

CVPR 2021

High-Resolution Image Synthesis with Latent Diffusion Models

김대현

2023.06.01

목차

- **Abstract**
- Prior Approaches
- Proposed Solution
- Evaluation
- Conclusion

01 | Abstract



*"An astronaut riding a horse
in a photorealistic style"*

DALLE 2



*"A brain riding a rocketship
heading towards the moon"*

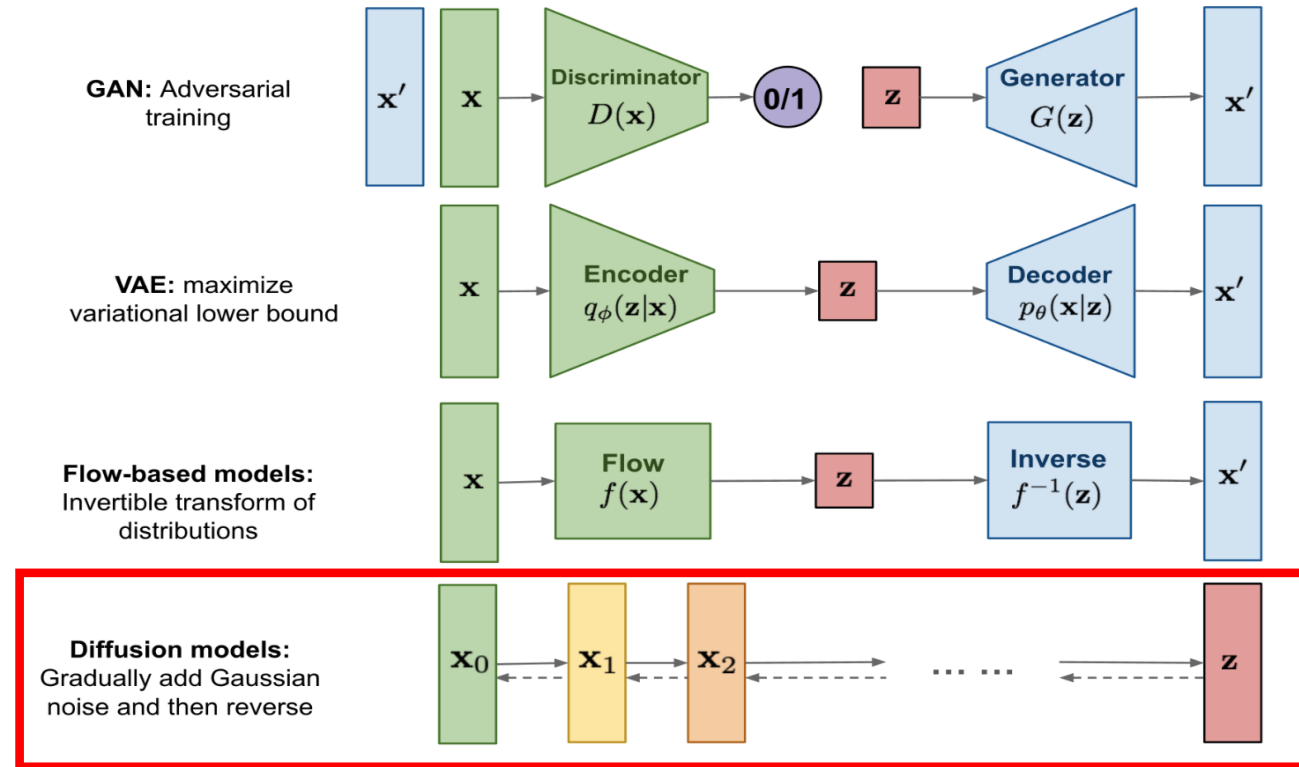
Imagen



- Contributions
 - 컴퓨터 사용 리소스 대폭 감소
 - 최근에는 이상치 탐지 backbone 모델로 활용
 - 이미지, 오디오 등 다양한 양식의 벤치마크 데이터에서 리더보드 상위권에 오르며 generation 성능을 확인함

목차

- Abstract
- **Prior Approaches**
- Proposed Solution
- Evaluation
- Conclusion

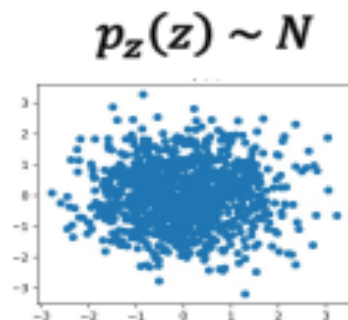


- 반복적 변화를 활용한다는 점에서 Flow-based models와 유사
- 분포에 대한 변분법적 추론 (Loss Function : ELBO 활용) 을 통한 학습을 진행한다는 점은 VAE와 유사
- 최근에는 Diffusion 모델의 학습에 Adversarial training을 활용하기도 함

02

Prior Approaches

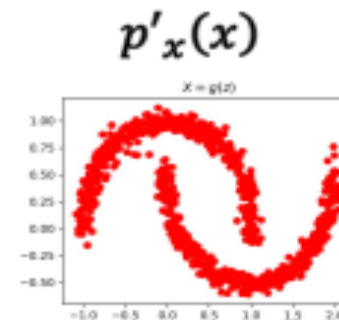
Generative model – Overview



- ✓ *Simple distribution*
- ✓ *Tractable(Gaussian)*

Input

Trained model



- ✓ *Complex distribution*
- ✓ *Visual, Audio patterns*

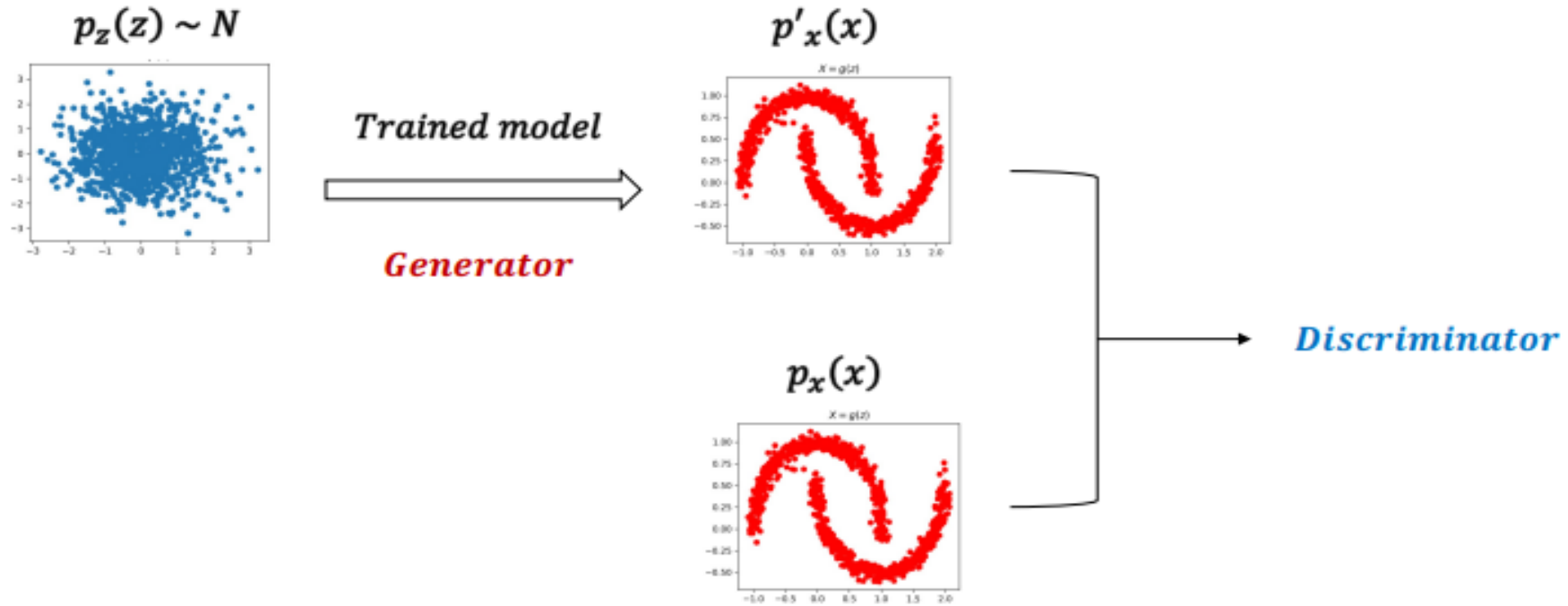
Output

- 결국 생성 모델로부터 원하는 것은 매우 간단한 분포(Z)를 특정한 패턴을 갖는 분포로 변환(mapping, transformation, sampling) 하는 것
- 대부분 생성모델의 목적은 주어진 입력 데이터로부터 latent variable(Z)을 얻어내고, 이를 변환하는 역량을 학습하는 것임

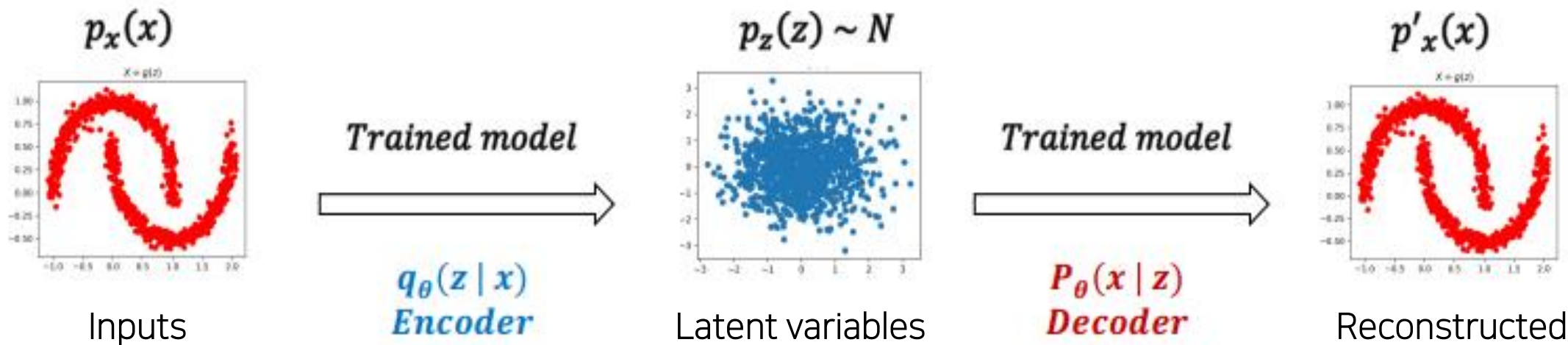
02

Prior Approaches

GAN - Overview



- 학습된 **Generator**를 통해 latent variable을 특정한 패턴의 분포로 mapping
- **Discriminator**를 모델 구조에 추가하여, **Generator**를 학습



- 배경 : 'Latent Space로부터 새로운 데이터를 생성해 낼 수는 없을까?' 에 대한 의문
- 문제 : $p_\theta(x)$ 의 값이 높아지는 방향으로 training 시킨다면, θ 는 시스템을 잘 표현하는 값으로 최적화 되지 않을까?
- 목표 : Latent vector Z 를 활용하여, x 와 유사한 새로운 데이터(x')를 생성하는 것
- Encoder를 모델 구조에 추가하여, Latent variable/Encoder/Decoder를 모두 학습

$$p_\theta(x) = \int p_\theta(z) \cdot p_\theta(x|z) dz$$

$$P_\theta(x) = \frac{P_\theta(x|z)P_\theta(z)}{P_\theta(z|x)} = \text{likelihood}$$

02

Prior Approaches

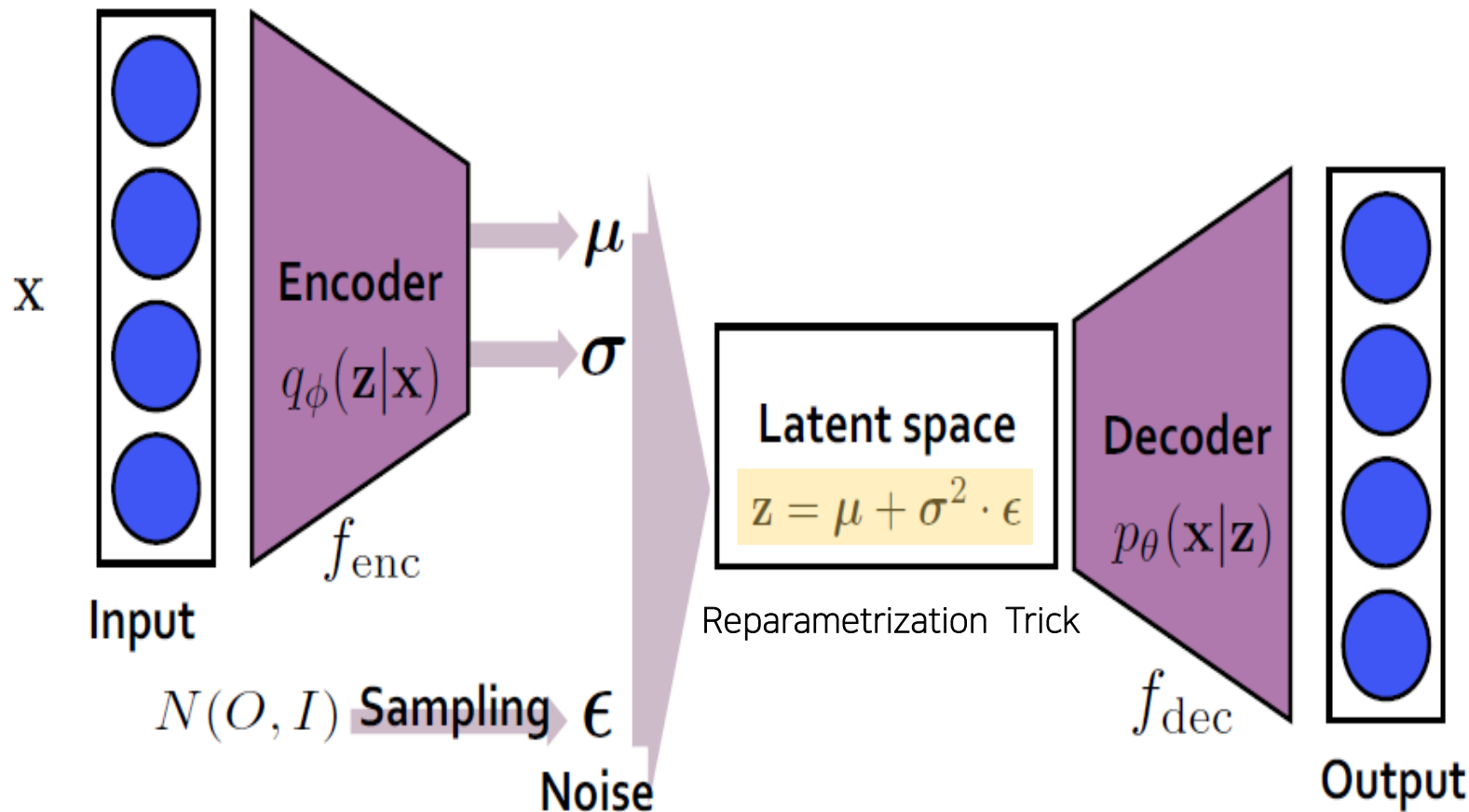
Variational Auto Encoder (VAE) – Architecture

- VAE 구성요소

- Encoder : Input을 Latent Space로 변환
- Decoder : Latent Space을 Input로 변환
- Latent Space : 숨겨진 vector
 - Noise를 Sampling 하여 이로부터 latent space를 만듦

- VAE 특성

- 이미지와 같은 고차원 데이터를 Encoding을 통해 저차원 Hidden 공간으로 변환
- 알기 어려운 사후확률(Posterior) 분포 $p(z|x)$ 를 다루기 쉬운 분포로 근사하는 변분법적 추론(variational Inference, $q(z|x)$)을 활용한 딥러닝 생성 모델



02

Prior Approaches

Variational Auto Encoder (VAE) – Loss Function

- VAE ELBO (Evidence Lower Bound) : 주어진 샘플 x 의 확률 분포를 잘 표현해보자

$$\begin{aligned}
 \log p_{\theta}(\mathbf{x}) &= \int q_{\phi}(\mathbf{z}|\mathbf{x}) \log p_{\theta}(\mathbf{x}) d\mathbf{z} && \text{1} && \text{Bayes rule 적용} \\
 &= \int q_{\phi}(\mathbf{z}|\mathbf{x}) \log \frac{p_{\theta}(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p_{\theta}(\mathbf{z}|\mathbf{x})} d\mathbf{z} && \text{2} && \text{분자, 분모에 동일하게 } q \text{항 곱} \\
 &= \int q_{\phi}(\mathbf{z}|\mathbf{x}) \log \frac{p_{\theta}(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p_{\theta}(\mathbf{z}|\mathbf{x})} \frac{q_{\phi}(\mathbf{z}|\mathbf{x})}{q_{\phi}(\mathbf{z}|\mathbf{x})} d\mathbf{z} && \text{3} && \text{Kullback-Leibler divergence 적용} \\
 &= \int q_{\phi}(\mathbf{z}|\mathbf{x}) \log p_{\theta}(\mathbf{x}|\mathbf{z}) d\mathbf{z} - KL(q_{\phi}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + KL(q_{\phi}(\mathbf{z}|\mathbf{x})||p_{\theta}(\mathbf{z}|\mathbf{x})) && \text{4}
 \end{aligned}$$

VAE의 maximum likelihood 증명

- Kullback-Leibler divergence : 두 확률 분포 간 다름의 정도(=relative entropy)

- 항상 양수
- 두 분포가 같을 때, 값 = 0
- 분포가 다를수록 값은 커짐

$$\begin{aligned}
 D_{KL}(p||q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \\
 &= \int p(x) \log p(x) dx - \int p(x) \log q(x) dx
 \end{aligned}$$

- VAE ELBO : Evidence lower Bound

$$\begin{aligned}
 \log p_{\theta}(\mathbf{x}) &= \int \overset{\textcircled{1}}{q_{\phi}(\mathbf{z}|\mathbf{x})} \log \overset{\textcircled{2}}{p_{\theta}(\mathbf{x}|\mathbf{z})} d\mathbf{z} - \overset{\textcircled{3}}{KL(q_{\phi}(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))} + KL(q_{\phi}(\mathbf{z}|\mathbf{x})||p_{\theta}(\mathbf{z}|\mathbf{x})) \\
 &\geq \int q_{\phi}(\mathbf{z}|\mathbf{x}) \log p_{\theta}(\mathbf{x}|\mathbf{z}) d\mathbf{z} - KL(q_{\phi}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \longleftarrow \text{Evidence lower bound (ELBO)} \\
 &\quad \text{Reconstruction error} \quad \text{Regularization error}
 \end{aligned}$$

$p_{\theta}(\mathbf{x}|\mathbf{z})$: \mathbf{z} 가 주어졌을 때 \mathbf{x} 를 근사하는 확률 (Decoder)

$p_{\theta}(\mathbf{z}|\mathbf{x})$: \mathbf{x} 가 주어졌을 때 \mathbf{z} 를 근사하는 확률 (Encoder)

① : Decoder 의미 (항의 값이 클수록 모델의 likelihood 값이 커짐)

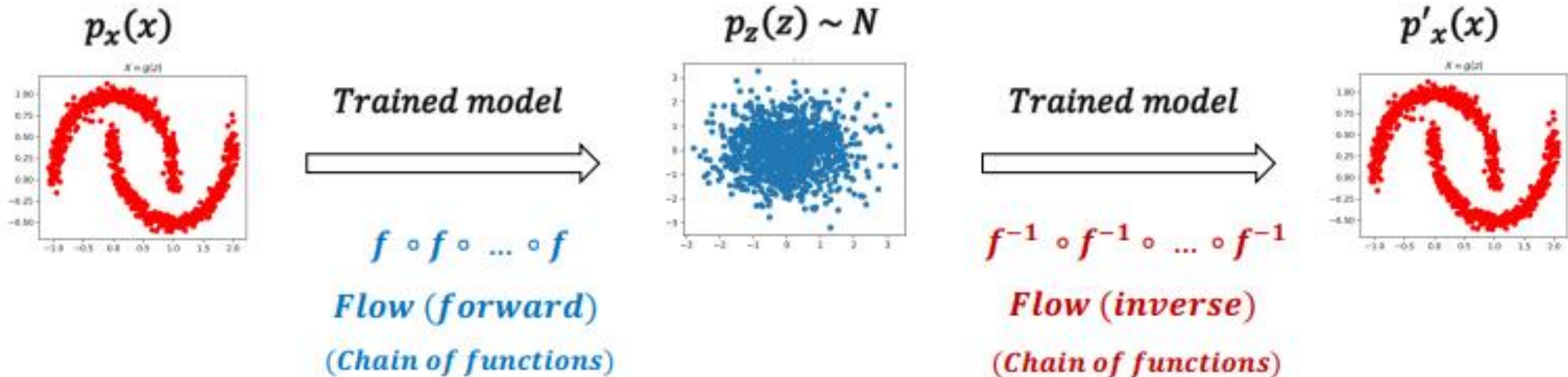
② : Encoder와 Gaussian 분포가 얼마나 유사한지 의미 (항의 값이 작을수록 모델의 likelihood 값이 작아짐)

③ : Intractable 함에 따라 구할 수 없음 (다만, KL divergence은 항상 양수 값 임에 따라 무시)

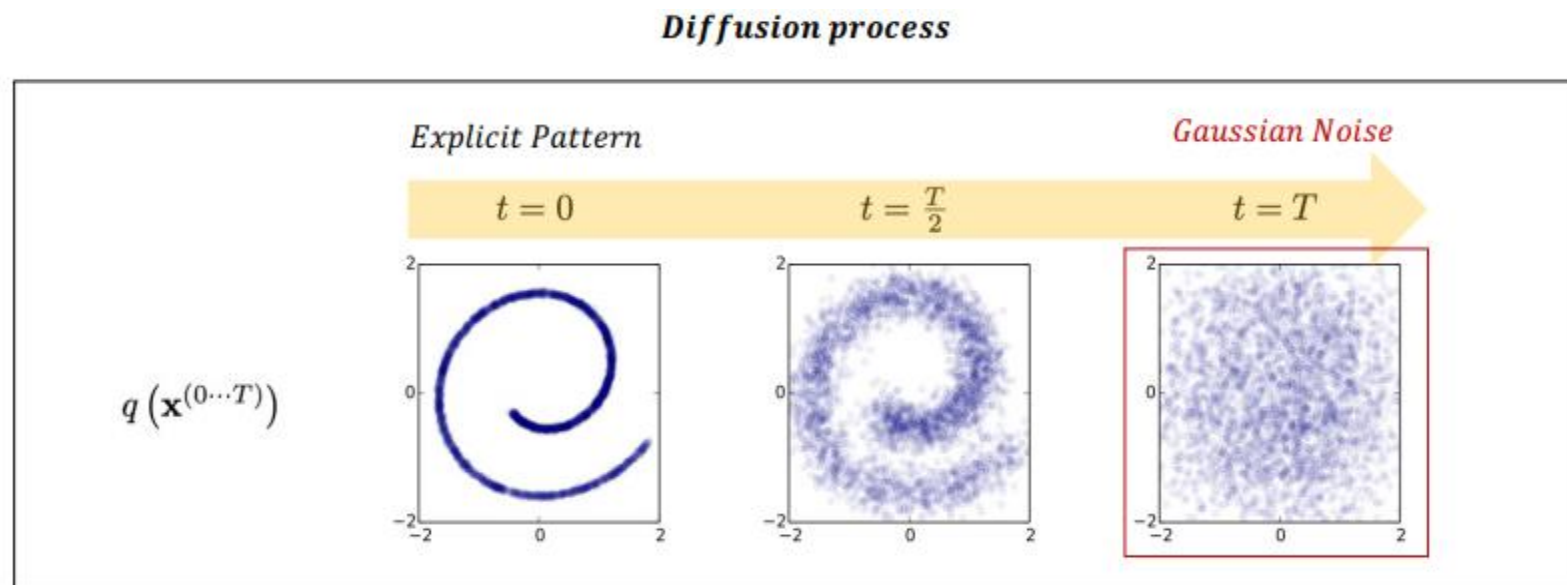
02

Prior Approaches

Flow-based Model – Overview



- 학습된 Flow model의 Inverse mapping을 통해 latent variable을 특정한 패턴의 분포로 mapping
- 생성에 활용되는 Inverse mapping을 학습하기 위해 Invertible function을 학습



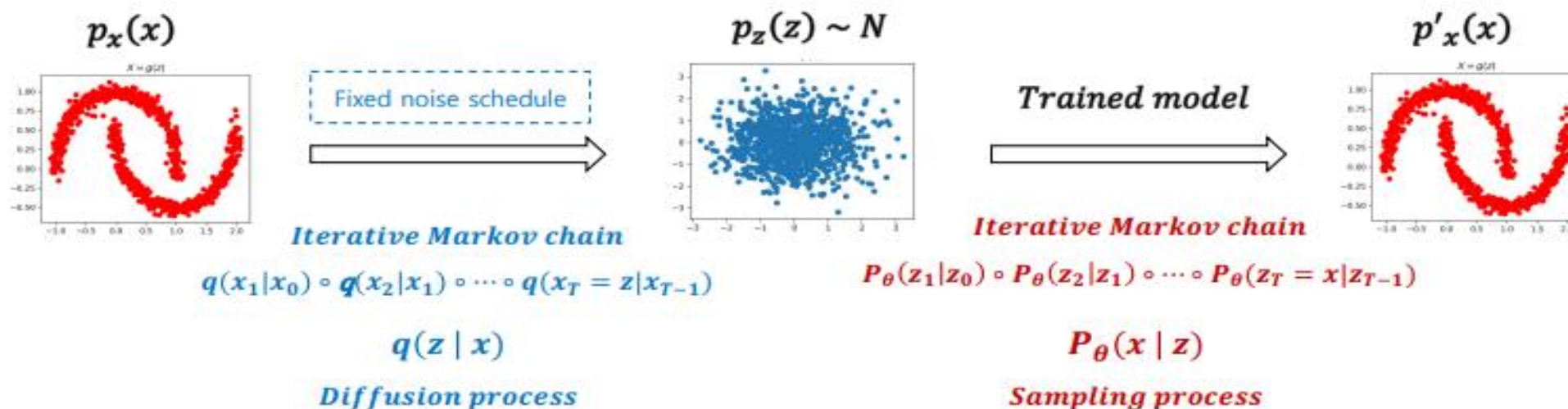
- 특정한 데이터의 패턴이 서서히 반복적인 과정을 거쳐 와해되는 과정('농도가 균일해지는')을 'Diffusion process'라 명명
- Deep Unsupervised Learning using Nonequilibrium Thermodynamics(2015, ICML)에서 비지도학습을 위한 방법론으로 첫 활용
- 대표적인 비지도학습 방법론인 이미지 생성 task에서 높은 성능을 보이며 주목을 받음

- Markov 성질을 갖는 이산 확률과정
 - Markov 성질 : 특정 상태의 확률($t+1$)은 오직 현재 (t)의 상태에 의존함
 - 이산 확률과정 : 이산적인 시간(0초, 1초, 2초, ...) 속에서의 확률적 현상

$$P[s_{t+1}|s_t] = P[s_{t+1}|s_1, \dots, s_t]$$



- ✓ 예 : “내일의 날씨는 오늘의 날씨만 보고 알 수 있다.” (내일의 날씨는 오로지 오늘의 날씨 만을 조건으로 하는 확률적 과정)

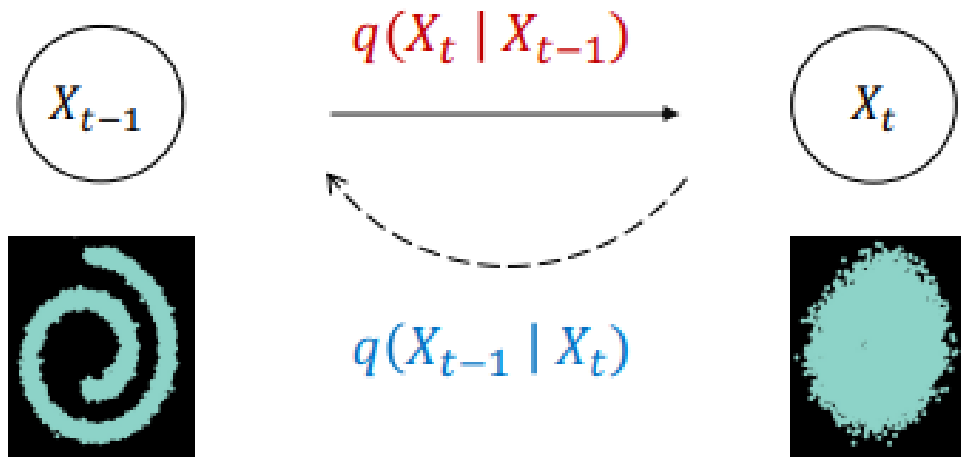


- 학습된 **Diffusion model**의 조건부 확률 분포 $P_\theta(x|z)$ 를 통해 특정한 패턴의 분포 획득
- 생성에 활용되는 조건부 확률 분포 $P_\theta(x|z)$ 을 학습하기 위해 Diffusion process $q(z|x)$ 를 활용

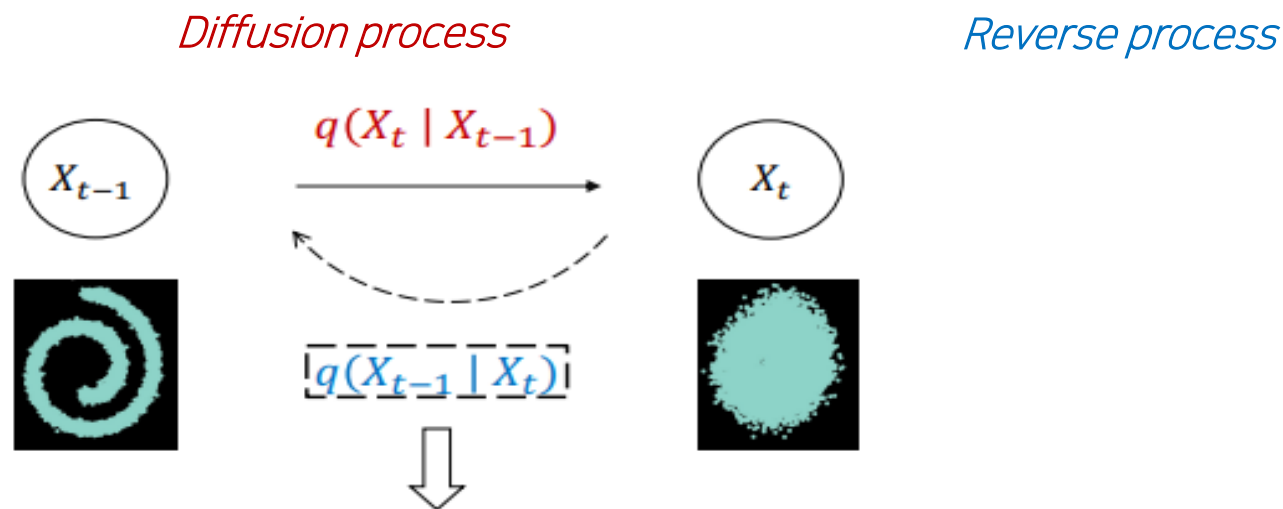
- Diffusion model은 Generative model로서 학습된 데이터의 패턴을 생성해내는 역할을 함
- 패턴 생성 과정을 학습하기 위해 고의적으로 패턴을 무너트리고(Noising), 이를 다시 복원하는 조건부 PDF를 학습(Denoising)

Diffusion process

Reverse process



- Diffusion model은 Generative model로서 학습된 데이터의 패턴을 생성해내는 역할을 함
- 패턴 생성 과정을 학습하기 위해 고의적으로 패턴을 무너트리고(Noising), 이를 다시 복원하는 조건부 PDF를 학습(Denoising)



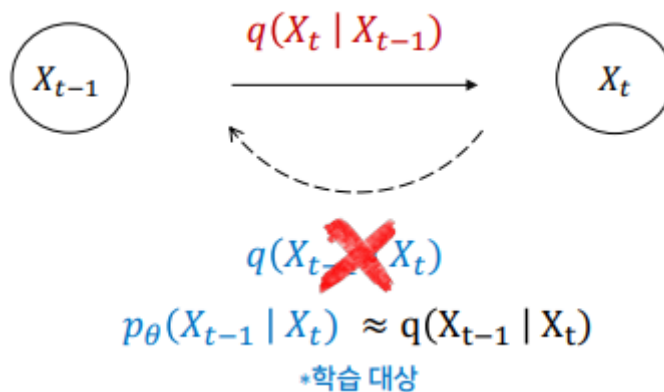
$q(X_t | X_{t-1})$ 로부터 바뀐 $q(X_{t-1} | X_t)$ 는 Inference과정에서 활용할 수 없음 (매우 복잡한 연산임)

다만, $q(X_t | X_{t-1})$ 이 Gaussian이라면 $q(X_{t-1} | X_t)$ 도 Gaussian이라는 점은 이미 증명됨* (β_t (노이즈 주입 정도)가 매우 작을 때)

- Diffusion model은 Generative model로서 학습된 데이터의 패턴을 생성해내는 역할을 함
- 패턴 생성 과정을 학습하기 위해 고의적으로 패턴을 무너트리고(Noising), 이를 다시 복원하는 조건부 PDF를 학습(Denoising)

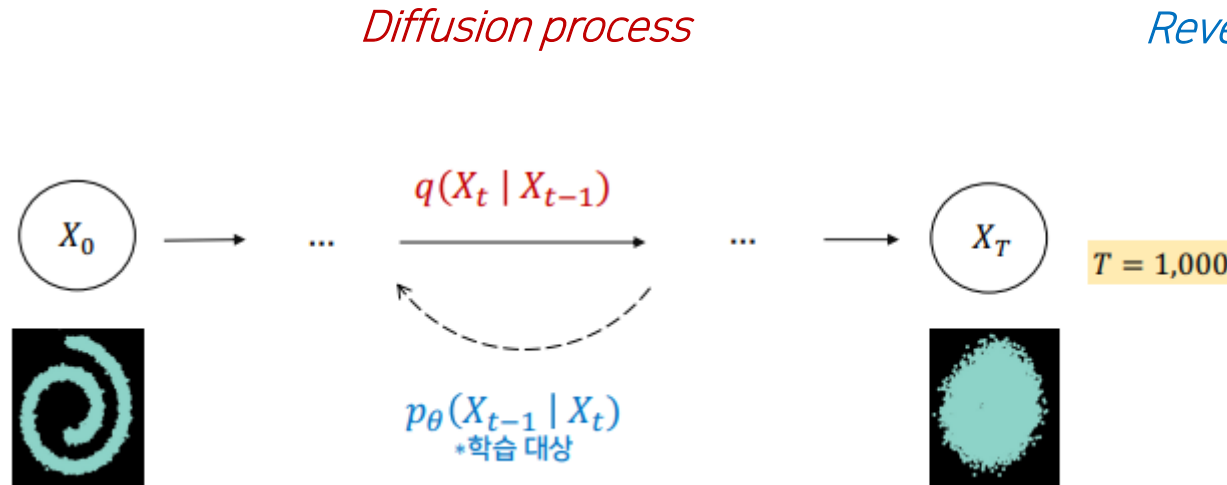
Diffusion process

Reverse process



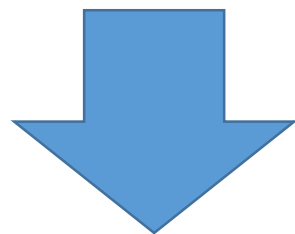
- $p_\theta(X_{t-1} | X_t)$ 를 상정하여 $q(X_{t-1} | X_t)$ 를 approximation
- 따라서, Diffusion model은 $p_\theta(X_{t-1} | X_t) \approx q(X_{t-1} | X_t)$ 되도록 학습

- Diffusion model은 Generative model로서 학습된 데이터의 패턴을 생성해내는 역할을 함
- 패턴 생성 과정을 학습하기 위해 고의적으로 패턴을 무너트리고(Noising), 이를 다시 복원하는 조건부 PDF를 학습(Denoising)



- 하지만, 이 변화(Noising, Denoising)를 하나의 단일 step transformation으로 학습하는 것은 매우 어려운 과제
- 이에, 2개의 각 Process(Diffusion, Reverse) 내 변화 과정은 Markov Chain으로 매우 많은 단계로 쪼개져 구성됨
- $p_\theta(X_t | X_{t-1})$ 의 학습은 결국 "large number of small perturbations"를 추정(estimate) 하는 것

Diffusion Model은 매우 큰 Computing Resource*를 요구함



DM의 성능을 유지하면서, Computing Demand를 감소시켜야 함

목차

- Abstract
- Prior Approaches
- **Proposed Solution**
 - **Introduction**
 - Method
- Evaluation
- Conclusion

3.1

Introduction

What's your purpose?

목표 : DM의 성능을 유지하면서, Computing Cost를 줄이는 것



- Denoising 과정에서 AutoEncoder를 사용
- Pixel 공간이 아닌, Latent Space에서 Denosing 함에따라, Computing Cost을 줄임
- 모델의 복잡성 감소 및 세부적인 표현 능력 상승 효과 달성
- Architecture에 Cross - Attention을 활용함으로써, 다른 도메인(Text, Audio ..) 을 모델상에서 함께 사용 가능

3.1

Introduction


IDEA : Departure to Latent Space

Importantly, the encoder downsampling the image by a factor

$$f = H/h = W/w,$$



Figure 1. Boosting the upper bound on achievable quality with **less aggressive downsampling**. Since diffusion models offer excellent inductive biases for spatial data, we do not need the heavy spatial downsampling of related generative models in latent space, but can still greatly reduce the dimensionality of the data via suitable autoencoding models, see Sec. 3. Images are from the DIV2K [1] validation set, evaluated at 512^2 px. We denote the spatial downsampling factor by f . Reconstruction FIDs [28] and PSNR are calculated on ImageNet-val. [12]; see also Tab. 8.

 : 원본이미지

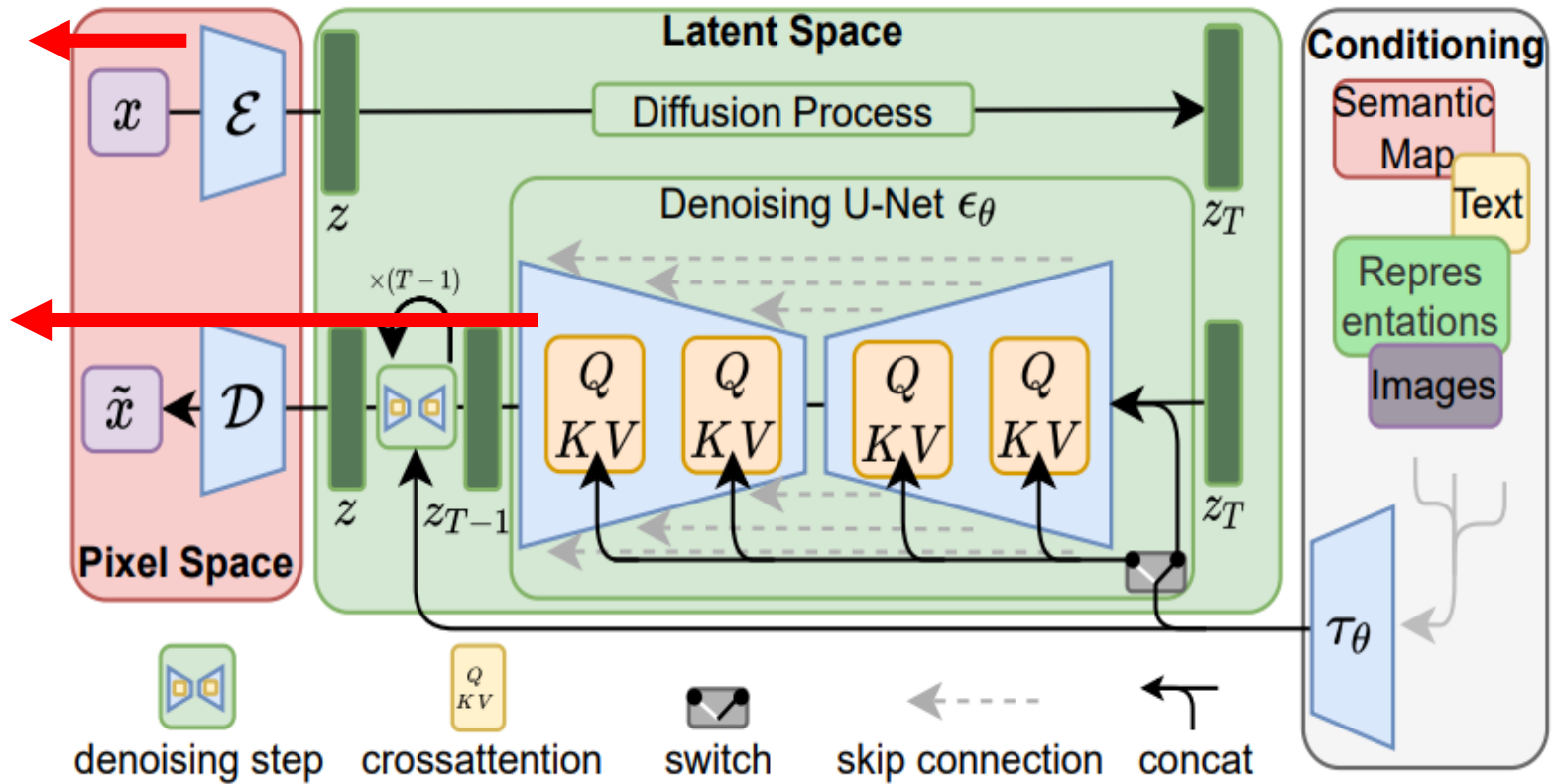
 : 이미지 복원 생성 결과

3.1

Introduction

Architecture

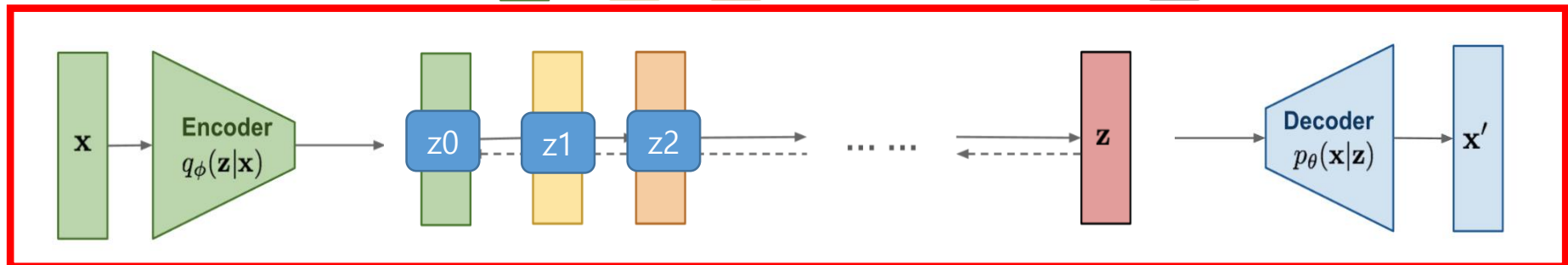
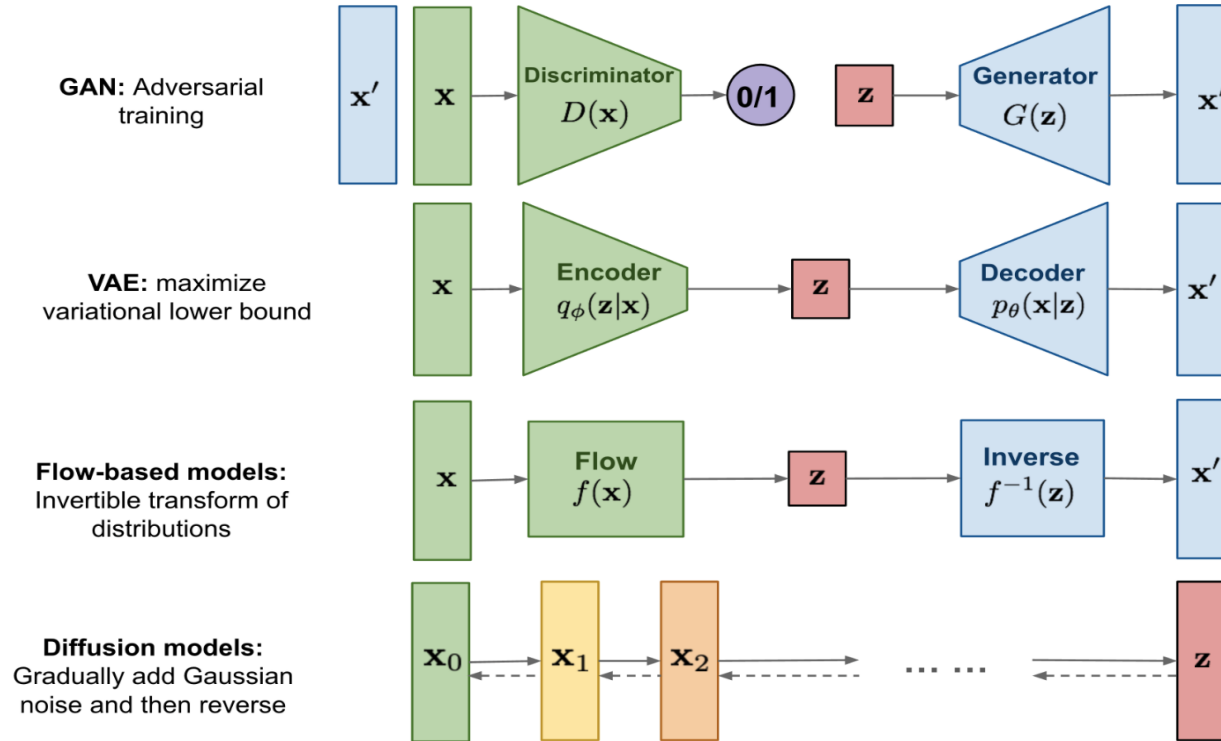
- First, we train an autoencoder which provides a lower-dimensional (and thereby efficient)
- We do not need to rely on excessive spatial compression, as we train DMs in the learned latent space, which exhibits better scaling properties with respect to the spatial dimensionality.
- We dub the resulting model class **Latent Diffusion Models (LDMs*)**
LDMs* : latent 공간에서 학습을 진행



3.1

Introduction

Architecture

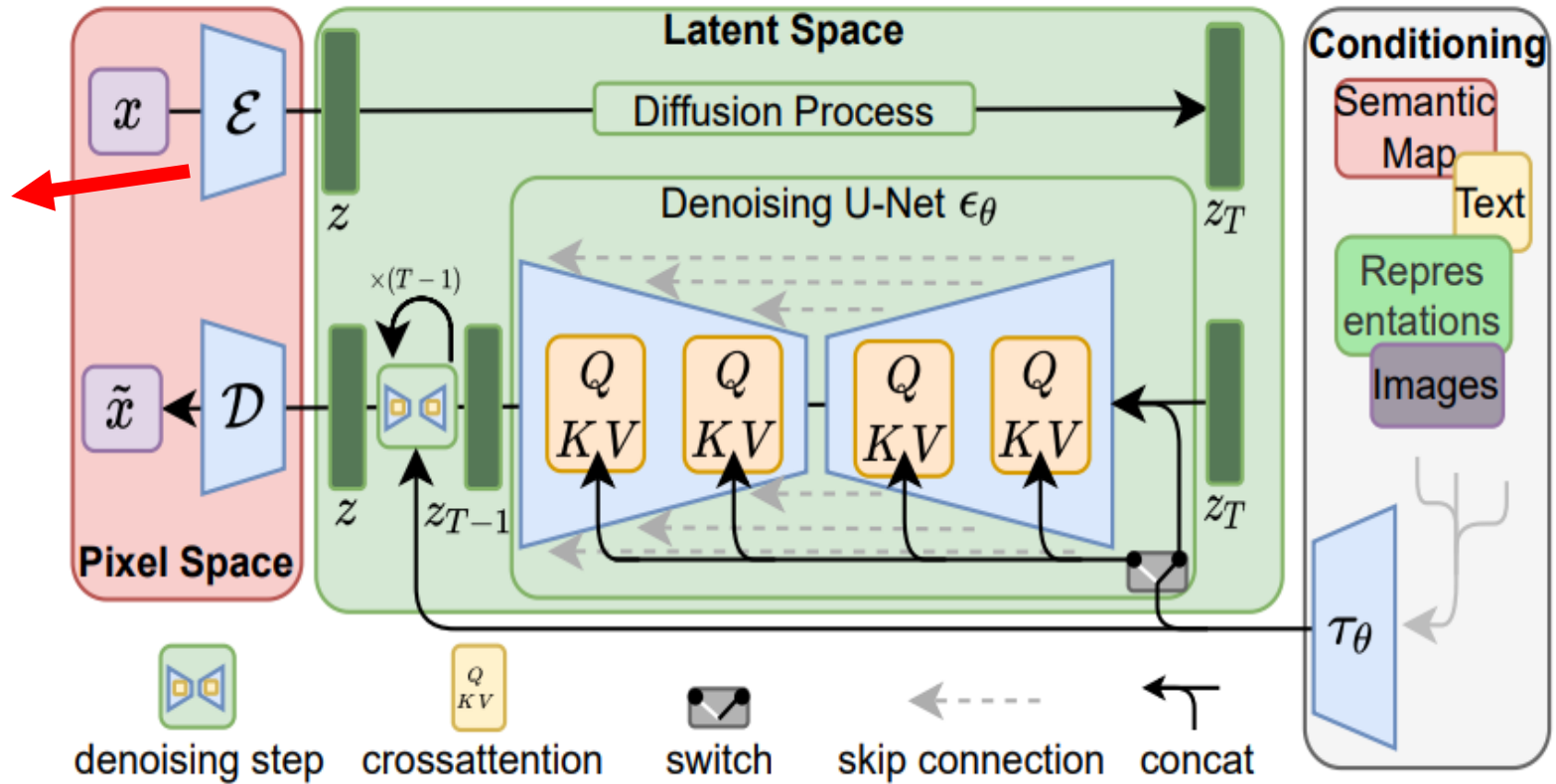


3.1

Introduction

Architecture

- A notable advantage of this approach is that we need to train the universal autoencoding stage only once and can therefore **reuse it for multiple DM trainings** or to explore possibly completely different tasks.

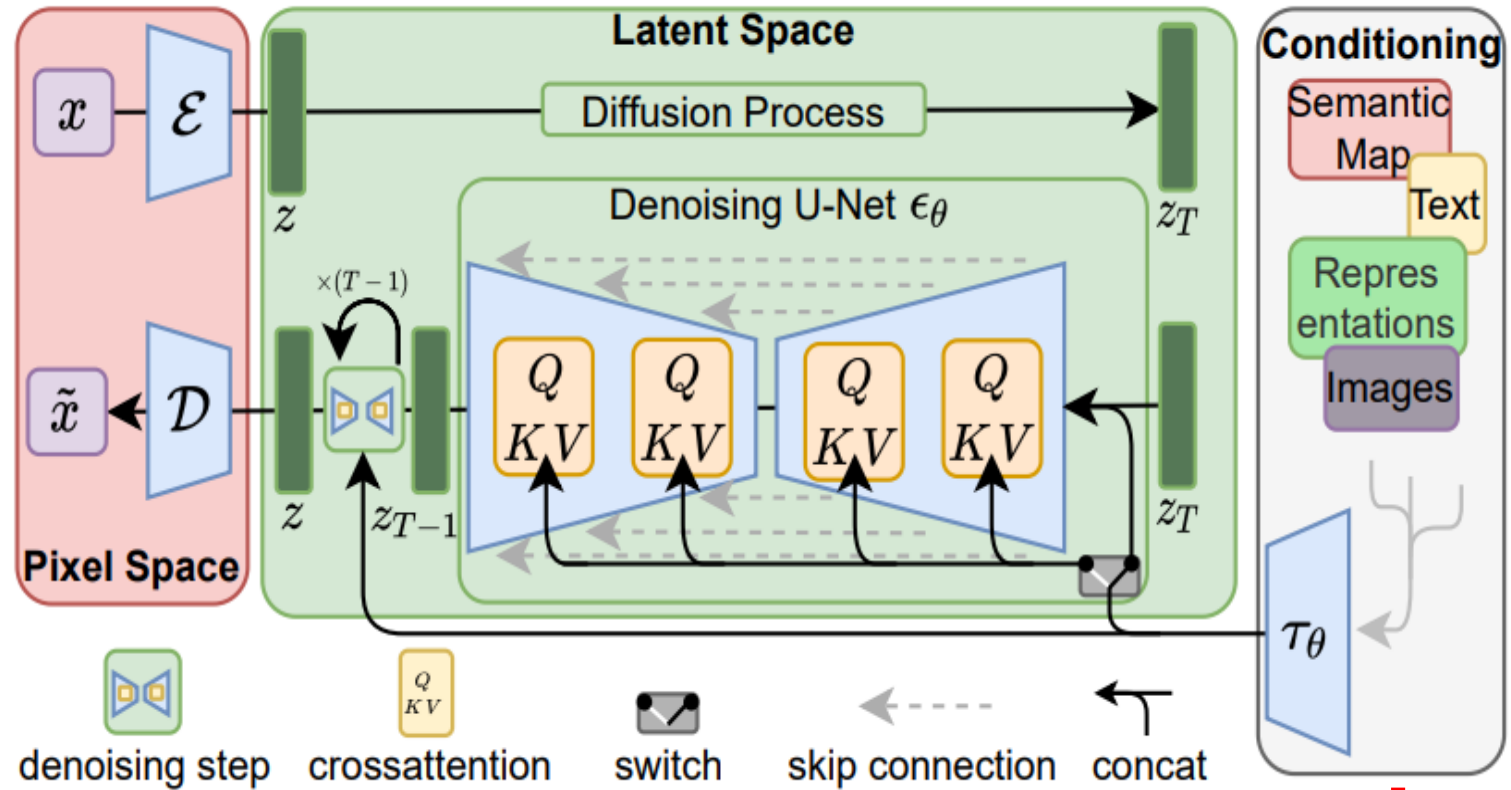


3.1

Introduction

Architecture

- We achieve competitive performance on multiple tasks (unconditional image synthesis, inpainting, stochastic super-resolution) and datasets while significantly lowering computational costs.
- Moreover, we design a general-purpose conditioning mechanism based on cross-attention, enabling multi-modal training. We use it to train class-conditional, text-to-image and layout-to-image models.



목차

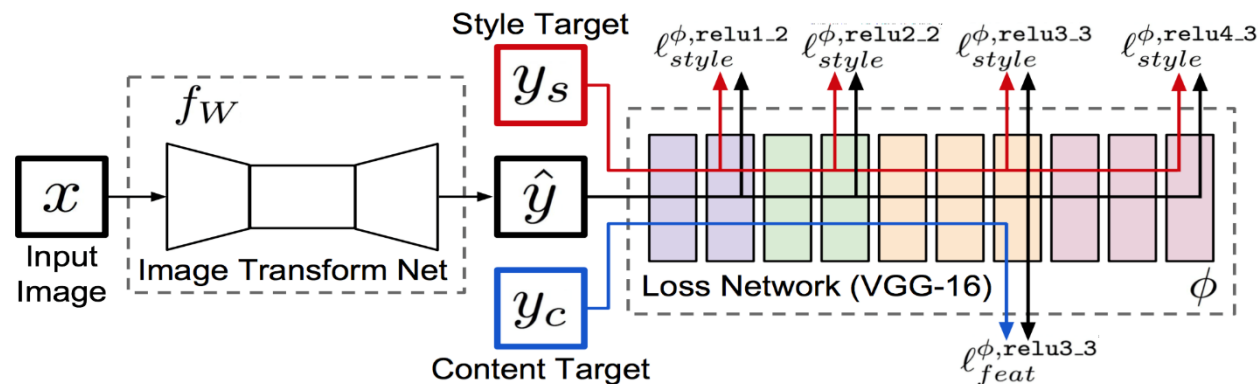
- Abstract
- Prior Approaches
- **Proposed Solution**
 - Introduction
 - **Method**
- Evaluation
- Conclusion

3.2

Method

Perceptual Image Compression (Pixel Space \leftrightarrow Latent Space)

Perceptual Loss

A patch-based
Adversarial objective

$$\ell_{feat}^{\phi,j}(\hat{y}, y) = \frac{1}{C_j H_j W_j} \|\phi_j(\hat{y}) - \phi_j(y)\|_2^2 \quad (= \text{Feature map마다 거리 계산})$$

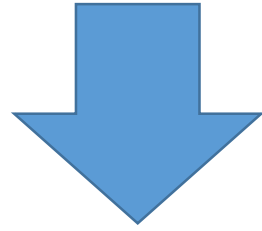
- 이미지 전체를 이용해서 판별하지 않고, 이미지 내의 패치를 특정 단위로 잘라서 T / F 판별
- 지역적인 사실성을 살릴 수 있음 (Enforcing local realism)
- L1, L2 Loss 처럼 픽셀 단위의 Loss에 집중했을 때, 나타날 수 있는 흐림 현상(Blurriness) 완화

3.2

Method

Perceptual Image Compression (Pixel Space \leftrightarrow Latent Space)

Autoencoder에서, latent space의 High-variance 최소화를 위한 두 가지 Regulation 방안

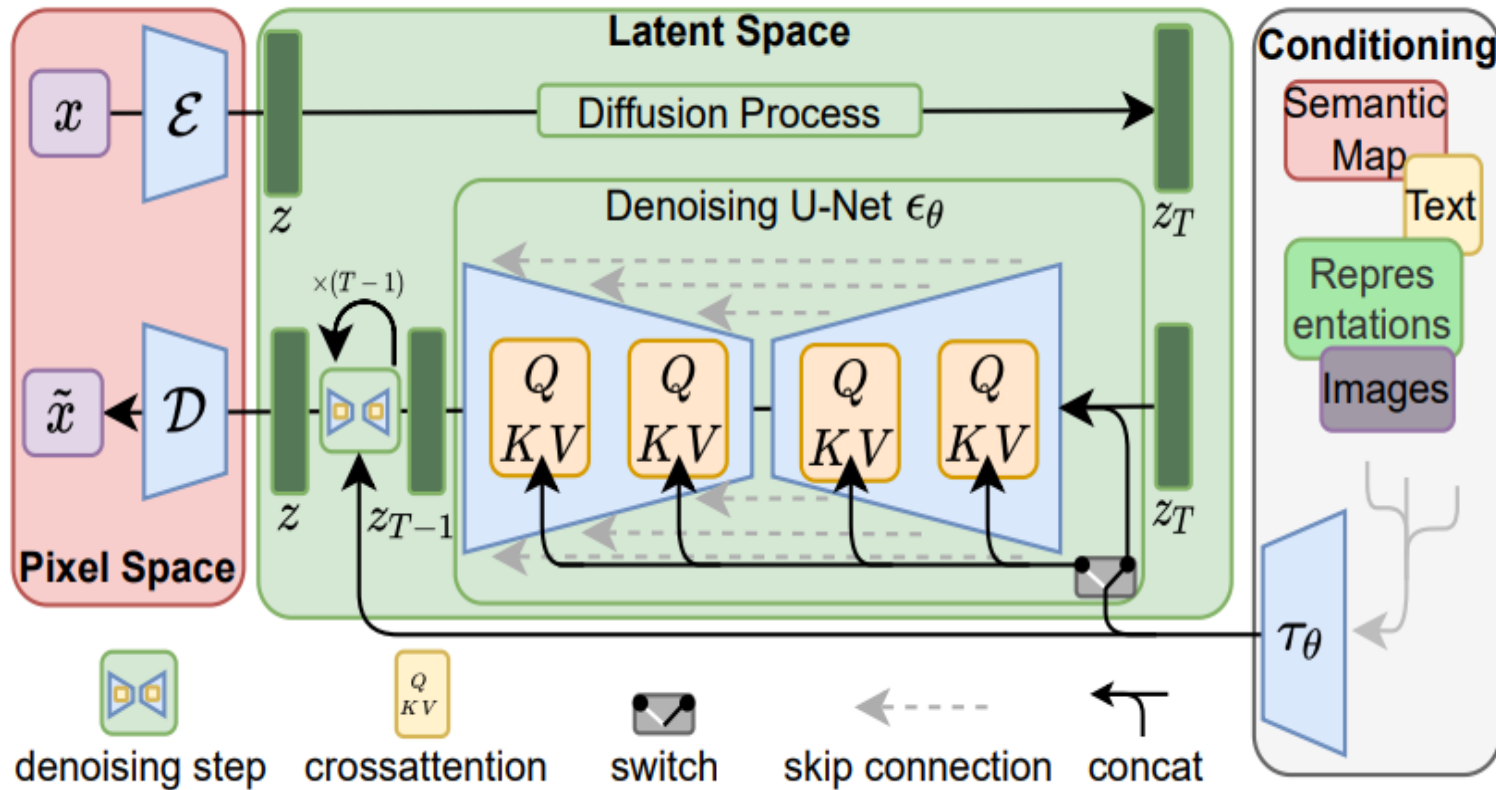


1. **KL-reg** : A small KL penalty towards a standard normal distribution over the learned latent, similar to VAE
2. **VQ-reg** : Uses a vector quantization layer within the decoder, like VQVAE but the quantization layer is absorbed by the decoder

3.2

Method

Latent Diffusion Models



- Latent Space(Low dimensional)

저자들은 High Dimensional Pixel Space보다 Low Dimension인 Latent Space에서의 연산이 훨씬 유리하다고 판단했으며, 아래와 같은 benefit을 기대

1. 데이터의 중요한 semantic bit에 집중할 수 있음
2. 훨씬 더 효율적인 계산가능

- Underlying Unet

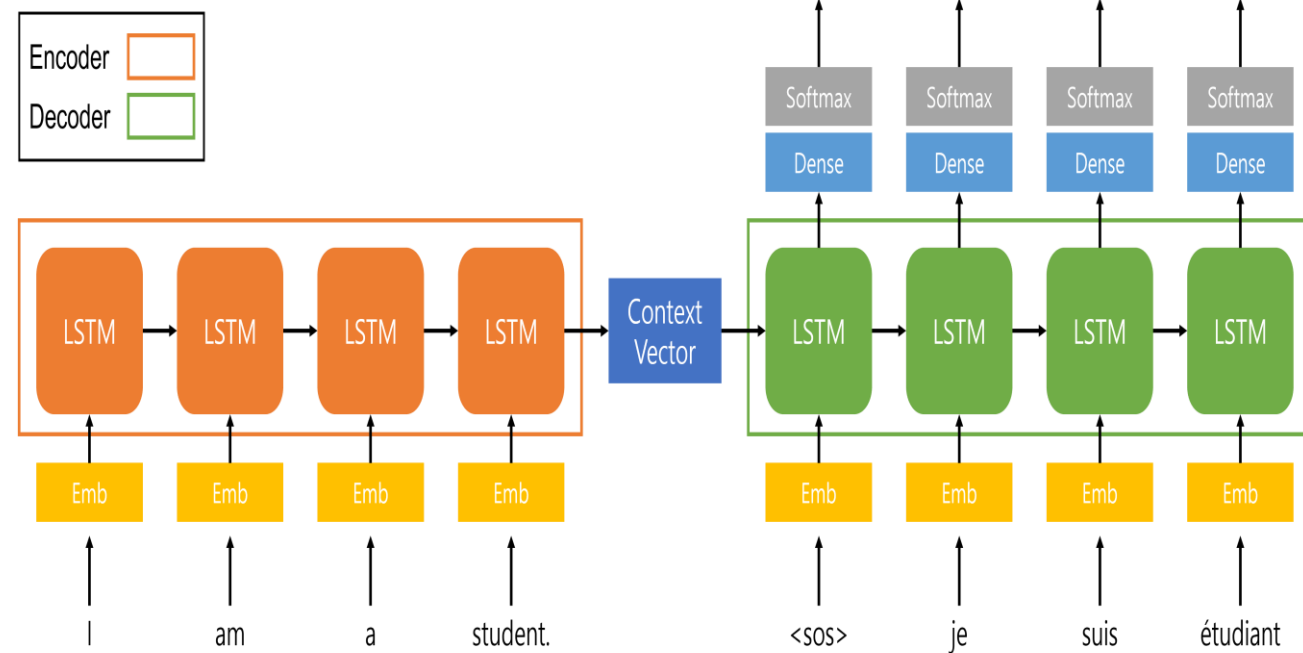
이미지별 Inductive Bias*를 활용할 수 있음
Time-conditional Unet 구조 활용 가능

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2 \right]$$

3.2

Method

Conditioning Mechanisms – Prerequisite (Cross Attention)



- 입력 시퀀스를 하나의 고정된 크기 벡터(컨텍스트 벡터)에 모두 압축 하는 과정에서 정보의 손실이 생길 수밖에 없음
- RNN 구조로 만들어진 모델이다 보니, 필연적으로 gradient vanishing/exploding 현상이 발생함

Cross Attention!

Decoder에서 다음 단어 예측을 위해 Encoder의 마지막 은닉 상태(Context Vector)뿐만 아니라,
Encoder의 매 시점 은닉 상태들을 모두 사용하자

3.2

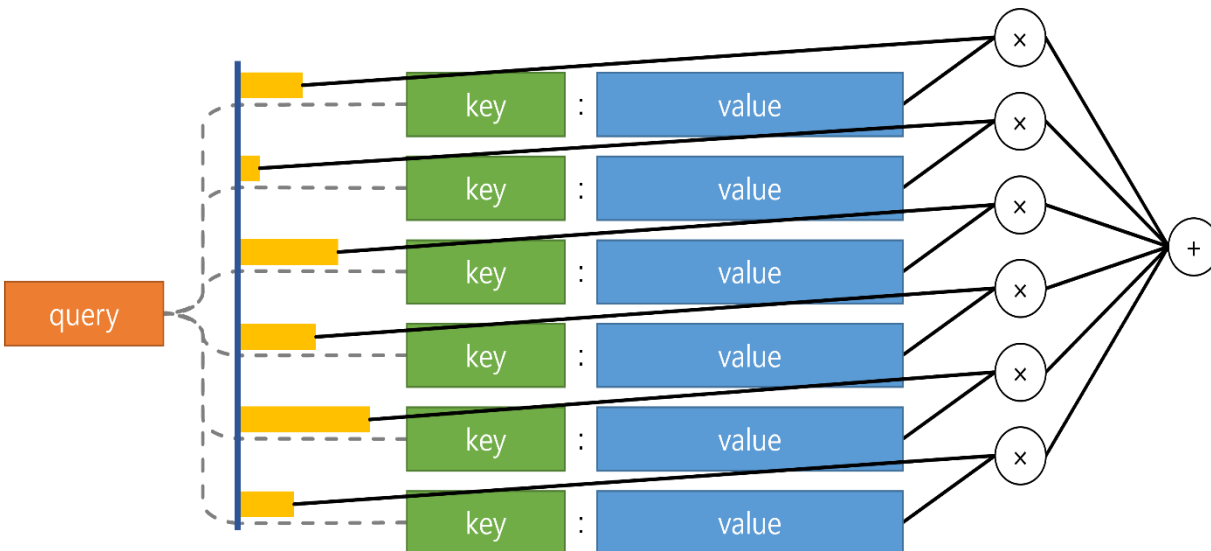
Method

Conditioning Mechanisms – Prerequisite (Cross Attention)

- Assumption

디코더가 단어 X 를 출력하기 직전의 디코더 은닉 상태는, 인코더가 입력 시퀀스에서 X 와 연관이 깊은 단어를 읽은 직후의 인코더 은닉 상태와 유사할 것이다.

- 목적 : query에 대해 value들에 집중할 지 결정하고자 함
- 방법 : query와 유사도가 높은 key를 찾음으로써, 어떤 value에 집중해야 하는지 파악



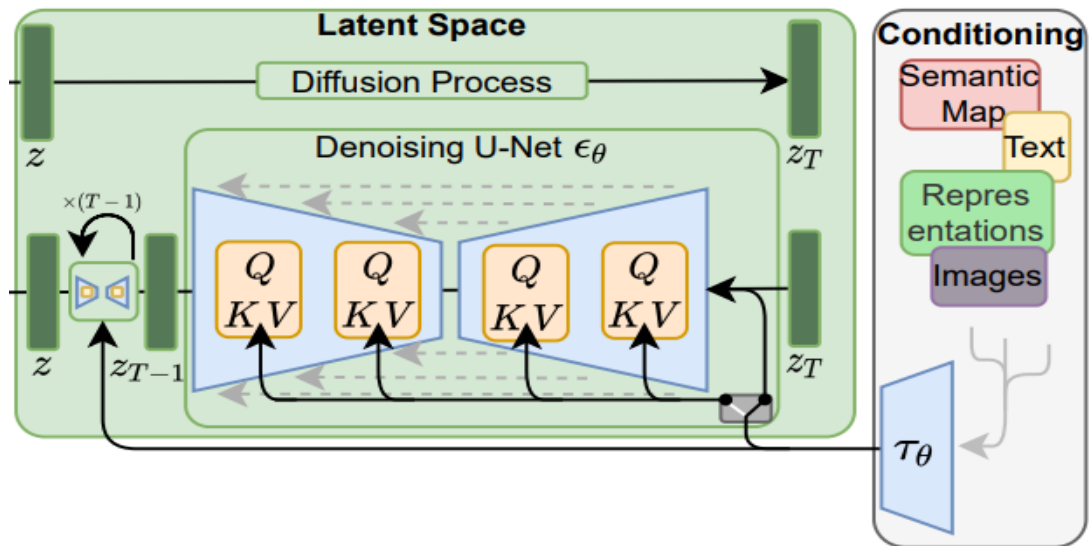
$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i$$

- 위 계산으로 얻어지는 어텐션 값(attention value)은 크기 (l, d_v) 인 행렬이 된다.
- 실제 구현할 때는 각 head별로 따로따로 h 번의 연산을 하는 것이 아니라 벡터 연산을 통해 한 번에 계산한다.

3.2

Method

Conditioning Mechanisms



- 다른 타입의 gen 모델과 유사하게 조건부 분포를 $p(z|y)$ 로 모델링
- 본 논문에서는 Unet backbone을 다양한 modality에 대한 conditioning을 가능케하기 위해 cross-attention mechanism을 도입
- τ_ϵ (Domain specific Encoder) : 다양한 종류의 y 를 전처리 하기 위한 도구(중간 representation) $\tau_\epsilon(y) \in \mathbb{R}^{M \times d_r}$.

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d}} \right) \cdot V$$

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t), \quad K = W_K^{(i)} \cdot \tau_\theta(y), \quad V = W_V^{(i)} \cdot \tau_\theta(y)$$

$z_t(\text{Image})$ $y(\text{text})$ $y(\text{text})$

- Q와 K를 dot product 연산한 뒤, softmax로 weight 형태 도출
- 이후, 이를 다시 V와 내적 연산

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2 \right]$$

목차

- Abstract
- Prior Approaches
- Proposed Solution
- **Evaluation**
- Conclusion

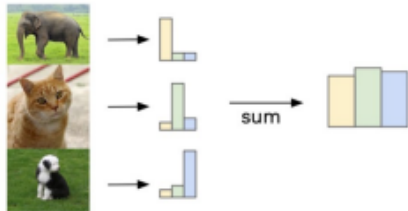
04

Evaluation

평가 척도(1) – Inception Score : the higher, the better

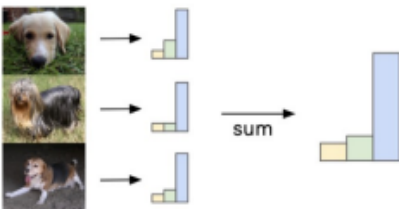
- Improved Techniques for Training GANs(2016)에서 제안된 generated image sample quality 평가 지표
- Inception image classifier를 활용해 'Inception'이라는 명칭을 사용함
- 2가지 특성을 통해 생성된 이미지의 품질을 평가. 아래 2가지를 만족할 수록 높은 score 획득

✓ Image quality : 생성된 이미지가 명확히 object를 표현할 수 있는지 (realistic?)



생성된 image들이 명확한 object를 갖는다면,
각각의 likelihood 분포의 summation은 균일한 분포를 가질 것

✓ Image diversity : 다양한 object가 생성되는지 ('dog' class : {Bulldog, Poodle, ... })



생성된 image들이 다양한 object를 가질 수 있다면,
각각의 likelihood 분포의 summation은 균일하지 않은 분포를 가질 것

둘 간의 KL divergence로
Inception Score 산출

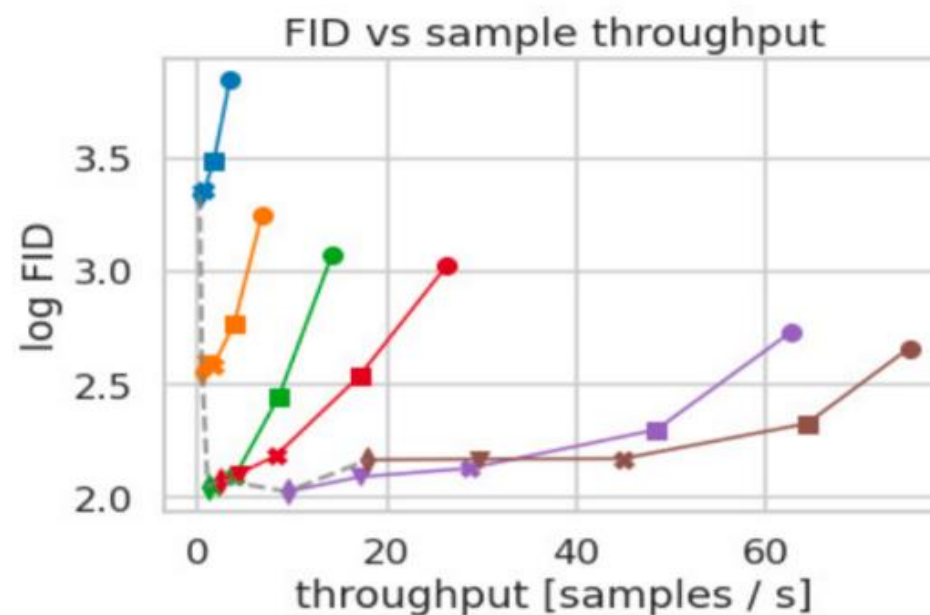
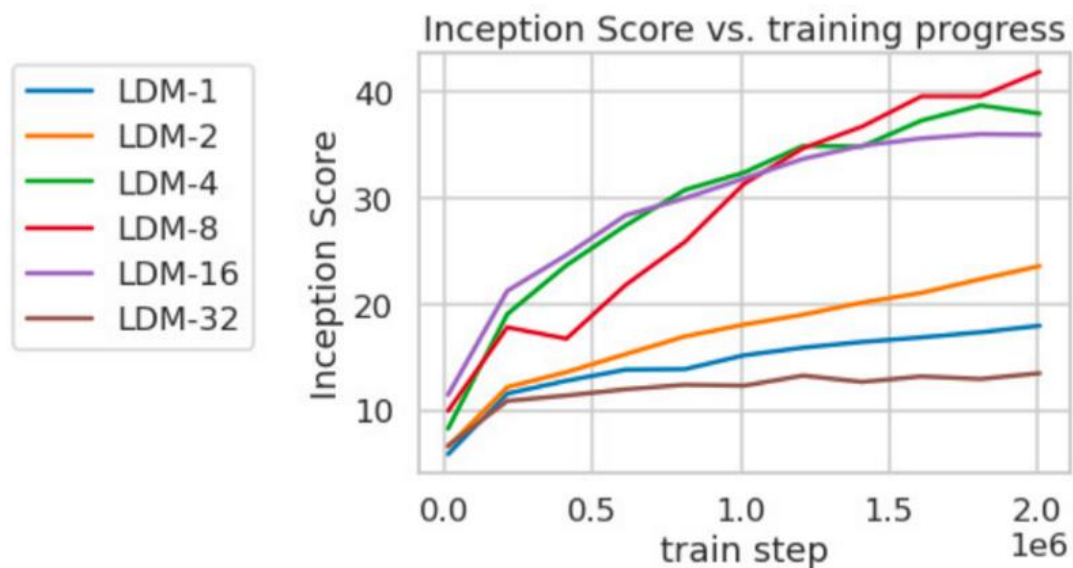
평가 척도(2) – FID(Frechet Inception Distance) score : the lower, the better

- Pre-trained Image classification model을 활용해 추출한 feature representation 간의 프레첷 거리(Frechet distance)를 score로 활용
 - Frechet Distance : 곡선을 이루는 points의 위치와 순서를 고려해 두 곡선 간 유사도를 측정하는 지표
- Feature 정보는 후반부 layer에서 추출한 high level representation을 활용
- FID score으로 계산은 아래와 같은 순서로 진행
 1. Pretrained image classification model define(ex: inception model)
 2. Compute embedding (feature representation)
 - 2.1 Embedding of real image
 - 2.2 Embedding of generated image
 3. Calculate 'Frechet distance' between 2.1 and 2.2

04

Evaluation

On Perceptual Compression Tradeoffs



$LDM-\{4-16\}$ 이 효율과 품질 간의 좋은 balance를 보였고, $LDM-4$, $LDM-8$ 이 high-quality 결과에 최적의 조건

04

Evaluation

Image Generation with Latent Diffusion (복원 test, FID 기준 SOTA 달성)



Figure 4. Samples from *LDMs* trained on CelebAHQ [39], FFHQ [41], LSUN-Churches [102], LSUN-Bedrooms [102] and class-conditional ImageNet [12], each with a resolution of 256×256 . Best viewed when zoomed in. For more samples *cf.* the supplement.

04

Evaluation

Conditional Latent Diffusion – (Text2Image)

Text-to-Image Synthesis on LAION. 1.45B Model.

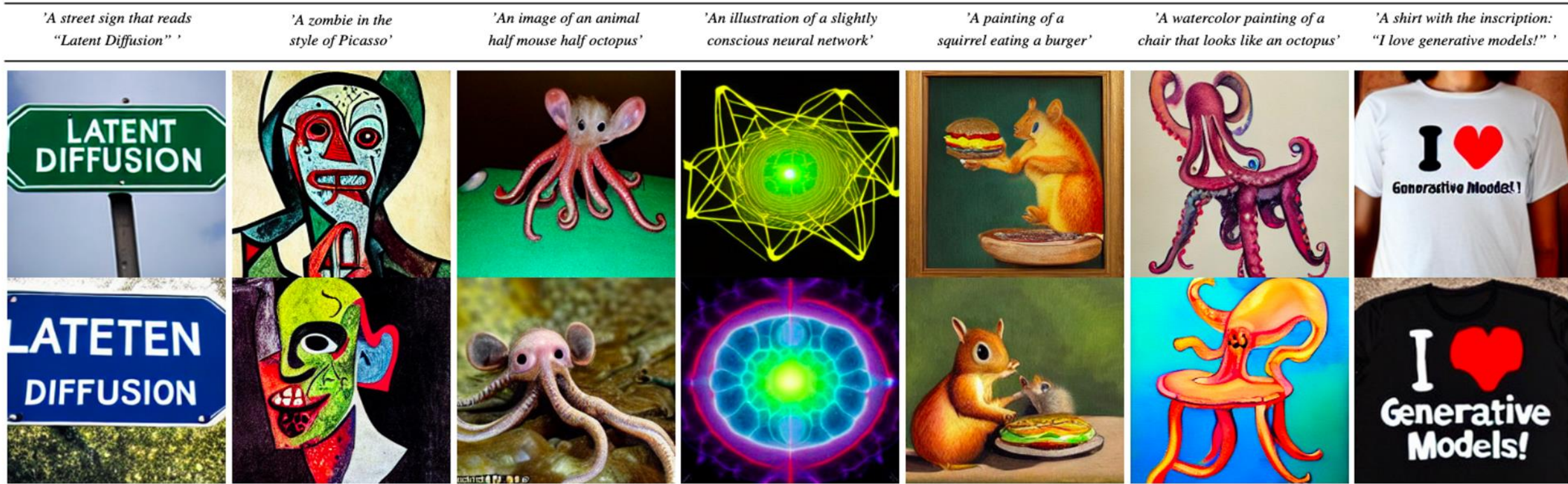


Figure 5. Samples for user-defined text prompts from our model for text-to-image synthesis, *LDM-8 (KL)*, which was trained on the LAION [78] database. Samples generated with 200 **DDIM** steps and $\eta = 1.0$. We use unconditional guidance [32] with $s = 10.0$.

04

Evaluation

Conditional Latent Diffusion – (Inpainting)



- LDM은 pixel based approach에 비해 computational demands를 크게 줄이지만, 샘플링 프로세스는 여전히 GAN보다 저조
- f=4 autoencoder 모델(*LDM-4*)에서 이미지 품질의 손실은 매우 작지만, pixel space에서 미세한 정확도가 요구되는 상황에서는 reconstruction capability에 대한 bottle-neck을 일으킬 수 있음

목차

- Abstract
- Prior Approaches
- Proposed Solution
- Evaluation
- **Conclusion**

05 Conclusion

- Quality를 저하시키지 않고, Denoising Diffusion Model의 학습 및 샘플링 효율성을 크게 향상
 - ✓ 이러한 Benefit을 간단하고 효율적인 방법인, LDM : Latent Diffusion Model 제안
- Cross-attention Conditioning Mechanism을 기반으로 Tasks 마다의 별도 architecture 없이 광범위한 Conditional image Synthesis tasks에 SOTA 모델들과 손색 없음
 - ✓ Unconditional Image Synthesis
 - ✓ Super-Resolution
 - ✓ Conditional Image Synthesis
 - ✓ Inpainting

Thank you!

- VAE ELBO (Evidence Lower Bound) : 주어진 샘플 x 의 확률 분포를 잘 표현해보자

$$\begin{aligned}
 \log p_{\theta}(x^{(i)}) &= \mathbb{E}_{z \sim q_{\phi}(z|x^{(i)})} \left[\log p_{\theta}(x^{(i)}) \right] \\
 &= \mathbb{E}_z \left[\log \frac{p_{\theta}(x^{(i)}|z) p_{\theta}(z)}{p_{\theta}(z|x^{(i)})} \right] \quad (\text{Bayes' Rule}) \\
 &= \mathbb{E}_z \left[\log \frac{p_{\theta}(x^{(i)}|z) p_{\theta}(z)}{p_{\theta}(z|x^{(i)})} \cdot \frac{q_{\phi}(z|x^{(i)})}{q_{\phi}(z|x^{(i)})} \right] \quad (\text{Multiply by constant}) \\
 &= \mathbb{E}_z \left[\log p_{\theta}(x^{(i)}|z) \div \frac{q_{\phi}(z|x^{(i)})}{p_{\theta}(z)} \times \frac{q_{\phi}(z|x^{(i)})}{p_{\theta}(z|x^{(i)})} \right] \quad (\text{Transformation}) \\
 &= \mathbb{E}_z \left[\log p_{\theta}(x^{(i)}|z) \right] - \mathbb{E}_z \left[\log \frac{q_{\phi}(z|x^{(i)})}{p_{\theta}(z)} \right] + \mathbb{E}_z \left[\log \frac{q_{\phi}(z|x^{(i)})}{p_{\theta}(z|x^{(i)})} \right] \quad (\text{Logarithms}) \\
 &= \mathbb{E}_z \left[\log p_{\theta}(x^{(i)}|z) \right] - D_{KL} \left(q_{\phi}(z|x^{(i)}) || p_{\theta}(z) \right) + D_{KL} \left(q_{\phi}(z|x^{(i)}) || p_{\theta}(z|x^{(i)}) \right)
 \end{aligned}$$