Denoising Diffusion Probabilistic Models (DDPM)

Presented by Chenwei Xu

Northwestern University

Part 1: Overview

Part 2: Diffusion Process

Part 3: Evidence Lower Bound (ELBO)

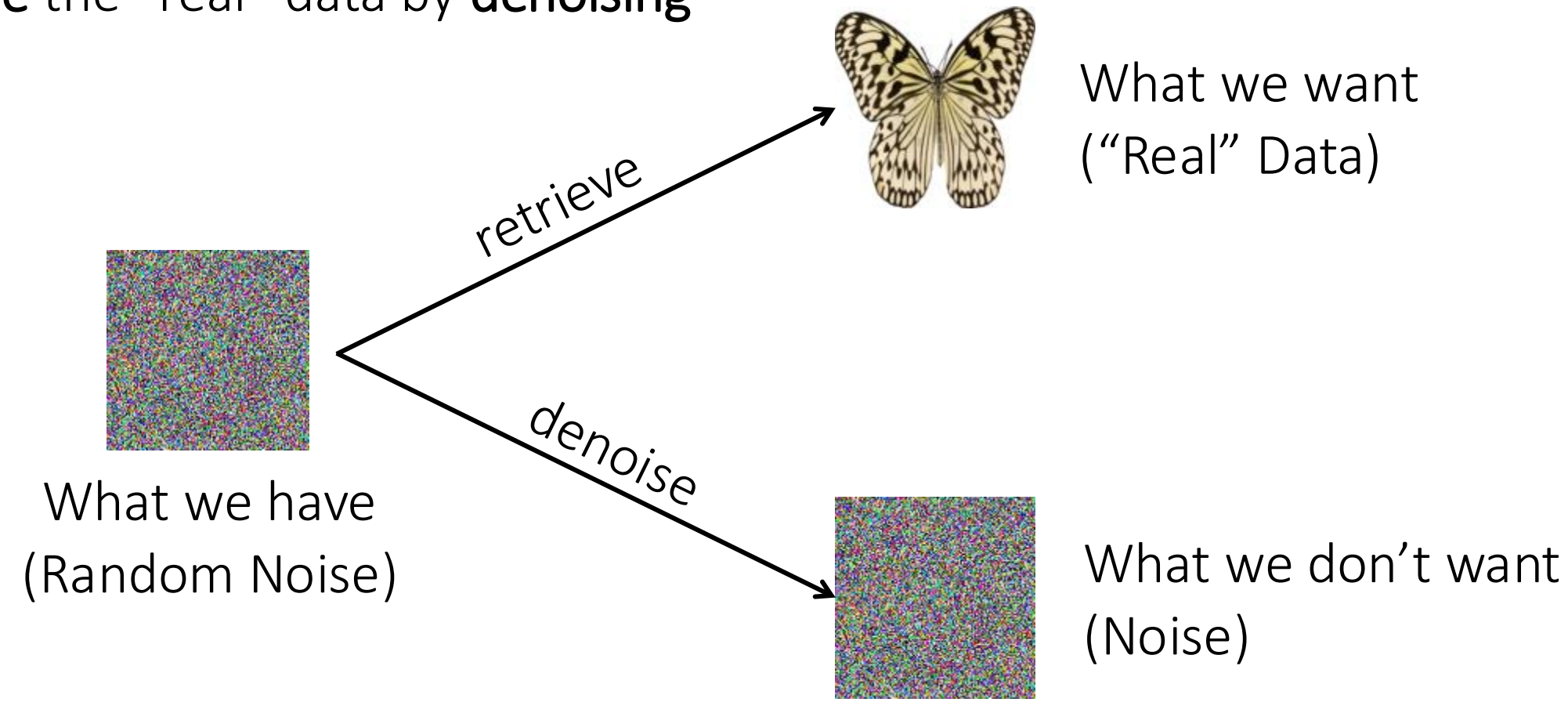
Part 4: ELBO of DDPM

Part 5: Training and Inference

Part 1: Overview

What is Diffusion Model?

Assumption: “Real data is already in Random Noise; Diffusion model **retrieve** the “real” data by **denoising**

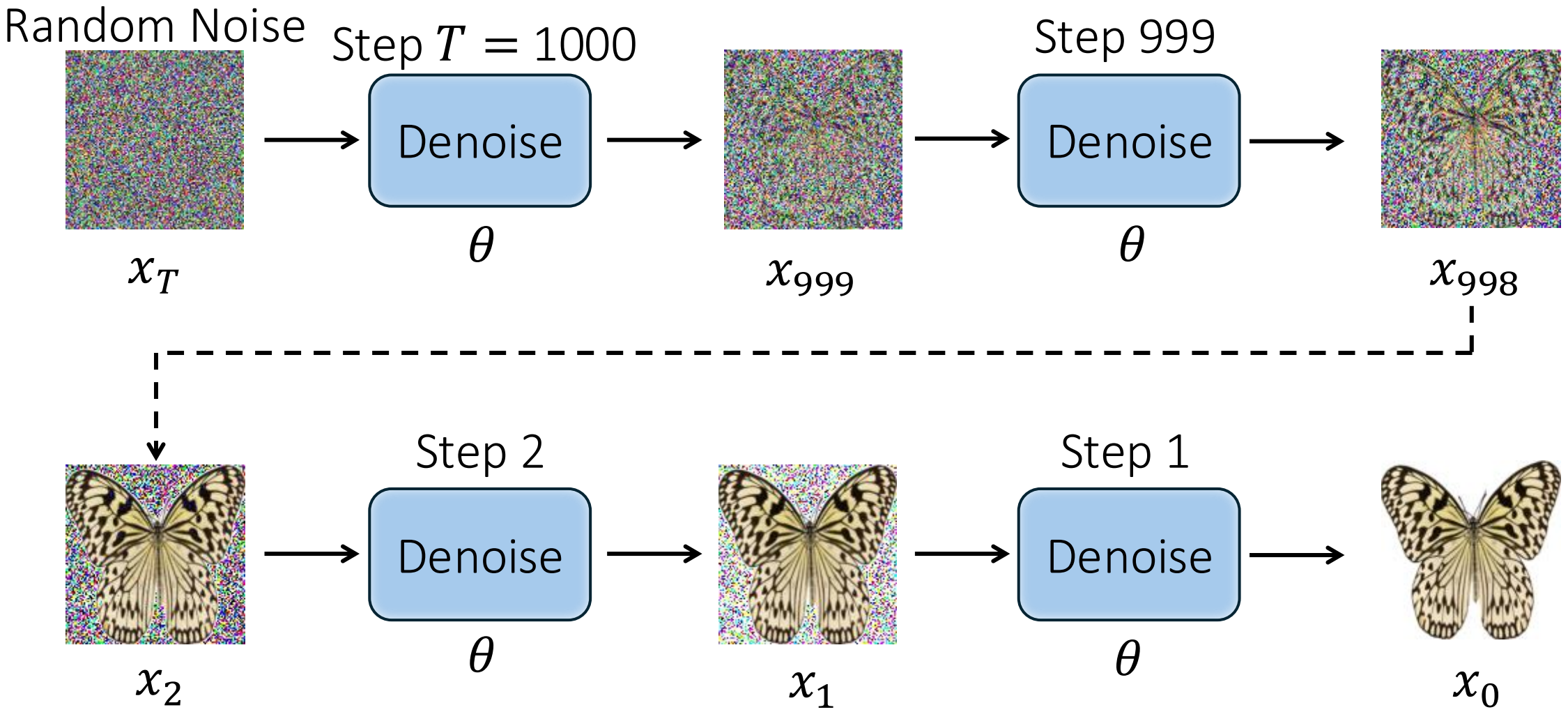


Basic Concepts

Forward (Diffusion) Process

Reverse (Generation) Process

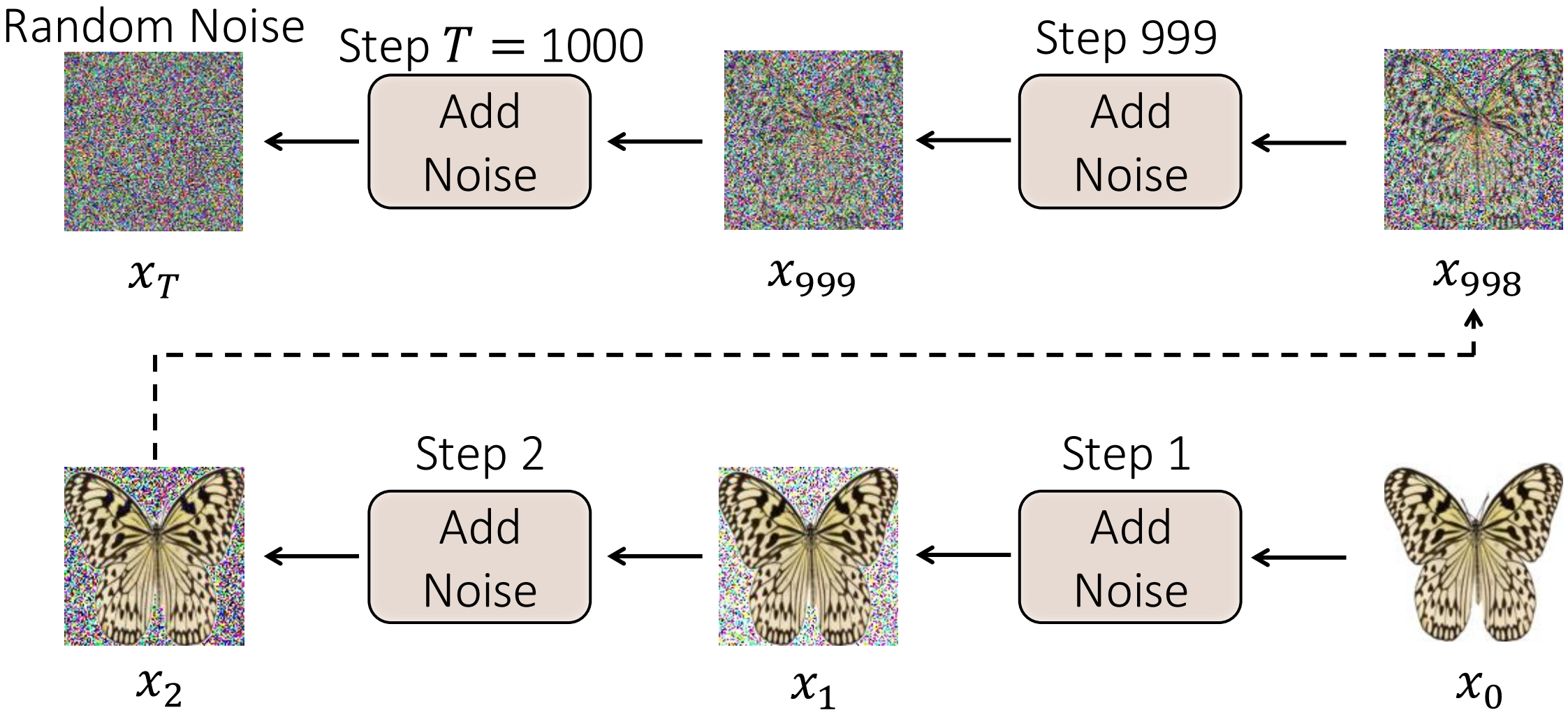
Reverse Process in DDPM



$$p_\theta(x_0) = \int_{x1:T} p(x_T)\, p_\theta(x_{T-1}|x_T) \cdots p_\theta(x_0|x_1) dx_1 : x_T$$

$p_\theta(x_{t-1}|x_t)$: Denoise process at t

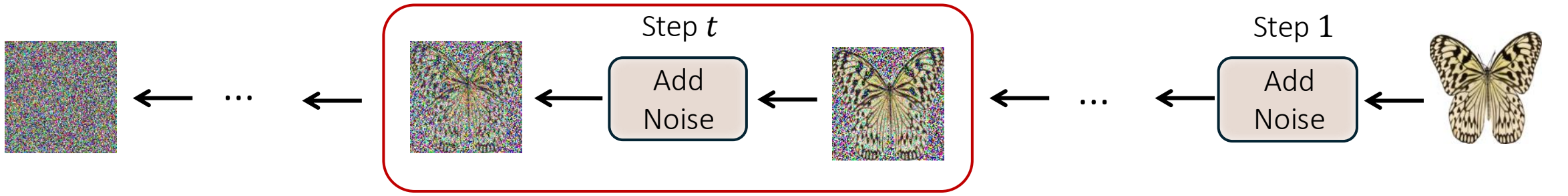
Diffusion Process in DDPM



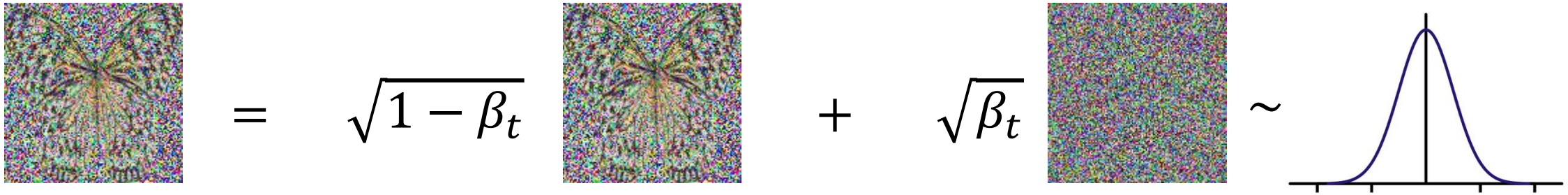
Also called: **Forward Process**

Part 2: Diffusion Process

$q(x_t|x_{t-1})$: Diffusion Process at t



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



$0 \leq \beta_1 \leq \beta_2 \leq \cdots \leq \beta_T \leq 1$: Noise Scheduler

$\sqrt{1-\beta_t}, \sqrt{\beta_t}$: $\mathrm{x}_t \rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I})$ as $t \rightarrow T$

Noise Scheduler Examples

Linear:

$$\beta_t = \beta_1 + (\beta_T - \beta_1) \times \frac{t}{T}$$

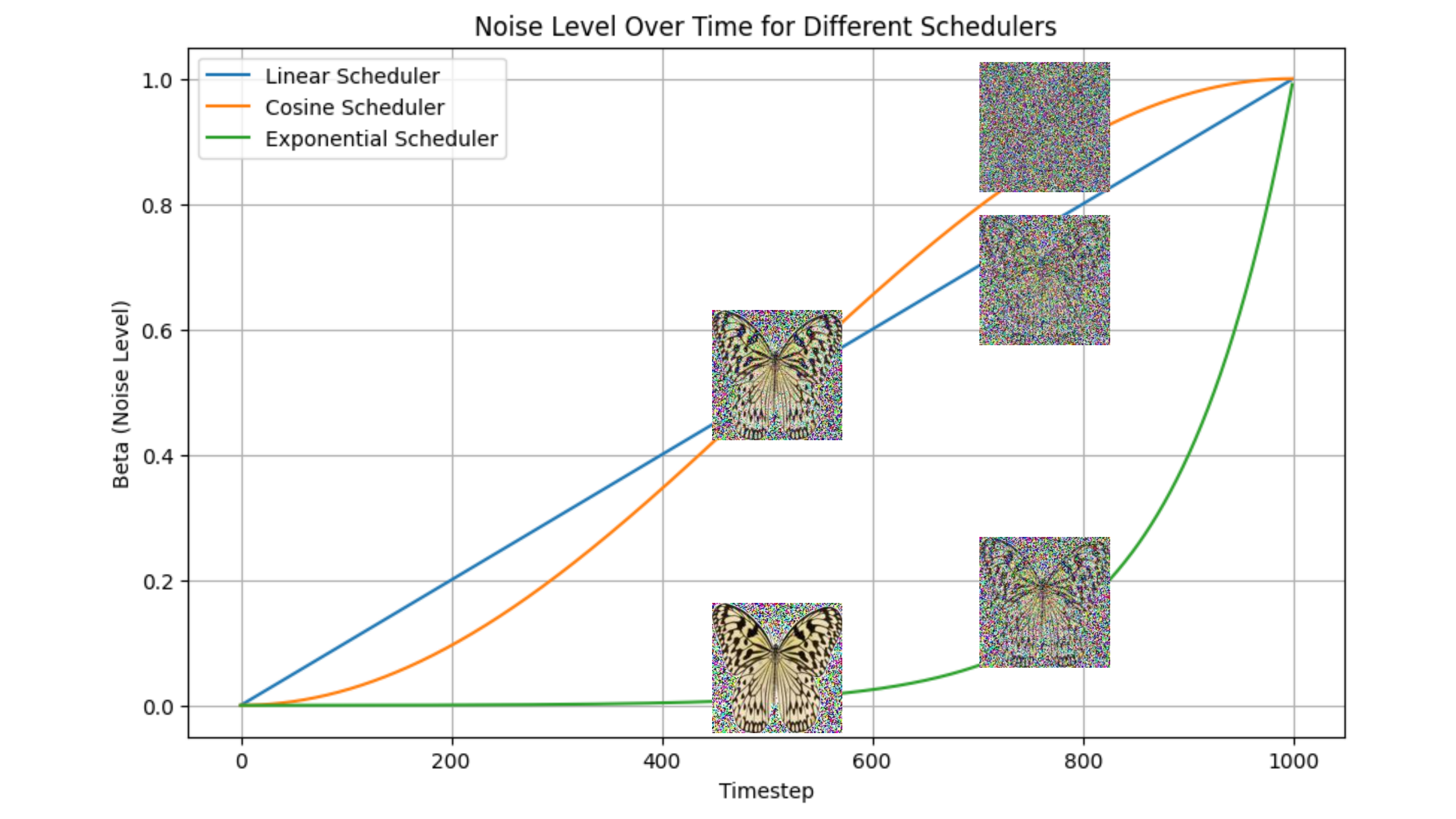
Cosine:

$$\beta_t = \beta_1 + (\beta_T - \beta_1) \times \frac{1-\cos(\pi \cdot t/T)}{2}$$

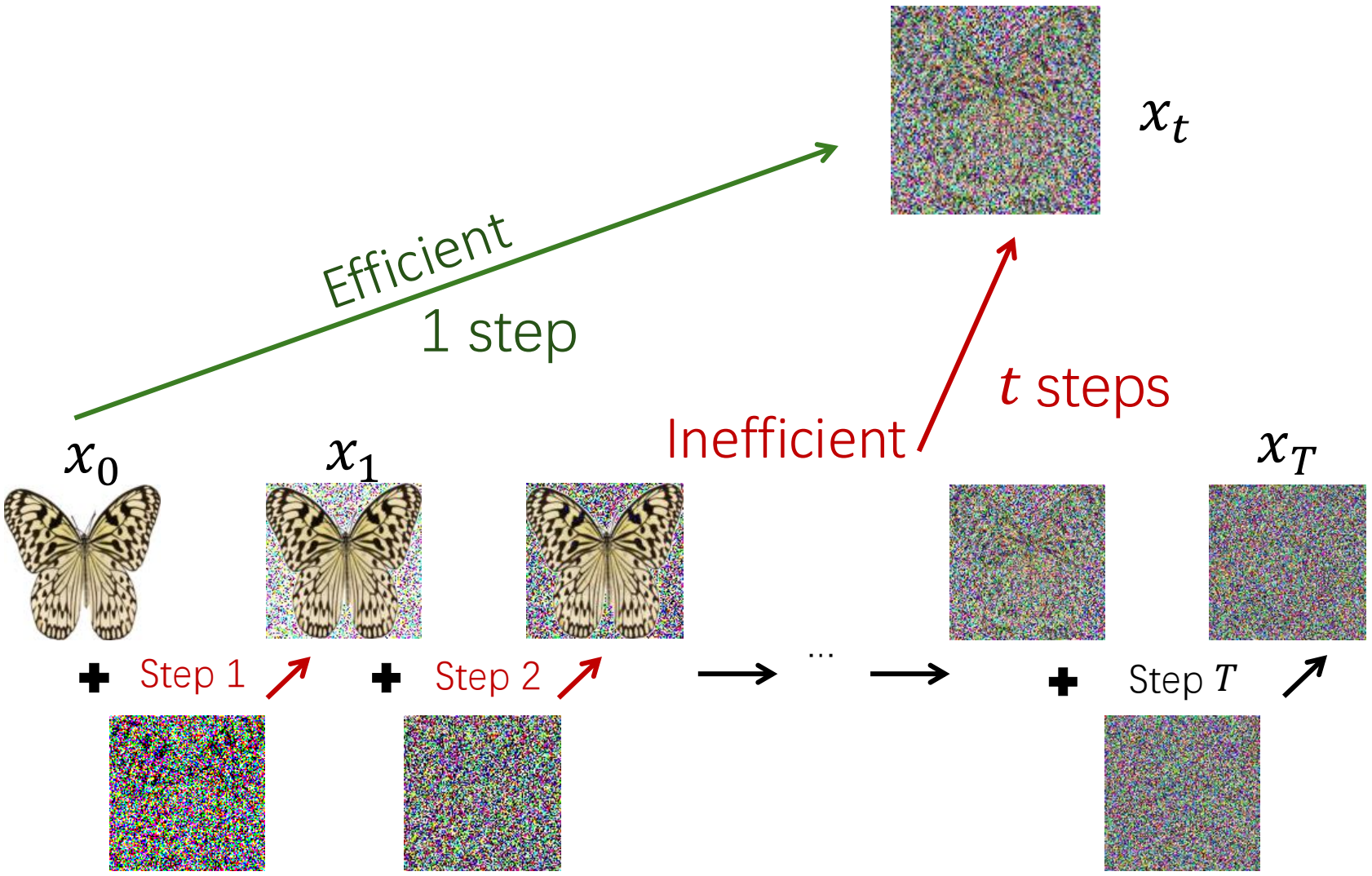
Geometric

$$\beta_t = \beta_1 \times \gamma^t, \gamma \geq 1$$

Others…

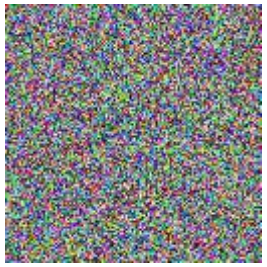

Noise Level Over Time for Different Schedulers
Linear Scheduler
Cosine Scheduler
Exponential Scheduler
Beta (Noise Level)
Timestep
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000

Efficiency in Adding Noise



$q(x_1|x_0)$:Drive x_1 by Adding Noise to $\mathbf{x_0}$

x_1 $= \sqrt{1-\beta_1}$  x_0 $+ \sqrt{\beta_1}$  ϵ_1: Noise $\sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

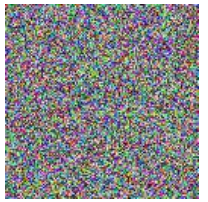
x_2 $= \sqrt{1-\beta_2}$  x_1 $+ \sqrt{\beta_2}$  ϵ_2: Noise $\sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$q(x_2|x_0)$: Drive x_2 by Adding Noise to x_0

x_2

$= \sqrt{1-\beta_2}\sqrt{1-\beta_1}$ 

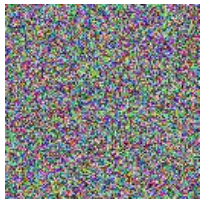
x_0: data

$+\sqrt{1-\beta_2}\sqrt{\beta_1}$ 

ϵ_1

$\sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

Ind.

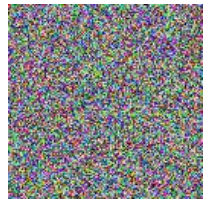
$+\sqrt{\beta_2}$ 

ϵ_2

$\sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$= \sqrt{1-\beta_2}\sqrt{1-\beta_1}$ 

x_0

$+\sqrt{1-(1-\beta_2)(1-\beta_1)}$ 

$\sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

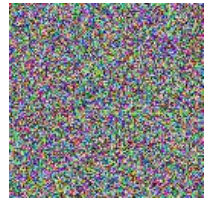
$q(\mathbf{x}_t|\mathbf{x}_0)$: Drive $\mathbf{x}_t$ by Adding Noise to $\mathbf{x}_0$ (Continued.)

x_t

$$= \sqrt{1-\beta_t}\cdots\sqrt{1-\beta_0}$$

x_0

$$+\sqrt{1-(1-\beta_t)\cdots(1-\beta_1)}$$

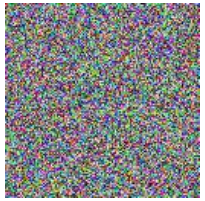


$\sim\mathcal{N}(\mathbf{0},\mathbf{I})$

$$= \sqrt{\bar{\alpha}_t}$$

x_0

$$+ \sqrt{1-\bar{\alpha}_t}$$



$\sim\mathcal{N}(0,\mathrm{I})$

$$\alpha_t = (1-\beta_t)$$
$$\bar{\alpha}_t = \alpha_1\alpha_2\cdots\alpha_t$$

Sample x_t from $\mathbf{x}_0$ in 1 time step

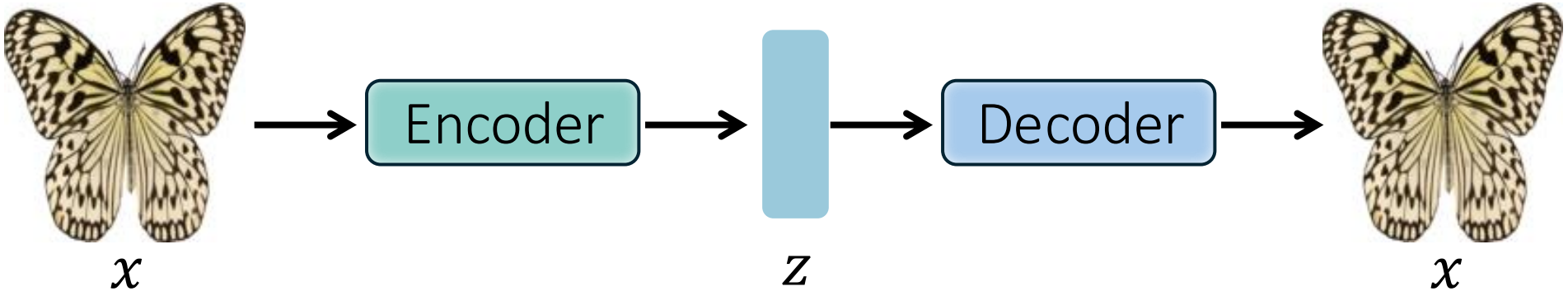
$$q(x_t|x_0) = \mathcal{N}(x_t|\sqrt{\bar{\alpha}_t}x_0, \sqrt{1-\bar{\alpha}_t}\mathbf{I})$$

Part 3: Evidence Lower Bond

VAE vs. Diffusion Model

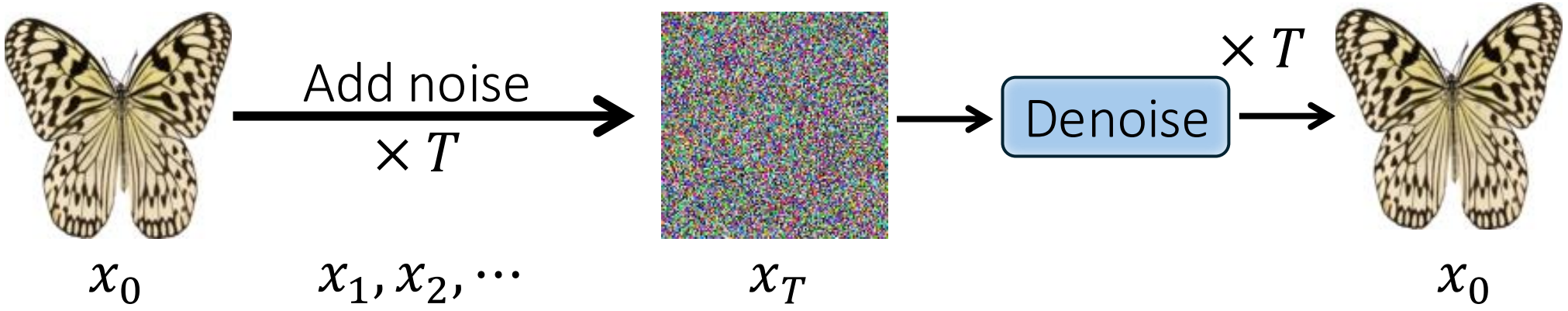
VAE: Variation Auto Encoder

VAE (another kind of generative model)

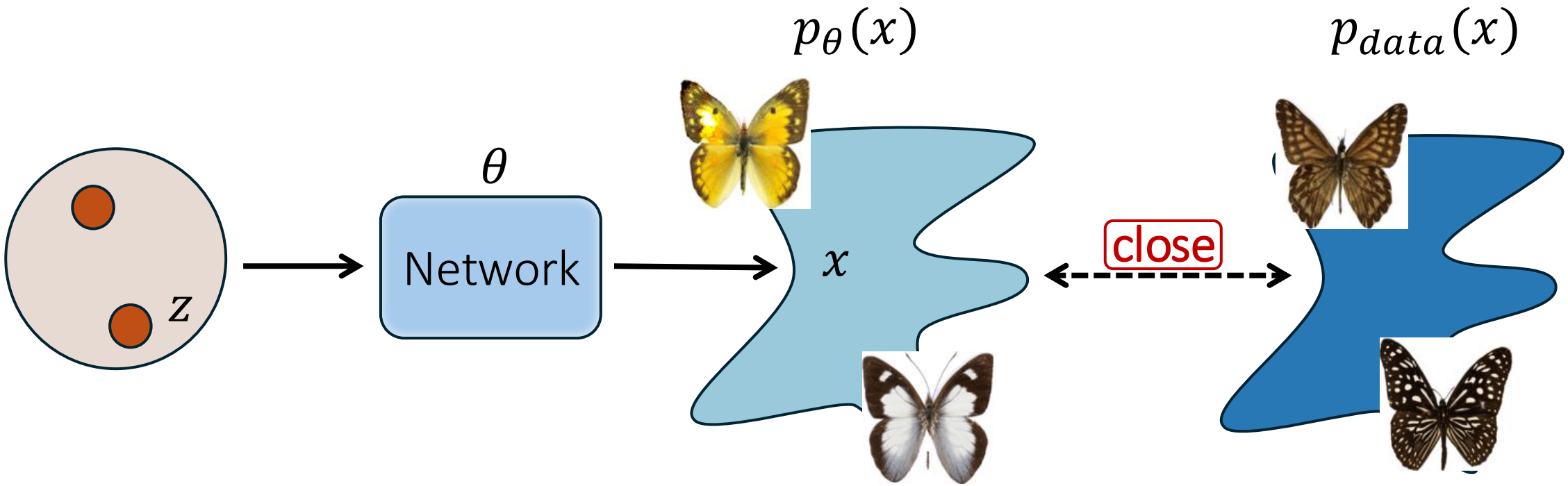


Use VAE as a trivial example for explaining diffusion

Diffusion

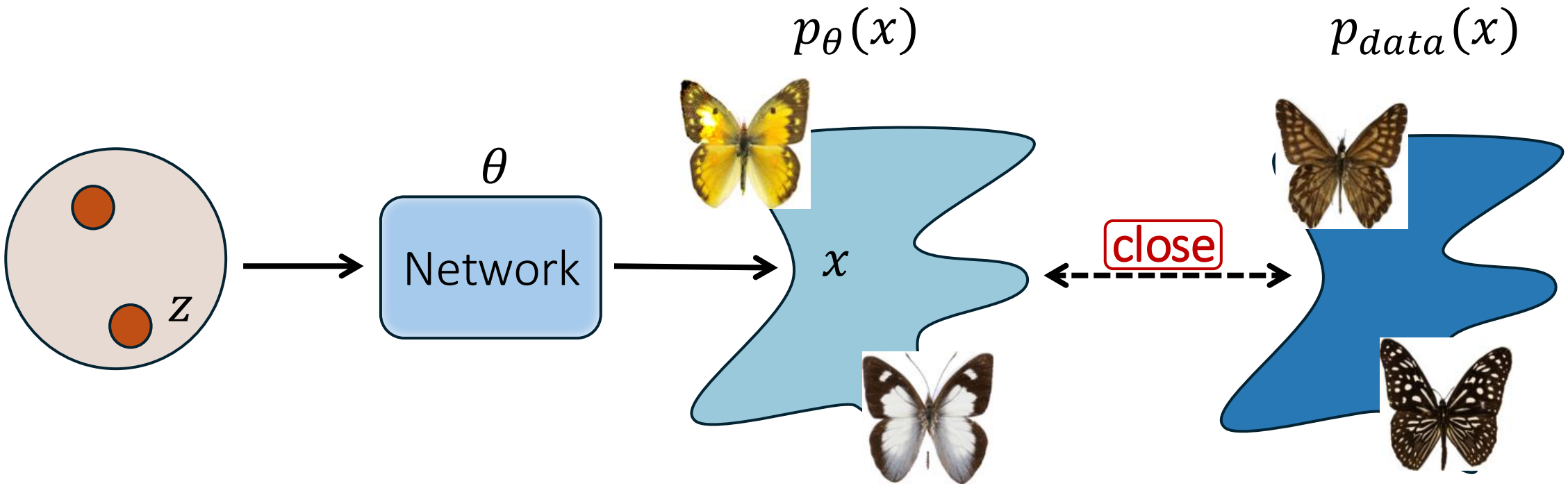


Common Objective in Generative Models



$$\theta^* = arg \min_\theta KL(p_{data}(x)||p_\theta(x))$$

Common Objective in Generative Models



$$\theta^* = arg \min_\theta KL(p_{data}(x)||p_\theta(x))$$

KL-Divergence and Maximum Likelihood Estimation

$$\theta^* = arg\min_\theta KL(p_{data}(x)||p_\theta(x)) = arg\min_\theta \mathbb{E}_{x\sim p_{data}}\left[\log\frac{p_{data}(x)}{p_\theta(x)}\right]$$

No θ here

$$= \boxed{arg\min_\theta \mathbb{E}_{x\sim p_{data}}[\log p_{data}(x)]} - arg\min_\theta \mathbb{E}_{x\sim p_{data}}[\log p_\theta(x)]$$

$$= arg\max_\theta \mathbb{E}_{x\sim p_{data}}[\log p_\theta(x)]$$

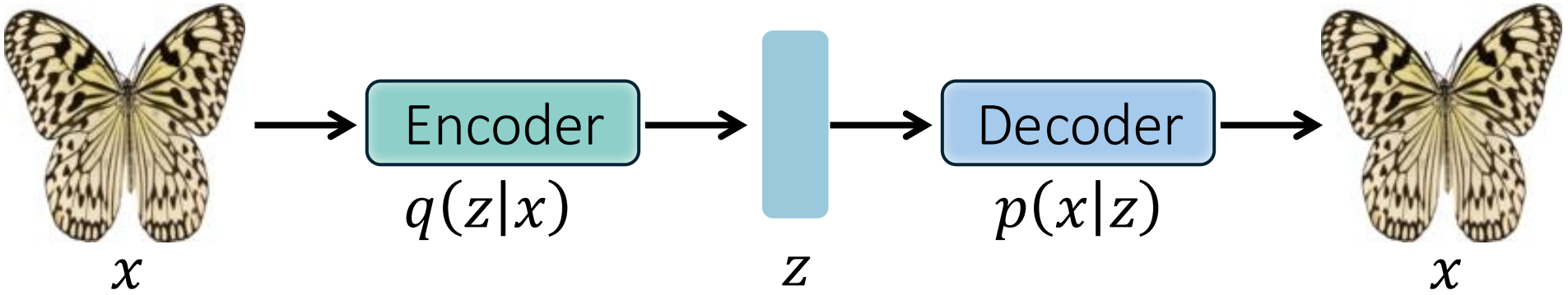
Suppose we sample $\{x^1, x^2, \cdots, x^m\}$ from $p_{data}(x)$

$$\approx \boxed{arg\max_\theta \sum_{i=1}^m \log p_\theta(x^i)}$$ Maximum Likelihood Estimation

Compute $\log p(x)$ is impossible

$$\log p(x) = \log \int p(x|z)\, p(z) dz$$

Consider all possible z



Evidence Lower Bound of $\log p(x)$

$$\log p(x) = \log p(x) \int_z q(z|x)\, dz = \log \left(\frac{p(z,x)}{p(z|x)} \right) \int_z q(z|x)\, dz$$

Encoder, can be any distribution

$$= \int_z q(z|x) \log \left(\frac{p(z,x)}{p(z|x)} \right) dz = \int_z q(z|x) \left(\frac{p(z,x)}{q(z|x)} \frac{q(z|x)}{p(z|x)} \right) dz$$

$$= \int_z q(z|x) \left(\frac{p(z,x)}{q(z|x)} \right) dz + \int_z q(z|x) \left(\frac{q(z|x)}{p(z|x)} \right) dz$$

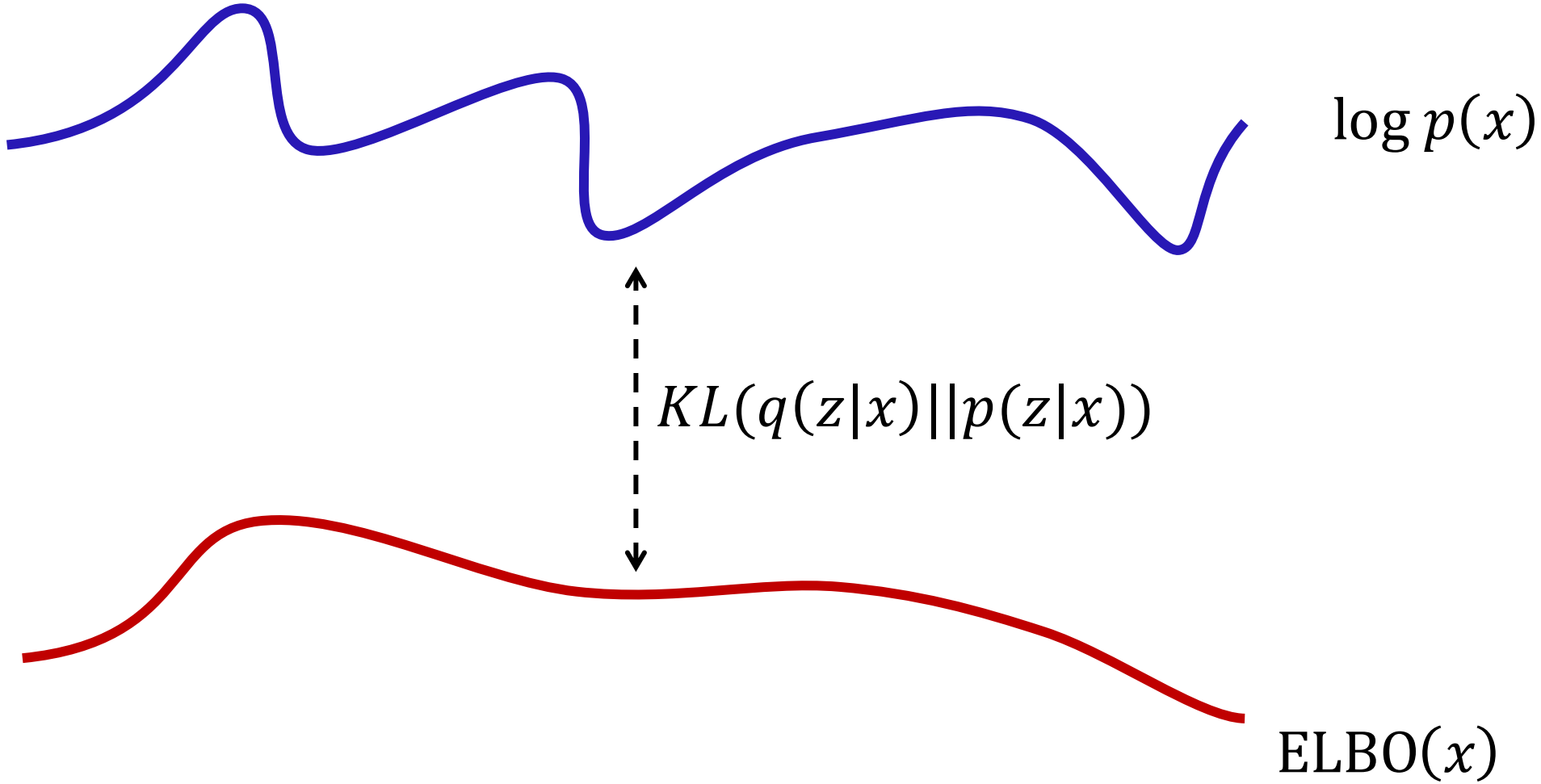
$$KL(q(z|x)||p(z|x)) \geq 0$$

$$\geq \int_z q(z|x) \left(\frac{p(z,x)}{q(z|x)} \right) dz = \mathbb{E}_{q(z|x)} \left[\log \frac{p(x,z)}{q(z|x)} \right]$$

Lower Bound of $\log p_\theta(x)$

ELBO

$\mathrm{ELBO}(x)$



Maximize $\mathrm{ELBO}(x)$

can help achieve the goal of maximizing $\log p(x)$

Further Derivation of $\mathbf{ELBO}(x)$

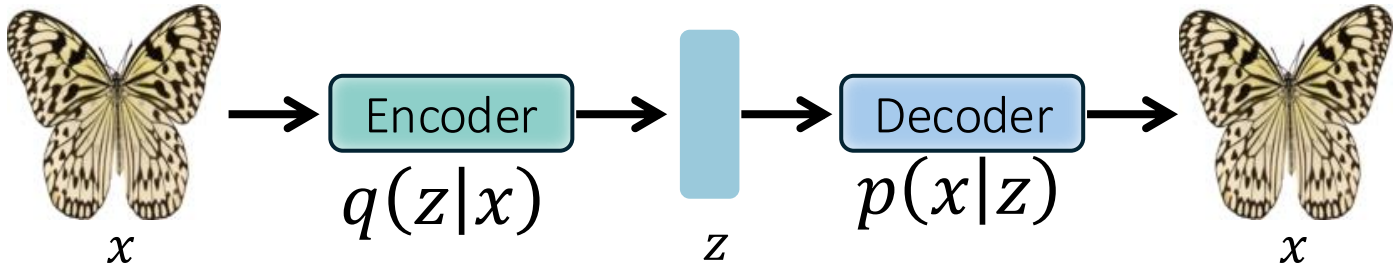
$$\mathbf{ELBO}(x) = \mathbb{E}_{q(z|x)}\left[\log \frac{p(x,z)}{q(z|x)}\right] = \mathbb{E}_{q(z|x)}\left[\log \frac{p(x|z)p(z)}{q(z|x)}\right]$$

$$= \mathbb{E}_{q(z|x)}[\log p(x|z)] + \mathbb{E}_{q(z|x)}\left[\log \frac{p(z)}{q(z|x)}\right]$$

$$= \mathbb{E}_{q(z|x)}[\log p(x|z)] - KL(q(z|x)||p(z))$$

How good the encoder is

how good the decoder is



Part 4: ELBO of DDPM

Optimization Goal of DDPM

VAE

$$\text{Maximize } p_\theta(x) \rightarrow \text{Maximize ELBO}(x): \mathbb{E}_{q(z|x)}\left[\log \frac{p(x,z)}{q(z|x)}\right],$$

latent variable z

Encoder

Diffusion

$$\text{Maximize } p_\theta(x) \rightarrow \text{Maximize ELBO}(x_0): \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{p(x_0, x_{1:T})}{q(x_{1:T}|x_0)}\right],$$

latent variable $x_{1:T}$

Diffusion (Forward) Process

Evidence Lower Bond of DDPM

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)\prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)\prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)}\right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}}\right]$$

$$\text{EBLO} = \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{p(x_{0:T})}{q(x_{1:T}|x_0)}\right]$$

Understanding Diffusion Models: A Unified Perspective
https://arxiv.org/pdf/2208.11970

Evidence Lower Bond of DDPM (Continued.)

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \cancel{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}}{\cancel{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}}} \right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}} + \log \frac{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \sum_{t=2}^{T} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \right] + \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \right]$$

$$= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \right] + \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \right]$$

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \right]}_{\text{denoising matching term}}$$

Also called “consistency term”

Optimization Goal

$$\text{Maximize}_\theta \text{ ELBO} =$$
$$\text{Maximize}_\theta \ \mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0|x_1)] - KL(q(x_T|x_0)||p(x_T))$$

$$- \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} \left[KL(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))\right]$$

Consistency Term

Maximize_θ ELBO

$\text{Maximize}_\theta \; \mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0|x_1)] - KL(q(x_T|x_0)||p(x_T))$

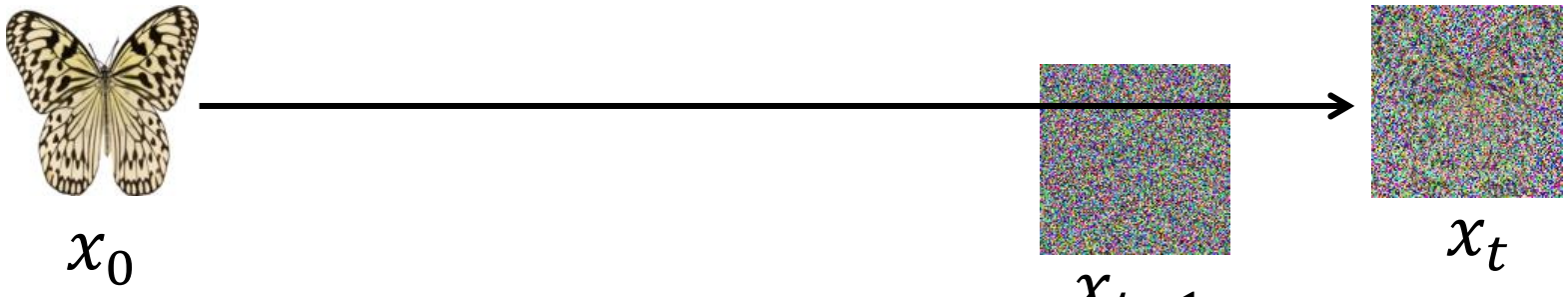
$$-\sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} \left[KL(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))\right]$$

True distribution of x_{t-1}

Predicted distribution of x_{t-1}

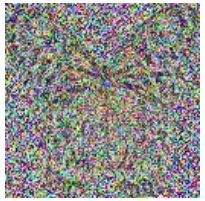
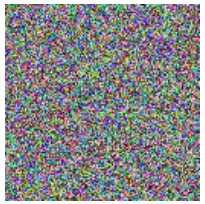
$$q(x_{t-1}|x_t, x_0)$$

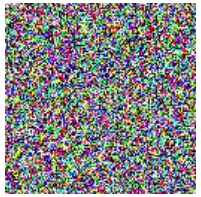
$\underline{q(x_{t-1}|x_t, x_0)}$

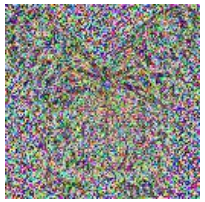
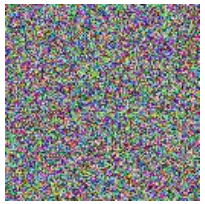


$$= \frac{q(x_{t-1}, x_t, x_0)}{q(x_t, x_0)} = \frac{q(x_t|x_{t-1})q(x_{t-1}|x_0)q(x_0)}{q(x_t|x_0)q(x_0)} = \boxed{\frac{q(x_t|x_{t-1})q(x_{t-1}|x_0)}{q(x_t|x_0)}}$$

$q(x_t|x_{t-1}), q(x_{t-1}|x_0), q(x_t|x_0)$:

Known, Gaussian $q(x_t|x_{t-1})$:  $= \sqrt{1-\beta_t}$  $+ \sqrt{\beta_t}$ 

Known, Gaussian $q(x_{t-1}|x_0)$:  $= \sqrt{\bar{\alpha}_{t-1}}$  $+ \sqrt{1-\bar{\alpha}_{t-1}}$ 

Known, Gaussian $q(x_t|x_0)$:  $= \sqrt{\bar{\alpha}_t}$  $+ \sqrt{1-\bar{\alpha}_t}$ 

Derivation of $q(x_{t-1}|x_t, x_0)$

$$
\begin{aligned}
q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) &= \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \\
&= \frac{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})\mathcal{N}(\boldsymbol{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0, (1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})} \\
&\propto \exp\left\{-\left[\frac{(\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{2(1-\alpha_t)} + \frac{(\boldsymbol{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_{t-1})} - \frac{(\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_t)}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{(\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{1-\alpha_t} + \frac{(\boldsymbol{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{1-\bar{\alpha}_t}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{(-2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1} + \alpha_t\boldsymbol{x}_{t-1}^2)}{1-\alpha_t} + \frac{(\boldsymbol{x}_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0)}{1-\bar{\alpha}_{t-1}} + C(\boldsymbol{x}_t, \boldsymbol{x}_0)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-\frac{2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1}}{1-\alpha_t} + \frac{\alpha_t\boldsymbol{x}_{t-1}^2}{1-\alpha_t} + \frac{\boldsymbol{x}_{t-1}^2}{1-\bar{\alpha}_{t-1}} - \frac{2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\bar{\alpha}_{t-1}})\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{\alpha_t(1-\bar{\alpha}_{t-1}) + 1 - \alpha_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\}
\end{aligned}
$$

https://arxiv.org/pdf/2208.11970

Derivation of $q(x_{t-1}|x_t, x_0)$ (Continued.)

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

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

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

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

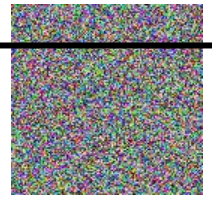
$$= \exp\left\{-\frac{1}{2}\left(\frac{1}{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}\right)\left[\boldsymbol{x}_{t-1}^2 - 2\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}\boldsymbol{x}_{t-1}\right]\right\}$$

$$\propto \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)})$$

$q(x_{t-1}|x_t, x_0)$: Gaussian

x_0

Gaussian



x_{t-1}

x_t

Mean

$$\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0 + \sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})x_t}{1-\bar{\alpha}_t}$$

$$\mu_q(x_t, x_0)$$

Variance

$$\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}$$

$$\Sigma_q(x_t)\mathbf{I} = \sigma_q^2(t)$$

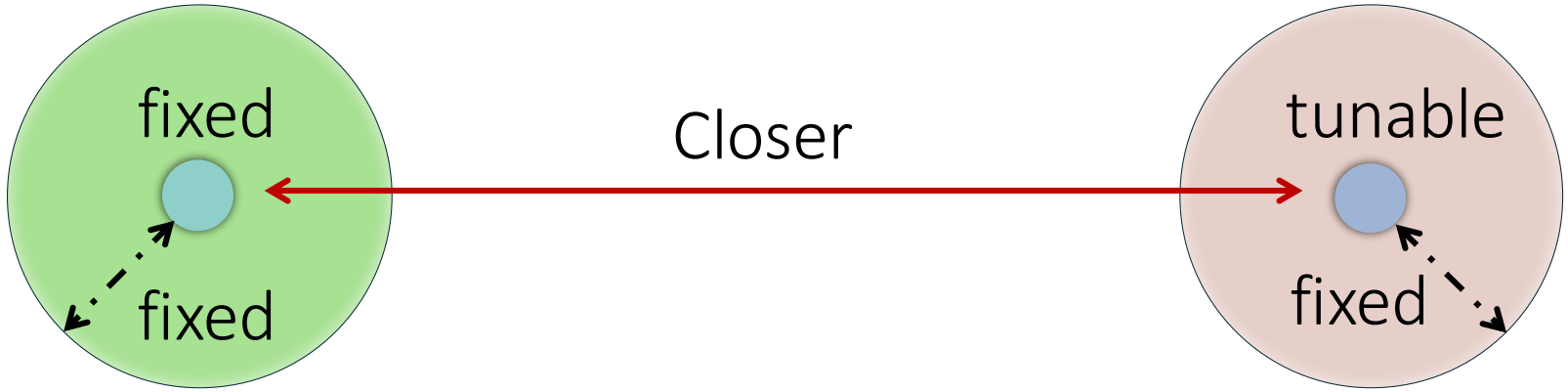
Optimization Goal in Consistency Term

A Gaussian

$$\mathbb{E}_{q(x_t|x_0)}[KL(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))]$$

How to minimize the KL divergence?

Another Gaussian, with variance: $\Sigma_q(x_t)\mathbf{I} = \sigma_q^2(t)\mathbf{I}$



Recall that the KL Divergence between two Gaussian distributions is:

$$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \parallel \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)) = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_y|}{|\boldsymbol{\Sigma}_x|} - d + \mathrm{tr}(\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_x) + (\boldsymbol{\mu}_y - \boldsymbol{\mu}_x)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x)\right]$$

Optimization Goal in Consistency Term (Continued)

$$\mu_q(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0 + \sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})x_t}{1-\bar{\alpha}_t}$$

x_t

Denoise

θ

$\mu_\theta(x_t)$

Optimization Goal in Consistency Term

$$
\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \\
&= \arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q(t)) \parallel \mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\Sigma}_q(t))) \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_q(t)|}{|\boldsymbol{\Sigma}_q(t)|} - d + \mathrm{tr}(\boldsymbol{\Sigma}_q(t)^{-1}\boldsymbol{\Sigma}_q(t)) + (\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)\right] \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\left[\log 1 - d + d + (\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)\right] \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\left[(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)\right] \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\left[(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)^T\left(\sigma_q^2(t)\mathbf{I}\right)^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)\right] \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\left[\|\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q\|_2^2\right]
\end{aligned}
$$

$$\arg\min_{\theta} \frac{1}{2\sigma_q^2(t)}\left[||\mu_\theta(x_t) - \mu_q(x_t, x_0)||_2^2\right]$$

Loss Function

https://arxiv.org/pdf/2403.18103

$\mu_\theta(x_t)$

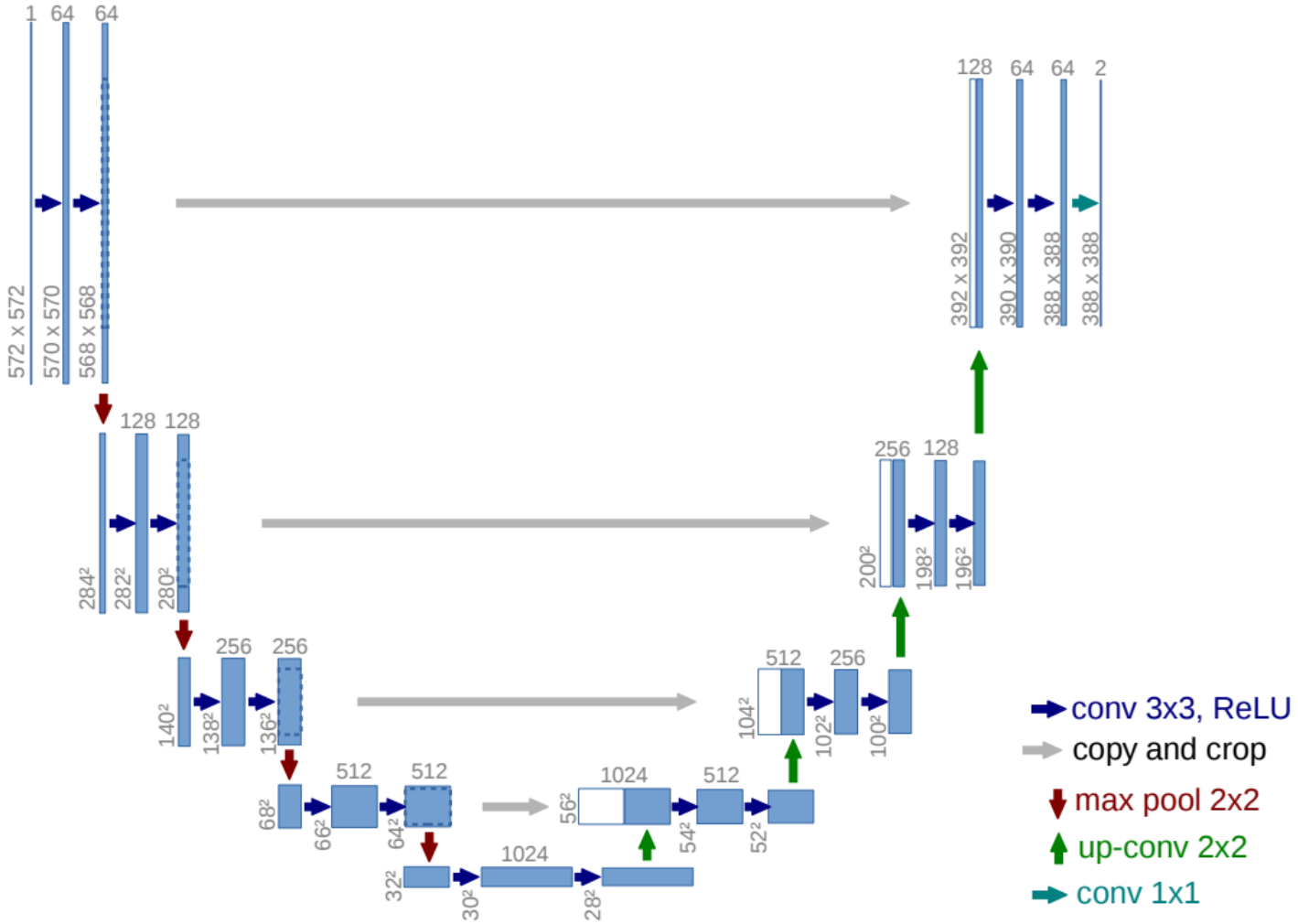
$$\mu_\theta(x_t) = \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)_t}{1-\bar{\alpha}_t} \hat{x}_\theta(x_t)$$

A neural Network,
i.e. U-Net

Align with

$$\mu_q(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} x_0$$

U–Net



U-Net: Convolutional Networks for Biomedical Image Segmentation
https://arxiv.org/pdf/1505.04597

$$p_\theta(x_{t-1}|x_t)$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu_\theta(x_t), \sigma_q^2(t)\mathbf{I})$$

$$\rightarrow x_{t-1} = \mu_\theta(x_t) + \sigma_q^2(t)z, \quad \text{where } z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$= \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)_t}{1-\bar{\alpha}_t} \hat{x}_\theta(x_t) + \sigma_q^2(t)z$$

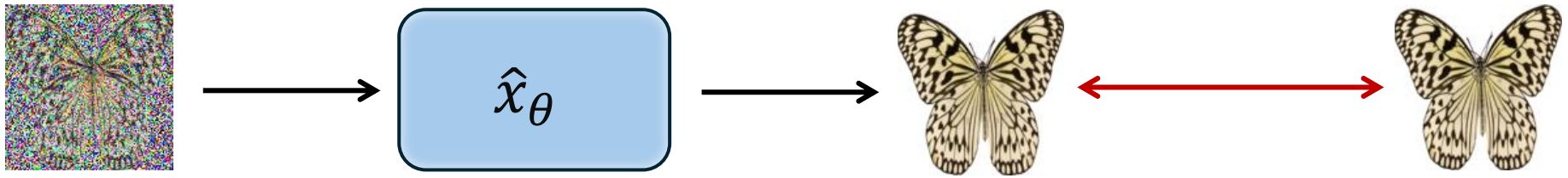
Will explain in the inference part

Rewrite Loss Function in Consistency Term

$$\frac{1}{2\sigma_q^2(t)}[||\mu_\theta(x_t) - \mu_q(x_t, x_0)||_2^2]$$

$$= \frac{1}{2\sigma_q^2(t)}[||\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{x}_\theta(x_t) - \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\ x_0||_2^2]$$

$$= \frac{1}{2\sigma_q^2(t)}\frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2}[||\hat{x}_\theta(x_t) - \ x_0||_2^2]$$



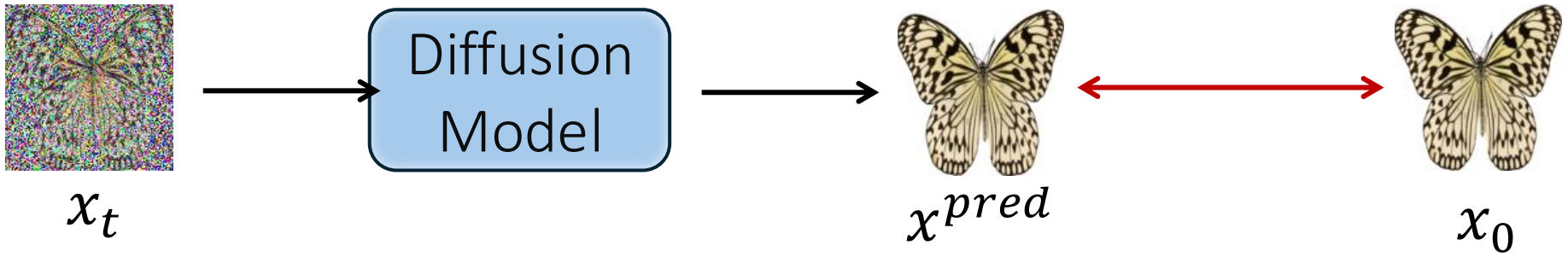
in consistency: $t \geq 2$

Rewrite Loss Function in Consistency Term

$$\frac{1}{2\sigma_q^2(t)}[||\mu_\theta(x_t) - \mu_q(x_t, x_0)||_2^2]$$

$$= \frac{1}{2\sigma_q^2(t)}[||\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{x}_\theta(x_t) - \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\ x_0||_2^2]$$

$$= \frac{1}{2\sigma_q^2(t)}\frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2}[||\hat{x}_\theta(x_t) - \ x_0||_2^2]$$



in consistency: $t \geq 2$

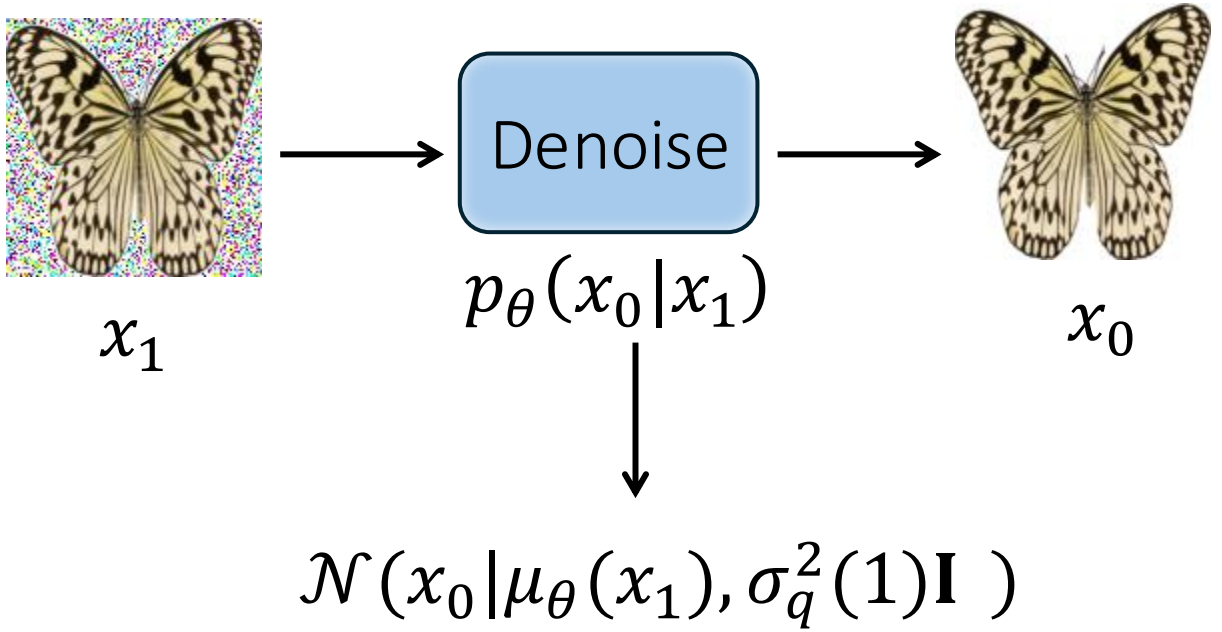
The Reconstruction Term

$$\text{Maximize}_\theta \text{ ELBO} =$$
$$\text{Maximize}_\theta \mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0|x_1)] - KL(q(x_T|x_0)||p(x_T))$$
$$- \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} [KL(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))]$$

Reconstruction Term

$p_\theta(x_0|x_1)$: Recover x_0 from x_1

Need: $\mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0|x_1)]$



Reconstruction Term Derivation

$$\log p_\theta(x_0|x_1) = \log \mathcal{N}\left(x_0 \middle| \mu_\theta(x_1), \sigma_q^2(1)\mathbf{I}\right) = \log \frac{1}{\sqrt{2\pi\sigma_q^2(1)}} e^{-\frac{(x_0-\mu_\theta(x_1))^2}{2\sigma_q^2(1)}}$$

$$= -\log\sqrt{2\pi\sigma_q^2(1)} - \frac{(x_0-\mu_\theta(x_1))^2}{2\sigma_q^2(1)}$$

$$\propto -\frac{||\mu_\theta(x_1) - x_0||_2^2}{2\sigma_q^2(1)}$$

$$\bar{\alpha}_0 = 1$$
$$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i \,, \bar{\alpha}_1 = \alpha_1$$

$$\mu_\theta(x_1) = \frac{\sqrt{\bar{\alpha}_1}(1-\bar{\alpha}_0)}{1-\bar{\alpha}_1} x_1 + \frac{\sqrt{\bar{\alpha}_0}(1-\alpha_1)}{1-\bar{\alpha}_1} \hat{x}_\theta(x_1)$$

$$= \boxed{-\frac{1}{2\sigma_q^2(1)} ||\hat{x}_\theta(x_1) - x_0||_2^2}$$

Match with Consistency Term

$$\log p_\theta(x_0|x_1)$$

$$\bar{\alpha}_0 = 1$$
$$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i \,, \bar{\alpha}_1 = \alpha_1$$

$$= \frac{1}{2\sigma_q^2(1)} \frac{\bar{\alpha}_0(1-\alpha_1)^2}{(1-\bar{\alpha}_1)^2} ||\hat{x}_\theta(x_1) - x_0||_2^2 \text{ for } t = 1$$

(Reconstruction Term)

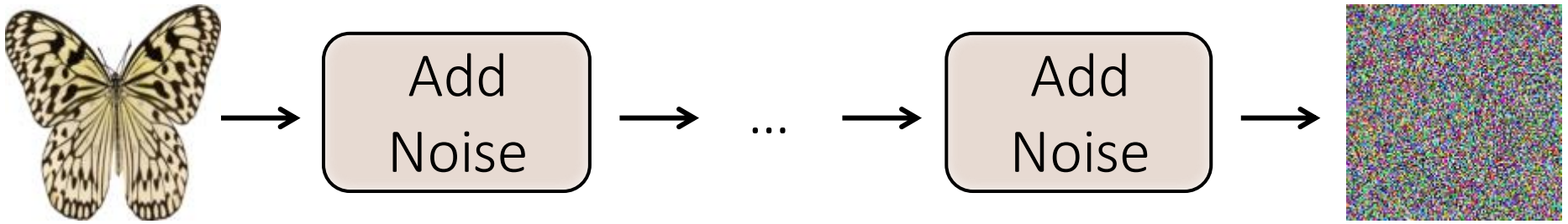
Recall: $\frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} ||\hat{x}_\theta(x_t) - x_0||_2^2 \text{ for } t \geq 2$

(Consistency Term)

Prior Matching Term

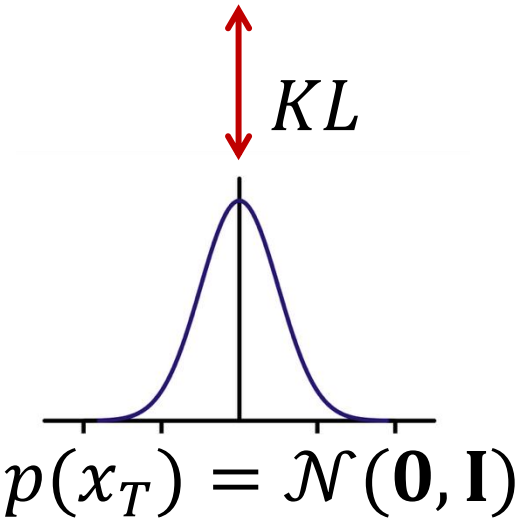
$$\mathbb{E}_{q(x_1|x_0)}[log p_\theta(x_0|x_1)] \boxed{- KL(q(x_T|x_0)||p(x_T))}$$

$$- \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} [KL(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))]$$



x_0

$q(x_T|x_0) \rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I})$



No $\boldsymbol{\theta}$ to train, measure how good the noise scheduler is

Rewrite the ELBO

$$\frac{1}{2\sigma_q^2(1)} \frac{\bar{\alpha}_0(1-\alpha_1)^2}{(1-\bar{\alpha}_1)^2} ||\hat{x}_\theta(x_1) - x_0||_2^2$$

$$\text{Maxmize}_\theta \text{ ELBO} =$$

$$\text{Maxmize}_\theta \mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0|x_1)] - \cancel{KL(q(x_T|x_0)||p(x_T))}$$

$$-\sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} [KL(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))]$$

$$\frac{1}{\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} ||\hat{x}_\theta(x_t) - x_0||_2^2$$

ELBO and Loss Function

$$\text{ELBO} = -\sum_{t=1}^{T} \mathbb{E}_{q(x_t|x_0)} \left[\frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} ||\hat{x}_\theta(x_t) - x_0||_2^2\right]$$

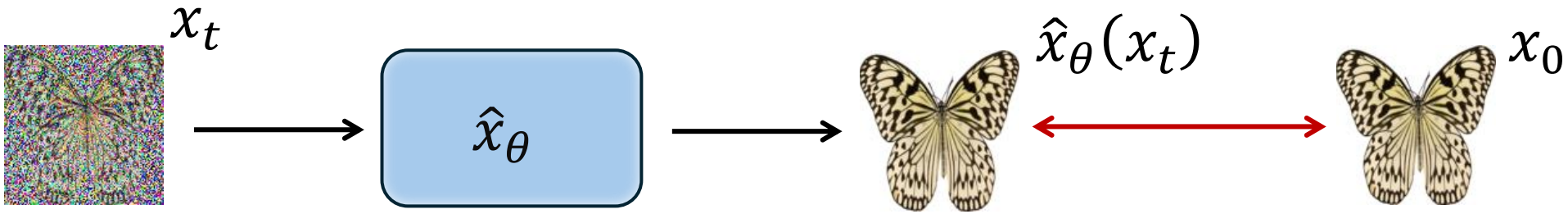
Loss Function:

$$\theta^* = \text{argmin}_\theta \sum_{t=1}^{T} \frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} \mathbb{E}_{q(x_t|x_0)} [||\hat{x}_\theta(x_t) - x_0||_2^2]$$

For a particular x_t:

Image Prediction

$$\text{argmin}_\theta ||\hat{x}_\theta(x_t) - x_0||_2^2$$



Part 5: Training and Inference

Training in Image Prediction

Repeat:

Pick x_0 from training dataset

x_0

Pick t from $\mathbf{Uniform}[1, T]$

Draw a sample $x_t \sim \mathcal{N}(x_t | \sqrt{\bar{\alpha}_t} x_0, \sqrt{1 - \bar{\alpha}_t}\mathbf{I})$

x_t

Take gradient descent step on:

$$\nabla_\theta ||\hat{x}_\theta(x_t) - x_0||_2^2$$

Training in Image Prediction (Batch Version)

Repeat:

batch size

Pick $x_0^{1:N}(x_0^1, x_0^2, \cdots, x_0^N)$ from training dataset

Pick $t^{1:N}(t^1, t^2, \cdots, t^N)$ with $t^n \sim \mathbf{Uniform}[1, T]$ for n in $1:T$

Draw $x^{1:N}$, for n^{th} samples:

$$x^n \sim \mathcal{N}(\sqrt{\bar{\alpha}_{t^n}} x_0^n, \sqrt{1 - \bar{\alpha}_{t^n}}\mathbf{I})$$

Take gradient descent step on:

$$\nabla_\theta ||\hat{x}_\theta(x^{1:N}) - x_0^{1:N}||_2^2$$

Recall: $p_\theta(x_{t-1}|x_t)$

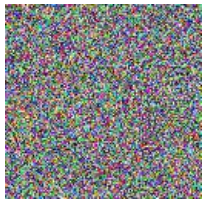
$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu_\theta(x_t), \sigma_q^2(t)\mathbf{I})$$

$$\rightarrow x_{t-1} = \mu_\theta(x_t) + \sigma_q^2(t)z, \qquad \text{where } z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$= \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)t}{1-\bar{\alpha}_t} \hat{x}_\theta(x_t) + \sigma_q^2(t)z$$

Inference in Image Prediction

Sample a Noise $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$



x_T

Repeat from $t = T$ to $t = 1$:

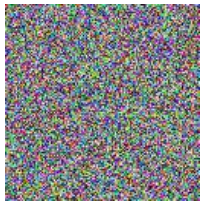
Update according to $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$

Generation Diversity

$$x_{t-1} = \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)_t}{1-\bar{\alpha}_t} \hat{x}_\theta(x_t) + \sigma_q^2(t)z$$

x_t

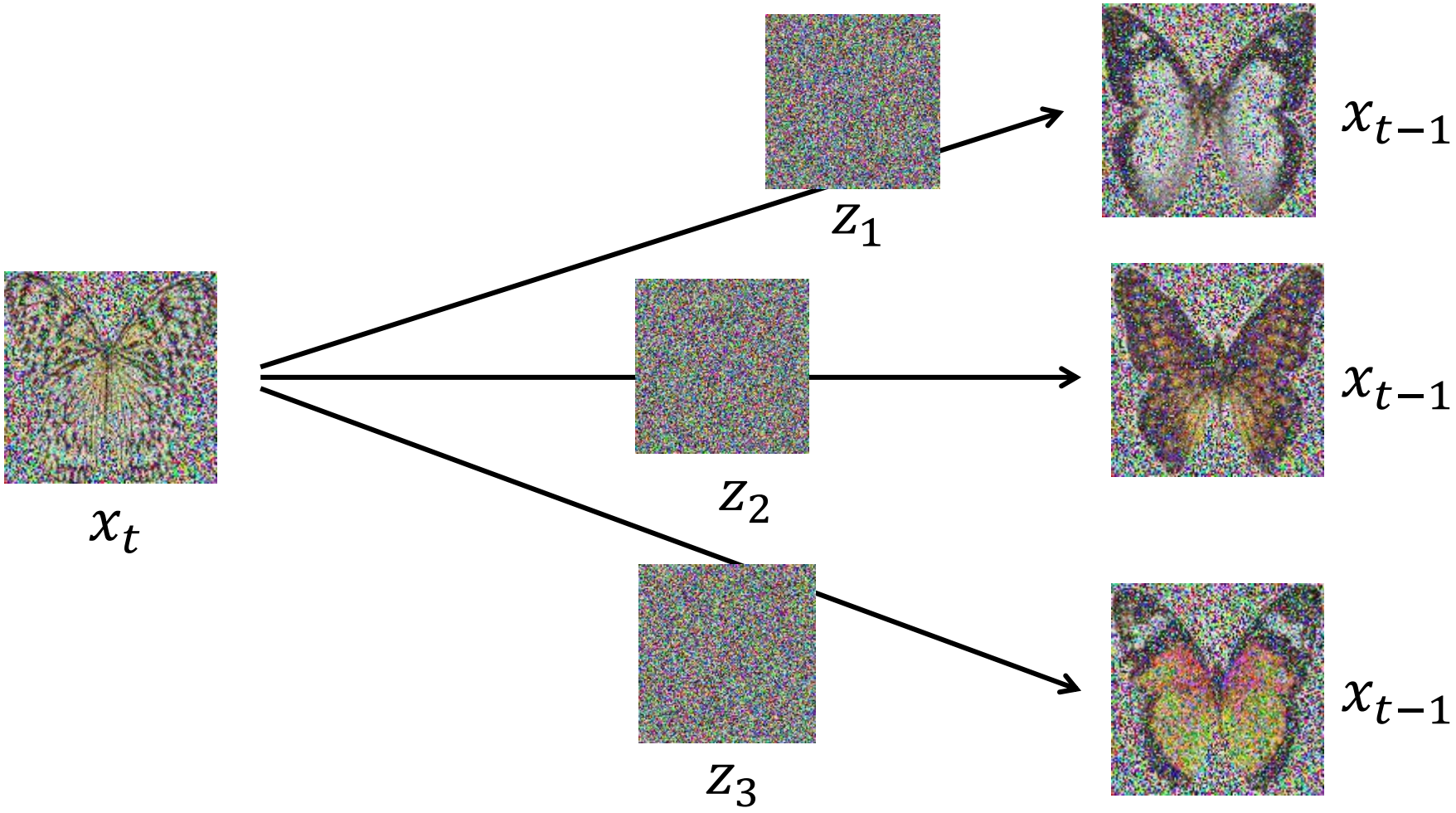
x_t



z

Add Noise in Generation

Non-deterministic Generation:



$$z_1, z_2, z_3 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Inference in Image Prediction (Batch Version)

Sample a Noise $x_T^{1:N} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

Repeat from $t = T$ to $t = 1$:

Sample $z^{1:N}$ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$

Update $x_{t-1}^{1:N}$, with each $x_{t-1}^{n} \sim p_\theta(x_{t-1}^{n} | x_t^{n})$

$$x_{t-1}^{1:N} = \frac{\sqrt{\overline{\alpha}_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t} x_t^{1:N} + \frac{\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)t}{1-\overline{\alpha}_t} \hat{x}_\theta(x_t^{1:N}) + \sigma_q^2(t) z^{1:N}$$

Noise Prediction

Forward Process:

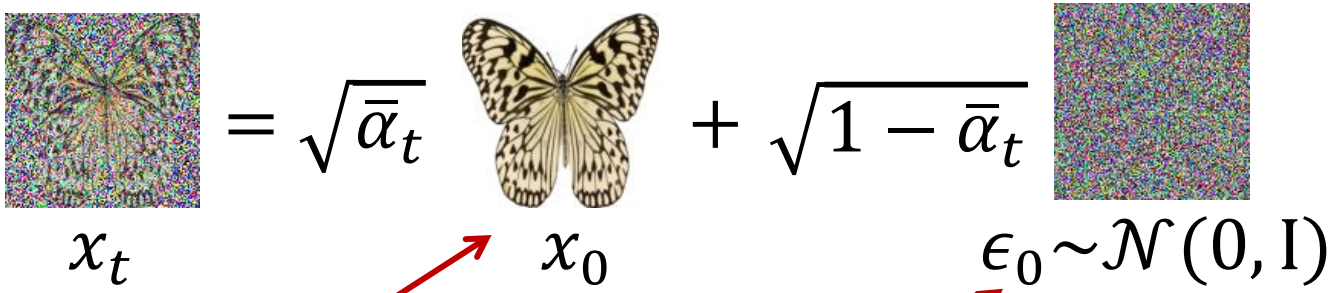
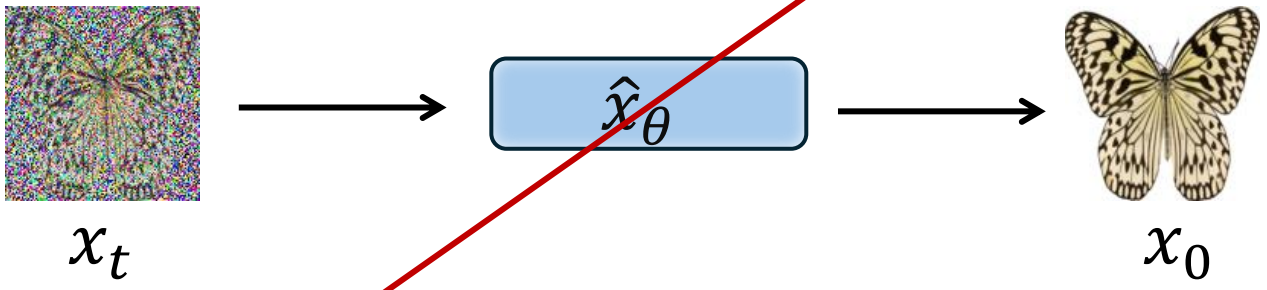
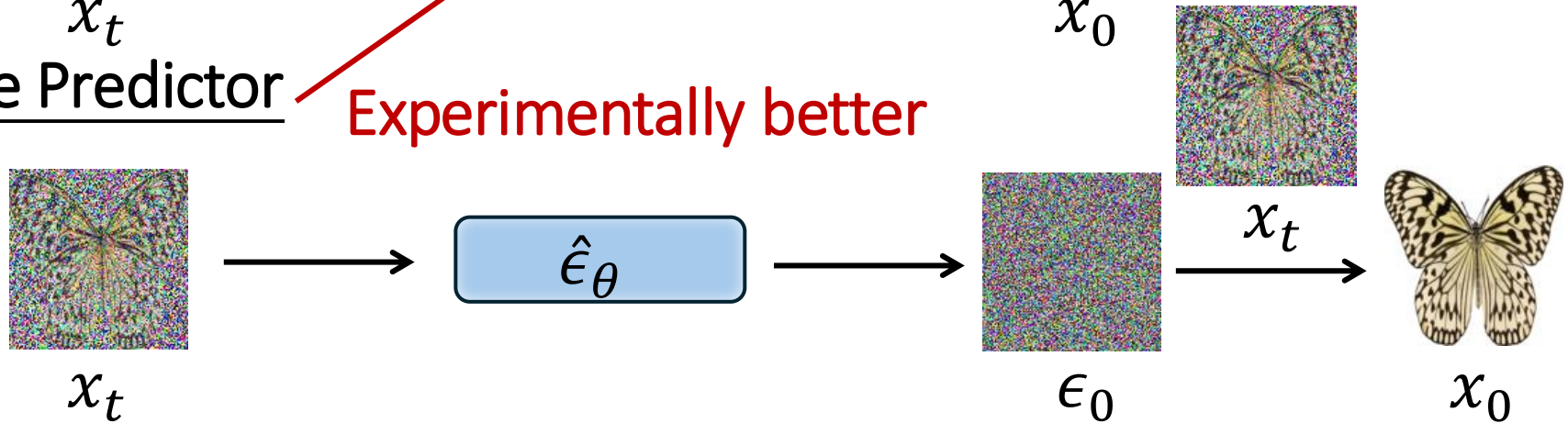


Image Predictor



Noise Predictor

Experimentally better

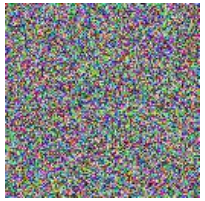


Optimization Goal Noise Prediction

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_0$$

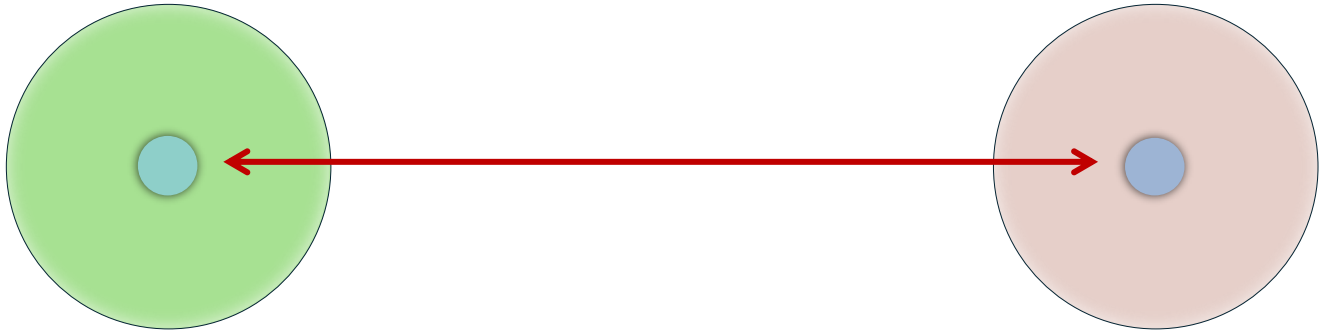
$$\downarrow$$

$$x_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_0}{\sqrt{\bar{\alpha}_t}}$$



$$\epsilon_0 \sim \mathcal{N}(0, \mathrm{I})$$

$$\mu_q(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t) x_0 + \sqrt{\bar{\alpha}_t}(1 - \bar{\alpha}_{t-1}) x_t}{1 - \bar{\alpha}_t}$$



$$\mu_\theta(x_t)$$

Optimization Goal Noise Prediction (Continued)

$$\mu_q(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_0}{\sqrt{\bar{\alpha}_t}} + \sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})x_t}{1-\bar{\alpha}_t}$$

$$= \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon_0\right)$$

$$\mu_\theta(x_t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}_\theta(x_t)\right)$$

Image Predictor

$$\mu_q(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0$$

$$\mu_\theta(x_t) = \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\hat{x}_\theta(x_t)$$

Rewrite ELBO

Image Predictor

$$\text{ELBO} = -\sum_{t=1}^{T} \mathbb{E}_{q(x_t|x_0)} \left[\frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} \boxed{||\hat{x}_\theta(x_t) - x_0||_2^2} \right]$$

Noise Predictor

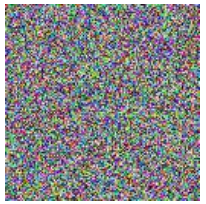
$$\text{ELBO} = -\sum_{t=1}^{T} \mathbb{E}_{q(x_t|x_0)} \left[\frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} \boxed{||\hat{\epsilon}_\theta(x_t) - \epsilon_0||_2^2} \right]$$

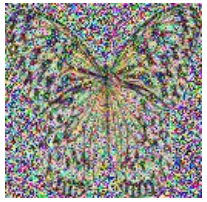
Training in Noise Prediction

Repeat:

Pick x_0 from training dataset 

Pick t from $\mathbf{Uniform}[1, T]$

Draw ϵ_0 noise from $\mathcal{N}(0, \mathrm{I})$ 

Drive $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_0$ 

x_t

Take gradient descent step on:

$$\nabla_\theta ||\hat{\epsilon}_\theta(x_t) - \epsilon_0||_2^2$$

Training in Noise Prediction (Batch Version)

Repeat:

Pick $x_0^{1:N}$ from training dataset

Pick $t^{1:N}$, with $t^n \sim \text{Uniform}[1, T]$

Draw $\epsilon_0^{1:N}$, with $\epsilon_0^n \sim \mathcal{N}(0, \mathrm{I})$

Drive $x^{1:N}$, for n^{th} samples:

$$x^n = \sqrt{\bar{\alpha}_{t^n}} x_0^n + \sqrt{1 - \bar{\alpha}_{t^n}} \epsilon_0^n$$

Take gradient descent step on:

$$\nabla_\theta || \hat{\epsilon}_\theta(x^{1:N}) - \epsilon_0^{1:N} ||_2^2$$

Rewrite $p_\theta(x_{t-1}|x_t)$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu_\theta(x_t), \sigma_q^2(t)\mathbf{I})$$

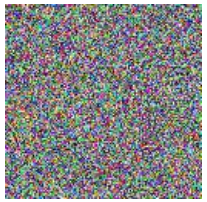
$$\mu_\theta(x_t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}_\theta(x_t)\right)$$

$$\rightarrow x_{t-1} = \mu_\theta(x_t) + \sigma_q^2(t)z, \quad \text{where } z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$= \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\epsilon}_\theta(x_t)\right) + \sigma_q^2(t)z$$

Inference in Noise Prediction

Sample a Noise $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$



x_T

Repeat from $t = T$ to $t = 1$:

Draw z from $\mathcal{N}(0, \mathrm{I})$

Update according to $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$

$$x_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\epsilon}_\theta(x_t)\right) + \sigma_q^2(t)z$$

Inference in Noise Prediction (Batch Version)

Sample n noise $x_T^{1:N}$, with $x_T^n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

Repeat from $t = T$ to $t = 1$:

Draw $z^{1:N}$, with $z^n \sim \mathcal{N}(0, \mathrm{I})$

Update $x_{t-1}^{1:N}$:

$$x_{t-1}^{1:N} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t^{1:N} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\epsilon}_\theta\left(x_t^{1:N}\right)\right) + \sigma_q^2(t) z^{1:N}$$