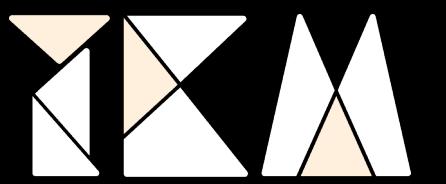


Paper Report

DDPM Denoising Diffusion Probabilistic Models, NeurIPS'20

@ 2022/09 Chia-Jen, Yeh



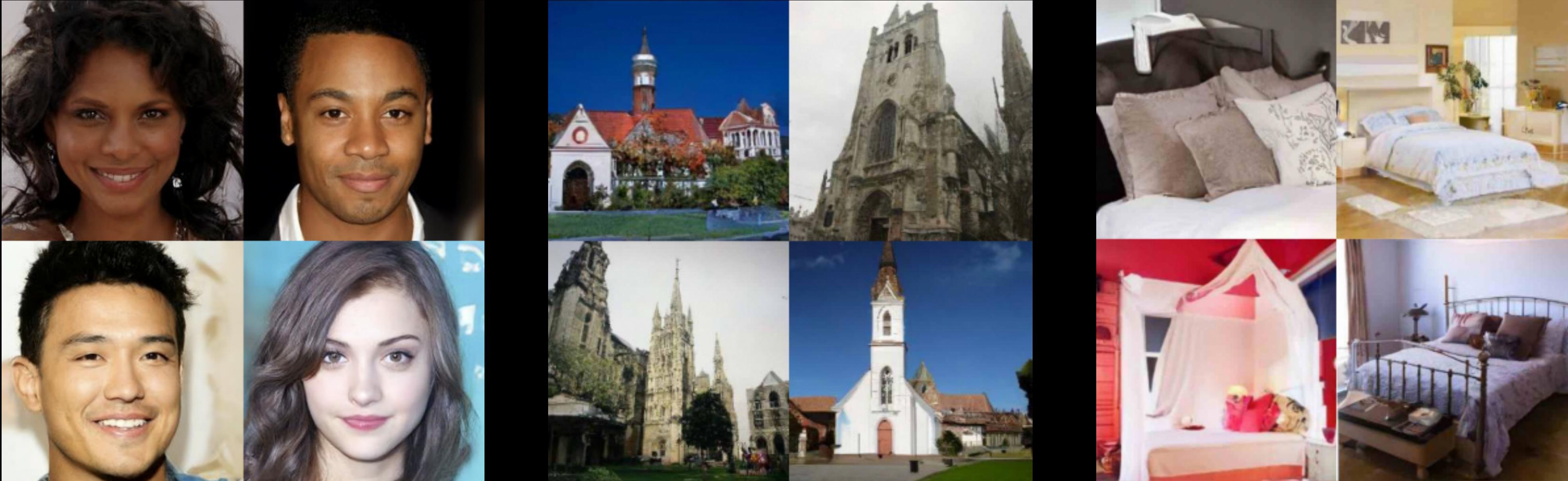
Report Structure

01 Introduction

02 Model Structure

03 Experiments

Introduction



- The diffusion model is a novel model that has beaten GAN in image generation in 2021.
- This paper presents high quality image synthesis results using diffusion probabilistic models.
- The diffusion model is based on Variational Inference.
- We use DDPM as short name of this paper propose model.

What is diffusion model?



x_0

What is diffusion model?



x_0

What is diffusion model?



x_0

What is diffusion model?



x_0



Gaussian Noise

x_T



What is diffusion model?



x_0



Markov chain

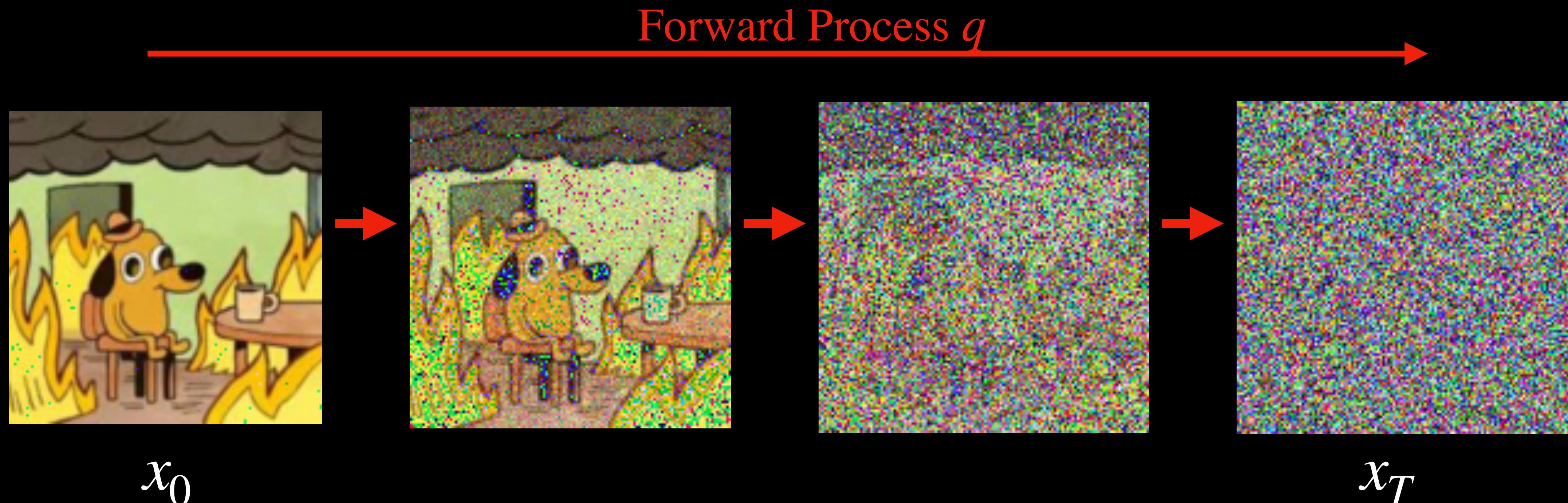


Gaussian Noise

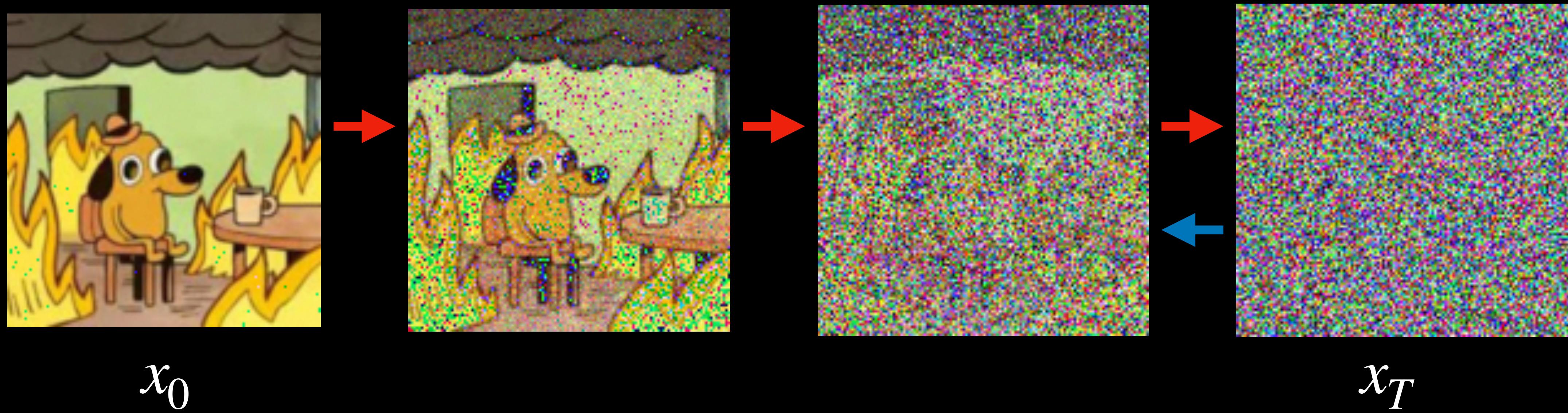


x_T

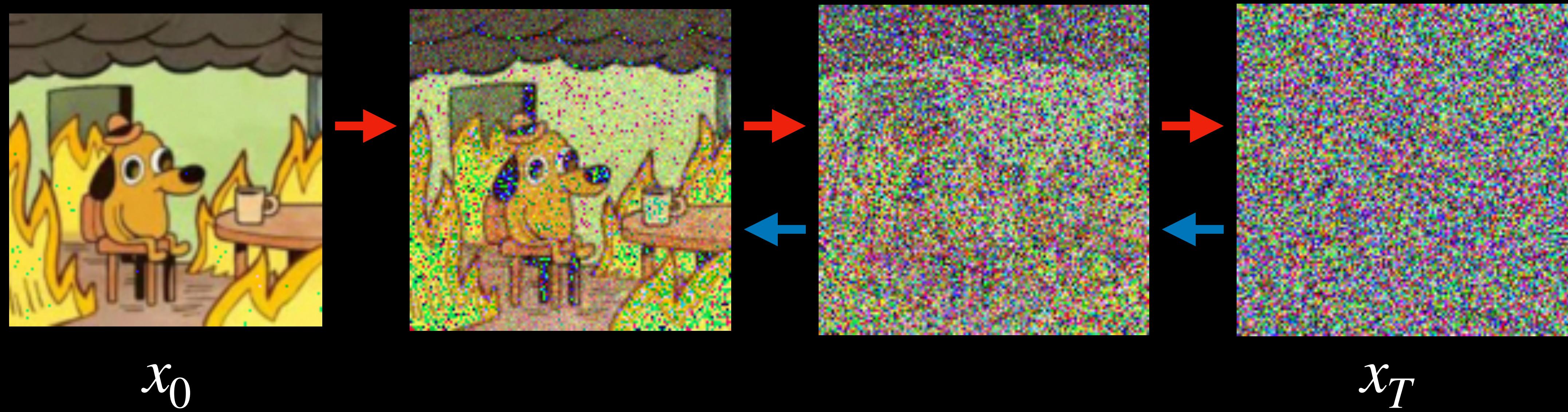
What is diffusion model?



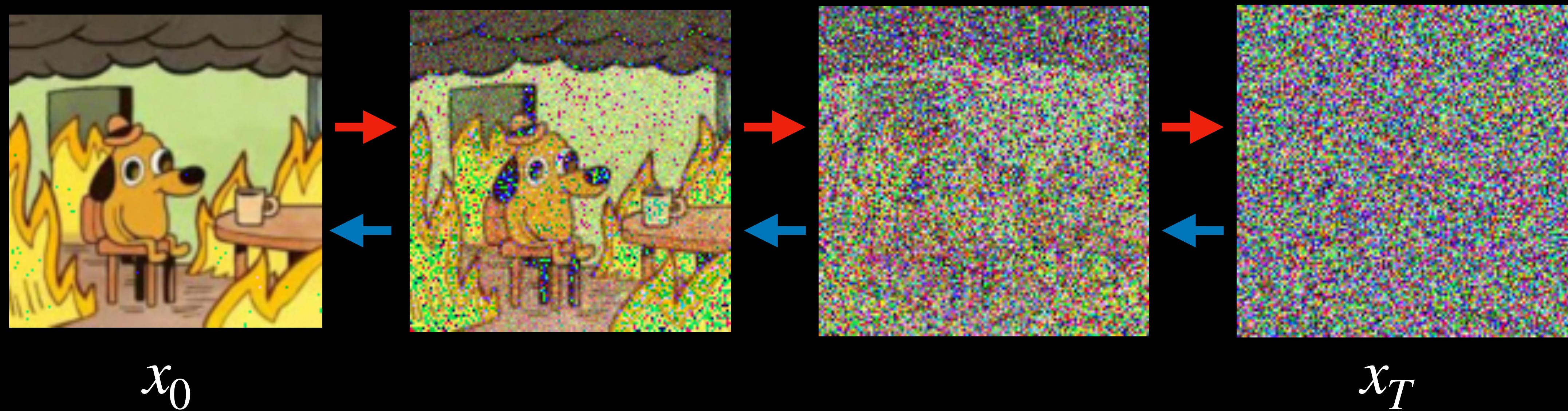
What is diffusion model?



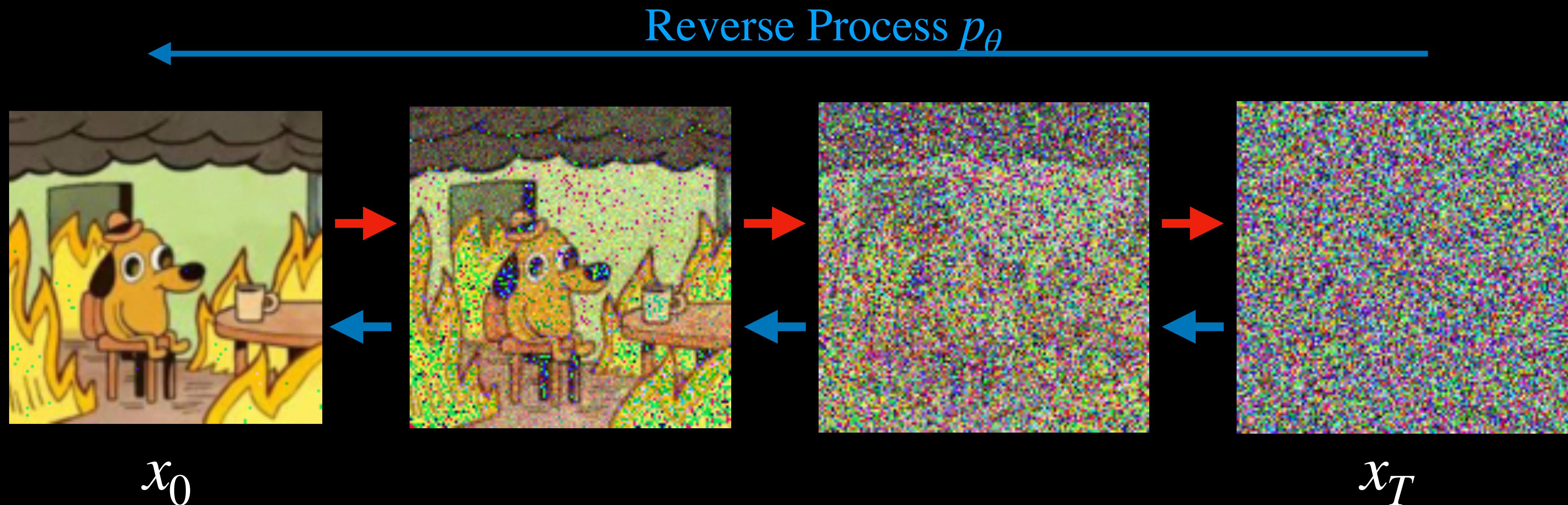
What is diffusion model?



What is diffusion model?



What is diffusion model?



What is diffusion model?



x_0

Timestep 1~T

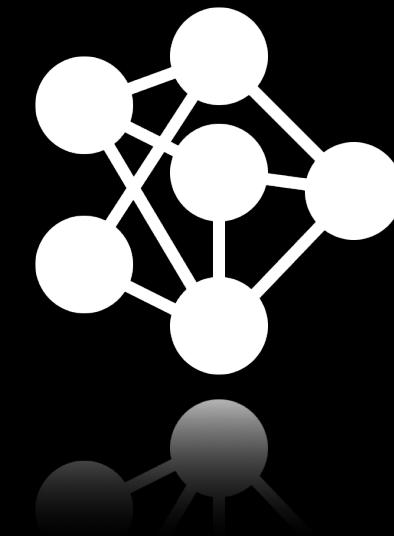


x_T

What is diffusion model?

Inference

Reverse Process



Sample Image

Diffusion Model

Initial Image

02

Model Structure



Define Question

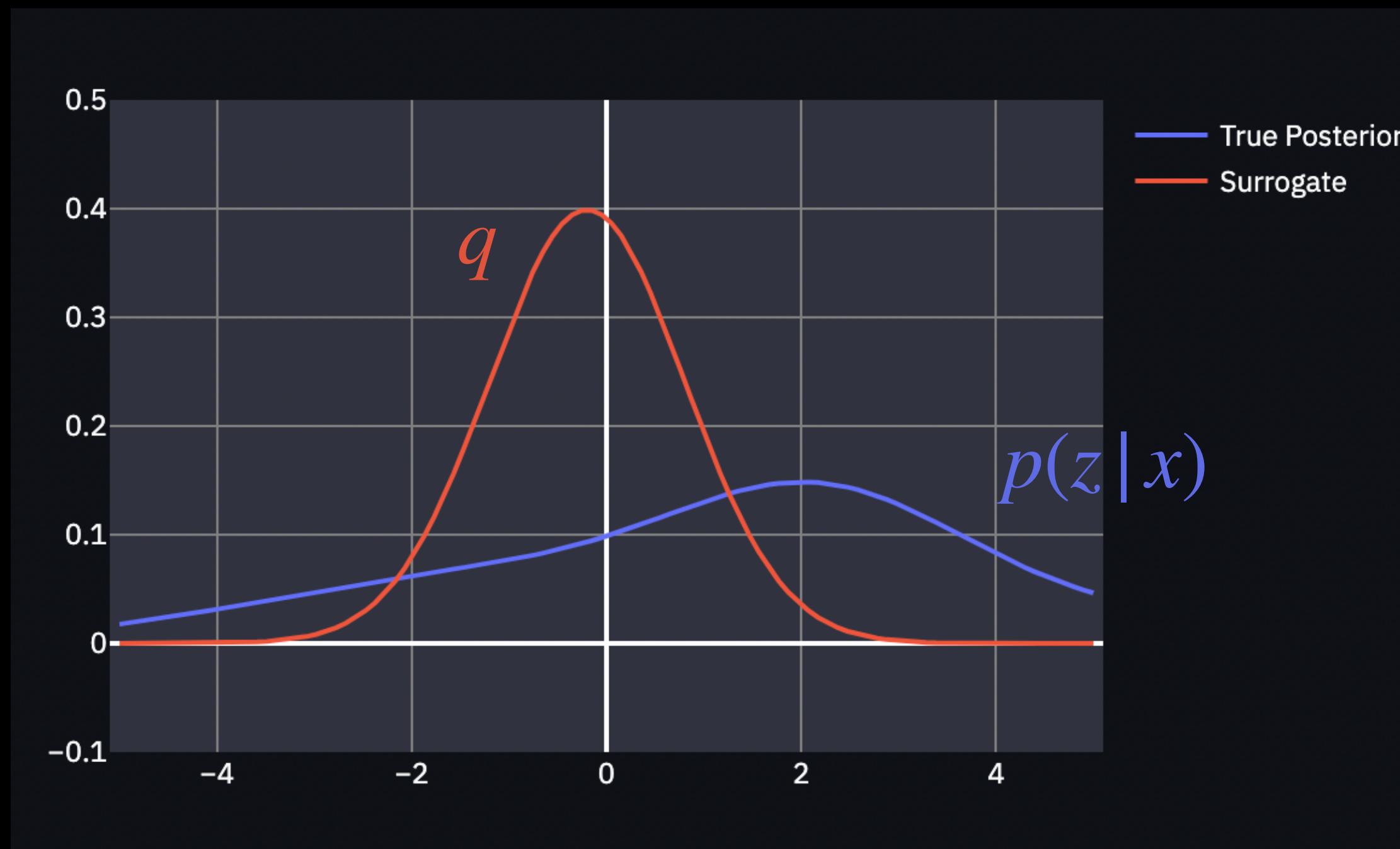
What is Variational Bayesian Inference?

What is Forward Process?

What is Reverse Process?

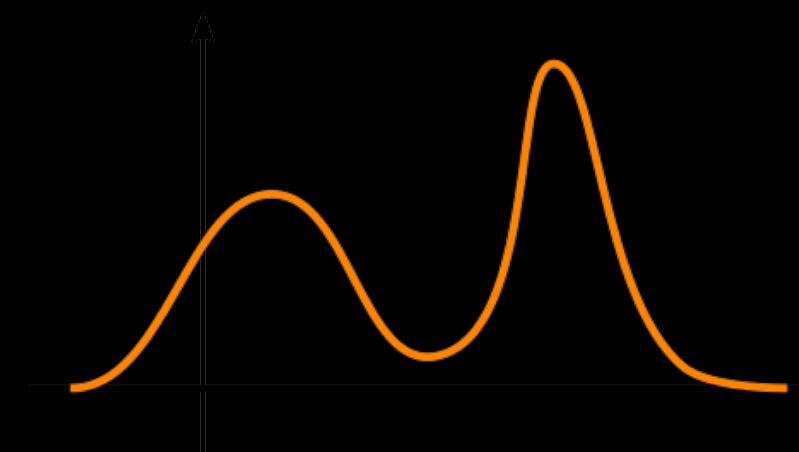
What is the Training Objective of the Diffusion Model?

What is Variational Bayesian Inference?

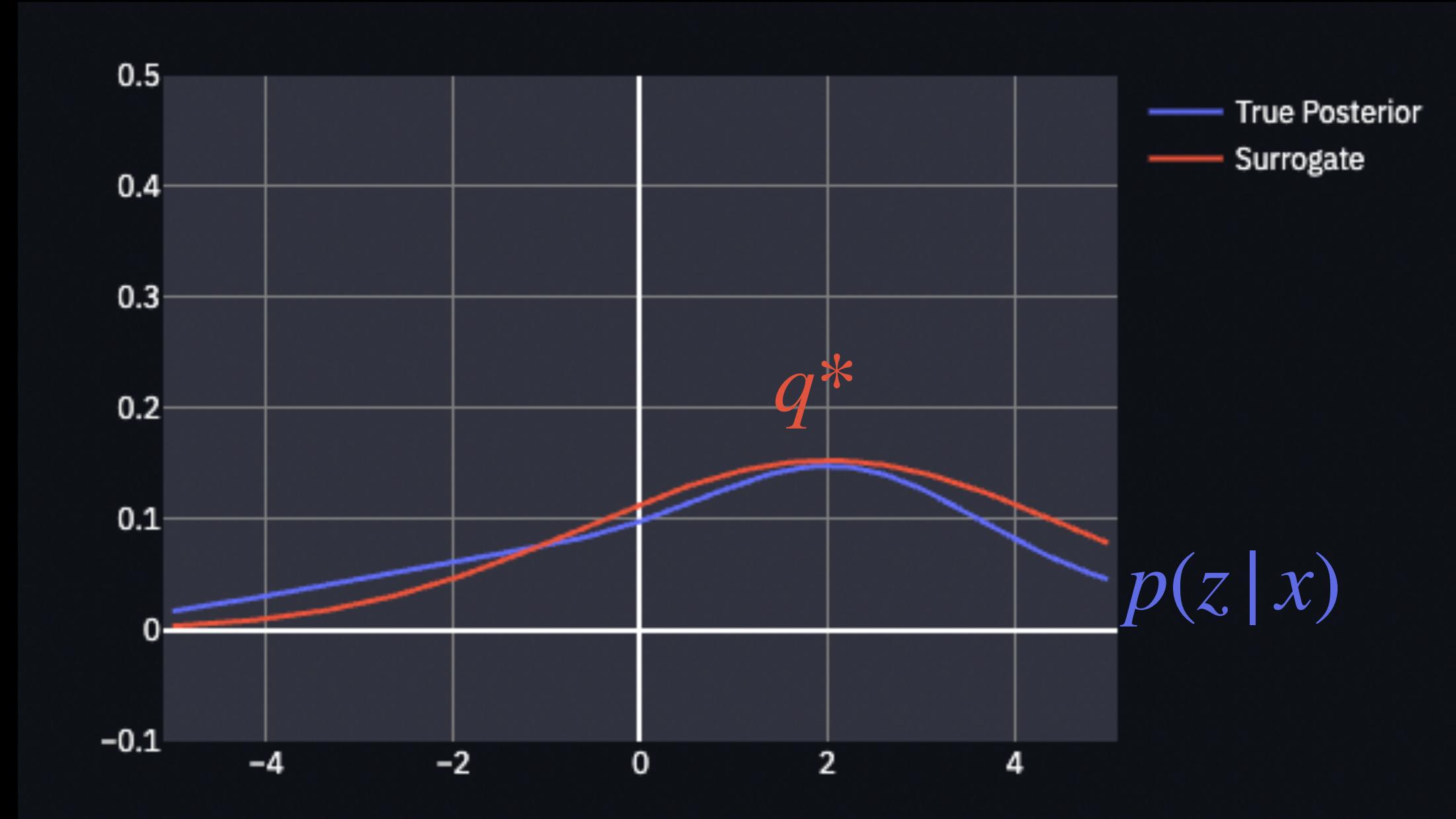


- Observed Variable x
- Latent Variable z
- Posterior Distribution $p(z|x)$

Posterior distribution $p(z|x)$ might be complex, so we use a Surrogate Normal Distribution q to fit $p(z|x)$



What is Variational Bayesian Inference?



- Observed Variable x
- Latent Variable z
- Posterior Distribution $p(z | x)$
- Surrogate Distribution q

We want q as similar to $p(z | x)$ as possible

$$q^* = \arg \min D_{KL}(q || p(z | x))$$

Variational Bayesian Inference in VAE

$$q(z|x) \approx p(z|x)$$

$$\begin{aligned}
D_{\text{KL}}(q_{\phi}(z|x) \| p_{\theta}(z|x)) &= \int q_{\phi}(z|x) \log \frac{q_{\phi}(z|x)}{p_{\theta}(z|x)} dz \\
&= \int q_{\phi}(z|x) \log \frac{q_{\phi}(z|x)p_{\theta}(x)}{p_{\theta}(z,x)} dz \\
&= \int q_{\phi}(z|x) \left(\log p_{\theta}(x) + \log \frac{q_{\phi}(z|x)}{p_{\theta}(z,x)} \right) dz \\
&= \log p_{\theta}(x) + \int q_{\phi}(z|x) \log \frac{q_{\phi}(z|x)}{p_{\theta}(z,x)} dz \\
&= \log p_{\theta}(x) + \int q_{\phi}(z|x) \log \frac{q_{\phi}(z|x)}{p_{\theta}(x|z)p_{\theta}(z)} dz \\
&= \log p_{\theta}(x) + \mathbb{E}_{z \sim q_{\phi}(z|x)} \left[\log \frac{q_{\phi}(z|x)}{p_{\theta}(z)} - \log p_{\theta}(x|z) \right] \\
&= \log p_{\theta}(x) + D_{\text{KL}}(q_{\phi}(z|x) \| p_{\theta}(z)) - \mathbb{E}_{z \sim q_{\phi}(z|x)} \log p_{\theta}(x|z)
\end{aligned}$$

Variational Bayesian Inference in VAE

$$q(z|x) \approx p(z|x)$$

$$\begin{aligned}
D_{\text{KL}}(q_{\phi}(z|x) \| p_{\theta}(z|x)) &= \int q_{\phi}(z|x) \log \frac{q_{\phi}(z|x)}{p_{\theta}(z|x)} dz \\
&\geq 0 \\
&= \int q_{\phi}(z|x) \log \frac{q_{\phi}(z|x)p_{\theta}(x)}{p_{\theta}(z,x)} dz \\
&= \int q_{\phi}(z|x) \left(\log p_{\theta}(x) + \log \frac{q_{\phi}(z|x)}{p_{\theta}(z,x)} \right) dz \\
&= \log p_{\theta}(x) + \int q_{\phi}(z|x) \log \frac{q_{\phi}(z|x)}{p_{\theta}(z,x)} dz \\
&= \log p_{\theta}(x) + \int q_{\phi}(z|x) \log \frac{q_{\phi}(z|x)}{p_{\theta}(x|z)p_{\theta}(z)} dz \\
&= \log p_{\theta}(x) + \mathbb{E}_{z \sim q_{\phi}(z|x)} \left[\log \frac{q_{\phi}(z|x)}{p_{\theta}(z)} - \log p_{\theta}(x|z) \right] \\
&= \log p_{\theta}(x) + D_{\text{KL}}(q_{\phi}(z|x) \| p_{\theta}(z)) - \mathbb{E}_{z \sim q_{\phi}(z|x)} \log p_{\theta}(x|z)
\end{aligned}$$

Variational Bayesian Inference in VAE

$$q(z|x) \approx p(z|x)$$

$$\begin{aligned} & \rightarrow \log p_\theta(x) + D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z)) - \mathbb{E}_{z \sim q_\phi(z|x)} \log p_\theta(x|z) \geq 0 \\ & \log p_\theta(x) \geq \mathbb{E}_{z \sim q_\phi(z|x)} \log p_\theta(x|z) - D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z)) \end{aligned}$$

Variational Bayesian Inference in VAE

$$q(z|x) \approx p(z|x)$$

$$\rightarrow \log p_\theta(x) + D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z)) - \mathbb{E}_{z \sim q_\phi(z|x)} \log p_\theta(x|z) \geq 0$$

$$\boxed{\log p_\theta(x) \geq \mathbb{E}_{z \sim q_\phi(z|x)} \log p_\theta(x|z) - D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z))}$$

Log-likelihood

Evidence/Variational Lower Bound, ELBO

Variational Bayesian Inference in VAE

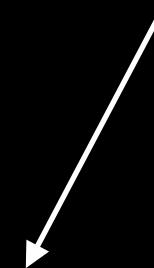
$$q(z|x) \approx p(z|x)$$

$$\log p_\theta(x) + D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z)) - \mathbb{E}_{z \sim q_\phi(z|x)} \log p_\theta(x|z) \geq 0$$

$$\log p_\theta(x) \geq \mathbb{E}_{z \sim q_\phi(z|x)} \log p_\theta(x|z) - D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z))$$

$$= \mathbb{E}_{q_\phi} \left[\log p_\theta(x|z) - D_{\text{KL}}(q_\phi(z|x) || p_\theta(z)) \right]$$

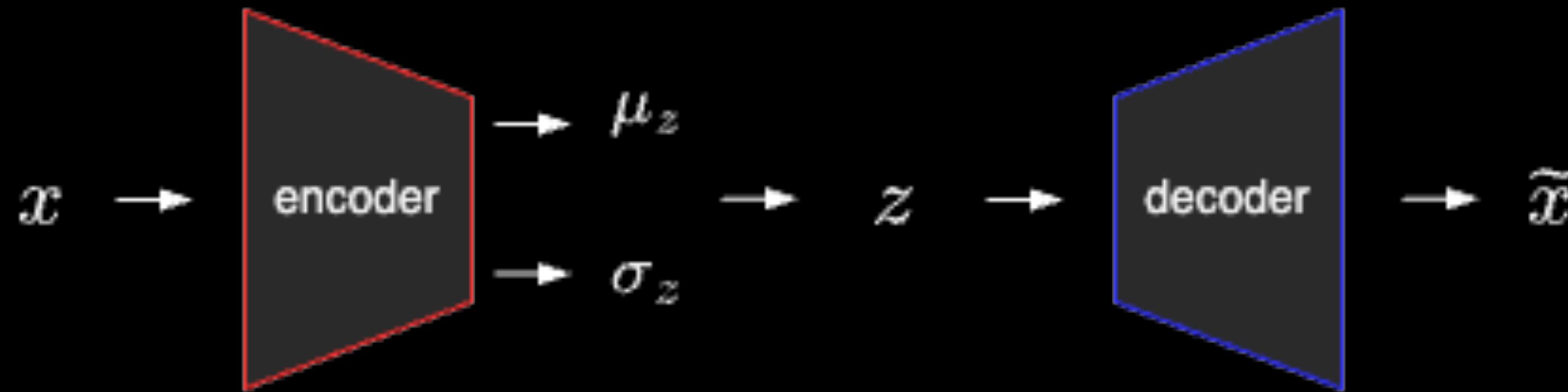
VAE's loss function



use Monte Carlo (sampling) method. In VAE, the sample num is 1.

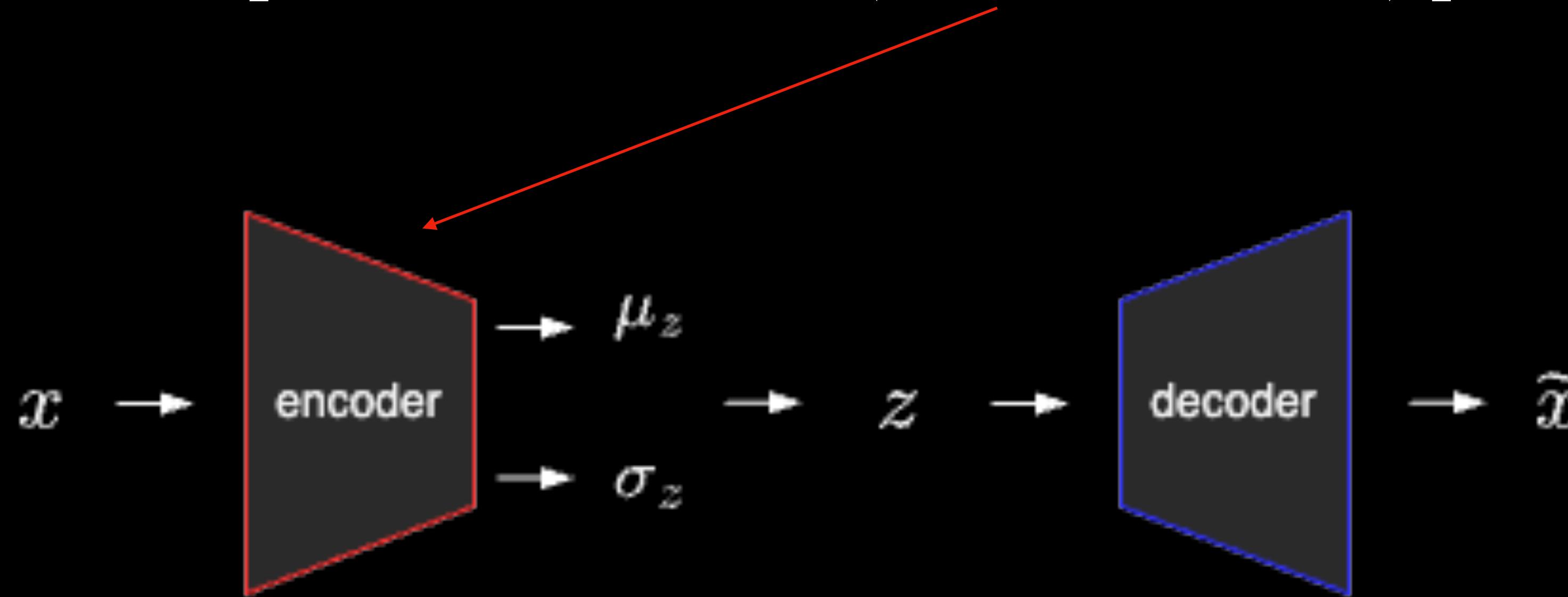
Variational Bayesian Inference in VAE

$$\mathbb{E}_{q_\phi} \left[\log p_\theta(x | z) - D_{KL} \left(q_\phi(z | x) || p_\theta(z) \right) \right]$$

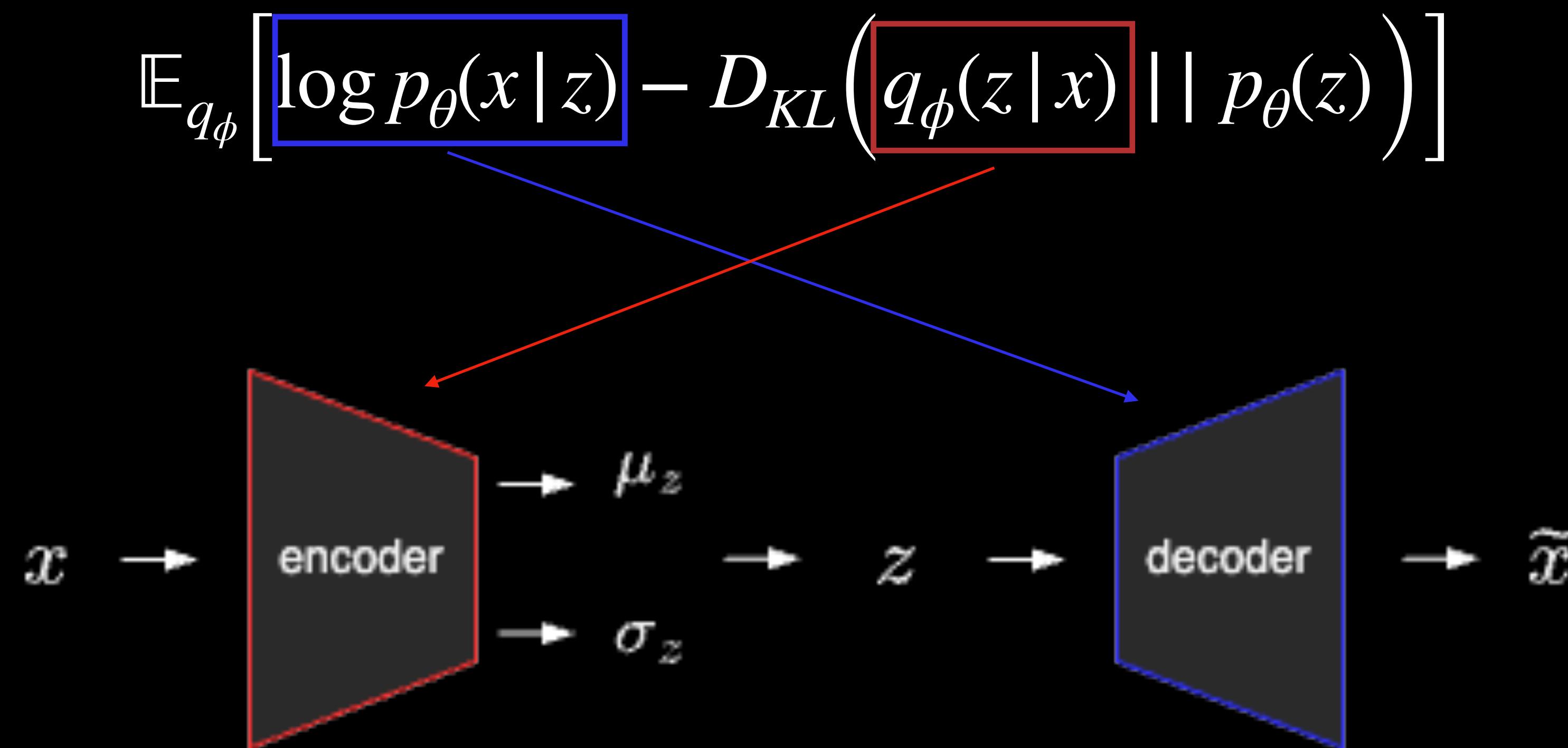


Variational Bayesian Inference in VAE

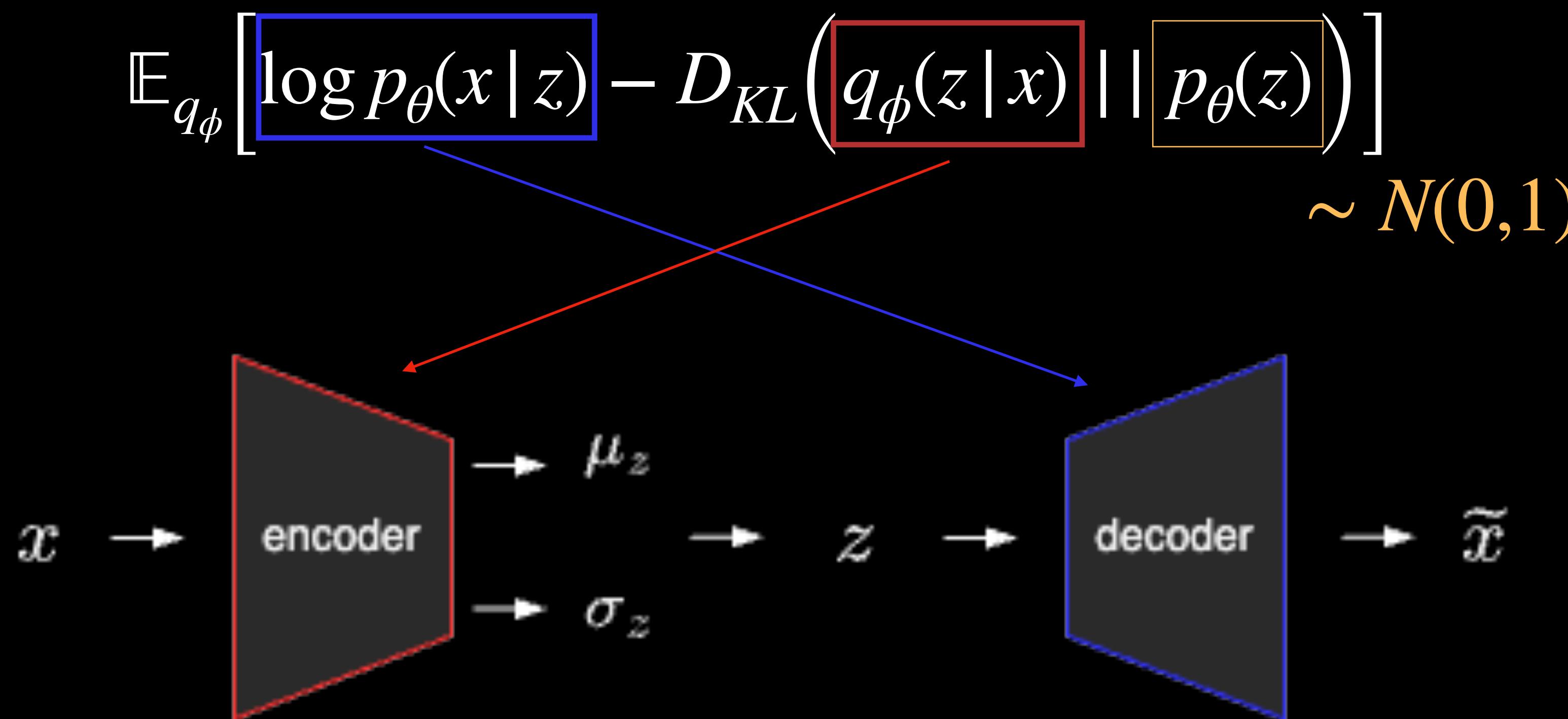
$$\mathbb{E}_{q_\phi} \left[\log p_\theta(x | z) - D_{KL} \left(q_\phi(z | x) || p_\theta(z) \right) \right]$$



Variational Bayesian Inference in VAE

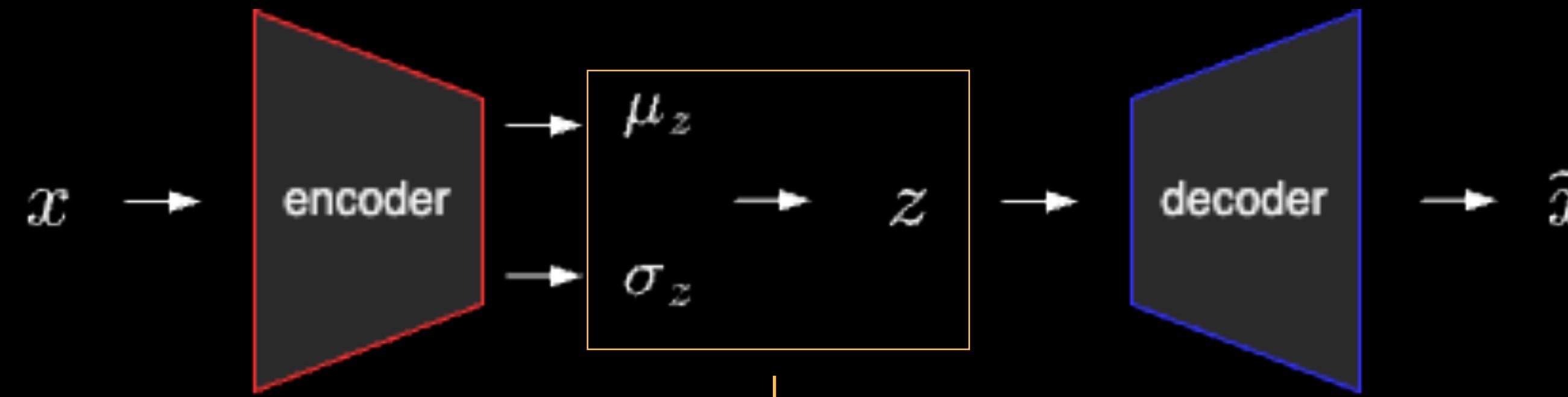


Variational Bayesian Inference in VAE

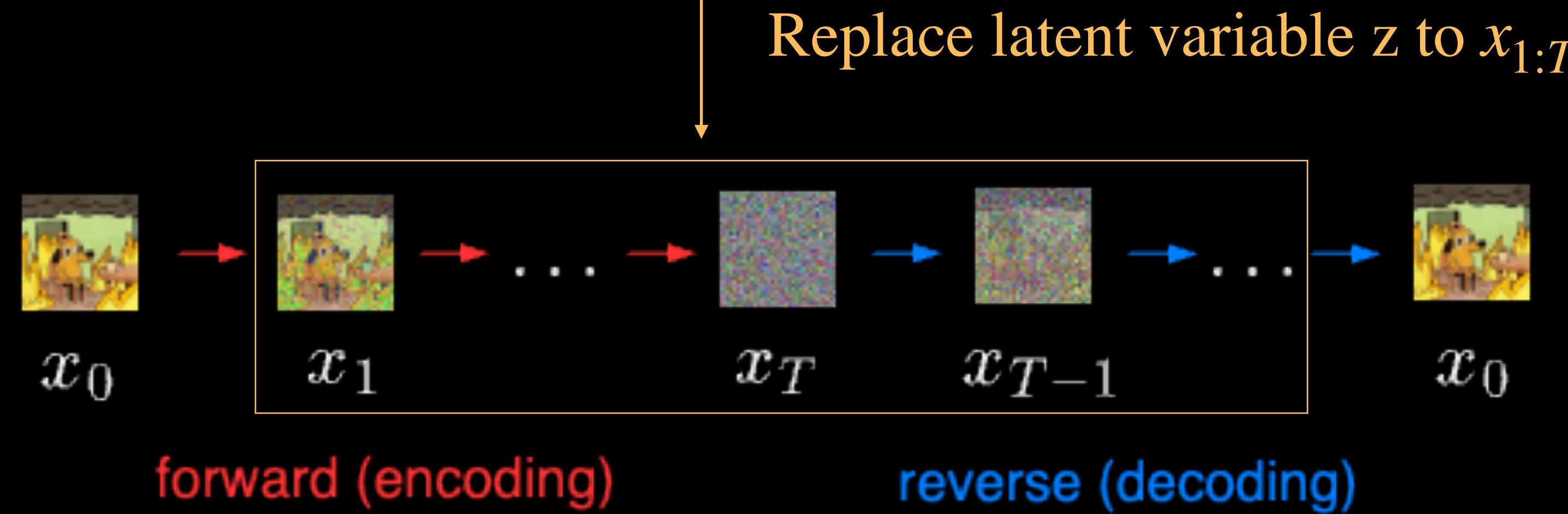


From VAE to Diffusion

VAE

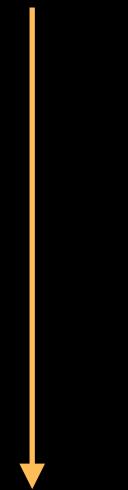


DDPM



From VAE to Diffusion

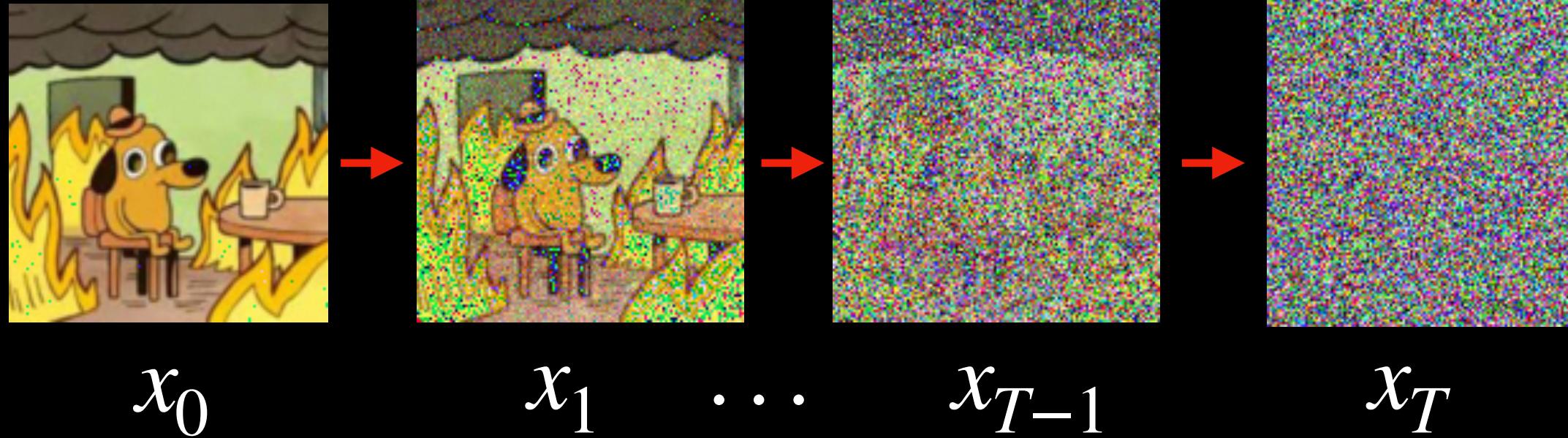
VAE $\mathbb{E}_{q_\phi} \left[\log p_\theta(x | z) - D_{KL} \left(q_\phi(z | x) || p_\theta(z) \right) \right]$



Replace latent variable
from z to $x_{1:T}$

DDPM $\mathbb{E}_q \left[\log p_\theta(\underline{x_0} | \underline{x_{1:T}}) - D_{KL} \left(q(\underline{x_{1:T}} | \underline{x_0}) || p_\theta(\underline{x_{1:T}}) \right) \right]$

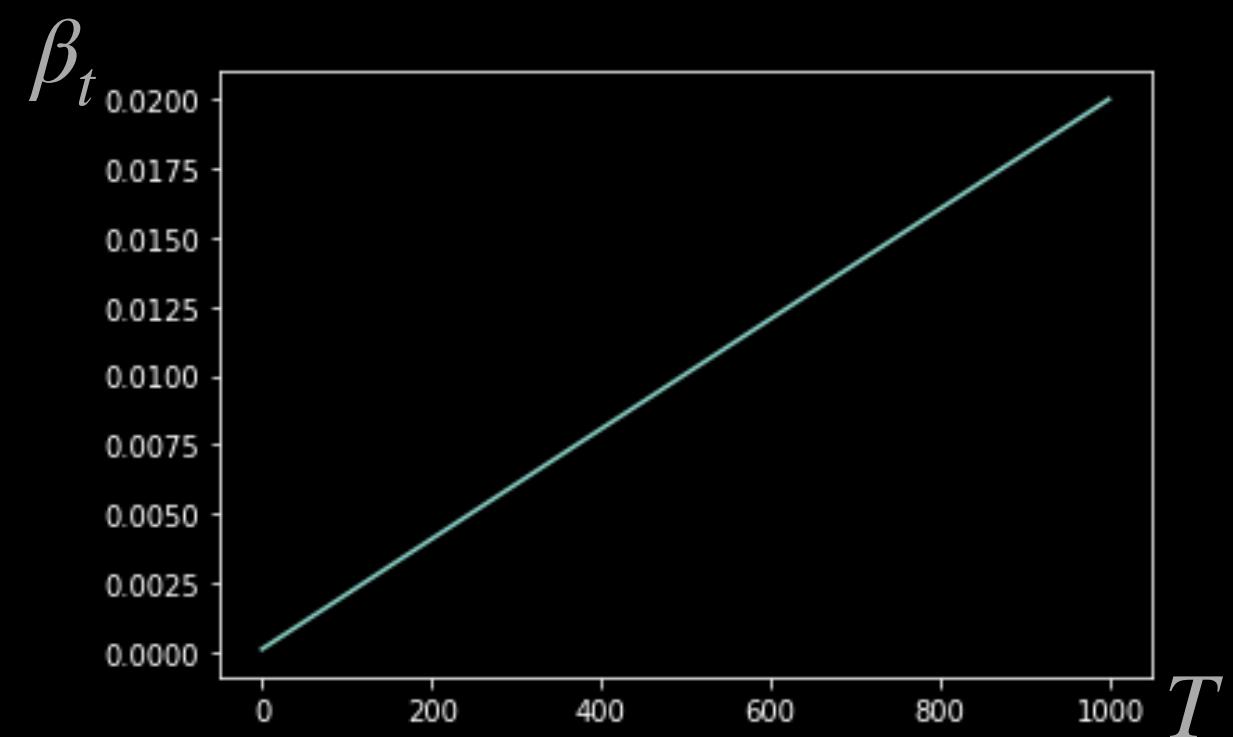
What is forward process?



$$q(x_{1:T} | x_0) := \prod_{t=1}^T q(x_t | x_{t-1}),$$
$$q(x_t | x_{t-1}) := N(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I)$$

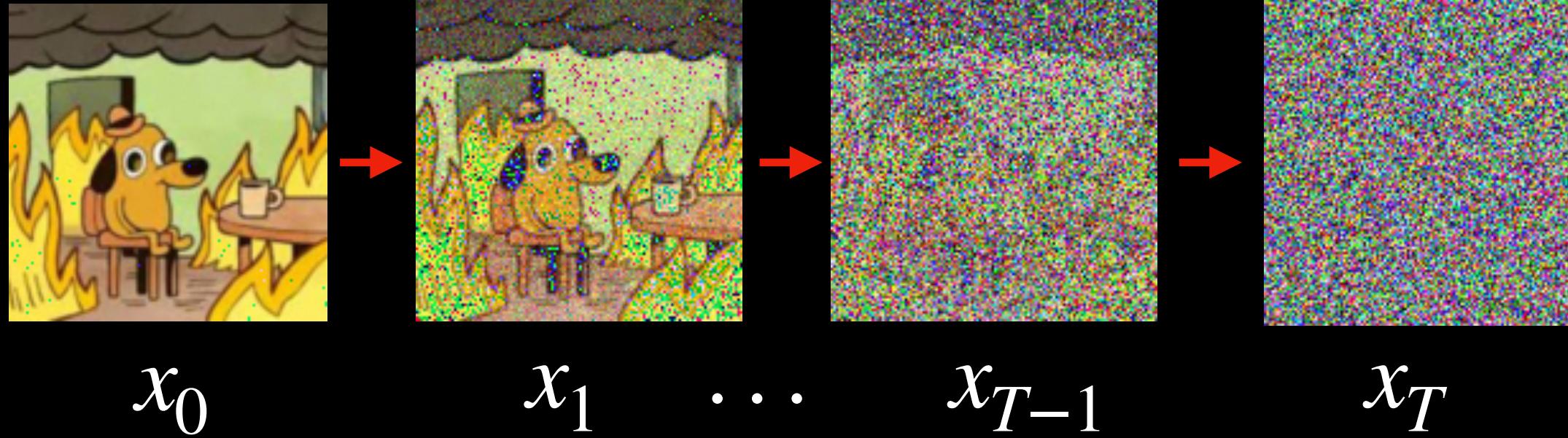
- Define a **Markov chain** of diffusion steps to slowly add random noise to the data according to a **variance schedule** β_1, \dots, β_T

$$\beta_1 = 0.0001$$
$$\beta_T = 0.02$$



When T is bigger enough ($T \rightarrow \infty$), the final x_T would become a random gaussian distribution

What is forward process?



$$q(x_{1:T} | x_0) := \prod_{t=1}^T q(x_t | x_{t-1}),$$

$$q(x_t | x_{t-1}) := N(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I)$$

- Define a **Markov chain** of diffusion steps to slowly add random noise to the data according to a **variance schedule** β_1, \dots, β_T
- The process is **fixed** throughout the process, has nothing to do with model and training

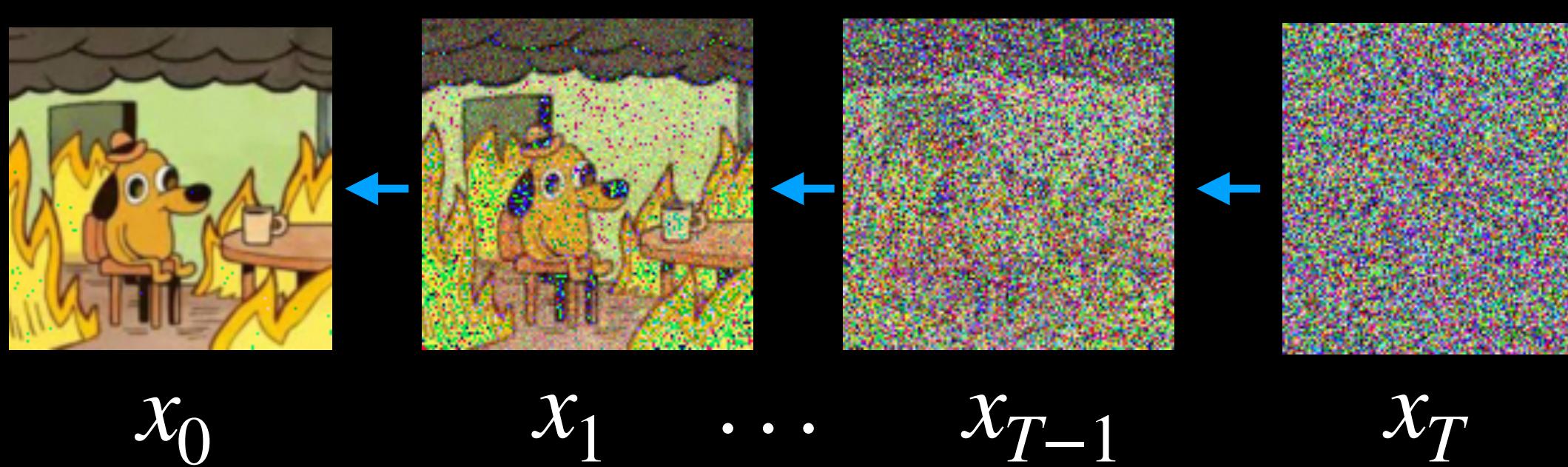
VAE's ELBO

$$\mathbb{E}_{q_\phi} \left[\log p_\theta(x | z) - D_{KL} \left(q_\phi(z | x) || p_\theta(z) \right) \right]$$

Diffusion Model's ELBO

$$\mathbb{E}_q \left[\log p_\theta(x_0 | x_{1:T}) - D_{KL} \left(q(x_{1:T} | x_0) || p_\theta(x_{1:T}) \right) \right]$$

What is reverse process?

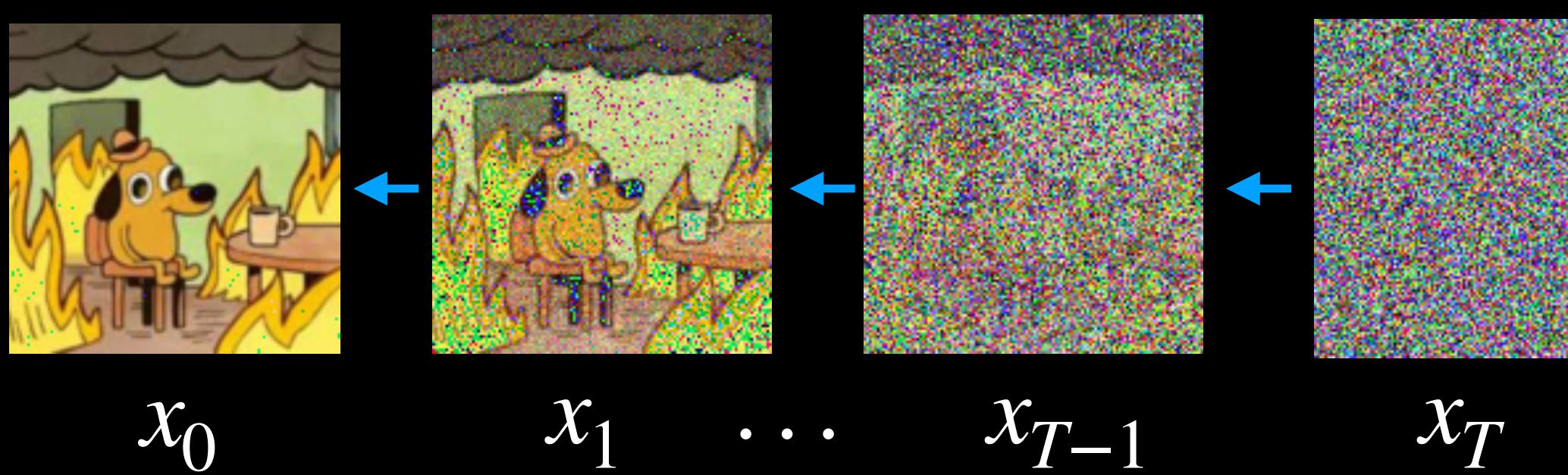


$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^T p_\theta(x_{t-1} | x_t),$$
$$p_\theta(x_{t-1} | x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

- The reverse process is a model learn to reverse the diffusion process to construct desired data from the noise.
- Unlike VAE just use Linear Layer, DDPM use U-net as predict model

* U-Net is a convolutional neural network that was developed for biomedical image segmentation

What is reverse process?



$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^T p_\theta(x_{t-1} | x_t),$$

$$p_\theta(x_{t-1} | x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

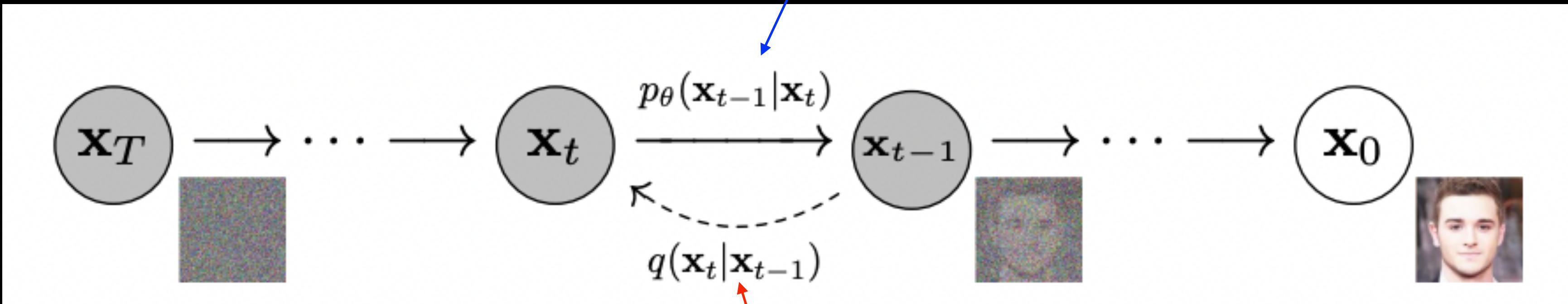
Can be fixed as forwarding's variance β_t ,
or been predicted by the model

- The reverse process is a model learn to reverse the diffusion process to construct desired data from the noise.
- Unlike VAE just use Linear Layer, DDPM use U-net as predict model

* U-Net is a convolutional neural network that was developed for biomedical image segmentation

The directed graphical model considered in this work.

$$p_{\theta}(x_{0:T}) := p(x_T) \prod_{t=1}^T p_{\theta}(x_{t-1} | x_t), \quad p_{\theta}(x_{t-1} | x_t) := N(x_{t-1}; \mu_{\theta}(x_t, t), \Sigma_{\theta}(x_t, t))$$



$$q(x_{1:T} | x_0) := \prod_{t=1}^T q(x_t | x_{t-1}), \quad q(x_t | x_{t-1}) := N(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I)$$

What is the Training Objective of the Diffusion Model?

ELBO

$$\mathbb{E}_q \left[\log p_\theta(x_0 | x_{1:T}) - D_{\text{KL}}(q(x_{1:T} | x_0) || p_\theta(x_{1:T})) \right]$$



Learned Σ_θ :

Fixed Σ as σ_t , $\sigma_t^2 = \beta_t$ or $\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$:

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2 \|\Sigma_\theta\|_2^2} \|\tilde{\mu}(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right]$$

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\mu}(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right]$$

Predict μ_θ



Predict ϵ_θ

Noise

$$\mathbb{E}_{x_0, \epsilon} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1 - \bar{\alpha}_t)} \|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right] \xrightarrow{\text{Simplify}} \mathbb{E}_{x_0, \epsilon, t} [\|\epsilon - \epsilon_\theta(x_t, t)\|^2]$$

(MSE of noise)

What is the Training Objective of the Diffusion Model?

ELBO

$$\mathbb{E}_q \left[\log p_\theta(x_0 | x_{1:T}) - D_{\text{KL}}(q(x_{1:T} | x_0) || p_\theta(x_{1:T})) \right]$$



Learned Σ_θ :

Fixed Σ as σ_t , $\sigma_t^2 = \beta_t$ or $\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$:

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2 \|\Sigma_\theta\|_2^2} \|\tilde{\mu}(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right]$$

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\mu}(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right]$$

Predict μ_θ



Predict ϵ_θ

Noise

$$\mathbb{E}_{x_0, \epsilon} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1 - \bar{\alpha}_t)} \|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right]$$

Simplify



$$\mathbb{E}_{x_0, \epsilon, t} [\|\epsilon - \epsilon_\theta(x_t, t)\|^2] \quad (\text{MSE of noise})$$

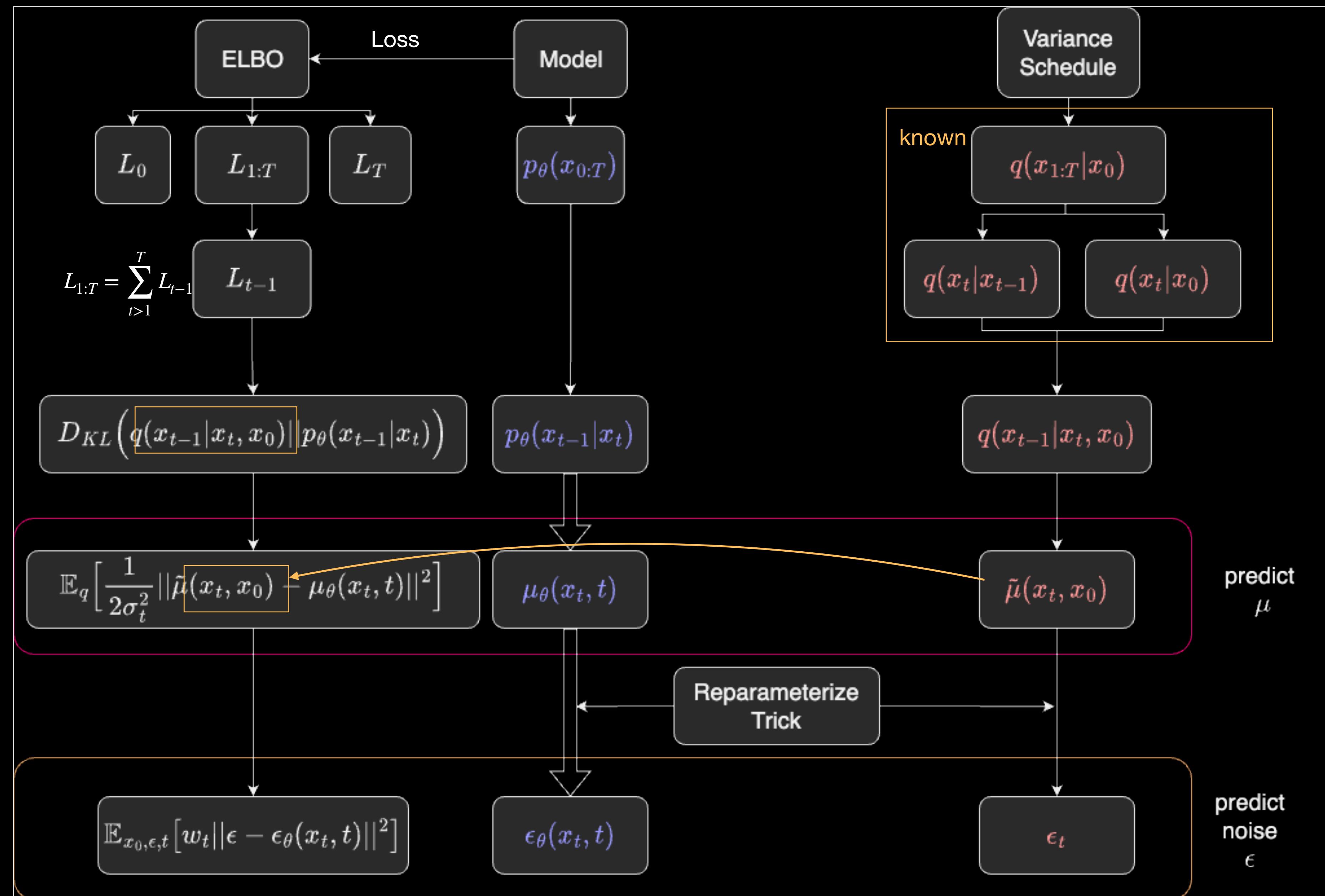
Derivation flow chart

1. Split ELBO loss to different time step, and ignore L_0 and L_T , we can focus on the loss between each step.

2. Use Bayes' Rule and Probability Density Function to derive $q(x_{t-1} | x_t)$ when conditioned on x_0 .

3. We can represent the distribution in μ and σ format when compute the KLD between two normal distribution

4. Use Reparameterize Trick
Convert $\tilde{\mu}_t$ to ϵ_t form



03

Experiments



Measure Method

Inception Score, IS

Frechet Inception Distance, FID

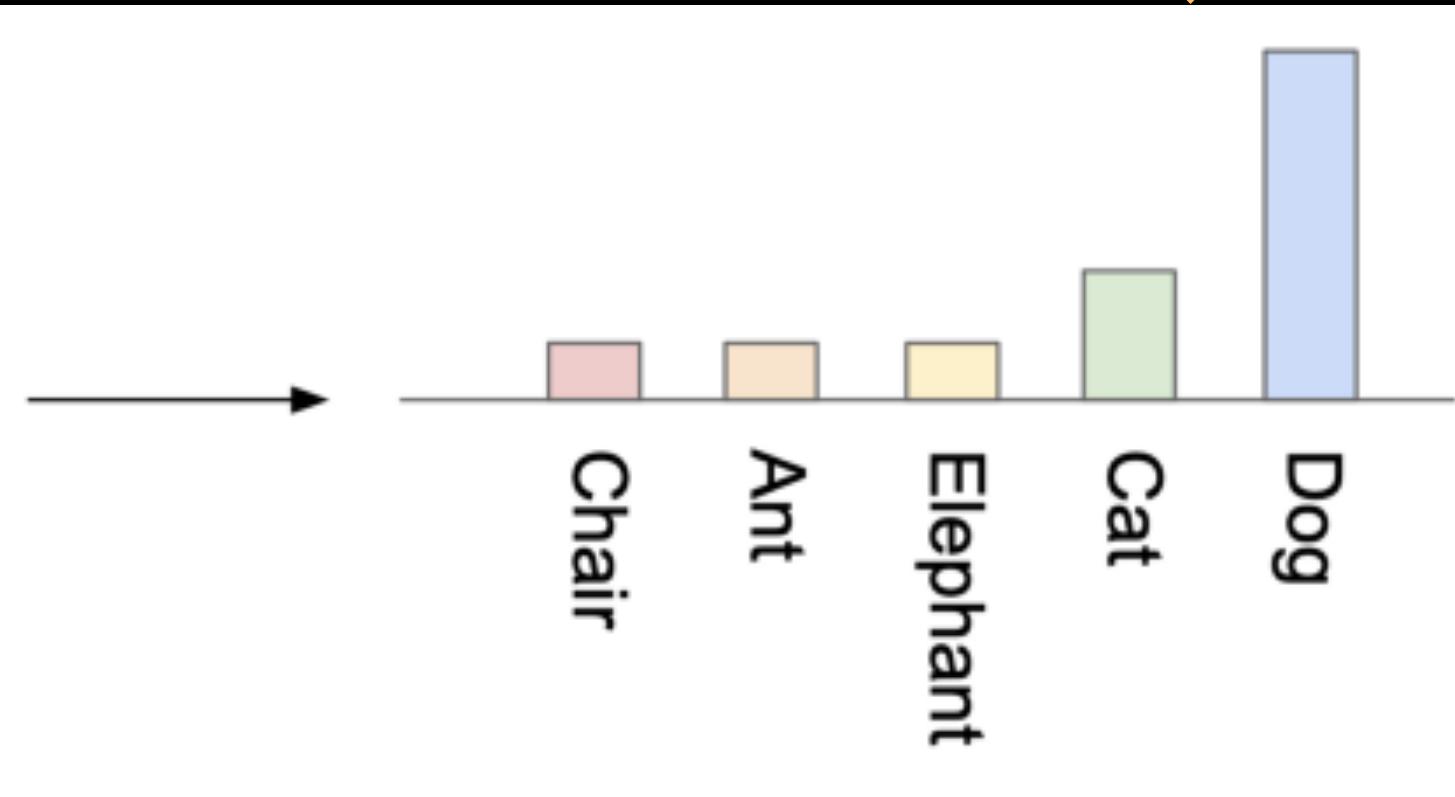
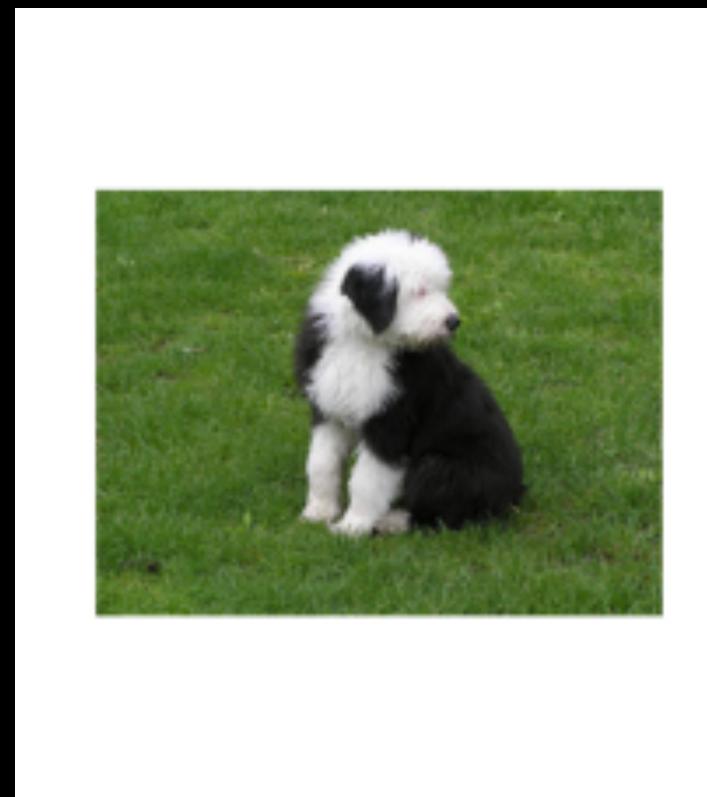
Both:

- Measure of how realistic a generated image is.
- Use Inception V3 (Google's model trained on ImageNet Dataset)

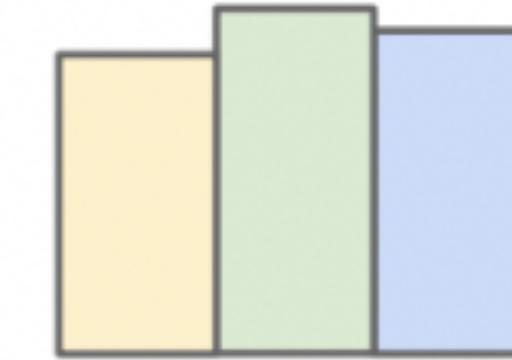
Inception Score

Inception v3 model p
Generated image x

$$\text{IS} = \exp\left(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|x) || p(y))\right)$$



Distinct

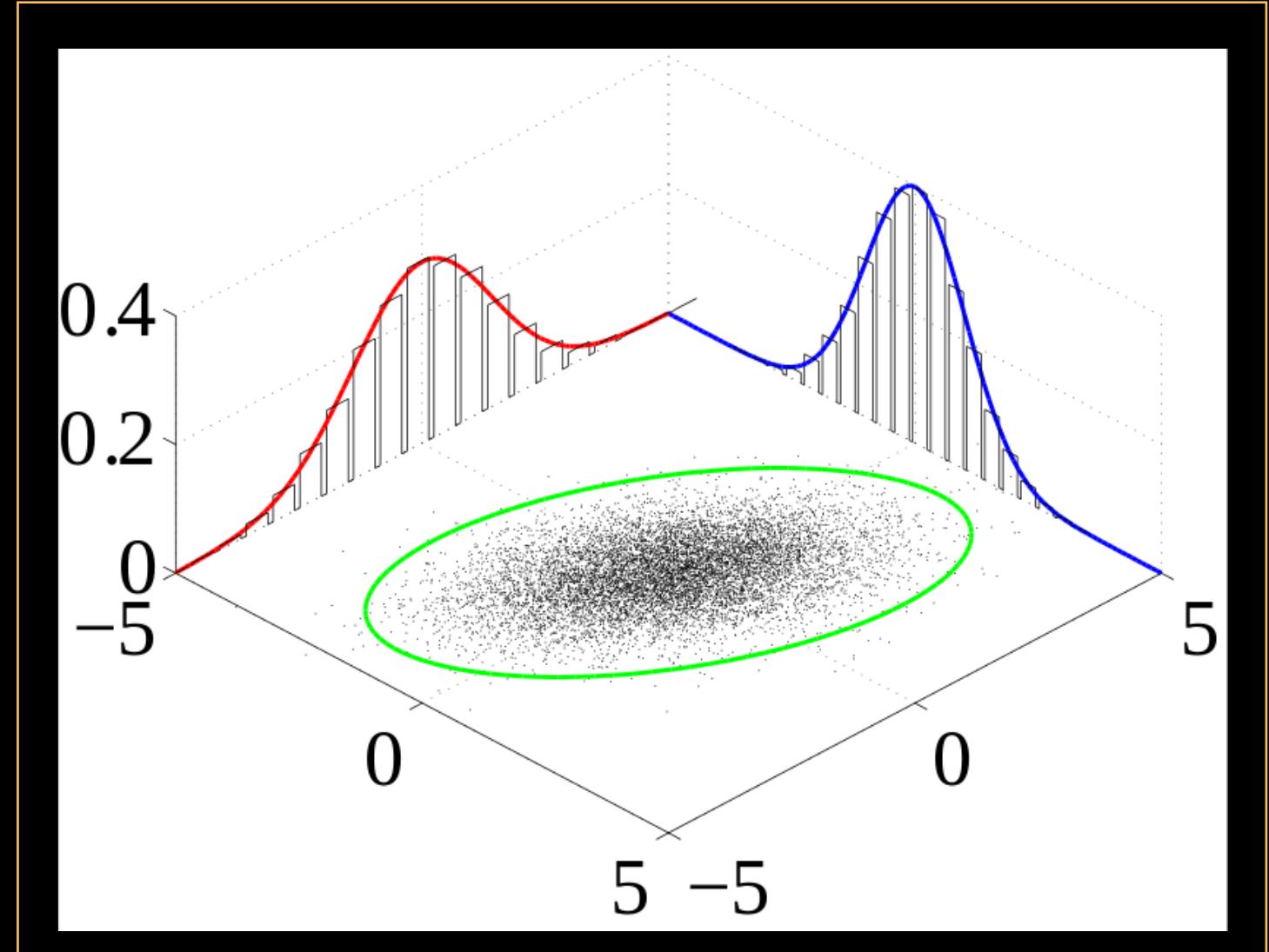
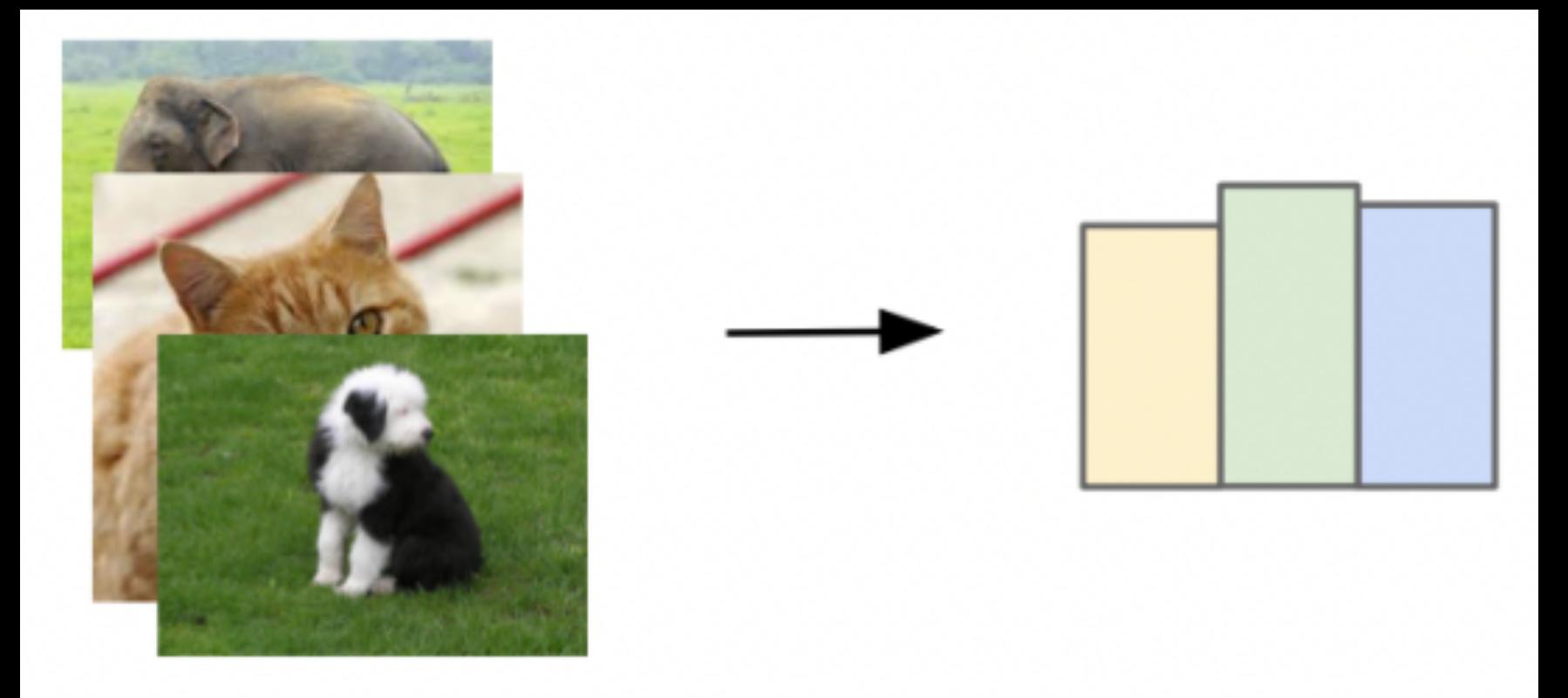
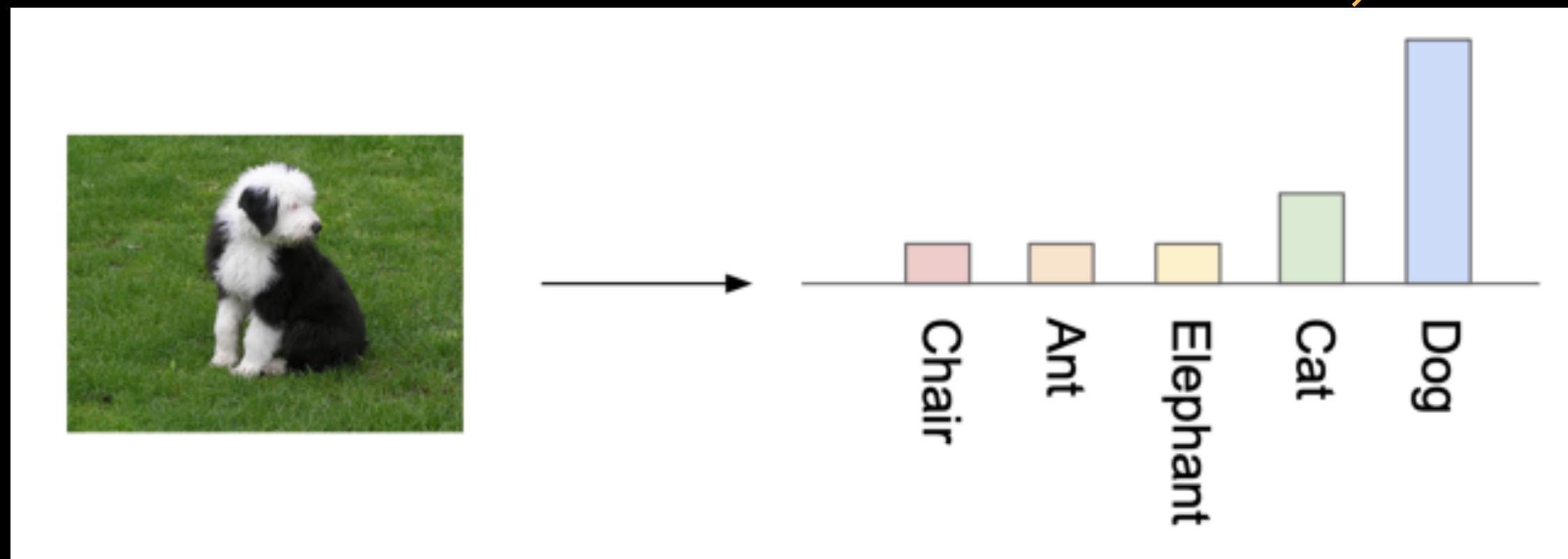


Diversity (look all input x)

The bigger the better

Inception Score

$$\text{IS} = \exp\left(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|x) || p(y))\right)$$



Marginal distribution when seen all x

The bigger the better

Frechet Inception Distance

$$\text{FID} = \| \mu_x - \mu_y \|^2 - \text{Tr}(\Sigma_X + \Sigma_Y - 2\Sigma_X \Sigma_Y)$$

Real Images Embeddings X

Fake Images Embeddings Y

(Assumed X, Y to be two multivariate normal distributions)

Trace of the matrix Tr

The smaller the better

Ablation study

Objective	IS	FID
$\tilde{\mu}$ prediction (baseline)		
L , learned diagonal Σ	7.28 ± 0.10	23.69
L , fixed isotropic Σ	8.06 ± 0.09	13.22
$\ \tilde{\mu} - \tilde{\mu}_\theta\ ^2$	-	-
ϵ prediction (ours)		
L , learned diagonal Σ	-	-
L , fixed isotropic Σ	7.67 ± 0.13	13.51
$\ \tilde{\epsilon} - \epsilon_\theta\ ^2$ (L_{simple})	9.46 ± 0.11	3.17

- $\tilde{\mu}$ only works well on the true ELBO
- ϵ prediction approximately as well as prediction $\tilde{\mu}$ when Σ is fixed
- ϵ prediction without w_t get the best results

Blank entries were unstable to train and generated poor samples

What is the Training Objective of the Diffusion Model?

ELBO

$$\mathbb{E}_q \left[\log p_\theta(x_0 | x_{1:T}) - D_{\text{KL}} \left(q(x_{1:T} | x_0) || p_\theta(x_{1:T}) \right) \right]$$

Learned Σ_θ :

Fixed Σ as σ_t , $\sigma_t^2 = \beta_t$ or $\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$:

Predict μ_θ

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\mu}(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right]$$

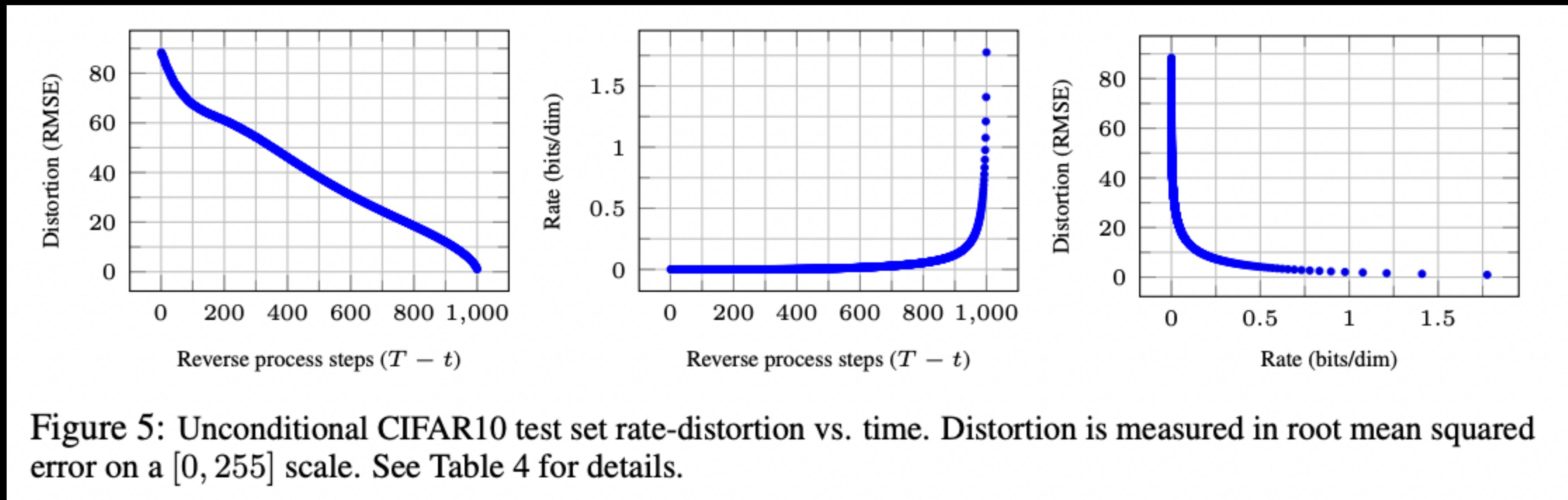
$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\mu}(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right]$$

Predict ϵ_θ

$$\mathbb{E}_{x_0, \epsilon} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1 - \bar{\alpha}_t)} \|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right] \xrightarrow{\text{Simplify}} \mathbb{E}_{x_0, \epsilon, t} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right]$$

(MSE of noise)

The influence of Timesteps



Ablation study

Objective	IS	FID
$\tilde{\mu}$ prediction (baseline)		
L , learned diagonal Σ	7.28 ± 0.10	23.69
L , fixed isotropic Σ	8.06 ± 0.09	13.22
$\ \tilde{\mu} - \tilde{\mu}_\theta\ ^2$	-	-
ϵ prediction (ours)		
L , learned diagonal Σ	-	-
L , fixed isotropic Σ	7.67 ± 0.13	13.51
$\ \tilde{\epsilon} - \epsilon_\theta\ ^2$ (L_{simple})	9.46 ± 0.11	3.17

- $\tilde{\mu}$ only works well on the true ELBO
- ϵ prediction approximately as well as prediction $\tilde{\mu}$ when Σ is fixed
- ϵ prediction without w_t get the best results

Blank entries were unstable to train and generated poor samples

Ablation study

Objective	IS	FID
$\tilde{\mu}$ prediction (baseline)		
L , learned diagonal Σ	7.28 ± 0.10	23.69
L , fixed isotropic Σ	8.06 ± 0.09	13.22
$\ \tilde{\mu} - \tilde{\mu}_\theta\ ^2$	-	-
ϵ prediction (ours)		
L , learned diagonal Σ	-	-
L , fixed isotropic Σ	7.67 ± 0.13	13.51
$\ \tilde{\epsilon} - \epsilon_\theta\ ^2$ (L_{simple})	9.46 ± 0.11	3.17

- $\tilde{\mu}$ only works well on the true ELBO
- ϵ prediction approximately as well as prediction $\tilde{\mu}$ when Σ is fixed
- ϵ prediction without w_t get the best results

Blank entries were unstable to train and generated poor samples

Ablation study

Objective	IS	FID
$\tilde{\mu}$ prediction (baseline)		
L , learned diagonal Σ	7.28 ± 0.10	23.69
L , fixed isotropic Σ	8.06 ± 0.09	13.22
$\ \tilde{\mu} - \tilde{\mu}_\theta\ ^2$	-	-
ϵ prediction (ours)		
L , learned diagonal Σ	-	-
L , fixed isotropic Σ	7.67 ± 0.13	13.51
$\ \tilde{\epsilon} - \epsilon_\theta\ ^2$ (L_{simple})	9.46 ± 0.11	3.17

- $\tilde{\mu}$ only works well on the true ELBO
- ϵ prediction approximately as well as prediction $\tilde{\mu}$ when Σ is fixed
- ϵ prediction without w_t get the best results

Blank entries were unstable to train and generated poor samples

Compare with SOTAs

Table 1: CIFAR10 results. NLL measured in bits/dim.

Model	IS	FID	NLL Test (Train)
Conditional			
EBM [11]	8.30	37.9	
JEM [17]	8.76	38.4	
BigGAN [3]	9.22	14.73	
StyleGAN2 + ADA (v1) [29]	10.06	2.67	
Unconditional			
Diffusion (original) [53]			≤ 5.40
Gated PixelCNN [59]	4.60	65.93	3.03 (2.90)
Sparse Transformer [7]			2.80
PixelIQN [43]	5.29	49.46	
EBM [11]	6.78	38.2	
NCSNv2 [56]			31.75
NCSN [55]	8.87 ± 0.12	25.32	
SNGAN [39]	8.22 ± 0.05	21.7	
SNGAN-DDLS [4]	9.09 ± 0.10	15.42	
StyleGAN2 + ADA (v1) [29]	9.74 ± 0.05	3.26	
Ours (L , fixed isotropic Σ)	7.67 ± 0.13	13.51	≤ 3.70 (3.69)
Ours (L_{simple})	9.46 ± 0.11	3.17	≤ 3.75 (3.72)

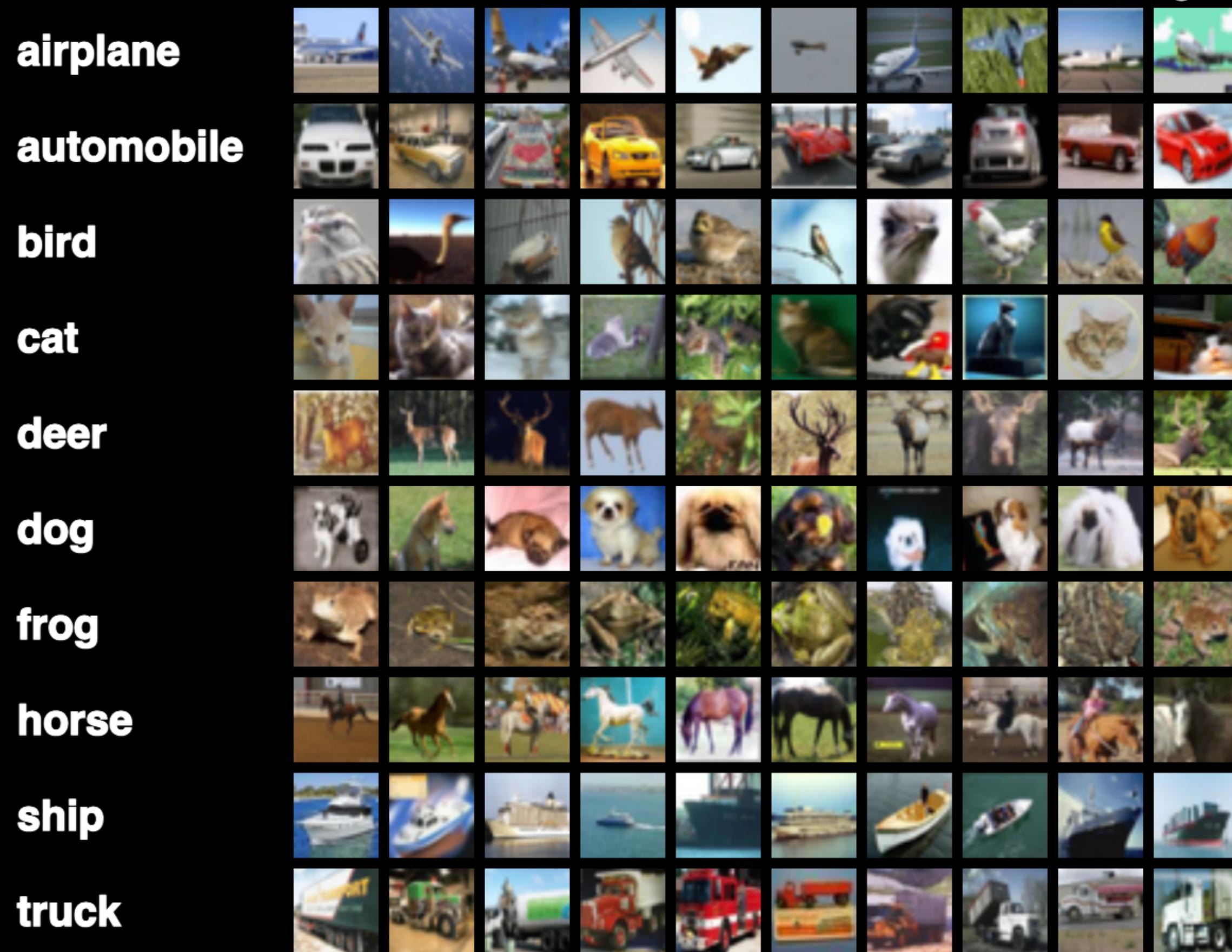
Conditional image synthesis

create an image according to some multi-modal guidance in the forms of textual descriptions, reference images, etc.

- L_{simple} only worse than the model proposed from Training generative adversarial networks with limited data, NeurIPS'20

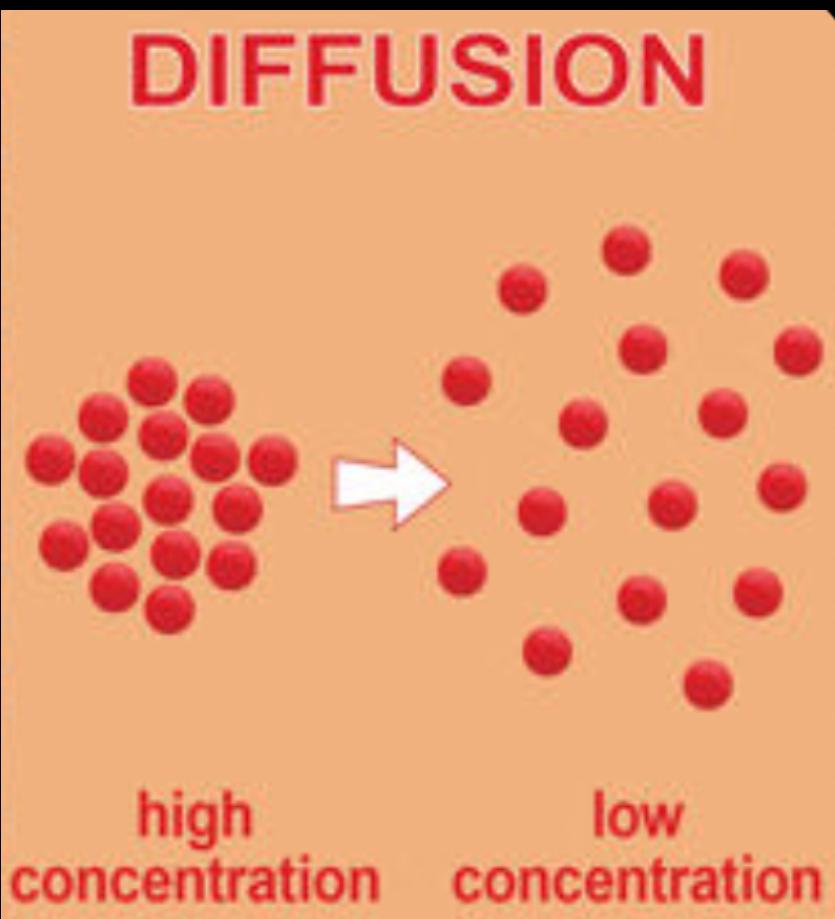
CIFAR-10 Datasets

CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes



Questions

- Q1. Why called diffusion?



Diffusion is a direct result of the second law of entropy.

The molecules spread out in all directions, lowering the molecules' concentrations in the original space.

The system state changed from Low entropy to High entropy.

- Q2. What's the difference between DDPM and Diffusion Model 2015?
 1. Use the ϵ as the training objective to explicit connection between diffusion models and denoising score matching.
 2. The model structure used in the reverse process is different:
DDPM use U-Net and Diffusion Model 2015 use the other CNN model propose in "Volumetric Semantic Segmentation Using Pyramid Context Features, IEEE'13 "

SWOT

Strengths

- Variational inference can have better performance in the case of fewer data and the case of data generation.
- The idea that use Markov chain as latent variable can achieve High performance on image synthesis.

Opportunities

- Any situation that can combined with Variational Inference.
- The idea of diffusion process may be used for recommender system

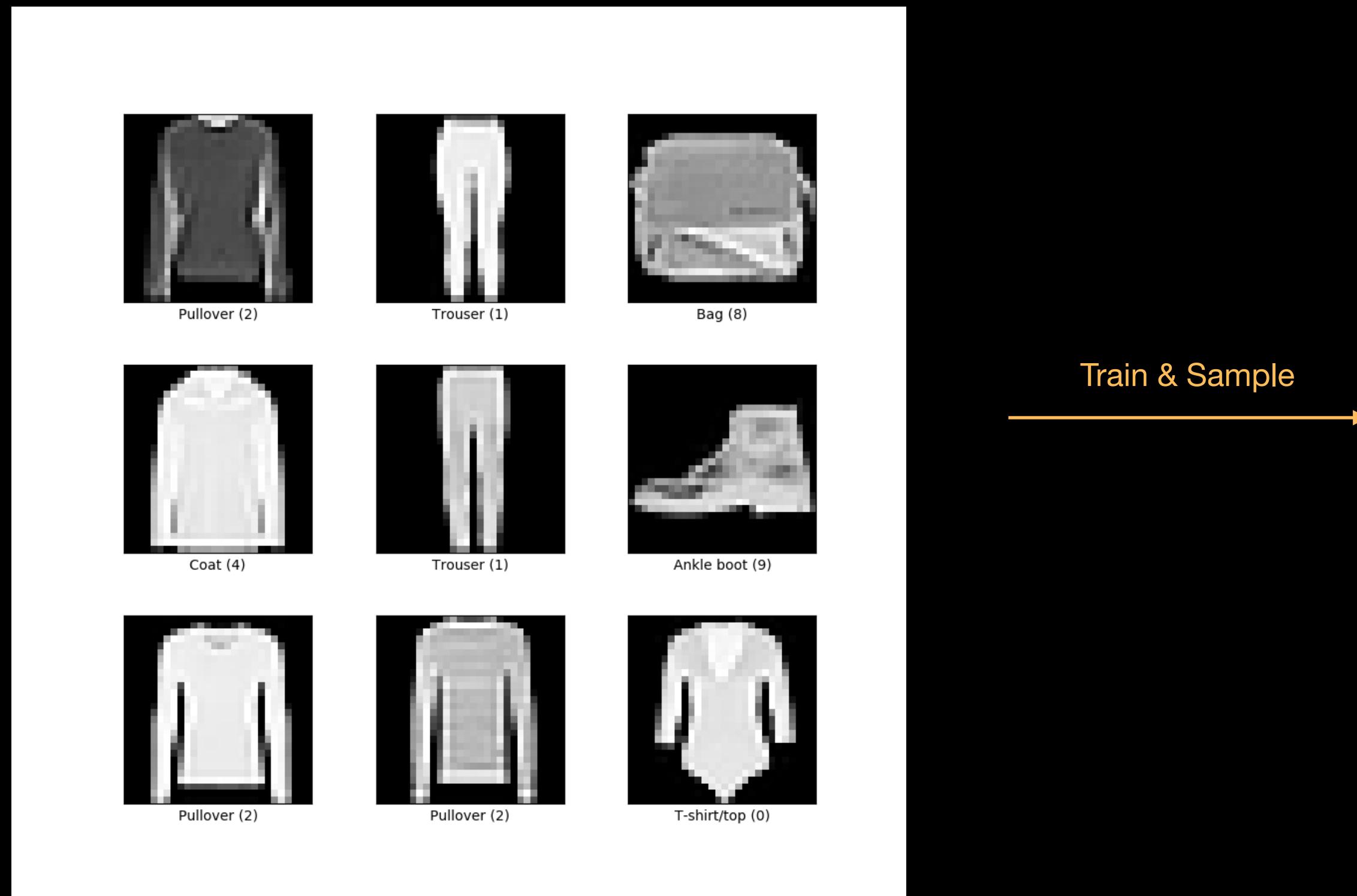
Weaknesses

- Sample Speed is low.
- Hard to use in Discrete Data.

Threats

- How to apply the diffusion model to discrete tradings data is a complex problem.

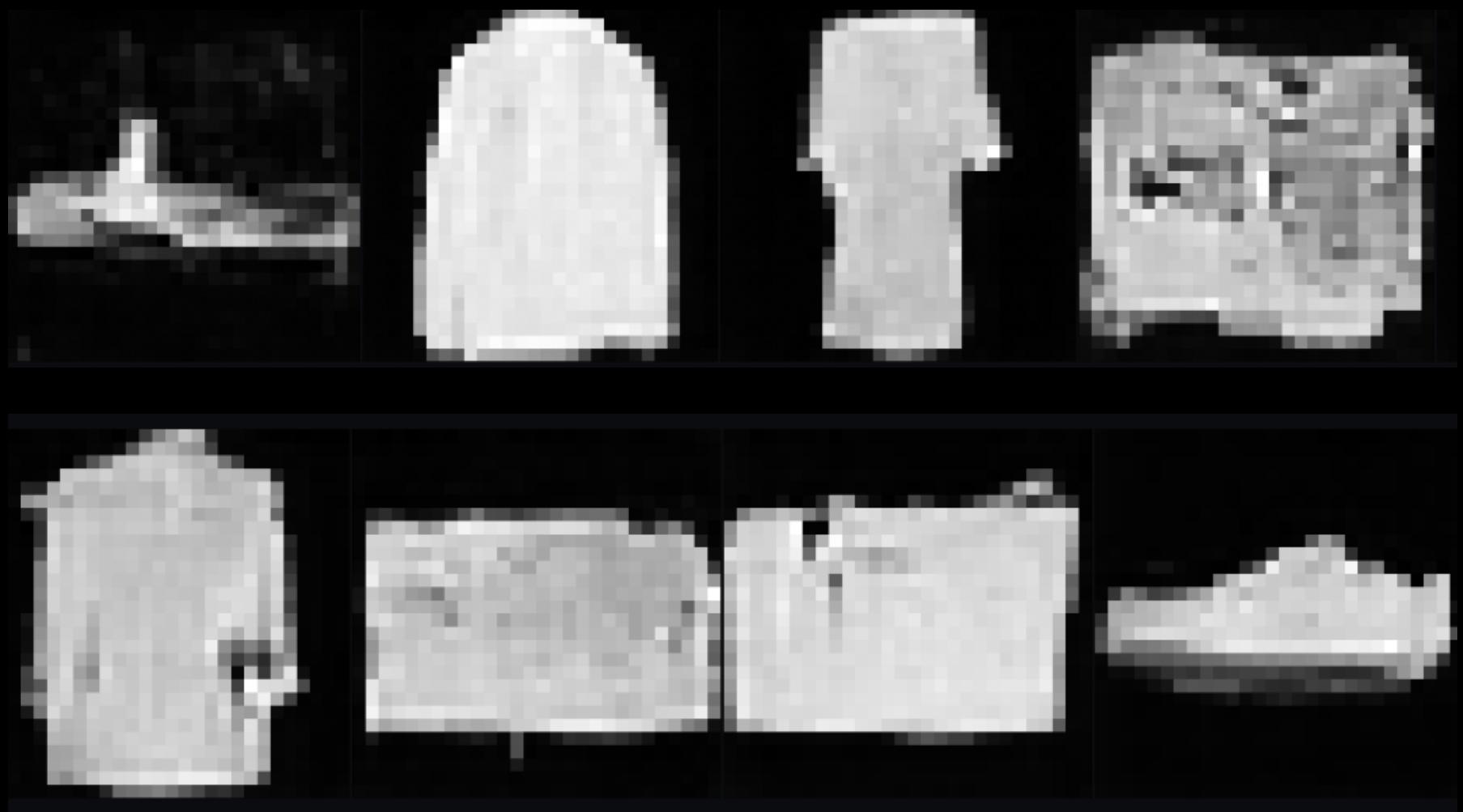
My Experiments



Fashion_mnist Dataset

Each example is a 28x28 grayscale image,
associated with a label from 10 classes.

(Pullover, Trouser, Bag, Coat...)



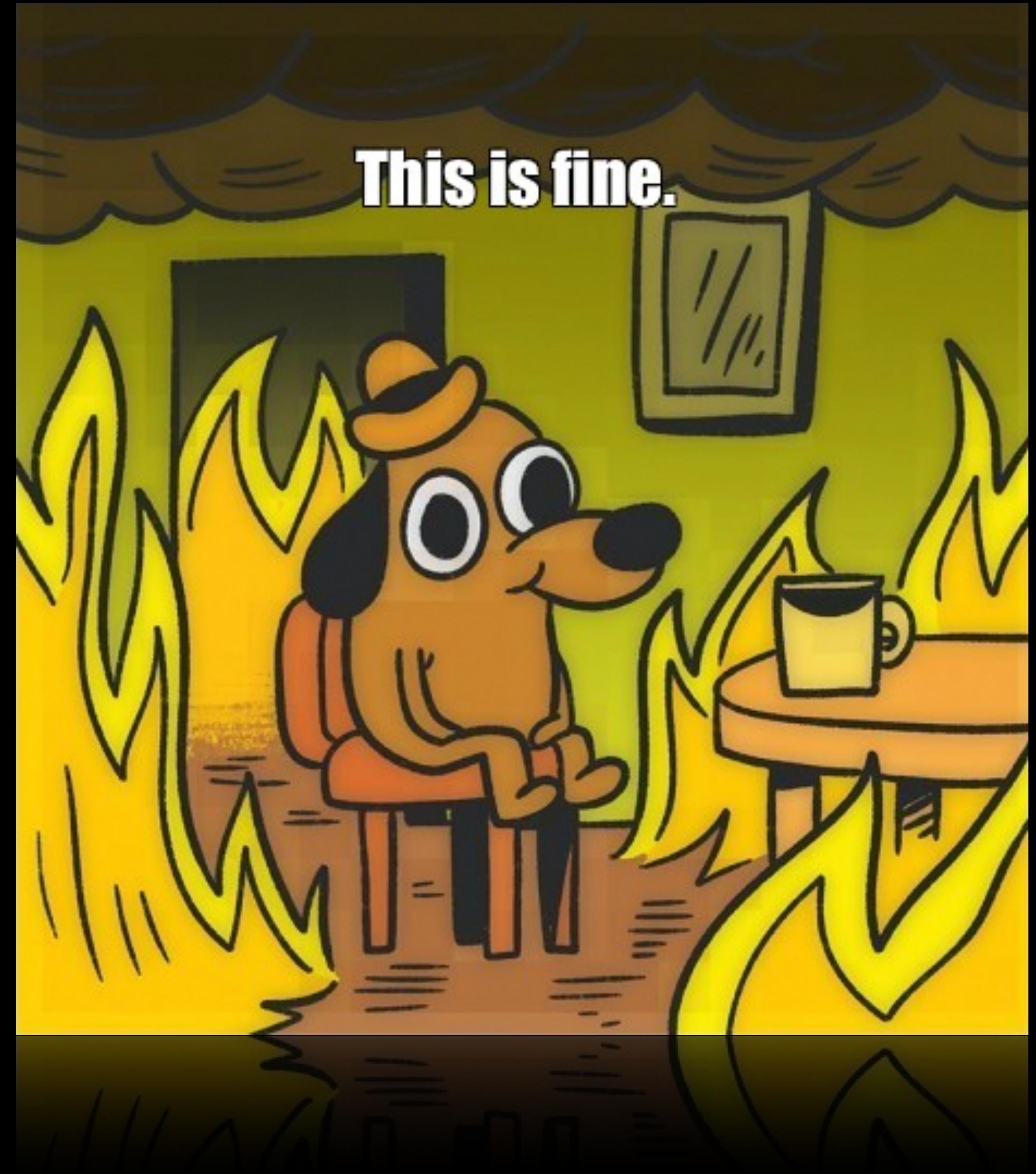
GeForce RTX 2070
Linear Variance schedule
epochs = 10
timesteps = 1000
training took 8m 28.2s
sampling 8 images took 10.2s
Memory usage 2749MiB

References

- Deep Unsupervised Learning using Nonequilibrium Thermodynamics, ICML'15
<https://arxiv.org/pdf/1503.03585.pdf>
- Denoising Diffusion Probabilistic Models, NIPS'20
<https://arxiv.org/pdf/2006.11239.pdf>
- Volumetric Semantic Segmentation using Pyramid Context Features, IEEE'13
<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6751540>
- From Autoencoder to Beta-VAE
<https://lilianweng.github.io/posts/2018-08-12-vae/>
- What are Diffusion Models?
<https://lilianweng.github.io/posts/2021-07-11-diffusion-models/>
- Hung-yiLee VAE Tutorial
https://www.youtube.com/watch?v=8zomhgKrsmA&ab_channel=Hung-yiLee

404

Appendix



Derivation flow chart

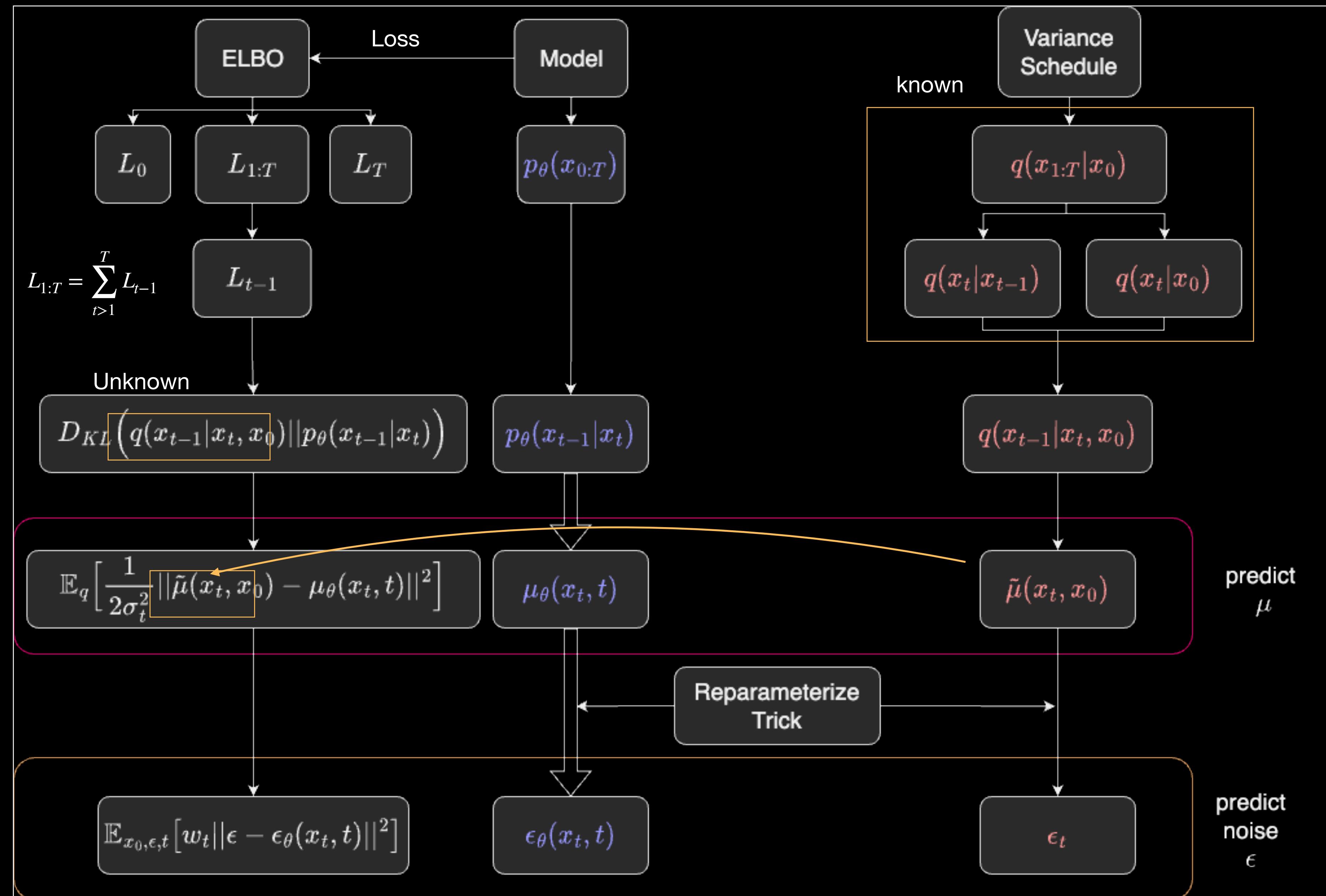
1. Split ELBO loss to different time step, and Ignore L_0 and L_T , we can focus on the loss between each step.

2. Use Bayes' Rule and Probability Density Function to derive $q(x_{t-1} | x_t)$ when conditioned on x_0 .

3. We can represent the distribution in μ and σ format when compute the KLD between two normal distribution

4. Use Reparameterize Trick

Convert $\tilde{\mu}_t$ to ϵ_t form



Step1. Split ELBO loss to different time step: $L = L_T + L_{1:T} + L_0$

From $x_{1:T}$ to the form of conditional probability

Reverse Process

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^T p_\theta(x_{t-1} | x_t),$$

$$p_\theta(x_{t-1} | x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

$$\begin{aligned} & \mathbb{E}_q \left[\log p_\theta(x_0 | x_{1:T}) - D_{\text{KL}}(q(x_{1:T} | x_0) || p_\theta(x_{1:T})) \right] \\ &= \mathbb{E}_q \left[\log p_\theta(x_0 | x_{1:T}) - \log \frac{q(x_{1:T} | x_0)}{p_\theta(x_{1:T})} \right] \\ &= \mathbb{E}_q \left[\log p_\theta(x_0 | x_{1:T}) + \log \frac{p_\theta(x_{1:T})}{q(x_{1:T} | x_0)} \right] \\ &= \mathbb{E}_q \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T} | x_0)} \right] \\ &= \mathbb{E}_q \left[\log \frac{p_\theta(x_T)p_\theta(x_0 | x_1)\dots p_\theta(x_{T-1} | x_T)}{q(x_1 | x_0)q(x_2 | x_1)\dots q(x_T | x_{T-1})} \right] \\ &= \mathbb{E}_q \left[\log p_\theta(x_T) + \sum_{t \geq 1} \log \frac{p_\theta(x_{t-1} | x_t)}{q(x_t | x_{t-1})} \right] \end{aligned}$$

Step1. Split ELBO loss to different time step: $L = L_T + L_{1:T} + L_0$

According to the Paper, split L into different parts: $L_T, L_{1:T}, L_0$

$$\begin{aligned}
L &= -\mathbb{E}_q \left[\log p(\mathbf{x}_T) + \sum_{t \geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_t | \mathbf{x}_{t-1})} \right] \\
&= -\mathbb{E}_q \left[\log p(\mathbf{x}_T) + \sum_{t>1} \log \frac{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_t | \mathbf{x}_{t-1})} + \log \frac{p_\theta(\mathbf{x}_0 | \mathbf{x}_1)}{q(\mathbf{x}_1 | \mathbf{x}_0)} \right] \\
&= -\mathbb{E}_q \left[\log p(\mathbf{x}_T) + \sum_{t>1} \log \frac{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_{t-1} | \mathbf{x}_t)} \cdot \frac{q(\mathbf{x}_{t-1} | \mathbf{x}_0)}{q(\mathbf{x}_t | \mathbf{x}_0)} + \log \frac{p_\theta(\mathbf{x}_0 | \mathbf{x}_1)}{q(\mathbf{x}_1 | \mathbf{x}_0)} \right] \\
&= -\mathbb{E}_q \left[\log p(\mathbf{x}_T) + \sum_{t>1} \log \frac{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_{t-1} | \mathbf{x}_t)} + \log \frac{q(\mathbf{x}_1 | \mathbf{x}_0)}{q(\mathbf{x}_2 | \mathbf{x}_0)} \cdot \frac{q(\mathbf{x}_2 | \mathbf{x}_0)}{q(\mathbf{x}_3 | \mathbf{x}_0)} \cdot \dots \cdot \frac{q(\mathbf{x}_{T-1} | \mathbf{x}_0)}{q(\mathbf{x}_T | \mathbf{x}_0)} + \log \frac{p_\theta(\mathbf{x}_0 | \mathbf{x}_1)}{q(\mathbf{x}_1 | \mathbf{x}_0)} \right] \\
&= -\mathbb{E}_q \left[\log p(\mathbf{x}_T) + \sum_{t>1} \log \frac{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_{t-1} | \mathbf{x}_t)} + \log \frac{q(\mathbf{x}_1 | \mathbf{x}_0)}{q(\mathbf{x}_T | \mathbf{x}_0)} + \log \frac{p_\theta(\mathbf{x}_0 | \mathbf{x}_1)}{q(\mathbf{x}_1 | \mathbf{x}_0)} \right] \\
&= -\mathbb{E}_q \left[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T | \mathbf{x}_0)} + \sum_{t>1} \log \frac{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_{t-1} | \mathbf{x}_t)} + \log p_\theta(\mathbf{x}_0 | \mathbf{x}_1) \right] \\
&= \mathbb{E}_q \left[D_{\text{KL}}(q(\mathbf{x}_T | \mathbf{x}_0) || p(\mathbf{x}_T)) + \sum_{t>1} D_{\text{KL}}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) || p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)) - \log p_\theta(\mathbf{x}_0 | \mathbf{x}_1) \right] \\
&= L_T + L_{1:T} + L_0
\end{aligned}$$

Step1. Split ELBO loss to different time step: $L = L_T + L_{1:T} + L_0$

Ignore L_T and L_0

$$\begin{aligned}&= \mathbb{E}_q \left[D_{\text{KL}}(q(\mathbf{x}_T | \mathbf{x}_0) || p(\mathbf{x}_T)) + \sum_{t>1} D_{\text{KL}}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) || p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)) - \log p_\theta(\mathbf{x}_0 | \mathbf{x}_1) \right] \\&= \cancel{L_T} + L_{1:T} + \cancel{L_0}\end{aligned}$$

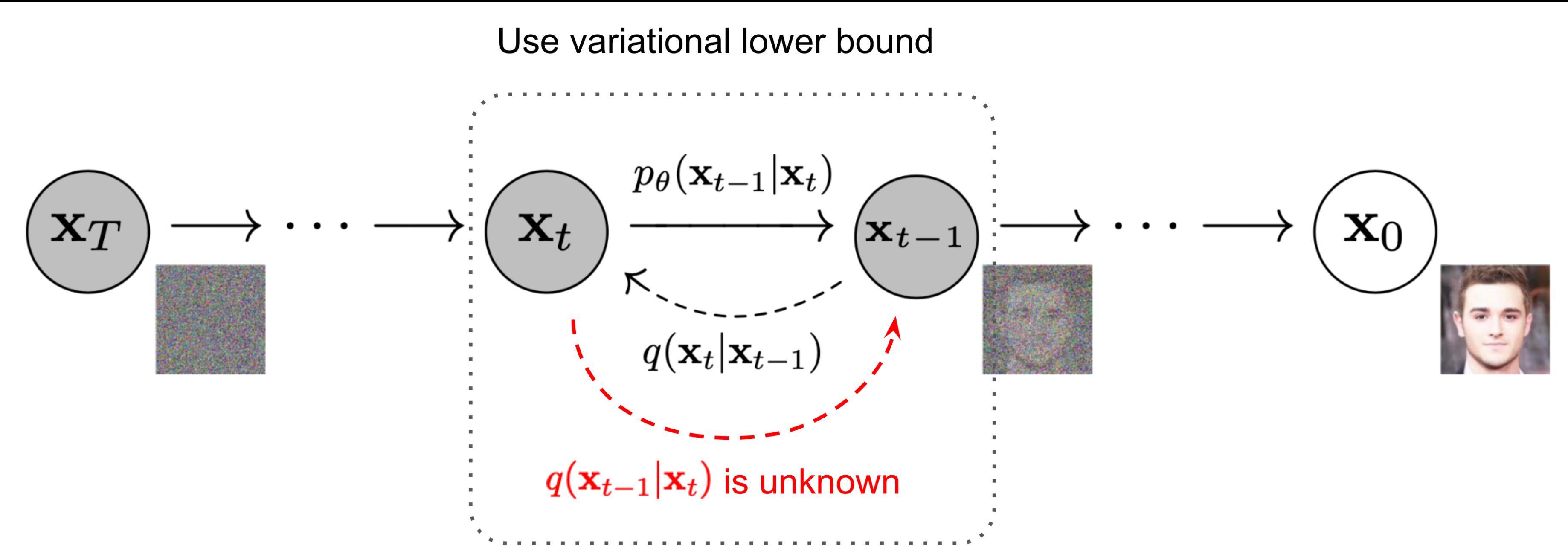
$q(x_T | x_0)$ & $p(x_T)$ is fixed, and no parameters θ need to be trained, so L_T is a const and can be ignore.



$$L_{t-1} = D_{KL}(q(x_{t-1} | x_t, x_0) || p_\theta(x_{t-1} | x_t))$$

Step1. Split ELBO loss to different time step: $L = L_T + L_{1:T} + L_0$

To compute the L_{t-1} , we have to know $q(x_{t-1} | x_t)$ first

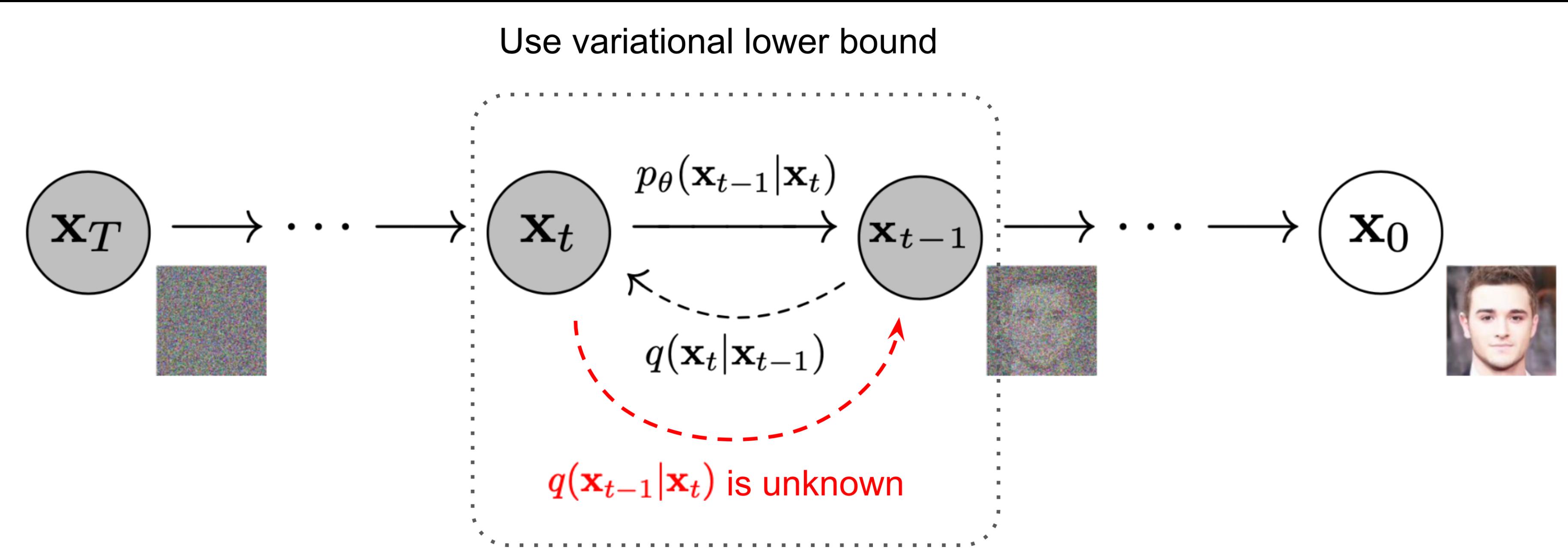


$$L_{t-1} = D_{KL}(q(x_{t-1} | x_t) || p_\theta(x_{t-1} | x_t))$$

unknown

Step1. Split ELBO loss to different time step: $L = L_T + L_{1:T} + L_0$

In forward process, we know that $q(x_{t-1} | x_t)$ can be tractable when conditioned on x_0



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

tractable when conditioned on x_0

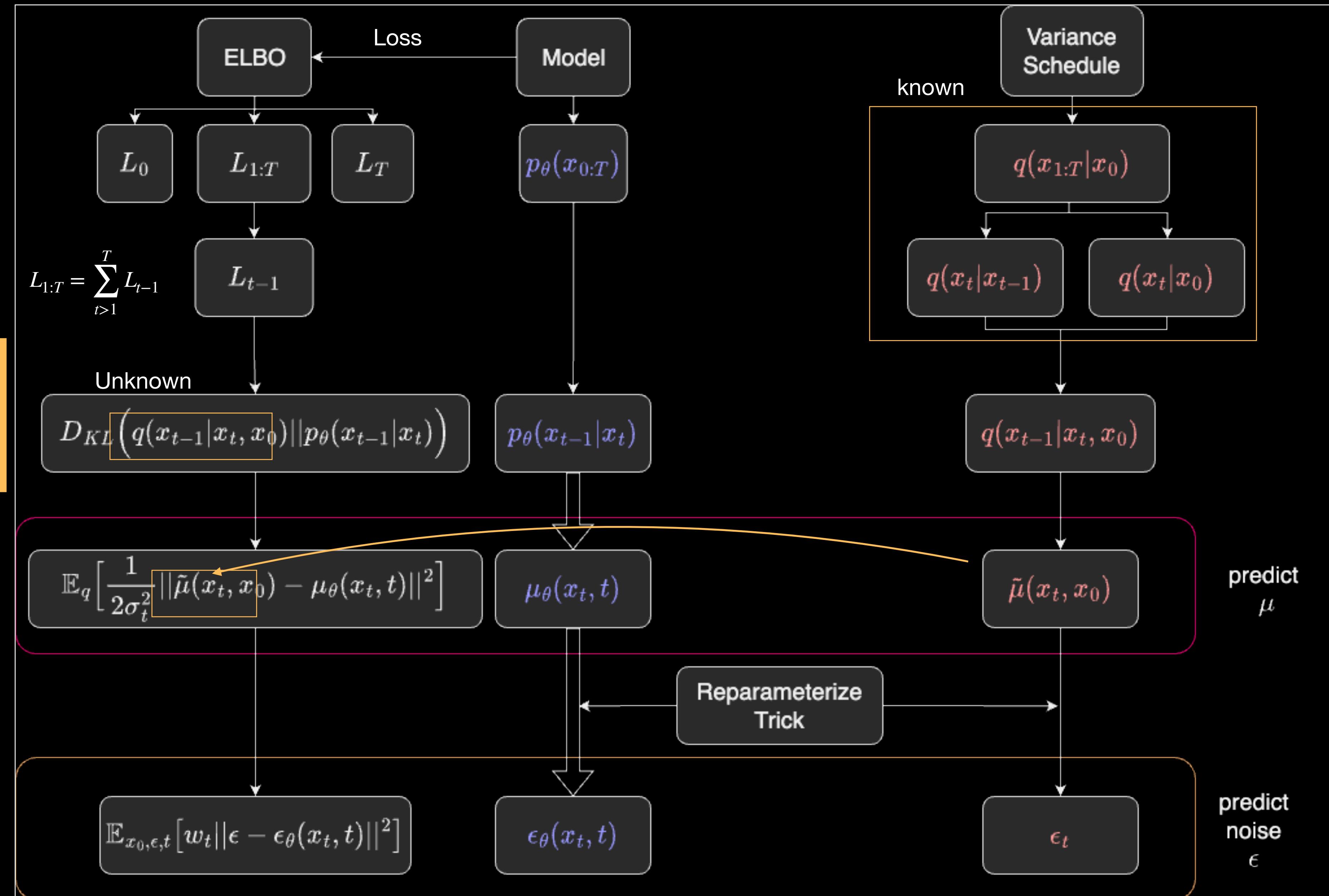
Derivation flow chart

1. Split ELBO loss to different time step, and ignore L_0 and L_T , we can focus on the loss between each step.

2. Use Bayes' Rule and Probability Density Function to tract $q(x_{t-1} | x_t)$ when conditioned on x_0 .

3. We can represent the distribution in μ and σ format when compute the KLD between two normal distribution

4. Use Reparameterize Trick Convert $\tilde{\mu}_t$ to ϵ_t form



Step2. Use Bayes' Rule and Probability density function to derive $q(x_{t-1} | x_t)$ when conditioned on x_0

Assume

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\boldsymbol{\beta}}_t \mathbf{I})$$

Known $q(x_t | x_{t-1}) := N(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I)$

$$\alpha_t := 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{s=1}^t \alpha_s$$

$$q(x_t | x_0) = N(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) I)$$

PDF: $\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

$$q(x_{t-1} | x_t, x_0) = q(x_t | x_{t-1}, x_0) \times \frac{q(x_{t-1} | x_0)}{q(x_t | x_0)}$$

$$\propto \exp \left(-\frac{1}{2} \left(\frac{(x_t - \sqrt{\alpha_t} x_{t-1})^2}{\beta_t} + \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}} x_0)^2}{1 - \bar{\alpha}_{t-1}} - \frac{(x_t - \sqrt{\bar{\alpha}_t} x_0)^2}{1 - \bar{\alpha}_t} \right) \right)$$

$$= \exp \left(-\frac{1}{2} \left(\frac{\mathbf{x}_t^2 - 2\sqrt{\alpha_t} \mathbf{x}_t \mathbf{x}_{t-1} + \alpha_t \mathbf{x}_{t-1}^2}{\beta_t} + \frac{\mathbf{x}_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0 \mathbf{x}_{t-1} + \bar{\alpha}_{t-1} \mathbf{x}_0^2}{1 - \bar{\alpha}_{t-1}} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0)^2}{1 - \bar{\alpha}_t} \right) \right)$$

$$= \exp \left(-\frac{1}{2} \left(\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1}^2 - \left(\frac{2\sqrt{\alpha_t}}{\beta_t} \mathbf{x}_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} \mathbf{x}_0 \right) \mathbf{x}_{t-1} + C(\mathbf{x}_t, \mathbf{x}_0) \right) \right)$$

Step2. Use Bayes' Rule and Probability density function to derive $q(x_{t-1} | x_t)$ when conditioned on x_0

Assume

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\boldsymbol{\beta}}_t \mathbf{I})$$

Known $q(x_t | x_{t-1}) := N(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I)$

$$\alpha_t := 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{s=1}^t \alpha_s$$

$$q(x_t | x_0) = N(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) I)$$

PDF: $\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

$$q(x_{t-1} | x_t, x_0) = q(x_t | x_{t-1}, x_0) \times \frac{q(x_{t-1} | x_0)}{q(x_t | x_0)}$$

Left term of PDF

$$\begin{aligned} & \propto \exp \left(-\frac{1}{2} \left(\frac{(x_t - \sqrt{\alpha_t} x_{t-1})^2}{\beta_t} + \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}} x_0)^2}{1 - \bar{\alpha}_{t-1}} - \frac{(x_t - \sqrt{\bar{\alpha}_t} x_0)^2}{1 - \bar{\alpha}_t} \right) \right) \\ & = \exp \left(-\frac{1}{2} \left(\frac{\mathbf{x}_t^2 - 2\sqrt{\alpha_t} \mathbf{x}_t \mathbf{x}_{t-1} + \alpha_t \mathbf{x}_{t-1}^2}{\beta_t} + \frac{\mathbf{x}_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0 \mathbf{x}_{t-1} + \bar{\alpha}_{t-1} \mathbf{x}_0^2}{1 - \bar{\alpha}_{t-1}} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0)^2}{1 - \bar{\alpha}_t} \right) \right) \\ & = \exp \left(-\frac{1}{2} \left(\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1}^2 - \left(\frac{2\sqrt{\alpha_t}}{\beta_t} \mathbf{x}_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} \mathbf{x}_0 \right) \mathbf{x}_{t-1} + C(\mathbf{x}_t, \mathbf{x}_0) \right) \right) \end{aligned}$$

$$\begin{aligned} & \frac{1}{\sigma_{q(x_t | x_{t-1}, x_0)} \sqrt{2\pi}} \frac{1}{\sigma_{q(x_{t-1} | x_0)} \sqrt{2\pi}} \frac{1}{\sigma_{q(x_t | x_0)} \sqrt{2\pi}} \\ & = \frac{1}{\beta_t \sqrt{2\pi}} \frac{1}{(1 - \bar{\alpha}_{t-1}) \sqrt{2\pi}} \frac{1}{(1 - \bar{\alpha}_t) \sqrt{2\pi}} \end{aligned}$$

Step2. Use Bayes' Rule and Probability density function to derivate $q(x_{t-1} | x_t)$ when conditioned on x_0

Assume

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\boldsymbol{\beta}}_t \mathbf{I})$$

Known

PDF $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

$$\exp \left(-\frac{1}{2} \left(\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1}^2 - \left(\frac{2\sqrt{\alpha_t}}{\beta_t} \mathbf{x}_t + \frac{2\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} \mathbf{x}_0 \right) \mathbf{x}_{t-1} + C(\mathbf{x}_t, \mathbf{x}_0) \right) \right)$$

$$\tilde{\boldsymbol{\beta}}_t = 1 / \left(\frac{\alpha_t}{\beta_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} \right) = 1 / \left(\frac{\alpha_t - \bar{\alpha}_t + \beta_t}{\beta_t(1 - \bar{\alpha}_{t-1})} \right) = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \cdot \beta_t$$

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) &= \left(\frac{\sqrt{\alpha_t}}{\beta_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} \mathbf{x}_0 \right) / \left(\frac{\alpha_t}{\beta_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} \right) \\ &= \left(\frac{\sqrt{\alpha_t}}{\beta_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} \mathbf{x}_0 \right) \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \cdot \beta_t \\ &= \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} \mathbf{x}_0 \end{aligned}$$

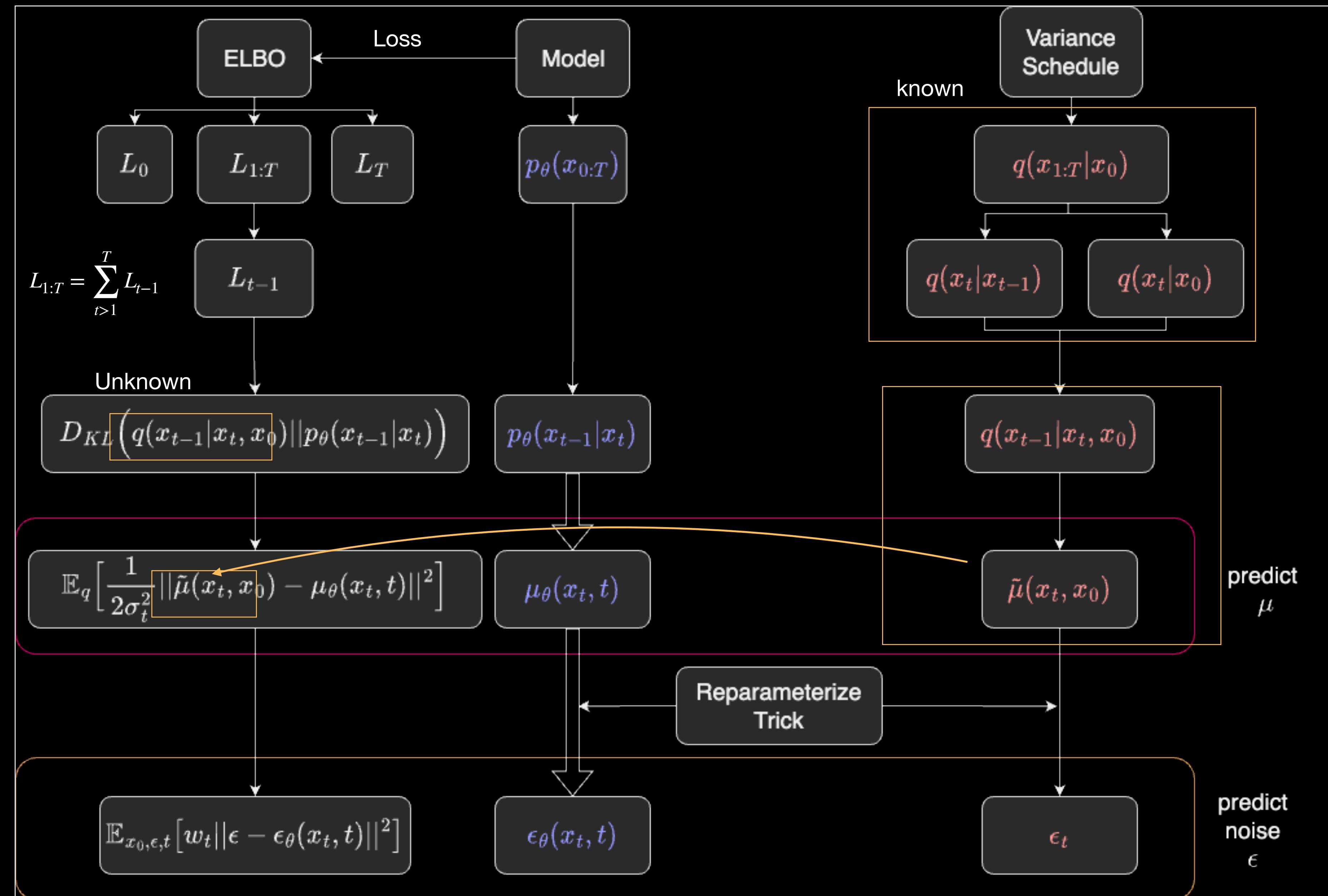
Derivation flow chart

1. Split ELBO loss to different time step, and ignore L_0 and L_T , we can focus on the loss between each step.

2. Use Bayes' Rule and Probability Density Function to derive $q(x_{t-1} | x_t)$ when conditioned on x_0 .

3. We can represent the distribution in μ and σ format when compute the KLD between two normal distribution

4. Use Reparameterize Trick Convert $\tilde{\mu}_t$ to ϵ_t form



Step3. KLD between two normal distribution

Use formula of KLD between two normal distribution to derivate KLD into μ form

Formula of KLD between two Normal Distribution

when $\sigma_1 = \sigma_2$

$$\log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

$$\log \frac{\sigma_1}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_1^2} - \frac{1}{2} = \boxed{\frac{1}{2\sigma_1^2}(\mu_1 - \mu_2)^2}$$

$$L_{t-1} = D_{KL}(q(x_{t-1} | x_t, x_0) || p_\theta(x_{t-1} | x_t))$$

$$= \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} || \tilde{\mu}(x_t, x_0) - \mu_\theta(x_t, t) ||^2 \right]$$

Fixed Σ as σ_t , $\sigma_t^2 = \beta_t$ or $\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$:

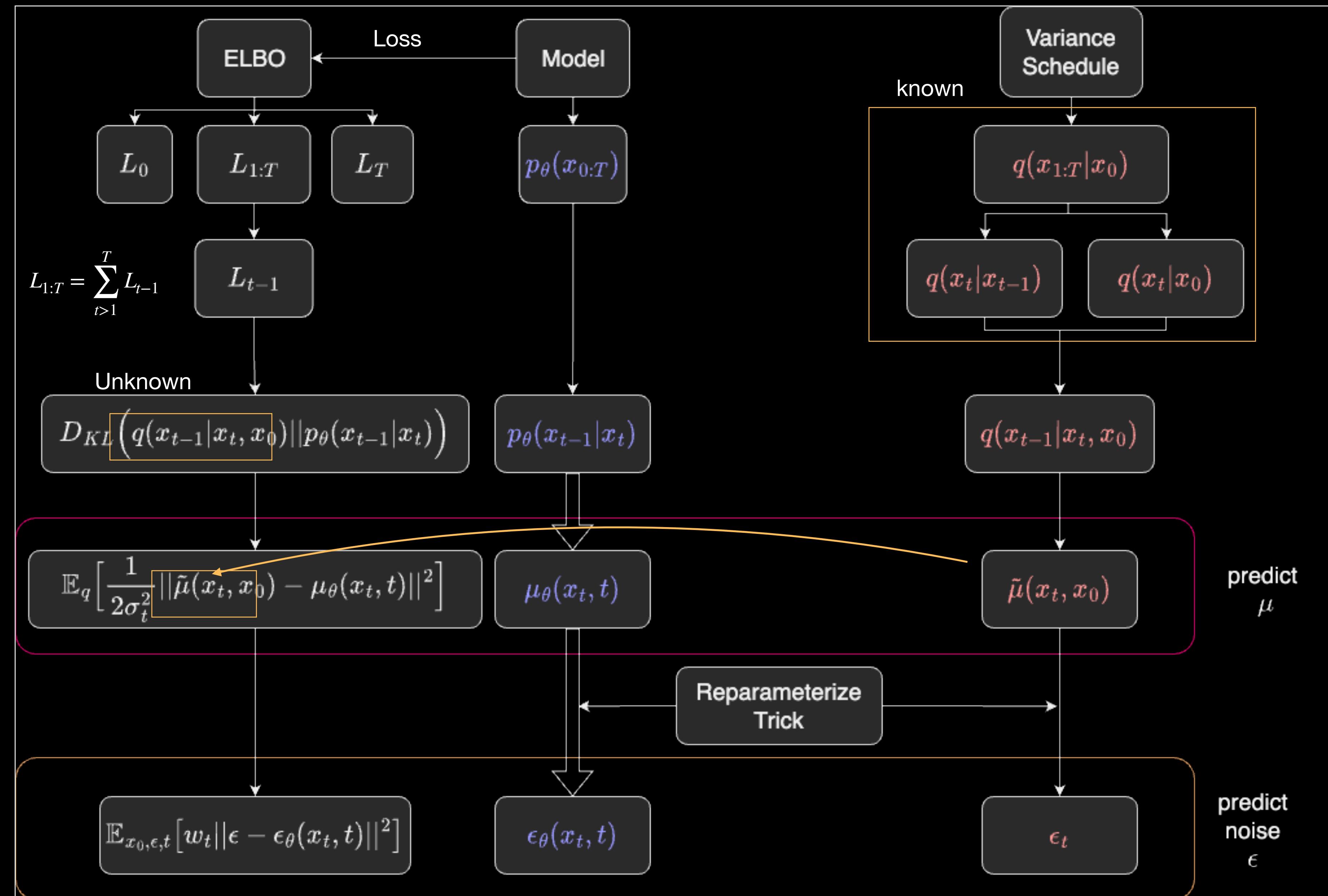
Derivation flow chart

1. Split ELBO loss to different time step, and ignore L_0 and L_T , we can focus on the loss between each step.

2. Use Bayes' Rule and Probability Density Function to derive $q(x_{t-1} | x_t)$ when conditioned on x_0 .

3. We can represent the distribution in μ and σ format when compute the KLD between two normal distribution

4. Use Reparameterize Trick
Convert $\tilde{\mu}_t$ to ϵ_t form



Use Reparameterize Trick Convert $\tilde{\mu}_t$ to ϵ_t form

Reparameterize Trick

$$z \sim q_{\phi}(z|x) = N(z; \mu, \sigma^2 I)$$

$$z = \mu + \sigma \odot \epsilon$$

$$\epsilon \sim N(0, I)$$

Use this trick,
we can simply use ϵ to represent μ and ϵ

Forward process's property

$$q(x_{1:T}|x_0) := \prod_{t=1}^T q(x_t|x_{t-1}),$$

$$q(x_t|x_{t-1}) := N(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I)$$

$$\begin{aligned}\mathbf{x}_t &= \sqrt{1 - \beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \epsilon_{t-1} \\ &= \sqrt{\alpha_t} \mathbf{x}_{t-1} + \sqrt{1 - \alpha_t} \epsilon_{t-1} \\ &= \sqrt{\alpha_t \alpha_{t-1}} \mathbf{x}_{t-2} + \sqrt{1 - \alpha_t \alpha_{t-1}} \bar{\epsilon}_{t-2} \\ &= \dots \\ &= \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon \\ \mathbf{x}_0 &= \frac{1}{\sqrt{\bar{\alpha}_t}} (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon)\end{aligned}$$

Use Reparameterize Trick Convert $\tilde{\mu}_t(x_t, x_0)$ to ϵ_t form

Now we have:

$$\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon)$$

Put x_0 into $\tilde{\mu}_t(x_t, x_0)$ to represent $\tilde{\mu}_t(x_t, x_0)$ in ϵ from

$$\begin{aligned}\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} \mathbf{x}_0 \\ &= \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} \frac{1}{\sqrt{\bar{\alpha}_t}} (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_t) \\ &= \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon \right)\end{aligned}$$

Use Reparameterize Trick Convert $\tilde{\mu}_t(x_t, x_0)$ to ϵ_t form

$$\tilde{\mu}_t(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_t \right)$$

Then set our training objective from $\mu_\theta(x_t, x_0)$ to $\epsilon_\theta(x_t, t)$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right)$$

Use Reparameterize Trick Convert $\tilde{\mu}_t(x_t, x_0)$ to ϵ_t form

Rewrite L_{t-1} with ϵ

$$\begin{aligned}
 L_{t-1} &= \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\mu}(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right] \\
 &= \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[\frac{1}{2\sigma_t^2} \left\| \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon \right) - \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) \right\|^2 \right] \\
 &= \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[\frac{(1 - \alpha_t)^2}{2\sigma_t^2 \alpha_t (1 - \bar{\alpha}_t)} \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2 \right] \\
 &= \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[\frac{(1 - \alpha_t)^2}{2\sigma_t^2} \alpha_t (1 - \bar{\alpha}_t) \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|^2 \right]
 \end{aligned}$$

Simplified step-specific weighting from ELBO

$$\text{loss} = \mathbb{E}_{\mathbf{x}_0, \epsilon, t} \left[\mathbf{w}_t \|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right]$$

- DDPM find that a simpler version of the variational bound that discards the term weights that appear in the original bound led to better sample quality.

$$\text{loss}_{\text{simple}} = \mathbb{E}_{\mathbf{x}_0, \epsilon, t} [\|\epsilon - \epsilon_\theta(x_t, t)\|^2] \#$$

Use Reparameterize Trick Convert $\tilde{\mu}_t(x_t, x_0)$ to ϵ_t form

Rewrite L_{t-1} with ϵ

$$\begin{aligned}
 L_{t-1} &= \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\mu}(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right] \\
 &= \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[\frac{1}{2\sigma_t^2} \left\| \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon \right) - \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) \right\|^2 \right] \\
 &= \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[\frac{(1 - \alpha_t)^2}{2\sigma_t^2 \alpha_t (1 - \bar{\alpha}_t)} \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2 \right] \\
 &= \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[\frac{(1 - \alpha_t)^2}{2\sigma_t^2} \alpha_t (1 - \bar{\alpha}_t) \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|^2 \right]
 \end{aligned}$$

Simplified step-specific weighting from ELBO

$$\text{loss} = \mathbb{E}_{\mathbf{x}_0, \epsilon, t} [\mathbf{w}_t \|\epsilon - \epsilon_\theta(x_t, t)\|^2]$$

Objective	IS	FID
L , fixed isotropic Σ	7.67 ± 0.13	13.51
$\ \tilde{\epsilon} - \epsilon_\theta\ ^2$ (L_{simple})	9.46 ± 0.11	3.17

- DDPM find that a simpler version of the variational bound that discards the term weights that appear in the original bound led to better sample quality.

$$\text{loss}_{\text{simple}} = \mathbb{E}_{\mathbf{x}_0, \epsilon, t} [\|\epsilon - \epsilon_\theta(x_t, t)\|^2] \quad \#$$