Diffusion Model

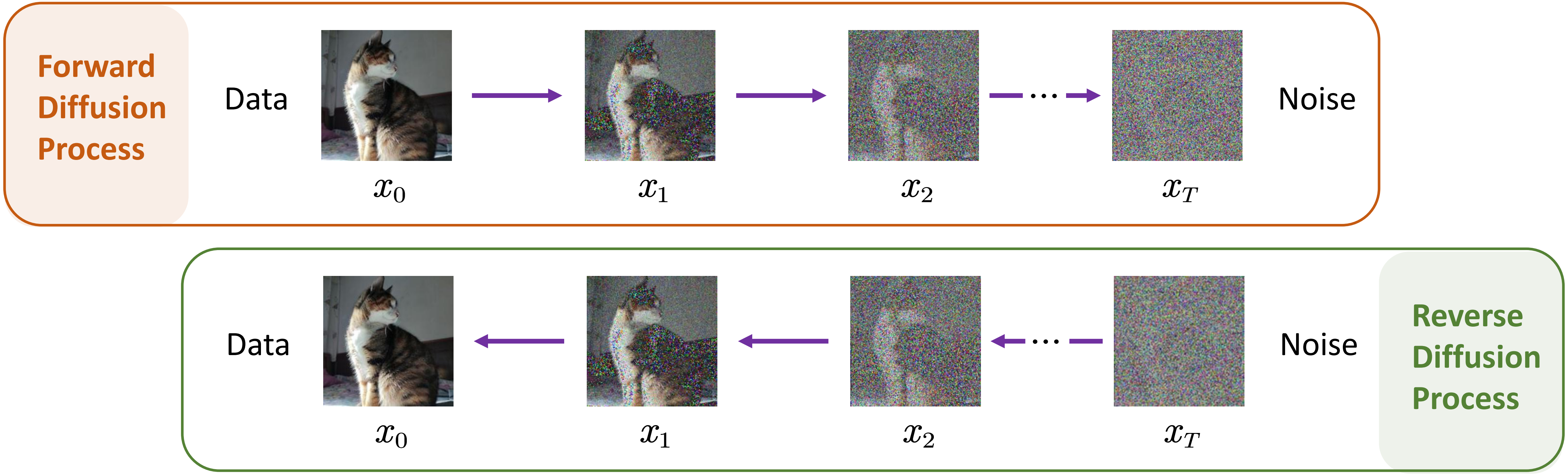
Denoising Diffusion Probabilistic Models

- What is the diffusion model?
- How to visually understand the diffusion model?
- How to derive the diffusion model mathematically?
- How to train a diffusion model and infer it?

Xin Zhang

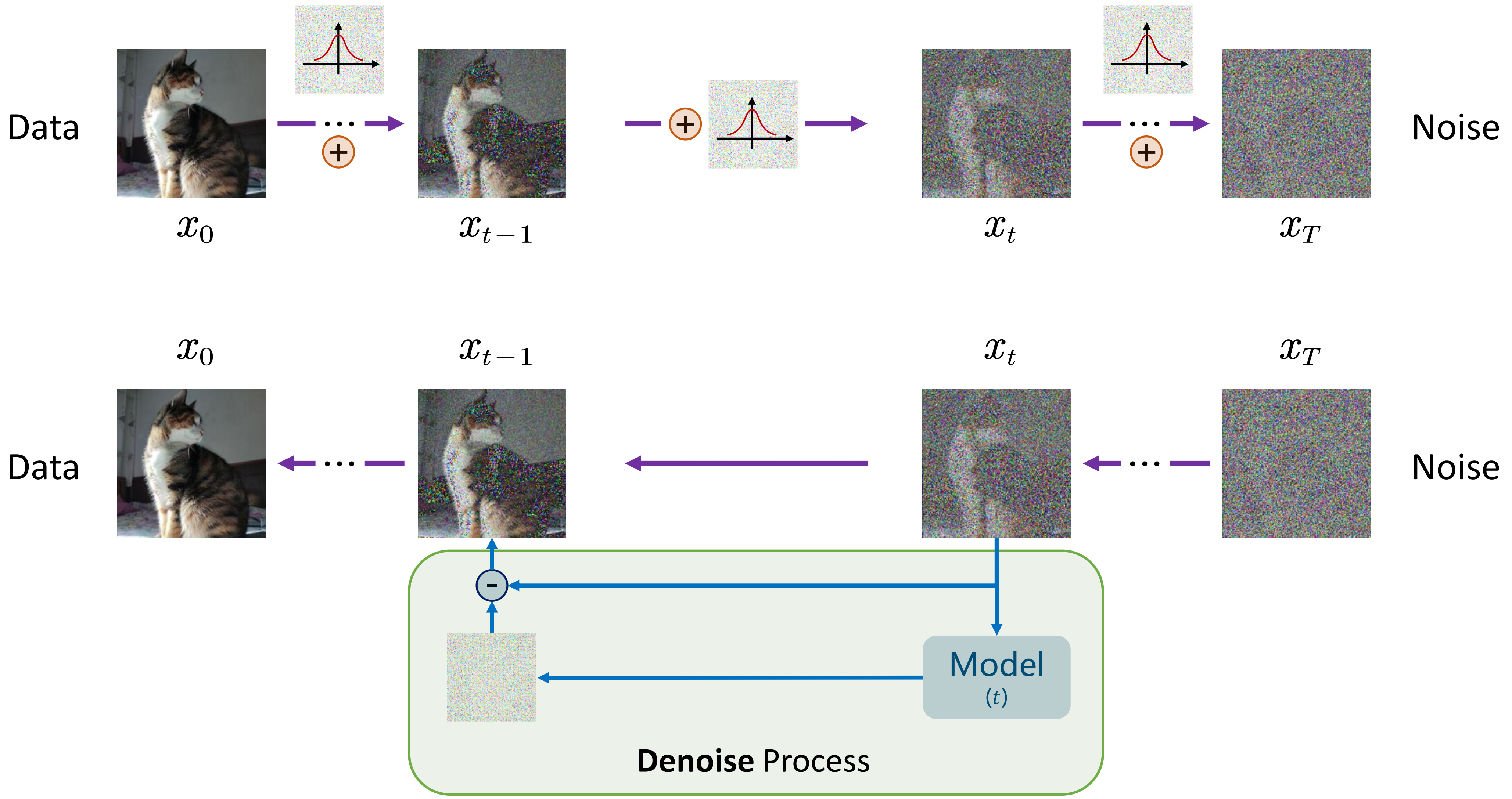
What is Diffusion Model ?

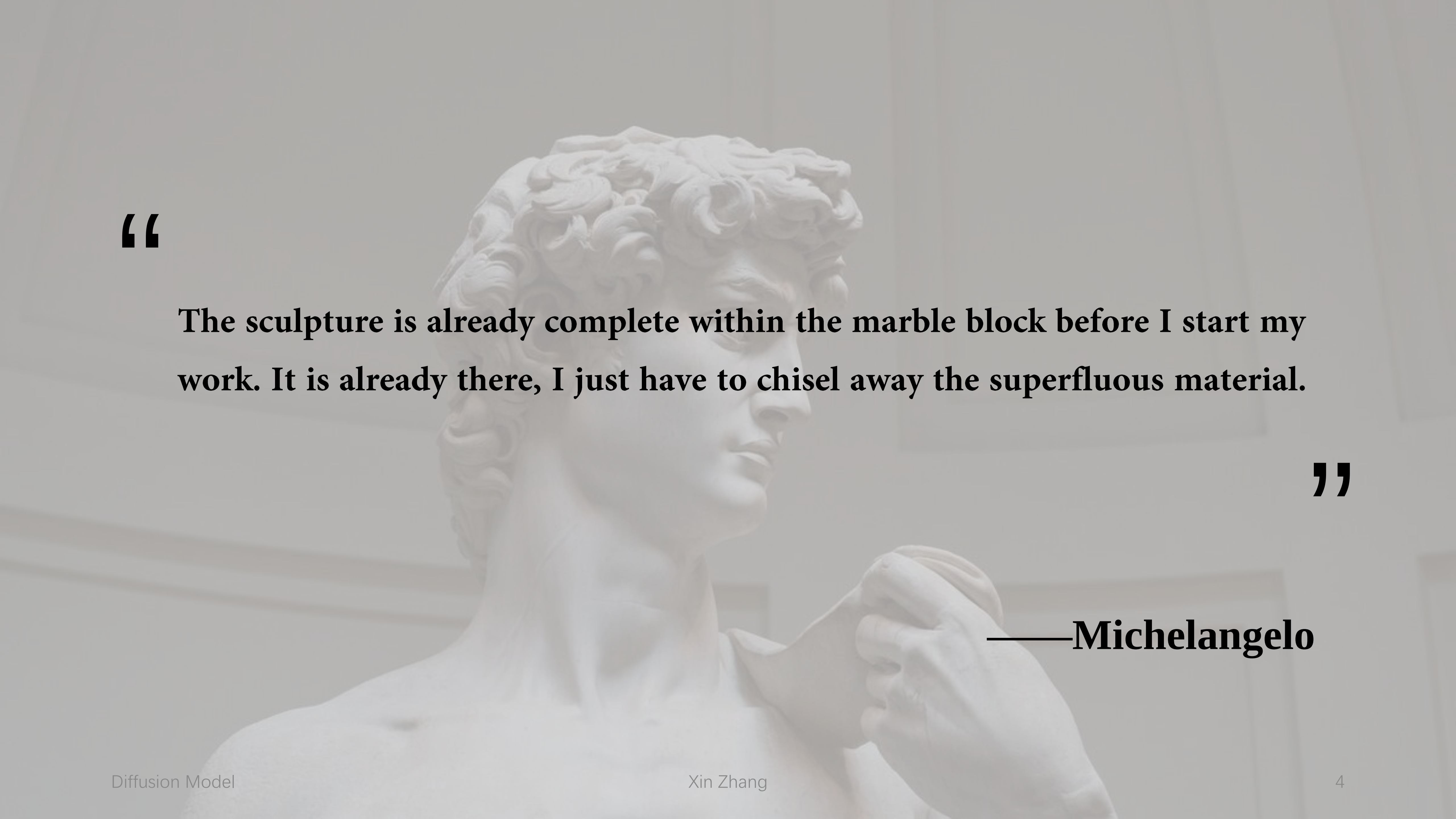
Denoising **D**iffusion **P**robabilistic **M**odels[1]



[1] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. NeuralPS, 2020.

What is Diffusion Model ?





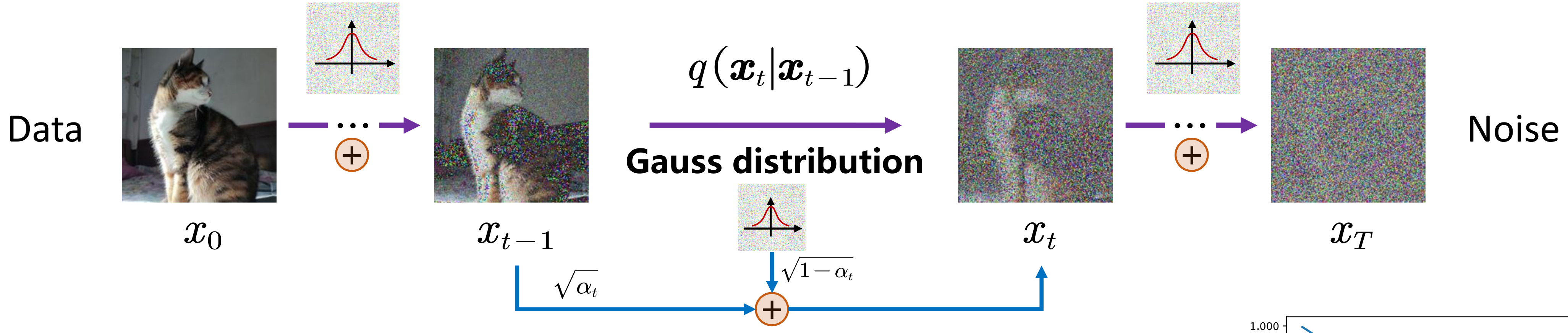
"

The sculpture is already complete within the marble block before I start my work. It is already there, I just have to chisel away the superfluous material.

"

——Michelangelo

Forward Diffusion Process

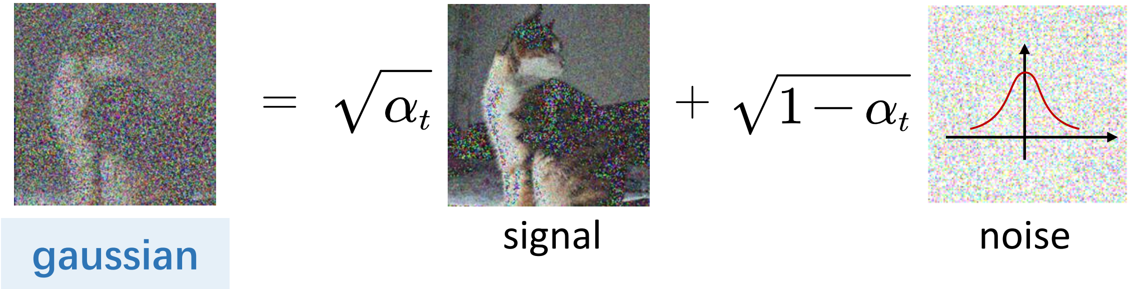


$$z \sim \mathcal{N}(\mu,\ \sigma^2)$$

$$\frac{z-\mu}{\sigma} \sim \mathcal{N}(0,\ I)$$

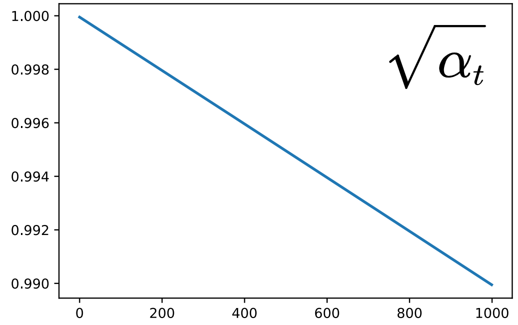
$$z = \mu + \sigma \cdot \varepsilon$$

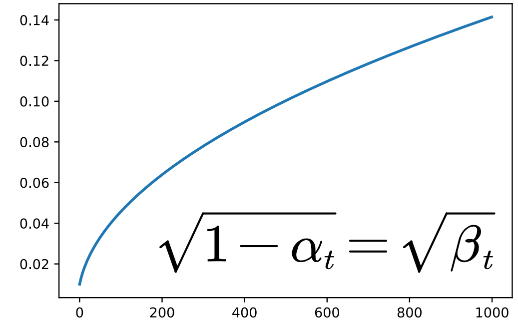
$$\varepsilon \sim \mathcal{N}(0, I)$$



$$\boldsymbol{x}_t = \sqrt{\alpha_t}\,\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\,\boldsymbol{\varepsilon}_{t-1}$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\Big(\boldsymbol{x}_t; \underbrace{\sqrt{\alpha_t}\,\boldsymbol{x}_{t-1}}_{\text{mean}}, \underbrace{(1-\alpha_t)\mathbf{I}}_{\text{variance}}\Big)$$





Forward Diffusion Process

① forward (close-form)

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\,\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\,\boldsymbol{\varepsilon}_{t-1}$$

$$= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\,x_{t-2} + \sqrt{1-\alpha_{t-1}}\,\boldsymbol{\varepsilon}_{t-2}\right) + \sqrt{1-\alpha_t}\,\boldsymbol{\varepsilon}_{t-1}$$

$$= \sqrt{\alpha_t\alpha_{t-1}}\,x_{t-2} + \left(\sqrt{\alpha_t-\alpha_t\alpha_{t-1}}\,\boldsymbol{\varepsilon}_{t-2} + \sqrt{1-\alpha_t}\,\boldsymbol{\varepsilon}_{t-1}\right)$$

$\cdots$

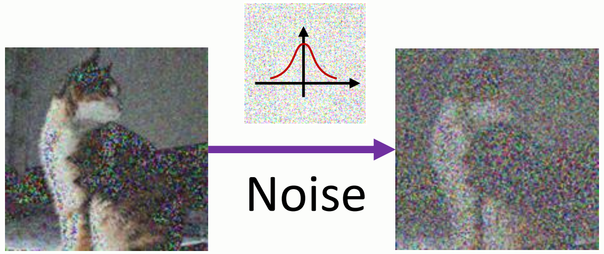
If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ $\quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$

$Z = X + Y$

Then $Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

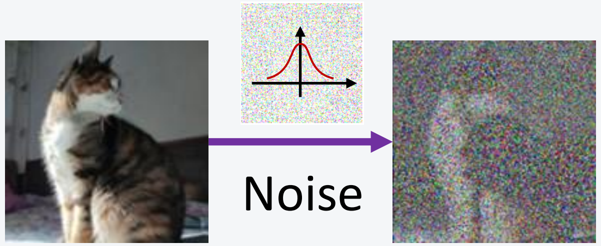
$$= \sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_1}\,\boldsymbol{x}_0 + \sqrt{1-\alpha_t\alpha_{t-1}\cdots\alpha_1}\,\boldsymbol{\varepsilon}$$

$$= \sqrt{\overline{\alpha}_t}\,\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\,\boldsymbol{\varepsilon}$$



x_{t-1} $\qquad$ x_t

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\,\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\,\boldsymbol{\varepsilon}_t$$

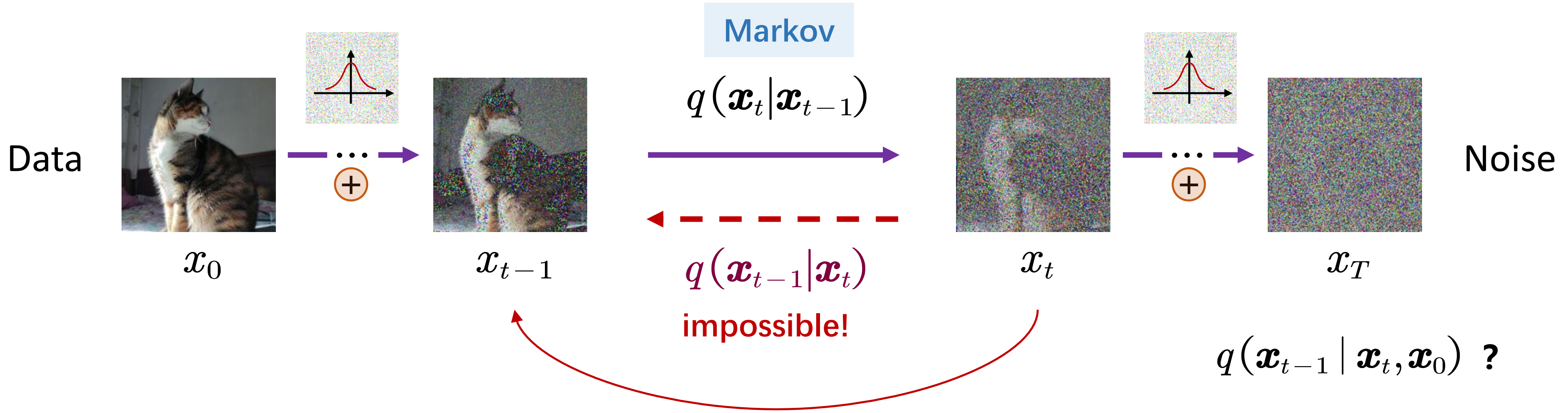


x_0 $\qquad$ x_t

$$\boldsymbol{x}_t = \sqrt{\overline{\alpha}_t}\,\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\,\boldsymbol{\varepsilon}_t$$

$$\overline{\alpha}_t = \alpha_1\alpha_2\ldots\alpha_t$$

Reverse Diffusion Process



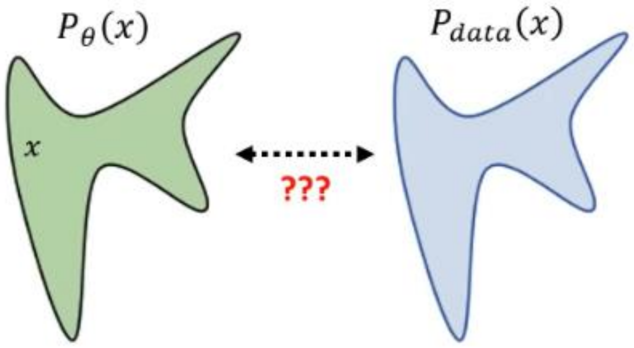
$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ Neural Network

Assume: the output is gaussian

Target Distribution $q(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_t(x_t), \Sigma_t(x_t))$

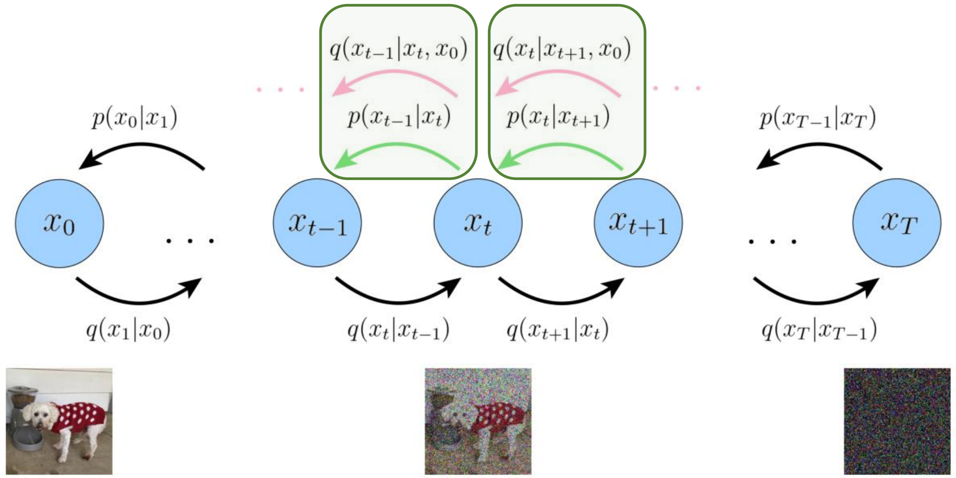
Approximated Distribution $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$

Maximum Likelihood Estimation



$$\arg\max_{\theta} \prod_{i=1}^{t} p_\theta(\boldsymbol{x}_i)$$

$$\arg\max_{\theta} \sum_{i=1}^{t} \log p_\theta(\boldsymbol{x}_i)$$



② optimization (view 1)

$$\min\ -\log p_\theta(x_0) \leqslant -\log p_\theta(x_0) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$\min\ -\log p_\theta(x_0) \leqslant \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\right]$$ **ELBO**

$$\min\ -\log p_\theta(x_0) \leqslant \mathbb{E}_{q(x_{1:T}|x_0)}[D_{KL}(q(x_T|x_0)||p_\theta(x_T))] + \boxed{\sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t))} - \log p_\theta(x_0|x_1)$$

[1] Luo C. Understanding diffusion models: A unified perspective. arXiv, 2022.

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

x_t $\quad$ x_{t-1} $\quad$ x_0

If we know x_0 and x_t

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

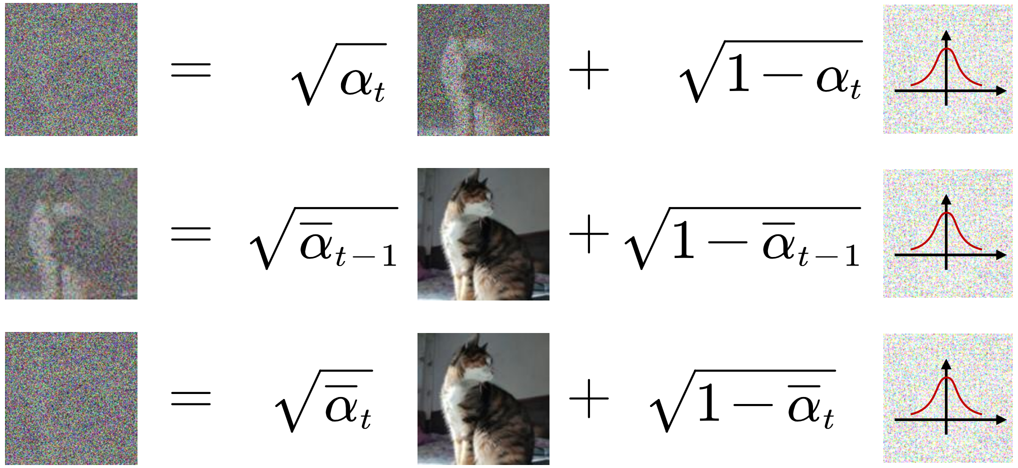
Assume: Markov

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

$$q(x_t \mid x_{t-1}) \sim \mathcal{N}\left(x_t; \sqrt{\alpha_t}\, x_{t-1}, 1 - \alpha_t\right)$$

$$q(x_{t-1} \mid x_0) \sim \mathcal{N}\left(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\, x_0, 1 - \bar{\alpha}_{t-1}\right)$$

$$q(x_t \mid x_0) \sim \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}\, x_0, 1 - \bar{\alpha}_t\right)$$



[1] Luo C. Understanding diffusion models: A unified perspective. arXiv, 2022.

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

$$q(x_t \mid x_{t-1}) \sim \mathcal{N}\left(x_t;\ \sqrt{\alpha_t}\, x_{t-1}, 1-\alpha_t\right)$$

$$q(x_{t-1} \mid x_0) \sim \mathcal{N}\left(x_{t-1};\ \sqrt{\bar{\alpha}_{t-1}}\, x_0, 1-\bar{\alpha}_{t-1}\right)$$

$$q(x_t \mid x_0) \sim \mathcal{N}\left(x_t;\ \sqrt{\bar{\alpha}_t}\, x_0, 1-\bar{\alpha}_t\right)$$

③ **reverse**

If we know x_0

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\Bigg(\boldsymbol{x}_{t-1};\ \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boxed{\boldsymbol{x}_0}}{1-\bar{\alpha}_t}}_{\boldsymbol{\mu}_q(\boldsymbol{x}_t, t)},\ \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{I}}_{\Sigma_q(t)}\Bigg)$$

Assume: fixed

Minimize the distance between two Gaussian distributions (refer to PRML)

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

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) || p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{2\sigma_q^2(t)}[||\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q||_2^2]$$

[1] Luo C. Understanding diffusion models: A unified perspective. arXiv, 2022.

Remove x_0

③ **reverse** If we know x_0

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boxed{\boldsymbol{x}_0}}{1-\bar{\alpha}_t}}_{\boldsymbol{\mu}_q(\boldsymbol{x}_t, t)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{I}}_{\Sigma_q(t)}\right)$$

① **forward (close-form)**

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\varepsilon}_t \quad \Rightarrow \quad \boldsymbol{x}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\varepsilon}_t}{\sqrt{\bar{\alpha}_t}}$$

$$\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t} \quad \Rightarrow \quad \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_t\right)$$

noise predictor

$$\boldsymbol{\varepsilon}_t(\boldsymbol{x}_0 \to \boldsymbol{x}_t)$$

Why not predict $\boldsymbol{x}_0$ directly?

[1] Luo C. Understanding diffusion models: A unified perspective. arXiv, 2022.

Training and Sampling

③ reverse

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boxed{\boldsymbol{x}_0}}{1-\bar{\alpha}_t}}_{\boldsymbol{\mu}_q(\boldsymbol{x}_t,t)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}_{\Sigma_q(t)}\boldsymbol{I}\right)$$

$$\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t} \Rightarrow \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_t\right)$$

④ training

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{2\sigma_q^2(t)}[||\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q||_2^2]$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t}[||\boldsymbol{\varepsilon}_t - \hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t,t)||_2^2]$$

⑤ sampling

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

[1] Luo C. Understanding diffusion models: A unified perspective. arXiv, 2022.

Training and Sampling

Algorithm 1 Training

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

Algorithm 2 Sampling

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

④ training

$$\underset{\boldsymbol{\theta}}{\arg\min} \frac{1}{2\sigma_q^2(t)}\left[\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q\|_2^2\right]$$

$$\underset{\boldsymbol{\theta}}{\arg\min} \boxed{\frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t}}\left[\|\boldsymbol{\varepsilon}_t - \hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)\|_2^2\right]$$

simplify

⑤ sampling

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

[1] Luo C. Understanding diffusion models: A unified perspective. arXiv, 2022.

The variance

③ reverse

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)=\mathcal{N}\left(\boldsymbol{x}_{t-1};\underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t+\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boxed{\boldsymbol{x}_0}}{1-\bar{\alpha}_t}}_{\boldsymbol{\mu}_q(\boldsymbol{x}_t,t)},\underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}_{\Sigma_q(t)}\boldsymbol{I}\right)$$

$$\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}=\frac{\bar{\beta}_{t-1}}{\bar{\beta}_t}\beta_t$$

Fixed-small $\quad {\sigma_t}^2=\frac{\bar{\beta}_{t-1}}{\bar{\beta}_t}\beta_t=\tilde{\beta}_t$ [GLIGEN]

Fixed-large $\quad {\sigma_t}^2=\beta_t$ [DDPM]

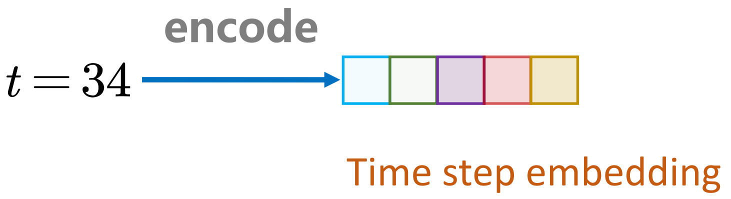
hybrid $\quad {\sigma_t}^2=\exp\left(v\log\beta_t+(1-v)\log\tilde{\beta}_t\right)$ [IDDPM]

optimal [Analytic DPM]

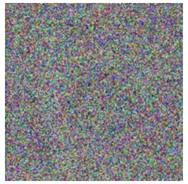
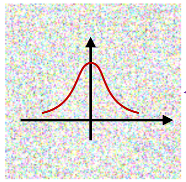
[1] Nichol A Q, Dhariwal P. Improved denoising diffusion probabilistic models. ICML, 2021.
[2] Li Y, Liu H, Wu Q, et al. Gligen: Open-set grounded text-to-image generation. CVPR, 2023.

Illustration of training

1. Randomly select a time step and encode it.



2. Add noise to image.

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

Noisy image | Original image | Gaussian noise

3. Train the UNet.

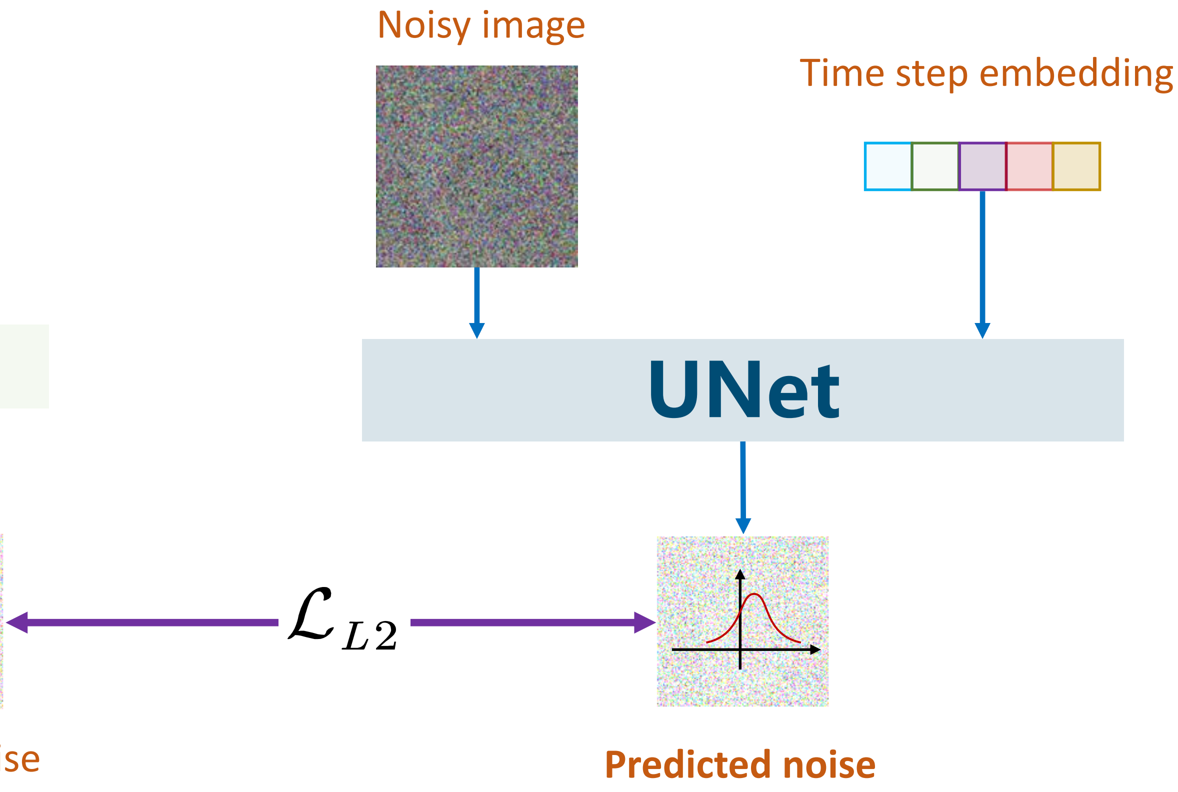
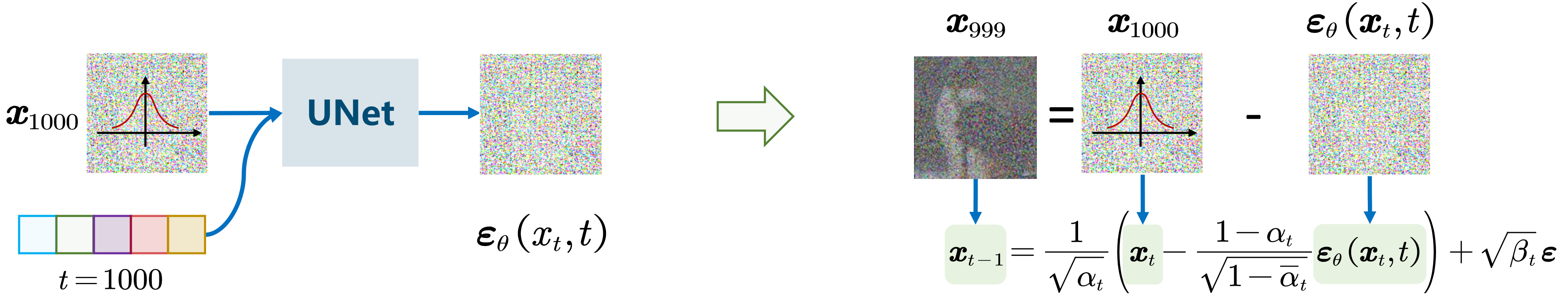


Illustration of sampling

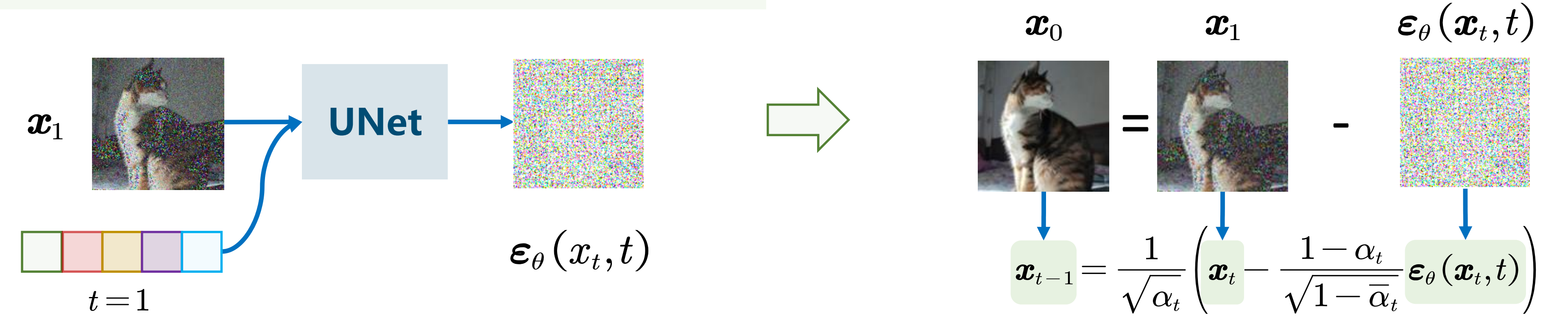
1. Iteratively denoise the image ($T = 1000$)



$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)\right) + \sqrt{\beta_t}\,\boldsymbol{\varepsilon}$$

2. Iteratively denoise the image ($T = 999..2$)

3. Iteratively denoise the image ($T = 1$)



$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)\right)$$

Reference

1. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. NeuraIPS, 2020.
2. Luo C. Understanding diffusion models: A unified perspective. arXiv, 2022.
3. Nichol A Q, Dhariwal P. Improved denoising diffusion probabilistic models. ICML, 2021.
4. https://jalammar.github.io/illustrated-stable-diffusion/
5. https://lilianweng.github.io/posts/2021-07-11-diffusion-models/ (What are diffusion models?)
6. https://www.youtube.com/watch?v=ifCDXFdeaaM&t=210s (李宏毅,【生成式AI】Diffusion Model 原理剖析 (1/4) (optional))
7. https://kexue.fm/archives/9119. (苏剑林, 生成扩散模型漫谈（一）：DDPM = 拆楼 + 建楼)
8. https://www.bilibili.com/video/BV19H4y1G73r (SY_007,【较真系列】讲人话-Diffusion Model全解(原理+代码+公式))
9. https://www.bilibili.com/video/BV1b541197HX (deep_thoughts, 54、Probabilistic Diffusion Model概率扩散模型理论与完整PyTorch代码详细解读)
10. https://www.bilibili.com/video/BV1p24y1K7Pf (Nik_Li, 一个视频看懂扩散模型DDPM原理推导|AI绘画底层模型)

The end

非常感谢你能看到这，希望该课件对你有帮助，视频讲解版在B站。

课件中出现的是我家的猫咪的照片，她已经陪伴了我很多年了，感谢她的友情出镜。o(*￣▽￣*)ブ

Thank you so much for seeing this, I hope the slide is helpful, the video explanation version is on Bilibili.

The picture on the slide is of my cat, who has been with me for many years now, thanks for her friendly appearance! o(*￣▽￣*)ブ

> **“ What I cannot create, I do not understand. ”**
>
> **——Richard Feynman**