KUBIG

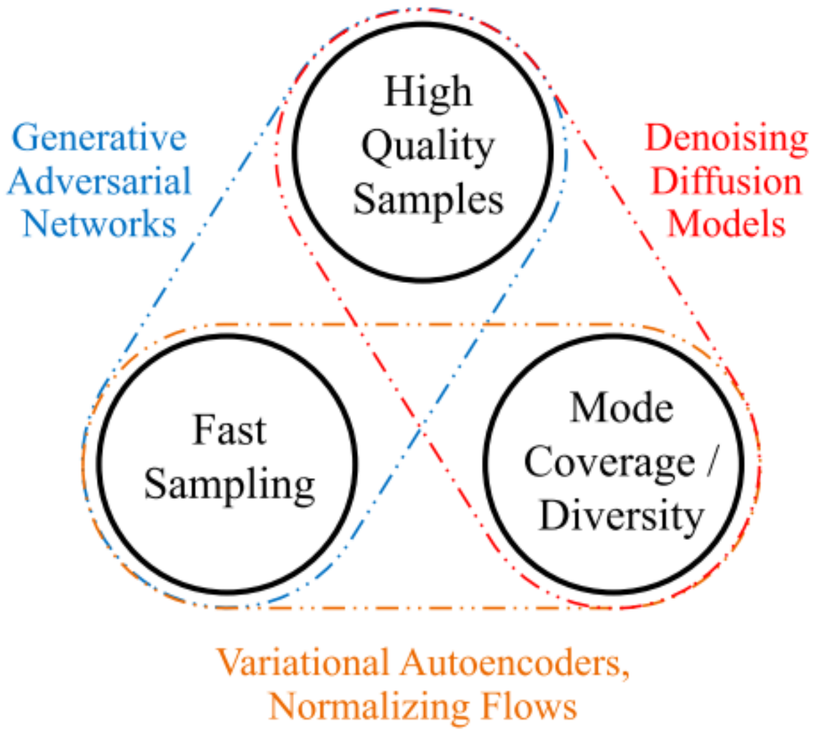
2023 Summer CV Paper Review Study

TACKLING THE GENERATIVE LEARNING TRILEMMA WITH DENOISING DIFFUSION GANS

15기 장수혁

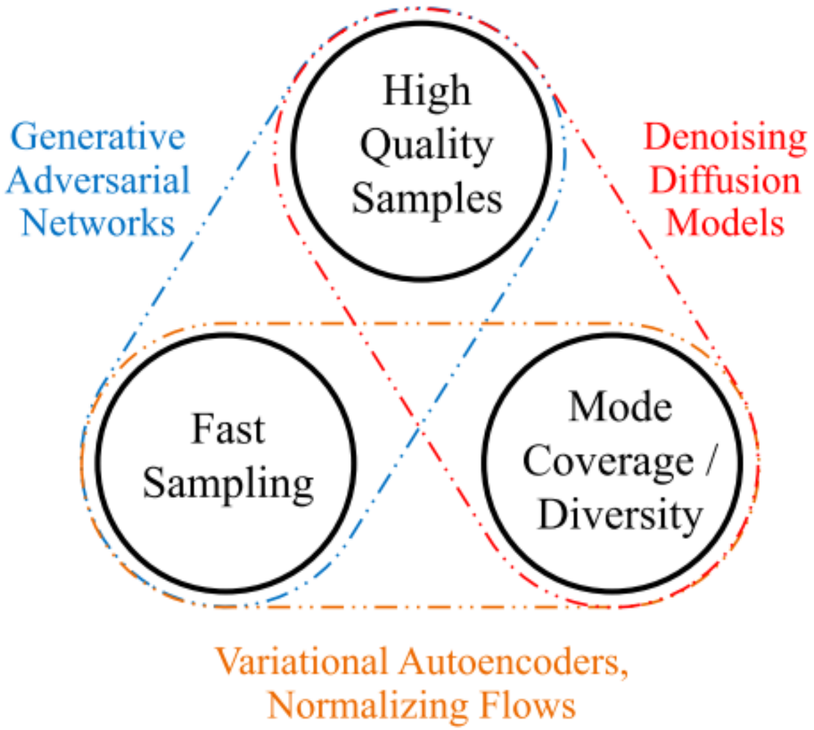
Contents

Generative Learning Trilemma



- **Current generative learning framework : 3개의 요소 동시에 만족 X**
- Most current works in image synthesis : High-Quality에 집중
- Mode Coverage / Diversity도 중요한 요소 중 하나
  - Better representing minorities
  - Reducing the negative social impacts
- Fast Sampling 또한 빼놓을 수 없는 요소
  - Applications such as interactive image editing or real-time speech synthesis

Generative Learning Trilemma



Model	High Quality	Mode Coverage/ Diversity	Fast Sampling
GANs	O	X	O
VAEs	X	O	O
Normalizing Flows	X	O	O
Diffusion Models	O (beat GANS)	O (High Likelihood)	X (Thousands of network evaluations)

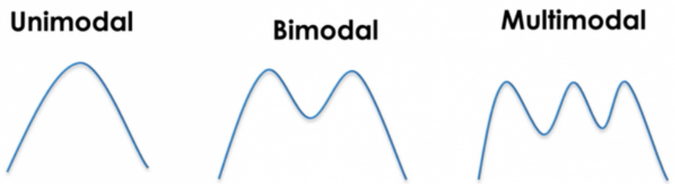
Denoising Diffusion Models for Fast Sampling

- **Denoising Distribution can be approximated by Gaussian Distributions**

- Gaussian assumption은 small denoising step의 infinitesimal limit에만 의존
  ➡ Large number of denoising steps

- **만약 reverse process에서 larger step size를 사용한다면? (= fewer denoising steps)**
  - **Modeling Denoising Distribution ➡ Non-Gaussian Multimodal Distribution**

  - **Multimodal Distribution : Multiple plausible clean images ➡ same noisy image**



Denoising Diffusion Models for Fast Sampling

- Parametrize the denoising distribution : Multimodal distribution
  ➡ **Enable denoising for large steps**

- Denoising Diffusion GAN
  - **Denoising distributions : conditional GANs 사용해 모델링**
  - **Sample Quality, Mode Coverage 부문에서는 다른 diffusion model과 비슷함**
  - **Sampling 속도는 2000배이상 빨라짐 on CIFAR-10 (predictor-corrector sampling과 비교)**
    - **Only two denoising steps 사용**
  - **Sample diversity도 SOTA GANs에 비해 훨씬 뛰어남 (Sample fidelity는 비슷함)**

Contributions

- Diffusion model의 slow sampling은 denoising distribution의 Gaussian assumption에서 기인
  ➡ **Complex & Multimodal denoising distribution을 제안**

- **Denoising Diffusion GANs**를 제안 : reverse process가 **conditional GANs**에 의해 parametrized 됨

- Denoising Diffusion GANs는 current diffusion model에 비해 **수십 배 이상 speed-up** 되는 것을 확인함
  - For both image generation and editing

DDPM

Forward Process

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t\geq 1} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}),$$

Reverse Process

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t\geq 1} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}),$$

Denoising model

$$\mathcal{L} = -\sum_{t\geq 1} \mathbb{E}_{q(\mathbf{x}_t)} \left[D_{\mathrm{KL}} \left(q(\mathbf{x}_{t-1}|\mathbf{x}_t) \| p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \right) \right] + C,$$

The goal of training : Maximize the ELBO

Matching the true denoising distribution q(X(t-1)|X(t)) & denoising model p(X(t-1)|X(t))

Denoising Score Matching 방식으로 trained된 score-based model의 형태와도 동일함
(입실론 추정의 문제 ➡ score function 추정의 문제로 치환)

Key Assumptions in Diffusion Models

- The denoising distribution p(X(t-1)|X(t)) is modeled with a Gaussian Distribution
- The number of denoising steps T = hundreds to thousands of steps

Multimodal Denoising Distributions for Large Denoising Steps

- Common assumption : Approximate q(X(t-1)|X(t)) with Gaussian Distribution

Is it really True?

$$q(x_{t-1}|x_t) \propto q(x_t|x_{t-1})q(x_{t-1})$$

Bayes' Rule

q(X(t)|X(t-1)) = Forward Gaussian Diffusion
q(X(t-1)) = Marginal Data Distribution

2가지 상황에서만 True denoising distribution이 Gaussian Form을 취한다고 할 수 있음!

Two case when the true denoising distribution is Gaussian

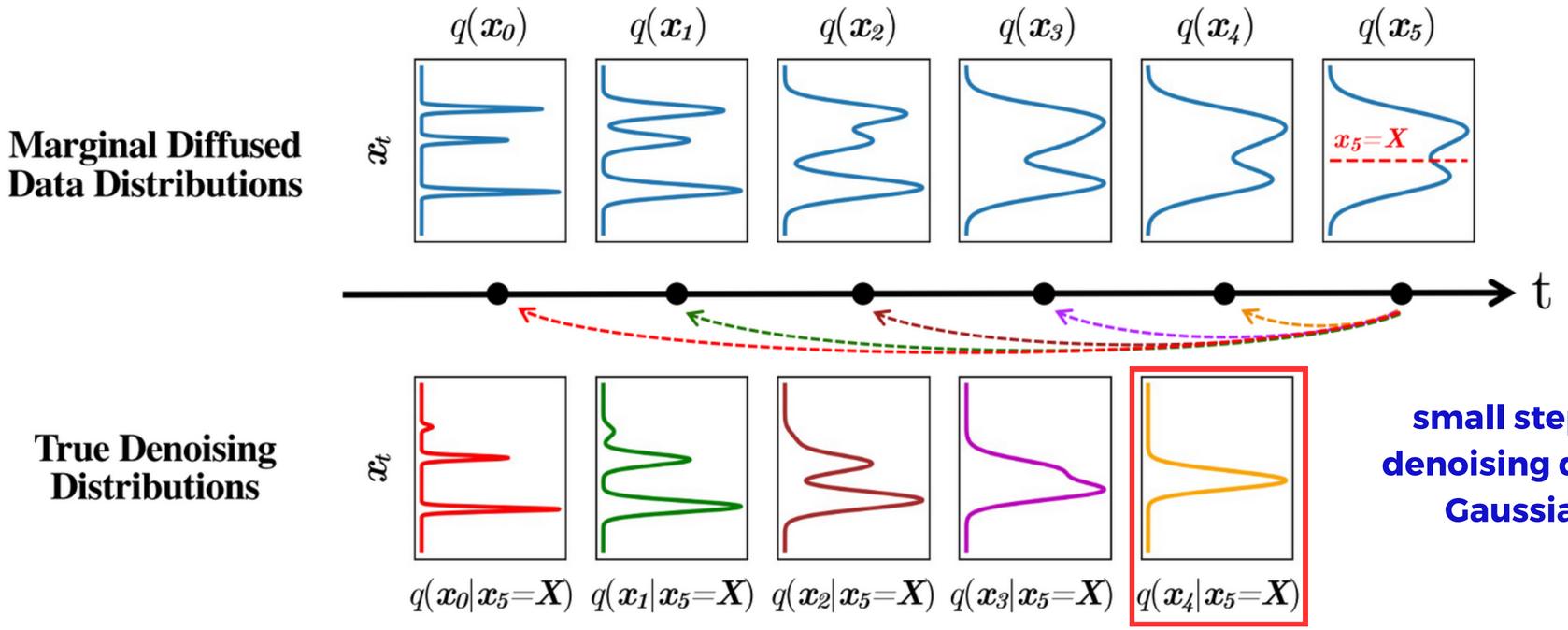
$$q(x_{t-1}|x_t) \propto q(x_t|x_{t-1})q(x_{t-1})$$

1. **Limit of infinitesimal step size B(t) ➡ q(X(t)|X(t-1))의 영향력 지배적**
   - reverse process가 곧 forward process
   - forward process는 Gaussian noise를 더하는 행위이므로 Gaussian distribution을 따를 수밖에 없고, denoising distribution도 Gaussian distribution을 따른다

2. **Data marginal q(X(t)) is Gaussian ➡ denoising distribution q(X(t-1)|X(t)) is Gaussian**

결론적으로, denoising step이 large하고 data distribution이 non-Gaussian이라면 denoising distribution이 Gaussian을 따른다고 보장할 수 없다!!

Multimodal Denoising Distributions for Large Denoising Steps



small step size에서는
denoising distribution이
Gaussian에 가깝다

- Denoising step gets larger ➡ true denoising distribution becomes more complex and multimodal

Modeling Denoising Distributions with Conditional GANs

- Goal : Reduce the number of denoising diffusion steps T = Increase the denoising step size
  ➡ Model the denoising distribution with an expressive multimodal distribution

- Conditional GANs이 image domain에서 복잡한 conditional distributions을 모델링하는데 사용되어서 True denoising distribution인 q(X(t-1)|X(t))를 근사하는데 적용

- 새로운 diffusion process
  - Forward diffusion 과정은 동일, 가정만 추가
    - Main assumption : T is assumed to be small (T<=8) & each diffusion step has larger B(t)
  - Backward Training : matching the conditional GAN generator (P(X(t-1)|X(t)) and q(X(t-1)|X(t))

$$\min_{\theta} \sum_{t \geq 1} \mathbb{E}_{q(\mathbf{x}_t)} \left[D_{\text{adv}} \left(q(\mathbf{x}_{t-1} | \mathbf{x}_t) \| p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) \right) \right]$$

Loss of Denoising Diffusion GANs

$$\min_{\theta} \sum_{t \geq 1} \mathbb{E}_{q(\mathbf{x}_t)} \left[D_{\text{adv}} \left(q(\mathbf{x}_{t-1} | \mathbf{x}_t) \| p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) \right) \right]$$

- **Adversarial loss : Minimize a divergence D_adv per denoising step**
  - Successful GAN framework 중 하나인 StyleGANs에서 널리 사용되는 **non-saturating GANs**을 사용함
  - D_adv는 여러 divergence가 될 수 있지만, 논문에서는 **softened reverse KL (f-divergence 일종) 사용**

Loss of Denoising Diffusion GANs

Adversarial Training : Time- dependent Discriminator

➡ X(t-1)과 X(t)를 input으로 받고 X(t-1)이 X(t)의 plausible denoised version인지 판별

Fake sample : from P(X(t-1)|X(t)) / Real sample : from q(X(t-1)|X(t))

$$\min_{\phi} \sum_{t \geq 1} \mathbb{E}_{q(\mathbf{x}_t)} \left[\mathbb{E}_{q(\mathbf{x}_{t-1}|\mathbf{x}_t)} [-\log(D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, t)] + \mathbb{E}_{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} [-\log(1 - D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, t))] \right],$$

q(X(t-1)|X(t)) is unknown!

$$\mathbb{E}_{q(\mathbf{x}_t) q(\mathbf{x}_{t-1}|\mathbf{x}_t)} [-\log(D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, t))] = \mathbb{E}_{q(\mathbf{x}_0) q(\mathbf{x}_{t-1}|\mathbf{x}_0) q(\mathbf{x}_t|\mathbf{x}_{t-1})} [-\log(D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, t))]$$

Markov Property 이용

Adversarial Training : Time- dependent Generator

➡ Updates the generator with the non-saturating GAN objective

$$\max_\theta \sum_{t \geq 1} \mathbb{E}_{q(\mathbf{x}_t)} \mathbb{E}_{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} [\log(D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, t))]$$

Parametrizing the Implicit Denoising Model

Instead of directly predicting X(t-1), parametrizing the Implicit denoising Model (DDIM)

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0 = f_\theta(\mathbf{x}_t, t))$$

Denoising Model

Posterior Distribution q(X(t-1)|X(t), X(0)) always has a Gaussian Form

- **Distribution over X(t-1) when denoising from X(t) to X(0)**
  (결론적으로, X(0)로부터 forward process 진행한다고 생각할 수 있음)
- **Independent of the step size & complexity of the data distribution**

Parametrizing the Implicit Denoising Model

기존 DDPM Denoising Model

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \int p_\theta(\mathbf{x}_0|\mathbf{x}_t) q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) d\mathbf{x}_0 = \int p(\mathbf{z}) q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0 = G_\theta(\mathbf{x}_t, \mathbf{z}, t)) d\mathbf{z},$$

DDIM Denoising Model imposed by GAN Generator

Sample made by Generator

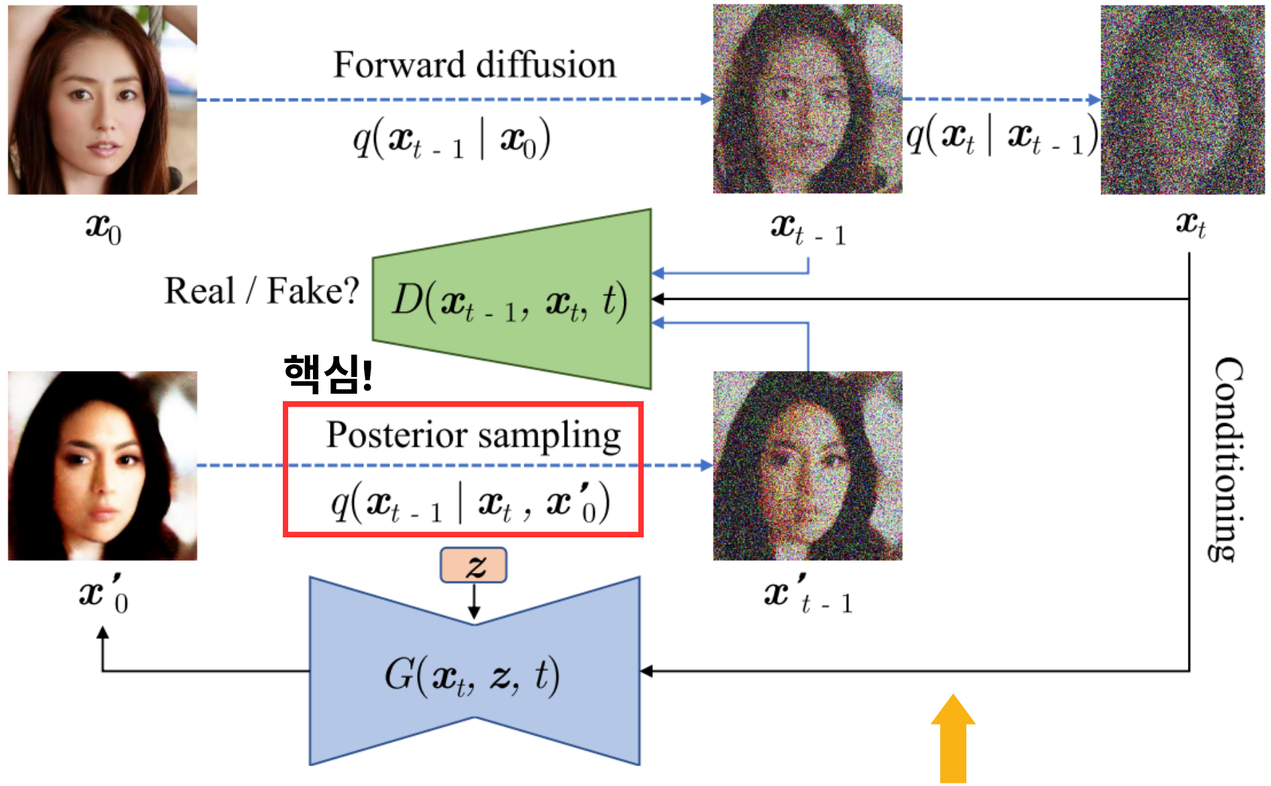
Latent Variable Z~N(0,I)과 X(t)를 받아 X(0) 생성

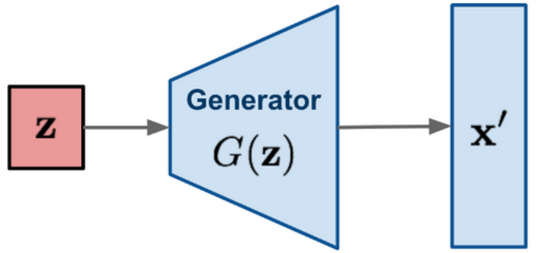
Then Why Parametrizing?

1. P(X(t-1)|X(t))가 DDPM과 유사하기 때문에 network structure에서 오는 inductive bias 가져올 수 있음
   - 차이점: DDPM은 x(0)가 x(t)의 deterministic한 mapping으로 예측됨 ➡ Unimodal Denoising Model
     하지만 Diffusion GANs에서는 random latent variable z와 Generator로부터 생성됨 ➡ Multimodal Denoising Model

2. Generator가 X(0)만 예측하면 되기 때문에, 시점 t가 달라져도 x(t-1)을 예측하는데 문제없다

Training Process of Denoising Diffusion GAN



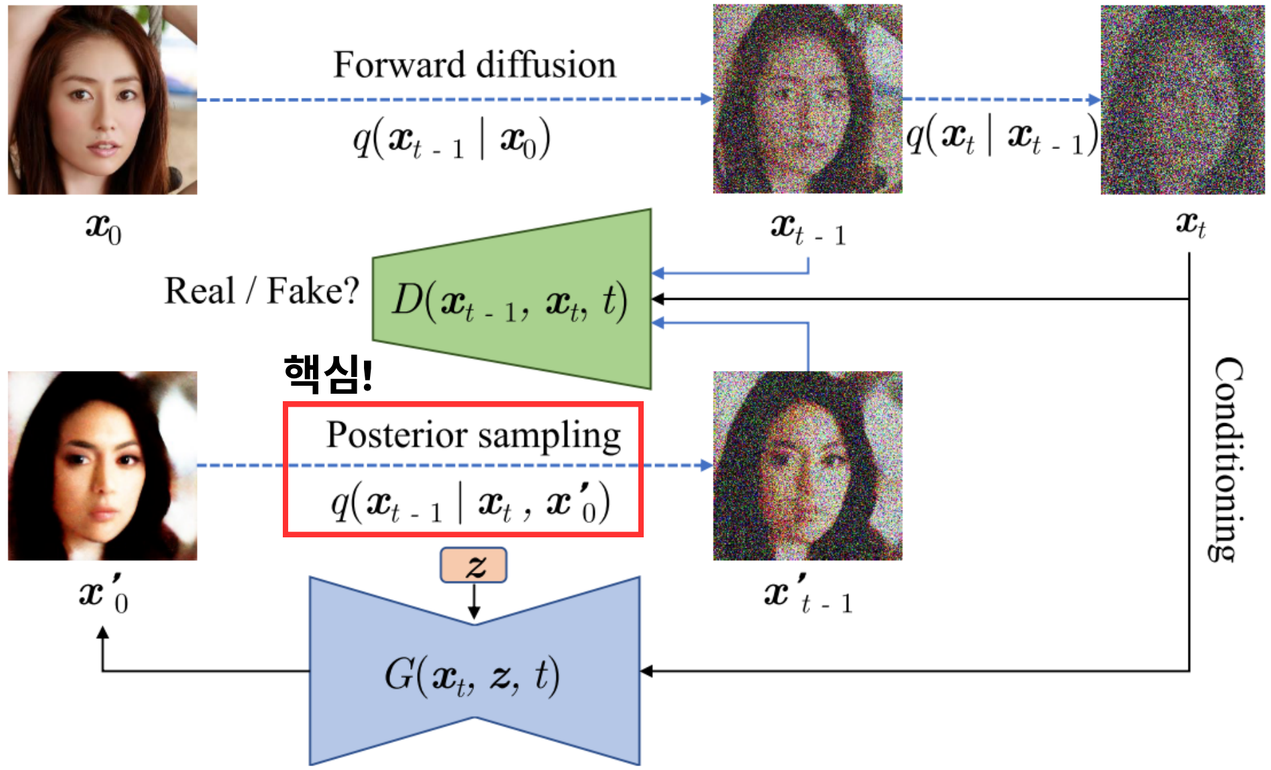


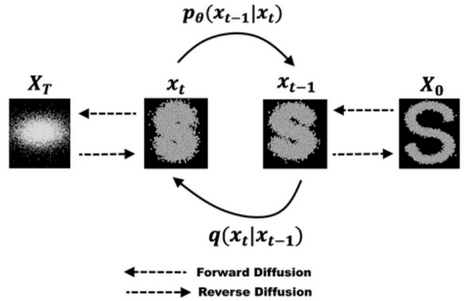
Why not one-shot generation?

- **GAN의 고질적 문제 : training instability and mode collapse**
- **Generating samples from a complex distribution in one-shot is difficult**
- **Overfitting : when the discriminator only looks at clean samples**

Several Conditional Denoising Diffusion Step 존재

Training Process of Denoising Diffusion GAN





Why iterative diffusion process?

- Each step in generation process is relatively simple
  ➡ strong conditioning on X(t)
- Smooth the data distribution
  ➡ prevent overfitting
- Better training stability and mode coverage

Results

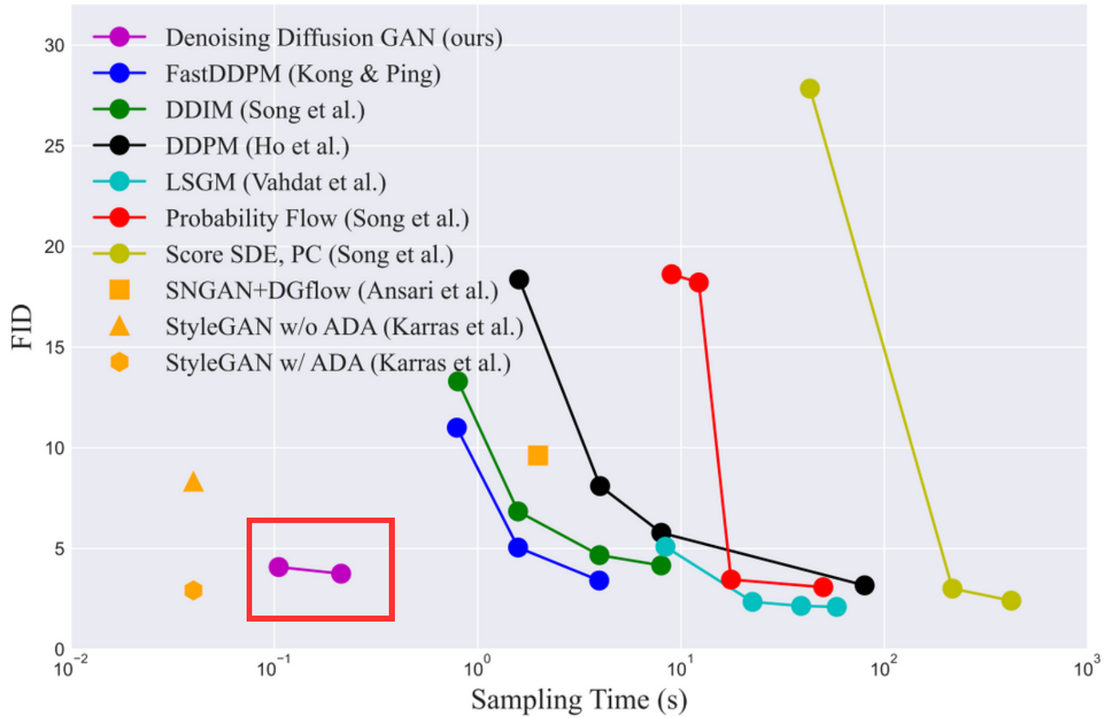


Figure 4: Sample quality vs sampling time trade-off.

Sample Diversity: Higher recall score than StyleGAN

Table 1: Results for unconditional generation on CIFAR-10.

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
Glow (Kingma & Dhariwal, 2018)	3.92	48.9	-	1	-
PixelCNN (Oord et al., 2016b)	4.60	65.9	-	1024	-
NVAE (Vahdat & Kautz, 2020)	7.18	23.5	0.51	1	0.36
IGEBM (Du & Mordatch, 2019)	6.02	40.6	-	60	-
VAEBM (Xiao et al., 2021)	8.43	12.2	0.53	16	8.79

Sample Quality : Best diffusion model, Best GAN과 유사

Sampling Time: 타 Diffusion model에 비해 압도적 속도 (PC와 2000배 차이)

THANK YOU