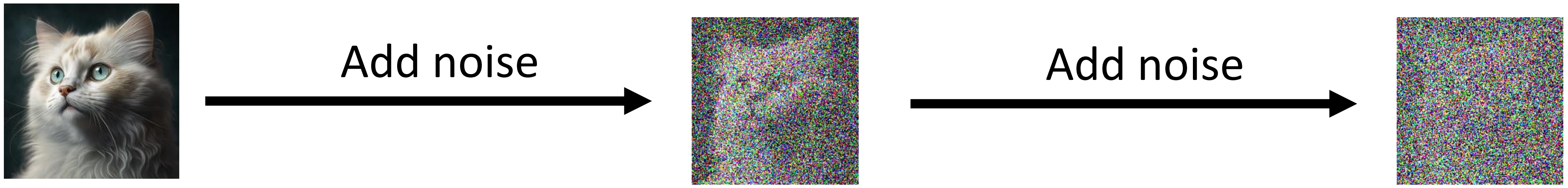
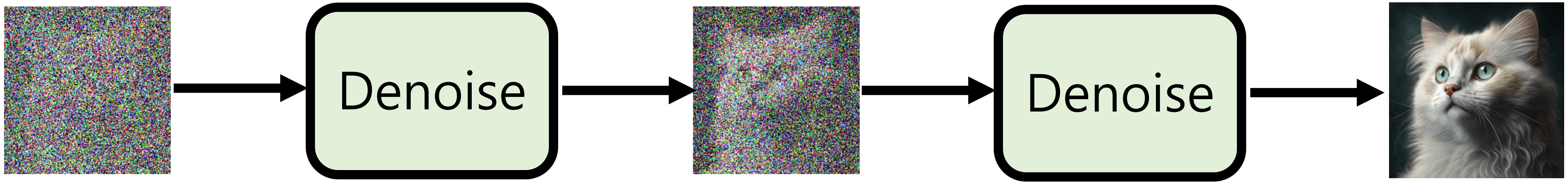
Diffusion Model
背後的數學原理

基本概念

Forward Process



Reverse Process



Denoising Diffusion Probabilistic Models

Algorithm 1 Training

1: **repeat**
2: $\quad \mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $\quad t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
4: $\quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: $\quad$ 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** $t = T, \ldots, 1$ **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** $\mathbf{x}_0$

暗藏玄機!

Training

x_0: clean image

ε: noise

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ ← sample clean image
3: $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ← sample a noise
5: Take gradient descent step on

$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \underbrace{\boldsymbol{\epsilon}_\theta(\underbrace{\boxed{\sqrt{\bar\alpha_t}\mathbf{x}_0 + \sqrt{1-\bar\alpha_t}\boldsymbol{\epsilon}}}_{\text{Noisy image}}, t)}_{\text{Noise predictor}} \right\|^2$$

6: **until** converged

$\bar{\alpha}_1, \bar{\alpha}_2, \ldots \bar{\alpha}_T$

Target Noise

Training

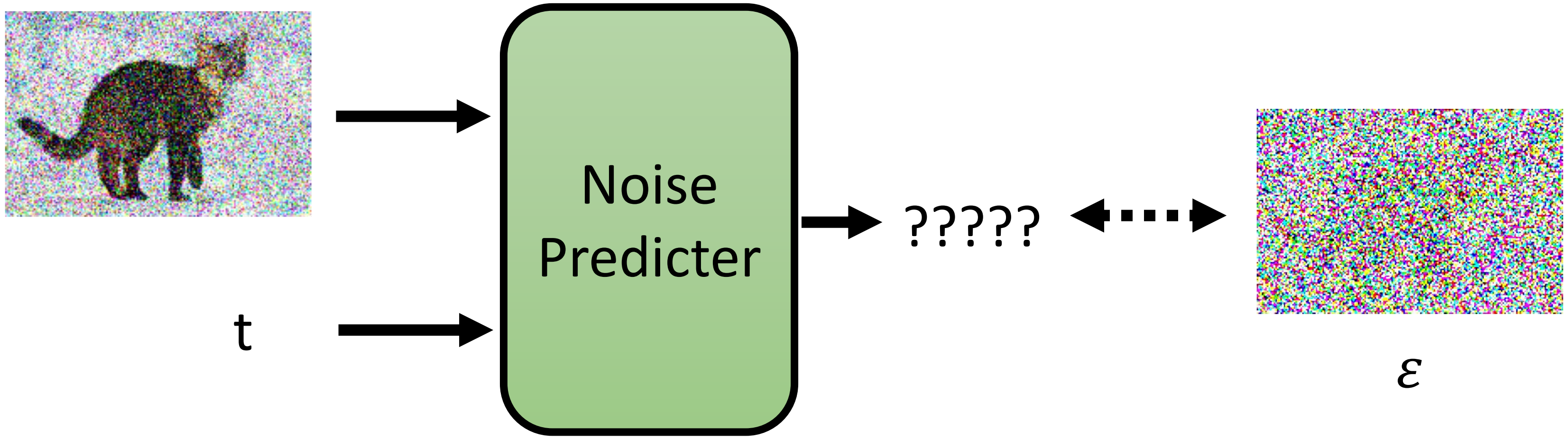
$\bar{\alpha}_1, \bar{\alpha}_2, \ldots \bar{\alpha}_T$

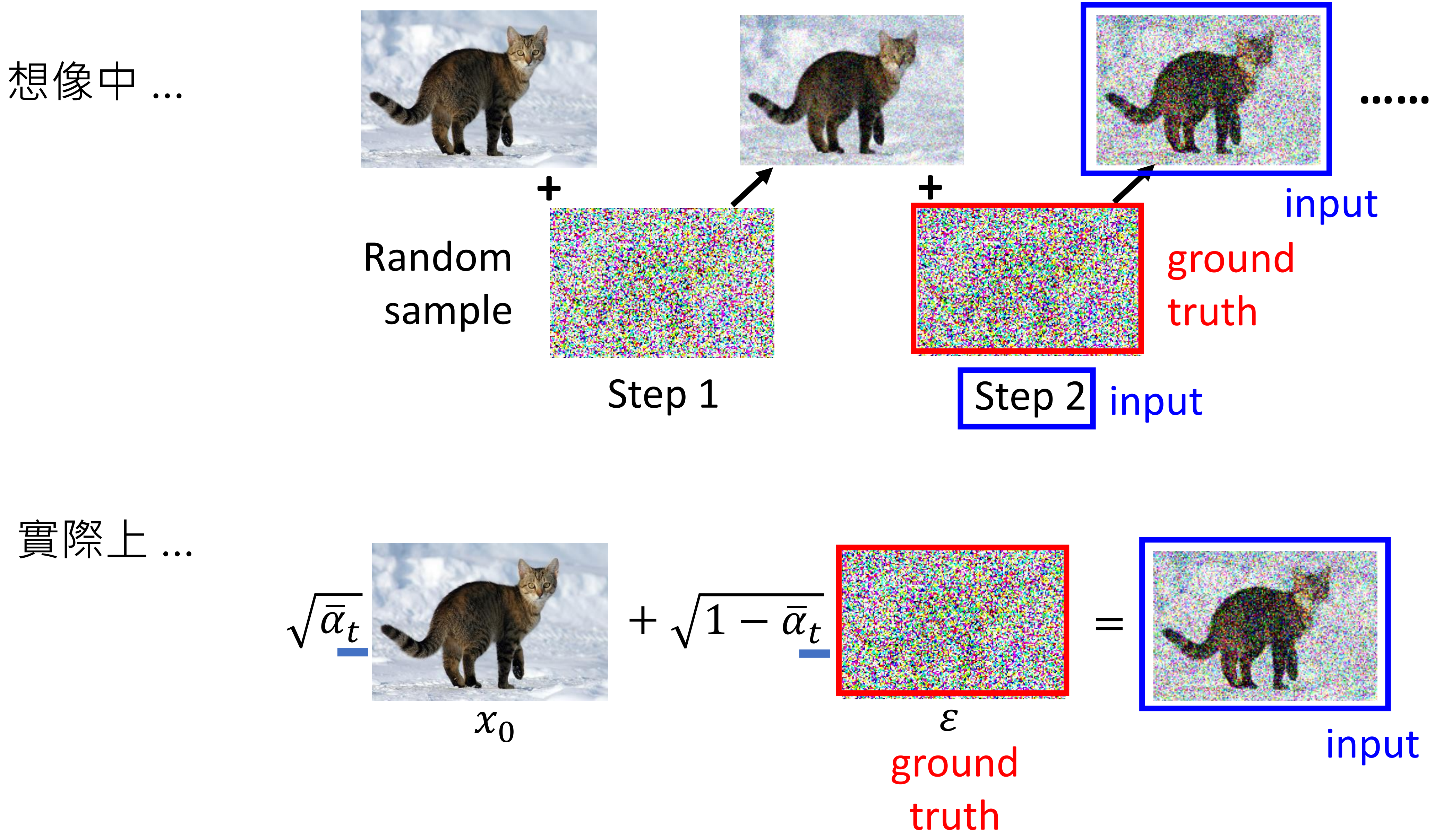
 

x_0 ε

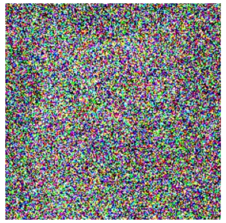
$\sqrt{\bar{\alpha}_t}$  $+ \sqrt{1-\bar{\alpha}_t}$  $=$ 

x_0 ε




想像中 ...
Random sample
Step 1
Step 2
input
ground truth
input
......
實際上 ...
$\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon =$
x_0
ε
ground truth
input

Inference

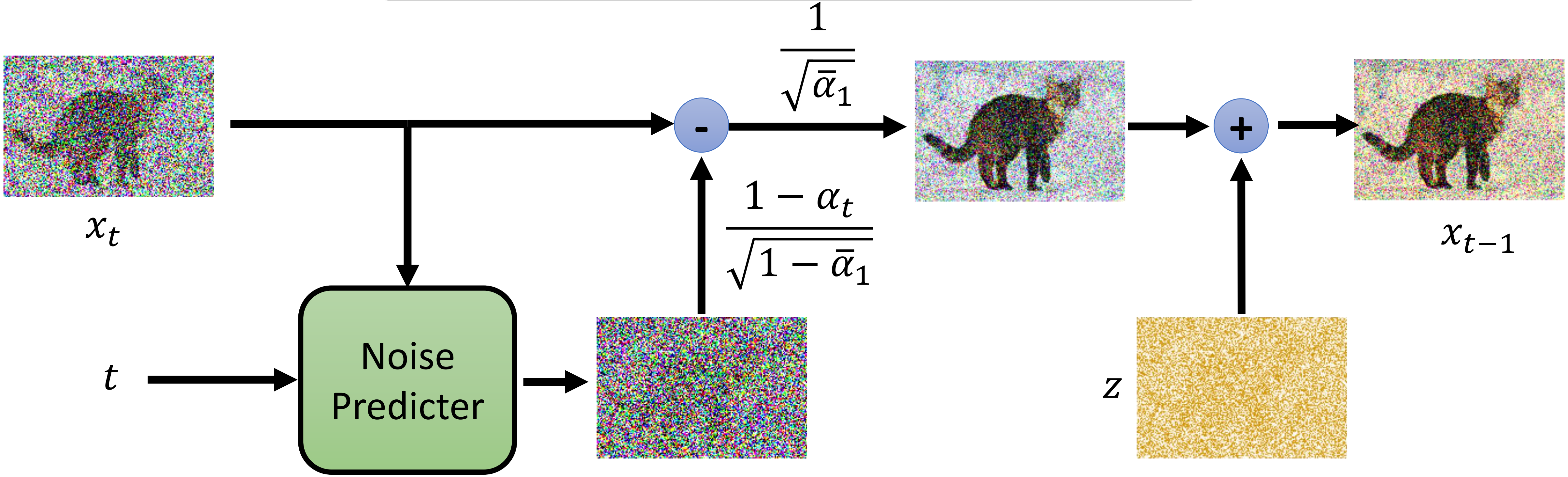


x_T

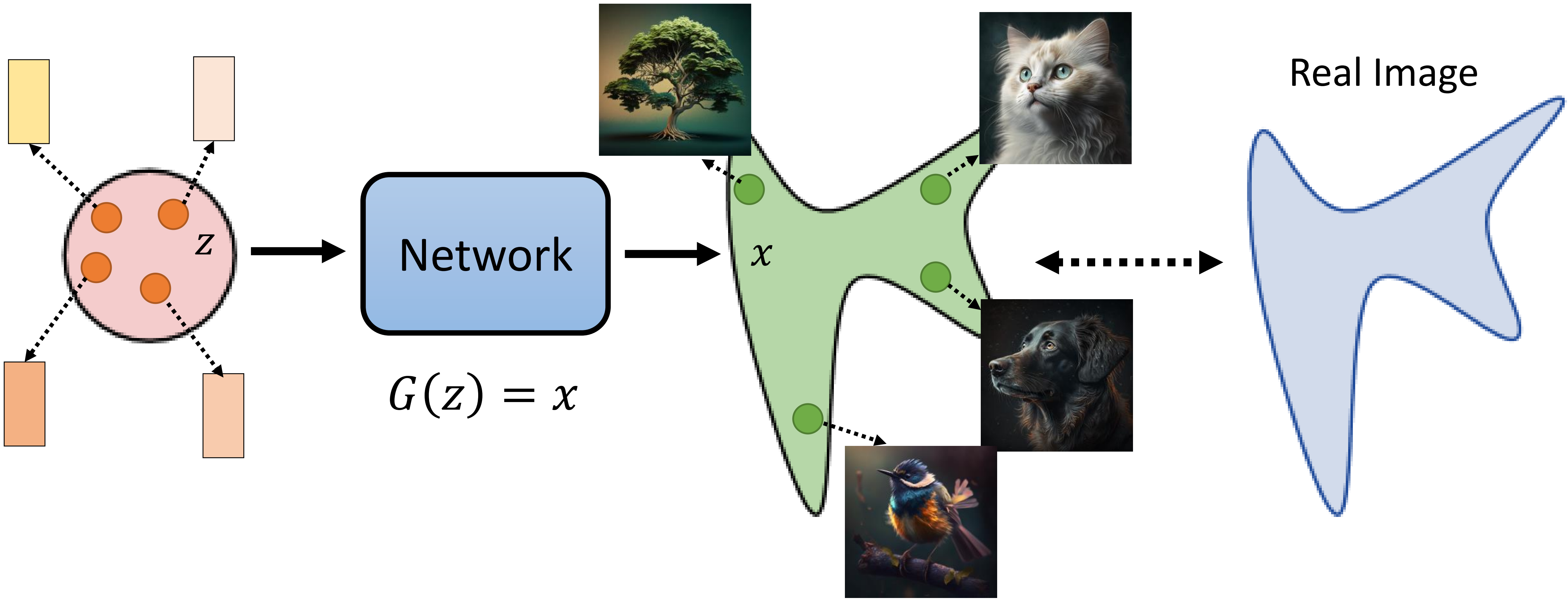
Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\quad \boxed{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}$ 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** $\mathbf{x}_0$

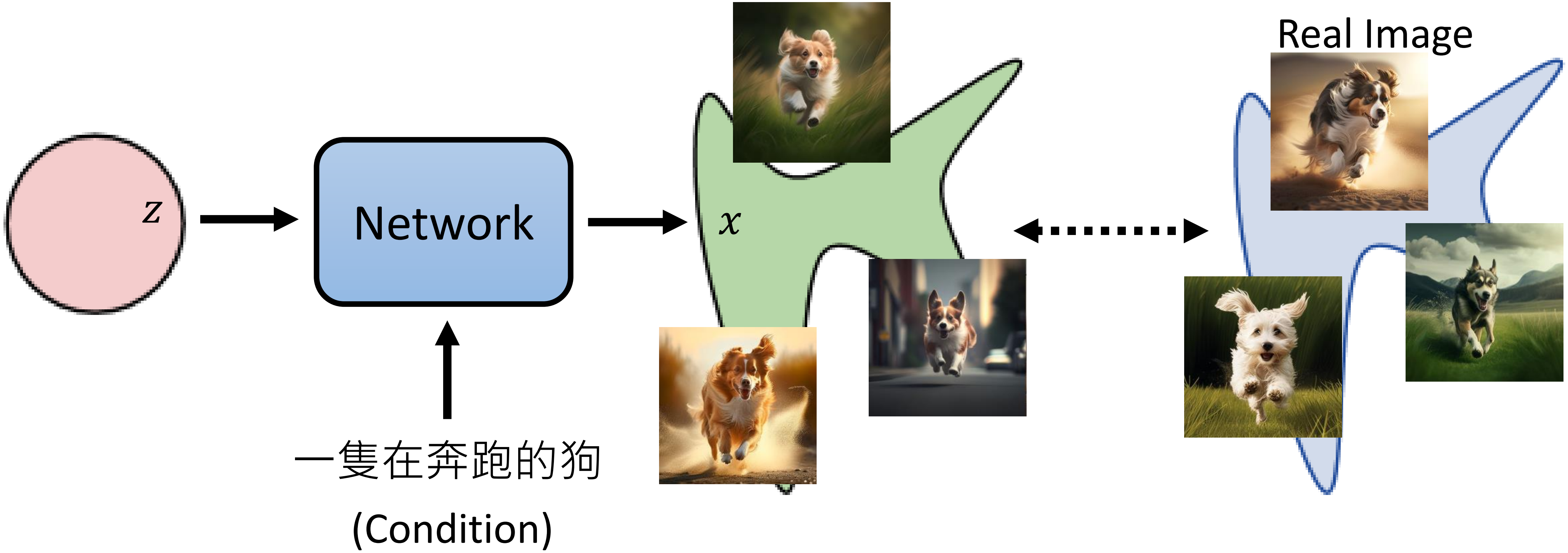
sample a noise?!



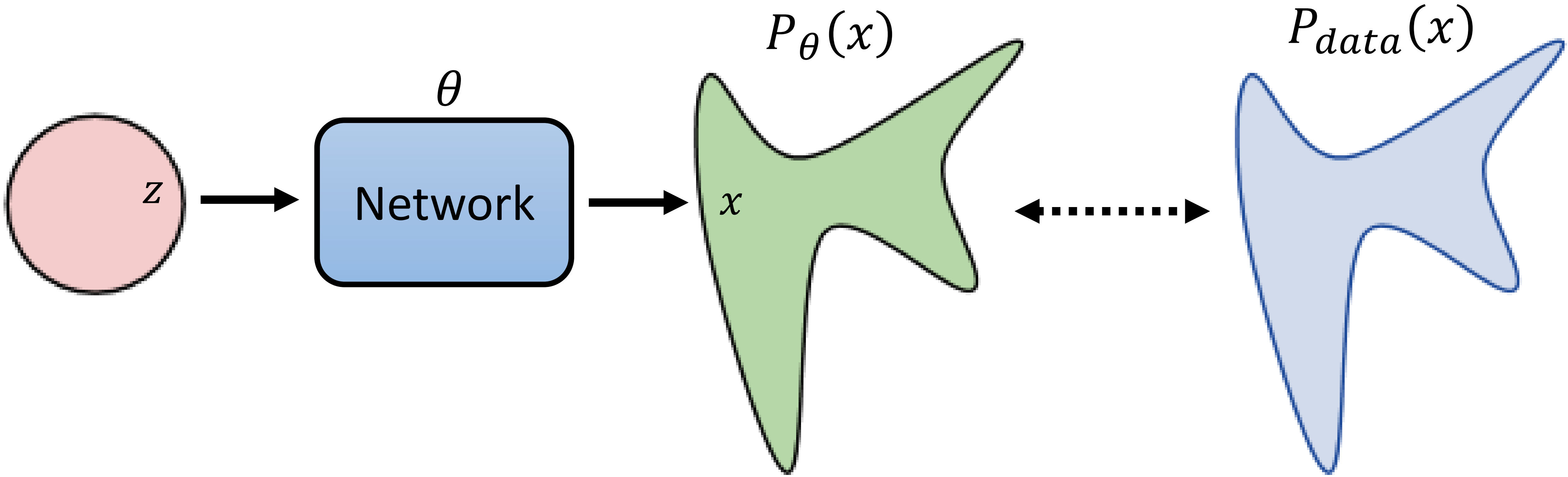
影像生成模型本質上的共同目標



影像生成模型本質上的共同目標



Maximum Likelihood Estimation



Sample $\{x^1, x^2, \dots, x^m\}$ from $P_{data}(x)$

We can compute $P_\theta(x^i)$

???

$$\theta^* = arg \max_\theta \prod_{i=1}^{m} P_\theta(x^i)$$

Sample $\{x^1, x^2, \ldots, x^m\}$ from $P_{data}(x)$

$$\theta^* = arg\max_\theta \prod_{i=1}^{m} P_\theta(x^i) \quad = arg\max_\theta log \prod_{i=1}^{m} P_\theta(x^i)$$

$$= arg\max_\theta \sum_{i=1}^{m} logP_\theta(x^i) \quad \approx arg\max_\theta E_{x\sim P_{data}}[logP_\theta(x)]$$

(not related to θ)

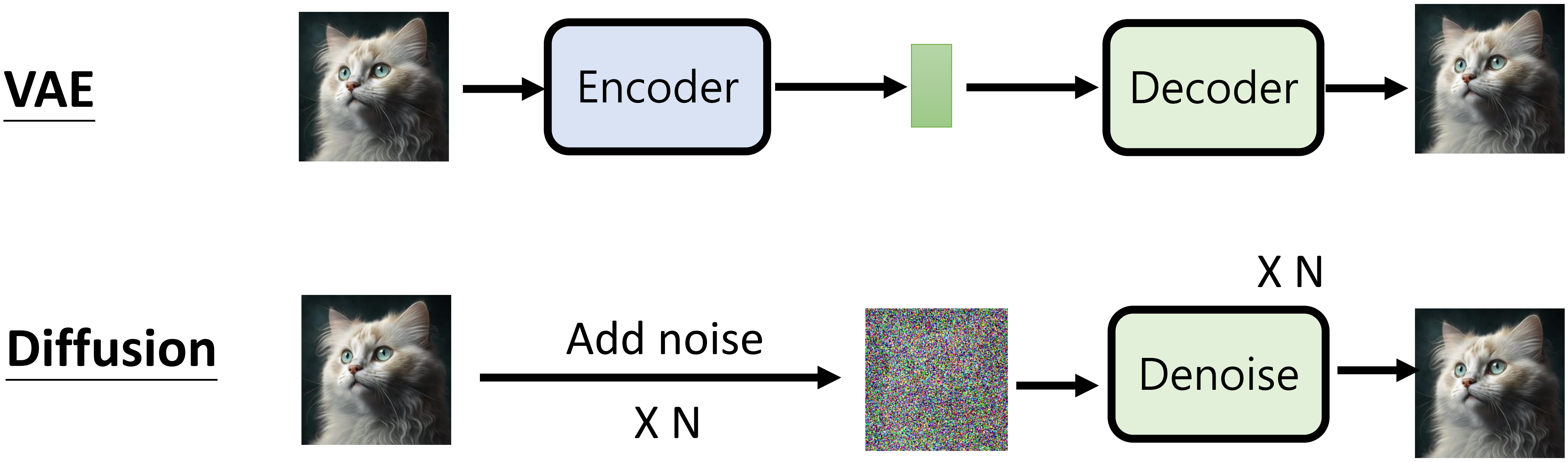
$$= arg\max_\theta \int_x P_{data}(x)logP_\theta(x)dx \quad - \int_x P_{data}(x)logP_{data}(x)dx$$

Difference between P_{data} and P_θ

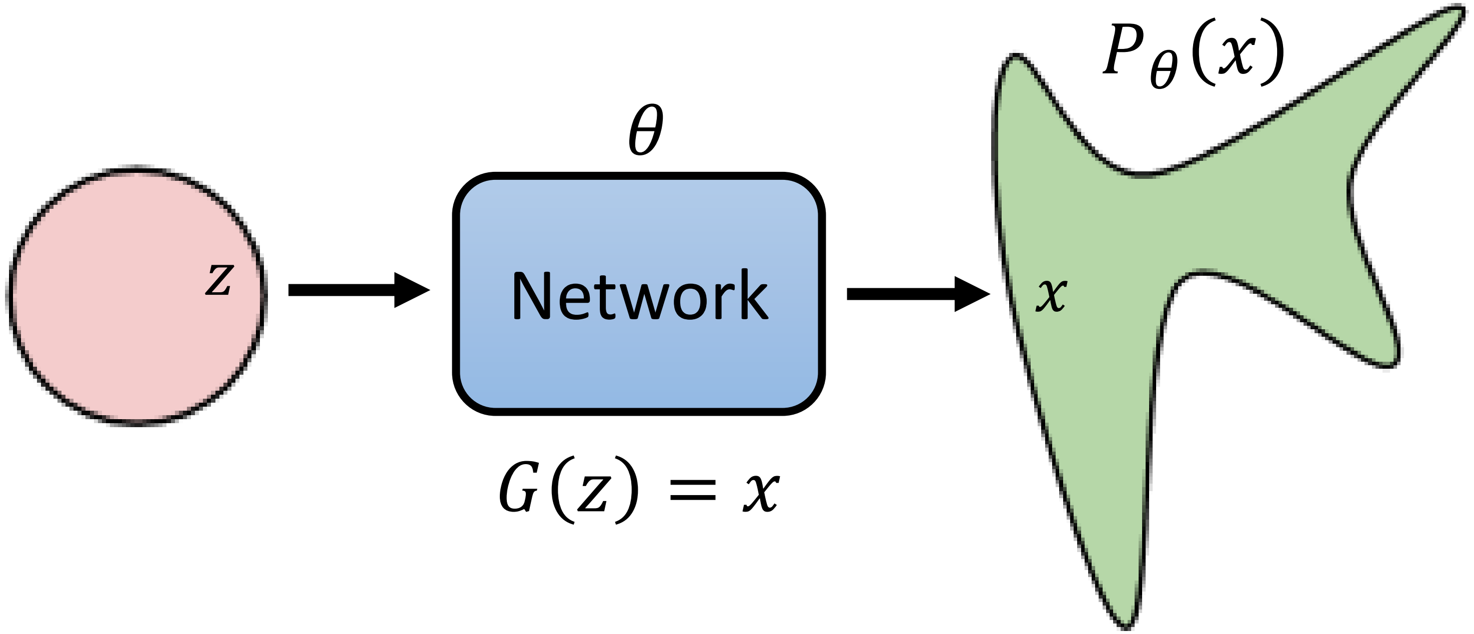
$$= arg\max_\theta \int_x P_{data}(x)log\frac{P_\theta(x)}{P_{data}(x)}dx \quad = arg\min_\theta KL(P_{data}||P_\theta)$$

Maximum Likelihood = Minimize KL Divergence

VAE vs. Diffusion Model



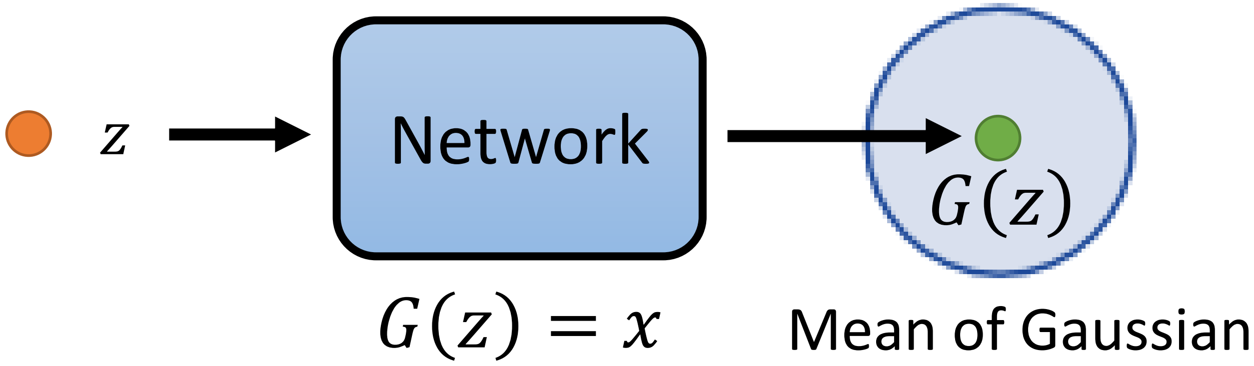
VAE: Compute $P_\theta(x)$



$$P_\theta(x) = \int_z P(z)P_\theta(x|z)dz$$

$$P_\theta(x|z) = \begin{cases} 1, & G(z) = x \\ 0, & G(z) \neq x \end{cases}$$

可能會幾乎都是 0 ☹



$$P_\theta(x|z) \propto exp(\|G(z) - x\|_2)$$

VAE: Lower bound of $logP(x)$

$$logP_\theta(x) = \int_z q(z|x)logP(x)dz$$

$q(z|x)$ can be any distribution

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

$$= \int_z q(z|x)log\left(\frac{P(z,x)}{q(z|x)}\right)dz + \underbrace{\int_z q(z|x)log\left(\frac{q(z|x)}{P(z|x)}\right)dz}_{KL\left(q(z|x)||P(z|x)\right)}$$

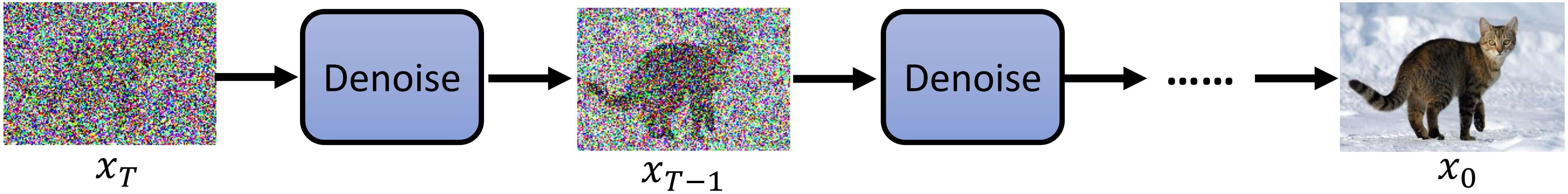
≥ 0

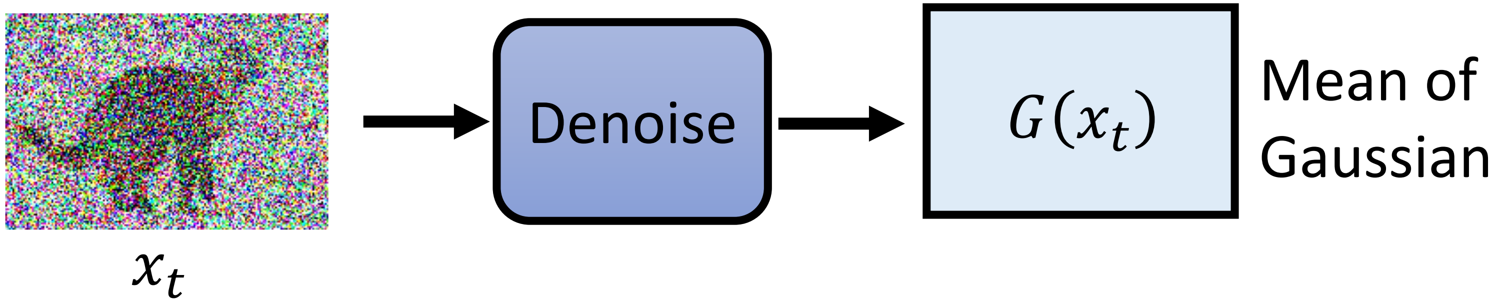
$$\geq \int_z q(z|x)log\left(\frac{P(z,x)}{q(z|x)}\right)dz \;\; = E_{\underline{q(z|x)}}[log\left(\frac{P(x,z)}{q(z|x)}\right)]$$

lower bound

Encoder

DDPM: Compute $P_\theta(x)$





$$P_\theta(x_{t-1}|x_t)$$
$$\propto \exp(\|G(x_t) - x_{t-1}\|_2)$$

$$P_\theta(x_0) = \int_{x_1:x_T} P\,(x_T) P_\theta(x_{T-1}|x_T) \ldots P_\theta(x_{t-1}|x_t) \ldots P_\theta(x_0|x_1) dx_1\!:x_T$$

DDPM: Lower bound of $logP(x)$

VAE

Maximize $logP_\theta(\underline{x})$ $\longrightarrow$ Maximize $\mathrm{E}_{\boxed{q(\underline{z}|x)}}[log\left(\frac{P(\underline{x,z})}{q(\underline{z}|x)}\right)]$

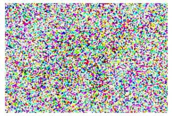
Encoder

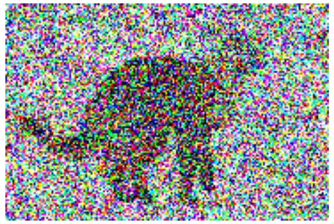
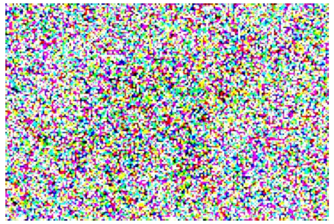
DDPM

Maximize $logP_\theta(\underline{x_0})$ $\longrightarrow$ Maximize $\mathrm{E}_{\boxed{q(\underline{x_1:x_T}|x_0)}}[log\left(\frac{P(\underline{x_0:x_T})}{q(\underline{x_1:x_T}|x_0)}\right)]$

Forward Process
(Diffusion Process)

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

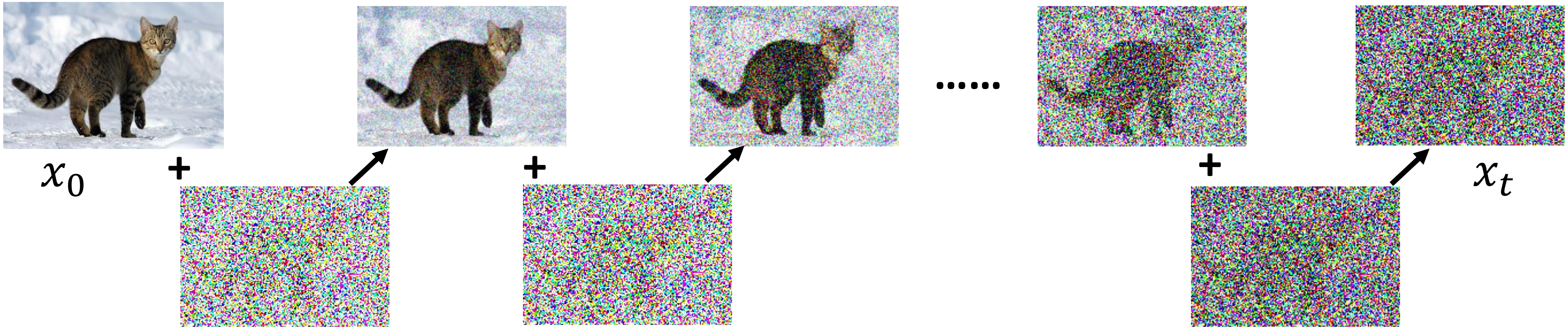
 $\sim\mathcal{N}(\mathbf{0}, I)$

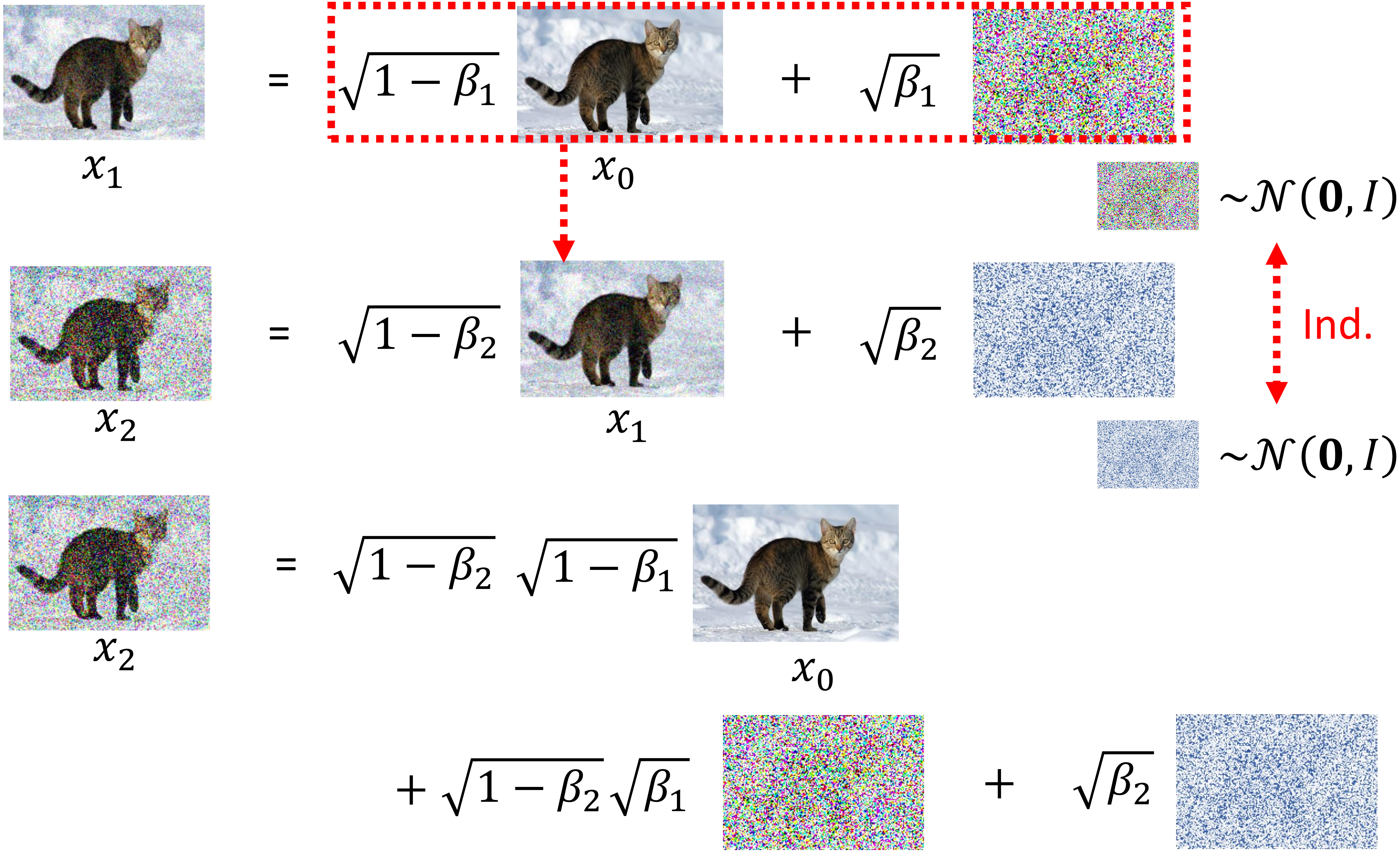
 $= \sqrt{1-\beta_t}$  $+ \sqrt{\beta_t}$ 

x_{t-1} $\qquad$ x_t

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

$q(x_t|x_0) = q(x_1|x_0)q(x_2|x_1)\dots q(x_t|x_{t-1})$ 不用!



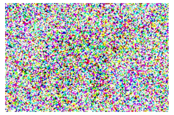


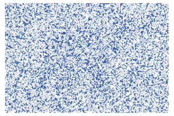
$x_1 = \sqrt{1-\beta_1}\, x_0 + \sqrt{\beta_1}$
x_1
x_0
$\sim \mathcal{N}(\mathbf{0}, I)$
Ind.
$x_2 = \sqrt{1-\beta_2}\, x_1 + \sqrt{\beta_2}$
x_2
x_1
$\sim \mathcal{N}(\mathbf{0}, I)$
$x_2 = \sqrt{1-\beta_2}\sqrt{1-\beta_1}\, x_0$
x_2
x_0
$+ \sqrt{1-\beta_2}\sqrt{\beta_1} + \sqrt{\beta_2}$

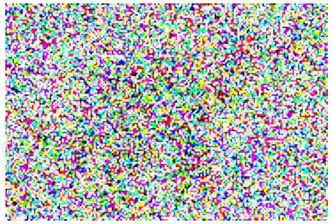
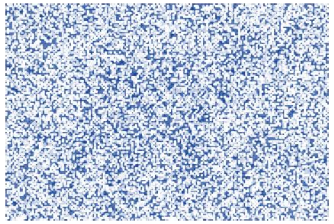
x_2

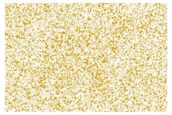
$= \sqrt{1-\beta_2}\,\sqrt{1-\beta_1}$ 

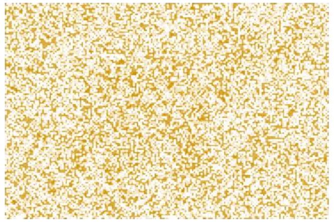
x_0

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

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

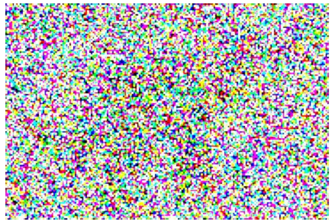
$+\sqrt{1-\beta_2}\sqrt{\beta_1}$  $+ \sqrt{\beta_2}$ 

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

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

$q(x_t|x_0)$

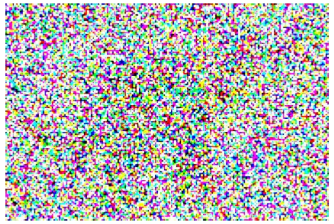
$\beta_1, \beta_2, \dots, \beta_T$

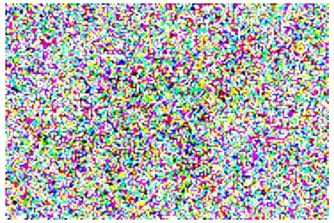


$\sim\mathcal{N}(\mathbf{0}, I)$

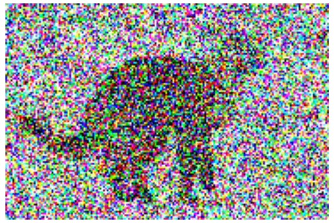
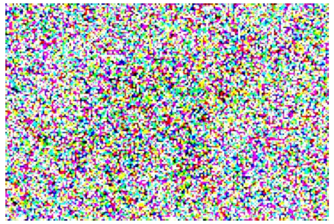
$\alpha_t = 1 - \beta_t$

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

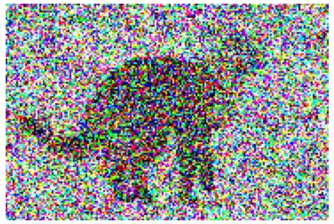
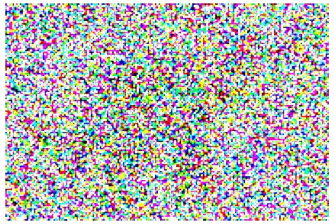
 $= \sqrt{1-\beta_1}$  $+ \sqrt{\beta_1}$ 

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

$\vdots$

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

|||

 $= \underbrace{\boxed{\sqrt{1-\beta_1}\dots\sqrt{1-\beta_t}}}_{\sqrt{\bar{\alpha}_t}}$  $+ \underbrace{\boxed{\sqrt{1-(1-\beta_1)\dots(1-\beta_t)}}}_{\sqrt{1-\bar{\alpha}_t}}$ 

$$
\begin{aligned}
\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] && (47)\\
&= \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] && (48)\\
&= \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] && (49)\\
&= \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] && (50)\\
&= \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] && (51)\\
&= \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] && (52)\\
&= \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] && (53)\\
&= \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] && (54)\\
&= \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] && (55)\\
&= \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] && (56)\\
&= \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] && (57)\\
&= \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}} && (58)
\end{aligned}
$$

Understanding Diffusion Models:
A Unified Perspective

https://arxiv.org/pdf/2208.11970.pdf

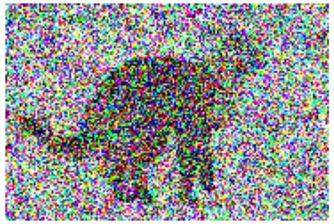
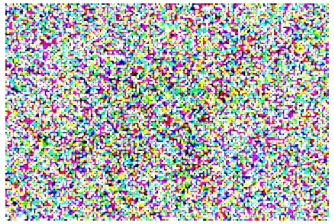
DDPM: Lower bound of $logP(x)$

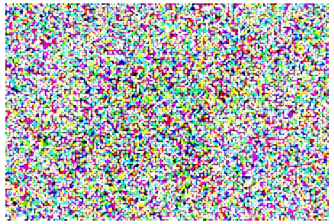
$$\mathrm{E}_{q(x_1|x_0)}[logP(x_0|x_1)] \; -KL\big(q(x_T|x_0)||P(x_T)\big)$$

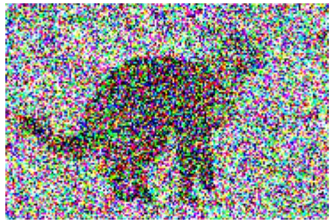
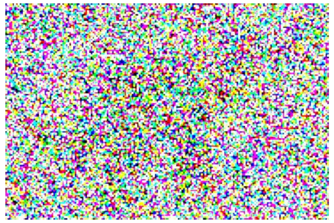
$$-\sum_{t=2}^{T} \mathrm{E}_{q(x_t|x_0)}\big[KL\big(q(x_{t-1}|x_t, x_0)||P(x_{t-1}|x_t)\big)\big]$$

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

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

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

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

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

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

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

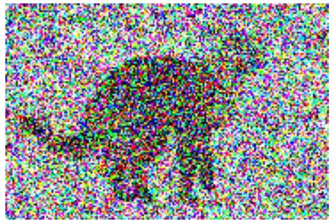
x_0

$q(x_t|x_0)$

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

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

x_{t-1}



x_t

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

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

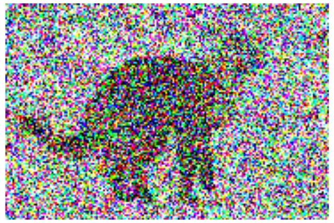
$$
\begin{aligned}
q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) &= \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} && (71)\\
&= \frac{\mathcal{N}(\boldsymbol{x}_t;\sqrt{\alpha_t}\boldsymbol{x}_{t-1},(1-\alpha_t)\mathbf{I})\mathcal{N}(\boldsymbol{x}_{t-1};\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})} && (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\} && (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\} && (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\} && (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\} && (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\} && (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\} && (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\} && (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\} && (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\} && (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\} && (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\} && (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)}) && (84)
\end{aligned}
$$

https://arxiv.org/pdf/2208.11970.pdf

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

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

Gaussian



Mean

Variance

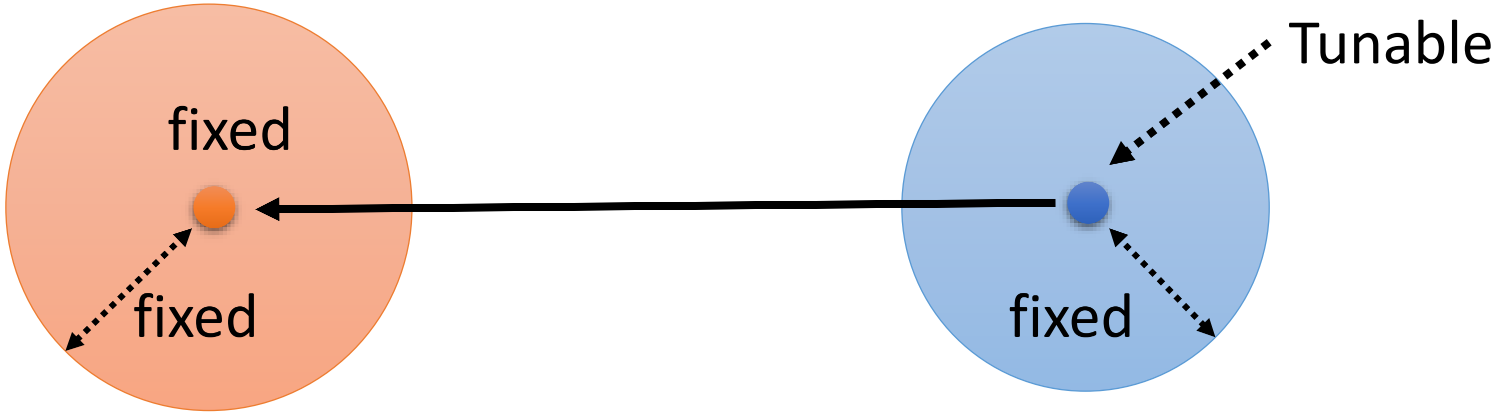
$$\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t x_0 + \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t}{1-\bar{\alpha}_t}$$

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

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

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

How to minimize KL divergence?



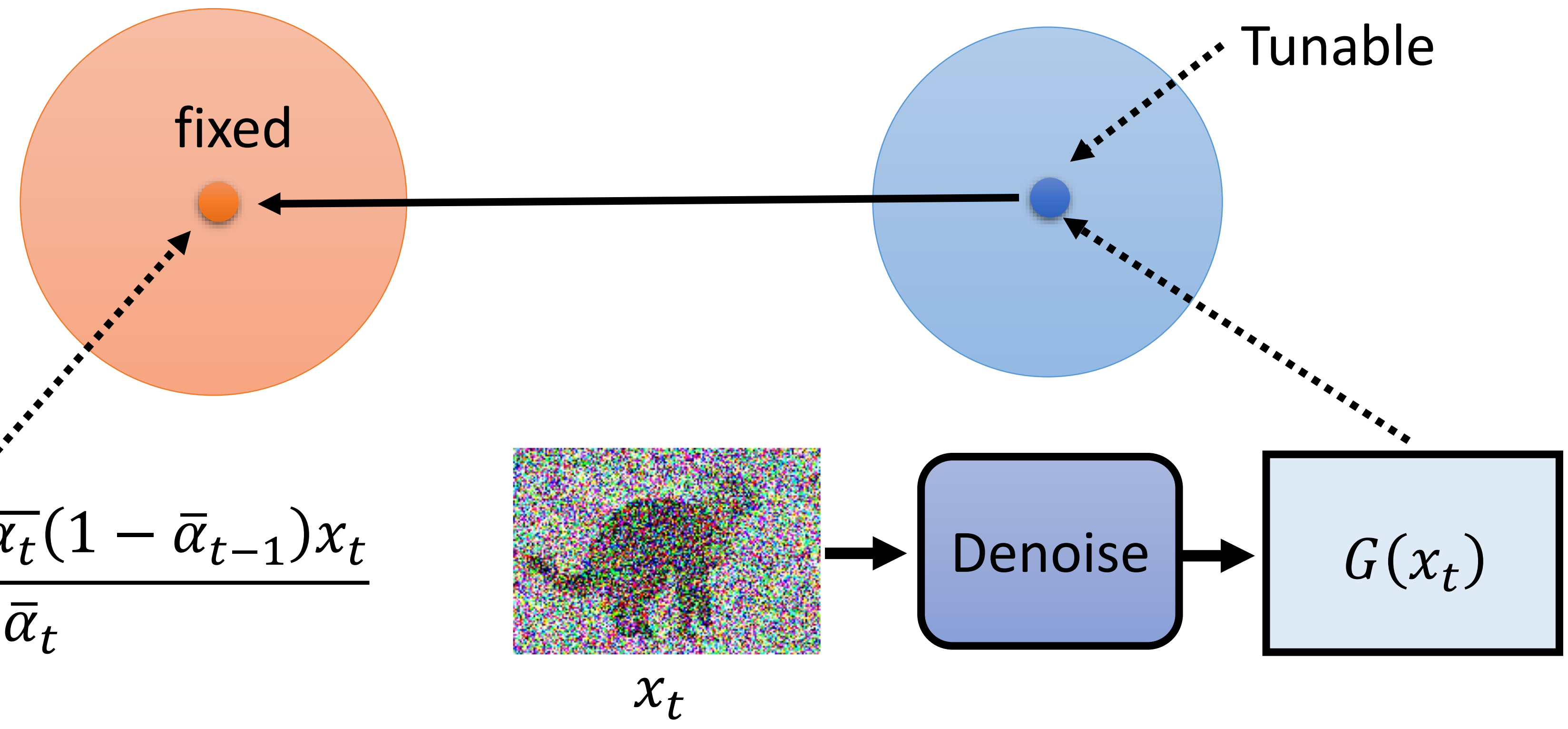
Recall that the KL Divergence between two Gaussian distributions is:

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

$$E_{q(x_1|x_0)}[\log P(x_0|x_1)] - KL(q(x_T|x_0)||P(x_T))$$

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

How to minimize KL divergence?



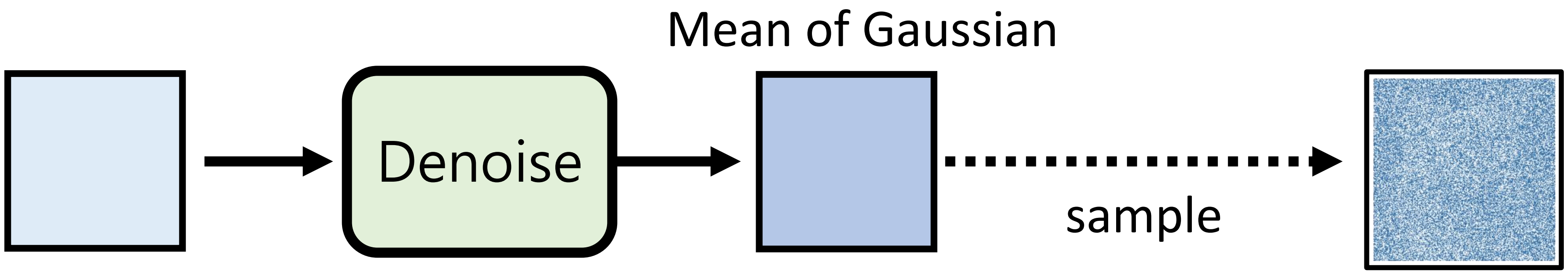
$$\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t x_0 + \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t}{1-\bar{\alpha}_t}$$

Algorithm 2 Sampling

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

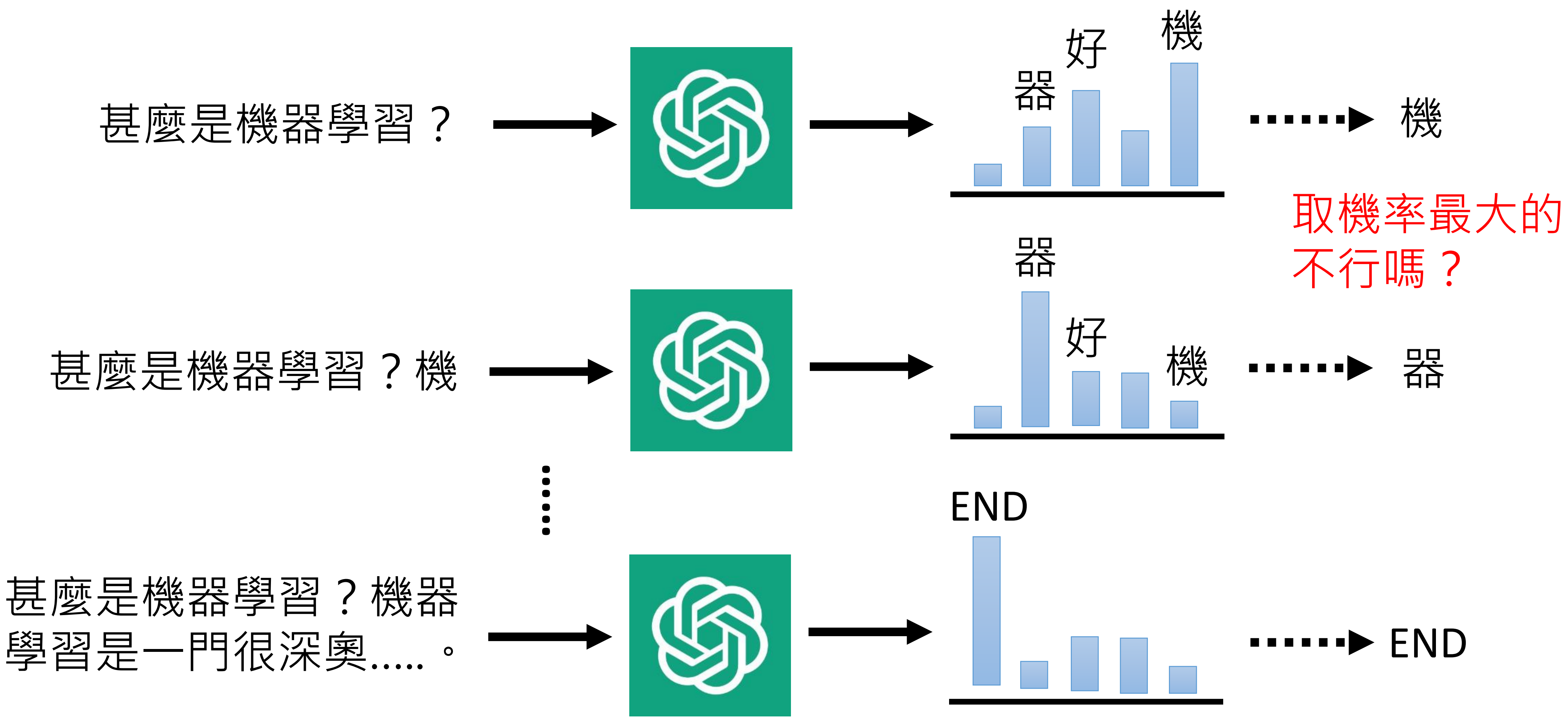
$\sigma_t = 0$

感謝伏宇寬助教提供結果



為什麼不直接取 Mean ?

為什麼生成文句時需要 Sample？



為什麼生成文句時需要 Sample？

- The Curious Case of Neural Text Degeneration

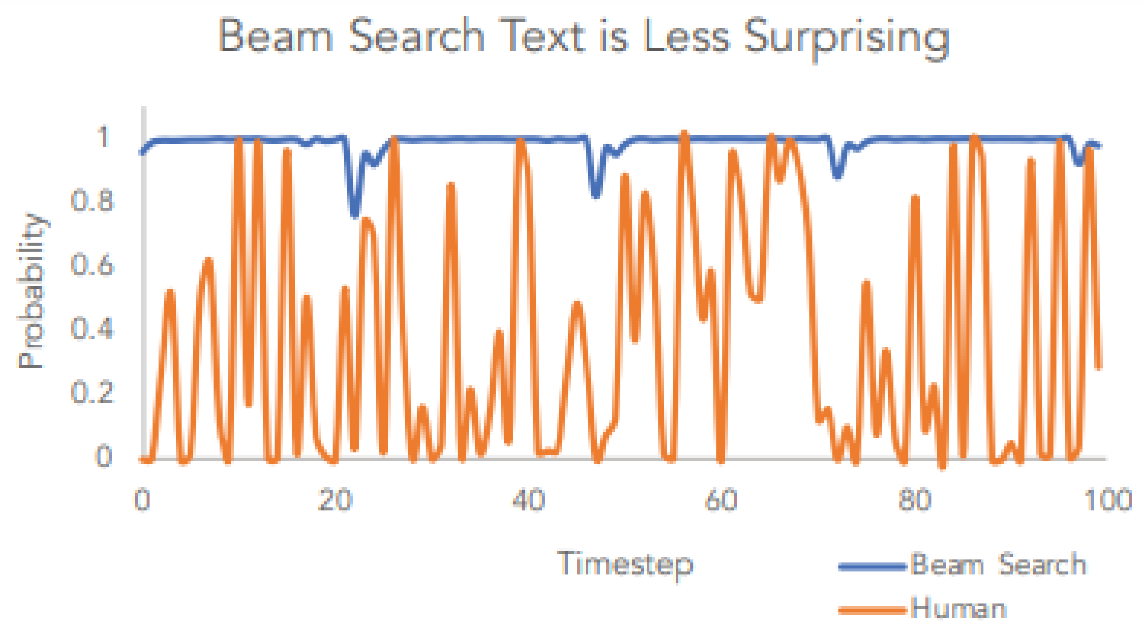
  https://arxiv.org/abs/1904.09751

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

https://arxiv.org/abs/1904.09751



Beam Search

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

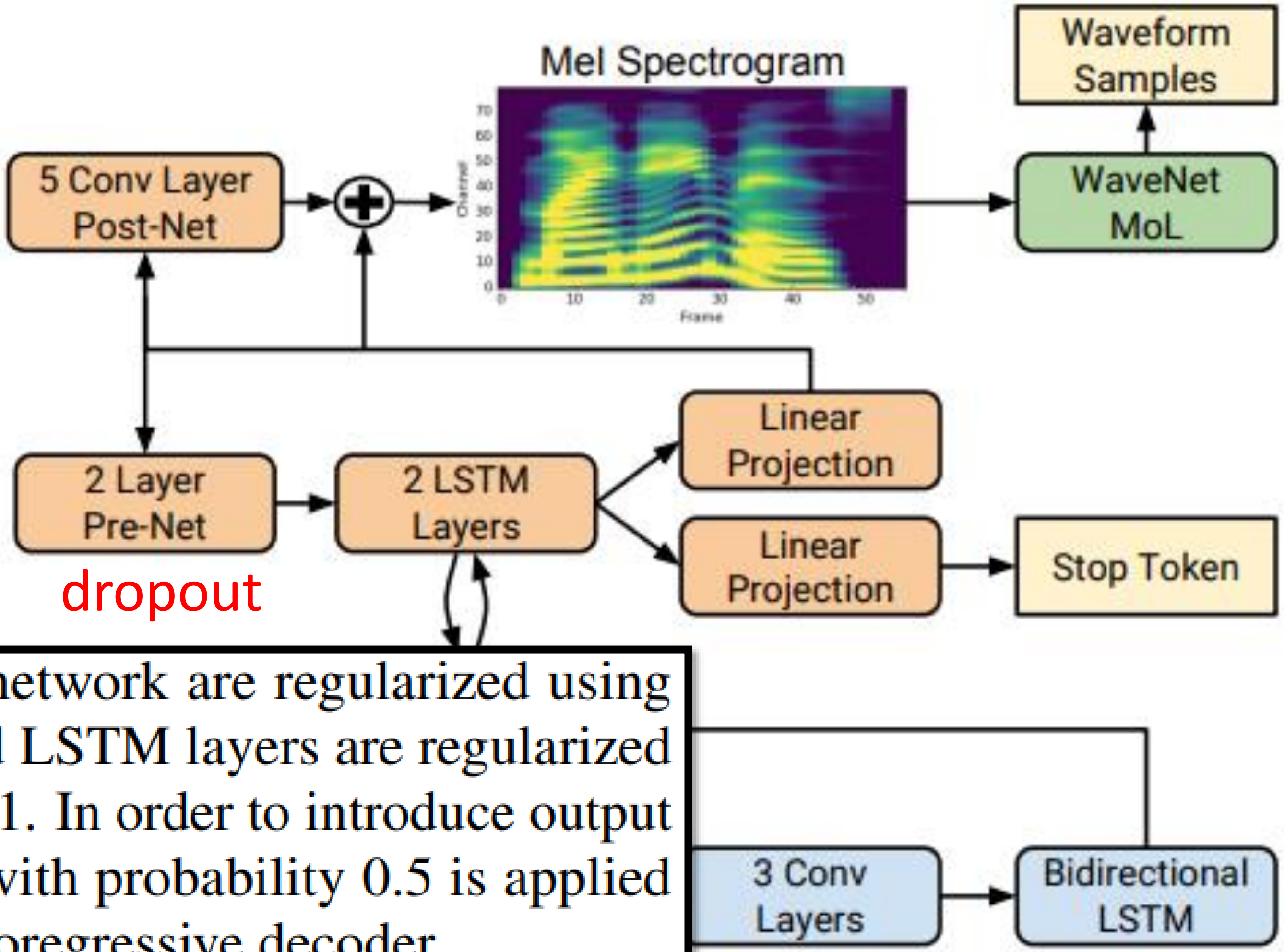
語音合成也需要 Sampling !

https://arxiv.org/abs/1712.05884

with dropout

without dropout

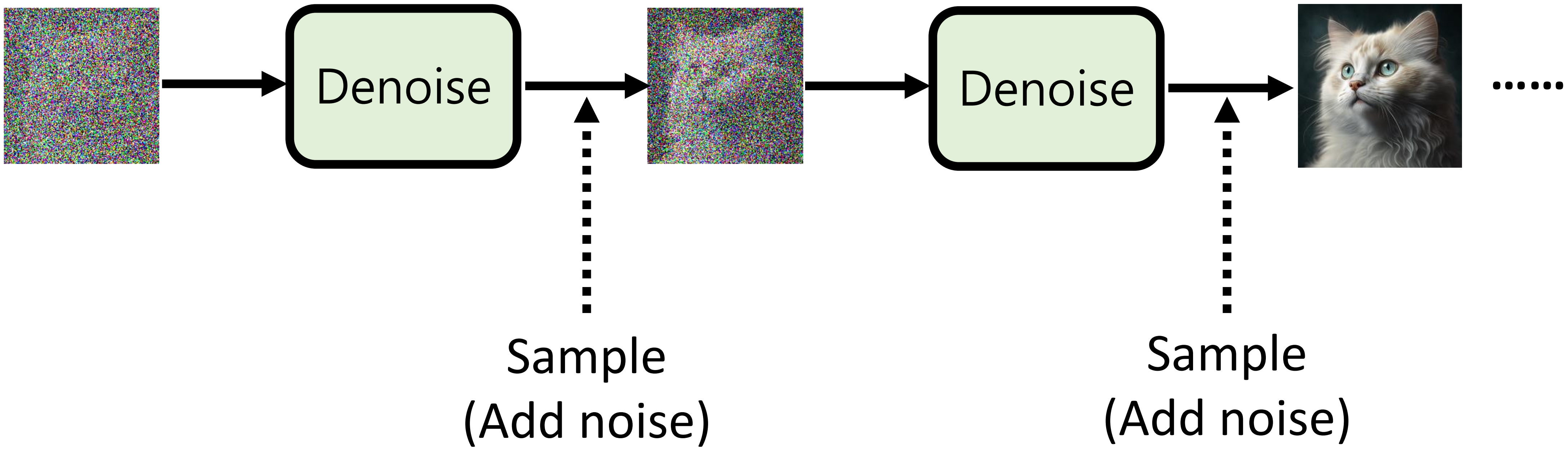
感謝杜濤同學提供實驗結果



The convolutional layers in the network are regularized using dropout [25] with probability 0.5, and LSTM layers are regularized using zoneout [26] with probability 0.1. In order to introduce output variation at inference time, dropout with probability 0.5 is applied only to layers in the pre-net of the autoregressive decoder.

Diffusion Model 是一種 Autoregressive

「一次到位」改成「 N次到位」



Denoising Diffusion Probabilistic Models

Algorithm 1 Training

repeat

  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$

  $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$

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

  Take gradient descent step on

$$\nabla_\theta \left\| \boxed{\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)} \right\|^2$$

until converged

Algorithm 2 Sampling

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

for $t = T, \ldots, 1$ **do**

  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$

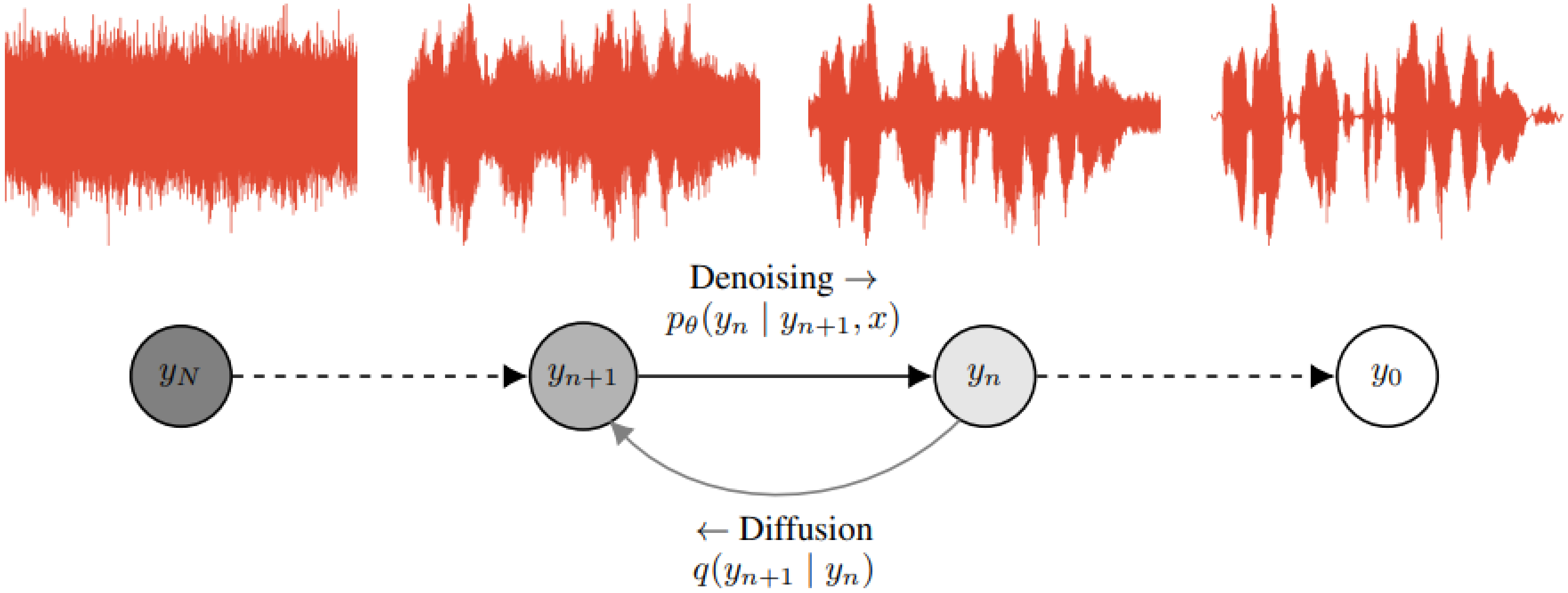
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \boxed{\sigma_t \mathbf{z}}$

end for

return $\mathbf{x}_0$

Diffusion Model for Speech

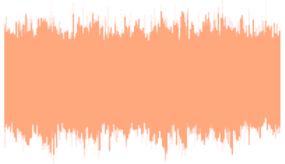
- WaveGrad



https://arxiv.org/abs/2009.00713

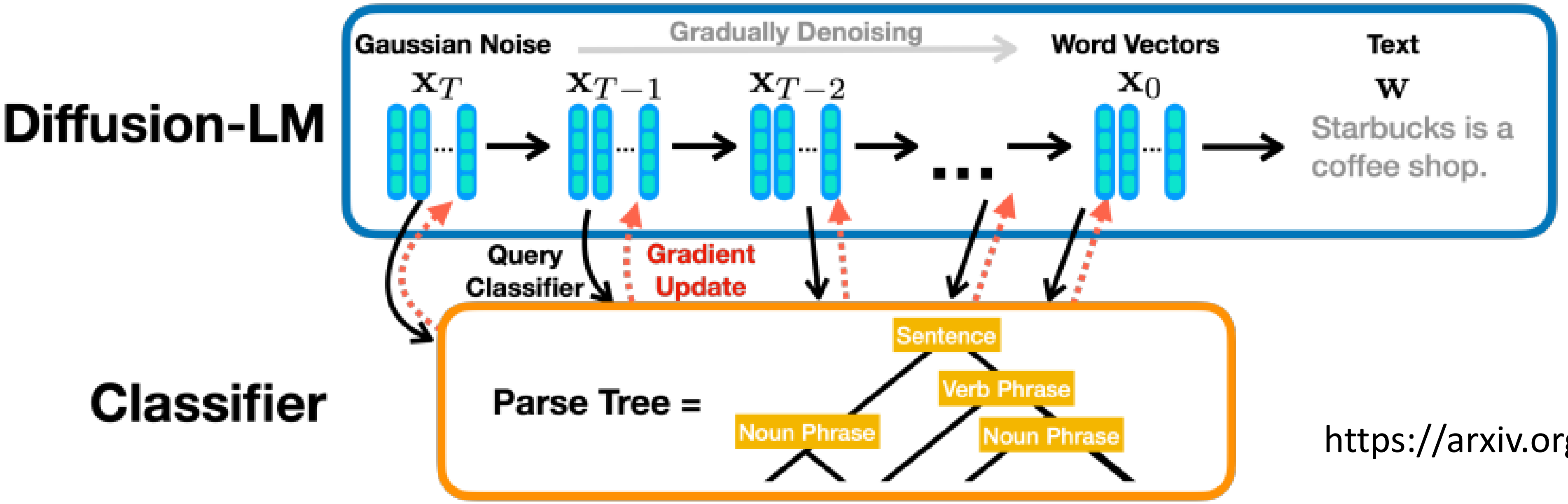
Diffusion Model for Text

- Difficulty: 你 好 嗎 ? → … → … →



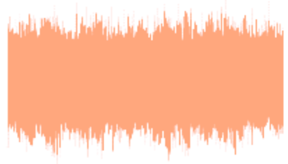
Add noise

- Solution: Noise on latent space



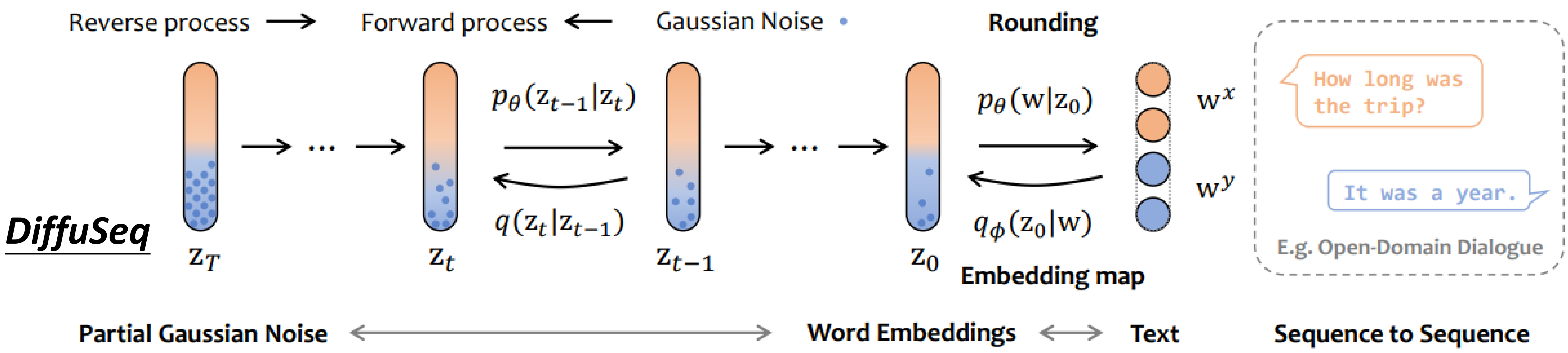
https://arxiv.org/abs/2205.14217

Diffusion Model for Text

- Difficulty: 你 好 嗎 ? → … → … → 

  Add noise

- Solution: Noise on latent space



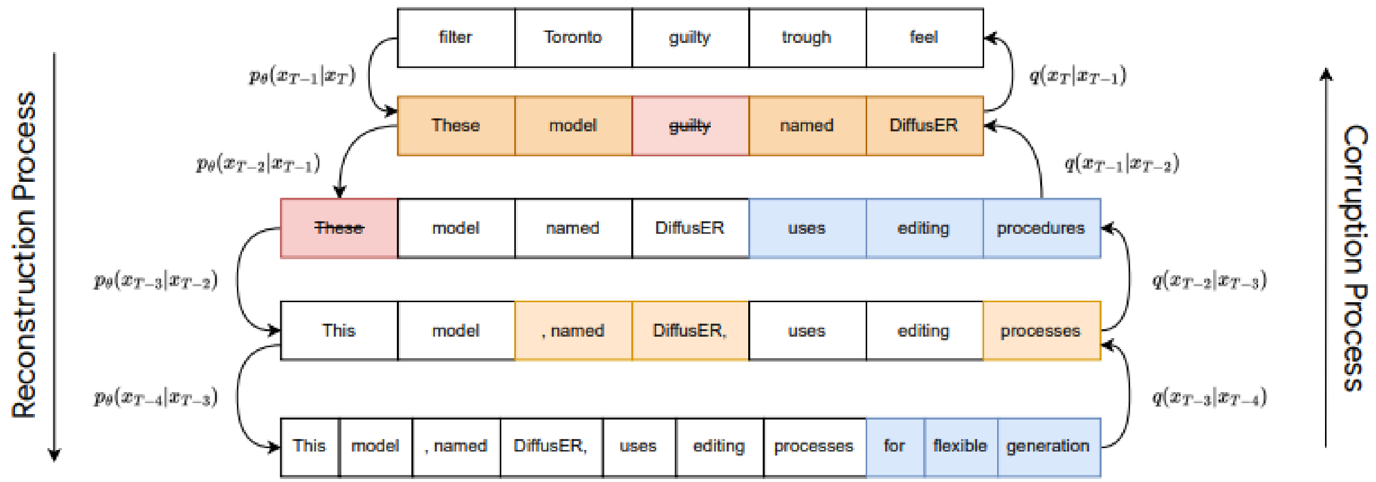
https://arxiv.org/abs/2210.08933

Diffusion Model for Text

- Solution: Don’t add Gaussian noise

https://arxiv.org/abs/2210.16886

Diffusion via Edit-based Reconstruction (DiffusER)



t = 128 **[MASK] [MASK] [MASK] [MASK] [MASK] [MASK]**...

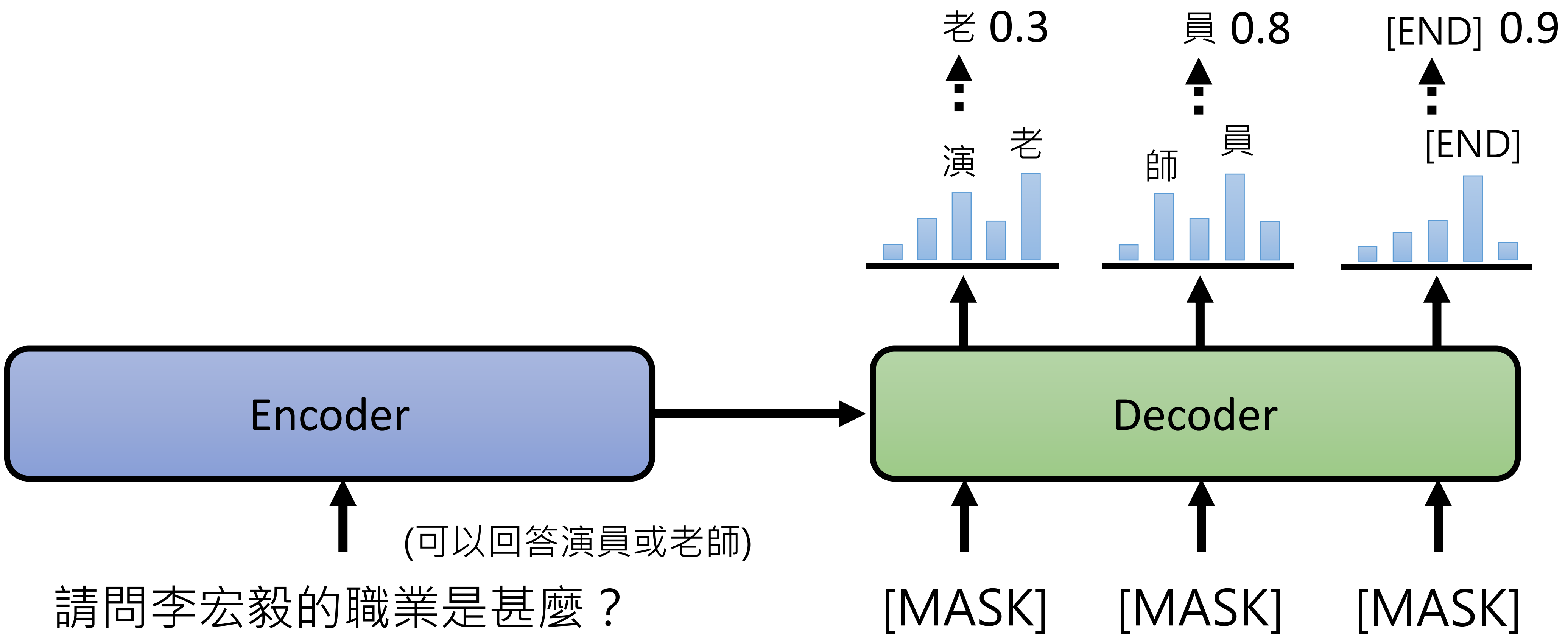
t = 25 In response **[MASK]** the demands , **[MASK] [MASK]**y Workers union said **[MASK]** backflow fund **[MASK]**s would face further investigation and a fine.

t = 0 In response to the demands , the Community Workers union said the backflow fund managers would face further investigation and a fine .

https://arxiv.org/abs/2107.03006

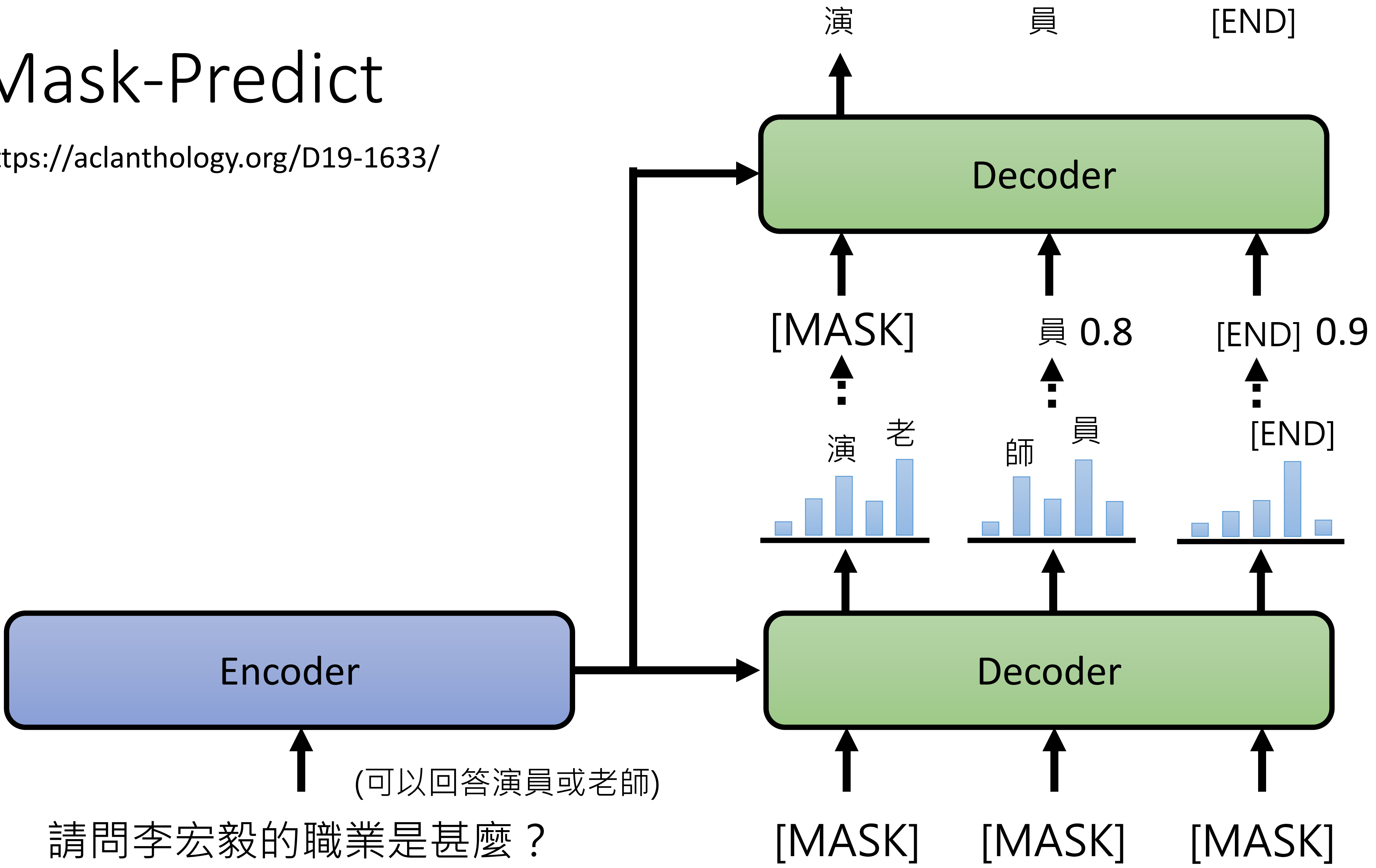
Mask-Predict

https://aclanthology.org/D19-1633/



Mask-Predict

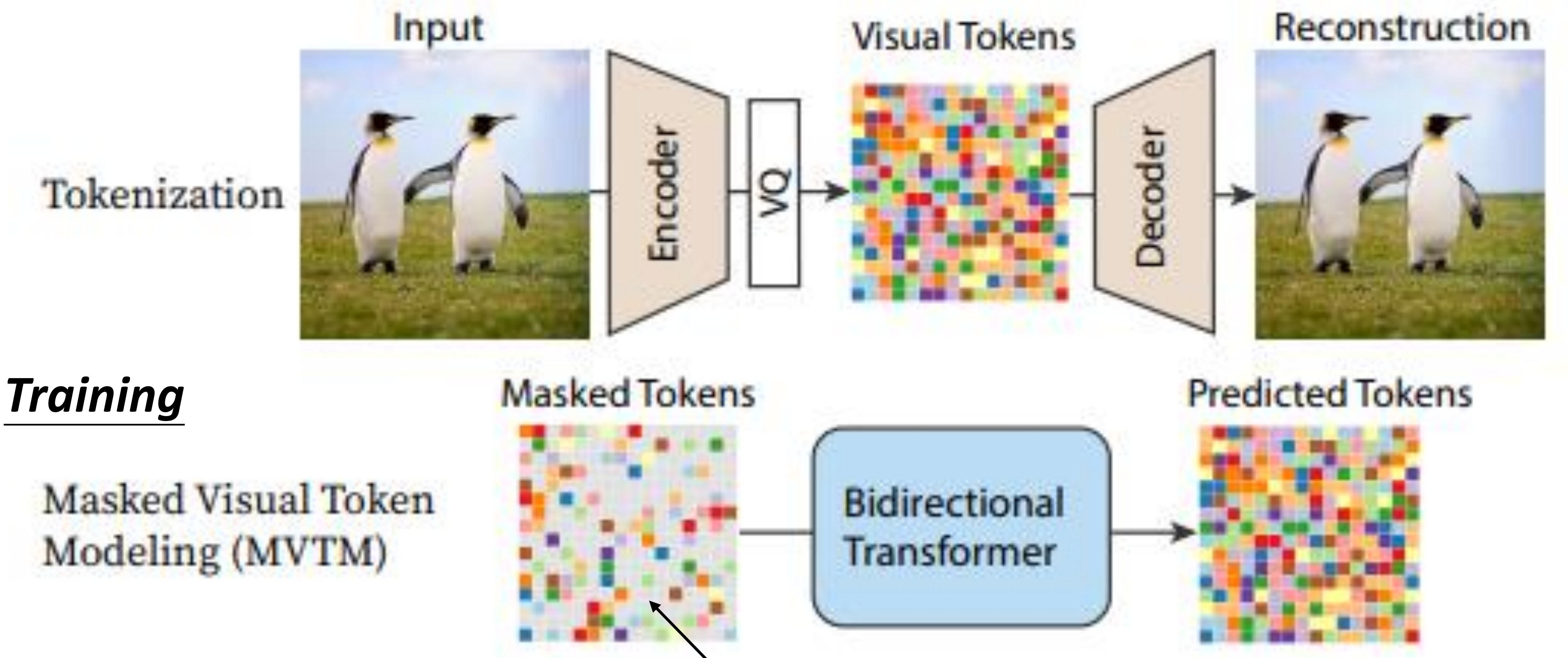
https://aclanthology.org/D19-1633/



Mask-Predict

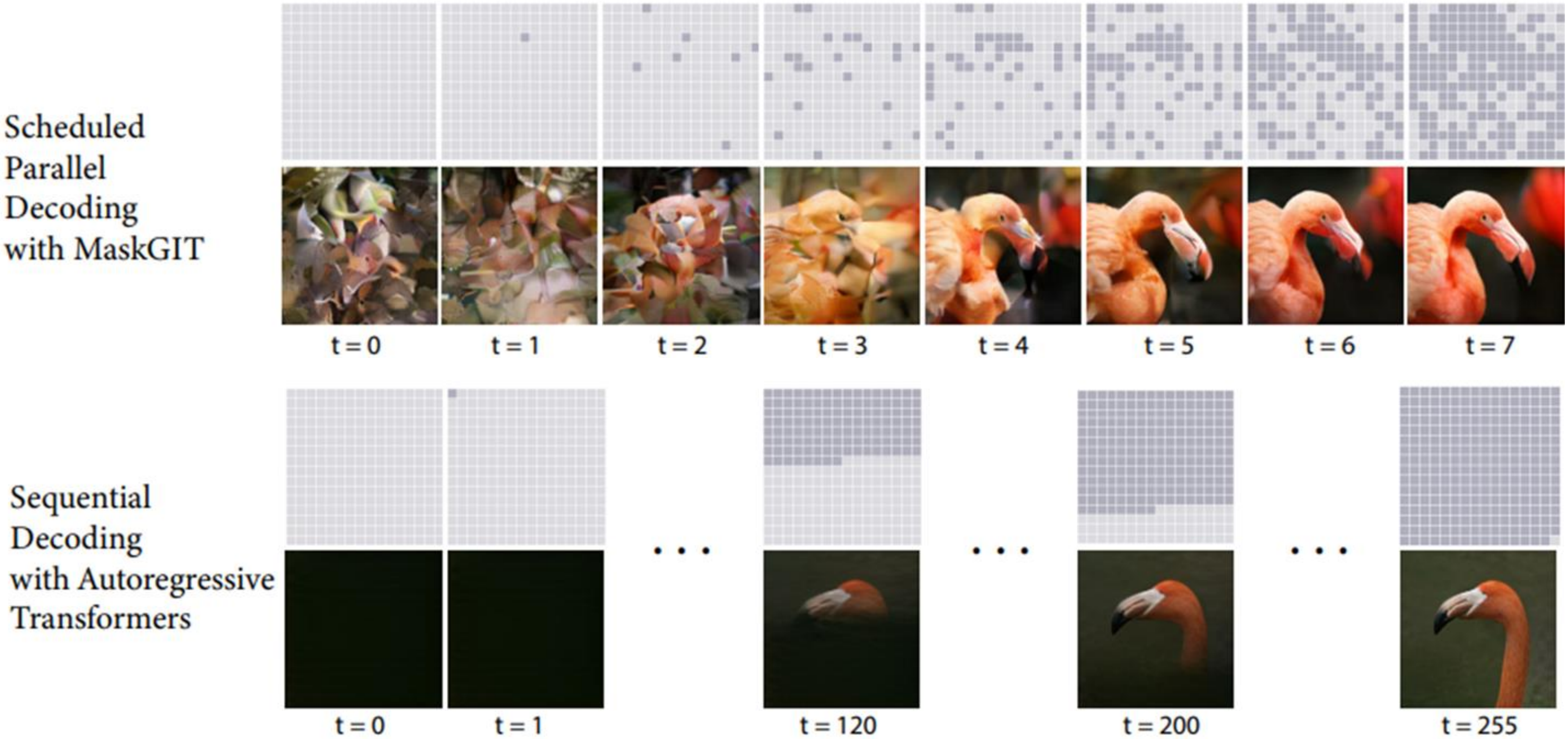
https://arxiv.org/abs/2202.04200

https://arxiv.org/abs/2301.00704



Training

Gray: [mask] token



Scheduled Parallel Decoding with MaskGIT
t = 0
t = 1
t = 2
t = 3
t = 4
t = 5
t = 6
t = 7
Sequential Decoding with Autoregressive Transformers
t = 0
t = 1
t = 120
t = 200
t = 255