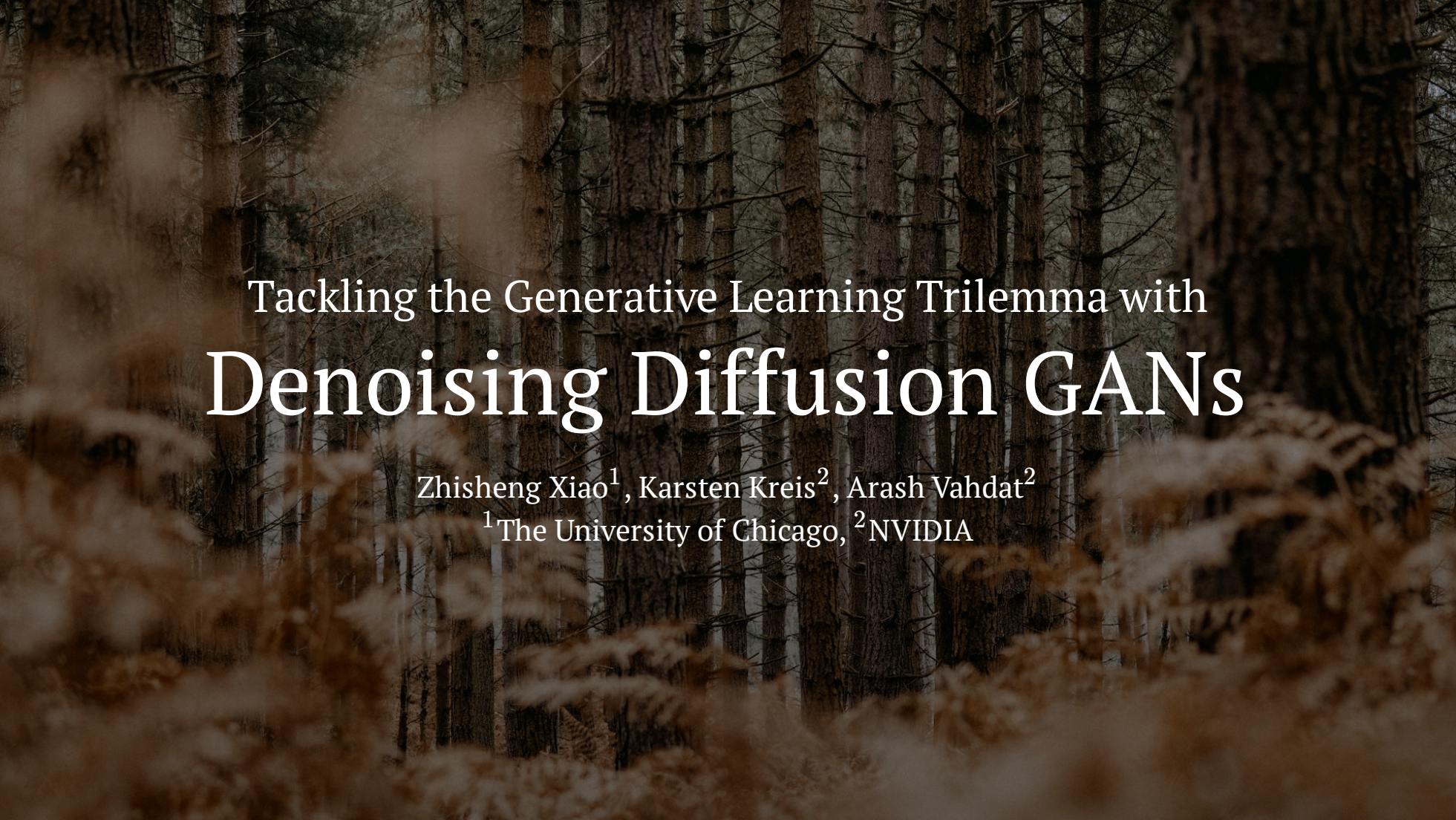


Tackling the Generative Learning Trilemma with

Denoising Diffusion GANs

Zhisheng Xiao[1], Karsten Kreis[2], Arash Vahdat[2]
[1]The University of Chicago, [2]NVIDIA

The Generative Learning Trilemma

In the past decade, a plethora of deep generative models has been developed for various domains such as images, audios, point clouds and graphs.

However, current generative learning frameworks cannot yet simultaneously satisfy three key requirements,

- High-Quality Sampling,

  Variational Autoencoders (VAEs) and Normalizing Flows often suffer from low sample quality.

- Mode Coverage and Sample Diversity, and

  Generative Adversarial Networks (GANs) are known for poor mode coverage.

- Fast and Computationally Inexpensive Sampling.

  Sampling from Diffusion Model often requires thousands of neural network evaluations.

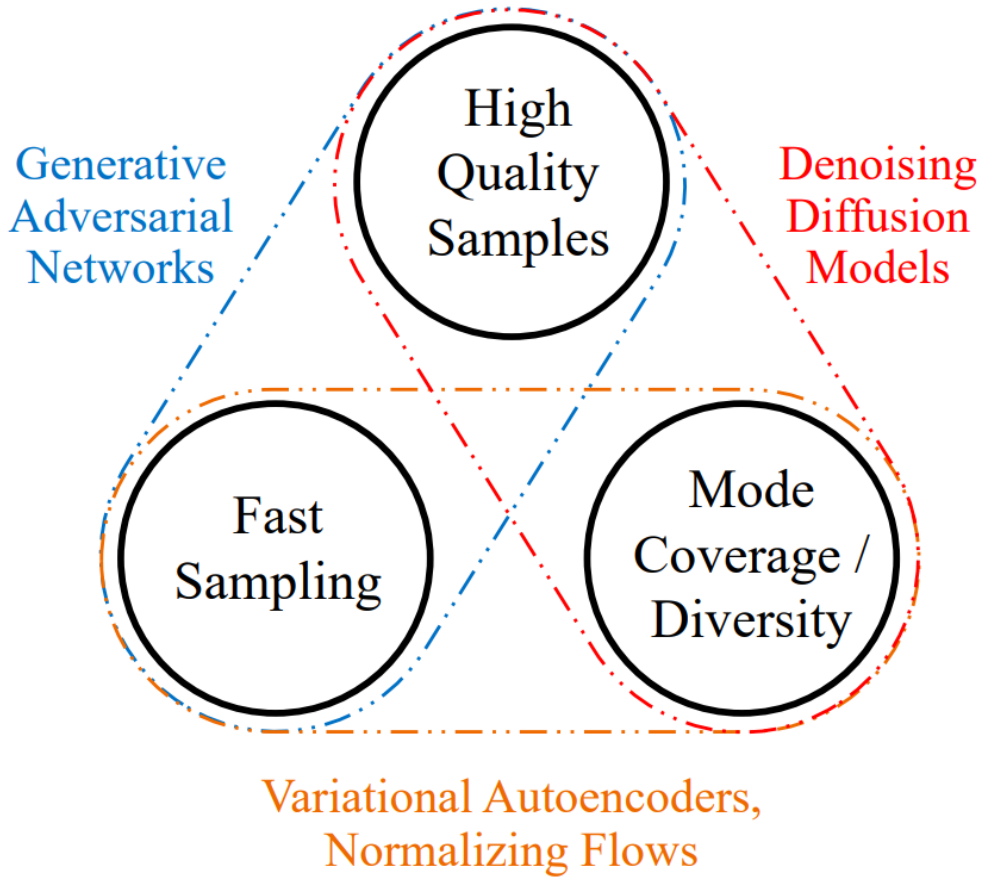
Diffusion Models

Normal Distribution Assumption

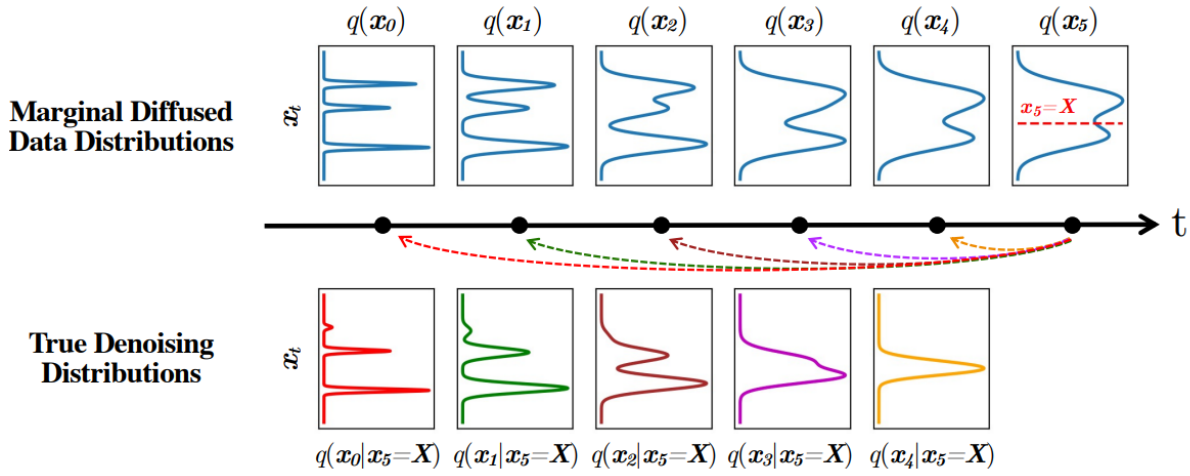
In the generation framework introduced above, Diffusion Models requires a lot of generation time in exchange for high-quality sampling because its normal distribution assumption is only established when the step size is extremely small.

Building upon this premise, this study addresses the trilemma of generative learning by **REDUCING THE SAMPLING TIME** of Diffusion Models through the **MODIFICATION OF THE DENOISING DISTRIBUTION**.

By using GAN to model the denoising distribution of Diffusion Models, it changes from a normal distribution to a **MULTIMODAL DISTRIBUTION**, thereby reducing the number of sampling steps required.


High Quality Samples
Generative Adversarial Networks
Denoising Diffusion Models
Fast Sampling
Mode Coverage / Diversity
Variational Autoencoders, Normalizing Flows

Denoising Distribution



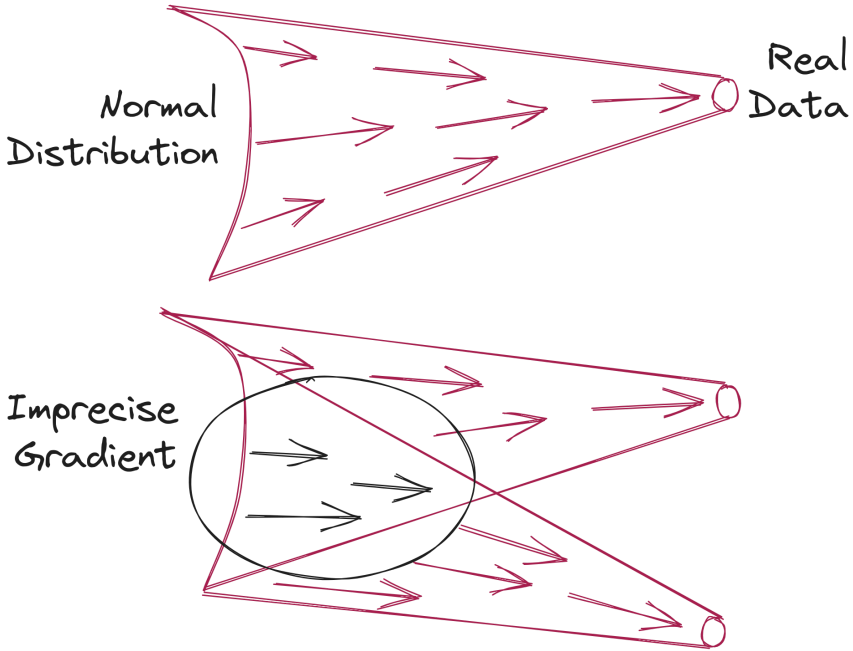
If diffusion process is

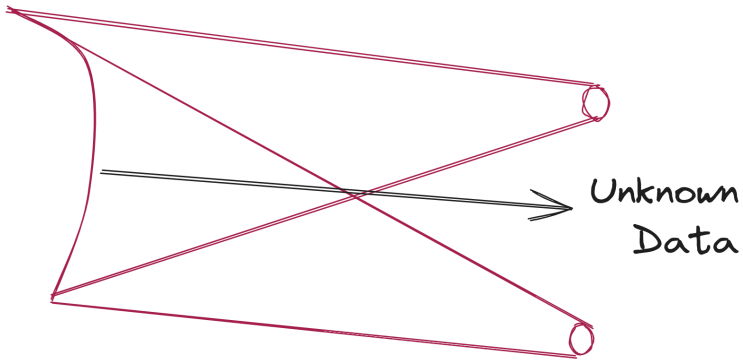
1. Markov chain
2. Normal distribution

And when the step size of the denoising process is **EXTREMELY SMALL**, the denoising process will also be normally distributed.

Imprecise Estimation

When close to the normal distribution





When the current state is close to the normal distribution, there will be multiple clean data corresponding to the same noisy data.

Denoising Diffusion GAN

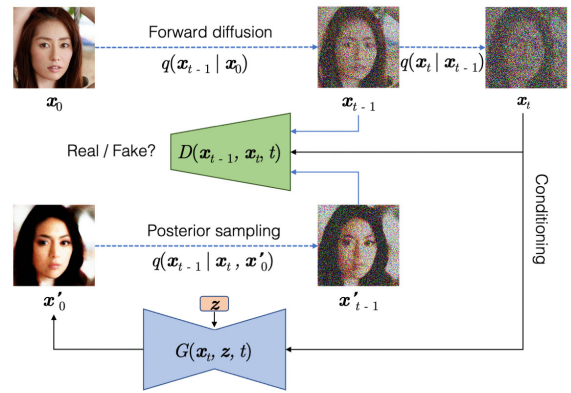
Implicitly Modeling the Denoising Process with GANs

$$p_\theta(x_{t-1}|x_t) := \int p_\theta(x_0|x_t)q(x_{t-1}|x_t, x_0)dx_0$$
$$= \int p(z)q(x_{t-1}|x_t, x_0 = G_\theta(x_t, z, t))dz$$

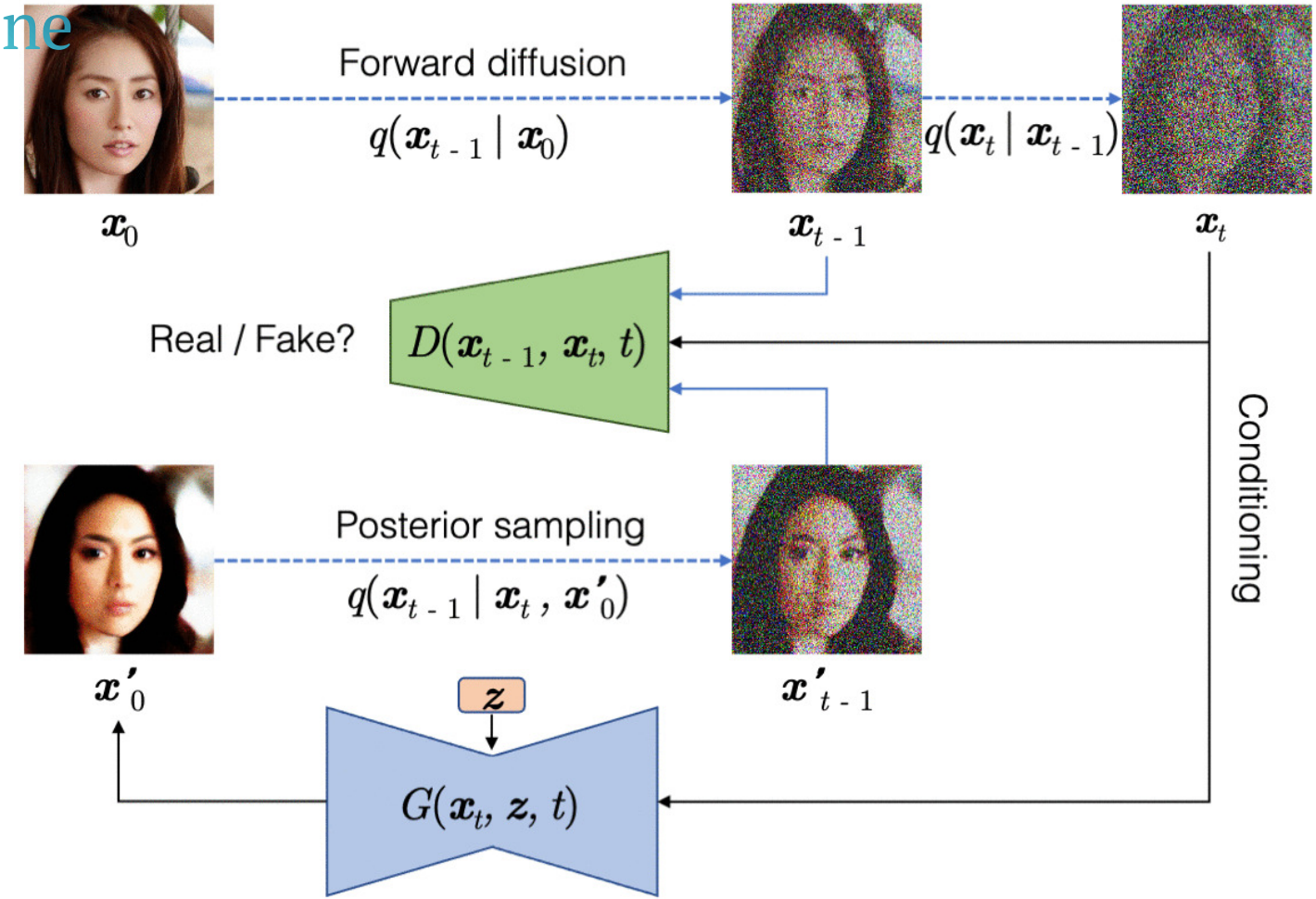
$$\sum_{t\geq 1} \mathbb{E}_{q(x_t)}[D_{KL}(q(x_{t-1}|x_t)||p_\theta(x_{t-1}|x_t))] + C$$

$$\Downarrow$$

$$\min_\theta \max_\phi \sum_{t\geq 1} \mathbb{E}_{q(x_t)}[D_\phi(q(x_{t-1}|x_t)||p_\theta(x_{t-1}|x_t))]$$



Pipeline

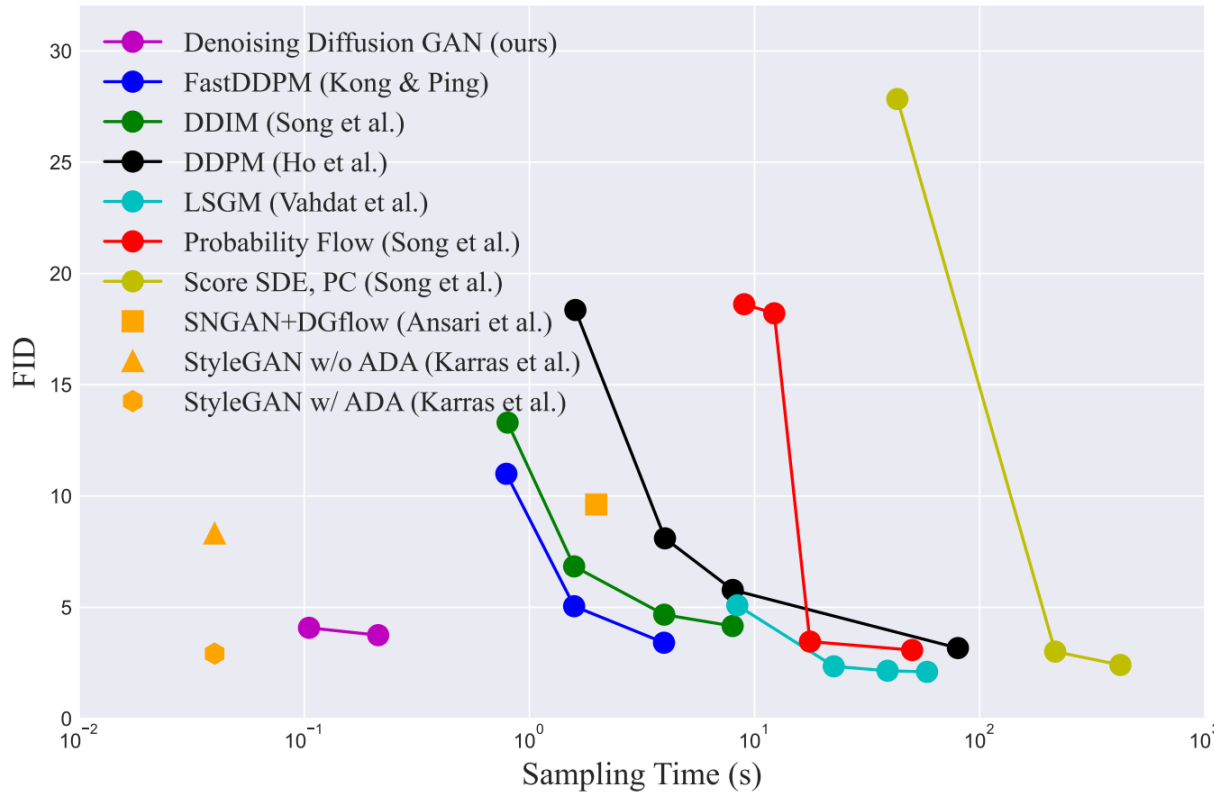

Forward diffusion
$q(x_{t-1} \mid x_0)$
$q(x_t \mid x_{t-1})$
x_0
x_{t-1}
x_t
Real / Fake?
$D(x_{t-1}, x_t, t)$
Conditioning
Posterior sampling
$q(x_{t-1} \mid x_t, x'_0)$
x'_0
x'_{t-1}
z
$G(x_t, z, t)$

Experiments

Model	IS↑	FID↓	Recall↑	NFE ↓	Time (s) ↓
Denoising Diffusion GAN (ours), T=4	9.63	3.75	0.57	4	0.21
DDPM (Ho et al., 2020)	9.46	3.21	0.57	1000	80.5
NCSN (Song & Ermon, 2019)	8.87	25.3	-	1000	107.9
Adversarial DSM (Jolicoeur-Martineau et al., 2021b)	-	6.10	-	1000	-
Likelihood SDE (Song et al., 2021b)	-	2.87	-	-	-
Score SDE (VE) (Song et al., 2021c)	9.89	2.20	0.59	2000	423.2
Score SDE (VP) (Song et al., 2021c)	9.68	2.41	0.59	2000	421.5
Probability Flow (VP) (Song et al., 2021c)	9.83	3.08	0.57	140	50.9
LSGM (Vahdat et al., 2021)	9.87	2.10	0.61	147	44.5
DDIM, T=50 (Song et al., 2021a)	8.78	4.67	0.53	50	4.01
FastDDPM, T=50 (Kong & Ping, 2021)	8.98	3.41	0.56	50	4.01
Recovery EBM (Gao et al., 2021)	8.30	9.58	-	180	-
Improved DDPM (Nichol & Dhariwal, 2021)	-	2.90	-	4000	-
VDM (Kingma et al., 2021)	-	4.00	-	1000	-
UDM (Kim et al., 2021)	10.1	2.33	-	2000	-
D3PMs (Austin et al., 2021)	8.56	7.34	-	1000	-
Gotta Go Fast (Jolicoeur-Martineau et al., 2021a)	-	2.44	-	180	-
DDPM Distillation (Luhman & Luhman, 2021)	8.36	9.36	0.51	1	-
SNGAN (Miyato et al., 2018)	8.22	21.7	0.44	1	-
SNGAN+DGflow (Ansari et al., 2021)	9.35	9.62	0.48	25	1.98
AutoGAN (Gong et al., 2019)	8.60	12.4	0.46	1	-
TransGAN (Jiang et al., 2021)	9.02	9.26	-	1	-
StyleGAN2 w/o ADA (Karras et al., 2020a)	9.18	8.32	0.41	1	0.04
StyleGAN2 w/ ADA (Karras et al., 2020a)	9.83	2.92	0.49	1	0.04
StyleGAN2 w/ Diffaug (Zhao et al., 2020)	9.40	5.79	0.42	1	0.04

Experiments

Sample Quality vs Sampling Time Trade-off.



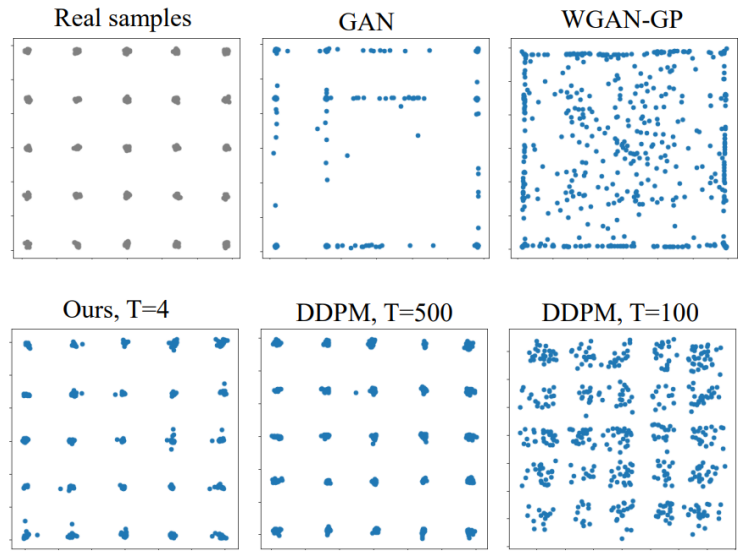
Experiments

Ablation Studies on CIFAR-10.

Model Variants	IS↑	FID↓	Recall↑
T = 1	8.93	14.6	0.19
T = 2	**9.80**	4.08	0.54
T = 4	9.63	**3.75**	**0.57**
T = 8	9.43	4.36	0.56
One-shot w/ aug	8.96	13.2	0.25
Direct denoising	9.10	6.03	0.53
Noise generation	8.79	8.04	0.52
No latent variable	8.37	20.6	0.42

- **direct denoising**: G_θ directly output denoised samples x_{t-1}
- **noise generation**: G_θ output the noise ϵ_t

Mode Coverage



Conclusions

Address Generative Learning Trilemma by Denoising Diffusion GAN

- High Quality & Mode Coverage
  Advantages of Inheriting from Diffusion Models.

- Fast Sampling than Diffusion Model
  Uses GAN to model the denoising distribution of the diffusion model as a multimodal distribution.

Denoising Diffusion GAN removes the normal distribution assumption and greatly reduces the required sampling steps.