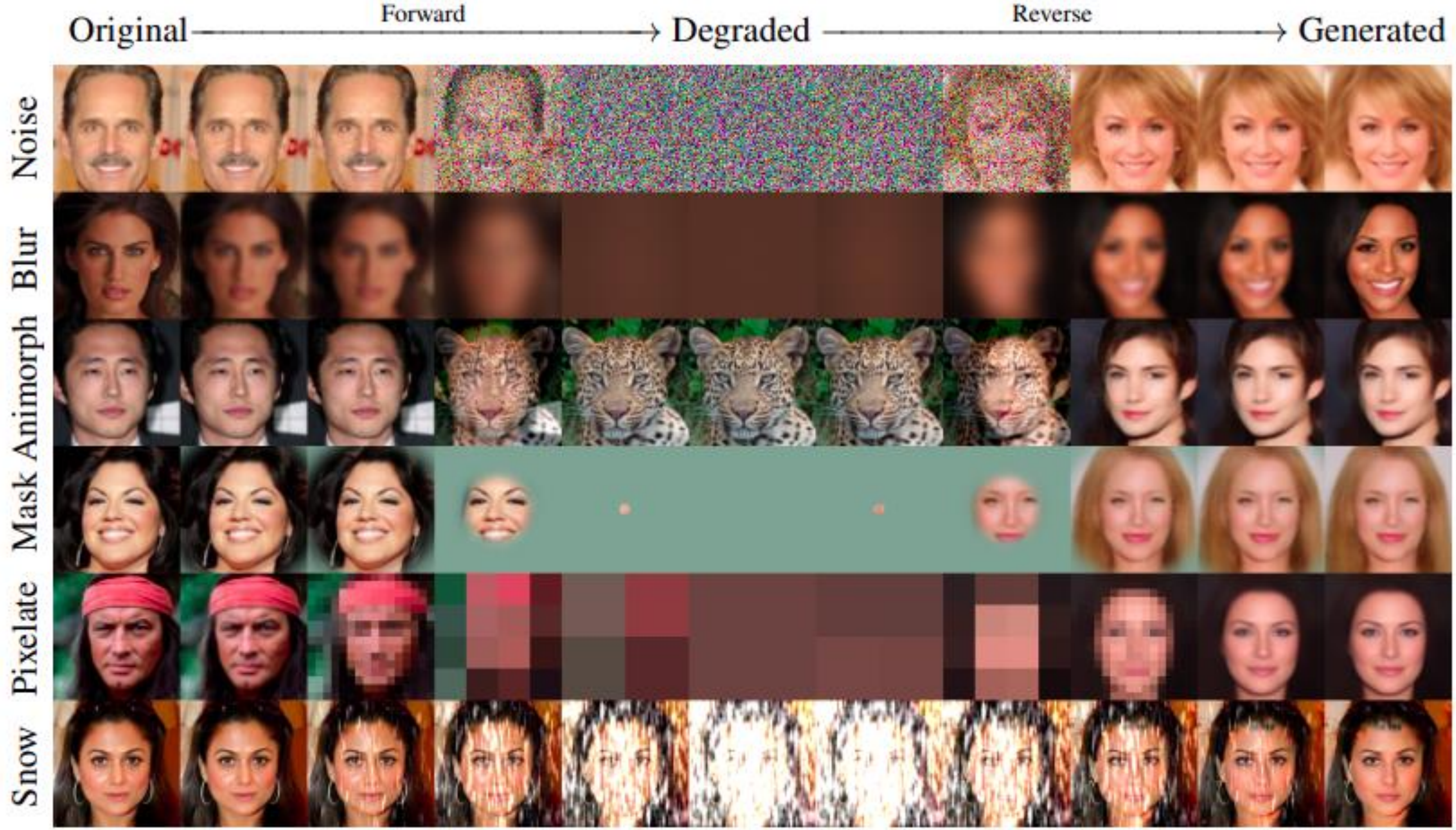
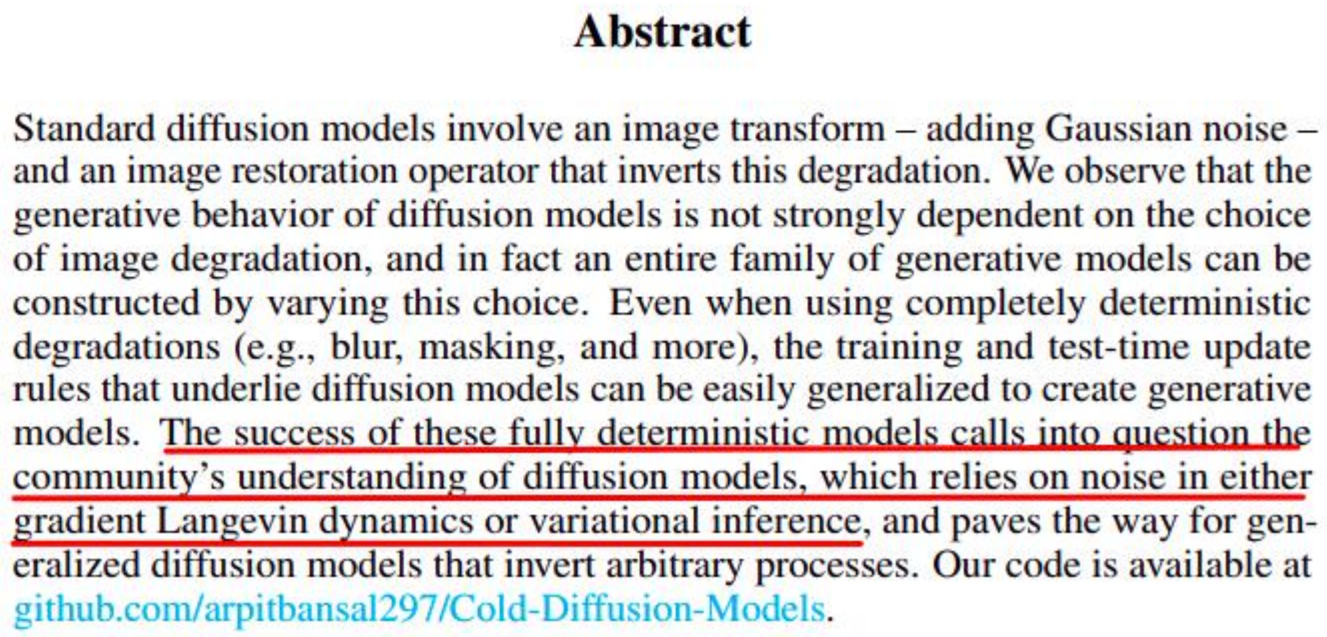
Cold Diffusion





Abstract

Standard diffusion models involve an image transform – adding Gaussian noise – and an image restoration operator that inverts this degradation. We observe that the generative behavior of diffusion models is not strongly dependent on the choice of image degradation, and in fact an entire family of generative models can be constructed by varying this choice. Even when using completely deterministic degradations (e.g., blur, masking, and more), the training and test-time update rules that underlie diffusion models can be easily generalized to create generative models. The success of these fully deterministic models calls into question the community's understanding of diffusion models, which relies on noise in either gradient Langevin dynamics or variational inference, and paves the way for generalized diffusion models that invert arbitrary processes. Our code is available at github.com/arpitbansal297/Cold-Diffusion-Models.

3.1 Model components and training

Given an image $x_0 \in \mathbb{R}^N$, consider the *degradation* of x_0 by operator D with severity t, denoted $x_t = D(x_0, t)$. The output distribution $D(x_0, t)$ of the degradation should vary continuously in t, and the operator should satisfy

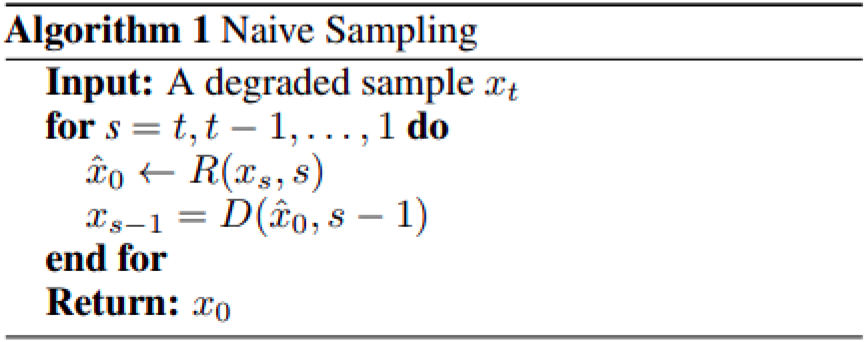
$$D(x_0, 0) = x_0.$$

We also require a *restoration* operator R that (approximately) inverts D. This operator has the property that

$$R(x_t, t) \approx x_0.$$

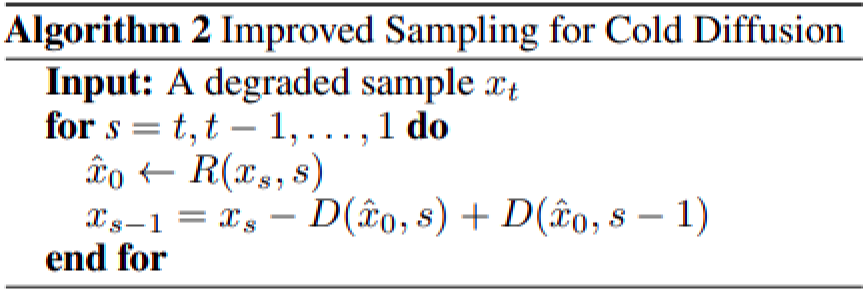
In practice, this operator is implemented via a neural network parameterized by θ. The restoration network is trained via the minimization problem

$$\min_{\theta} \mathbb{E}_{x \sim \mathcal{X}} \| R_\theta(D(x, t), t) - x \|, \tag{1}$$



Algorithm 1 Naive Sampling

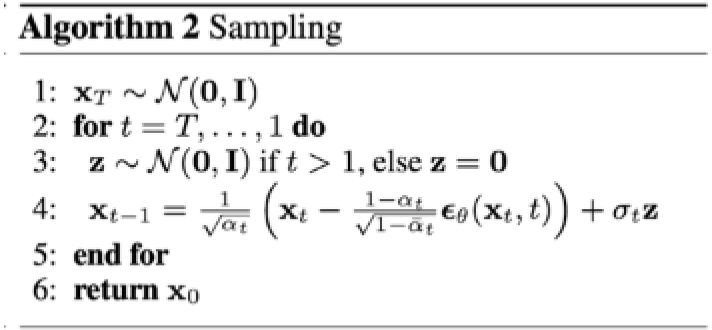
```
Input: A degraded sample $x_t$
for $s = t, t-1, \dots, 1$ do
    $\hat{x}_0 \leftarrow R(x_s, s)$
    $x_{s-1} = D(\hat{x}_0, s-1)$
end for
Return: $x_0$
```



Algorithm 2 Improved Sampling for Cold Diffusion

```
Input: A degraded sample $x_t$
for $s = t, t-1, \dots, 1$ do
    $\hat{x}_0 \leftarrow R(x_s, s)$
    $x_{s-1} = x_s - D(\hat{x}_0, s) + D(\hat{x}_0, s-1)$
end for
```

DDPM



Algorithm 2 Sampling

```
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \dots, 1$ do
    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
    $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$
```

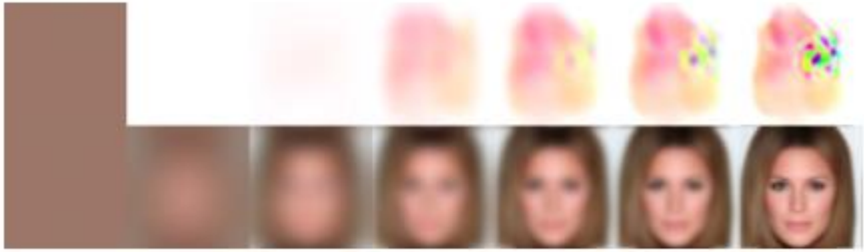


Figure 2: Comparison of sampling methods for cold diffusion on the CelebA dataset. **Top:** Algorithm 1 produces compounding artifacts and fails to generate a new image. **Bottom:** Algorithm 2 succeeds in sampling a high quality image without noise.

vector e. While this ansatz may seem rather restrictive, note that the Taylor expansion of any smooth degradation $D(x, s)$ around $x = x_0, s = 0$ has the form $D(x, s) \approx x + s \cdot e + \text{HOT}$ where HOT denotes higher order terms. Note that the constant/zeroth-order term in this Taylor expansion is zero because we assumed above that the degradation operator satisfies $D(x, 0) = x$.

For a degradation of the form (3.3) and any restoration operator R, the update in Algorithm 2 can be written

$$\begin{aligned} x_{s-1} &= x_s - D(R(x_s, s), s) + D(R(x_s, s), s-1) \\ &= D(x_0, s) - D(R(x_s, s), s) + D(R(x_s, s), s-1) \\ &= x_0 + s \cdot e - R(x_s, s) - s \cdot e + R(x_s, s) + (s-1) \cdot e \\ &= x_0 + (s-1) \cdot e \\ &= D(x_0, s-1) \end{aligned}$$

Table 1: Quantitative metrics for quality of image reconstruction using deblurring models.

	Degraded			**Sampled**			**Direct**		
Dataset	FID	SSIM	RMSE	FID	SSIM	RMSE	FID	SSIM	RMSE
MNIST	438.59	0.287	0.287	**4.69**	0.718	0.154	5.10	**0.757**	0.142
CIFAR-10	298.60	0.315	0.136	**80.08**	0.773	0.075	83.69	**0.775**	0.071
CelebA	382.81	0.254	0.193	**26.14**	0.568	0.093	36.37	**0.607**	0.083

Table 2: Quantitative metrics for quality of image reconstruction using inpainting models.

	Degraded			**Sampled**			**Direct**		
Dataset	FID	SSIM	RMSE	FID	SSIM	RMSE	FID	SSIM	RMSE
MNIST	108.48	0.490	0.262	**1.61**	0.941	0.068	2.24	**0.948**	0.060
CIFAR-10	40.83	0.615	0.143	**8.92**	0.859	0.068	9.97	**0.869**	0.063
CelebA	127.85	0.663	0.155	**5.73**	0.917	0.043	7.74	**0.922**	0.039

Table 3: Quantitative metrics for quality of image reconstruction using super-resolution models.

	Degraded			**Sampled**			**Direct**		
Dataset	FID	SSIM	RMSE	FID	SSIM	RMSE	FID	SSIM	RMSE
MNIST	368.56	0.178	0.231	4.33	0.820	0.115	**4.05**	**0.823**	0.114
CIFAR-10	358.99	0.279	0.146	**152.76**	0.411	0.155	169.94	**0.420**	0.152
CelebA	349.85	0.335	0.225	**96.92**	0.381	0.201	112.84	**0.400**	0.196

5.1 Generation using deterministic noise degradation

Here we discuss image generation using noise-based degradation. We consider "deterministic" sampling in which the noise pattern is selected and frozen at the start of the generation process, and then treated as a constant. We study two ways of applying Algorithm 2 with fixed noise. We first define

$$D(x,t) = \sqrt{\alpha_t}x + \sqrt{1-\alpha_t}z,$$

as the (deterministic) interpolation between data point x and a fixed noise pattern $z \in \mathcal{N}(0,1)$, for increasing $\alpha_t < \alpha_{t-1}$, $\forall\, 1 \leq t \leq T$ as in Song et al. [2021a]. Algorithm 2 can be applied in this case by fixing the noise z used in the degradation operator $D(x,s)$. Alternatively, one can deterministically calculate the noise vector z to be used in step t of reconstruction by using the formula

$$\hat{z}(x_t,t) = \frac{x_t - \sqrt{\alpha_t}R(x_t,t)}{\sqrt{1-\alpha_t}}.$$

Table 4: Quantitative metrics for quality of image reconstruction using *desnowification* models.

	Degraded Image			**Reconstruction**		
Dataset	FID	SSIM	RMSE	FID	SSIM	RMSE
CIFAR-10	125.63	0.419	0.327	31.10	0.074	0.838
CelebA	398.31	0.338	0.283	27.09	0.033	0.907

Table 5: FID scores for CelebA and AFHQ datasets using hot (using noise) and cold diffusion (using blur transformation). This table shows that This table also shows that breaking the symmetry withing pixels of the same channel further improves the FID scores.

	Hot Diffusion		**Cold Diffusion**	
Dataset	Fixed Noise	Estimated Noise	Perfect symmetry	Broken symmetry
CelebA	59.91	23.11	97.00	49.45
AFHQ	25.62	20.59	93.05	54.68