CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

CSDS 600: Deep Generative Models

Diffusion Models

Yu Yin (yu.yin@case.edu)

Case Western Reserve University

https://yin-yu.github.io/

Outline

- Introduction
- Theory of diffusion
- Tricks to improve image synthesis models
- Examples of recent diffusion models
  - Text-to-image generation
    - Stable diffusion
    - DALL-E series
    - Imagen
  - …

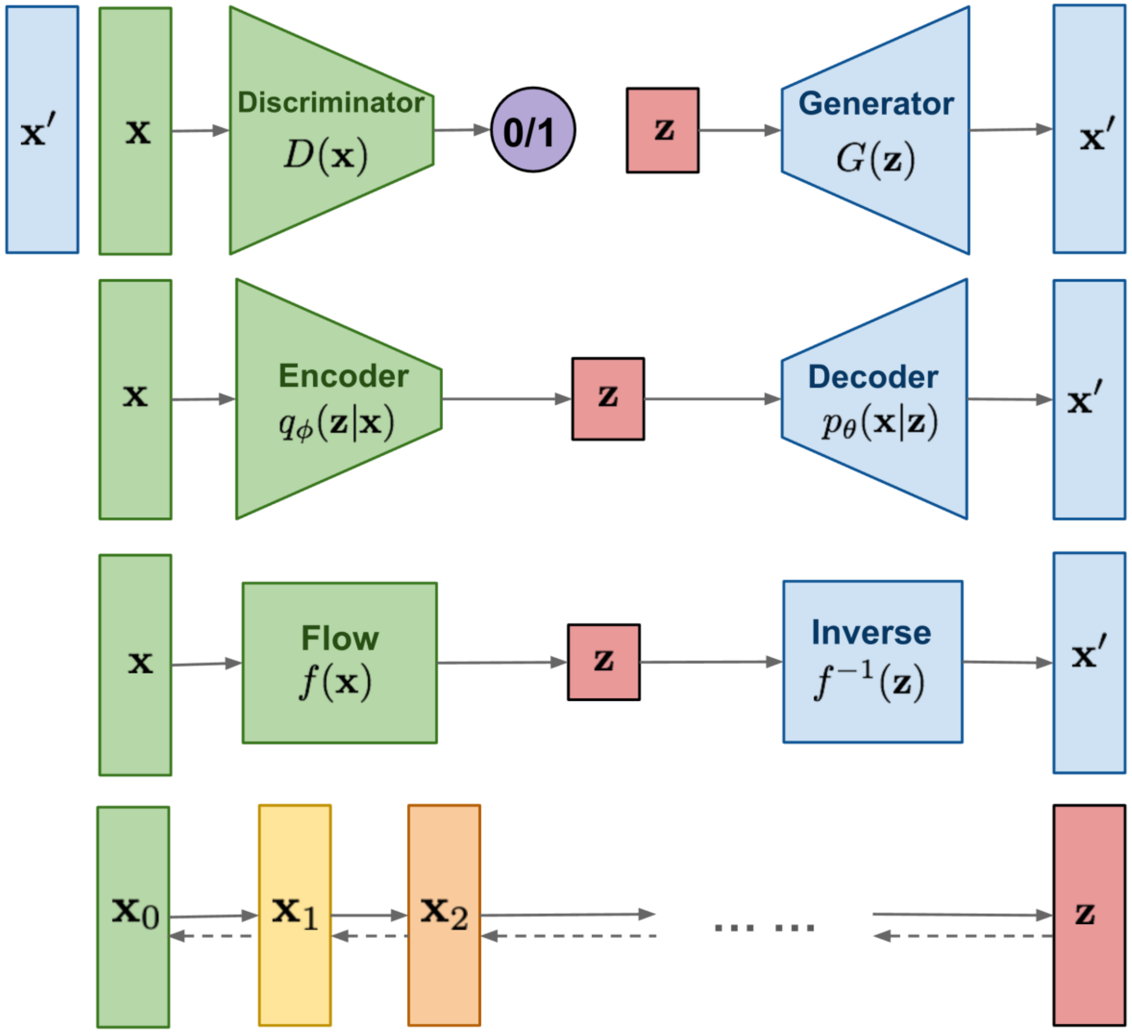
Sampling from Noise

GAN: Adversarial training

VAE: maximize variational lower bound

Flow-based models: Invertible transform of distributions

Diffusion models: Gradually add Gaussian noise and then reverse



Diffusion

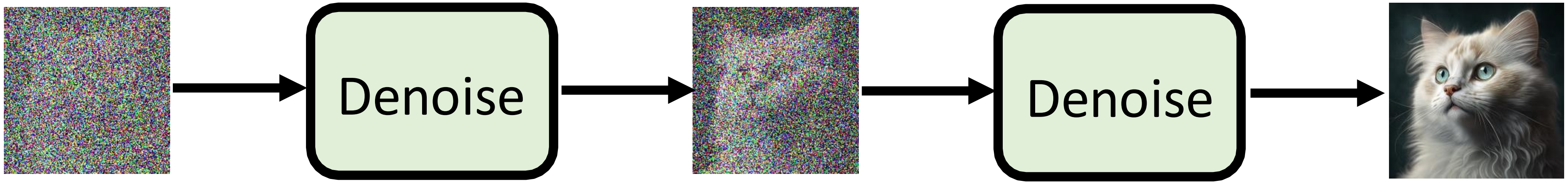
- A generative modeling technique that takes inspiration from physics
- Main idea:

  convert a well-known and simple base distribution (like a Gaussian) to the target (data) distribution iteratively, with small step sizes, via a Markov chain

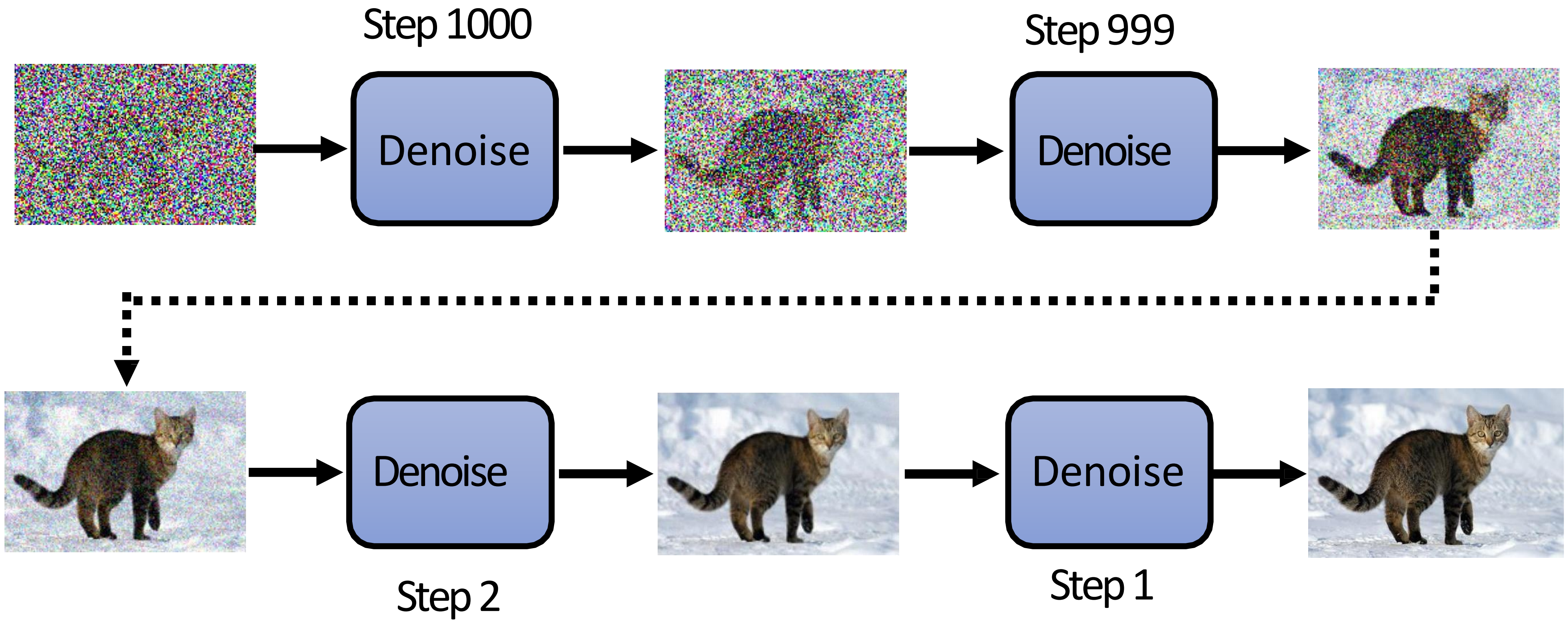
Introduction

Forward Process

Reverse Process



How does diffusion model work?



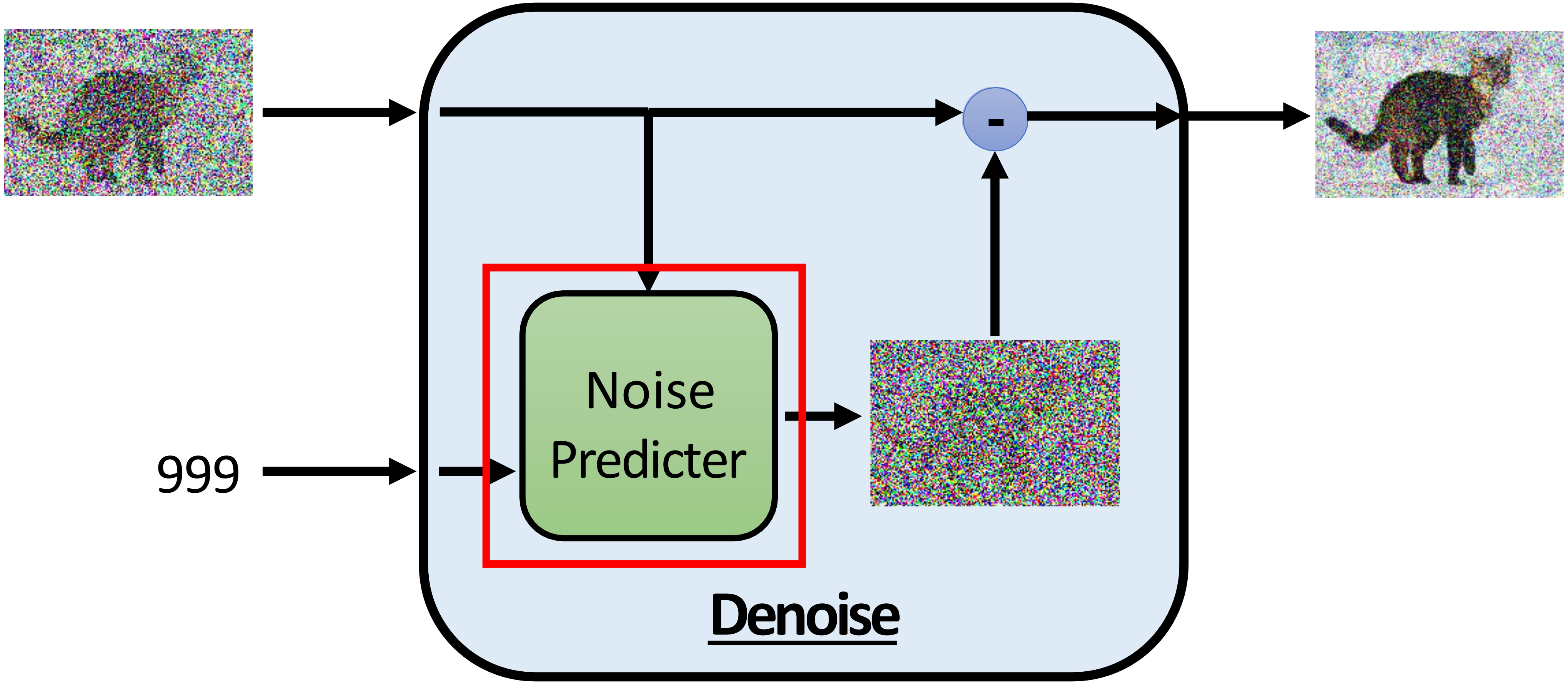
<u>Reverse Process</u>

How does diffusion model work?

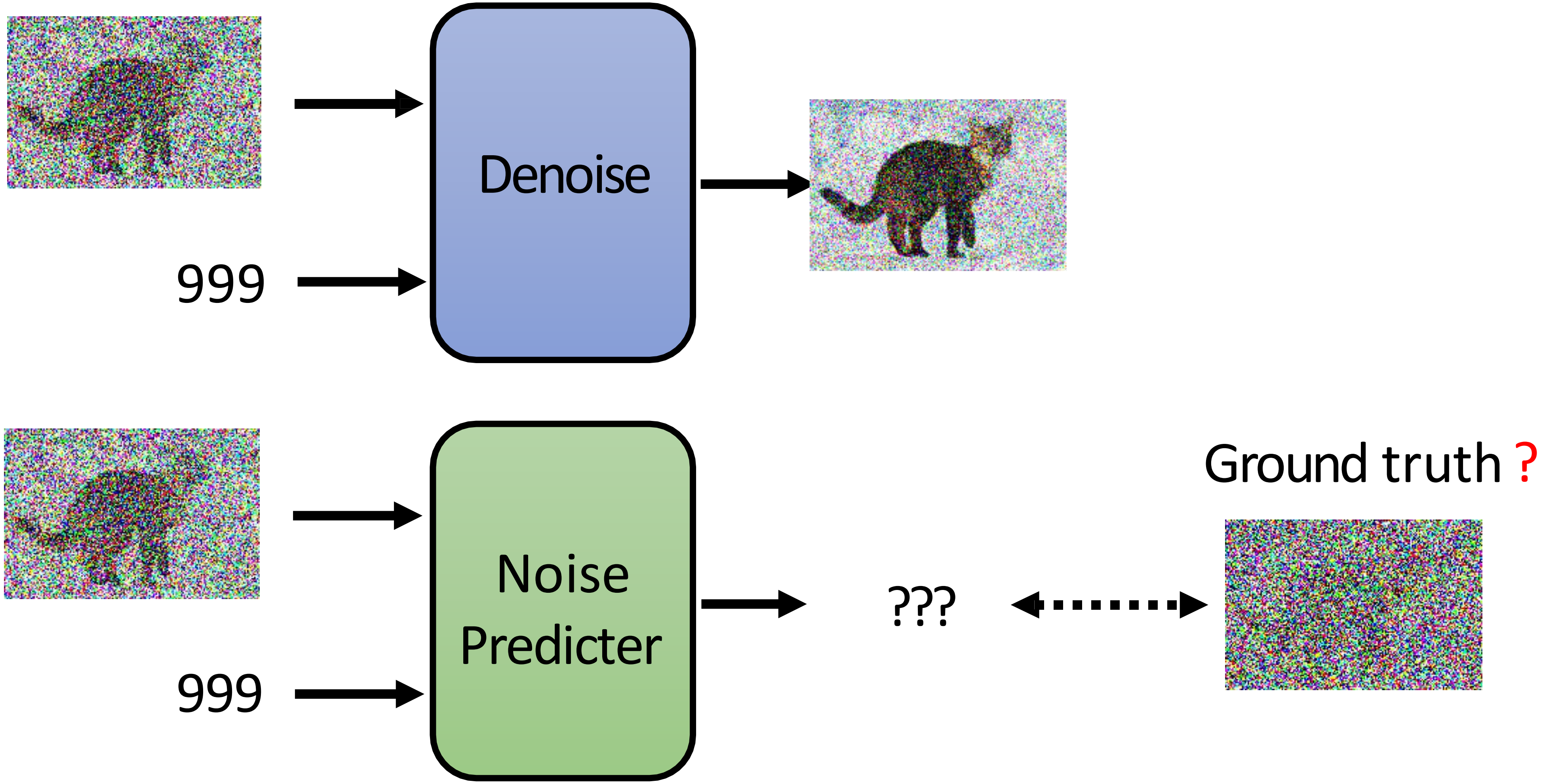
Denoise module is the same for each step.

<u>Reverse Process</u>

Denoise module

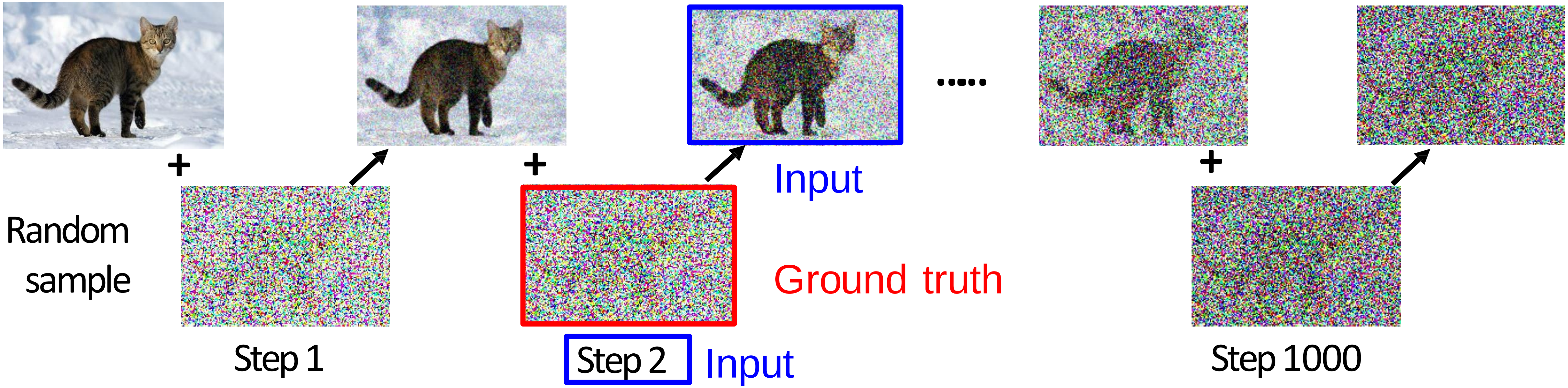


How to train noise predictor?



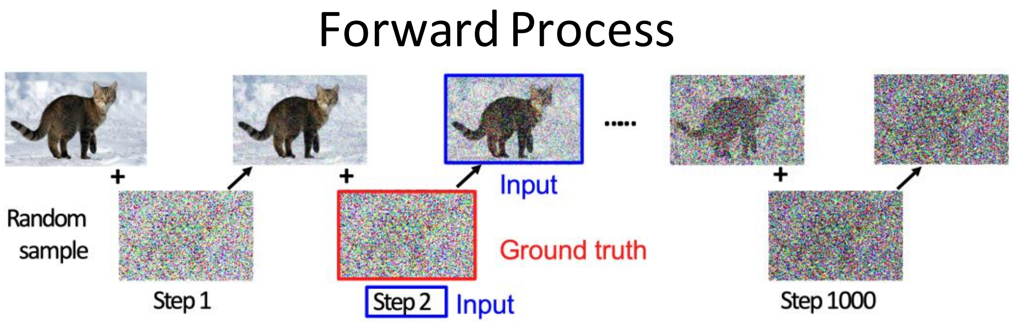
How to train noise predictor?

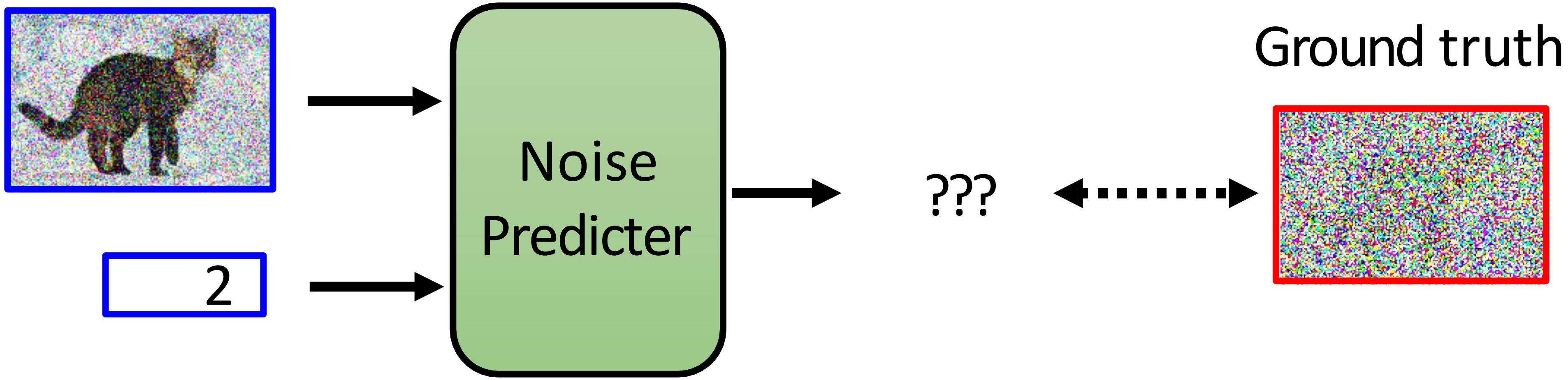
Create pair-wise training data



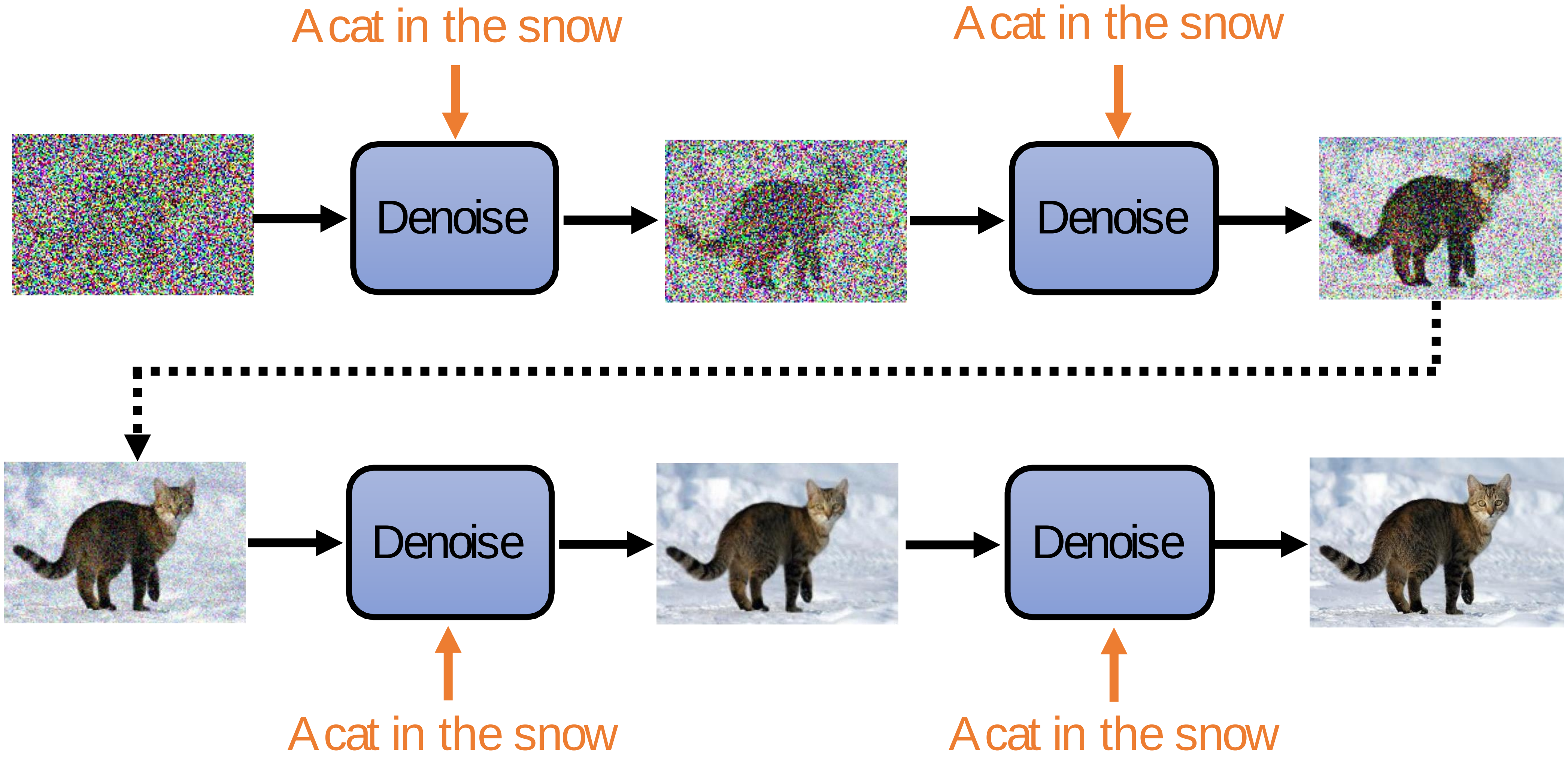
Forward Process (Diffusion Process)

How to train noise predictor?



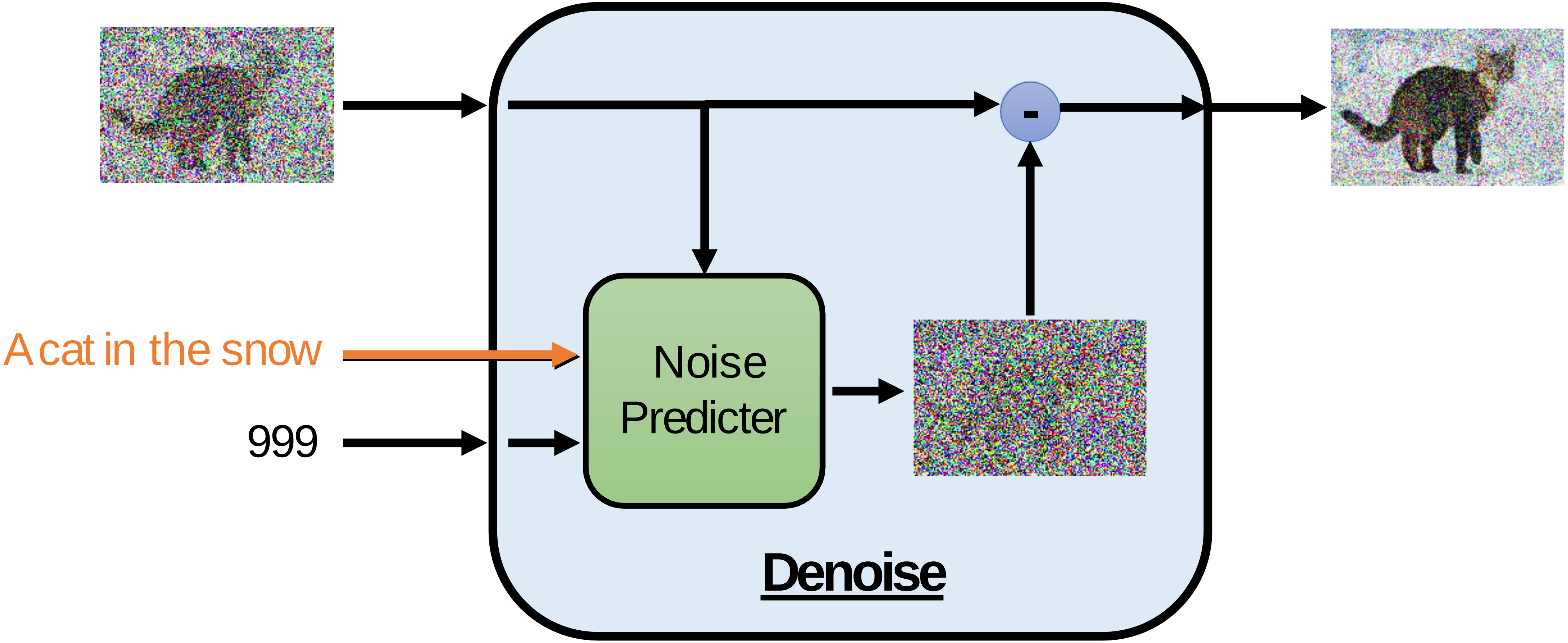


Text-to-image Generation



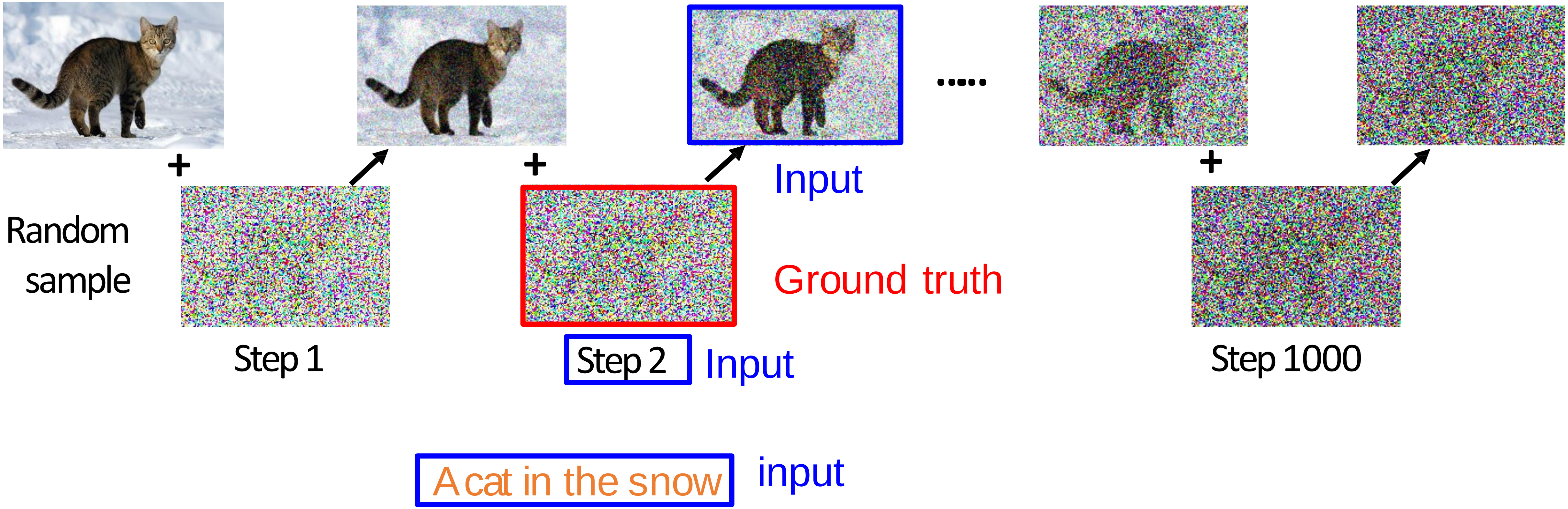
Text-to-image Generation

Denoise module



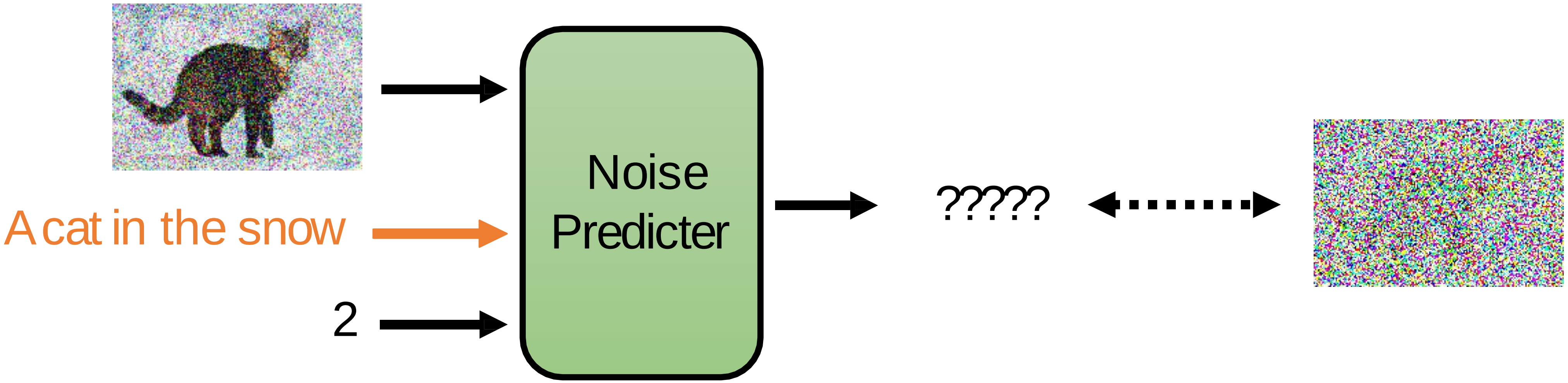
Text-to-image Generation

Forward Process



Text-to-image Generation

Noise predictor



Denoising Diffusion Probabilistic Models

Algorithm 1 Training

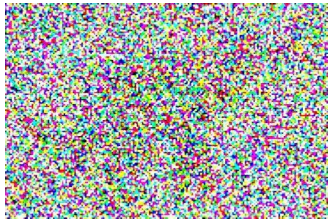
1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0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

Algorithm 2 Sampling

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

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

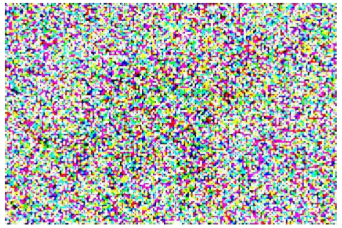
Target Noise

$\bar\alpha_1, \bar\alpha_2, \ldots \bar\alpha_T$ → smaller

Training

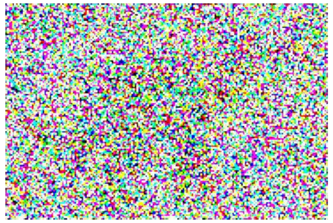
- Sampling

x_0



ε

Time step t

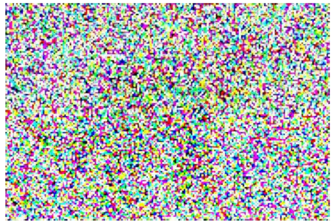
$\sqrt{\bar{\alpha}_t}$  $+ \sqrt{1 - \bar{\alpha}_t}$  $=$ 

x_0 ε

t

Noise Predicter

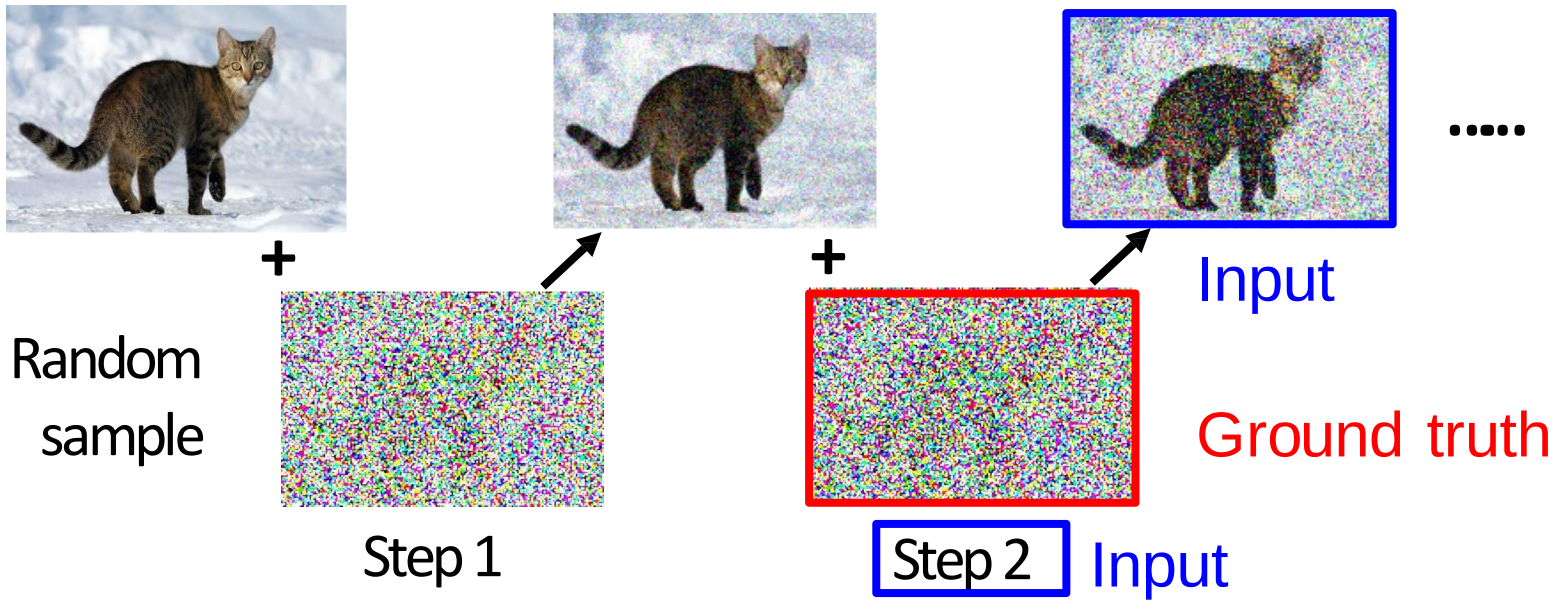
?????



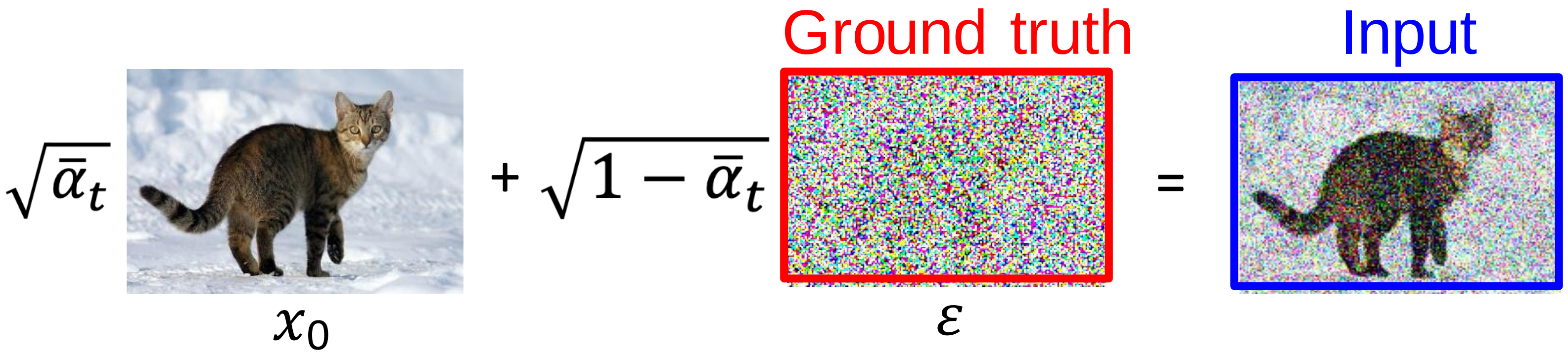
ε

Forward pass

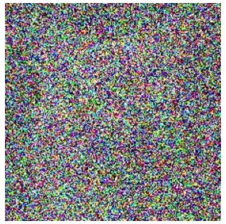
- Ideally



- DDPM



Inference



x_T

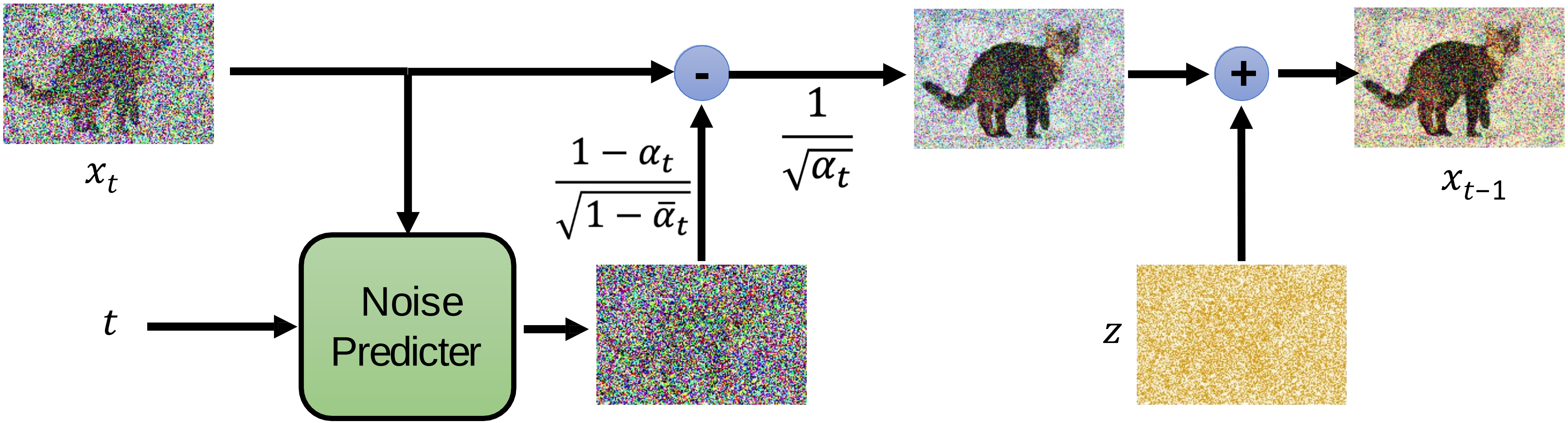
Algorithm 2 Sampling

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

Sample a noise?

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

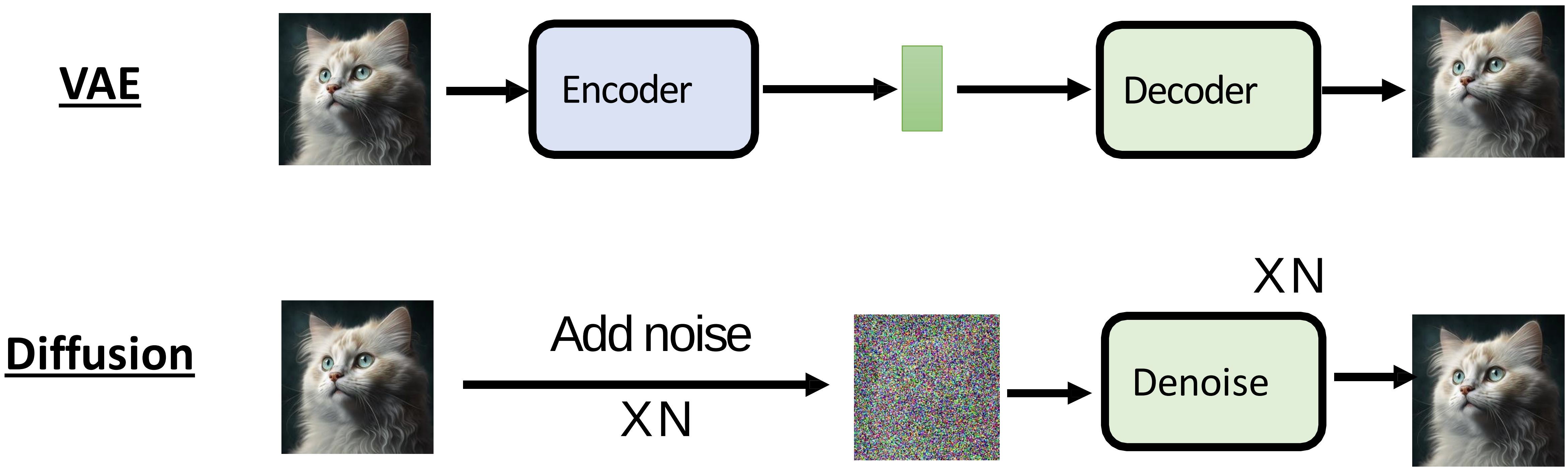
$\alpha_1, \alpha_2, \ldots \alpha_T$



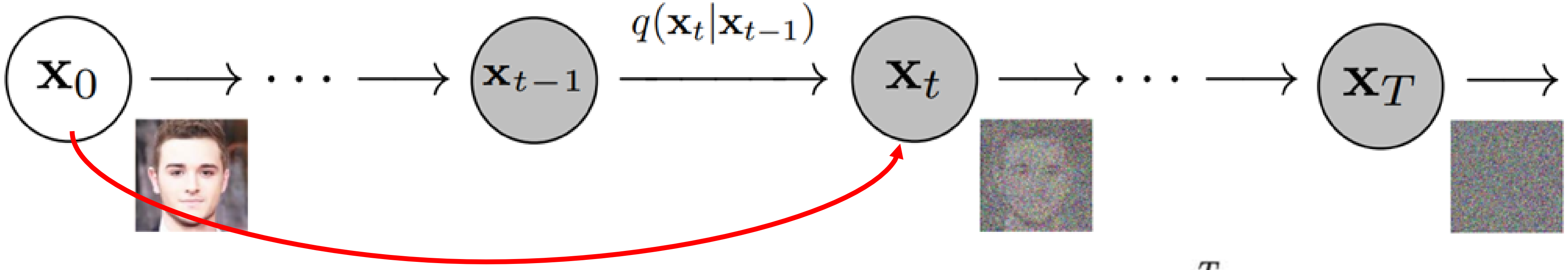
Outline

- Introduction
- Theory of diffusion
- Tricks to improve image synthesis models
- Examples of recent diffusion models
  - Text-to-image generation
    - Stable diffusion
    - DALL-E series
    - Imagen
  - …

VAE vs. Diffusion Model



Forward Process



$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

$$\{\beta_t \in (0,1)\}_{t=1}^{T}$$

- Take a datapoint x_0 and keep gradually adding small amounts of Gaussian noise
  - Vary the parameters of the Gaussian according to a noise schedule controlled by β_t
- Repeat this process for T steps — as the timesteps increase, the more features of the original input are destroyed

A neat (reparameterization) trick

Define

$$\alpha_t = 1 - \beta_t$$

$$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$$

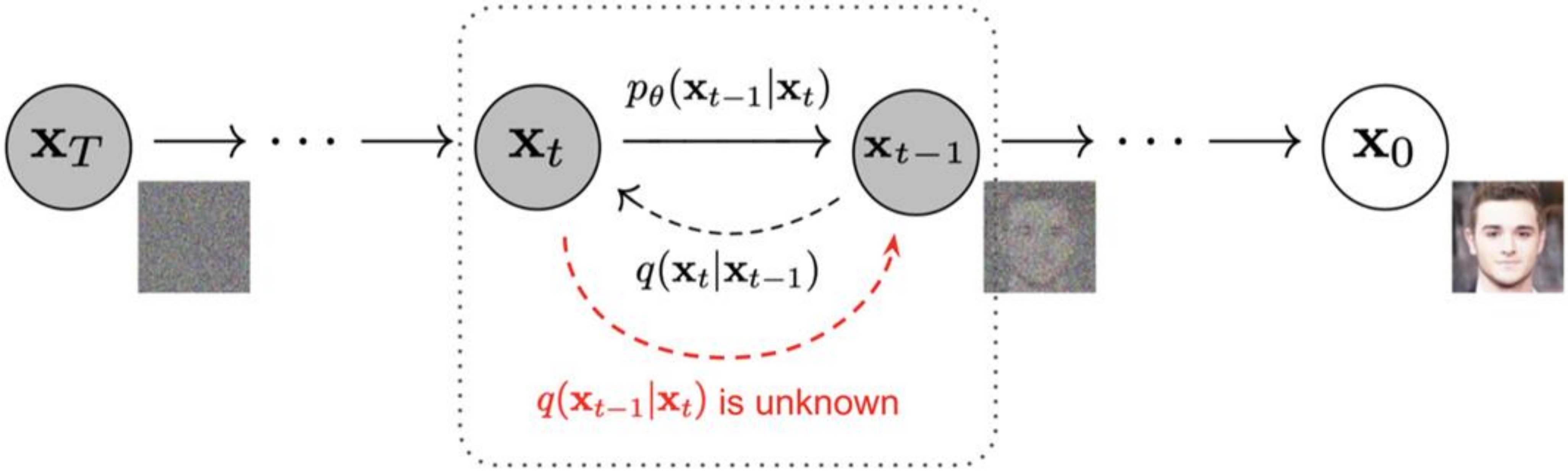
Then

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right)$$

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \\
&= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\epsilon \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\epsilon \\
&= \dots \\
&= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon
\end{aligned}$$

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right)$$

Reverse Process



- The goal of a diffusion model is to **learn** the reverse denoising process to iteratively **undo** the forward process
- In this way, the reverse process appears as if it is generating new data from random noise!

Finding the exact distribution is hard

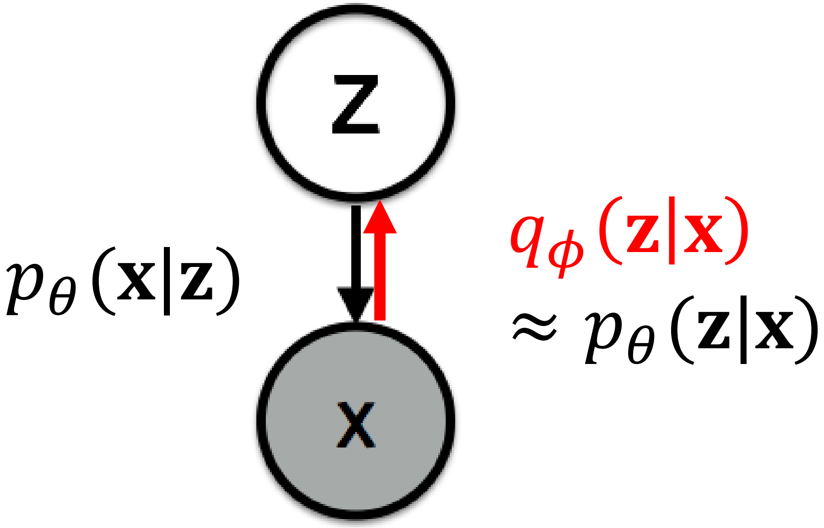
$$f(\theta \mid x) = \frac{f(\theta, x)}{f(x)} = \frac{f(\theta)\, f(x \mid \theta)}{f(x)} \Longrightarrow q(x_{t-1} \mid x_t) = q(x_t \mid x_{t-1}) \frac{q(x_{t-1})}{q(x_t)}$$

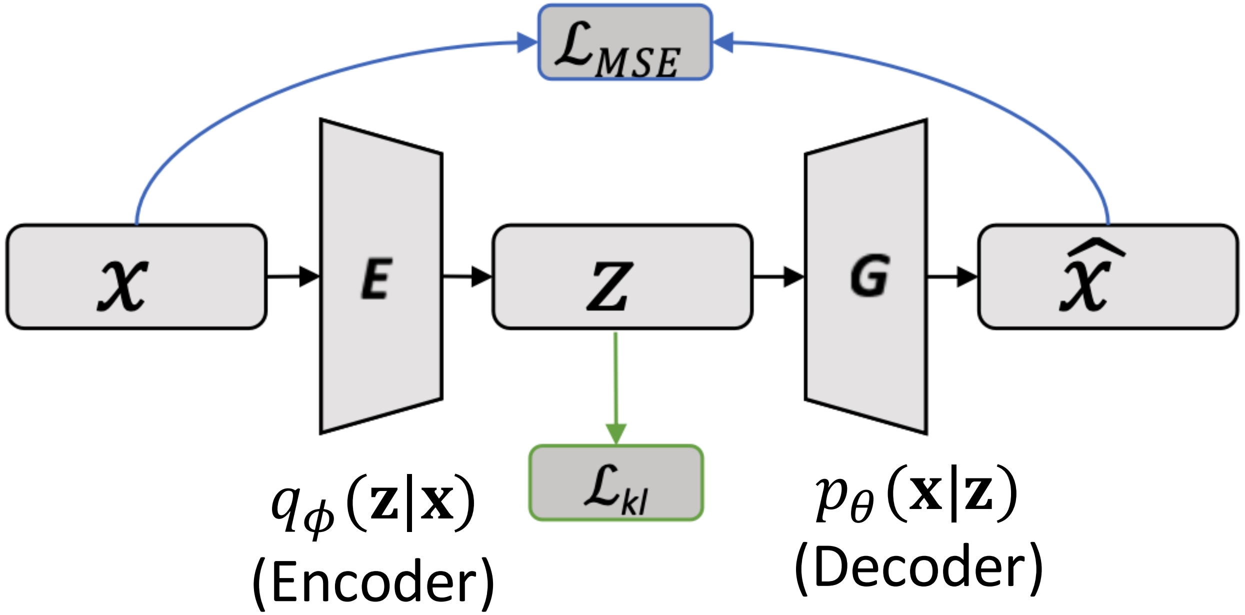
$$q(x_t) = \int q(x_t \mid x_{t-1}) q(x_{t-1}) \mathrm{d}x$$

- The distribution of each timestep and $q(x_t \mid x_{t-1})$ depends on the entire data distribution:
  - Computing this is computationally intractable (where else have we seen this dilemma?)
  - However, we still need those to describe the reverse process. Can we approximate them somehow?

VAE: Variational Autoencoder

- Expensive to compute $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z}) dz$
- Alternatively, we introduce a variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ to approximates the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$?





DDPM: Lower bound of $logP(x)$

VAE Maximize $\log(P_\theta(x))$ $\longrightarrow$ Maximize $\mathbb{E}_{\boxed{q(z|x)}}[\log(\frac{p(x,z)}{q(z|x)})]$

Encoder

Diffusion Maximize $\log(P_\theta(x_0))$ $\longrightarrow$ Maximize $\mathbb{E}_{\boxed{q(x_1:x_T|x_0)}}[\log(\frac{p(x_0:x_T)}{q(x_1:x_T|x_0)})]$

Forward Process
(Diffusion Process)

$$q(x_1{:}\,x_T|x_0) = q(x_1|x_0)q(x_2|x_1)\ldots q(x_T|x_{T-1})$$

What should the distribution look like?

Turns out that for small enough forward steps, i.e.

$$\{\beta_t \in (0,1)\}_{t=1}^{T}$$

the reverse process step $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ can be estimate as a Gaussian distribution too.

Therefore, we can parametrize the learned reverse process as

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

such that

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$

A preliminary objective

The VAE (ELBO) loss is a bound on the true log likelihood (also called the variational lower bound)

$$\boxed{-L_{\text{VAE}} = \log p_\theta(\mathbf{x}) - D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x})) \leq \log p_\theta(\mathbf{x})}$$

Apply the same trick to diffusion:

$$-\log p_\theta(\mathbf{x}_0) \leq \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}\right] = L_{VLB}$$

Expanding out,

$$\begin{aligned} L_{\text{VLB}} &= L_T + L_{T-1} + \cdots + L_0 \\ \text{where } L_T &= D_{\text{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T)) \\ L_t &= D_{\text{KL}}(\boxed{q(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{x}_0)} \parallel p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})) \text{ for } 1 \leq t \leq T-1 \\ L_0 &= -\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \end{aligned}$$

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

(Optional)

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

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] \tag{47}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] \tag{48}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)\prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] \tag{49}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)\prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)}\right] \tag{50}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)}\right] \tag{51}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}}\right] \tag{52}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\cancel{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}}{\cancel{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}}}\right] \tag{53}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}} + \log \frac{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right] \tag{54}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \sum_{t=2}^{T} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right] \tag{55}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\right] + \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right] \tag{56}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + \mathbb{E}_{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\right] + \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t,\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right] \tag{57}$$

$$\boxed{= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}} \quad (58)}$$

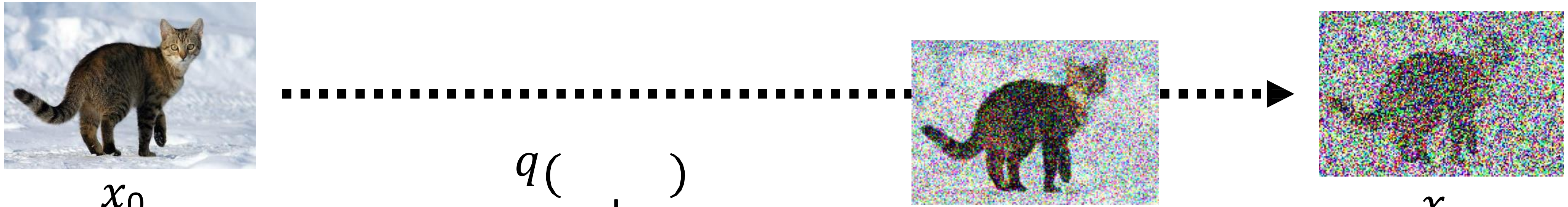
Understanding Diffusion Models: A Unified Perspective: https://arxiv.org/pdf/2208.11970.pdf

A simplified objective

The reverse step conditioned on x_0 is a Gaussian:

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

$$\text{where} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$



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

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

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

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

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

Known Gaussian

(Optional)

CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

$$
\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}
$$

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

A simplified objective

The reverse step conditioned on x_0 is a Gaussian:

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

$$\text{where} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

After doing some algebra, each loss term can be approximated by

$$\begin{aligned} L_{t-1} &= \mathbb{E}_{\mathbf{x}_0, \epsilon}\left[\frac{1}{2\|\boldsymbol{\Sigma}_\theta\|_2^2}\|\tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_{t,}, t)\|_2^2\right] \\ &= \mathbb{E}_{\mathbf{x}_0, \epsilon}\left[\frac{1}{2\|\boldsymbol{\Sigma}_\theta\|_2^2}\left\|\boxed{\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)} - \mu_\theta(\mathbf{x}_{t,}, t)\right\|_2^2\right] \end{aligned}$$

A simplified objective

Instead of predicting the mu, Ho et al. say that we should predict epsilon instead!

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) \Rightarrow \boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)$$

Thus, our loss becomes

$$\begin{aligned}
L_{t-1} &= \mathbb{E}_{\mathbf{x}_0, \epsilon}\left[\frac{1}{2\|\boldsymbol{\Sigma}_\theta\|_2^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right) - \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right)\right\|_2^2\right] \\
&= \mathbb{E}_{\mathbf{x}_0, \epsilon}\left[\frac{\beta_t^2}{2\alpha_t(1-\bar{\alpha}_t)\|\boldsymbol{\Sigma}_\theta\|_2^2}\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2\right] \\
&= \mathbb{E}_{\mathbf{x}_0, \epsilon}\left[\frac{\beta_t^2}{2\alpha_t(1-\bar{\alpha}_t)\|\boldsymbol{\Sigma}_\theta\|_2^2}\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|_2^2\right]
\end{aligned}$$

A simplified objective

- The authors of DDPM say that it's fine to drop all that baggage in the front and instead just use

$$L_{t-1} = \mathbb{E}_{\mathbf{x}_0, \epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|_2^2\right]$$

- Note that this is not a variational lower bound on the log-likelihood anymore: in fact, you can view it as a reweighted version of ELBO that emphasizes reconstruction quality!

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
    $\nabla_\theta \left\| \boxed{\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)} \right\|^2$
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 $\mathbf{z} = \mathbf{0}$
  $\boxed{\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)} + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$

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

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

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

Next

- Introduction
- Theory of diffusion
- Tricks to improve image synthesis models
- Examples of recent diffusion models
  - Text-to-image generation
    - Stable diffusion
    - DALL-E series
    - Imagen
  - …

Thank You

- Questions?

- Email: yu.yin@case.edu

Reference slides and papers

- Hung-Yi Lee. Machine Learning
- Aryan Jain. Machine Learning
- Lillian Weng's Blog: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
- Ho et al., Denoising Diffusion Probabilistic Models: https://arxiv.org/abs/2006.11239
- Understanding Diffusion Models: A Unified Perspective: https://arxiv.org/pdf/2208.11970.pdf