant, the probability of seeing the identity matrix is exponentially small in the dimension of $\boldsymbol{x}_t$. Hence, we should not expect our model to give good estimates of $\mathbb{E}[\boldsymbol{x}_0|\tilde{A}\boldsymbol{x}_t, \tilde{A}]$ for corruption matrices A that belong to the tails of the distribution $p(A)$.

The simplest idea is to sample a mask $\tilde{A} \sim p(\tilde{A})$ and approximate $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ with $\mathbb{E}[\boldsymbol{x}_0|\tilde{A}\boldsymbol{x}_t, \tilde{A}]$. Under this approximation, the discretized sampling rule becomes:

$$\boldsymbol{x}_{t-\Delta t} = \underbrace{\frac{\sigma_{t-\Delta t}}{\sigma_t}}_{\gamma_t}\boldsymbol{x}_t + \underbrace{\frac{\sigma_t - \sigma_{t-\Delta t}}{\sigma_t}}_{1-\gamma_t}\underbrace{\mathbb{E}[\boldsymbol{x}_0|\tilde{A}\boldsymbol{x}_t, \tilde{A}]}_{\hat{x}_0}. \tag{3.3}$$

This idea works surprisingly well. Unless mentioned otherwise, we use it for all the experiments in the main paper and we show that we can generate samples that are reasonably close to the true distribution (as shown by metrics such as FID and Inception) even with 90% of the pixels missing.

Sampling with Reconstruction Guidance. In the Fixed Mask Sampler, at any time t, the prediction is a convex combination of the current value and the predicted denoised image. As $t \to 0$, $\gamma_t \to 0$. Hence, for the masked pixels, the fixed mask sampler outputs the conditional expectation of their value given the observed pixels. This leads to averaging effects as the corruption gets higher. To correct this problem, we add one more term in the update: the Reconstruction Guidance term. The issue with the previous sampler is that the model never sees certain pixels. We would like to evaluate the model using different masks. However, the model outputs for the denoised image might be very different when evaluated with different masks. To account for this problem, we add an additional term that enforces updates that lead to consistency on the reconstructed image. The update of the sampler with Reconstruction Guidance becomes:

$$\boldsymbol{x}_{t-\Delta t} = \gamma_t \boldsymbol{x}_t + (1-\gamma_t)\mathbb{E}[\boldsymbol{x}_0|\tilde{A}\boldsymbol{x}_t, \tilde{A}] - w_t \nabla_{\boldsymbol{x}_t} \mathbb{E}_{A'} ||\mathbb{E}[\boldsymbol{x}_0|\tilde{A}\boldsymbol{x}_t, \tilde{A}] - \mathbb{E}[\boldsymbol{x}_0|\tilde{A}'\boldsymbol{x}_t, \tilde{A}']||^2. \quad (3.4)$$

This sampler is inspired by the Reconstruction Guidance term used in Imagen [23] to enforce consistency and correct for the sampling drift caused by imperfect score matching [13]. We see modest improvements over the Fixed Mask Sampler for certain corruption ranges. We ablate this sampler in the Appendix, Section E.3.

In the Appendix, Section A.1, we also prove that in theory, whenever it is possible to reconstruct $p_0(\boldsymbol{x}_0)$ from corrupted samples, it is also possible to reconstruct it using access to $\mathbb{E}[\boldsymbol{x}_0|A\boldsymbol{x}_t, A]$. However, as stated in the Limitations section, we were not able to find any practical algorithm to do so.

4 Theory

As elaborated in Section 3, one of our key goals is to learn the best restoration model for the measurements at all noise levels, i.e., the function $\boldsymbol{h}(A, \boldsymbol{y}_t, t) = \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{y}_t, A]$. We now show that under a certain assumption on the distribution of A and $\tilde{A}$, the true population minimizer of Eq. 3.2 is indeed essentially of the form above. This assumption formalizes the notion that even conditional on $\tilde{A}$, A has considerable variability, and the latter ensures that the best way to predict $A\boldsymbol{x}_0$ as a function of $\tilde{A}\boldsymbol{x}_t$ and $\tilde{A}$ is to optimally predict $\boldsymbol{x}_0$ itself. All proofs are deferred to the Appendix.

Theorem 4.1. *Assume a joint distribution of corruption matrices A and further corruption $\tilde{A}$. If for all $\tilde{A}$ in the support it holds that $\mathbb{E}_{A|\tilde{A}}[A^T A]$ is full-rank, then the unique minimizer of the objective in equation 3.2 is given by*

$$\boldsymbol{h}_{\theta^*}(\tilde{A}, \boldsymbol{y}_t, t) = \mathbb{E}[\boldsymbol{x}_0 \mid \tilde{A}\boldsymbol{x}_t, \tilde{A}] \quad (4.1)$$

Two simple examples that fit into this framework (see Corollaries A.1 and A.2 in the Appendix) are:

- Inpainting: $A \in \mathbb{R}^{n\times n}$ is a diagonal matrix where each entry $A_{ii} \sim \mathrm{Ber}(1-p)$ for some $p > 0$ (independently for each i), and the additional noise is generated by drawing $\tilde{A}|A$ such that $\tilde{A}_{ii} = A_{ii} \cdot \mathrm{Ber}(1-\delta)$ for some small $\delta > 0$ (again independently for each i).[1]
- Gaussian measurements: $A \in \mathbb{R}^{m\times n}$ consists of m rows drawn independently from $\mathcal{N}(0, I_n)$, and $\tilde{A} \in \mathbb{R}^{m\times n}$ is constructed conditional on A by zeroing out its last row.

Notice that the minimizer in Eq 4.1 is not entirely of the form we originally desired, which was $\boldsymbol{h}(A, \boldsymbol{y}_t, t) = \mathbb{E}[\boldsymbol{x}_0 \mid A\boldsymbol{x}_t, A]$. In place of A, we now have $\tilde{A}$, which is a further degraded matrix. Indeed, one trivial way to satisfy the condition in Theorem 4.1 is by forming $\tilde{A}$ completely independently of A, e.g. by always setting $\tilde{A} = 0$. However, in this case, the function we learn is not very useful. For this reason, we would like to add as little further noise as possible and ensure that $\tilde{A}$ is close to A. In natural noise models such as the inpainting noise model, by letting the additional corruption probability δ approach 0, we can indeed ensure that $\tilde{A}$ follows a distribution very close to that of A.

[1] $\mathrm{Ber}(q)$ indicates a Bernoulli random variable with a probability of q to equal 1 and $1-q$ for 0.

5 Experimental Evaluation

5.1 Training from scratch on corrupted data

Our first experiment is to train diffusion models from scratch using corrupted training data at different levels of corruption. The corruption model we use for these experiments is random inpainting: we form our dataset by deleting each pixel with probability p. To create the matrix $\tilde{A}$, we further delete each row of A with probability δ – this removes an additional δ-fraction of the surviving pixels. Unless mentioned otherwise, we use $\delta = 0.1$. We train models on CIFAR-10, AFHQ, and CelebA-HQ. All our models are trained with corruption level $p \in \{0.0, 0.2, 0.4, 0.6, 0.8, 0.9\}$. We use the EDM [30] codebase to train our models. We replace convolutions with Gated Convolutions [58] which are known to perform better for inpainting-type problems. To use the mask $\tilde{A}$ as an additional input to the model, we simply concatenate it with the image $\boldsymbol{x}$. The full training details can be found in the Appendix, Section C.

Dataset	Corruption Probability	Method	LPIPS	PSNR	NFE
CelebA-HQ	0.6	Ours	**0.037**	**31.51**	1
		DPS	0.053	28.21	100
		DDRM	0.139	25.76	35
			0.088	27.38	99
			0.069	28.16	199
	0.8	Ours	**0.084**	**26.80**	1
		DPS	0.107	24.16	100
		DDRM	0.316	20.37	35
			0.188	22.96	99
			0.153	23.82	199
	0.9	Ours	**0.152**	**23.34**	1
		DPS	0.168	20.89	100
		DDRM	0.461	15.87	35
			0.332	18.74	99
			0.242	20.14	199
AFHQ	0.4	Ours	0.030	33.27	1
		DPS	**0.020**	**34.06**	100
		DDRM	0.122	25.18	35
			0.091	26.42	99
			0.088	26.52	199
	0.6	Ours	0.062	29.46	1
		DPS	**0.051**	**30.03**	100
		DDRM	0.246	20.76	35
			0.166	22.79	99
			0.160	22.93	199
	0.8	Ours	0.124	**25.37**	1
		DPS	**0.107**	25.30	100
		DDRM	0.525	14.56	35
			0.295	18.08	99
			0.258	18.86	199

Table 1: Comparison of our model (trained on corrupted data) with state-of-the-art diffusion models on CelebA (DDIM [50] model) and AFHQ (EDM [30] model) for solving the random inpainting inverse problem. Our model performs on par with state-of-the-art diffusion inverse problem solvers, even though it has never seen uncorrupted training data. Further, this is achieved with a single score function evaluation. To solve this problem with a standard pre-trained diffusion model we need to use a reconstruction algorithm (such as DPS [11] or DDRM [34]) that typically requires hundreds of steps.

We first evaluate the restoration performance of our model for the task it was trained on (random inpainting and noise). We compare with state-of-the-art diffusion models that were trained on clean data. Specifically, for AFHQ we compare with the state-of-the-art EDM model [30] and for

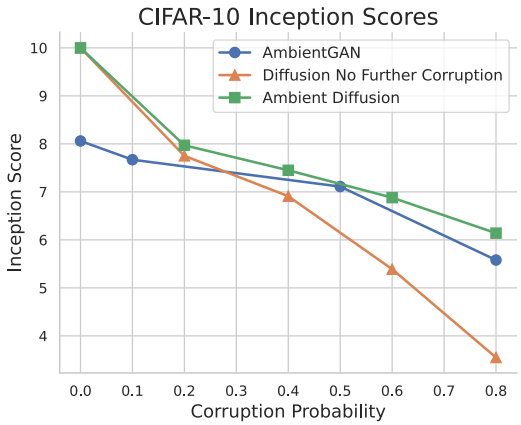

Figure 3: Performance on CIFAR-10 as a function of the corruption level. We compare our method with a diffusion model trained without our further corruption trick and AmbientGAN [7]. Ambient Diffusion outperforms both baselines for all ranges of corruption levels.

Dataset	Corruption Probability	FID	Inception Score
CelebA-HQ	0.0	3.26	N/A
	0.2	4.18	
	0.6	6.08	
	0.8	11.19	
	0.9	25.53	
AFHQ	0.0	2.41	N/A
	0.2	4.47	
	0.4	6.96	
	0.6	10.11	
	0.8	16.78	
	0.9	41.00	
CIFAR-10	0.0	1.85	9.94
	0.2	11.70	7.97
	0.4	18.85	7.45
	0.6	28.88	6.88
	0.8	46.27	6.14

Figure 4: Inception/FID results on random inpainting for models trained with our algorithm on CelebA-HQ, AFHQ and CIFAR-10.

CelebA we compare with DDIM [50]. These models were not trained to denoise, but we can use the prior learned in the denoiser as in [54, 29] to solve any inverse problem. We experiment with the state-of-the-art reconstruction algorithms: DDRM [34] and DPS [11].

We summarize the results in Table 1. Our model performs similarly to other diffusion models, even though it has never been trained on clean data. Further, it does so by requiring only one step, while all the baseline diffusion models require hundreds of steps to solve the same task with inferior or comparable performance. The performance of DDRM improves with more function evaluations at the cost of more computation. For DPS, we did not observe significant improvement by increasing the number of steps to more than 100. We include results with noisy inpainted measurements and comparisons with a supervised method in the Appendix, Section E, Tables 3, 4. We want to emphasize that all the baselines we compare against have an advantage: they are trained on *uncorrupted* data. Instead, our models were only trained on corrupted data. This experiment indicates that: i) our training algorithm for learning the conditional expectation worked and ii) that the choice of corruption that diffusion models are trained to reverse matters for solving inverse problems.

Next, we evaluate the performance of our diffusion models as generative models. To the best of our knowledge, the only generative baseline with quantitative results for training on corrupted data is AmbientGAN [7] which is trained on CIFAR-10. We further compare with a diffusion model trained without our further corruption algorithm. We plot the results in Figure 3. The diffusion model trained without our further corruption algorithm performs well for low corruption levels but collapses entirely for high corruption. Instead, our model trained with further corruption maintains reasonable corruption scores even for high corruption levels, outperforming the previous state-of-the-art AmbientGAN for all ranges of corruption levels.

For CelebA-HQ and AFHQ we could not find any generative baselines trained on corrupted data to compare against. Nevertheless, we report FID and Inception Scores and summarize our results in Table 4 to encourage further research in this area. As shown in the Table, for CelebA-HQ and AFHQ, we manage to maintain a decent FID score even with 90% of the pixels deleted. For CIFAR-10, the performance degrades faster, potentially because of the lower resolution of the training images.

5.2 Finetuning foundation models on corrupted data

We can apply our technique to finetune a foundational diffusion model. For all our experiments, we use Deepfloyd's IF model [2], which is one of the most powerful open-source diffusion generative models available. We choose this model over Stable Diffusion [46] because it works in the pixel space (and hence our algorithm directly applies).

To quantify the memorization, we follow the methodology of Somepalli et al. [49]. Specifically, we generate 10000 images from each model and we use DINO [9]-v2 [42] to compute top-1 similarity to the training images. Results are shown in Figure 6. Similarity values above 0.95 roughly correspond to the same person while similarities below 0.75 typically correspond to random faces. The standard finetuning (Red) often generates images that are near-identical with the training set. Instead, finetuning with corrupted samples (blue) shows a clear shift to the left. Visually we never observed a near-copy generated from our process – see also Figure 1.

We repeat this experiment for models trained on the full CelebA dataset and at different levels of corruption. We include the results in Figure 8 of the Appendix. As shown, the more we increase the corruption level the more the distribution of similarities shifts to the left, indicating less memorization. However, this comes at the cost of decreased performance, as reported in Table 4.

New domains and different corruption. We show that we can also finetune a pre-trained foundation model on a *new domain* given a limited-sized dataset in a few hours in a single GPU. Figure 5 shows generated samples from a finetuned model on a dataset containing 155 examples of brain tumor MRI images [26]. As shown, the model learns the statistics of full brain tumor MRI images while only trained on brain-tumor images that have a random box obfuscating 25% of the image. The training set was resized to 64×64 but the generated images are at 256×256 by simply leveraging the power of the cascaded Deepfloyd IF.

Limitations. Our work has several limitations. First, there is a tradeoff between generator quality and corruption levels. For higher corruption, it is less likely that our generator memorizes parts of training examples, but at a cost of degrading quality. Precisely characterizing this trade-off is an open research problem. Further, in this work, we only experimented with very simple approximation algorithms to estimate $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ using our trained models. Additionally, we cannot make any strict privacy claim about the protection of any training sample without making assumptions about the data distribution. We show in the Appendix that it is possible to recover $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ exactly using our restoration oracle, but we do not have an algorithm to do so. Finally, our method cannot handle measurements that also have noise. Future work could potentially address this limitation by exploiting SURE regularization as in [1].

Acknowledgements. The authors would like to thank Tom Goldstein for insightful discussions that benefited this work. This research has been supported by NSF Grants CCF 1763702, AF 1901292, CNS 2148141, Tripods CCF 1934932, NSF AI Institute for Foundations of Machine Learning (IFML) 2019844, the Texas Advanced Computing Center (TACC) and research gifts by Western Digital, WNCG IAP, UT Austin Machine Learning Lab (MLL), Cisco and the Archie Straiton Endowed Faculty Fellowship. Giannis Daras has been supported by the Onassis Fellowship (Scholarship ID: F ZS 012-1/2022-2023), the Bodossaki Fellowship and the Leventis Fellowship.

References

[1] Asad Aali, Marius Arvinte, Sidharth Kumar, and Jonathan I Tamir. "Solving Inverse Problems with Score-Based Generative Priors learned from Noisy Data". In: *arXiv preprint arXiv:2305.01166* (2023) (page 10).

[2] Stability AI. *Deepfloyd IF*. https://github.com/deep-floyd/IF. 2013 (pages 8, 17).

[3] Brian D.O. Anderson. "Reverse-time diffusion equation models". In: *Stochastic Processes and their Applications* 12.3 (1982), pp. 313–326 (page 3).

[4] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. "Cold diffusion: Inverting arbitrary image transforms without noise". In: *arXiv preprint arXiv:2208.09392* (2022) (page 4).

[5] Joshua Batson and Loic Royer. "Noise2self: Blind denoising by self-supervision". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 524–533 (page 2).

[6] Felix JB Bäuerlein and Wolfgang Baumeister. "Towards visual proteomics at high resolution". In: *Journal of Molecular Biology* 433.20 (2021), p. 167187 (page 2).

[7] Ashish Bora, Eric Price, and Alexandros G Dimakis. "AmbientGAN: Generative models from lossy measurements". In: *International conference on learning representations*. 2018 (pages 2, 3, 8).

[8] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. "Extracting Training Data from Large Language Models". In: *30th USENIX Security Symposium (USENIX Security 21)*. 2021, pp. 2633–2650. ISBN: 978-1-939133-24-3. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting (visited on 11/10/2022) (pages 1, 2, 9).

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660 (page 10).

[10] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. "Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions". In: *arXiv preprint arXiv:2209.11215* (2022) (pages 15, 16).

[11] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. "Diffusion Posterior Sampling for General Noisy Inverse Problems". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=OnD9zGAGT0k (pages 1, 7, 8, 20).

[12] The Event Horizon Telescope Collaboration et al. "First M87 Event Horizon Telescope Results. IV. Imaging the Central Supermassive Black Hole". In: *The Astrophysical Journal Letters* 875.1 (Apr. 2019), p. L4. DOI: 10.3847/2041-8213/ab0e85. URL: https://dx.doi.org/10.3847/2041-8213/ab0e85 (page 1).

[13] Giannis Daras, Yuval Dagan, Alexandros G Dimakis, and Constantinos Daskalakis. "Consistent diffusion models: Mitigating sampling drift by learning to be consistent". In: *arXiv preprint arXiv:2302.09057* (2023) (page 6).

[14] Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alex Dimakis, and Peyman Milanfar. "Soft Diffusion: Score Matching with General Corruptions". In: *Transactions on Machine Learning Research* (2023). ISSN: 2835-8856. URL: https://openreview.net/forum?id=W98rebBxlQ (pages 3, 4).

[15] Mauricio Delbracio and Peyman Milanfar. "Inversion by direct iteration: An alternative to denoising diffusion for image restoration". In: *arXiv preprint arXiv:2303.11435* (2023) (page 2).

[16] Jeffrey M Ede. "Deep learning in electron microscopy". In: *Machine Learning: Science and Technology* 2.1 (2021), p. 011004 (page 2).

[17] Bradley Efron. "Tweedie's formula and selection bias". In: *Journal of the American Statistical Association* 106.496 (2011), pp. 1602–1614 (page 4).

[18] Yonina C. Eldar. "Generalized SURE for Exponential Families: Applications to Regularization". In: *IEEE Transactions on Signal Processing* 57.2 (2009), pp. 471–481. DOI: 10.1109/TSP.2008.2008212 (page 2).

[19] Angela F Gao, Oscar Leong, He Sun, and Katherine L Bouman. "Image Reconstruction without Explicit Priors". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5 (page 1).

[20] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. "Toward convolutional blind denoising of real photographs". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1712–1722 (page 2).

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16000–16009 (page 18).

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in neural information processing systems* 30 (2017) (page 17).

[23] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. "Imagen video: High definition video generation with diffusion models". In: *arXiv preprint arXiv:2210.02303* (2022) (page 6).

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851 (pages 1, 17).

[25] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. "Cascaded Diffusion Models for High Fidelity Image Generation." In: *J. Mach. Learn. Res.* 23.47 (2022), pp. 1–33 (page 17).

[26] "Huggingface Brain Tumor MRI Dataset". In: (2020). URL: https://huggingface.co/datasets/miladfa7/Brain-MRI-Images-for-Brain-Tumor-Detection/ (pages 9, 10).

[27] Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Chiyuan Zhang. "Measuring Forgetting of Memorized Training Examples". In: *arxiv:2207.00099[cs]* (June 2022). DOI: 10.48550/arXiv.2207.00099. arXiv: 2207.00099 [cs]. URL: http://arxiv.org/abs/2207.00099 (visited on 11/10/2022) (pages 1, 2, 9).

[28] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. "Robust compressed sensing mri with deep generative priors". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 14938–14954 (page 1).

[29] Zahra Kadkhodaie and Eero P Simoncelli. "Solving linear inverse problems using the prior implicit in a denoiser". In: *arXiv preprint arXiv:2007.13640* (2020) (page 8).

[30] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. "Elucidating the design space of diffusion-based generative models". In: *arXiv preprint arXiv:2206.00364* (2022) (pages 1, 7, 16–18).

[31] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Alias-free generative adversarial networks". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 852–863 (page 16).

[32] Tero Karras, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4401–4410 (page 16).

[33] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and improving the image quality of stylegan". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8110–8119 (page 16).

[34] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. "Denoising Diffusion Restoration Models". In: *Advances in Neural Information Processing Systems* (pages 1, 7, 8, 20).

[35] Bahjat Kawar, Gregory Vaksman, and Michael Elad. "SNIPS: Solving noisy inverse problems stochastically". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 21757–21769 (page 1).

[36] Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. "Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 11201–11228 (page 16).

[37] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. "Noise2void-learning denoising from single noisy images". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2129–2137 (page 2).

[38] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. "Noise2Noise: Learning image restoration without clean data". In: *arXiv preprint arXiv:1803.04189* (2018) (page 2).

[39] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. "Coherent Semantic Attention for Image Inpainting". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Oct. 2019). DOI: 10.1109/iccv.2019.00427. URL: http://dx.doi.org/10.1109/ICCV.2019.00427 (page 2).

[40] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. "Pulse: Self-supervised photo upsampling via latent space exploration of generative models". In: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2020, pp. 2437–2445 (page 16).

[41] Alex Nichol, Aditya Ramesh, Pamela Mishkin, Prafulla Dariwal, Joanne Jang, and Mark Chen. *DALL·E 2 Pre-Training Mitigations*. June 2022. URL: https://openai.com/blog/dall-e-2-pre-training-mitigations/ (visited on 11/10/2022) (page 1).

[42] 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". In: *arXiv preprint arXiv:2304.07193* (2023) (page 10).

[43] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. "Context encoders: Feature learning by inpainting". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2536–2544 (page 2).

[44] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. "Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation". In: *arXiv preprint arXiv:2008.00951* (2020) (page 2).

[45] 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*. 2022, pp. 10684–10695 (page 1).

[46] 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*. 2022, pp. 10684–10695 (page 8).

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. "Photorealistic text-to-image diffusion models with deep language understanding". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 36479–36494 (page 1).

[48] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2256–2265 (page 1).

[49] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. "Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models". In: *arXiv preprint arXiv:2212.03860* (2022) (pages 1, 2, 9, 10).

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models". In: *arXiv preprint arXiv:2010.02502* (2020) (pages 7, 8, 18).

[51] Yang Song and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Advances in Neural Information Processing Systems* 32 (2019) (page 1).

[52] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. "Score-based generative modeling through stochastic differential equations". In: *arXiv preprint arXiv:2011.13456* (2020) (pages 1, 3, 17).

[53] Julián Tachella, Dongdong Chen, and Mike Davies. "Unsupervised Learning From Incomplete Measurements for Inverse Problems". In: *arXiv preprint arXiv:2201.12151* (2022) (page 2).

[54] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. "Plug-and-play priors for model based reconstruction". In: *2013 IEEE Global Conference on Signal and Information Processing*. IEEE. 2013, pp. 945–948 (page 8).

[55] Pascal Vincent. "A connection between score matching and denoising autoencoders". In: *Neural computation* 23.7 (2011), pp. 1661–1674 (page 4).

[56] Ge Wang, Jong Chul Ye, and Bruno De Man. "Deep learning for tomographic image reconstruction". In: *Nature Machine Intelligence* 2.12 (2020), pp. 737–748 (page 2).

[57] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. "Free-Form Image Inpainting With Gated Convolution". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Oct. 2019). DOI: 10.1109/iccv.2019.00457. URL: http://dx.doi.org/10.1109/ICCV.2019.00457 (page 2).

[58] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. "Free-form image inpainting with gated convolution". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 4471–4480 (pages 7, 16).

[59] Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smith, and Scott Howard. "A poisson-gaussian denoising dataset with real fluorescence microscopy images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 11710–11718 (page 2).