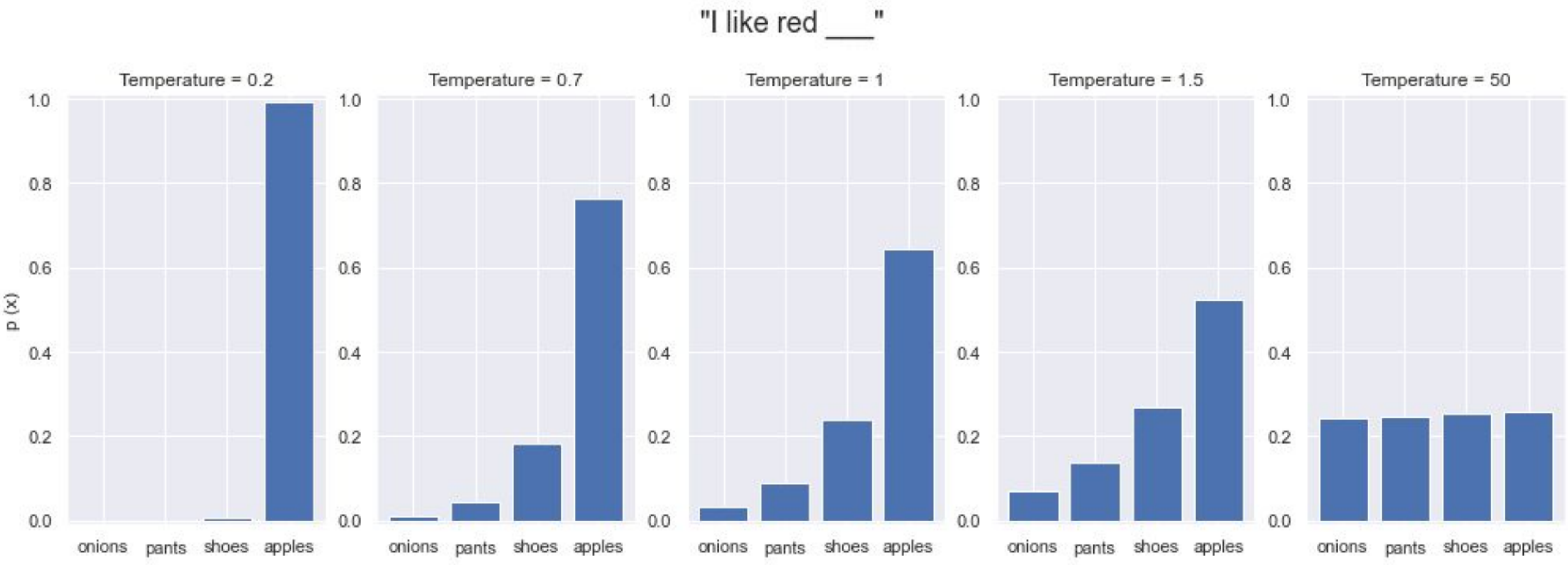
Classifier-Free Diffusion Guidance

Danil Sheshenya

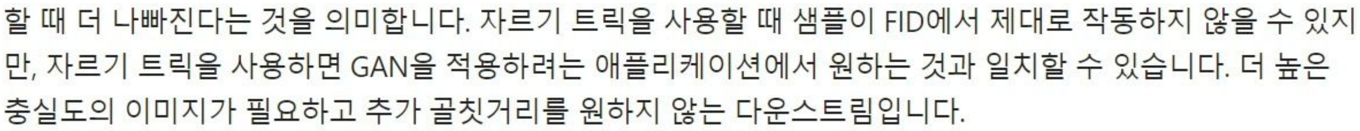
Problem

- We want to generate images with diffusion models based on some conditioning
- Individual samples in generative models can be not realistic enough
- What if we can trade off variety of samples for fidelity?

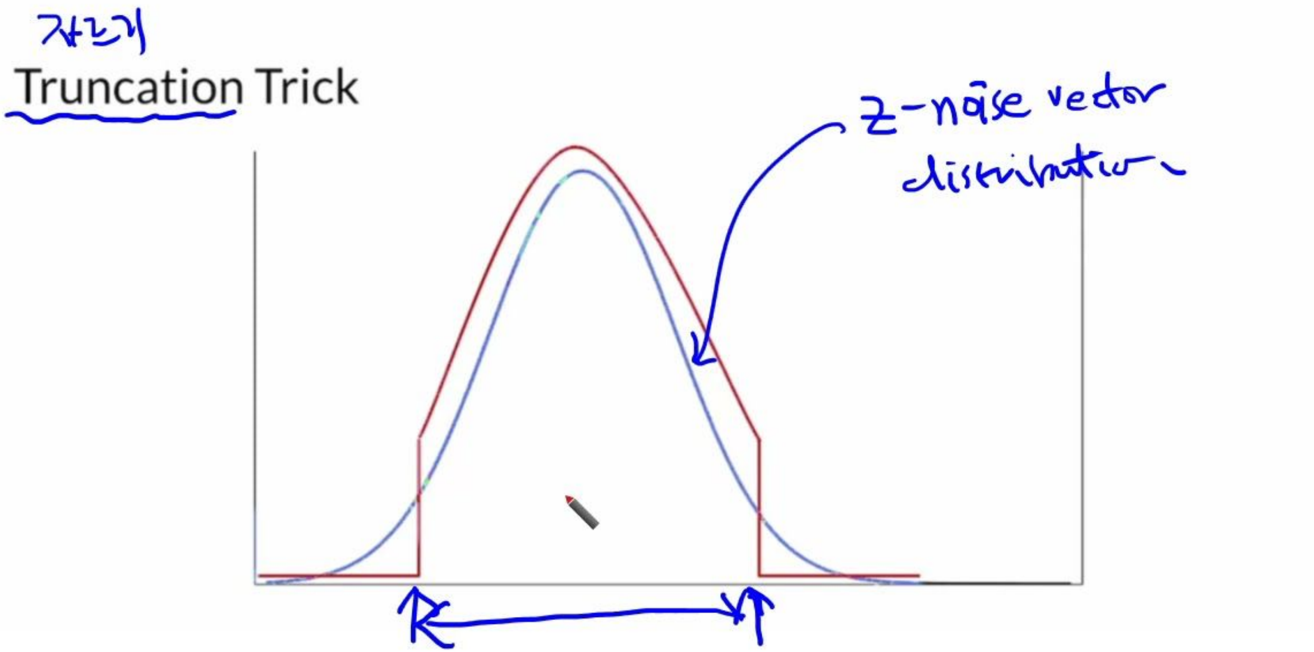
Low Temperature Sampling



Truncation Trick in GAN



할 때 더 나빠진다는 것을 의미합니다. 자르기 트릭을 사용할 때 샘플이 FID에서 제대로 작동하지 않을 수 있지만, 자르기 트릭을 사용하면 GAN을 적용하려는 애플리케이션에서 원하는 것과 일치할 수 있습니다. 더 높은 충실도의 이미지가 필요하고 추가 골칫거리를 원하지 않는 다운스트림입니다.



Truncation Trick in GAN

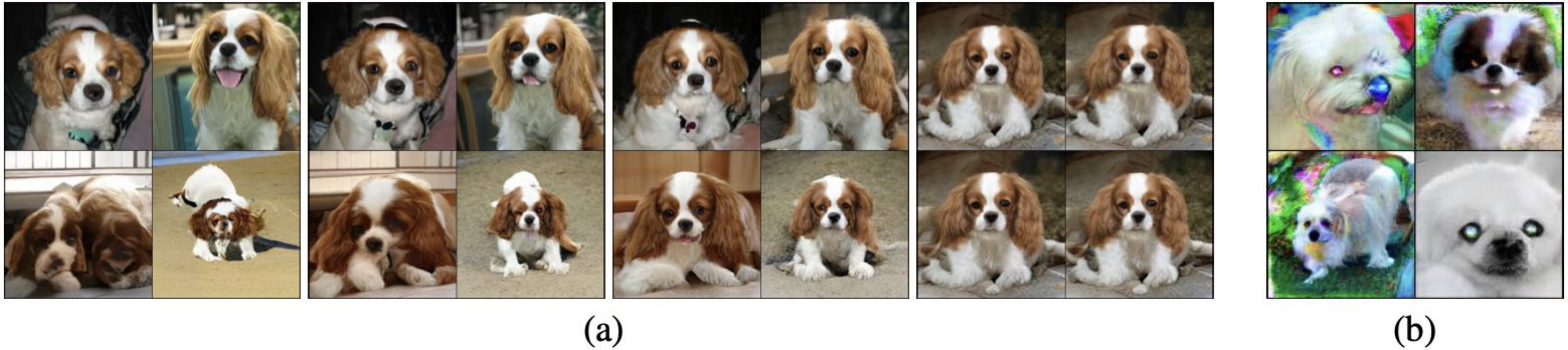
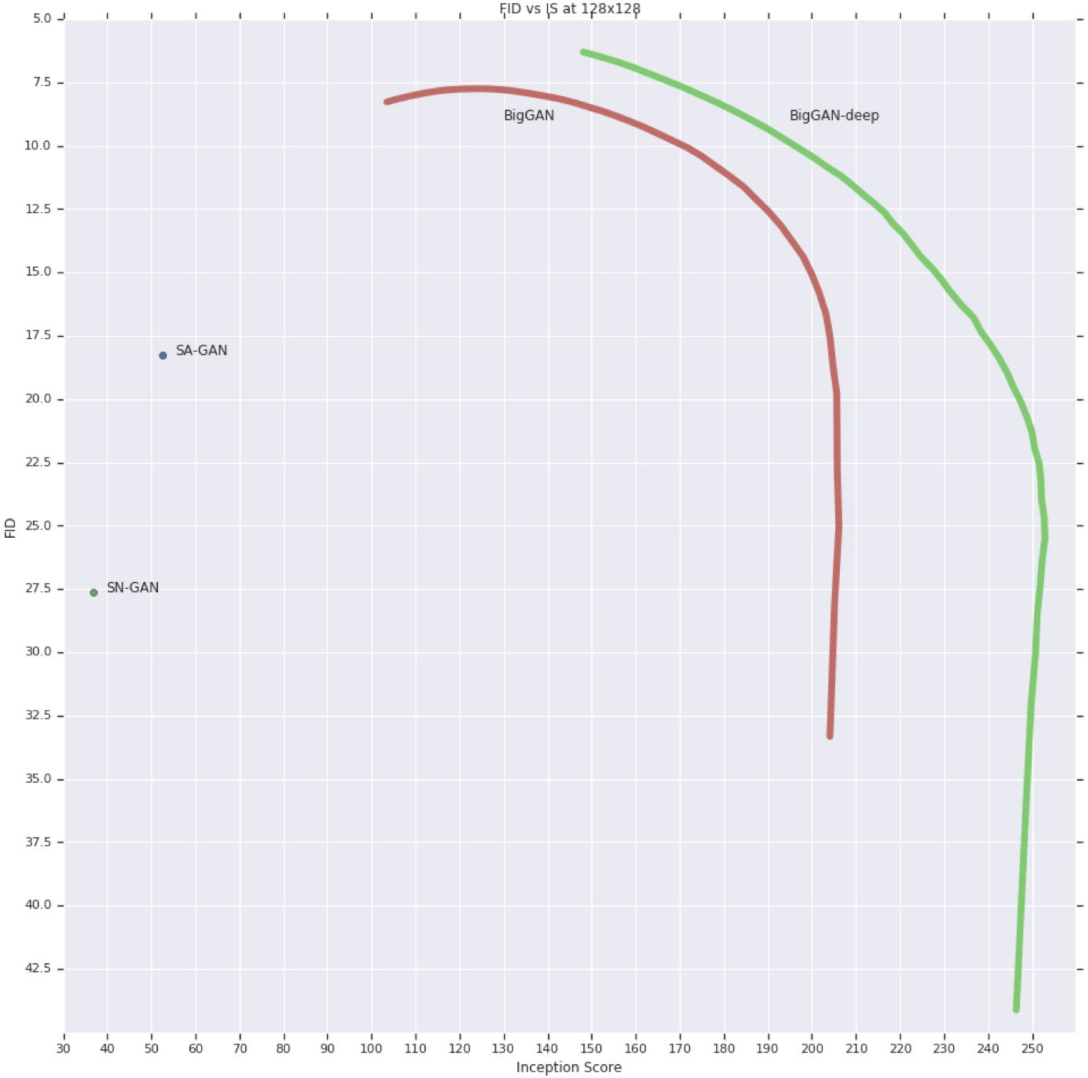


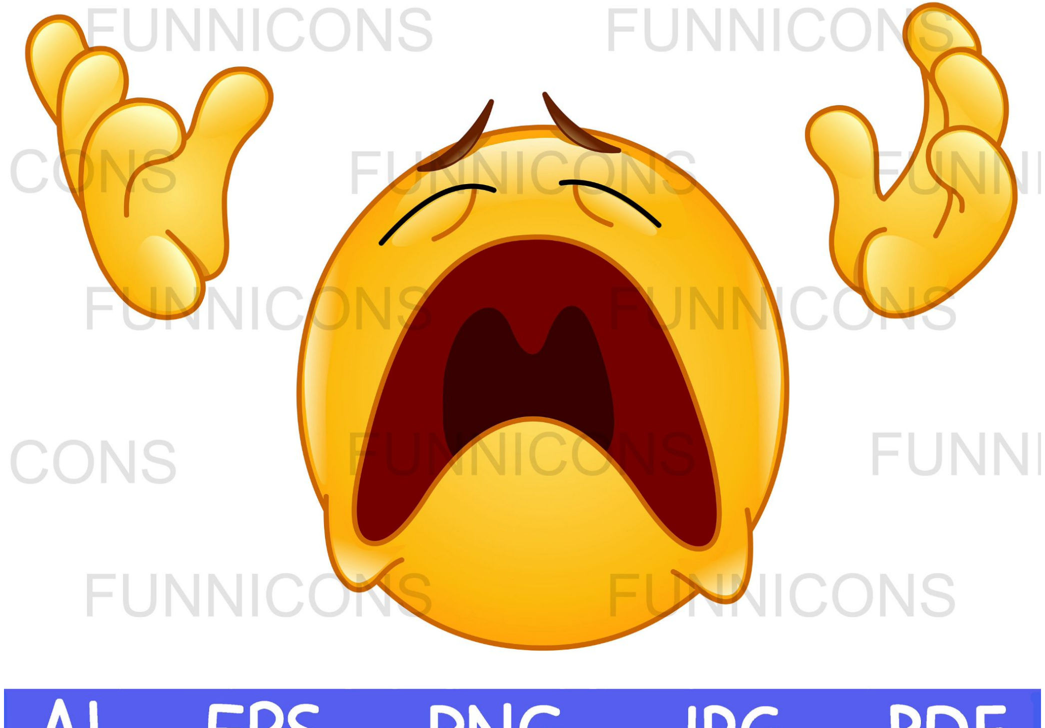
Figure 2: (a) The effects of increasing truncation. From left to right, the threshold is set to 2, 1, 0.5, 0.04. (b) Saturation artifacts from applying truncation to a poorly conditioned model.



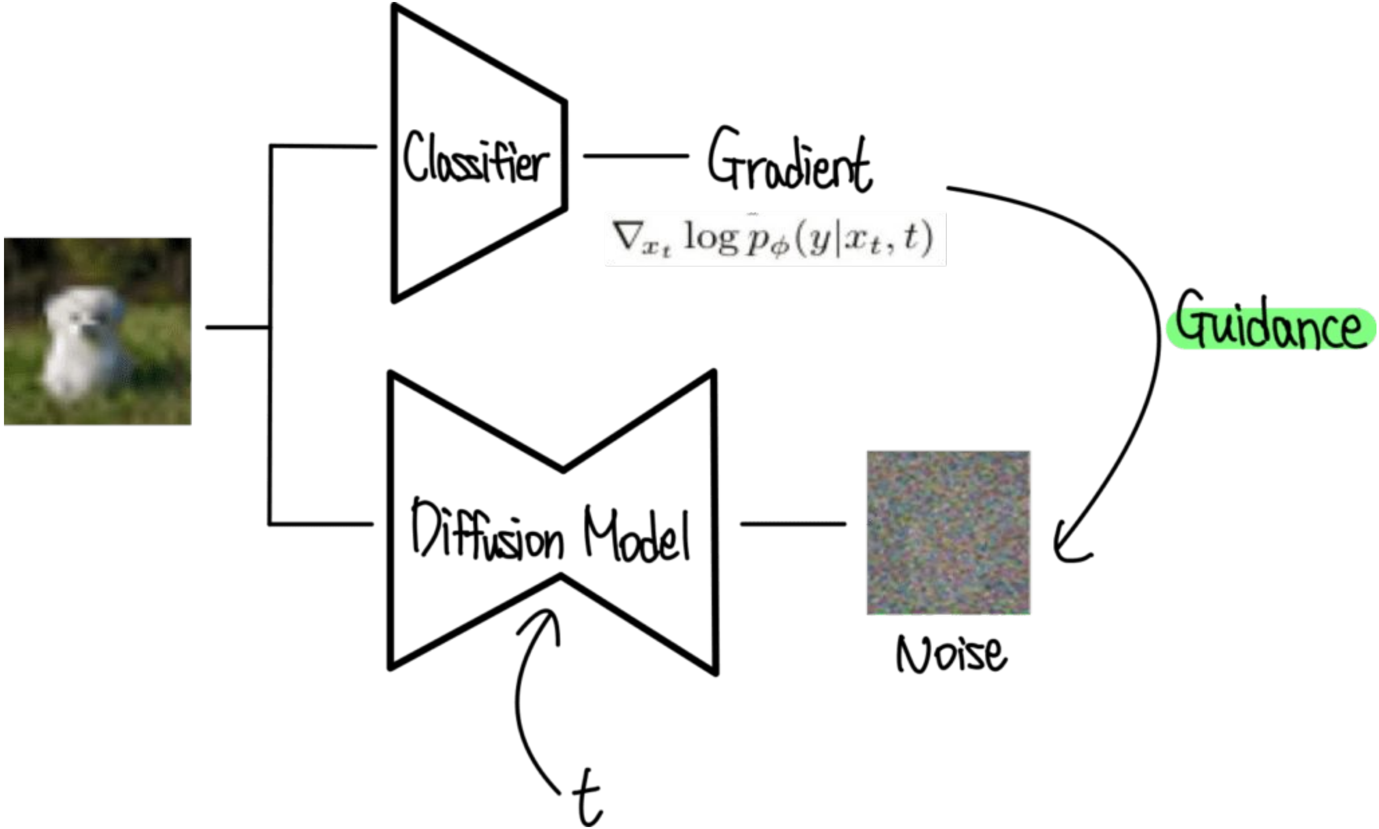
IS vs FID is just like

precision vs recall

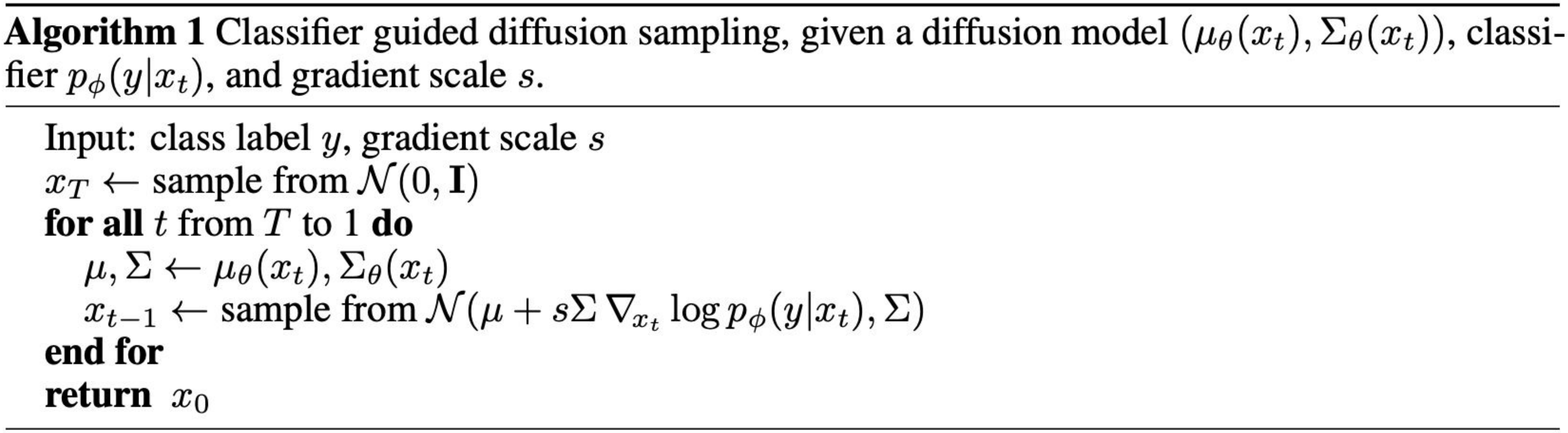
Unfortunately, straightforward attempts of implementing truncation or low temperature sampling in diffusion models are ineffective. For example, scaling model scores or decreasing the variance of Gaussian noise in the reverse process cause the diffusion model to generate blurry, low quality samples (Dhariwal & Nichol, 2021).



Classifier Guidance



Classifier Guidance



Algorithm 1 Classifier guided diffusion sampling, given a diffusion model $(\mu_\theta(x_t), \Sigma_\theta(x_t))$, classifier $p_\phi(y|x_t)$, and gradient scale s.

```
Input: class label $y$, gradient scale $s$
$x_T \leftarrow$ sample from $\mathcal{N}(0, \mathbf{I})$
for all $t$ from $T$ to 1 do
    $\mu, \Sigma \leftarrow \mu_\theta(x_t), \Sigma_\theta(x_t)$
    $x_{t-1} \leftarrow$ sample from $\mathcal{N}(\mu + s\Sigma \, \nabla_{x_t} \log p_\phi(y|x_t), \Sigma)$
end for
return $x_0$
```

Classifier Guidance

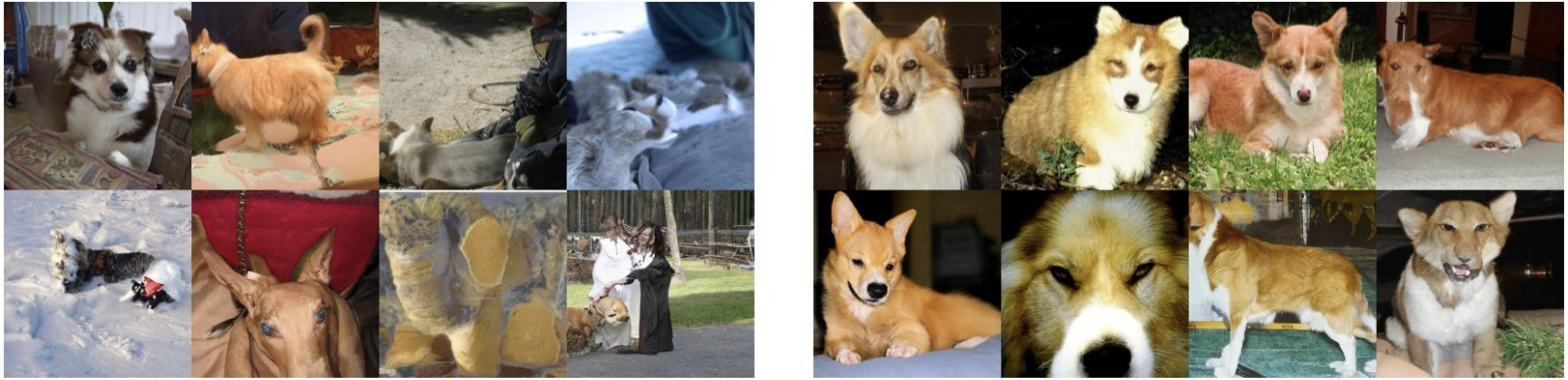


Figure 2: Samples from an unconditional diffusion model with classifier guidance to condition on the class "Pembroke Welsh corgi". Using classifier scale 1.0 (left; FID: 33.0) does not produce convincing samples in this class, whereas classifier scale 10.0 (right; FID: 12.0) produces much more class-consistent images.

Classifier Guidance

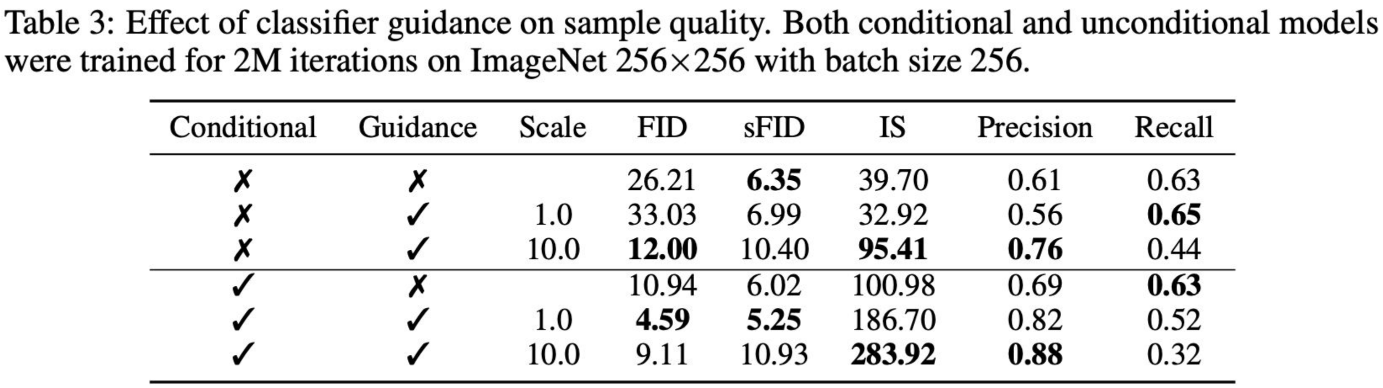
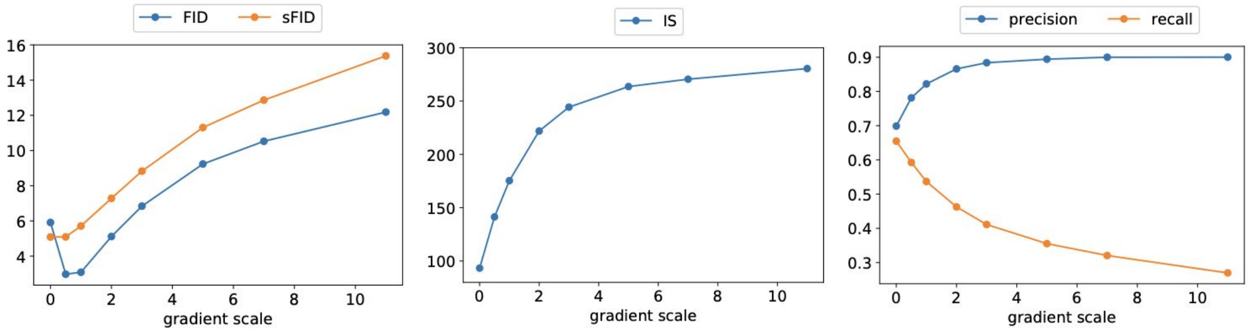


Table 3: Effect of classifier guidance on sample quality. Both conditional and unconditional models were trained for 2M iterations on ImageNet 256×256 with batch size 256.

Conditional	Guidance	Scale	FID	sFID	IS	Precision	Recall
✗	✗		26.21	**6.35**	39.70	0.61	0.63
✗	✓	1.0	33.03	6.99	32.92	0.56	**0.65**
✗	✓	10.0	**12.00**	10.40	**95.41**	**0.76**	0.44
✓	✗		10.94	6.02	100.98	0.69	**0.63**
✓	✓	1.0	**4.59**	**5.25**	186.70	0.82	0.52
✓	✓	10.0	9.11	10.93	**283.92**	**0.88**	0.32



Classifier Guidance

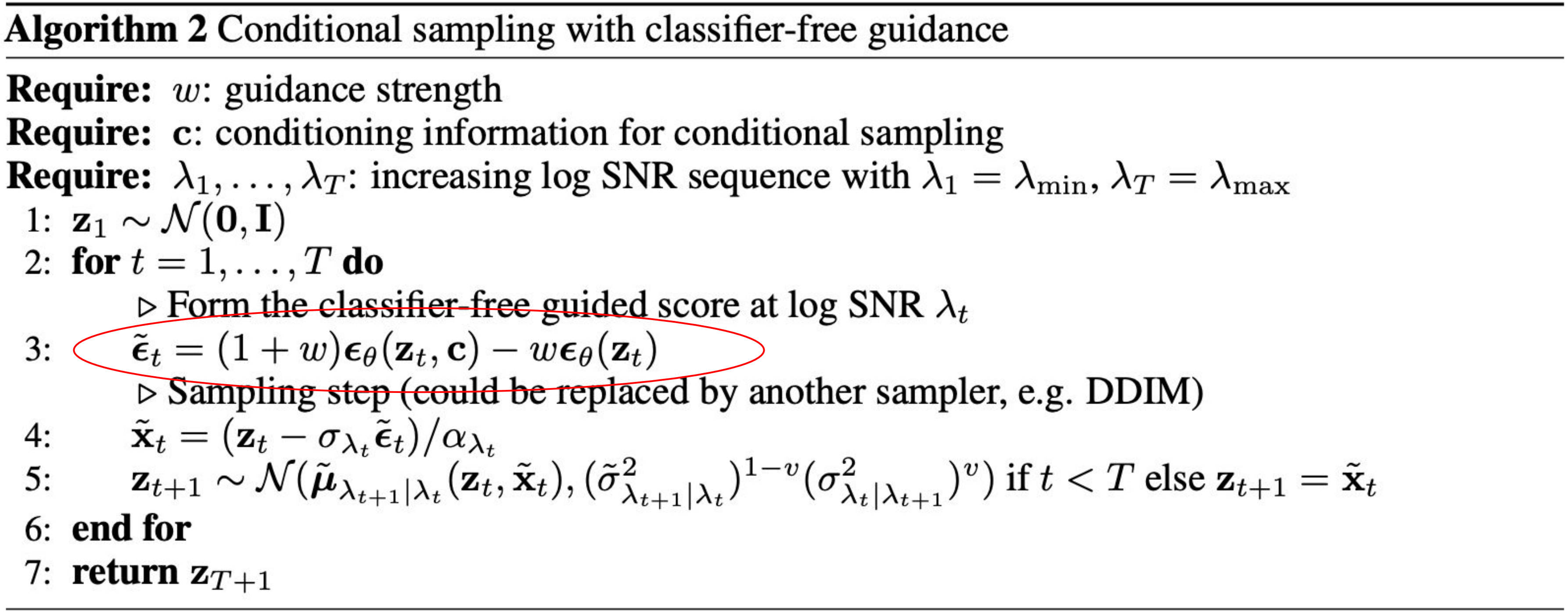
Pros:

- It works

Cons:

- Requires to train auxiliary classifier
- Can be interpreted as adversarial attack
- Metrics such as IS and FID become less representative, since they also use pretrained classifier

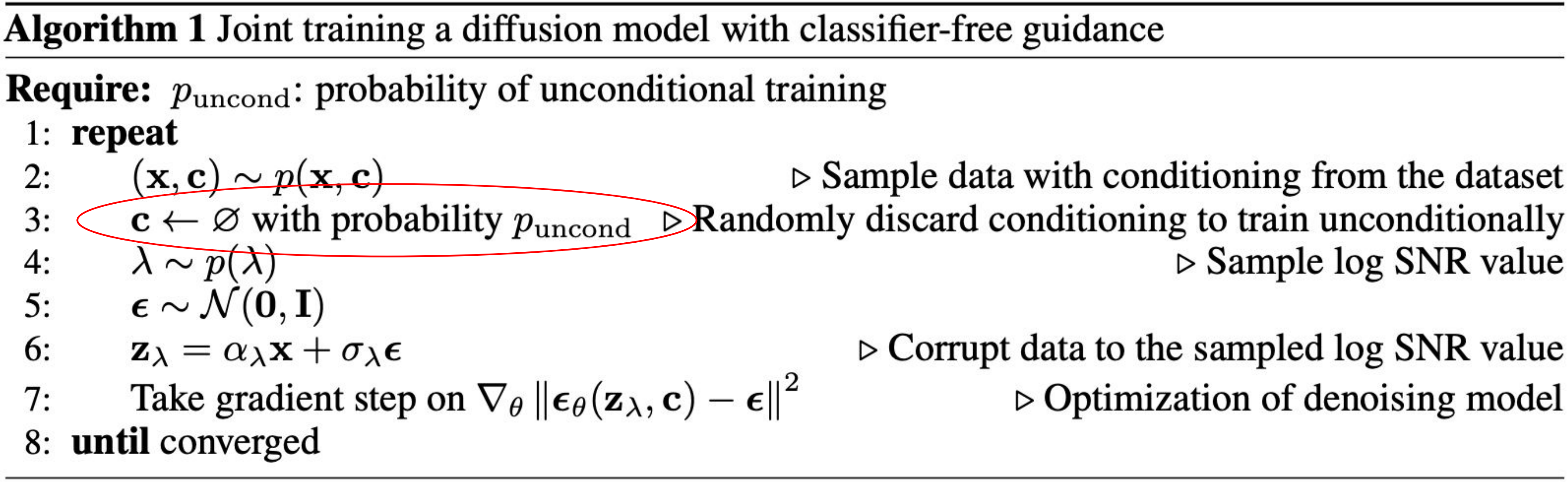
Classifier-Free Guidance



```
Algorithm 2 Conditional sampling with classifier-free guidance
Require: $w$: guidance strength
Require: $\mathbf{c}$: conditioning information for conditional sampling
Require: $\lambda_1, \ldots, \lambda_T$: increasing log SNR sequence with $\lambda_1 = \lambda_{\min}$, $\lambda_T = \lambda_{\max}$
$\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = 1, \ldots, T$ do
        ▷ Form the classifier-free guided score at log SNR $\lambda_t$
    $\tilde{\boldsymbol{\epsilon}}_t = (1 + w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{z}_t)$
        ▷ Sampling step (could be replaced by another sampler, e.g. DDIM)
    $\tilde{\mathbf{x}}_t = (\mathbf{z}_t - \sigma_{\lambda_t}\tilde{\boldsymbol{\epsilon}}_t)/\alpha_{\lambda_t}$
    $\mathbf{z}_{t+1} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}_{\lambda_{t+1}|\lambda_t}(\mathbf{z}_t, \tilde{\mathbf{x}}_t), (\tilde{\sigma}^2_{\lambda_{t+1}|\lambda_t})^{1-v}(\sigma^2_{\lambda_t|\lambda_{t+1}})^v)$ if $t < T$ else $\mathbf{z}_{t+1} = \tilde{\mathbf{x}}_t$
end for
return $\mathbf{z}_{T+1}$
```

Intuitive explanation is to decrease unconditional likelihood while increasing conditional likelihood

Classifier-Free Guidance



Algorithm 1 Joint training a diffusion model with classifier-free guidance

```
Require: $p_{\text{uncond}}$: probability of unconditional training
repeat
    $(\mathbf{x}, \mathbf{c}) \sim p(\mathbf{x}, \mathbf{c})$    ▷ Sample data with conditioning from the dataset
    $\mathbf{c} \leftarrow \varnothing$ with probability $p_{\text{uncond}}$    ▷ Randomly discard conditioning to train unconditionally
    $\lambda \sim p(\lambda)$    ▷ Sample log SNR value
    $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{z}_\lambda = \alpha_\lambda \mathbf{x} + \sigma_\lambda \boldsymbol{\epsilon}$    ▷ Corrupt data to the sampled log SNR value
    Take gradient step on $\nabla_\theta \left\| \boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) - \boldsymbol{\epsilon} \right\|^2$    ▷ Optimization of denoising model
until converged
```

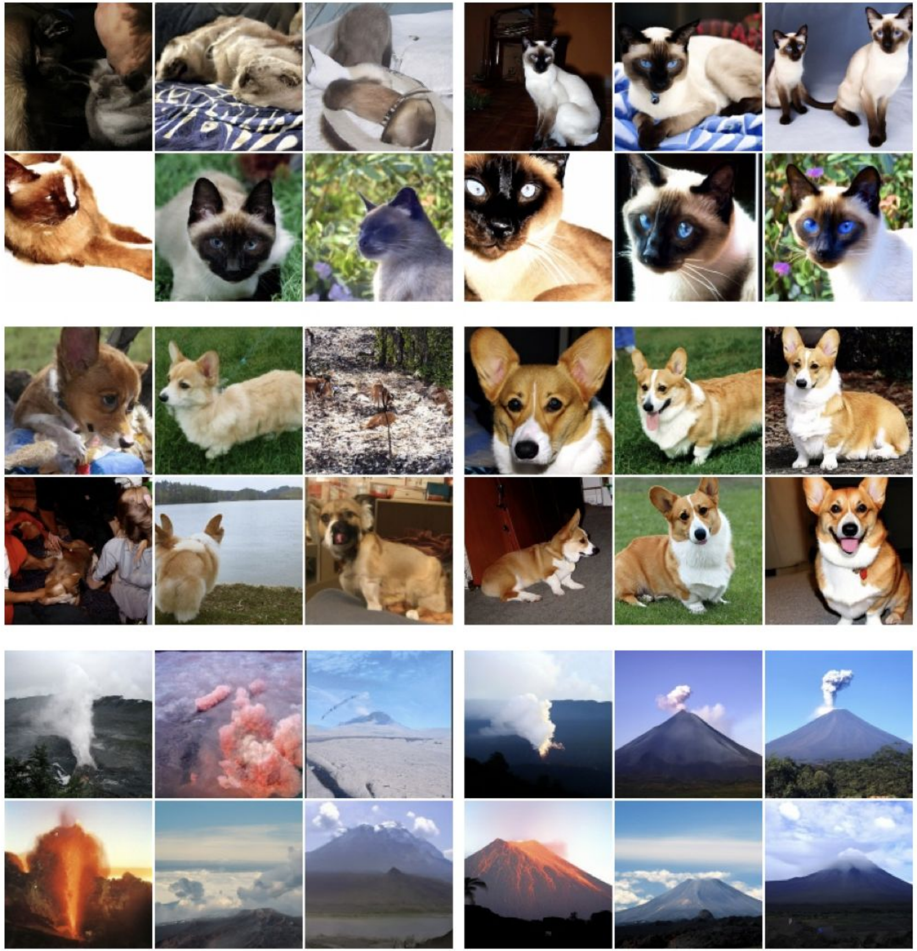
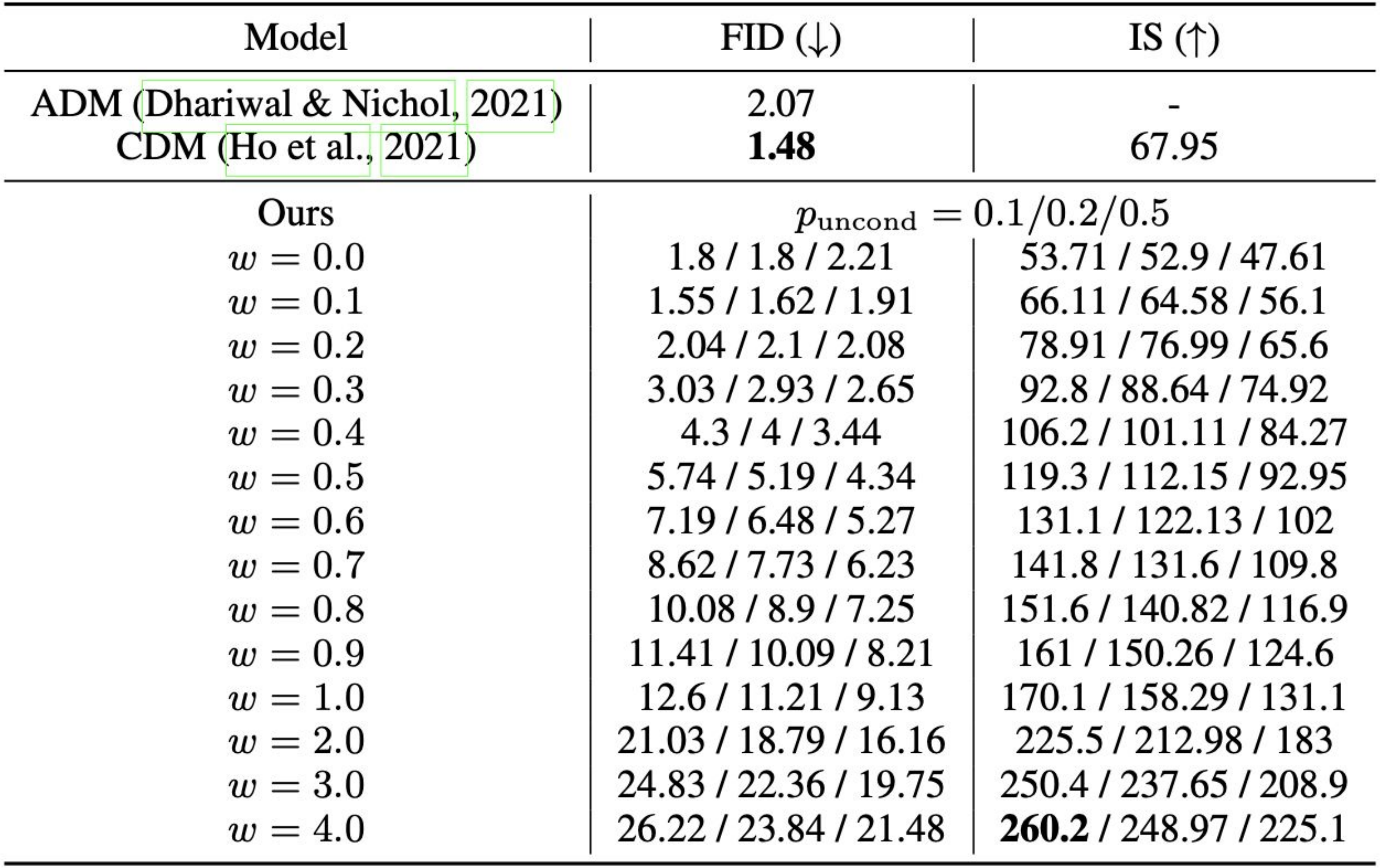


Figure 3: Classifier-free guidance on 128x128 ImageNet. Left: non-guided samples, right: classifier-free guided samples with $w = 3.0$. Interestingly, strongly guided samples such as these display saturated colors. See Fig. 8 for more.



Model	FID (↓)	IS (↑)
ADM (Dhariwal & Nichol, 2021)	2.07	-
CDM (Ho et al., 2021)	**1.48**	67.95
Ours	$p_{\text{uncond}} = 0.1/0.2/0.5$	
$w = 0.0$	1.8 / 1.8 / 2.21	53.71 / 52.9 / 47.61
$w = 0.1$	1.55 / 1.62 / 1.91	66.11 / 64.58 / 56.1
$w = 0.2$	2.04 / 2.1 / 2.08	78.91 / 76.99 / 65.6
$w = 0.3$	3.03 / 2.93 / 2.65	92.8 / 88.64 / 74.92
$w = 0.4$	4.3 / 4 / 3.44	106.2 / 101.11 / 84.27
$w = 0.5$	5.74 / 5.19 / 4.34	119.3 / 112.15 / 92.95
$w = 0.6$	7.19 / 6.48 / 5.27	131.1 / 122.13 / 102
$w = 0.7$	8.62 / 7.73 / 6.23	141.8 / 131.6 / 109.8
$w = 0.8$	10.08 / 8.9 / 7.25	151.6 / 140.82 / 116.9
$w = 0.9$	11.41 / 10.09 / 8.21	161 / 150.26 / 124.6
$w = 1.0$	12.6 / 11.21 / 9.13	170.1 / 158.29 / 131.1
$w = 2.0$	21.03 / 18.79 / 16.16	225.5 / 212.98 / 183
$w = 3.0$	24.83 / 22.36 / 19.75	250.4 / 237.65 / 208.9
$w = 4.0$	26.22 / 23.84 / 21.48	**260.2** / 248.97 / 225.1

Table 1: ImageNet 64x64 results ($w = 0.0$ refers to non-guided models).

Summary

Pros:

- Extremely easy to implement
- Does not require pretrained classifier
- Cannot be viewed as adversarial attack

Cons:

- Requires 2 sample passes

References

- Classifier-Free Diffusion Guidance
- Classifier Guidance
- Improved Precision and Recall Metric for Assessing Generative Models