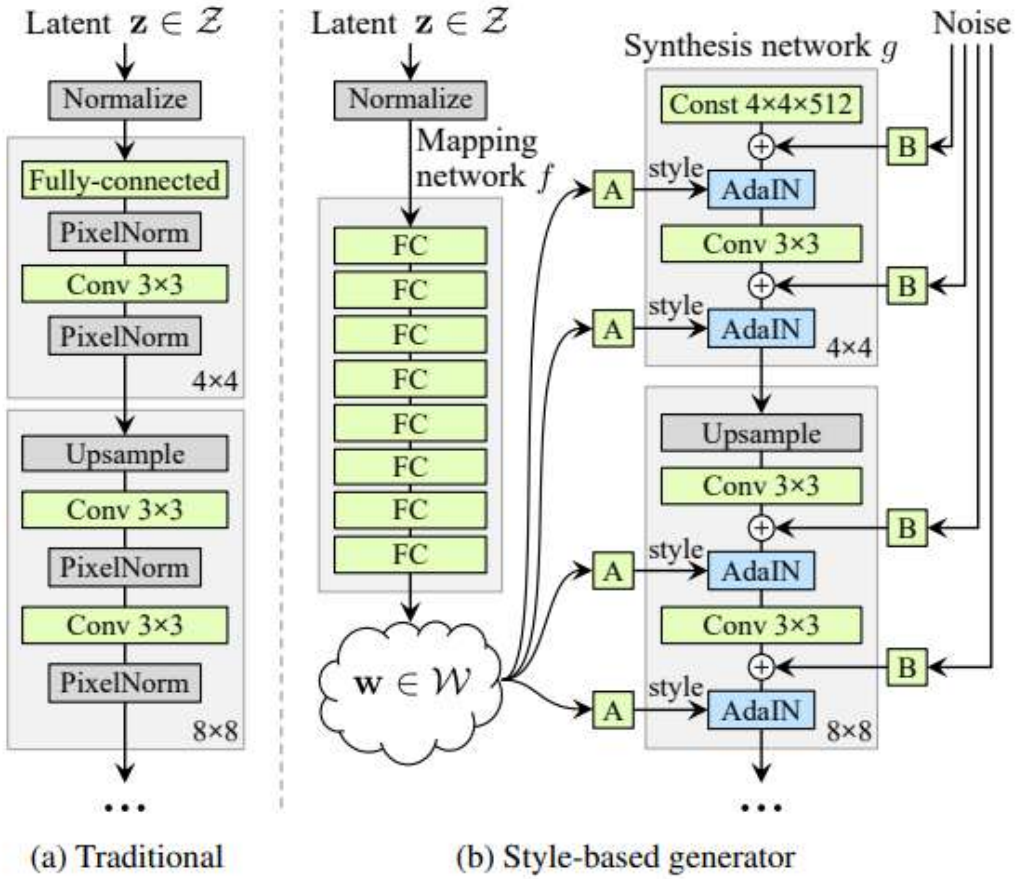
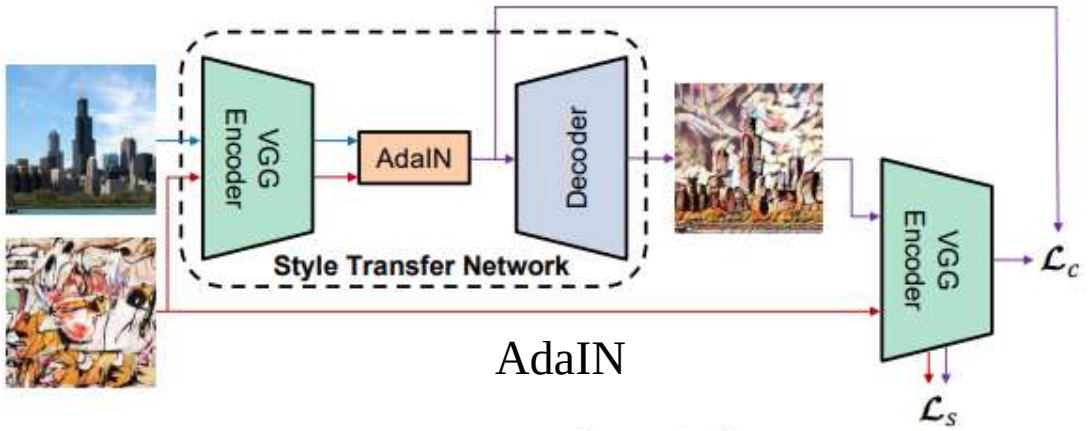
Recap.

STYLEGAN: A Style-Based Generator Architecture for Generative Adversarial Networks

- SoTA GAN before tokenizers.



(a) Traditional (b) Style-based generator



AdaIN

$$\text{AdaIN}(x, y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y)$$

(a) Distribution of features in training set (b) Mapping from $\mathcal{Z}$ to features (c) Mapping from $\mathcal{W}$ to features

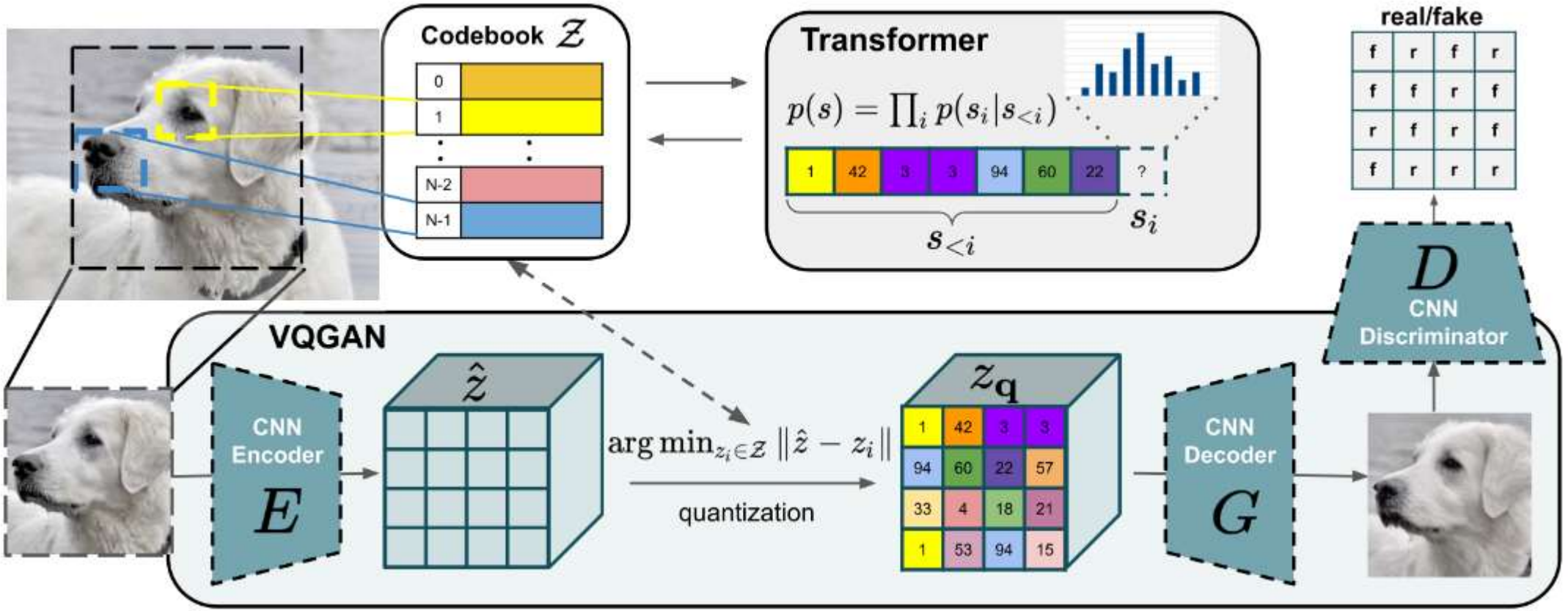
https://arxiv.org/pdf/1812.04948

STYLEGANv3: Alias-Free Generative Adversarial Networks

- Excellent interpolation.

https://arxiv.org/pdf/2106.12423

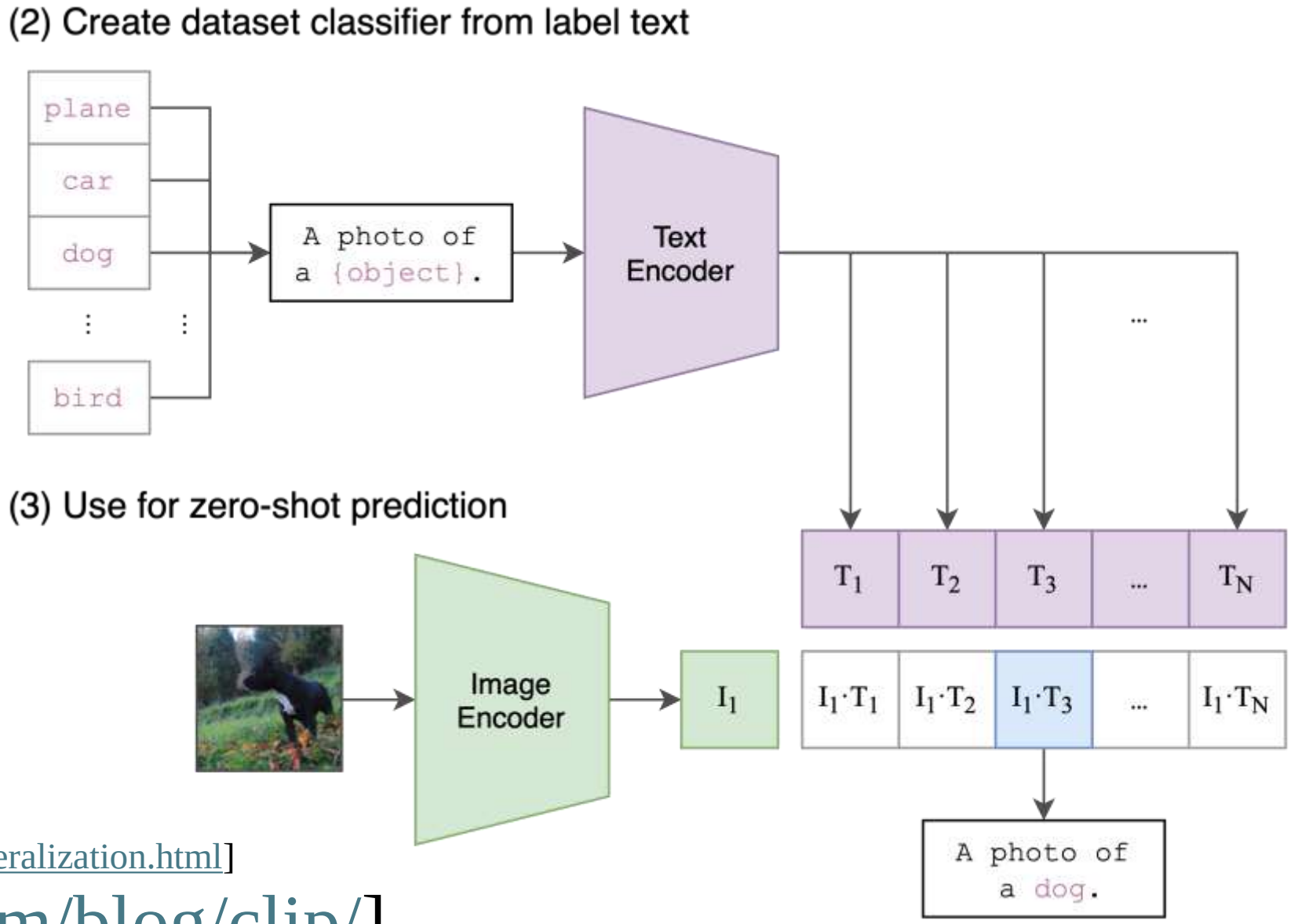
VQGAN



Esser, et al., "Taming Transformers for High-Resolution Image Synthesis", CVPR 2021

CLIP [Radford et al., 2021] https://arxiv.org/pdf/2103.00020.pdf

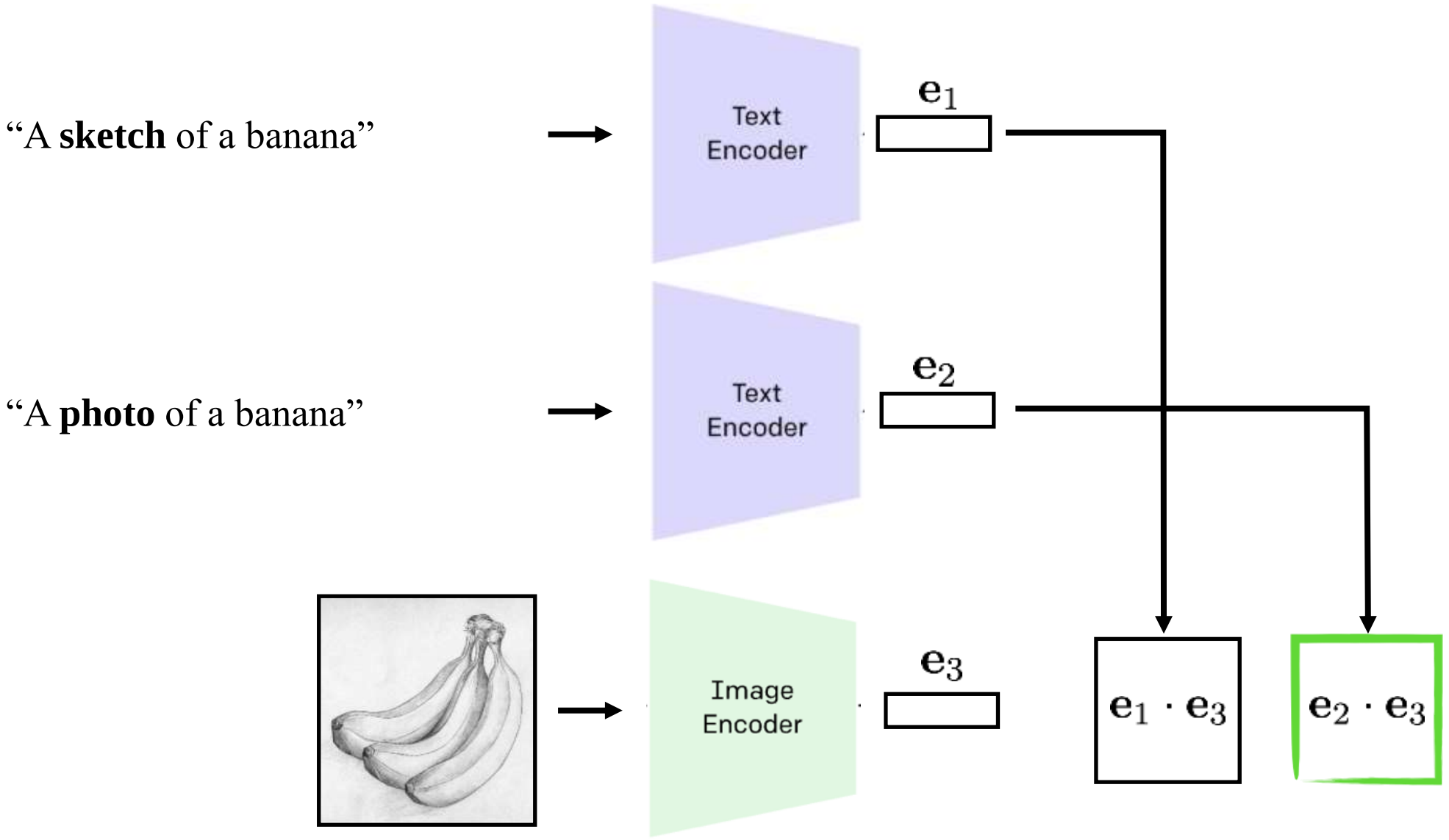
2. Adaptor: Just ask



[https://evjang.com/2021/10/23/generalization.html]

[https://openai.com/blog/clip/]

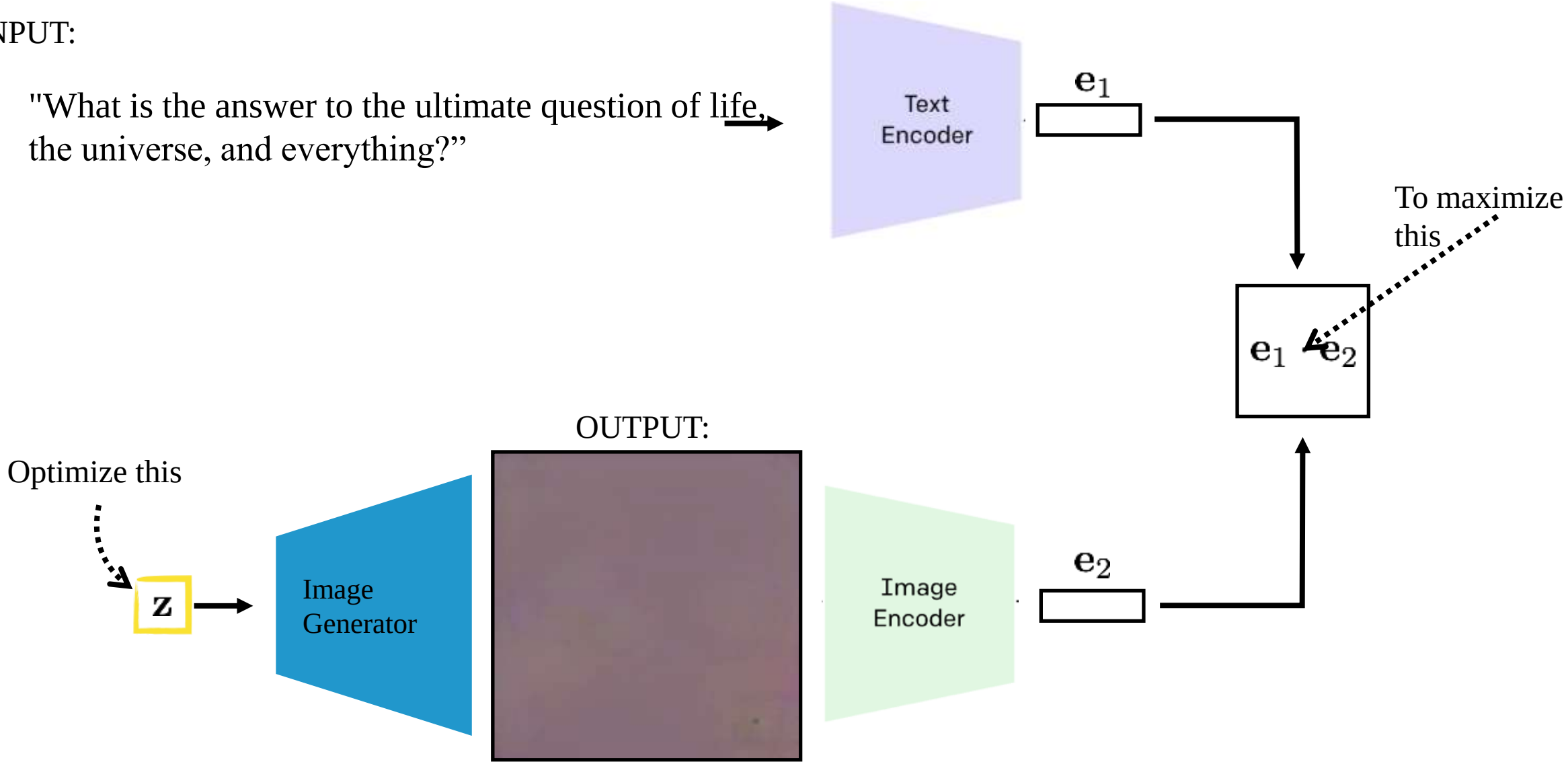
New capabilities by just asking



New capabilities by plugging pretrained models together: CLIP+GAN

INPUT:

"What is the answer to the ultimate question of life, the universe, and everything?"

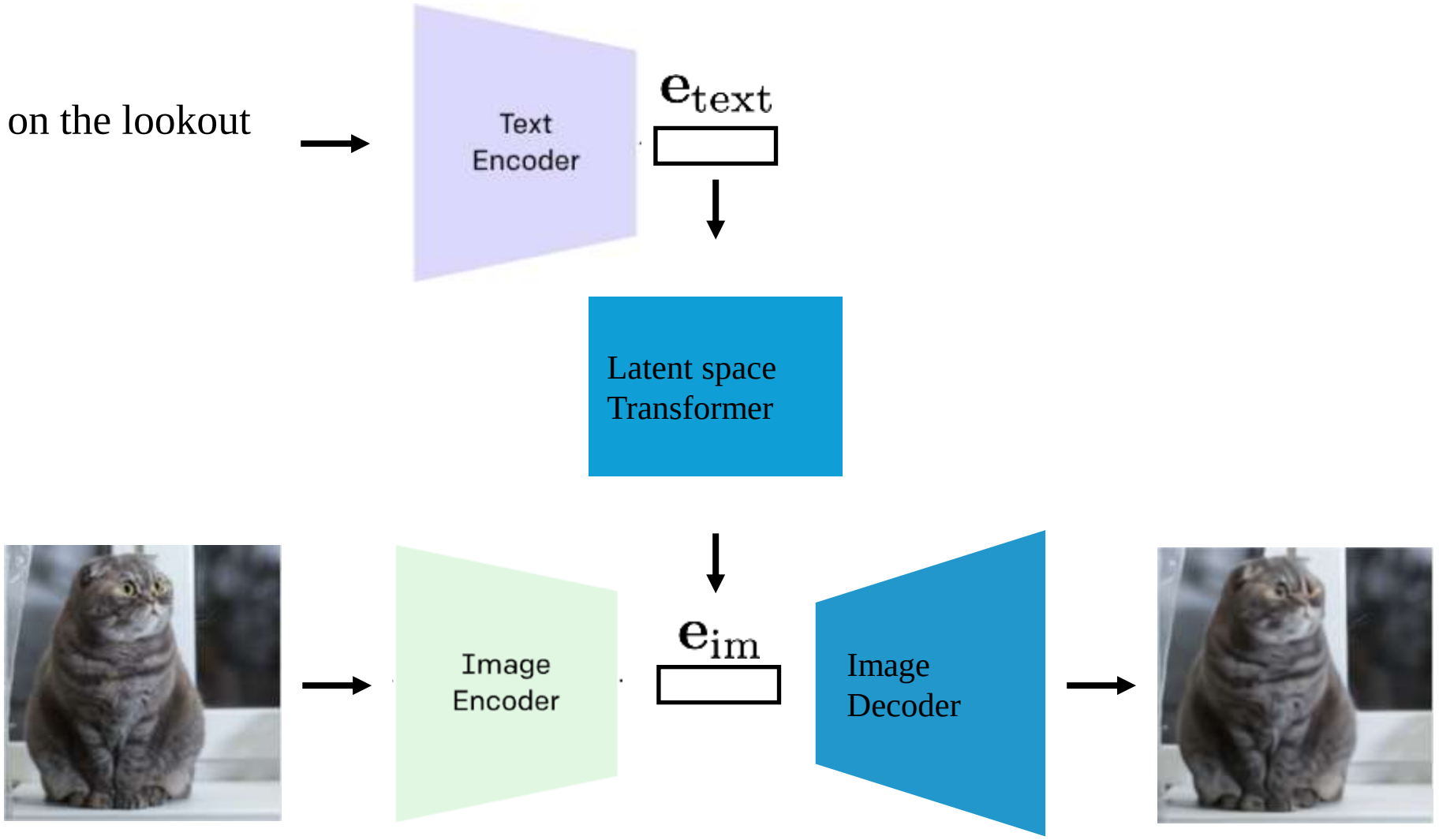


Code: https://colab.research.google.com/drive/1_4PQqzM_0KKytCzWtn-ZPi4cCa5bwK2F?usp=

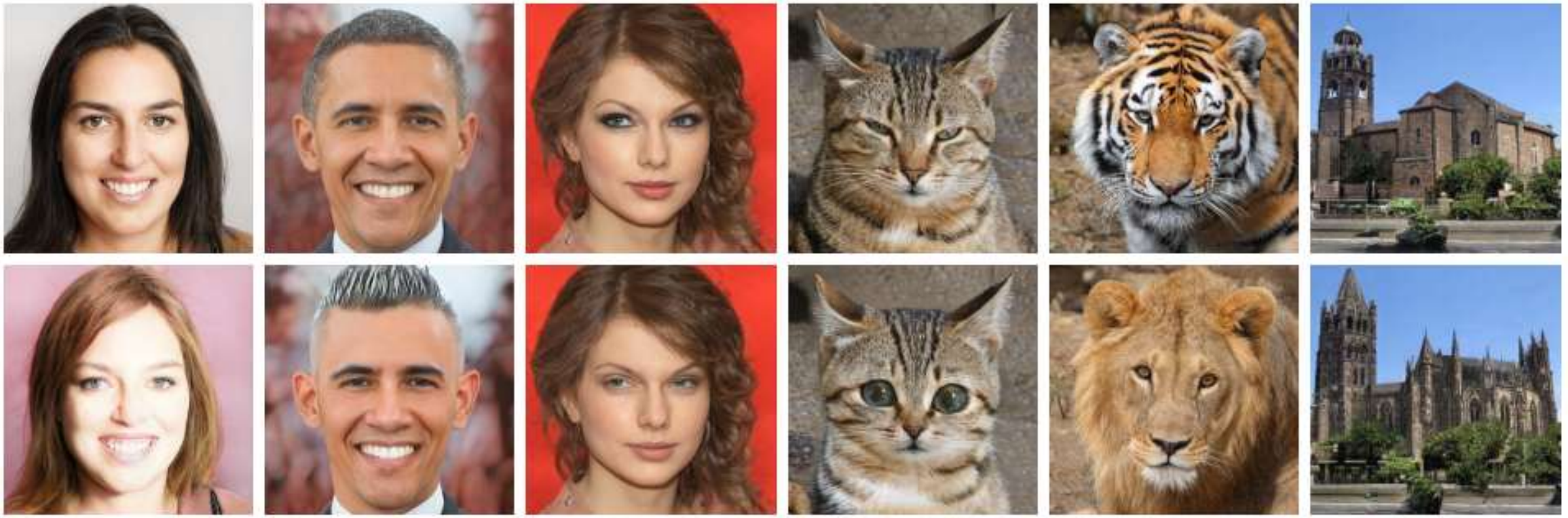
DALL-E [Ramesh et al. 2021] https://arxiv.org/pdf/2102.12092.pdf https://openai.com/blog/dall-e/

INPUT:

"A wide-eyed cat on the lookout for food"



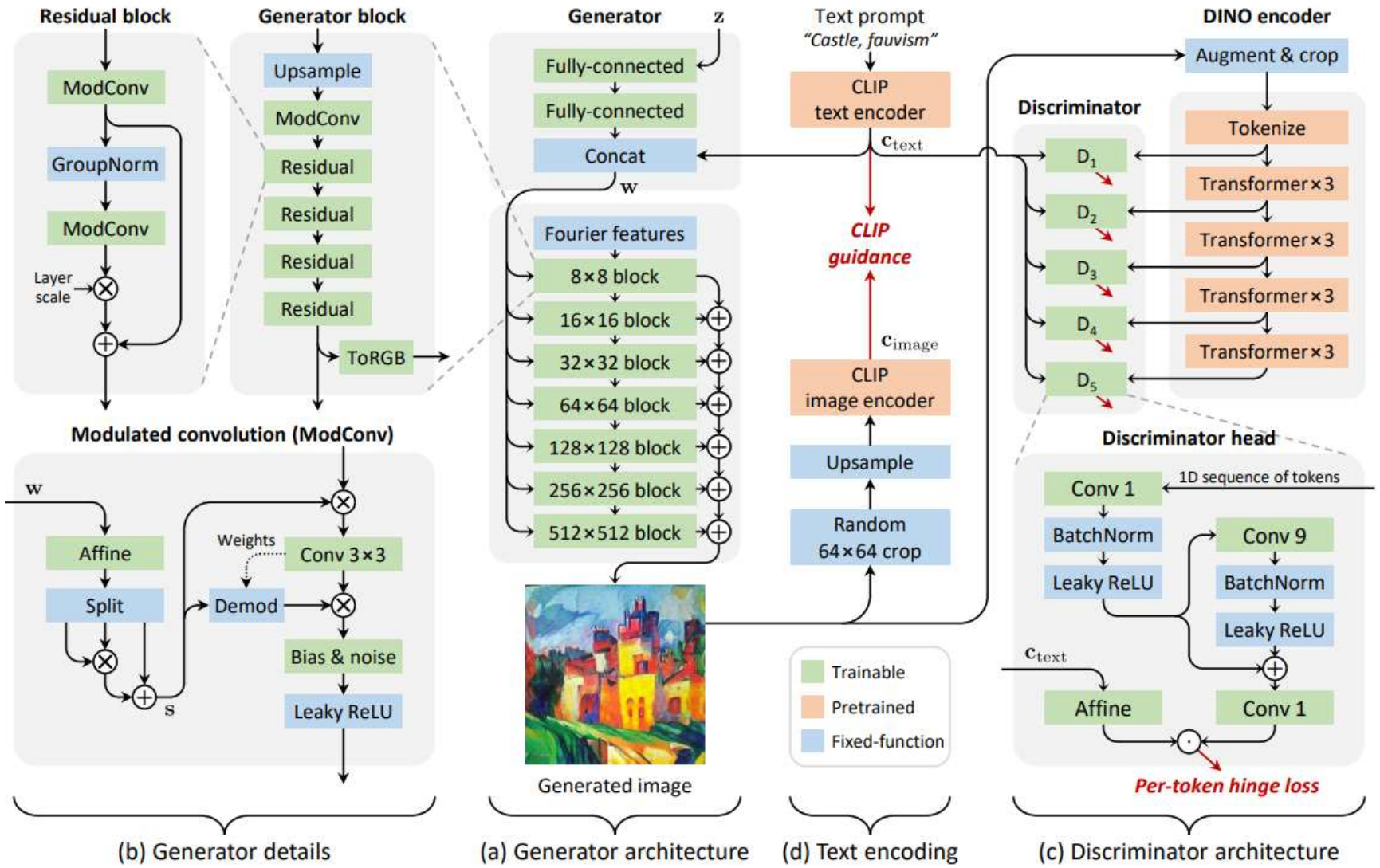
StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery



"Emma Stone" "Mohawk hairstyle" "Without makeup" "Cute cat" "Lion" "Gothic church"

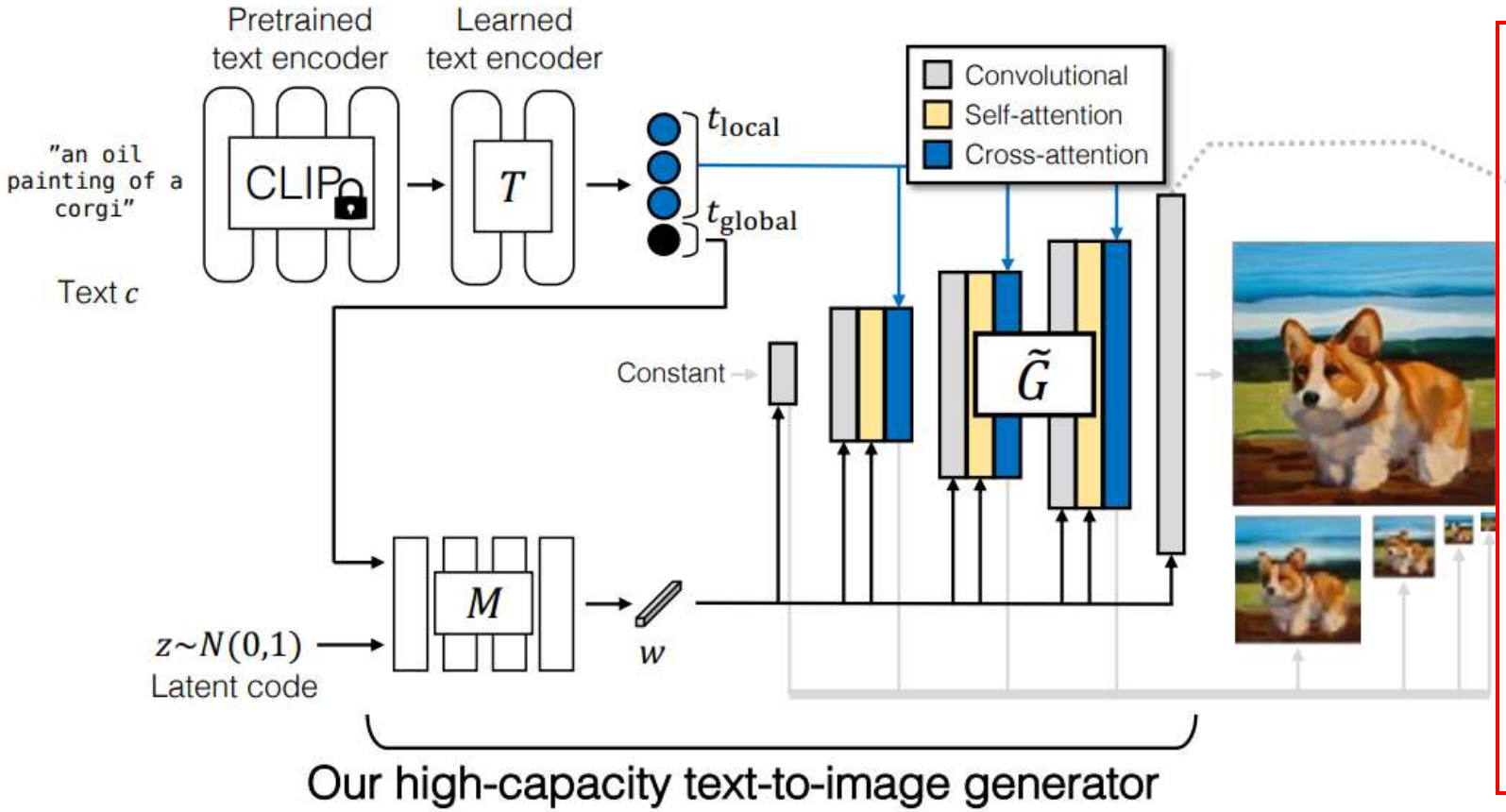
https://arxiv.org/pdf/2103.17249

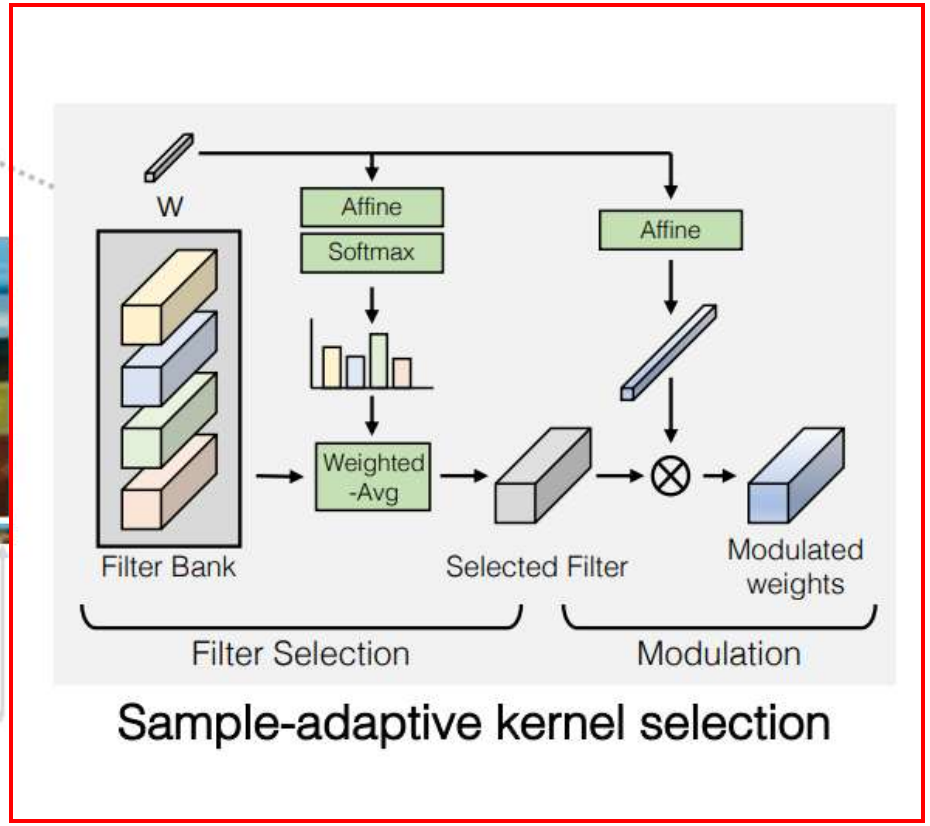
StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis



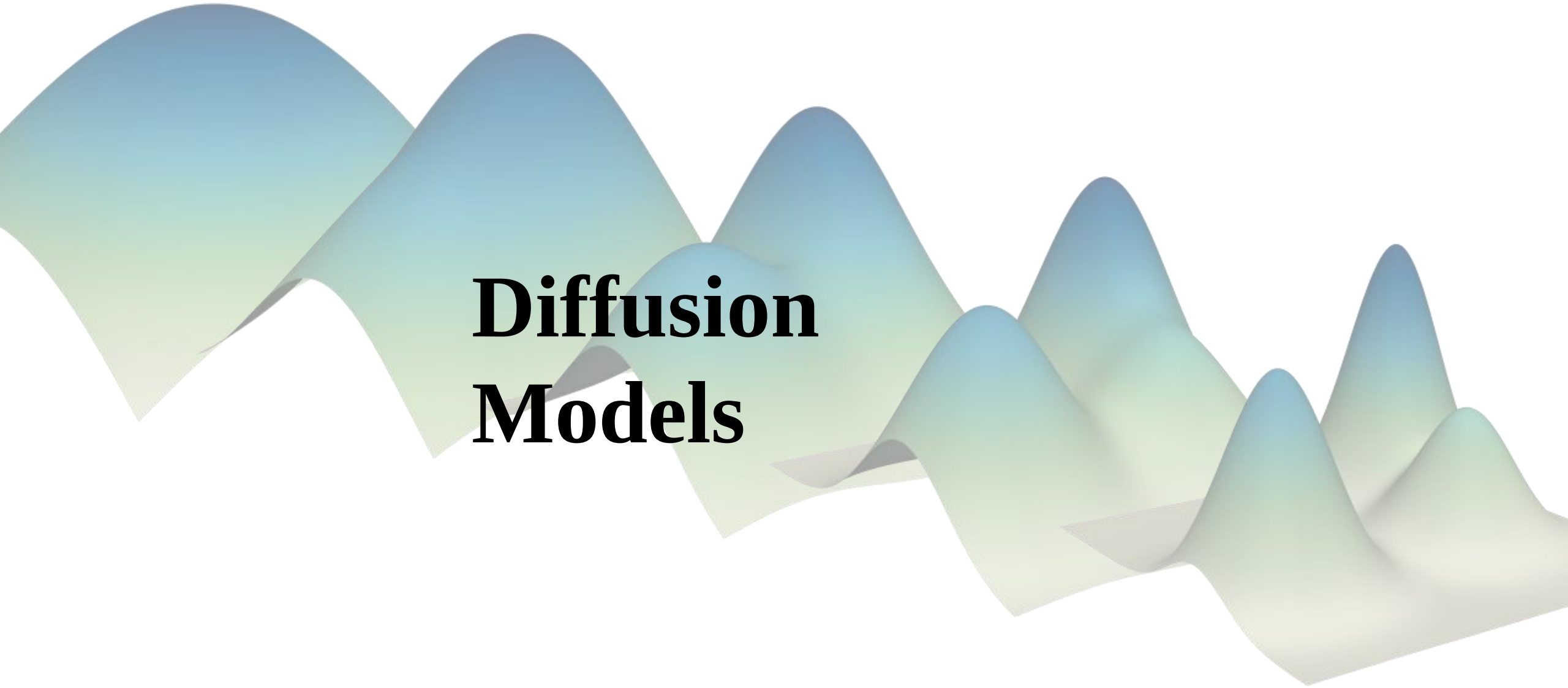
https://arxiv.org/pdf/2301.09515

GigaGAN: Scaling up GANs for Text-to-Image Synthesis





https://arxiv.org/pdf/2303.05511



Diffusion Models

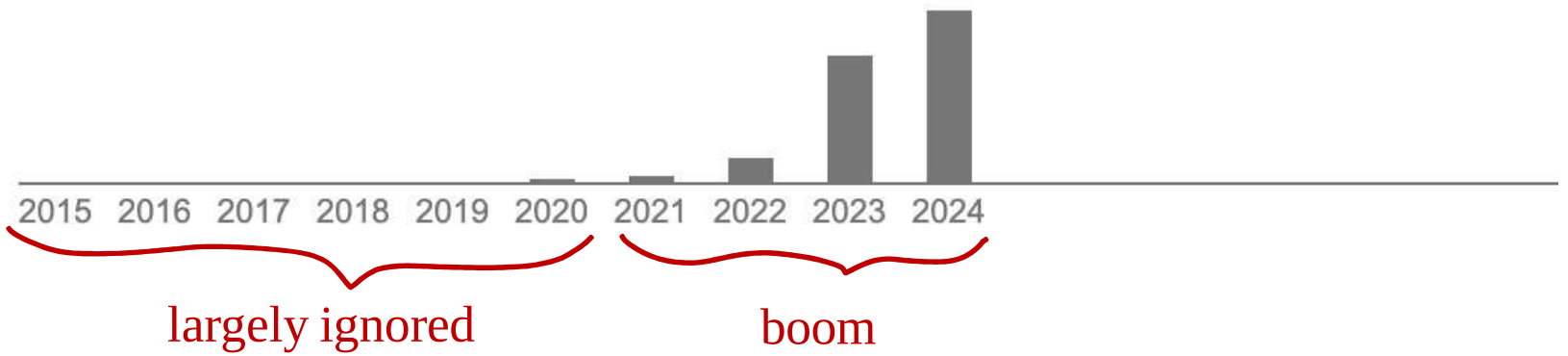
Overview

- Diffusion Models

- Energy-based Models and Score Matching

Deep unsupervised learning using nonequilibrium thermodynamics

Authors	Jascha Sohl-Dickstein, Eric A Weiss, Niru Maheswaranathan, Surya Ganguli
Publication date	2015/3/12
Journal	International Conference on Machine Learning
Total citations	Cited by 5630



Diffusion Models

Diffusion Models

- Forward process
- add noise to data

- Reverse process
- learn to denoise

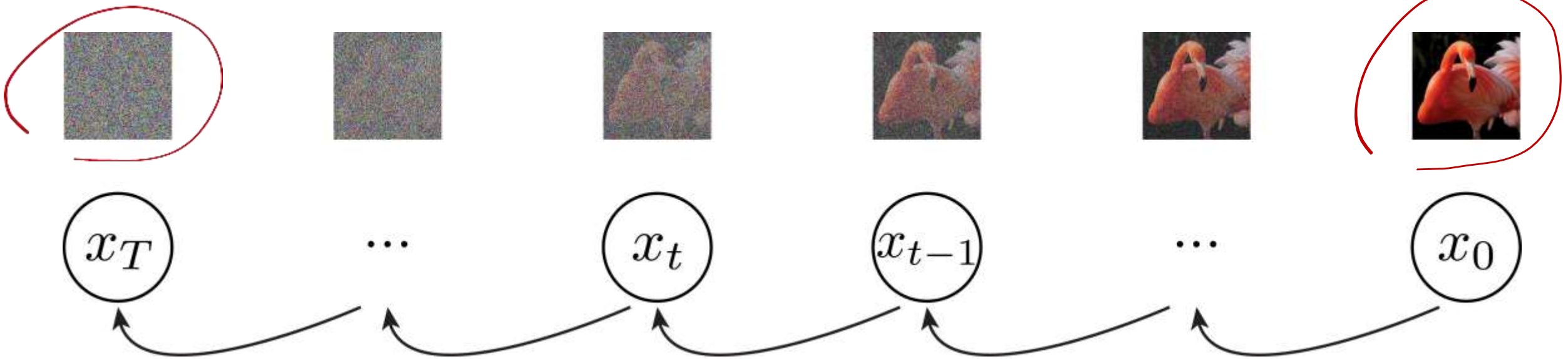
- Training objective
- from Hierarchical VAE to L2 loss

- Noise Conditional Network
- represent distributions

... in a nutshell

noise

data

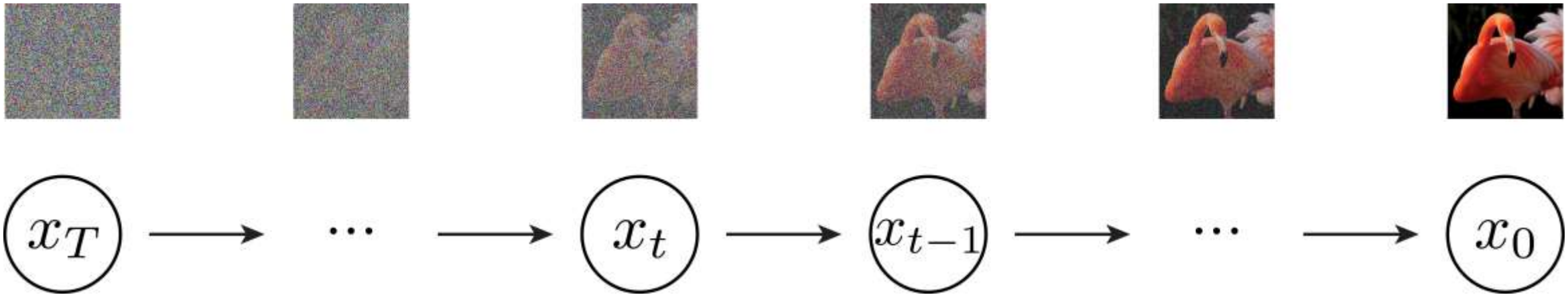


Forward process: add noise

... in a nutshell

noise

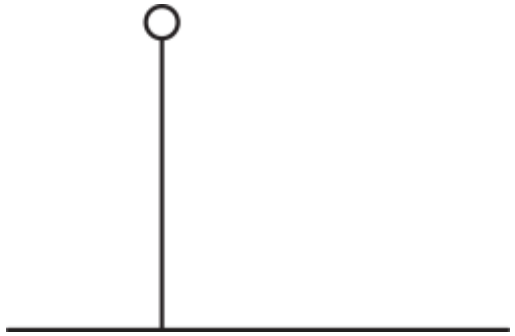
data



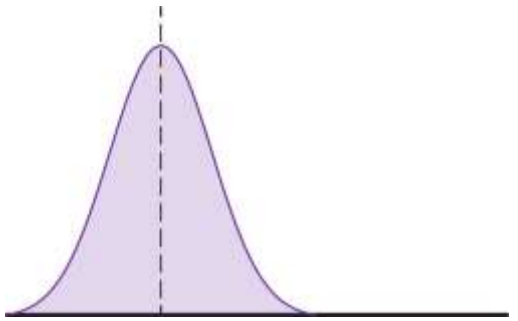
Reverse process: denoise

What is noise?

- Adding Gaussian noise ⇔ sampling $x \sim \mathcal{N}(x \mid x_0, \sigma)$



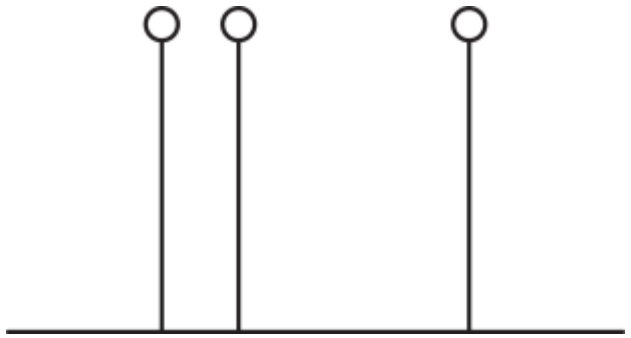
$$p(x) = \delta(x - x_0)$$



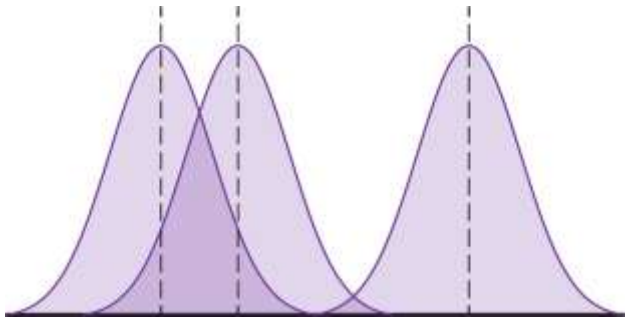
$$p(x) = \mathcal{N}(x \mid x_0, \sigma)$$

What is noise?

- Adding Gaussian noise ⇔ sampling $x \sim \mathcal{N}(x \mid x_0, \sigma)$



$p_{\text{data}}(x)$

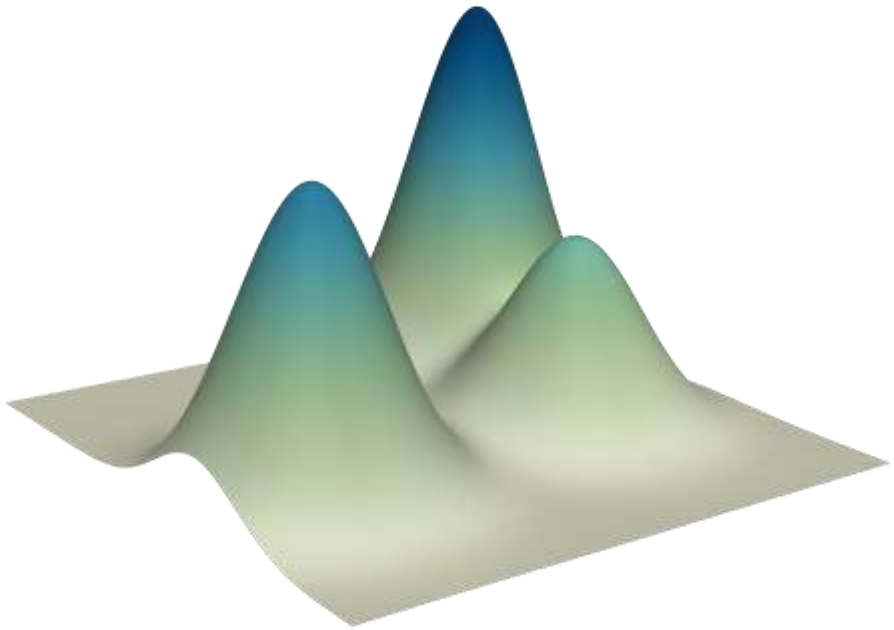


$p_{\text{data}}(x) * \mathcal{N}(x \mid 0, \sigma)$

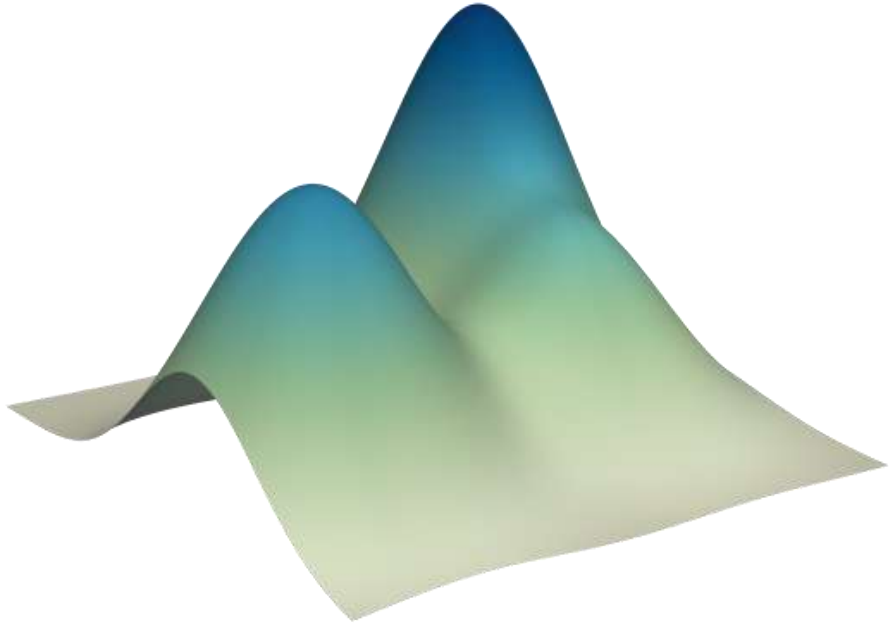
convolution
(of pdf)

What is noise?

- Adding Gaussian noise ⇔ $x \sim \mathcal{N}(x \mid x_0, \sigma)$ sampling



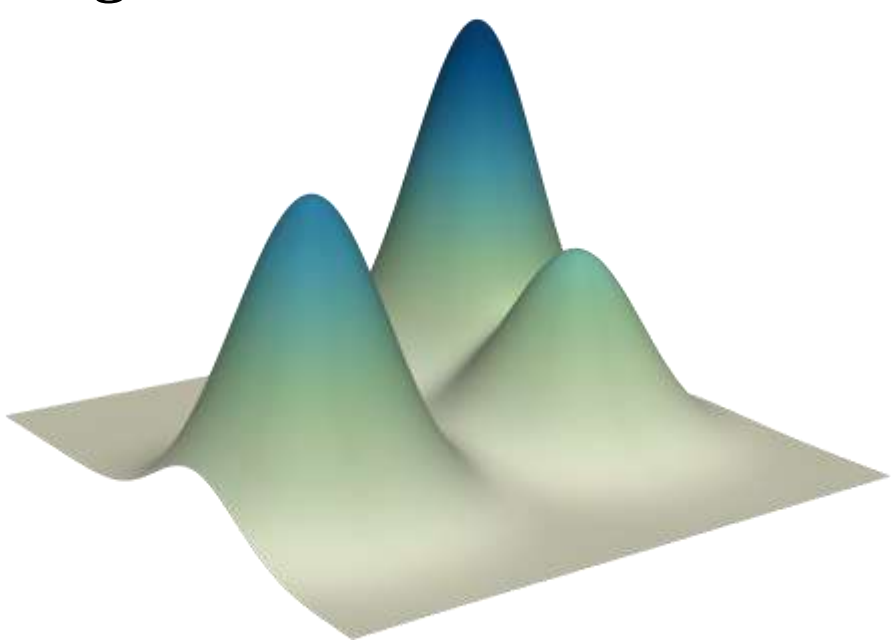
$p_{\text{data}}(x)$



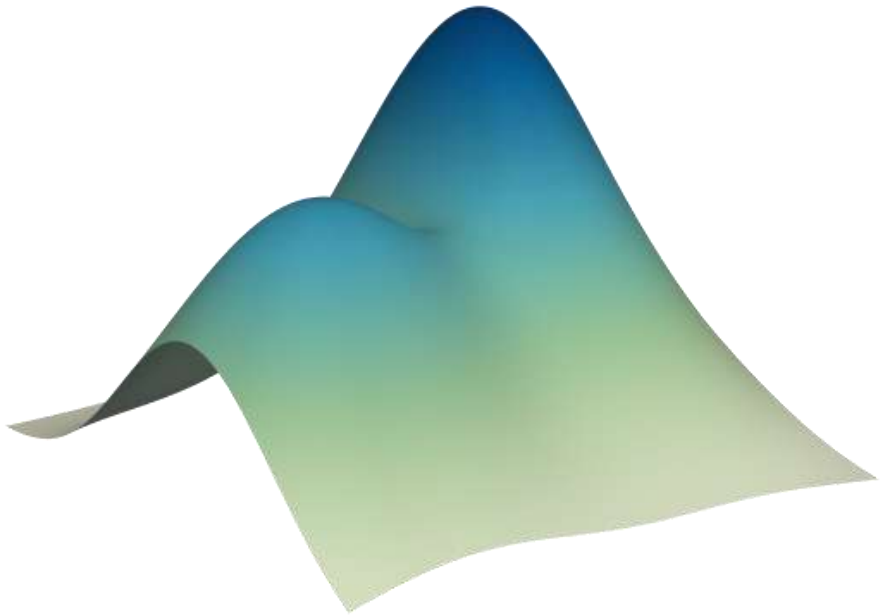
$p_{\text{data}}(x) * \mathcal{N}(x \mid 0, \sigma)$

What is noise?

- Adding Gaussian noise ⇔ $x \sim \mathcal{N}(x \mid x_0, \sigma)$ sampling



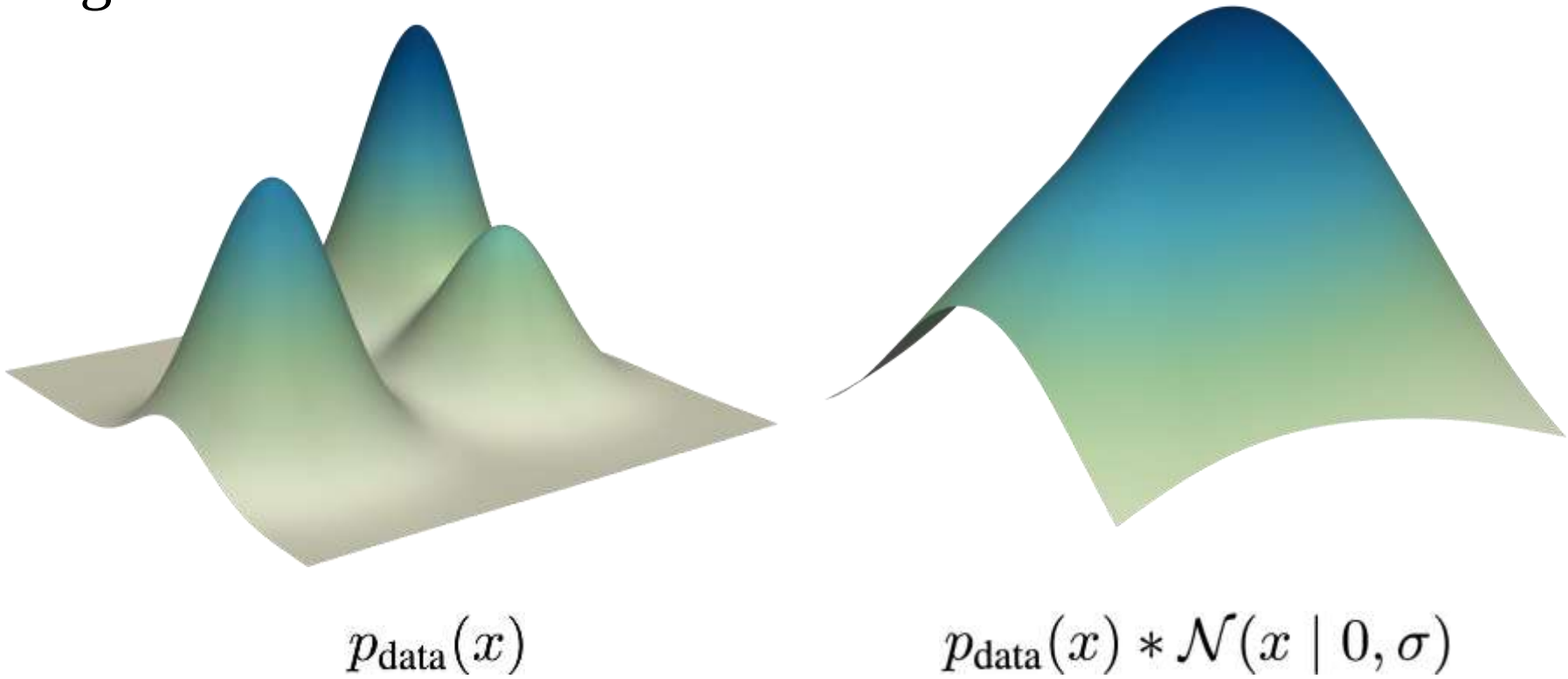
$p_{\text{data}}(x)$



$p_{\text{data}}(x) * \mathcal{N}(x \mid 0, \sigma)$

What is noise?

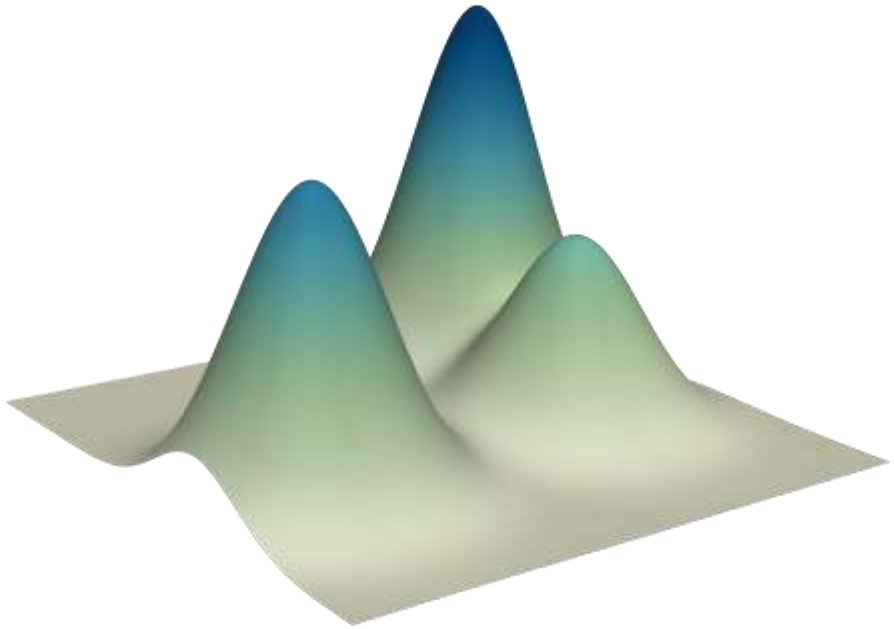
- Adding Gaussian noise ⇔ $x \sim \mathcal{N}(x \mid x_0, \sigma)$ sampling



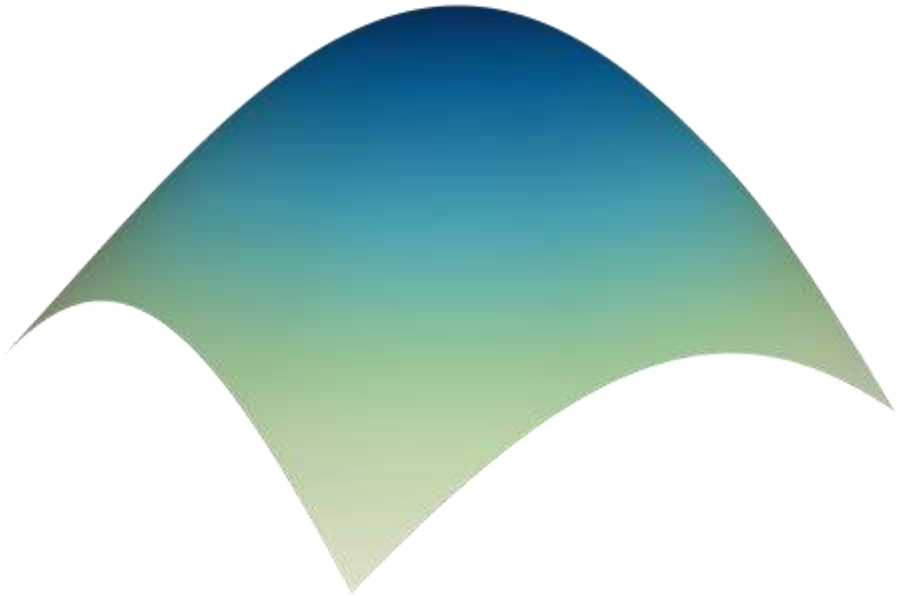
$p_{\text{data}}(x)$ $p_{\text{data}}(x) * \mathcal{N}(x \mid 0, \sigma)$

What is noise?

- Adding Gaussian noise ⇔ $x \sim \mathcal{N}(x \mid x_0, \sigma)$ sampling

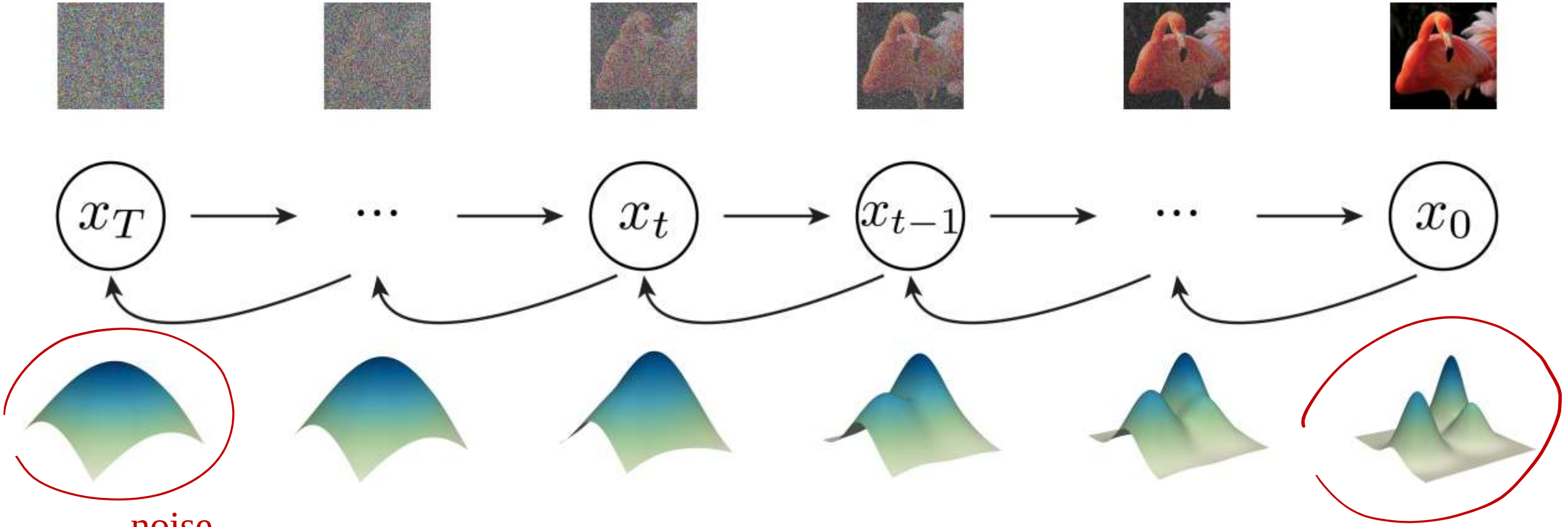


$p_{\text{data}}(x)$

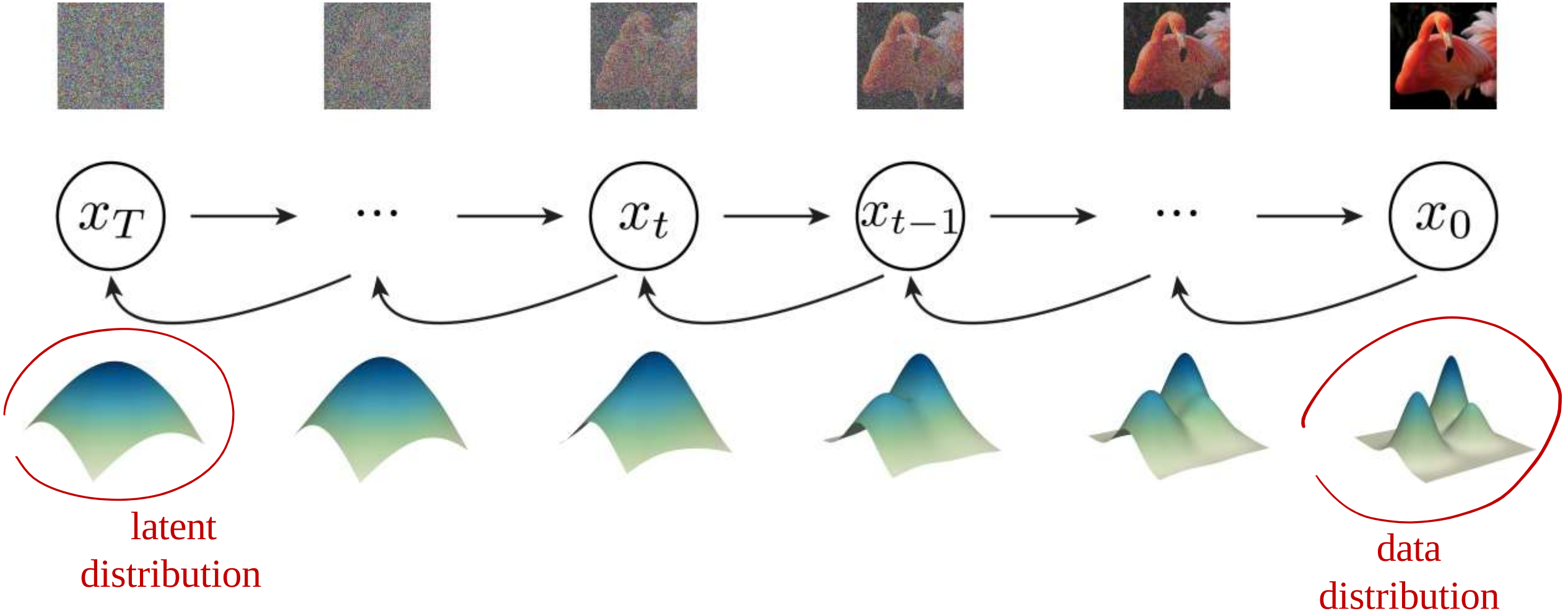


$p_{\text{data}}(x) * \mathcal{N}(x \mid 0, \sigma)$

What is noise?



What is noise?



Diffusion Models

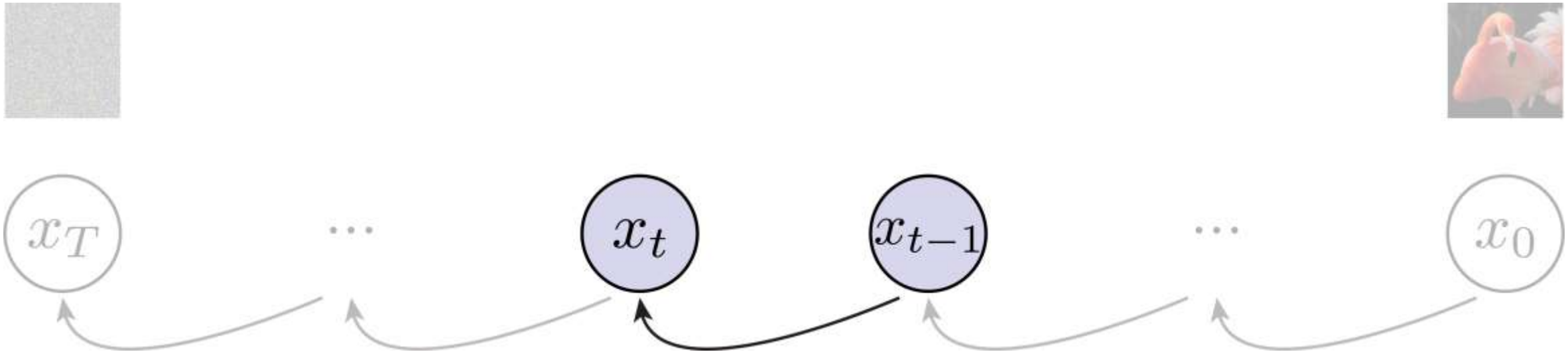
- Forward process
- add noise to data

- Reverse process
- learn to denoise

- Training objective
- from Hierarchical VAE to L2 loss

- Noise Conditional Network
- represent distributions

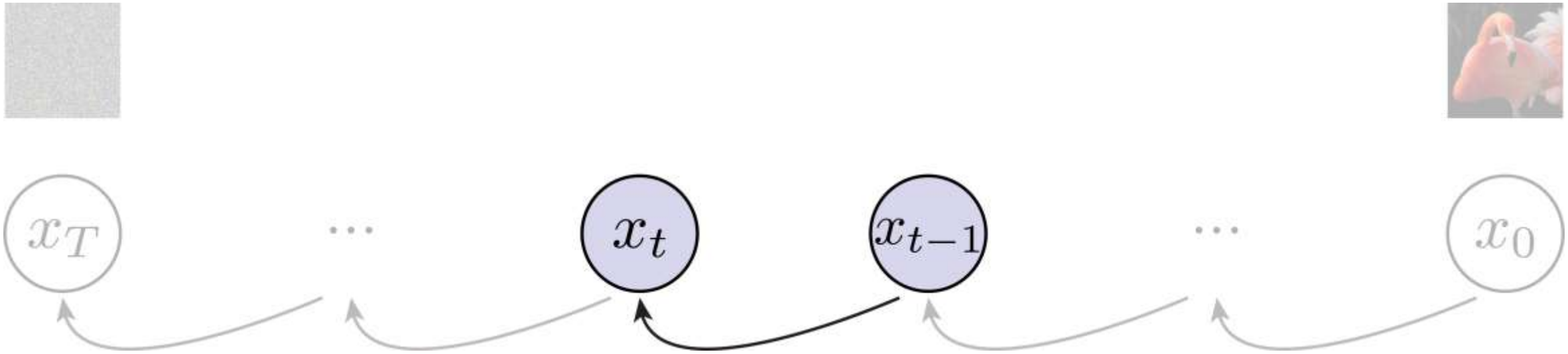
Forward Process



$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$$

coefficients:
variance preserving

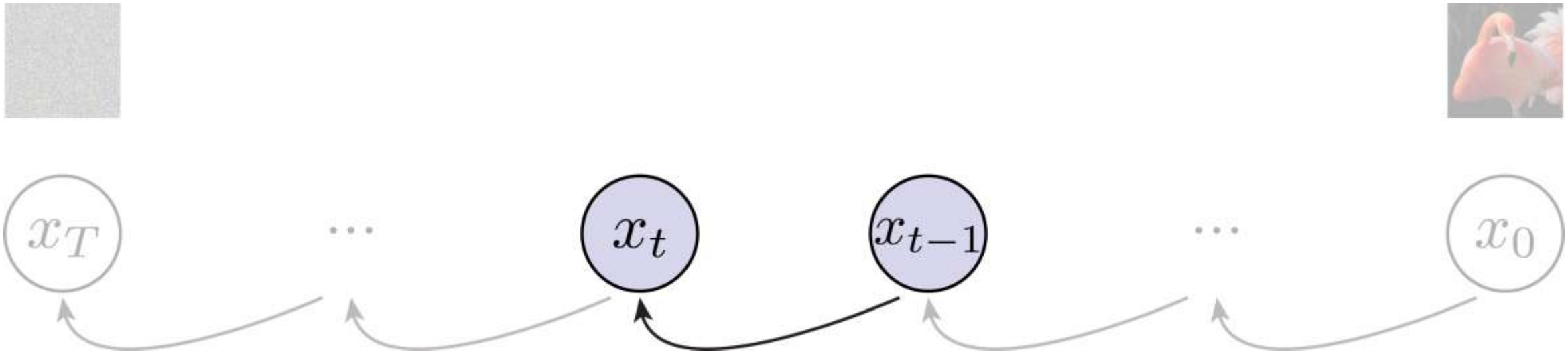
Forward Process



$$x_t = \sqrt{1-\beta_t}\,x_{t-1} + \sqrt{\beta_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$$

t: "schedule",
key to Diffusion Models' success

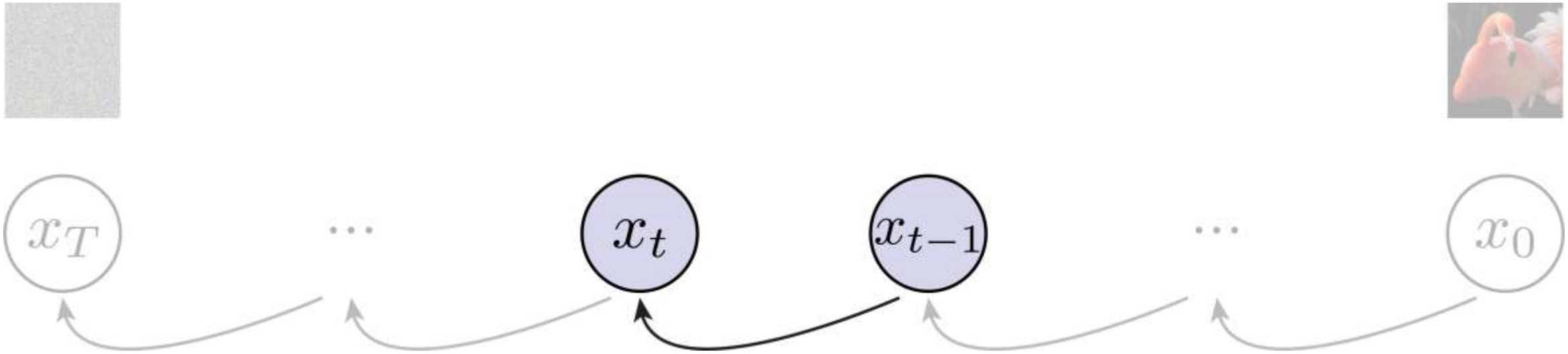
Forward Process



$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$$

mean of x_t std of x_t

Forward Process

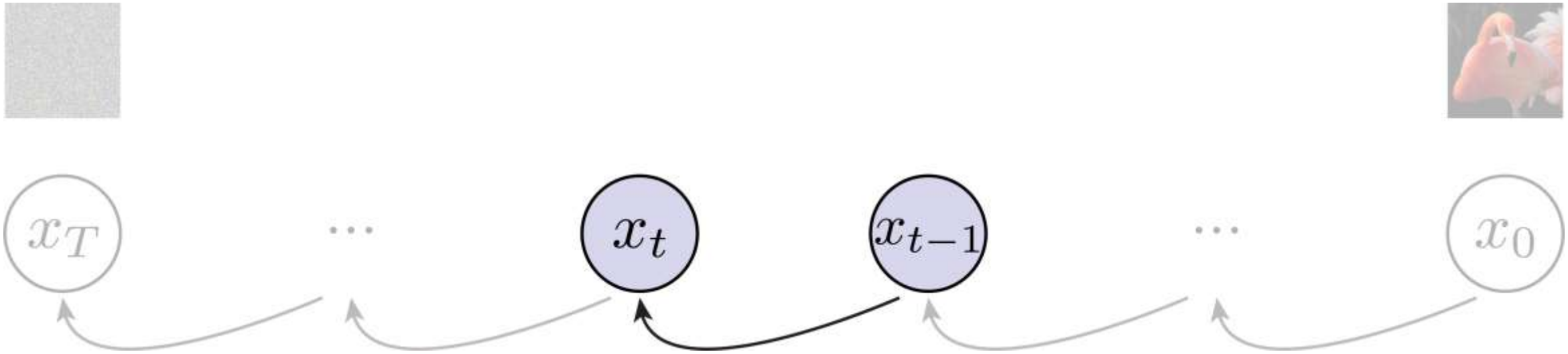


$$q(x_t \mid x_{t-1})$$

$$= \mathcal{N}(x_t \mid \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I})$$

mean of x_t var of x_t

Forward Process



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

$$= \mathcal{N}(x_t \mid \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I})$$

identity matrix

- sampling is i.i.d.
- dim = dim of data

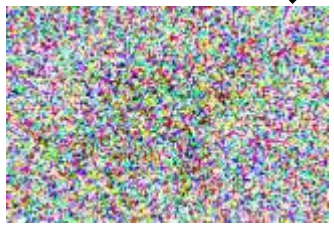
$q(x_t|x_{t-1})$

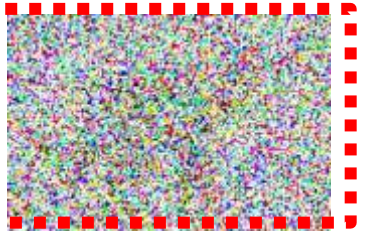
 $\sim \mathcal{N}(\mathbf{0}, I)$

 $= \sqrt{1-\beta_t}$  $+ \sqrt{\beta_t}$ 

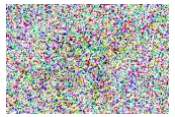
x_t x_{t-1} $\beta_1, \beta_2, \ldots, \beta_T$

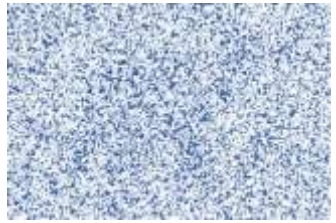
$q(x_t|x_0)$

 x_0 +   +   ... ...  +   x_t

x_1 $= \sqrt{1-\beta_1}$  $+ \sqrt{\beta_1}$ 

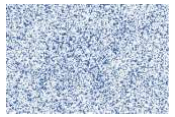
x_0

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

x_2 $= \sqrt{1-\beta_2}$  $+ \sqrt{\beta_2}$ 

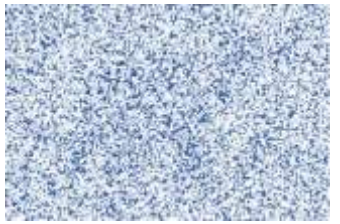
x_1

Ind.

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

x_2 $= \sqrt{1-\beta_2}\sqrt{1-\beta_1}$ 

x_0

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

x_2

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

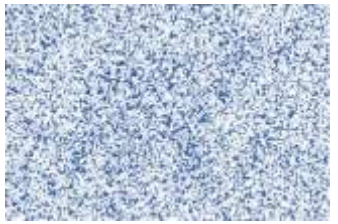
x_0

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

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

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

$+ \sqrt{\beta_2}$



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

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

$q(x_t|x_0)$

$\beta_1, \beta_2, \dots, \beta_T$

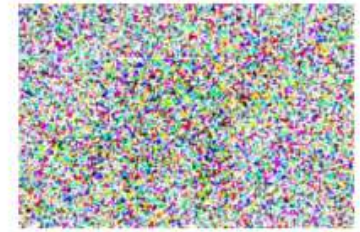
$\sim \mathcal{N}(\mathbf{0}, I)$

$\alpha_t = 1 - \beta_t$

$\bar{\alpha}_t = \alpha_1 \alpha_2 \dots \alpha_t$

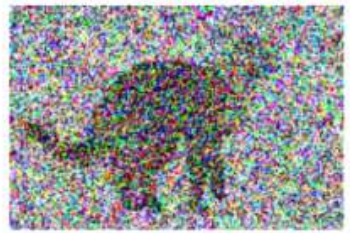
$= \sqrt{1-\beta_1}$

$+ \sqrt{\beta_1}$



$= \sqrt{1-\beta_2}$

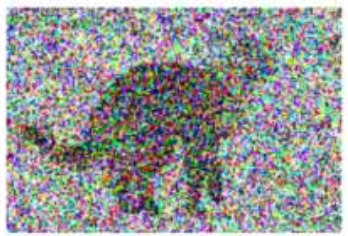
$+ \sqrt{\beta_2}$



$= \sqrt{1-\beta_t}$

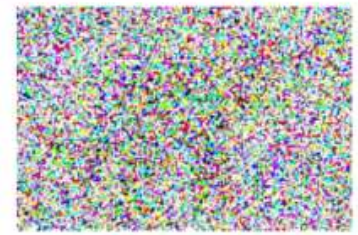
$+ \sqrt{\beta_t}$

$\Downarrow\Downarrow\Downarrow$

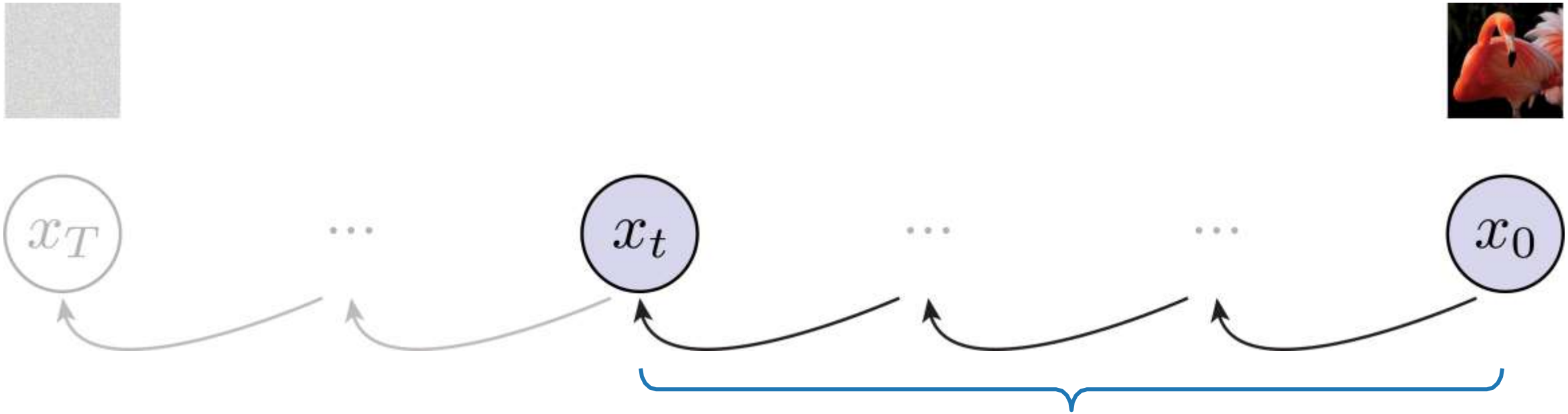


$= \dfrac{\sqrt{1-\beta_1} \dots \sqrt{1-\beta_t}}{\sqrt{\bar{\alpha}_t}}$

$+ \dfrac{\sqrt{1-(1-\beta_1)\dots(1-\beta_t)}}{\sqrt{1-\bar{\alpha}_t}}$



Forward Process



- sampling without simulation
- x_t from x_0 in closed form

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

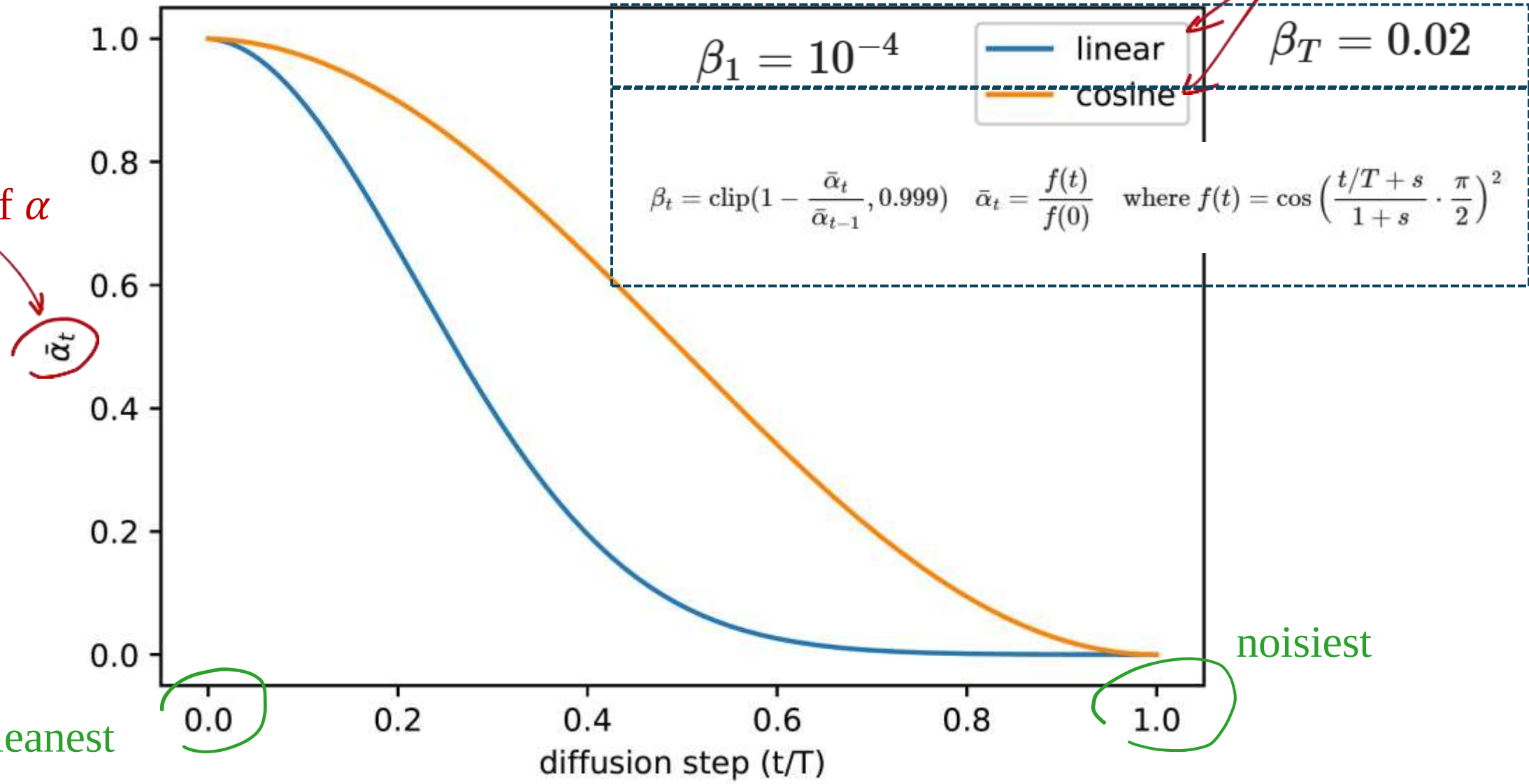
coefficients given by β

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

Forward Process: Noise Schedule

schedule of β

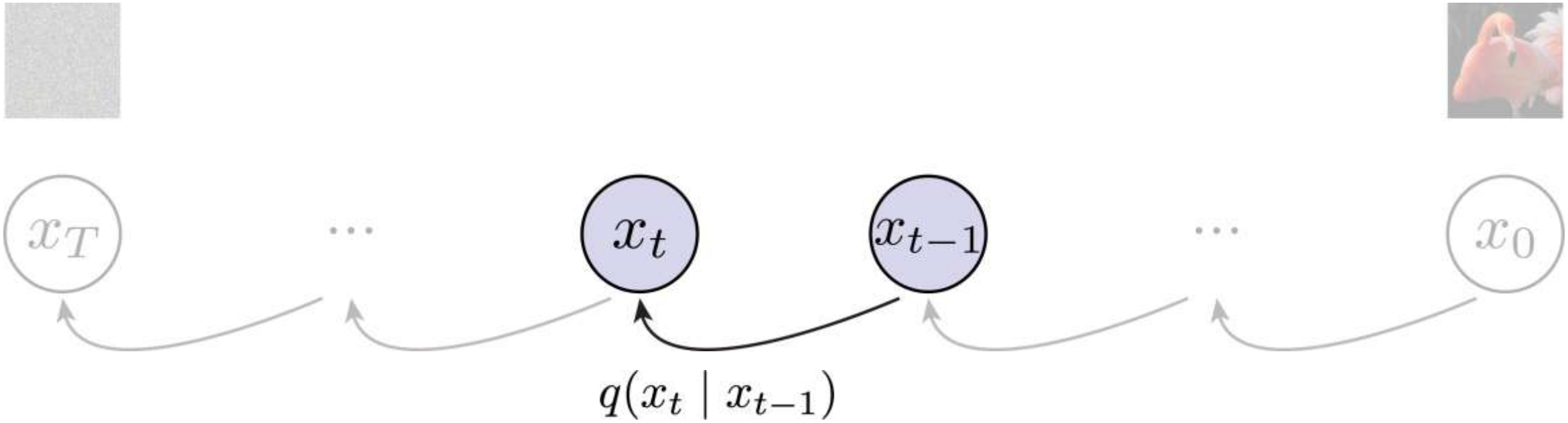
schedule of α



cleanest

noisiest

Forward Process



tl; dr:

- pre-defined conditional distributions
- Gaussian w/ controllable mean/std
- divide and conquer

Diffusion Models

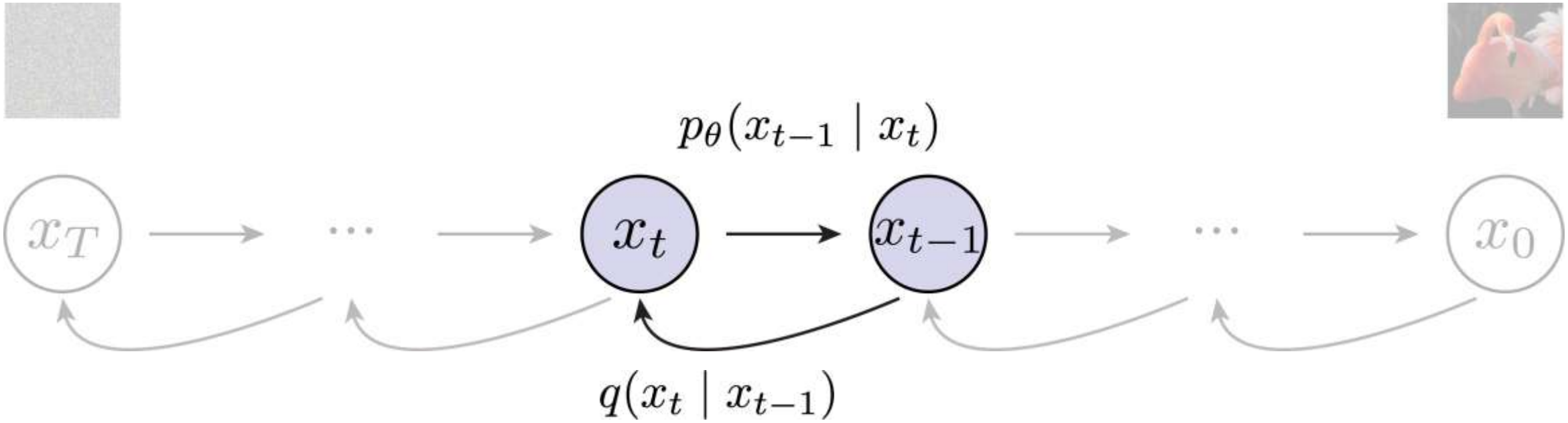
- Forward process
- add noise to data

- Reverse process
- learn to denoise

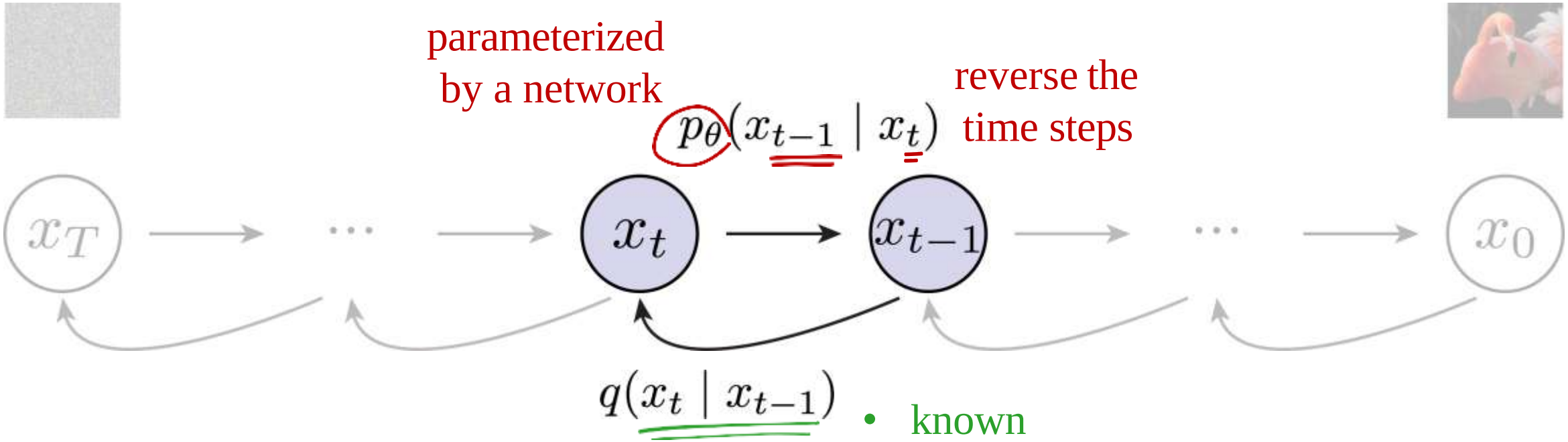
- Training objective
- from Hierarchical VAE to L2 loss

- Noise Conditional Network
- represent distributions

Reverse Process

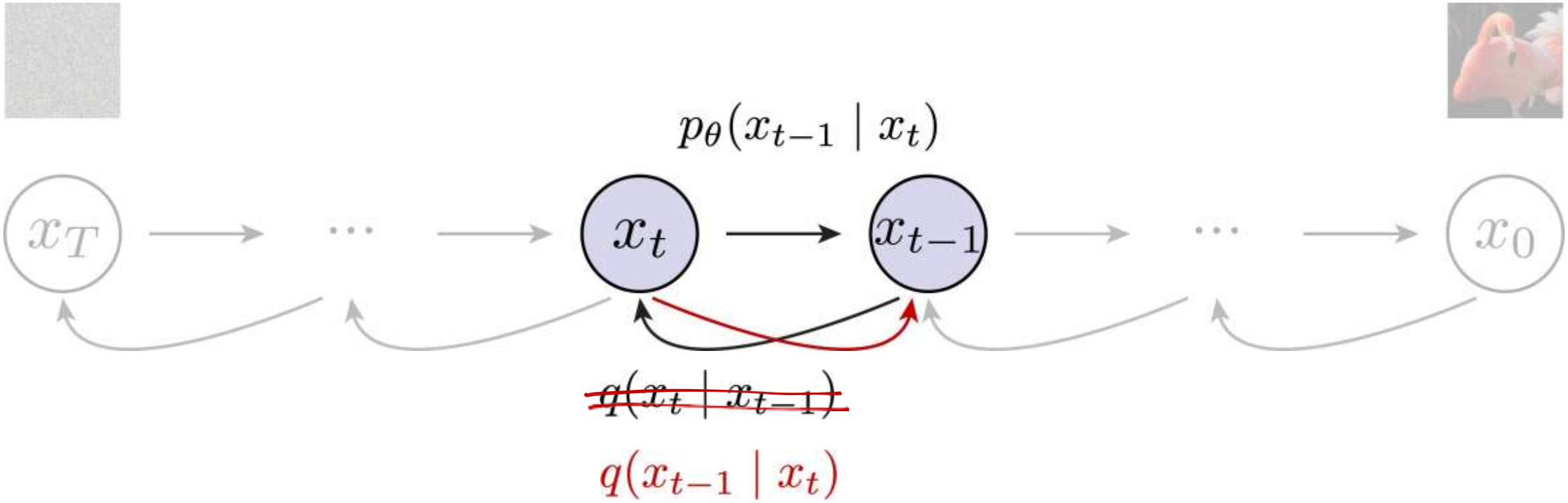


Reverse Process



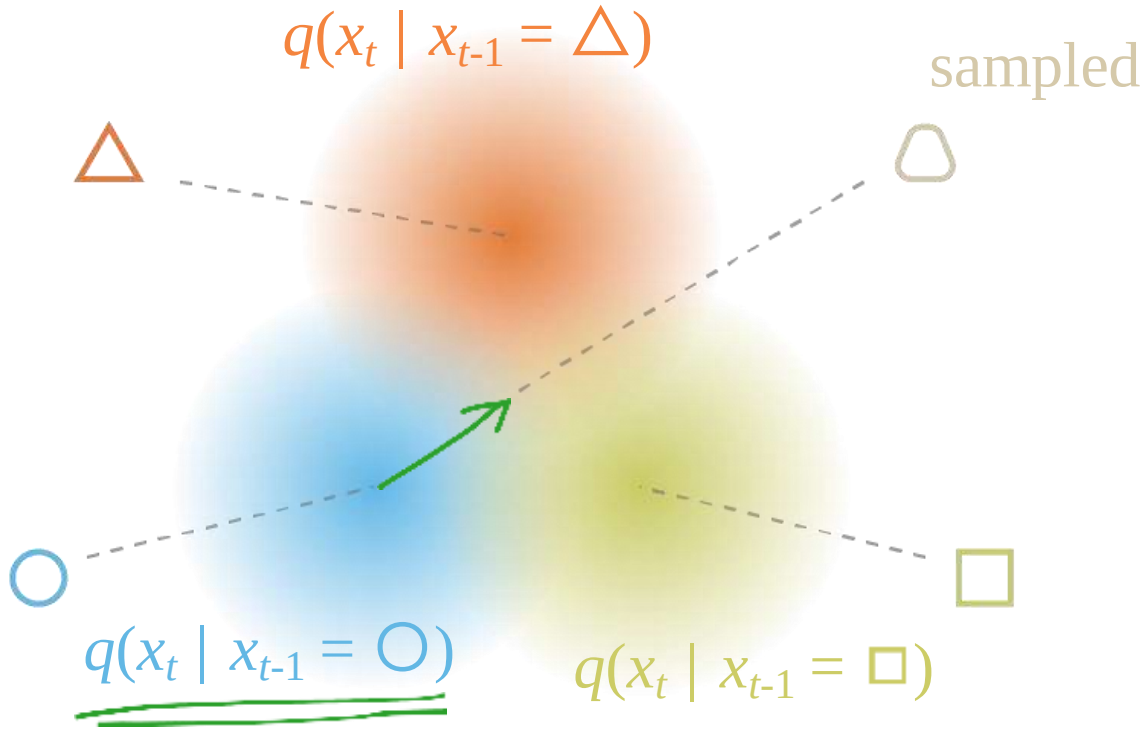
- known
- but not our target

Reverse Process



- our target
- but unknown

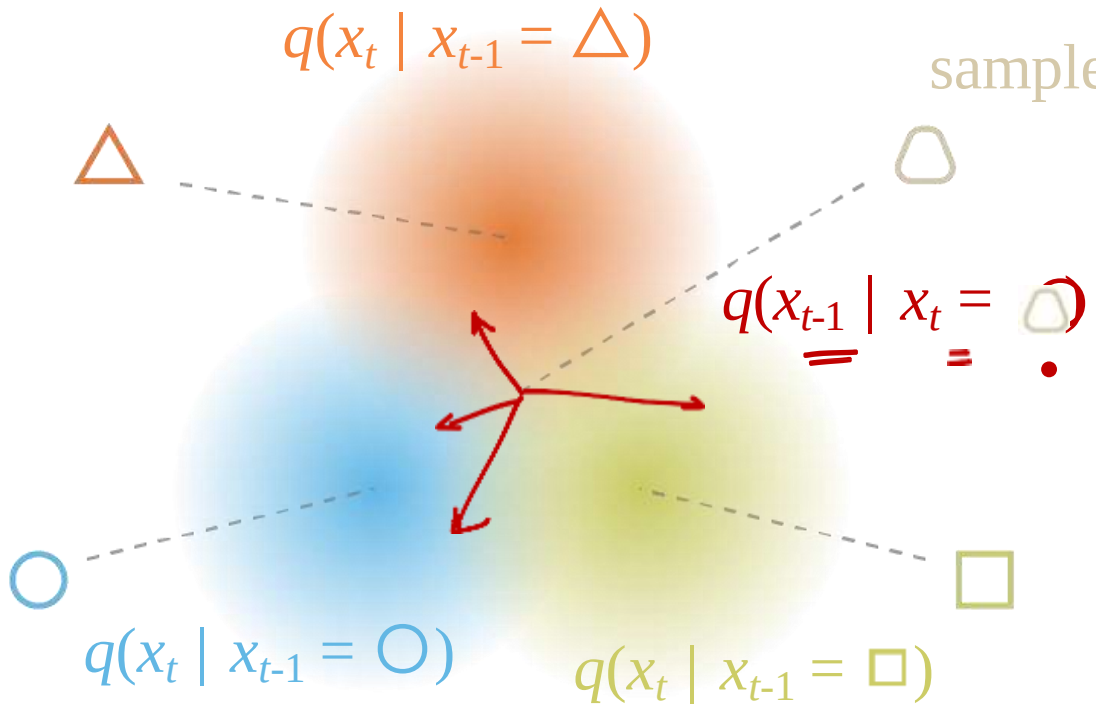
Why are the reverse conditionals unknown?



- known (Gaussian)

Figure adapted from: Joseph Rocca “Understanding Variational Autoencoders (VAEs)”
https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

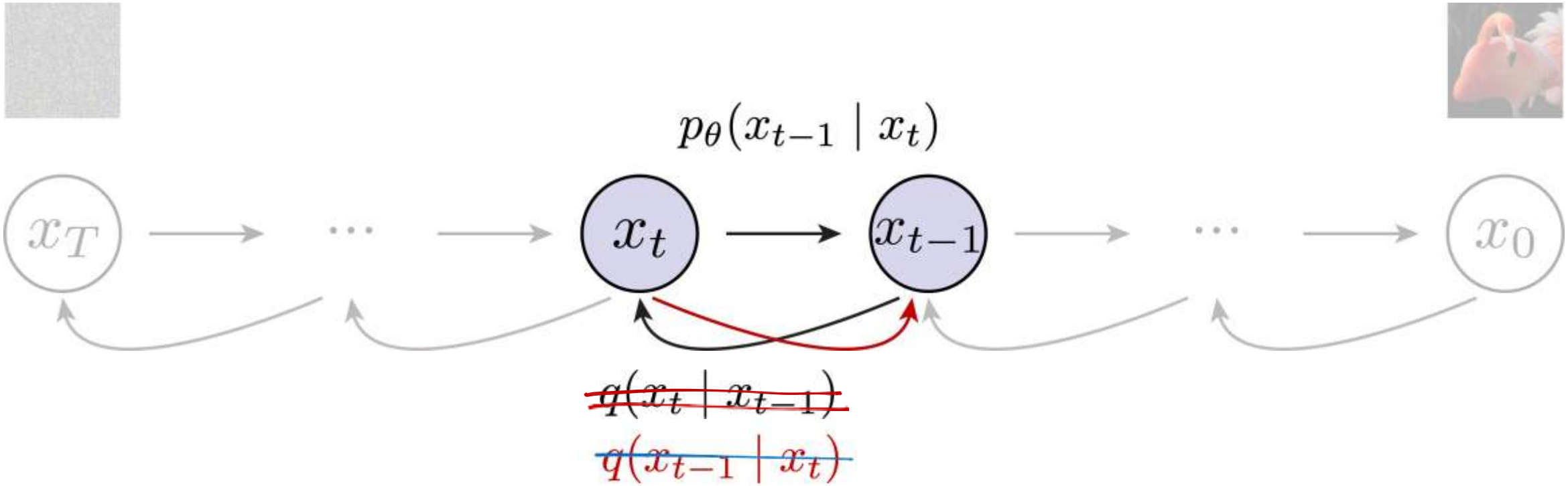
Why are the reverse conditionals unknown?



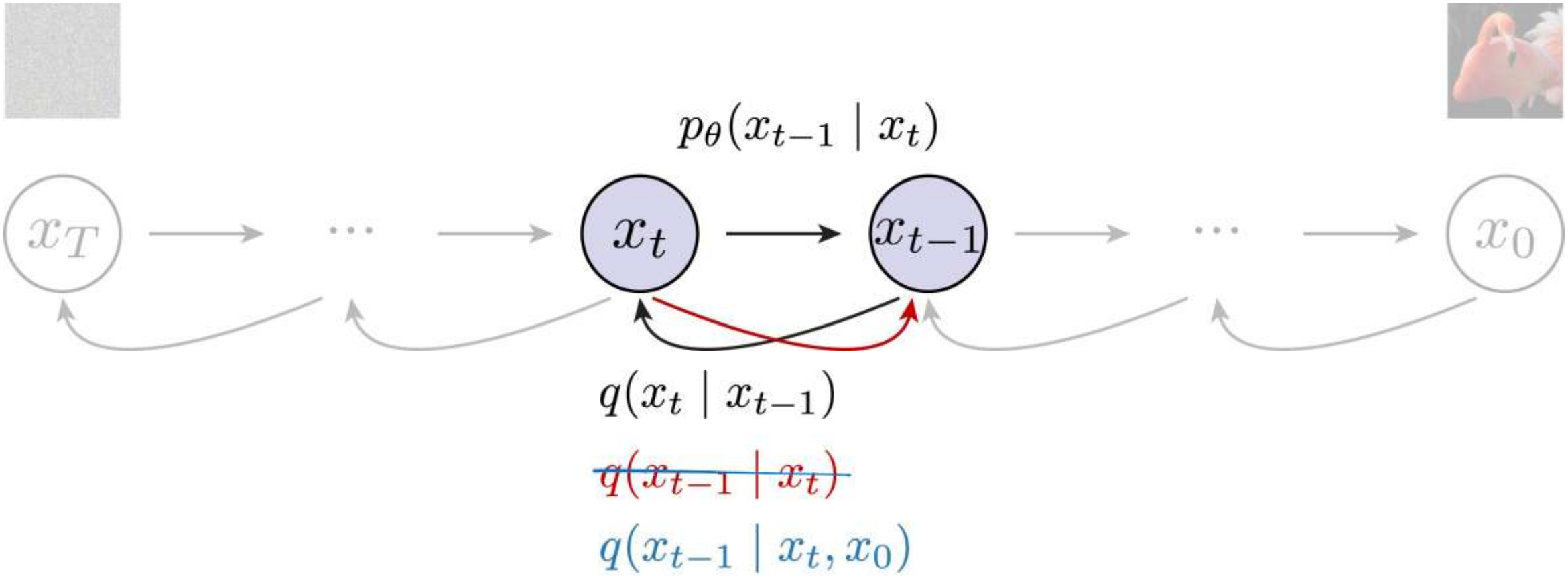
- unknown (depends on data distribution)

Figure adapted from: Joseph Rocca "Understanding Variational Autoencoders (VAEs)" https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

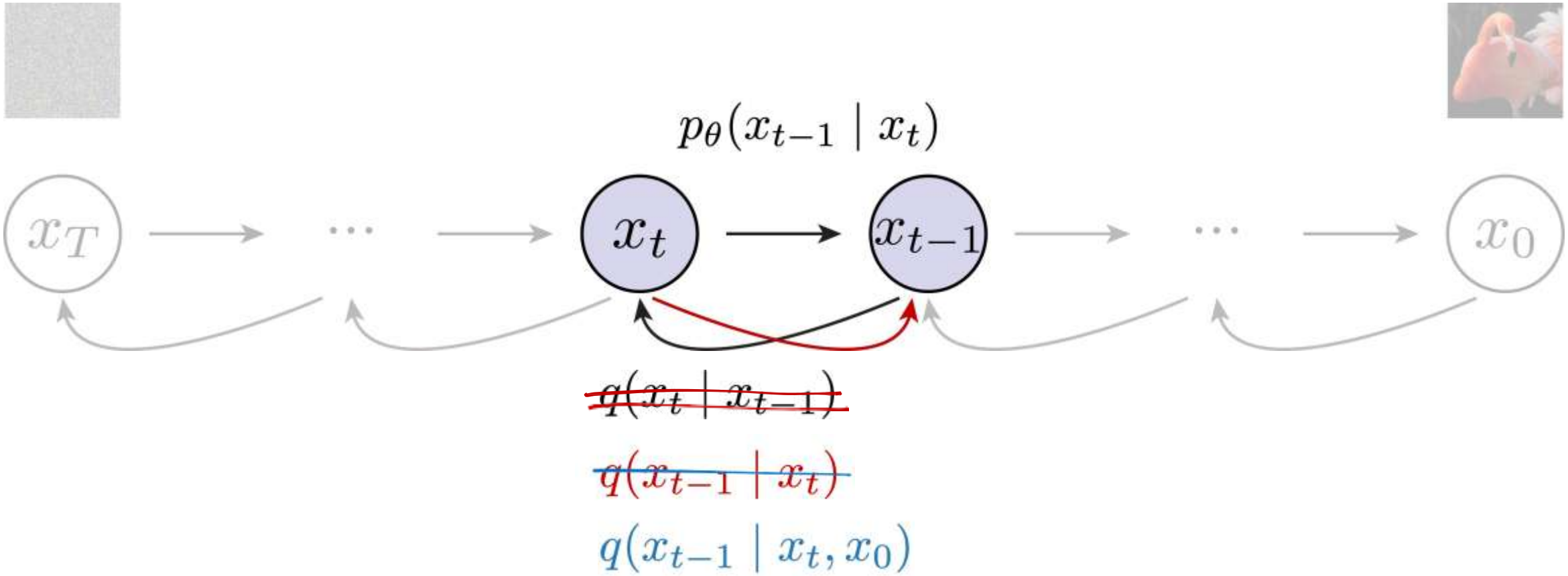
Reverse Process



Reverse Process



Reverse Process



- known
- Gaussian

Reverse Process

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

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

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

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

- known
- Gaussian

- known
- Gaussian

- known
- Gaussian

$$\log p(\boldsymbol{x}) = \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \tag{34}$$

$$= \log \int \frac{p(\boldsymbol{x}_{0:T}) q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} d\boldsymbol{x}_{1:T} \tag{35}$$

$$= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] \tag{36}$$

$$\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] \tag{37}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \tag{38}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \tag{39}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \prod_{t=1}^{T-1} p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \tag{40}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \prod_{t=1}^{T-1} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \tag{41}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\sum_{t=1}^{T-1} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \tag{42}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \tag{43}$$

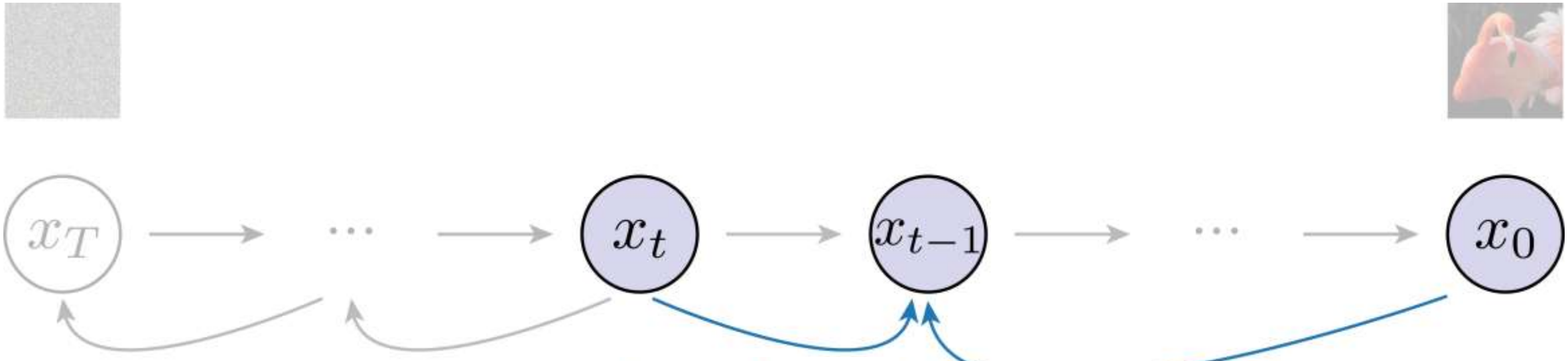
$$= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{T-1},\boldsymbol{x}_T|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_t,\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \tag{44}$$

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1) \right]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \parallel p(\boldsymbol{x}_T)) \right]}_{\text{prior matching term}} \tag{45}$$

$$- \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \parallel p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})) \right]}_{\text{consistency term}}$$

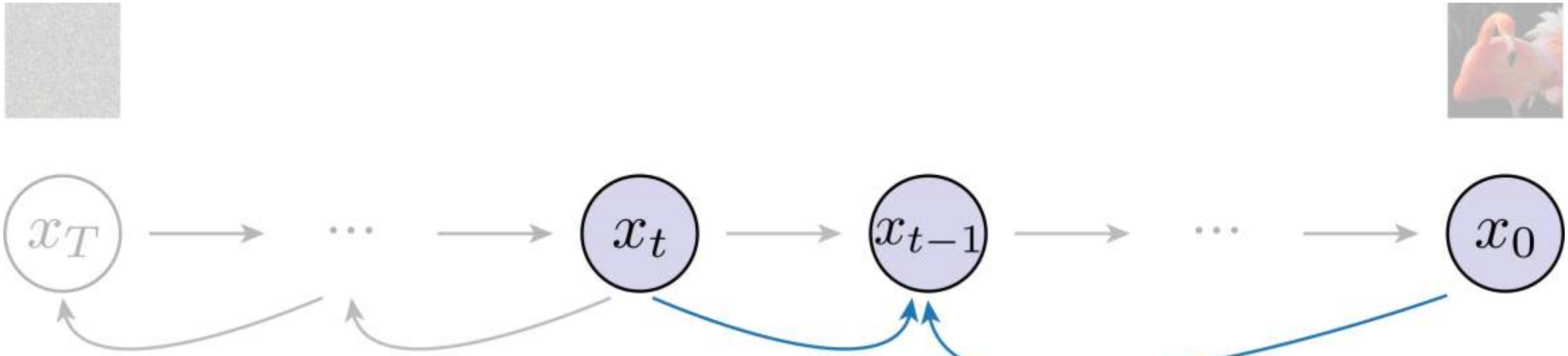
https://arxiv.org/pdf/2208.11970

Reverse Process



$$q(x_{t-1} \mid x_t, x_0)$$
$$= \mathcal{N}(x_{t-1} \mid \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I})$$

Reverse Process

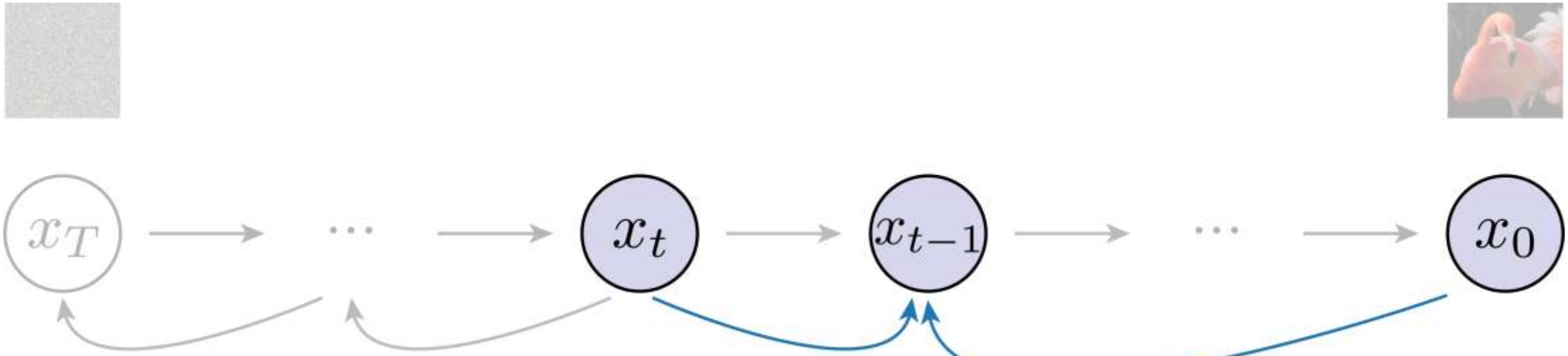


$$q(x_{t-1} \mid x_t, x_0)$$
$$= \mathcal{N}(x_{t-1} \mid \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I})$$

$$\tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$$

var

Reverse Process



$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1} \mid \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I})$$

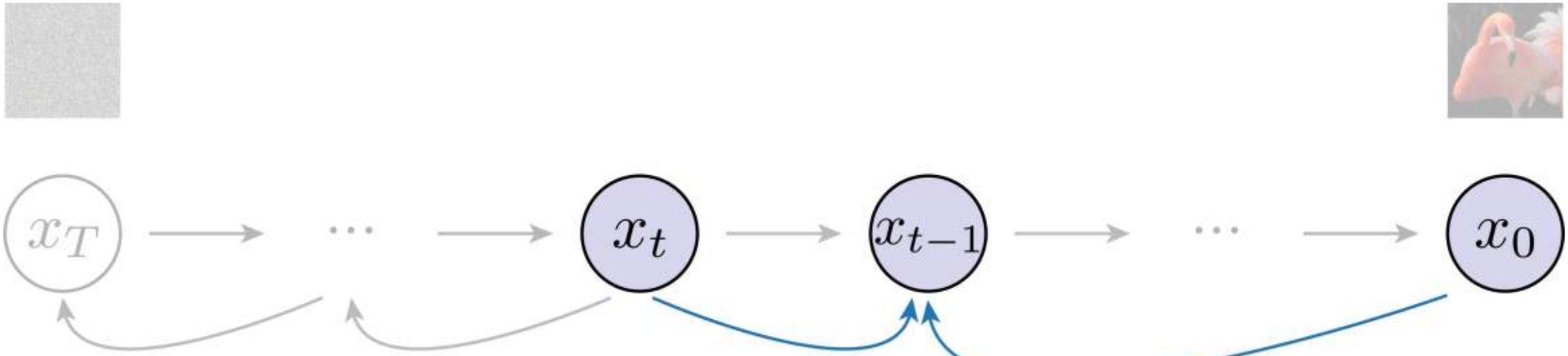
$$\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t$$

mean

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

var

Reverse Process



$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1} \mid \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I})$$

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

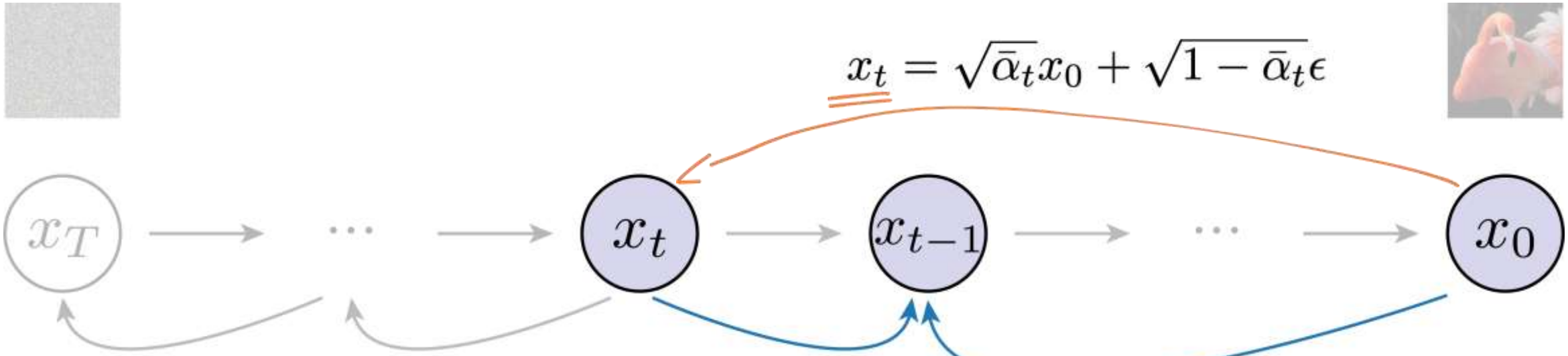
mean

linear combination

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

var

Reverse Process



$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1} \mid \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I})$$

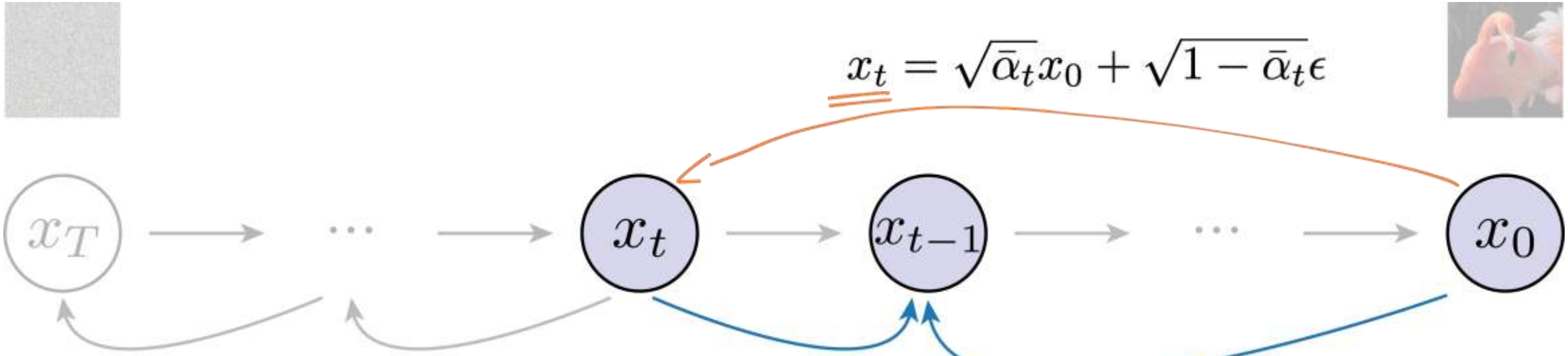
$$\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\alpha_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$$

mean

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

var

Reverse Process



$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1} \mid \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I})$$

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

mean

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

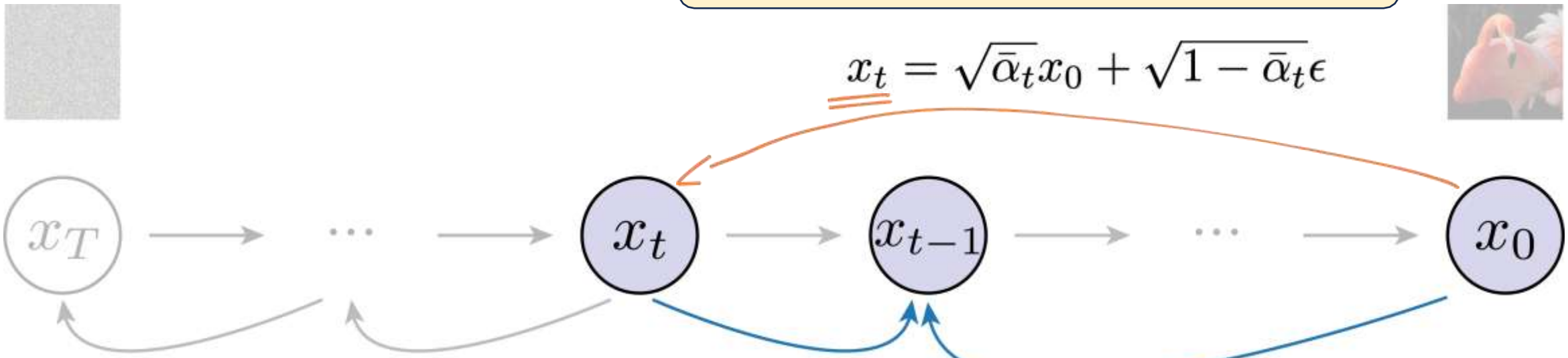
"noise"

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

var

Reverse Process

tl; dr:
- outcome of the dependency graph
- some linear combinations



$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1} \mid \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I})$$

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

mean

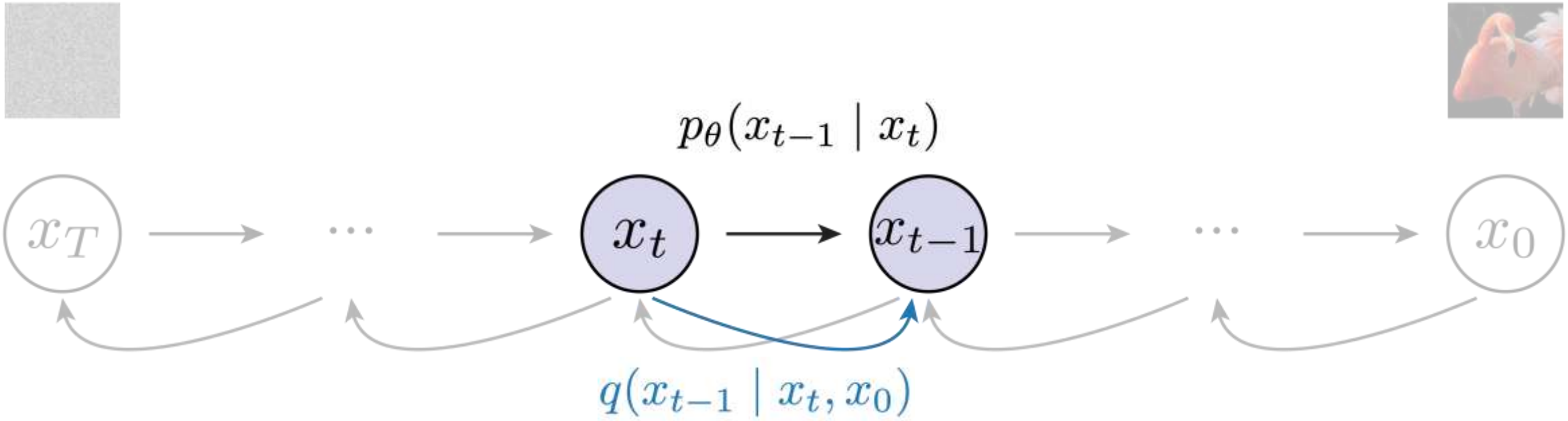
$$= \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)$$

"noise"

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

var

Reverse Process



- **tl; dr**: a known Gaussian
- we want to learn it by p_θ
- we can represent p_θ by a Gaussian
- minimize KL divergence

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \tag{71}$$

$$= \frac{\mathcal{N}(\boldsymbol{x}_t;\sqrt{\alpha_t}\boldsymbol{x}_{t-1},(1-\alpha_t)\mathbf{I})\mathcal{N}(\boldsymbol{x}_{t-1};\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0,(1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\boldsymbol{x}_t;\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0,(1-\bar{\alpha}_t)\mathbf{I})} \tag{72}$$

$$\propto \exp\left\{-\left[\frac{(\boldsymbol{x}_t-\sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{2(1-\alpha_t)} + \frac{(\boldsymbol{x}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_{t-1})} - \frac{(\boldsymbol{x}_t-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_t)}\right]\right\} \tag{73}$$

$$= \exp\left\{-\frac{1}{2}\left[\frac{(\boldsymbol{x}_t-\sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{1-\alpha_t} + \frac{(\boldsymbol{x}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(\boldsymbol{x}_t-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{1-\bar{\alpha}_t}\right]\right\} \tag{74}$$

$$= \exp\left\{-\frac{1}{2}\left[\frac{(-2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1}+\alpha_t\boldsymbol{x}_{t-1}^2)}{1-\alpha_t} + \frac{(\boldsymbol{x}_{t-1}^2-2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0)}{1-\bar{\alpha}_{t-1}} + C(\boldsymbol{x}_t,\boldsymbol{x}_0)\right]\right\} \tag{75}$$

$$\propto \exp\left\{-\frac{1}{2}\left[-\frac{2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1}}{1-\alpha_t} + \frac{\alpha_t\boldsymbol{x}_{t-1}^2}{1-\alpha_t} + \frac{\boldsymbol{x}_{t-1}^2}{1-\bar{\alpha}_{t-1}} - \frac{2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right]\right\} \tag{76}$$

$$= \exp\left\{-\frac{1}{2}\left[(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\bar{\alpha}_{t-1}})\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} \tag{77}$$

$$= \exp\left\{-\frac{1}{2}\left[\frac{\alpha_t(1-\bar{\alpha}_{t-1})+1-\alpha_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} \tag{78}$$

$$= \exp\left\{-\frac{1}{2}\left[\frac{\alpha_t-\bar{\alpha}_t+1-\alpha_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} \tag{79}$$

$$= \exp\left\{-\frac{1}{2}\left[\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} \tag{80}$$

$$= \exp\left\{-\frac{1}{2}\left(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right)\left[\boldsymbol{x}_{t-1}^2 - 2\frac{\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)}{\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}}\boldsymbol{x}_{t-1}\right]\right\} \tag{81}$$

$$= \exp\left\{-\frac{1}{2}\left(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right)\left[\boldsymbol{x}_{t-1}^2 - 2\frac{\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_{t-1}\right]\right\} \tag{82}$$

$$= \exp\left\{-\frac{1}{2}\left(\frac{1}{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}\right)\left[\boldsymbol{x}_{t-1}^2 - 2\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}\boldsymbol{x}_{t-1}\right]\right\} \tag{83}$$

$$\propto \mathcal{N}(\boldsymbol{x}_{t-1};\underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t,\boldsymbol{x}_0)},\underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)}) \tag{84}$$

https://arxiv.org/pdf/2208.11970

$$\mathrm{E}_{q(x_1|x_0)}[logP(x_0|x_1)] - KL\left(q(x_T|x_0)||P(x_T)\right)$$

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

Gaussian

Mean

Variance

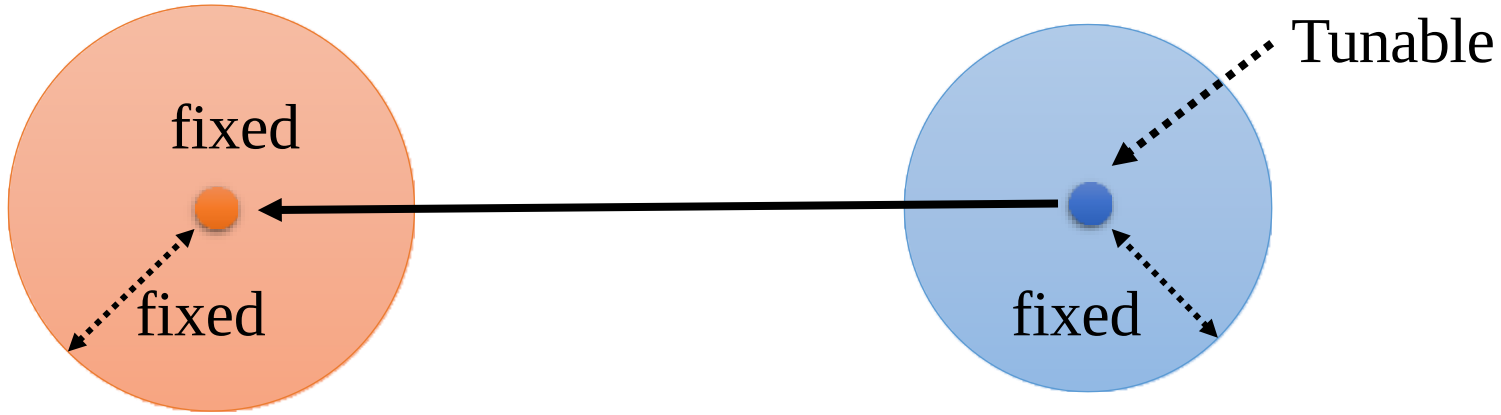
$$\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t x_0 + \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t}{1-\bar{\alpha}_t}$$

$$\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t I$$

$$\mathrm{E}_{q(x_1|x_0)}[logP(x_0|x_1)] - KL\left(q(x_T|x_0)||P(x_T)\right)$$

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

How to minimize KL divergence?



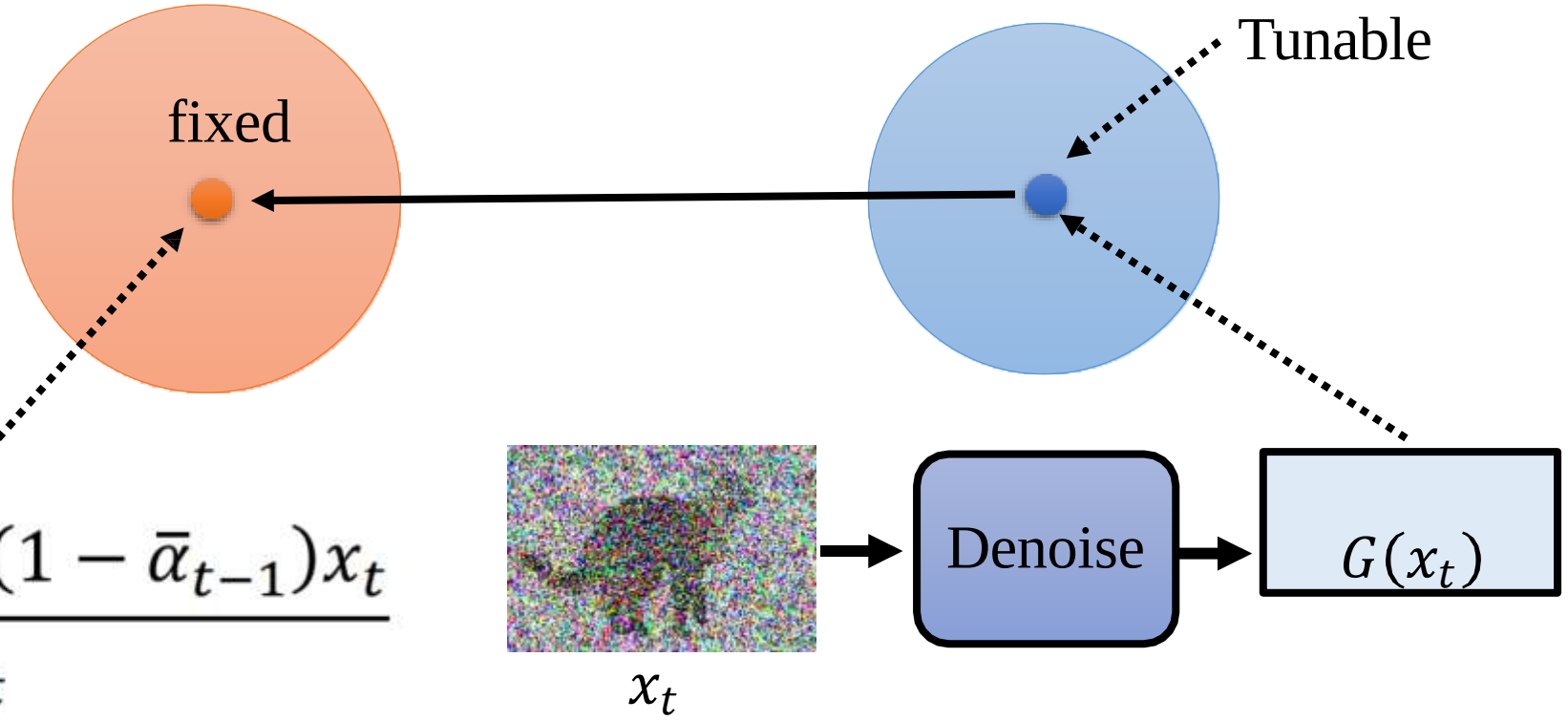
Recall that the KL Divergence between two Gaussian distributions is:

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

$$\mathrm{E}_{q(x_1|x_0)}[logP(x_0|x_1)] - KL\left(q(x_T|x_0)||P(x_T)\right)$$

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

How to minimize KL divergence?

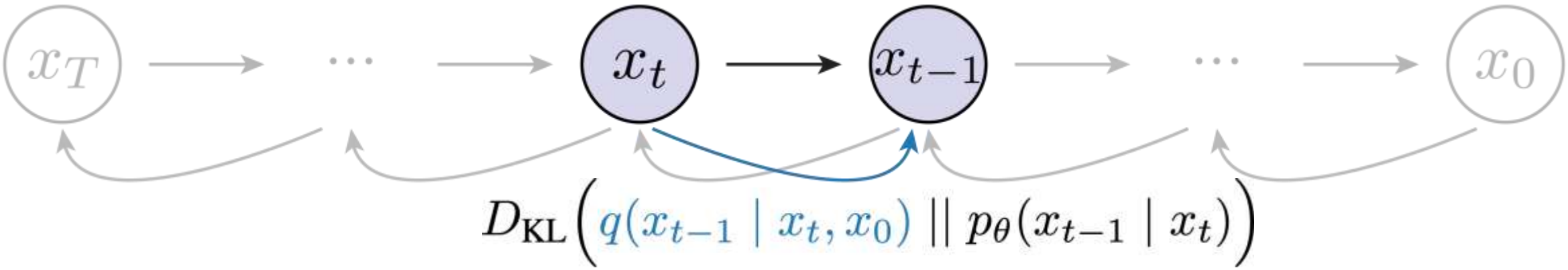


$$\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t x_0 + \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t}{1-\bar{\alpha}_t}$$

Reverse Process

D_{KL} of two Gaussians is like L2 loss:

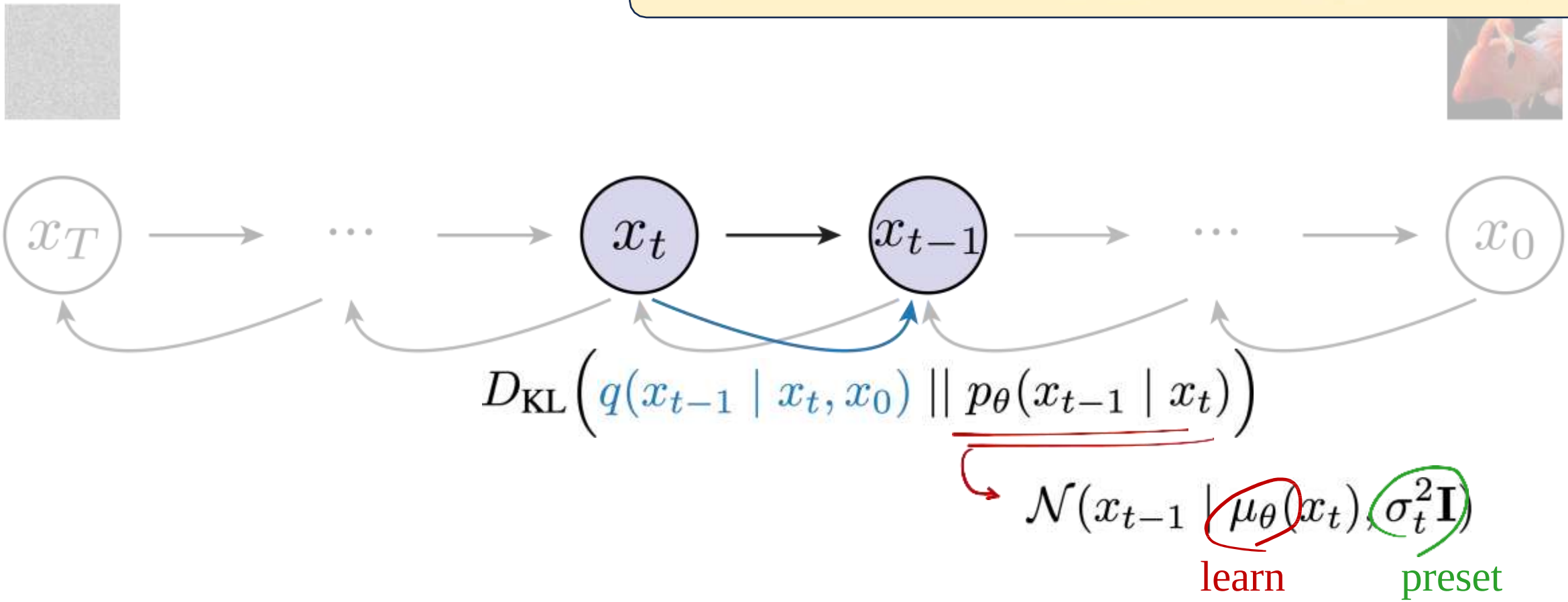
$$D_{KL}(\mathcal{N}_1 \parallel \mathcal{N}_2) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + \|\mu_1 - \mu_2\|^2}{2\sigma_2^2} - \frac{1}{2}$$



Reverse Process

D_{KL} of two Gaussians is like L2 loss:

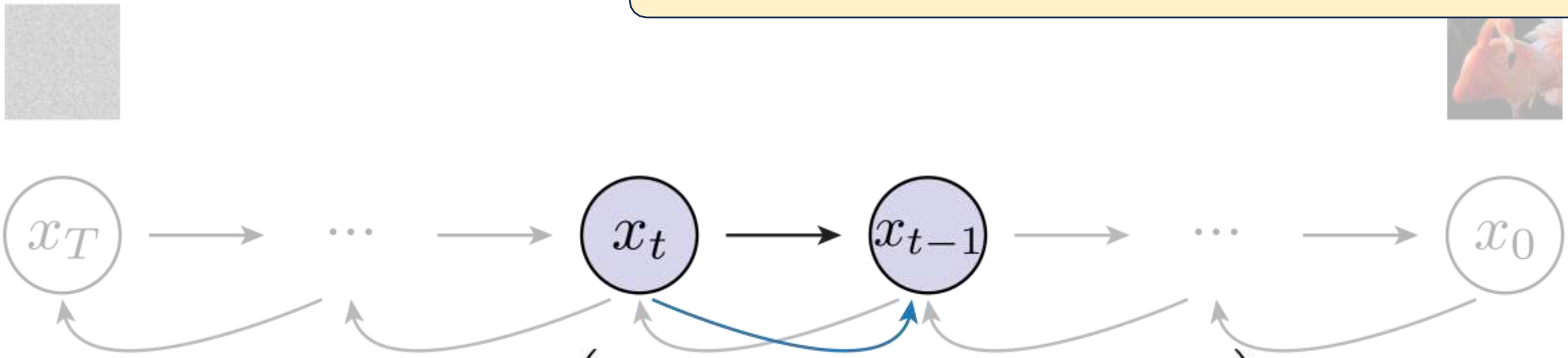
$$D_{KL}(\mathcal{N}_1 \parallel \mathcal{N}_2) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + \|\mu_1 - \mu_2\|^2}{2\sigma_2^2} - \frac{1}{2}$$



Reverse Process

D_{KL} of two Gaussians is like L2 loss:

$$D_{KL}(\mathcal{N}_1 \parallel \mathcal{N}_2) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + \|\mu_1 - \mu_2\|^2}{2\sigma_2^2} - \frac{1}{2}$$



$$D_{\mathrm{KL}}\Big(q(x_{t-1} \mid x_t, x_0) \;||\; p_\theta(x_{t-1} \mid x_t)\Big)$$

$$\mathcal{N}(x_{t-1} \mid \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I})$$

$$\mathcal{N}(x_{t-1} \mid \mu_\theta(x_t), \sigma_t^2 \mathbf{I})$$

learn

preset

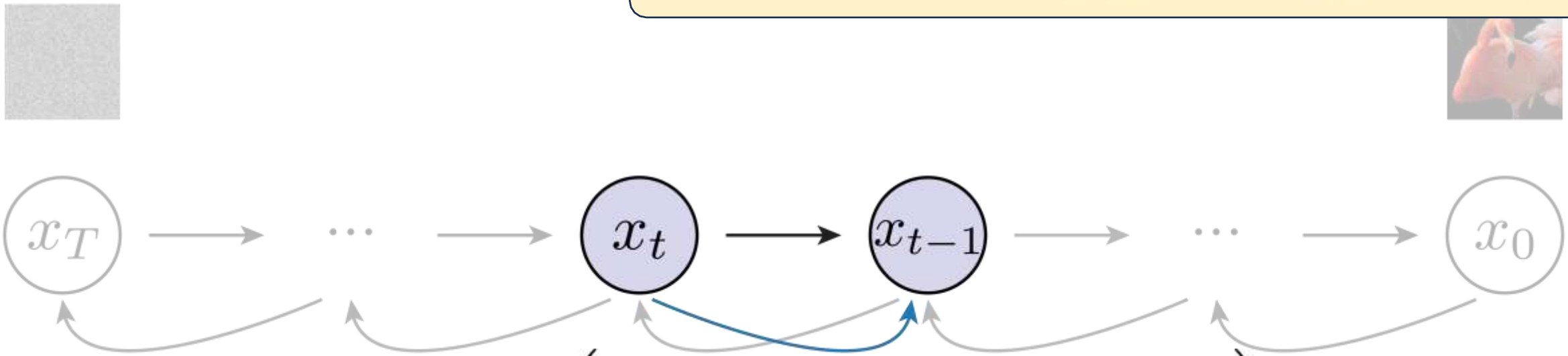
$$\frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)$$

"noise"

Reverse Process

D_{KL} of two Gaussians is like L2 loss:

$$D_{KL}(\mathcal{N}_1 \parallel \mathcal{N}_2) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + \|\mu_1 - \mu_2\|^2}{2\sigma_2^2} - \frac{1}{2}$$



$$D_{\mathrm{KL}}\left(q(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t)\right)$$

$$\mathcal{N}(x_{t-1} \mid \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I})$$

$$\mathcal{N}(x_{t-1} \mid \mu_\theta(x_t), \sigma_t^2 \mathbf{I})$$

learn

preset

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

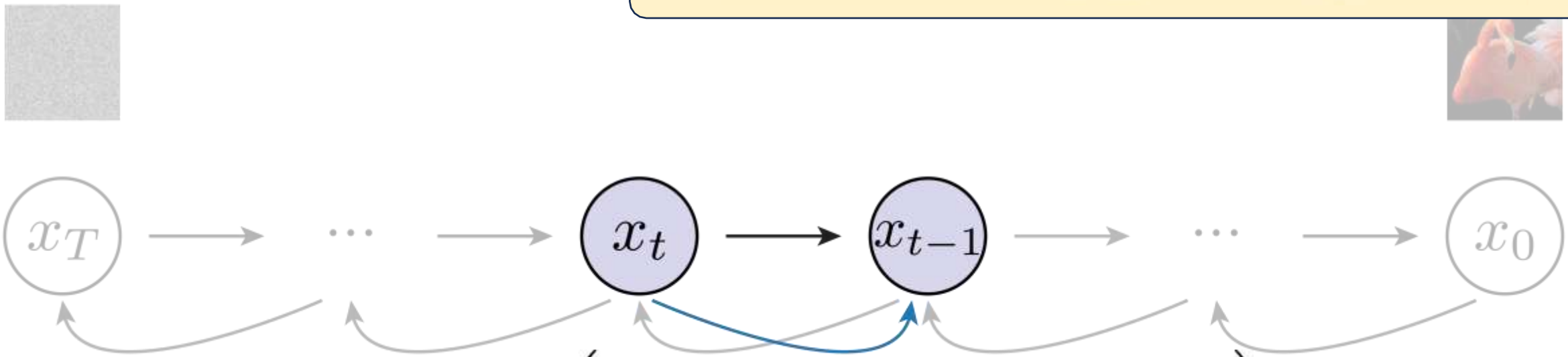
"noise"

estimate noise

Reverse Process

D_{KL} of two Gaussians is like L2 loss:

$$D_{KL}(\mathcal{N}_1 \parallel \mathcal{N}_2) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + \|\mu_1 - \mu_2\|^2}{2\sigma_2^2} - \frac{1}{2}$$



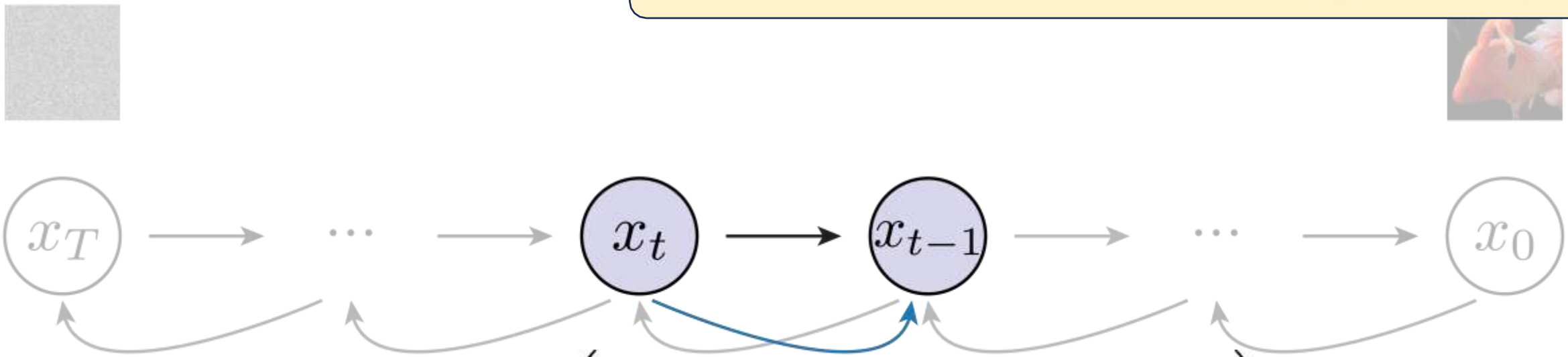
$$D_{\mathrm{KL}}\left(q(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t)\right)$$

$$w_t \|\epsilon - \epsilon_\theta(x_t)\|^2$$

Reverse Process

D_{KL} of two Gaussians is like L2 loss:

$$D_{KL}(\mathcal{N}_1 \parallel \mathcal{N}_2) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + \|\mu_1 - \mu_2\|^2}{2\sigma_2^2} - \frac{1}{2}$$



$$D_{\mathrm{KL}}\left(q(x_{t-1} \mid x_t, x_0) \parallel p_\theta(x_{t-1} \mid x_t)\right)$$

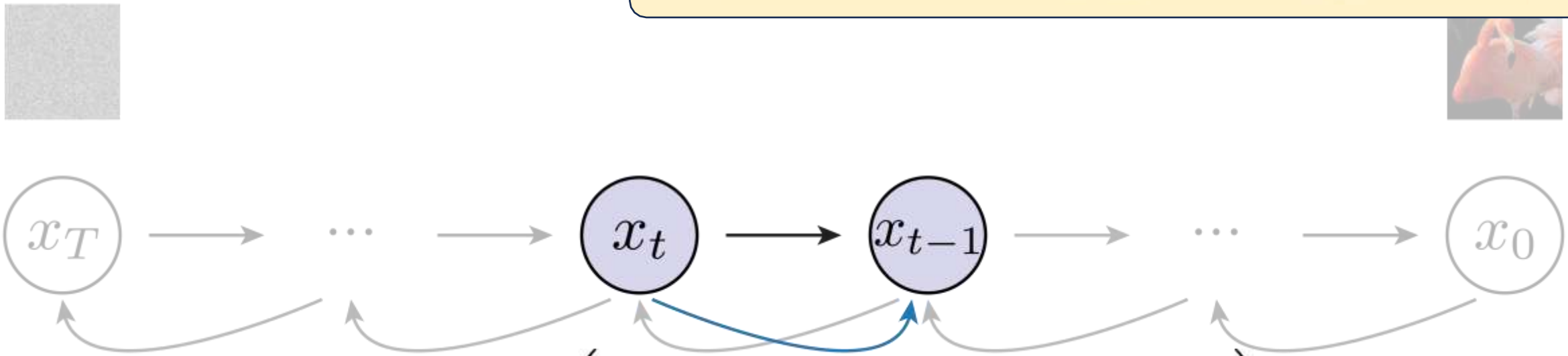
$$w_t \|\epsilon - \epsilon_\theta(x_t)\|^2$$

- a network to predict noise
- input: noisy image

Reverse Process

D_{KL} of two Gaussians is like L2 loss:

$$D_{KL}(\mathcal{N}_1 \parallel \mathcal{N}_2) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + \|\mu_1 - \mu_2\|^2}{2\sigma_2^2} - \frac{1}{2}$$



$$D_{\mathrm{KL}}\left(q(x_{t-1} \mid x_t, x_0) \parallel p_\theta(x_{t-1} \mid x_t)\right)$$

$$w_t \left\| \epsilon - \epsilon_\theta(x_t) \right\|^2$$

- weights due to α_t , β_t
- but set as 1 **(critical)**

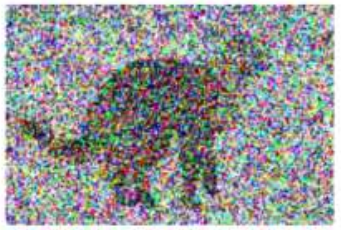
- a network to predict noise
- input: noisy image

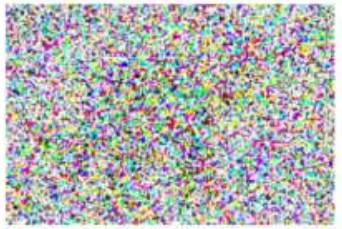
$$\mathrm{E}_{q(x_1|x_0)}[logP(x_0|x_1)] - KL\left(q(x_T|x_0)||P(x_T)\right)$$

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

Sample x_0

Sample x_t



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

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

$$\mathrm{E}_{q(x_1|x_0)}[logP(x_0|x_1)] - KL\left(q(x_T|x_0)||P(x_T)\right)$$

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

x_0

Sample x_t

x_t

t

Denoise

?

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

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

x_0

x_t

Sample x_t

t

Denoise → ? ⇢

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

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

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

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

$$= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t \dfrac{x_t - \sqrt{1-\bar{\alpha}_t}\varepsilon}{\sqrt{\bar{\alpha}_t}} + \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t}{1-\bar{\alpha}_t}$$

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

x_0

x_t

Sample x_t

t

Denoise → ? ⇢ $\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon\right)$

To predict

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

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

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

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\quad \mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

tl; dr

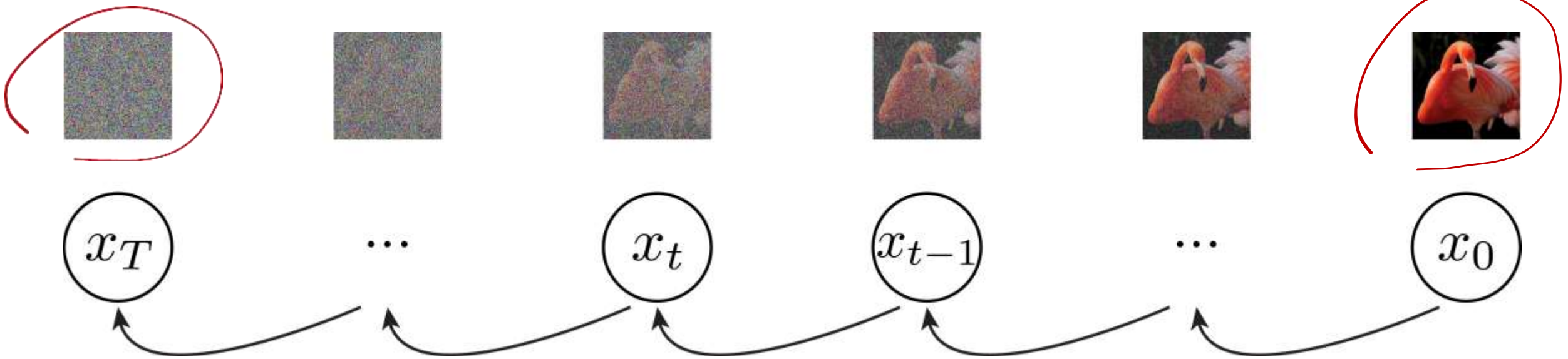
- some dependency graphs
- some linear combinations
- D_{KL}
- L2 loss of noise

Recap.

... in a nutshell

noise

data

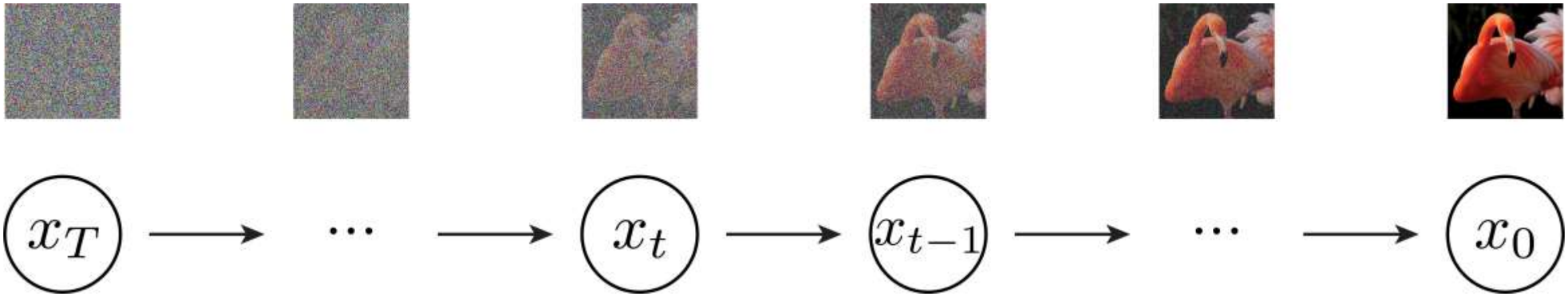


Forward process: add noise

... in a nutshell

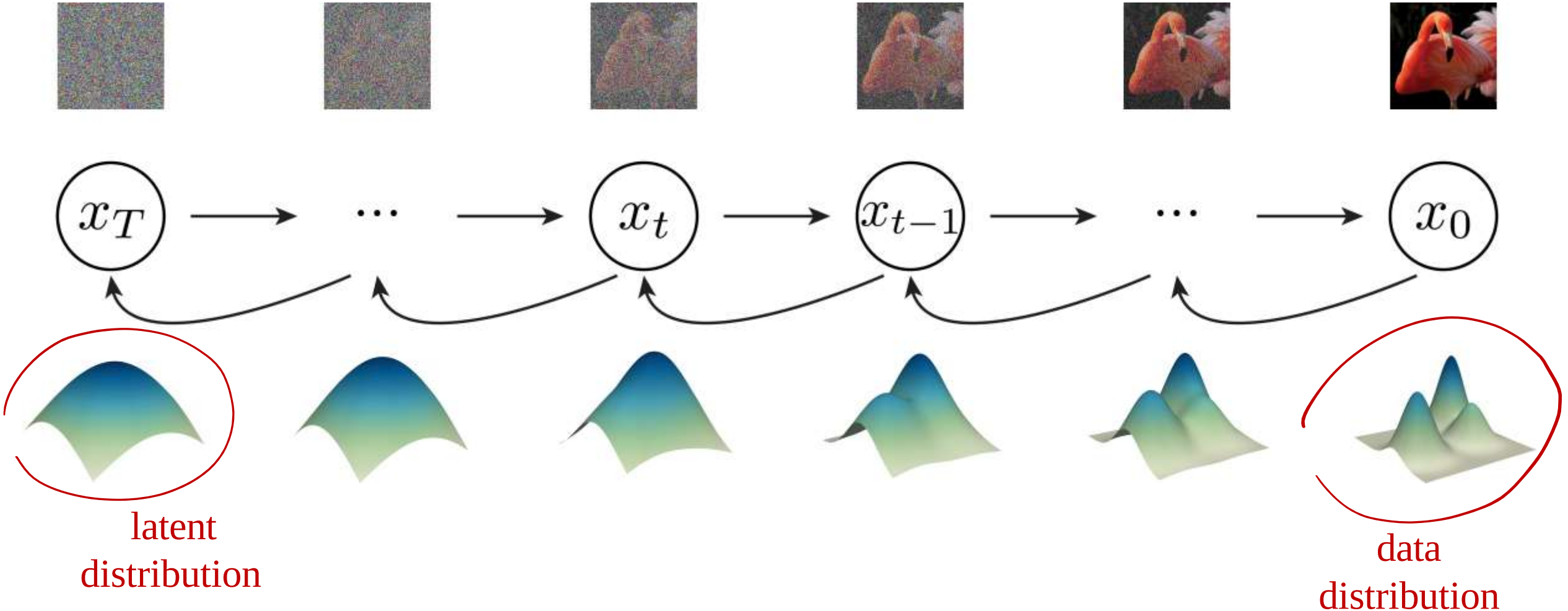
noise

data

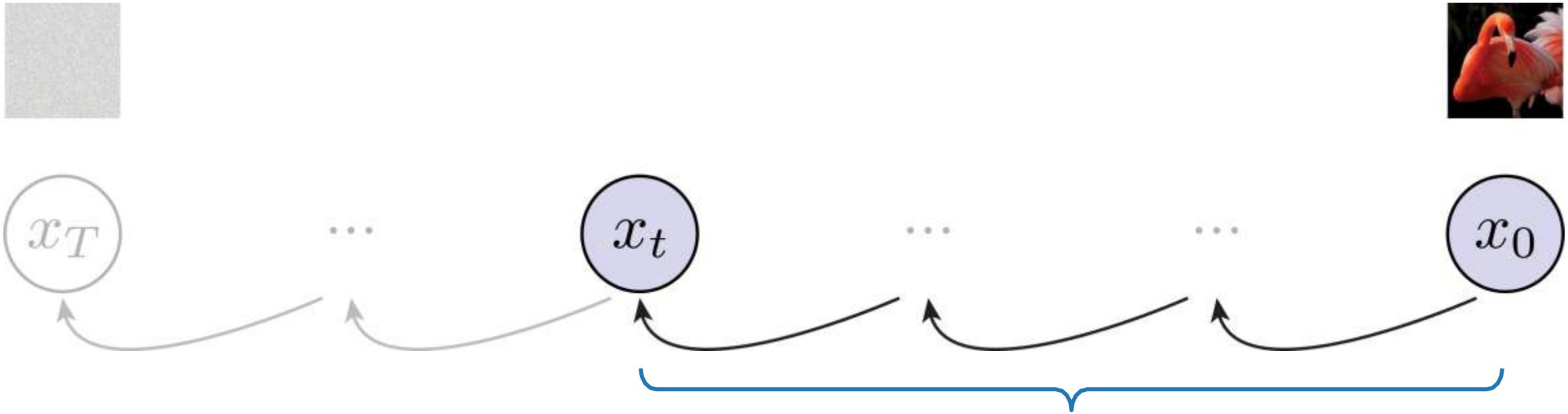


Reverse process: denoise

What is noise?



Forward Process



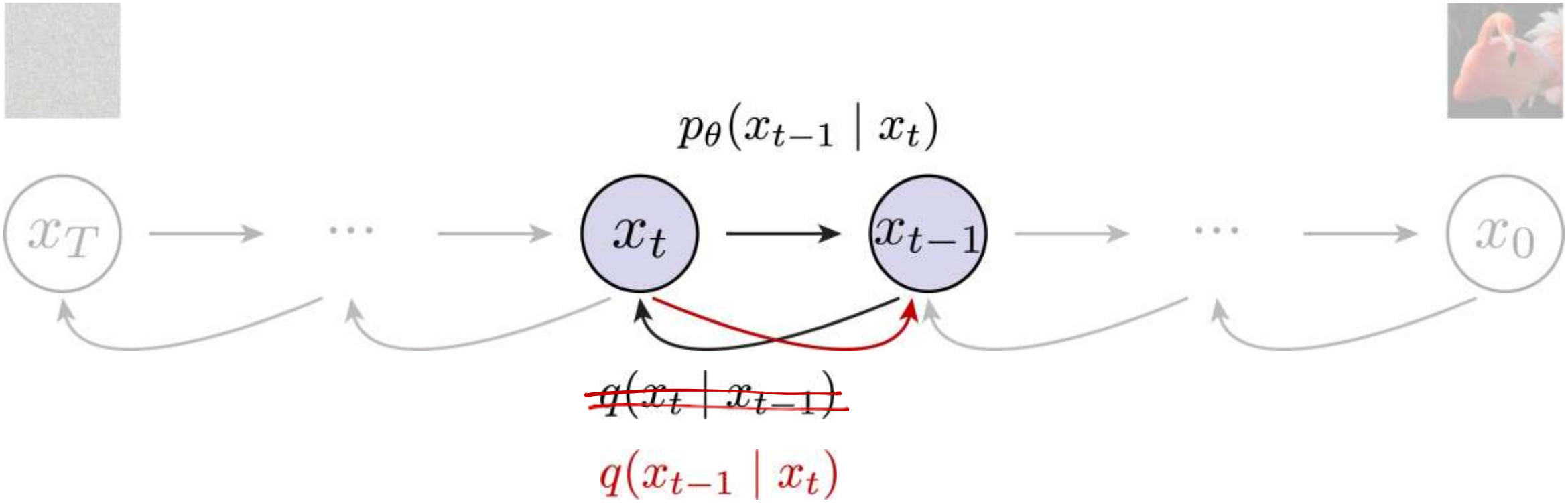
- sampling without simulation
- x_t from x_0 in closed form

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

coefficients given by β

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

Reverse Process



- our target
- but unknown

Reverse Process

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

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

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

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

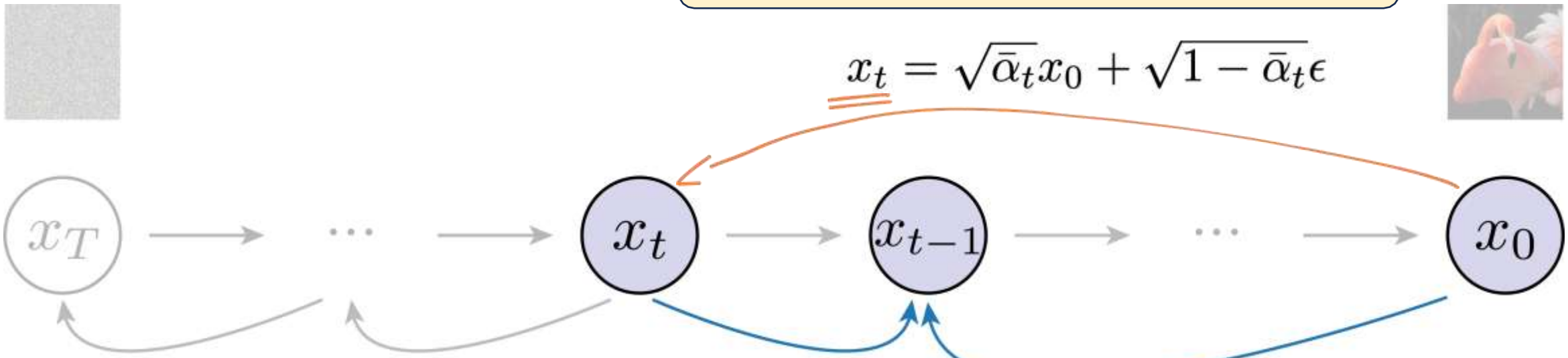
- known
- Gaussian

- known
- Gaussian

- known
- Gaussian

Reverse Process

tl; dr:
- outcome of the dependency graph
- some linear combinations



$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1} \mid \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \mathbf{I})$$

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

mean

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

"noise"

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

var

x_0

x_t

Sample x_t

t

Denoise → ? ⇢ $\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon\right)$

To predict

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

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

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

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\quad \mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Diffusion Models

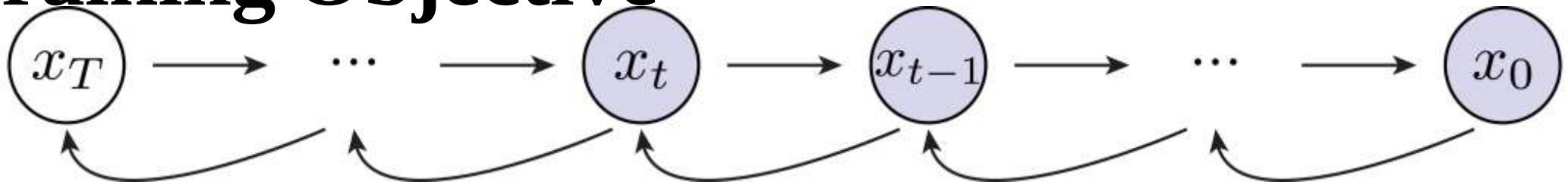
- Forward process
- add noise to data

- Reverse process
- learn to denoise

- Training objective
- from Hierarchical VAE to L2 loss

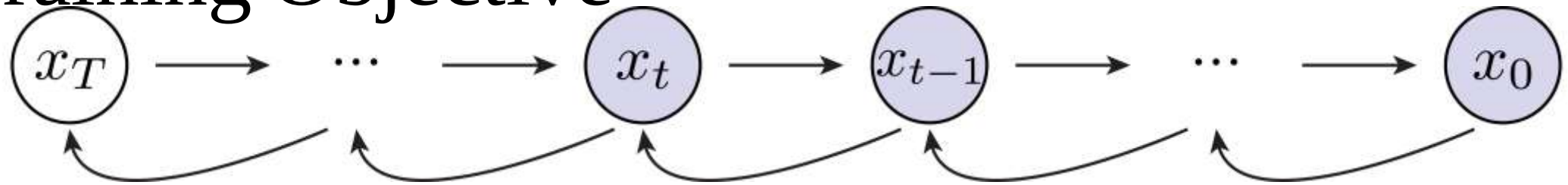
- Noise Conditional Network
- represent a distribution

Training Objective



$$
\begin{aligned}
\mathcal{L}_{\text{VLB}} &:= \mathcal{L}_T + \mathcal{L}_{T-1} + \ldots + \mathcal{L}_0 \\
\mathcal{L}_T &:= D_{\text{KL}}\Big(q(x_T \mid x_0) \,\|\, p_\theta(x_T)\Big) \\
\mathcal{L}_{t-1} &:= D_{\text{KL}}\Big(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)\Big) \\
\mathcal{L}_0 &:= -\log p_\theta(x_0 \mid x_1)
\end{aligned}
$$

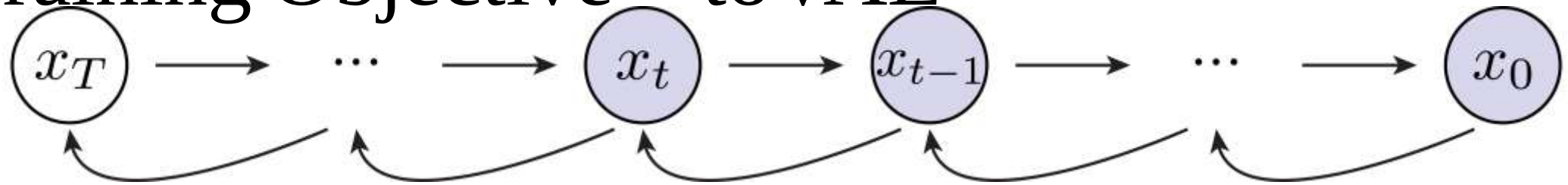
Training Objective



- variational lower bound
- like ELBO

$$
\begin{aligned}
\mathcal{L}_{\text{VLB}} &:= \mathcal{L}_T + \mathcal{L}_{T-1} + \ldots + \mathcal{L}_0 \\
\mathcal{L}_T &:= D_{\text{KL}}\Big(q(x_T \mid x_0) \,\|\, p_\theta(x_T)\Big) \\
\mathcal{L}_{t-1} &:= D_{\text{KL}}\Big(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)\Big) \\
\mathcal{L}_0 &:= -\log p_\theta(x_0 \mid x_1)
\end{aligned}
$$

Training Objective – toVAE



- variational lower bound
- like ELBO

$$\mathcal{L}_{\text{VLB}} := \mathcal{L}_T + \mathcal{L}_{T-1} + \ldots + \mathcal{L}_0$$

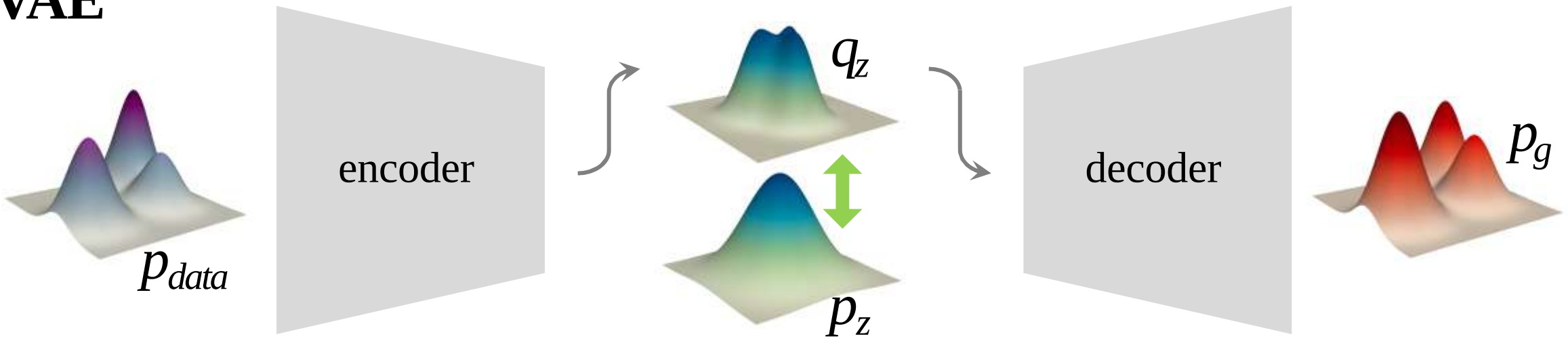
$$\mathcal{L}_T := D_{\text{KL}}\left(q(x_T \mid x_0) \,\|\, p_\theta(x_T)\right)$$

it's ELBO if one step

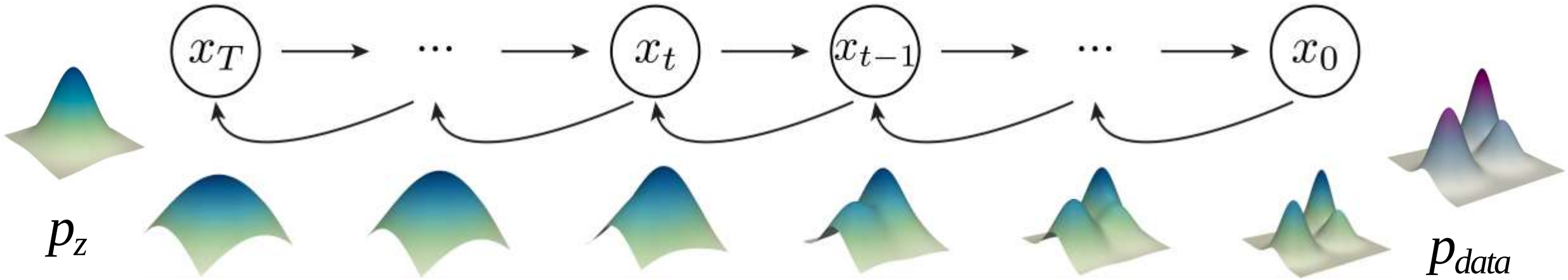
$$\mathcal{L}_{t-1} := D_{\text{KL}}\left(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)\right)$$

$$\mathcal{L}_0 := -\log p_\theta(x_0 \mid x_1)$$

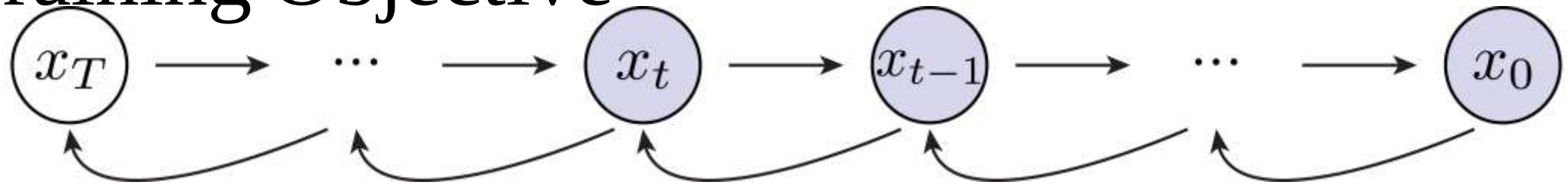
VAE



DM/Flow



Training Objective



$$\mathcal{L}_{\text{VLB}} := \mathcal{L}_T + \mathcal{L}_{T-1} + \ldots + \mathcal{L}_0$$

no parameter, unlike VAE's q_φ

$$\mathcal{L}_T := D_{\text{KL}}\Big(q(x_T \mid x_0) \,\|\, p_\theta(x_T)\Big)$$

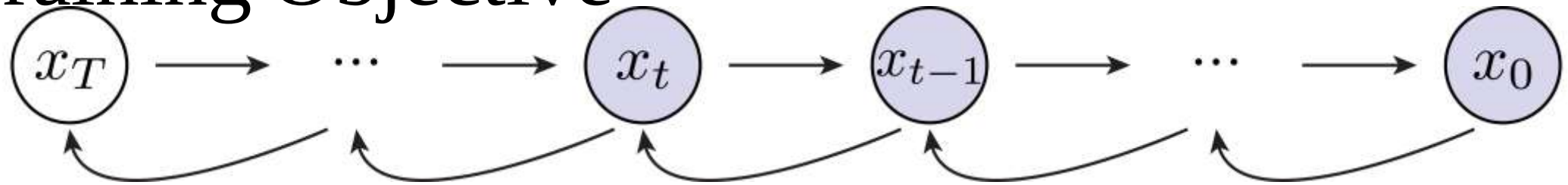
constant

Gaussian

$$\mathcal{L}_{t-1} := D_{\text{KL}}\Big(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)\Big)$$

$$\mathcal{L}_0 := -\log p_\theta(x_0 \mid x_1)$$

Training Objective



$$\mathcal{L}_{\text{VLB}} := \mathcal{L}_T + \mathcal{L}_{T-1} + \ldots + \mathcal{L}_0$$

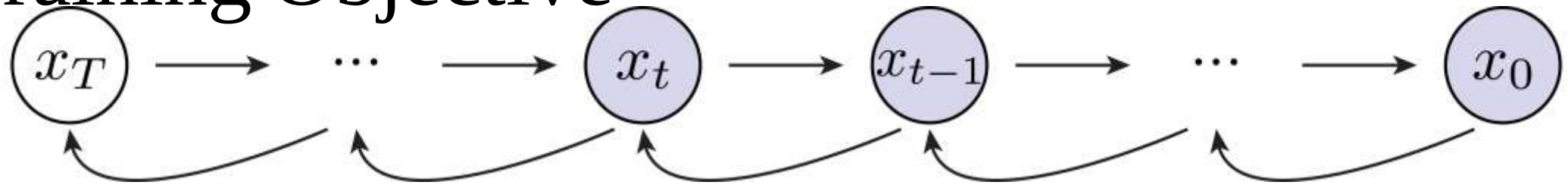
$$\mathcal{L}_T := D_{\text{KL}}\left(q(x_T \mid x_0) \,||\, p_\theta(x_T)\right)$$

$$\mathcal{L}_{t-1} := D_{\text{KL}}\left(q(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t)\right)$$

reconstruction loss, like VAE

$$\mathcal{L}_0 := -\log p_\theta(x_0 \mid x_1)$$

Training Objective



$$\mathcal{L}_{\text{VLB}} := \mathcal{L}_T + \mathcal{L}_{T-1} + \ldots + \mathcal{L}_0$$

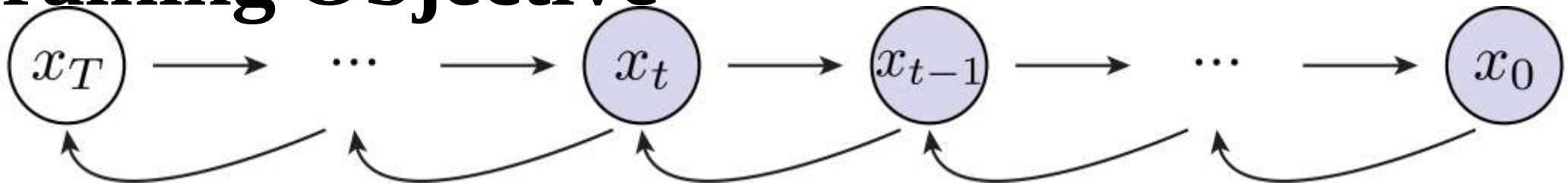
$$\mathcal{L}_T := D_{\text{KL}}\left(q(x_T \mid x_0) \,||\, p_\theta(x_T)\right)$$

L2 loss on noise

$$\mathcal{L}_{t-1} := D_{\text{KL}}\left(q(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t)\right)$$

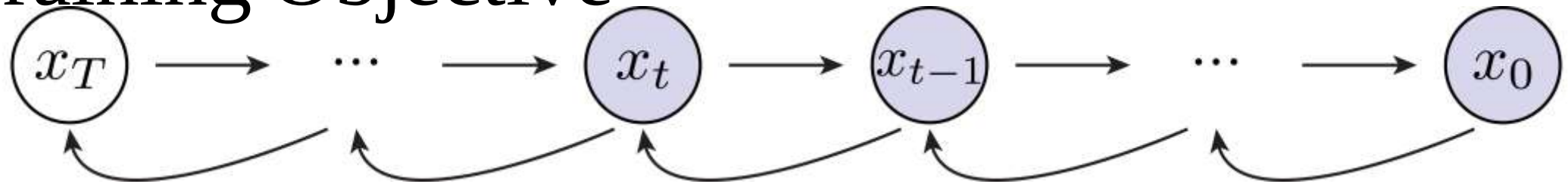
$$\mathcal{L}_0 := -\log p_\theta(x_0 \mid x_1)$$

Training Objective



$$\mathcal{L} = \mathbb{E}_{x_0,t,\epsilon}\left[w_t \|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]$$

Training Objective



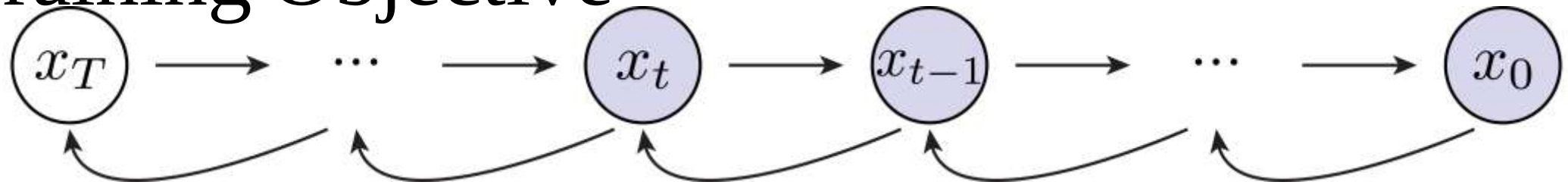
$$\mathcal{L} = \mathbb{E}_{x_0, t, \epsilon}\left[w_t \|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]$$

over p_{data}

over $[1, T]$

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

Training Objective



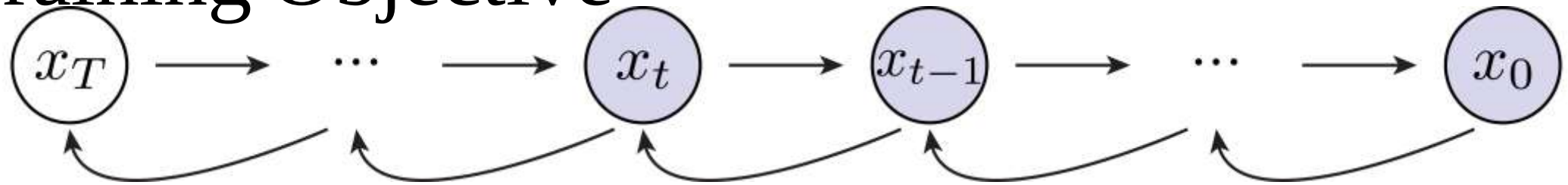
$$\mathcal{L} = \mathbb{E}_{x_0,t,\epsilon}\left[w_t \|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]$$

set as 1 **(critical)**

Objective	IS	FID
L, learned diagonal $\boldsymbol{\Sigma}$	–	–
L, fixed isotropic $\boldsymbol{\Sigma}$	7.67 ± 0.13	13.51
$\|\tilde{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_\theta\|^2$ (L_{simple})	$\mathbf{9.46 \pm 0.11}$	**3.17**

[Ho et al. 2020]; see more in [Salimans & Ho, 2022]

Training Objective



$$\mathcal{L} = \mathbb{E}_{x_0,t,\epsilon}\left[w_t \|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]$$

network to predict noise

conditioned on noise level **(critical)**

Diffusion Models

- Forward process
- add noise to data

- Reverse process
- learn to denoise

- Training objective
- from Hierarchical VAE to L2 loss

- Noise Conditional Network
- represent a distribution

Noise Conditional Network

- Diffusion models decompose a distribution into **many** simpler ones.
- We need the same # networks to fit **all** of them.
- We can **combine** all into one “powerful” network.
- This network is conditioned on noise level *t*.

- **Noise Conditional Network** [Song & Ermon 2019]: things made work

*It is called Noise Conditional Score Network (NCSN) in [Song & Ermon 2019] in the context of score matching.

Noise Conditional Network

How to represent $p_\theta(x_{t-1} \mid x_t)$

- network input: x_t
- network output: μ and σ of a distribution

- parametrize μ by: $\epsilon_\theta(x_t, t)$

noisy image:
- condition
- network input

noise level:
- condition
- network input

Noise Conditional Network

x_t

network input

t

also network input

$\epsilon_\theta(x_t, t)$

Unet, ViT, ...
(behaves like Denoising AE)

ϵ

output: same size

Diffusion algorithm annotated:

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

x_t

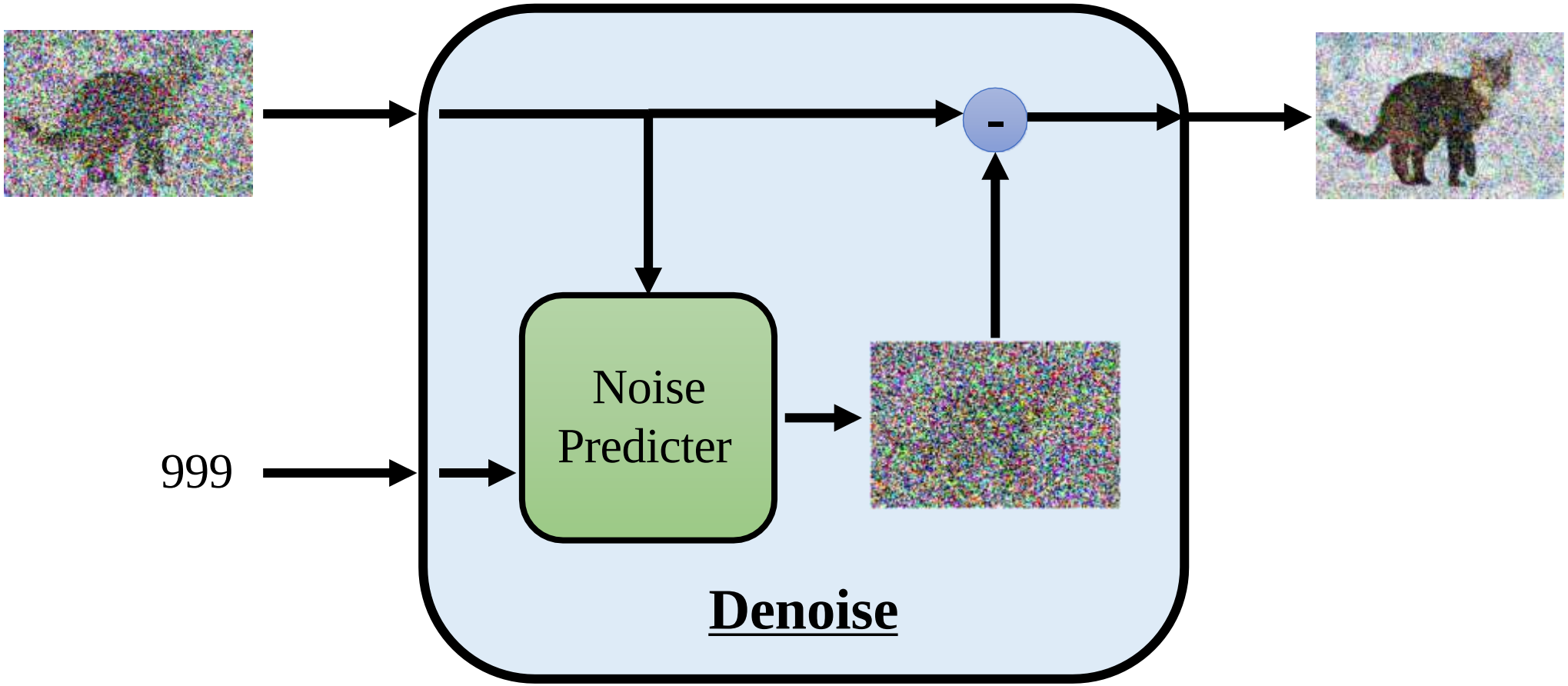
Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

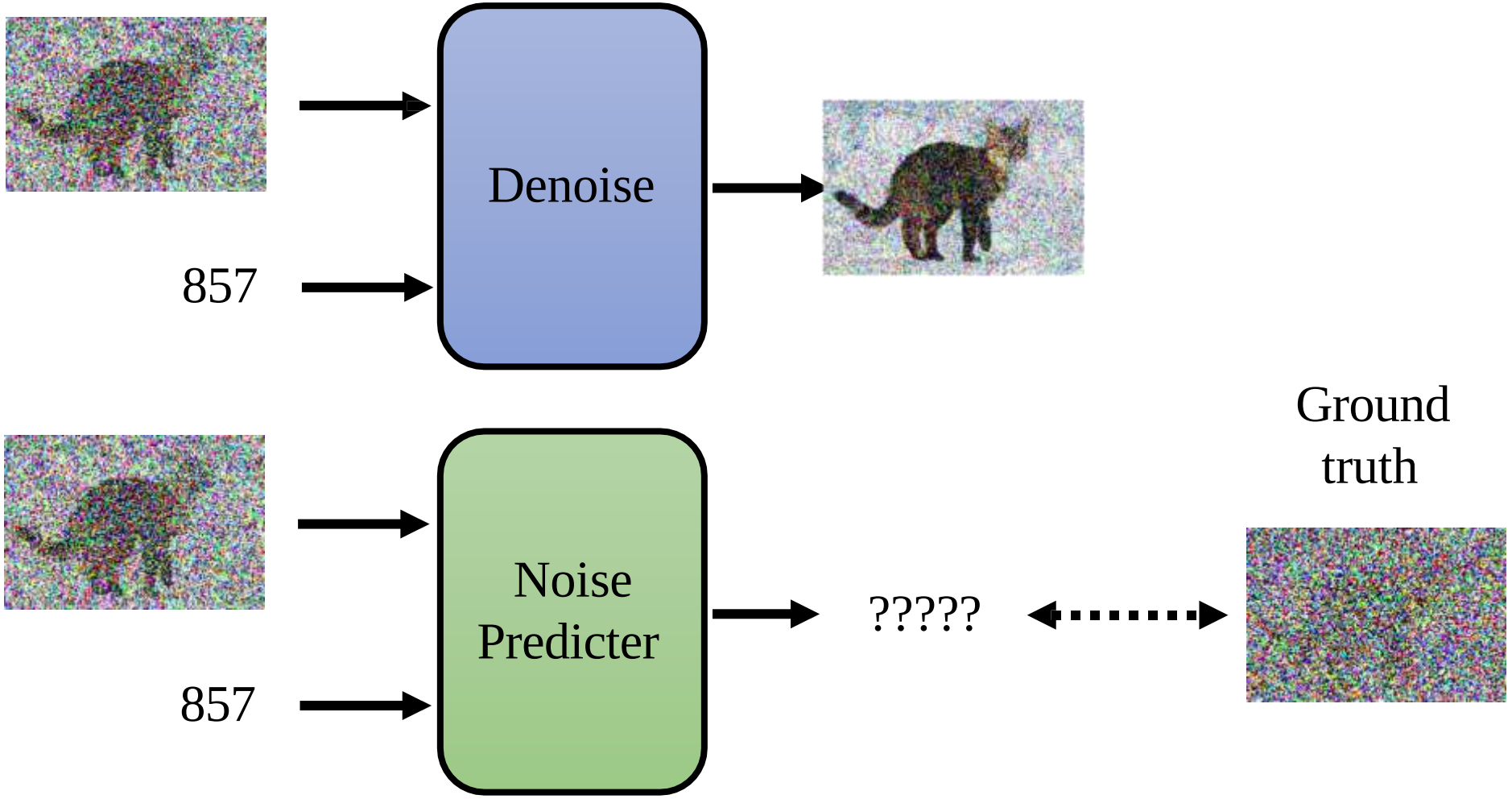
estimated μ

sampling from estimated distribution

Noise Conditional Network



Noise Conditional Network



Diffusion algorithm annotated:

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

tl; dr: noising and denoising

- Turns out to be extremely simple
- Being “simple and effective” moves the needle

Example: Unconditional Generation on CIFAR-10

noise
x_T

generated
x_0

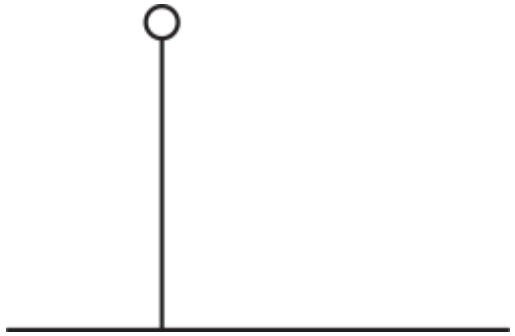
[Ho et al, 2020]

Example: shared intermediate latents

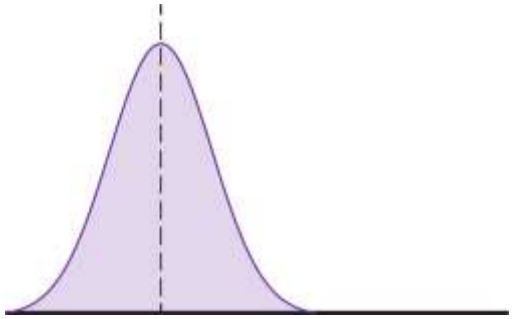
[Ho et al, 2020]

Recap.

- Adding Gaussian noise $\Leftrightarrow$ sampling $x \sim \mathcal{N}(x \mid x_0, \sigma)$

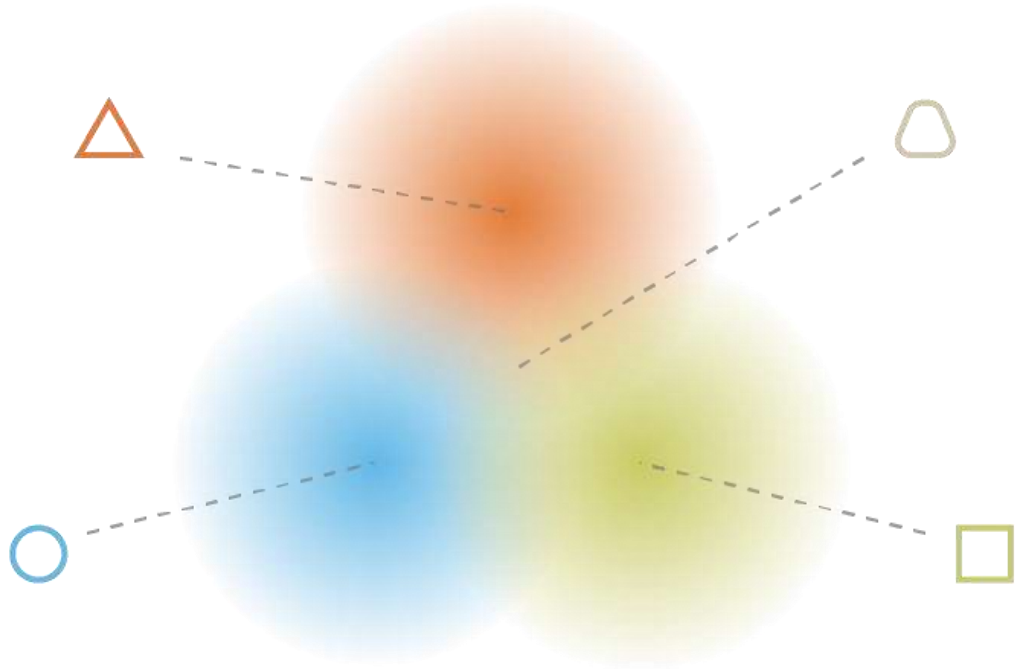


$p(x) = \delta(x - x_0)$



$p(x) = \mathcal{N}(x \mid x_0, \sigma)$

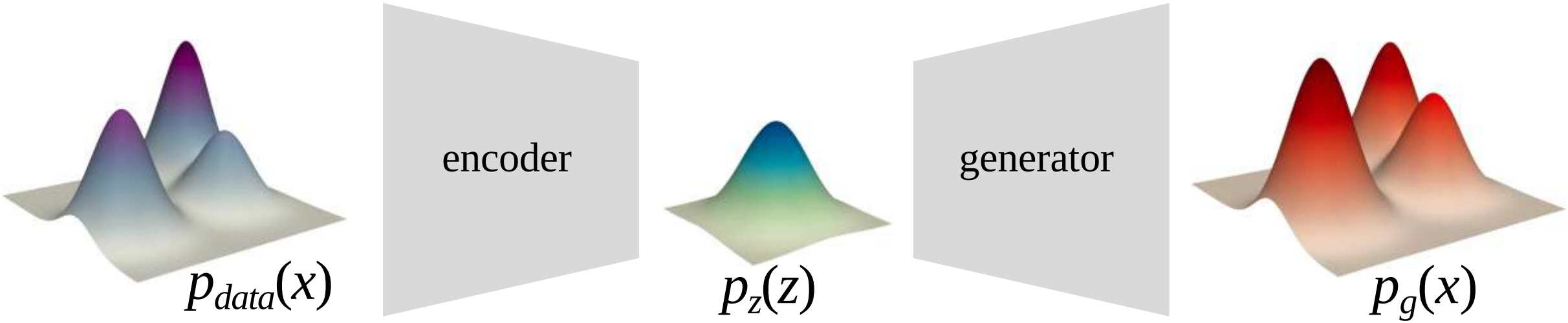
Problems



Recap: Variational Autoencoder (VAE)

Autoencoding distributions:

"Encoding" data distribution p_{data} into latent distribution p_z



"Reconstruction" loss

Summary

Forward process

- add noise to data

Reverse process

- learn to denoise

Training objective

- from Hierarchical VAE to L2 loss

Noise Conditional Network

- represent distributions