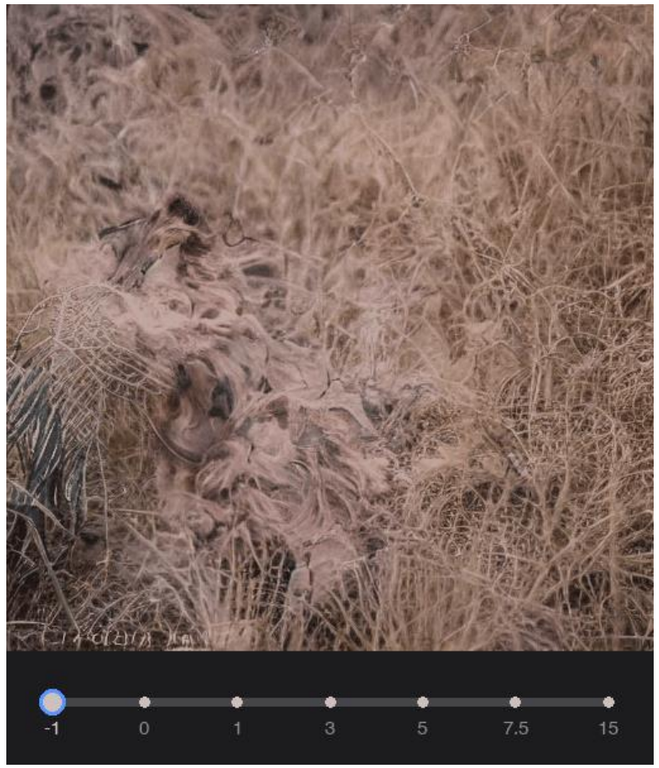
# Classifier-Free Diffusion Guidance

Presenter: Xiao Fang



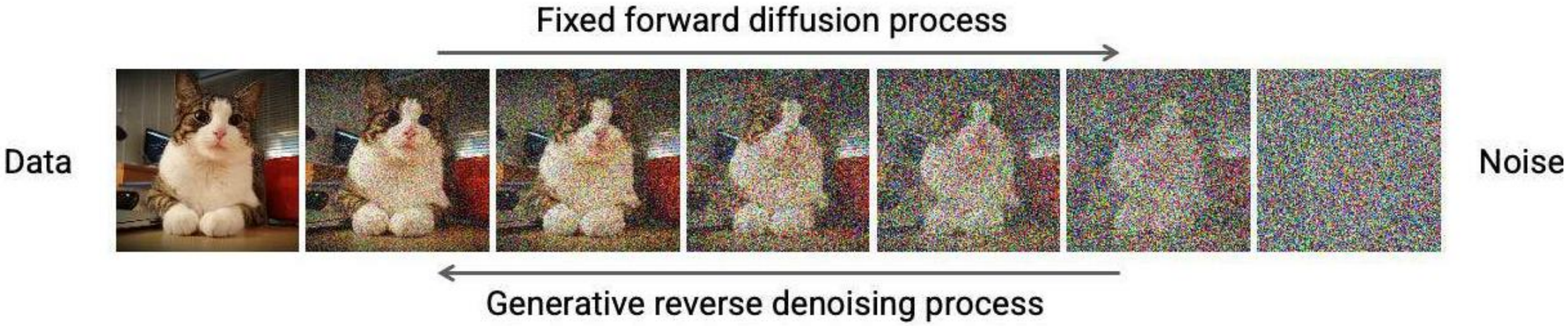
# Diffusion Models for Image Generation

*Strengths:*

- Distribution coverage, stationary training objective, easy scalability

*Weakness:*

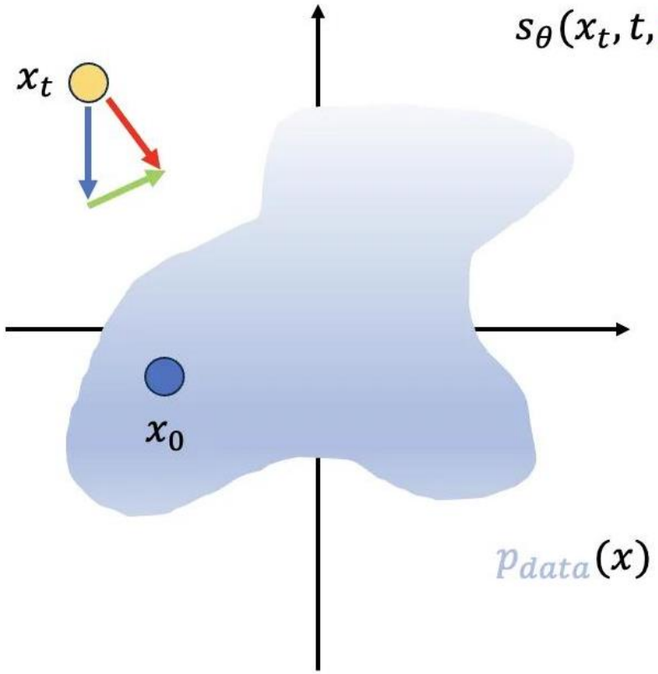
- Lags behind GANs on difficult generation datasets like ImageNet




Denoising Diffusion Probabilistic Models, NeurIPS 2020

# Classifier Guidance

How to condition diffusion models on class labels?




$$s_\theta(x_t, t, y) \approx \nabla \log p(x_t|y)$$

Conditional Score

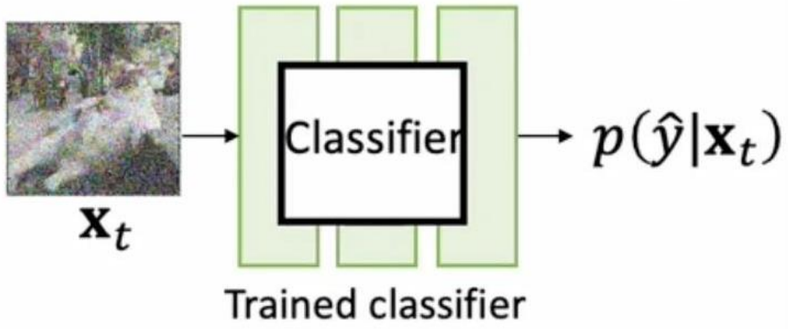
$$\begin{aligned}\nabla \log p(\boldsymbol{x}_t|y) &= \nabla \log \left( \frac{p(\boldsymbol{x}_t)p(y|\boldsymbol{x}_t)}{p(y)} \right) \\ &= \nabla \log p(\boldsymbol{x}_t) + \nabla \log p(y|\boldsymbol{x}_t) - \nabla \log p(y) \\ &= \quad \nabla \log p(\boldsymbol{x}_t) \quad + \quad \nabla \log p(y|\boldsymbol{x}_t)\end{aligned}$$

Unconditional Score

Adversarial Gradient

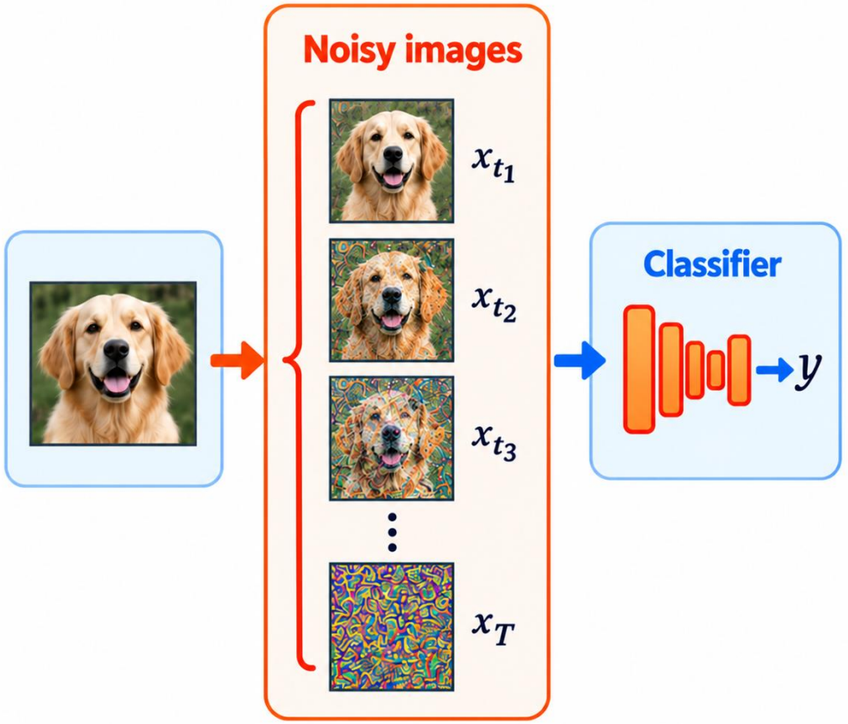
$$\nabla \log p_\gamma(x_t|y) = \nabla \log p(x_t) + \gamma \nabla \log p(y|x_t)$$

Conditional Score

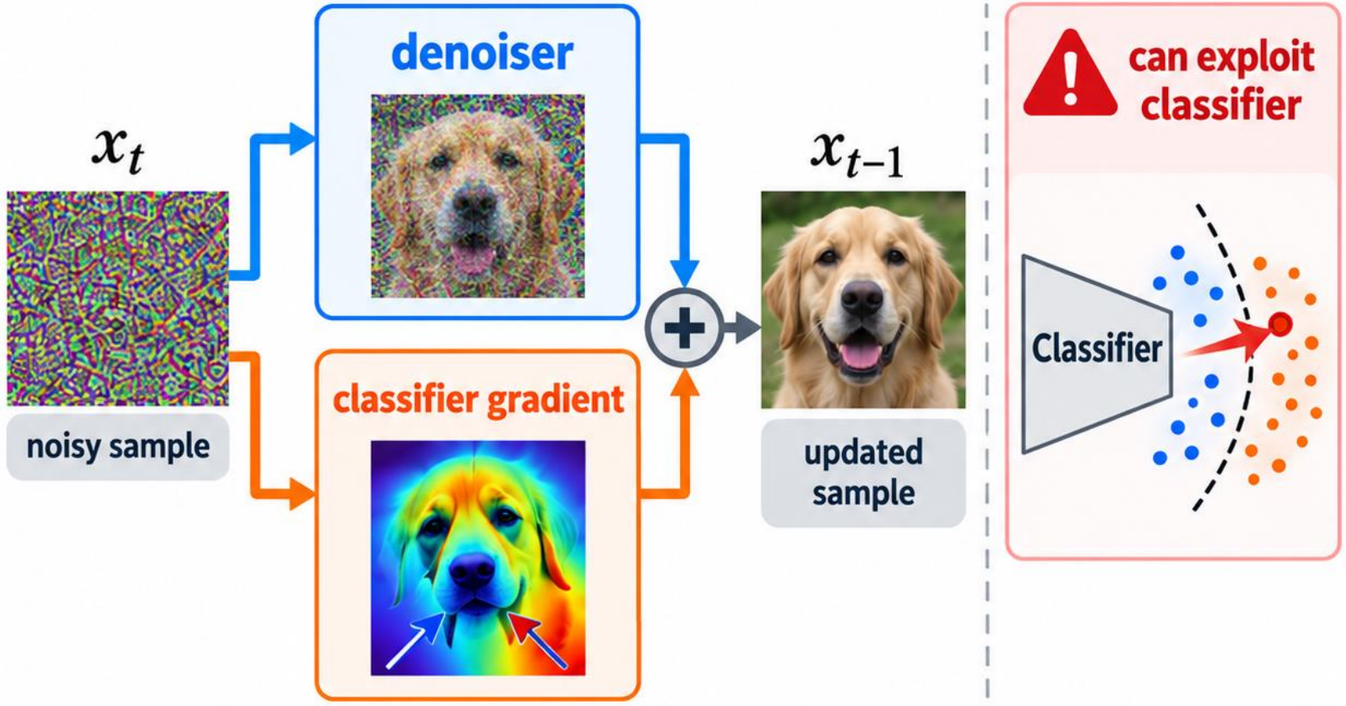



Diffusion Models Beat GANs on Image Synthesis, NeurIPS 2021

# Limitations of Classifier Guidance

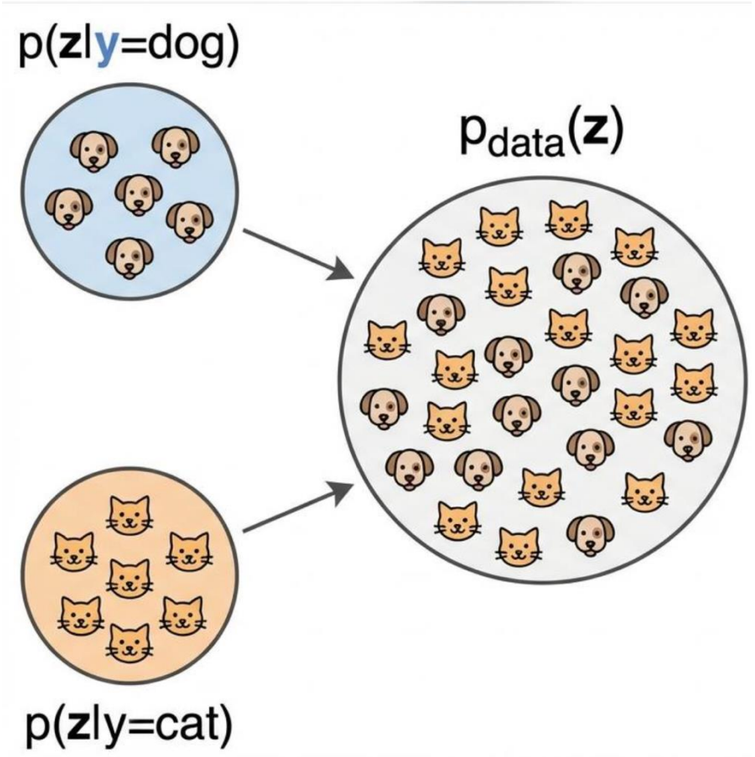



Extra Noisy classifier Training




Gradient-based classifier attack

# Classifier-Free Guidance




$$\nabla \log p(\boldsymbol{x}_t|y) = \nabla \log p(\boldsymbol{x}_t) + \gamma \nabla \log p(y|\boldsymbol{x}_t)$$

$$\nabla \log p(y|\boldsymbol{x}_t) = \nabla \log p(\boldsymbol{x}_t|y) - \nabla \log p(\boldsymbol{x}_t)$$

$$\begin{aligned}
\nabla \log p(\boldsymbol{x}_t|y) &= \nabla \log p(\boldsymbol{x}_t) + \gamma \left(\nabla \log p(\boldsymbol{x}_t|y) - \nabla \log p(\boldsymbol{x}_t)\right) \\
&= \nabla \log p(\boldsymbol{x}_t) + \gamma \nabla \log p(\boldsymbol{x}_t|y) - \gamma \nabla \log p(\boldsymbol{x}_t) \\
&= \underbrace{\gamma \nabla \log p(\boldsymbol{x}_t|y)}_{\text{conditional score}} + \underbrace{(1-\gamma) \nabla \log p(\boldsymbol{x}_t)}_{\text{unconditional score}}
\end{aligned}$$

# Algorithm: Training

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

**Require:** $p_{\text{uncond}}$: probability of unconditional training

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

# Lemma

$$
\begin{aligned}
\nabla_{x_t} \log q(x_t \mid x_0) &= \nabla_{x_t} \left( -\frac{\|x_t - \alpha_t x_0\|^2}{2\sigma_t^2} + C \right) \\
&= -\frac{x_t - \alpha_t x_0}{\sigma_t^2} \\
&= \boxed{-\frac{\epsilon}{\sigma_t}}
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{x_t} \log q_t(x_t) &= \frac{\nabla_{x_t} q_t(x_t)}{q_t(x_t)} \\
&= \frac{1}{q_t(x_t)} \int p_{\text{data}}(x_0) \nabla_{x_t} q(x_t \mid x_0)\, dx_0 \\
&= \int \underbrace{\frac{p_{\text{data}}(x_0) q(x_t \mid x_0)}{q_t(x_t)}}_{q(x_0 \mid x_t)} \nabla_{x_t} \log q(x_t \mid x_0)\, dx_0 \\
&= \mathbb{E}_{q(x_0 \mid x_t)} \left[ \nabla_{x_t} \log q(x_t \mid x_0) \right] \\
&= -\frac{1}{\sigma_t} \mathbb{E}[\epsilon \mid x_t]
\end{aligned}
$$

$$
\underbrace{s_\theta(x_t, t)}_{\text{Predicted score}} = -\frac{\epsilon_\theta(x_t, t)}{\sigma_t} \approx \underbrace{\nabla_{x_t} \log q_t(x_t)}_{\text{True marginal score}}
$$

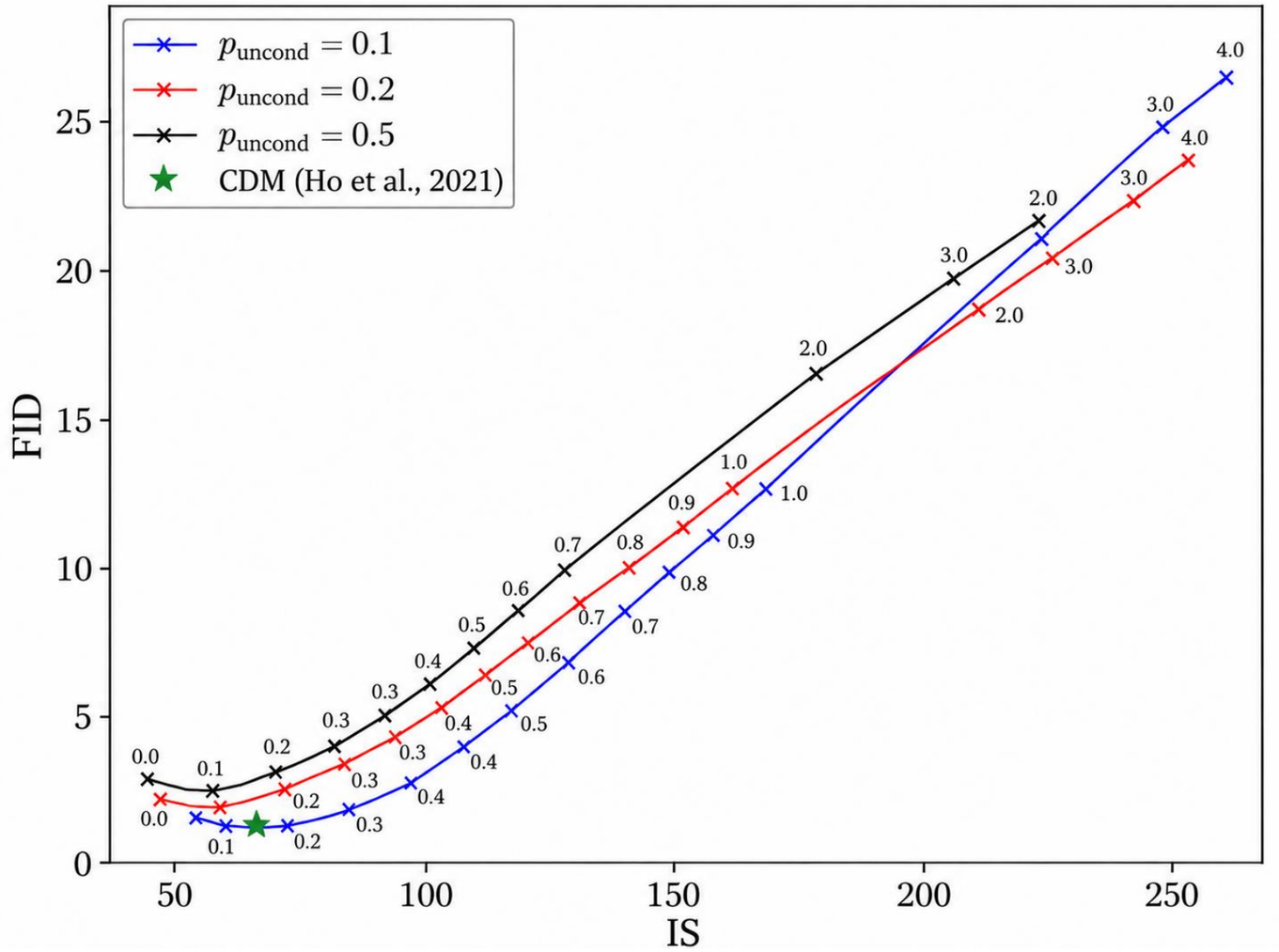
# Algorithm: Inference

**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 $\lambda_1 = \lambda_{\min}, \lambda_T = \lambda_{\max}$

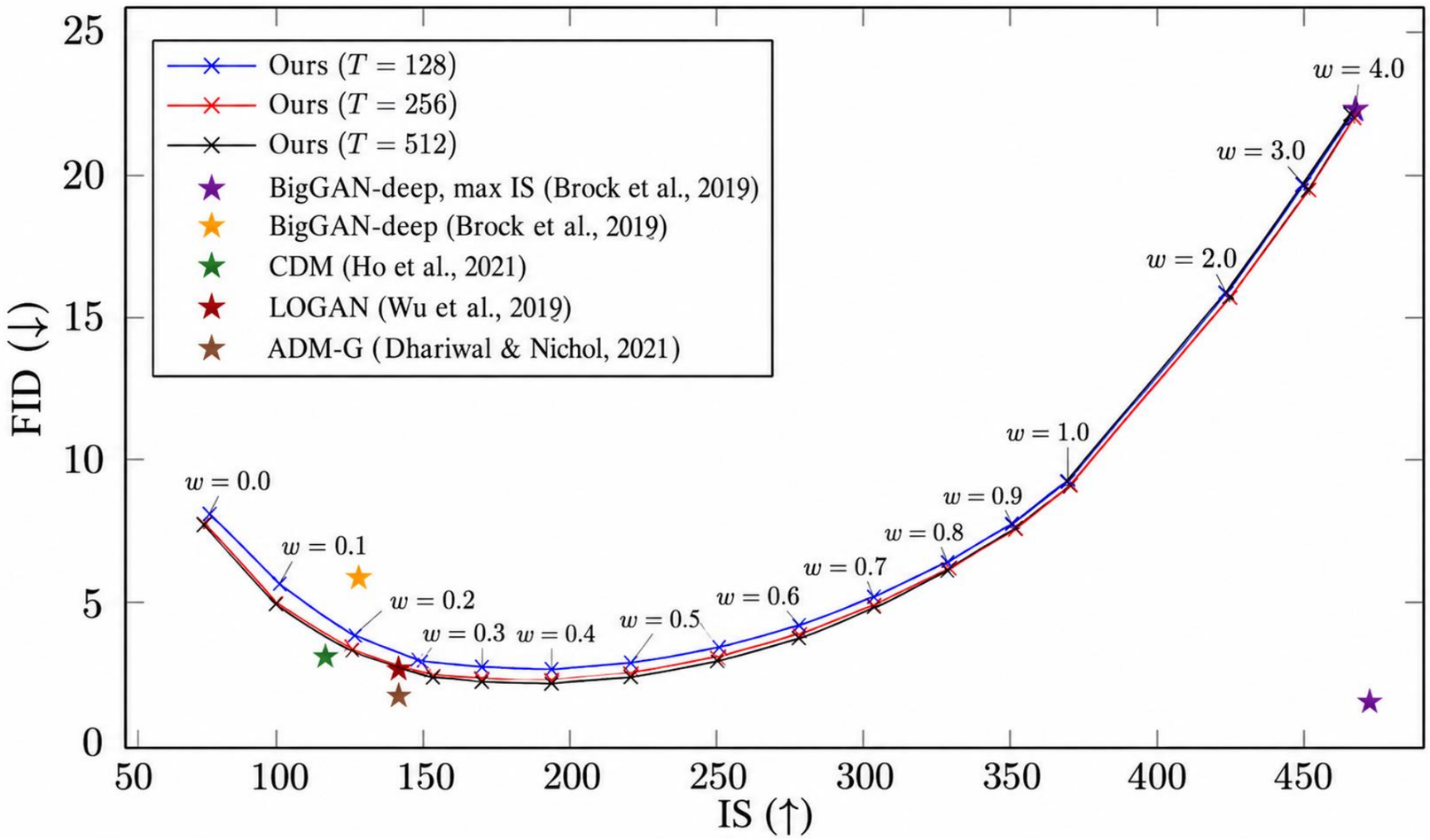
$\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
**for** $t = 1, \ldots, T$ **do**
    ▷ Form the classifier-free guided score at log SNR $\lambda_t$
    $\tilde{\boldsymbol{\epsilon}}_t = (1 + w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{z}_t)$
    ▷ 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** $\mathbf{z}_{T+1}$

# Comparison with Classifier Guidance



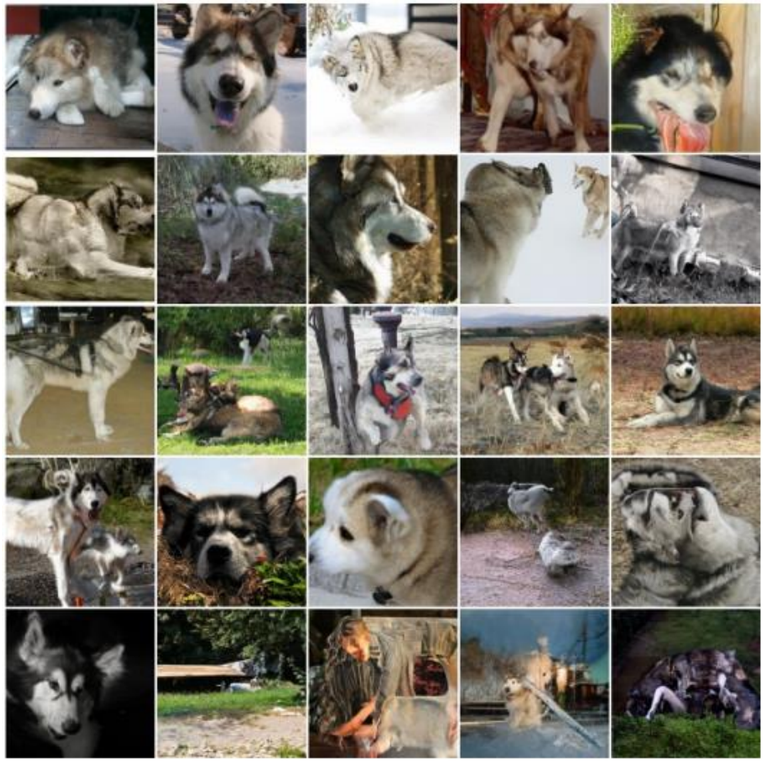

ImageNet 64x64 results

# Comparison with SOTA




ImageNet 128x128 results

# Visualizations



$\omega = 0$

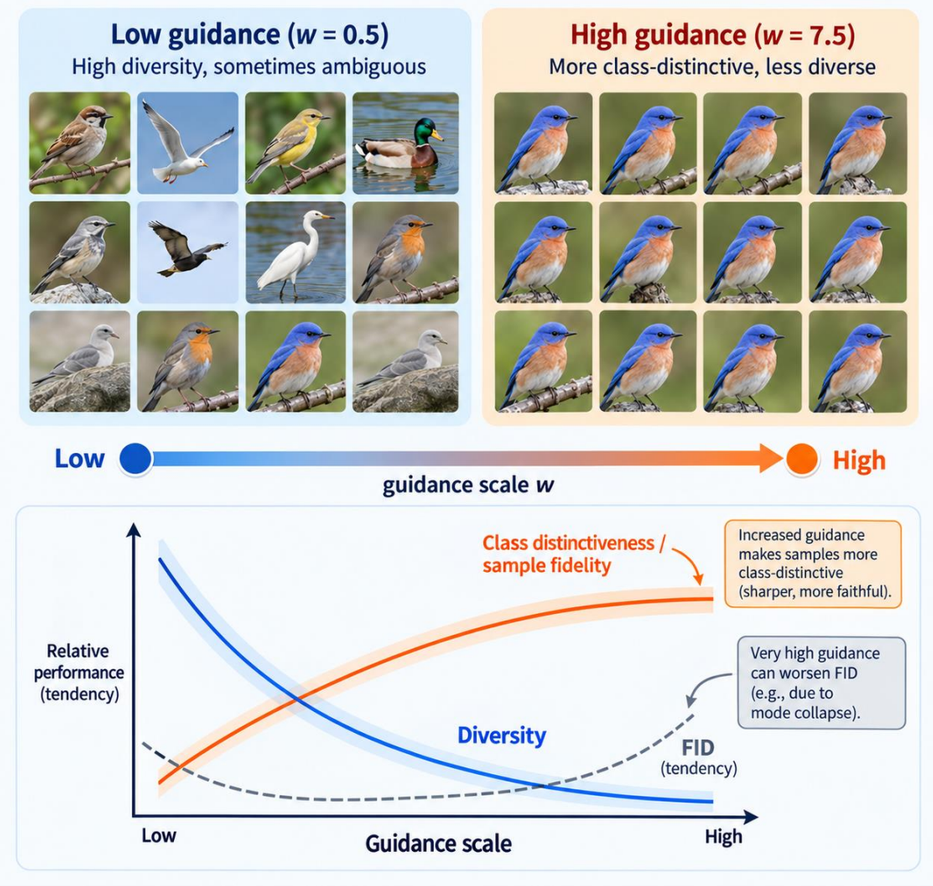


$\omega = 1$

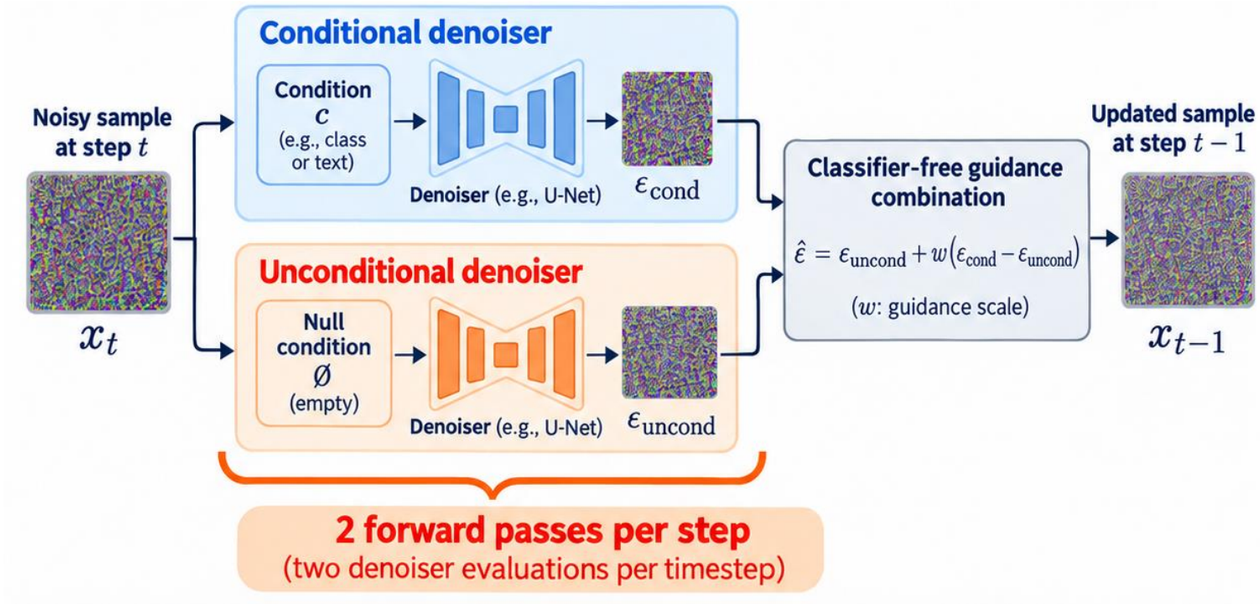


$\omega = 3$

# Limitations of Classifier-Free Guidance




Fidelity-Diversity Tradeoff




Sampling Cost

# Q & A

- Since $\varepsilon(z, c) - \varepsilon(z)$ is not the gradient of any real classifier, why does it still work so well as a guidance direction? Is there a theoretical explanation?

- How should we choose the guidance strength in practice since increasing it improves some metrics but can reduce diversity and make the samples less varied?

# Thank You!