Diffusion Model

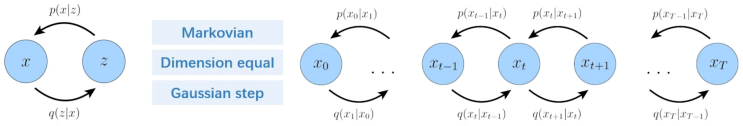
Algorithm 1 Training

repeat
  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
  $t \sim \text{Uniform}(\{1, \dots, T\})$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  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 $\mathbf{x}_0$

Relationship with diffusion



⑥ ELBO $\log p(\boldsymbol{x}) \geqslant \mathbb{E}_{q_\phi(z|\boldsymbol{x})} \log \left[\frac{p(z,\boldsymbol{x})}{q_\phi(z \mid \boldsymbol{x})} \right]$

② optimization (view 1)

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

⑧ optimization (view 2)

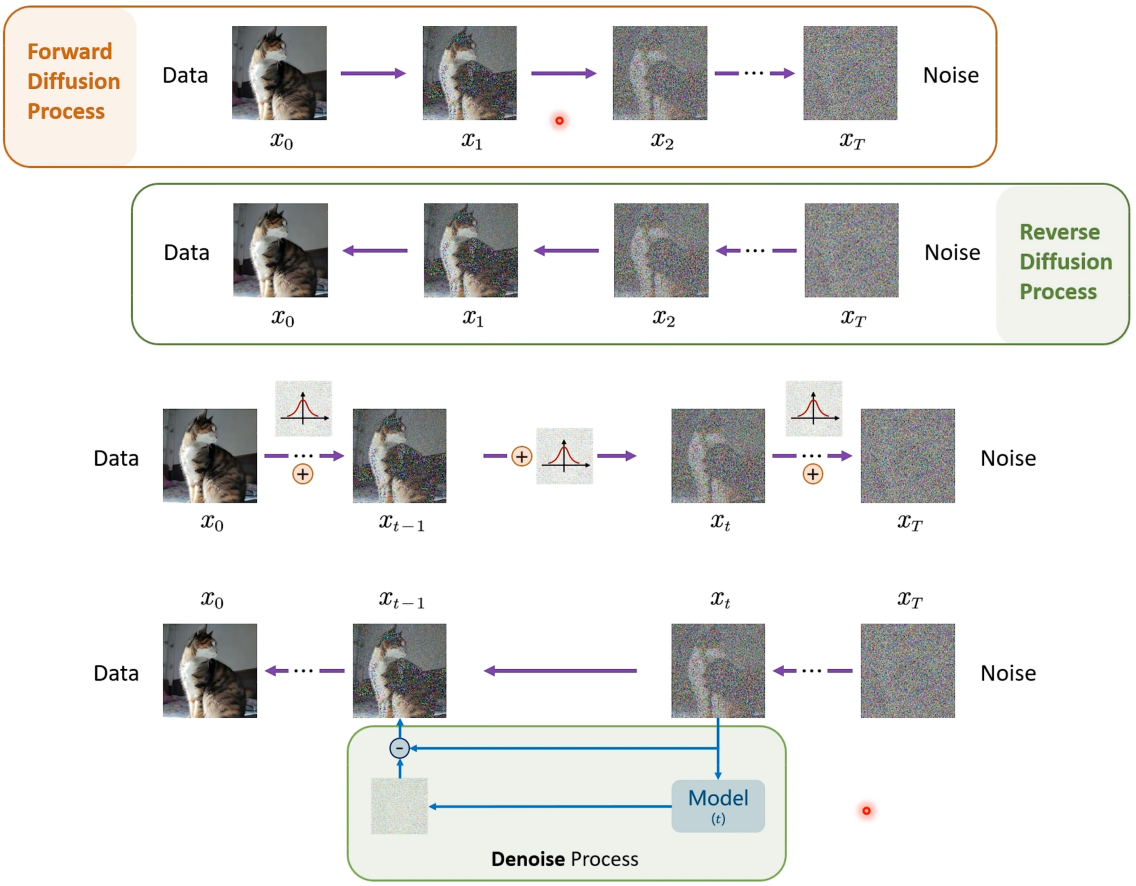
$\log p(\boldsymbol{x}) \geqslant \mathbb{E}_{q_\phi(z_{1:T}|\boldsymbol{x})} \log \left[\frac{p(z_{1:T},\boldsymbol{x})}{q_\phi(z_{1:T} \mid \boldsymbol{x})} \right]$

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

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

prior matching term | consistency term | reconstruction term

Denoising Diffusion Probabilistic Models[1]



Forward Process

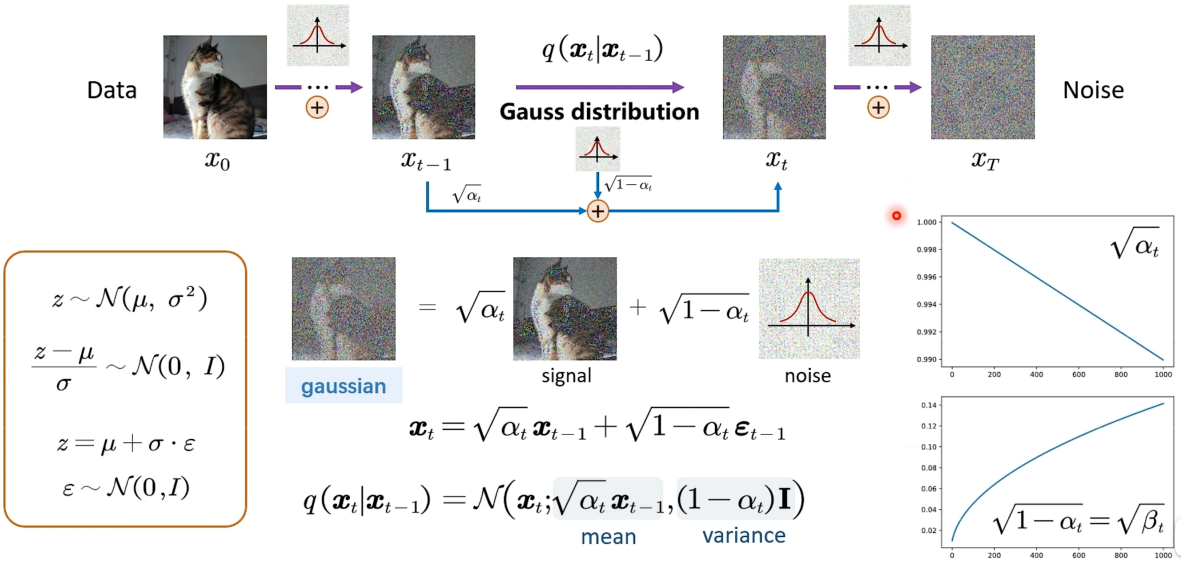
输入样本 x_0 通过加 T 次噪声达到 x_T, $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$。Markov 过程被定义为：

$$x_t \sim \mathcal{N}(\mu_t x_{t-1}, \sigma_t^2 \mathbf{I}) = \mu_t x_{t-1} + \epsilon \odot \sigma_t \mathbf{I}$$

Diffusion Noise 过程被定义为：

$$\begin{cases} \mu_t = \sqrt{\alpha_t} \\ \sigma_t = \sqrt{1-\alpha_t} \end{cases} \Longrightarrow \begin{cases} \mu_t = \sqrt{1-\beta_t} \\ \sigma_t = \sqrt{\beta_t} \end{cases}$$

$$\begin{aligned} x_t \sim \mathcal{N}\left(\sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}\right) \Rightarrow x_t &= \sqrt{1-\beta_t} x_{t-1} + \sqrt{\beta_t} \varepsilon_{t-1} \\ &= \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} \varepsilon' \end{aligned}$$



我们使用一个加噪音 Dist，其被定义为：

$$x_t = \underbrace{\sqrt{\alpha_t} x_{t-1}}_{\mu} + \underbrace{\overbrace{\sqrt{\beta_t}}^{\sqrt{1-\alpha_t}}\, \varepsilon_{t-1}}_{\sigma}, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

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

通过多次迭代，可发现：

$$\begin{aligned} x_t &= \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} \varepsilon_{t-1} \\ &= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_{t-1}} \varepsilon_{t-2}) + \sqrt{1-\alpha_t} \varepsilon_{t-1} \\ &= \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \underbrace{(\sqrt{\alpha_t(1-\alpha_t)} \varepsilon_{t-2} + \sqrt{1-\alpha_t} \varepsilon_{t-1})}_{\sqrt{\alpha_t(1-\alpha_t)+(1-\alpha_t)}\varepsilon'} \\ &= \cdots \end{aligned}$$

$$= \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_1} x_0 + \underbrace{\sqrt{1 - \alpha_t \alpha_{t-1} \cdots \alpha_1}}_{\Gamma_\varepsilon} \varepsilon'$$
$$= \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon'$$
$$\bar{\alpha}_t \triangleq \alpha_t \alpha_{t-1} \cdots \alpha_1$$

对于两个高斯分布 $X \sim \mathcal{N}(\mu_X, \sigma_X^2), Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. 若 $Z = X + Y$，则有：

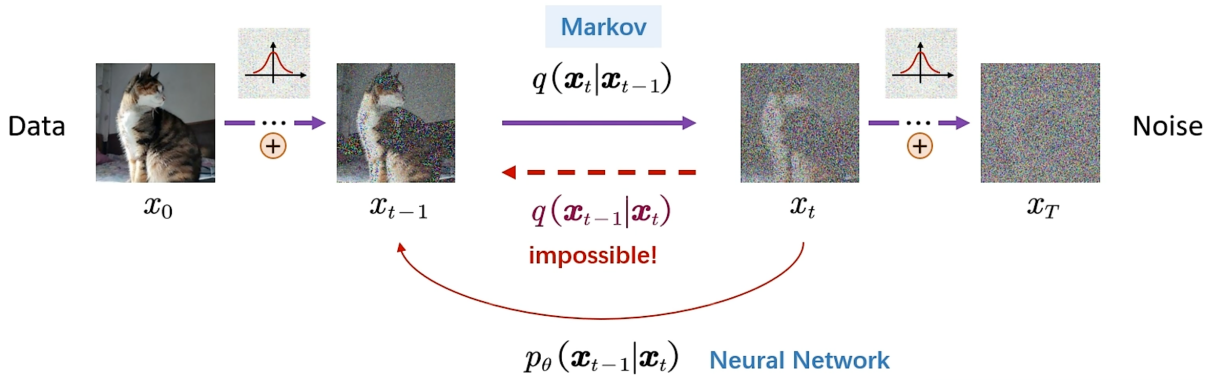
$$Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Γ_ε (ε 的系数)如下：

$$\begin{aligned}
&\varepsilon_{t-1} \Rightarrow \sqrt{1 - \alpha_t} \\
&\varepsilon_{t-2} \Rightarrow \sqrt{\alpha_t (1 - \alpha_{t-1})} \\
&\varepsilon_{t-2} \Rightarrow \sqrt{\alpha_t \alpha_{t-1} (1 - \alpha_{t-2})} \\
&\cdots \\
&\varepsilon_0 \Rightarrow \sqrt{\alpha_t \alpha_{t-1} \ldots \alpha_2 (1 - \alpha_1)}
\end{aligned}$$

$$\begin{aligned}
S &= (1 - \alpha_t) + [\alpha_t (1 - \alpha_{t-1})] + [\alpha_t \alpha_{t-1} (1 - \alpha_{t-2})] + \cdots + [\alpha_t \alpha_{t-1} \ldots \alpha_2 (1 - \alpha_1)] \\
&= 1 - \alpha_t + \alpha_t - \alpha_t \alpha_{t-1} + \alpha_t \alpha_{t-1} - \alpha_t \alpha_{t-1} \alpha_{t-2} + \cdots + \alpha_t \alpha_{t-1} \ldots \alpha_2 - \alpha_t \alpha_{t-1} \ldots \alpha_2 \alpha_1 \\
&= 1 - \alpha_t \alpha_{t-1} \ldots \alpha_2 \alpha_1
\end{aligned}$$

Reverse Process



当 β_t $(i.e., 1 - \alpha_t)$ 足够小时，Noise 的逆操作也符合正态分布，也就是：

$$x_{t-1} \sim \mathcal{N}(\tilde{\mu}_t, \tilde{\sigma}_t^2 \mathbf{I})$$

这里的 $\tilde{\mu}_t, \tilde{\beta}_t^2$ 是由当前时刻 t 和图像 x_t 决定的，因此我们需要 NN 去拟合：

$$\tilde{\mu}_t, \tilde{\sigma}_t^2 = p_\theta(x_{t-1} \mid t, x_t)$$

直接寻找反向概率分布 $q(x_{t-1} \mid x_t)$ 是不可能的，使用 NN $p_\theta(x_{t-1} \mid x_t)$以拟合

目标分布 $q(\,x_{t-1}\mid x_t\,)=\mathcal{N}\big(x_{t-1};\mu_t(x_t),\Sigma_t(x_t)\big)$

近似分布 $p_\theta(\,x_{t-1}\mid x_t\,)=\mathcal{N}\big(x_{t-1};\mu_\theta(x_t,t),\Sigma_\theta(x_t,t)\big)$

考虑 $p(x_t\mid x_{t-1})$ 是一个 Markov 过程，而$q(x_{t-1}\mid x_t)$ 也是一个 Markov 过程，因此

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

这个概率分布左侧是不可知的，但是右侧我们已知:

$$\begin{aligned}
q(\,x_t\mid x_{t-1},x_0\,)&=\mathcal{N}(x_t;\sqrt{\alpha_t}x_{t-1},(1-\alpha_t)\mathbf{I})\\
q(\,x_{t-1}\mid x_0\,)&=\mathcal{N}(x_{t-1};\sqrt{\bar\alpha_{t-1}}x_0,(1-\bar\alpha_{t-1})\mathbf{I})\\
q(\,x_t\mid x_0\,)&=\mathcal{N}(x_t;\sqrt{\bar\alpha_t}x_0,(1-\bar\alpha_t)\mathbf{I})
\end{aligned}$$

也就是右侧全都知道了。

$$\mathcal{N}(\tilde\mu_t,\tilde\sigma_t^2\mathbf{I})=\frac{\mathcal{N}(x_t;\sqrt{\alpha_t}x_{t-1},(1-\alpha_t)\mathbf{I})\mathcal{N}(x_{t-1};\sqrt{\bar\alpha_{t-1}}x_0,(1-\bar\alpha_{t-1})\mathbf{I})}{\mathcal{N}(x_t;\sqrt{\bar\alpha_t}x_0,(1-\bar\alpha_t)\mathbf{I})}$$

$$\begin{aligned}
&\log\mathcal{N}(\tilde\mu_t,\tilde\sigma_t^2\mathbf{I})\\
&\quad=\log\mathcal{N}(x_t;\sqrt{\alpha_t}x_{t-1},(1-\alpha_t)\mathbf{I})+\log\mathcal{N}(x_{t-1};\sqrt{\bar\alpha_{t-1}}x_0,(1-\bar\alpha_{t-1})\mathbf{I})\\
&\quad-\log\mathcal{N}(x_t;\sqrt{\bar\alpha_t}x_0,(1-\bar\alpha_t)\mathbf{I})
\end{aligned}$$

$$\log\mathcal{N}(x;\mu,\sigma^2)=-\frac{1}{2\sigma}(x-\mu)^2+C$$

$$\begin{aligned}
&\log\mathcal{N}(\tilde\mu_t,\tilde\sigma_t^2\mathbf{I})\\
&\quad=-\frac{1}{2}\Big\{\Big[\frac{1}{1-\alpha_t}(x_t-\sqrt{\alpha_t}x_{t-1})^2+C_1\Big]\\
&\quad+\Big[\frac{1}{1-\bar\alpha_{t-1}}(x_{t-1}-\sqrt{\bar\alpha_{t-1}}x_0)^2+C_2\Big]-\Big[\frac{1}{1-\bar\alpha_t}(x_t-\sqrt{\bar\alpha_t}x_0)^2+C_3\Big]\Big\}
\end{aligned}$$

可以求出

$$\begin{cases}
\tilde\mu_t=\dfrac{1}{\sqrt{\alpha_t}}\left(x_t-\dfrac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}}\varepsilon_t\right)\\
\tilde\sigma_t^2=\dfrac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\cdot\beta_t
\end{cases}$$

会发现 $\tilde\sigma_t^2$ 不依赖 x_t 是常量。模型去拟合 $\tilde\mu_t$ 即可。而 $\tilde\mu_t$ 中只有 ε_t不知道，因此只需要拟合 ε_t 即可。即 $\epsilon_\theta(x_t,t)\to\varepsilon_t$。而误差可以定义为：

$$L=\|\varepsilon_t-\epsilon_\theta(x_t,t)\|_2^2$$

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

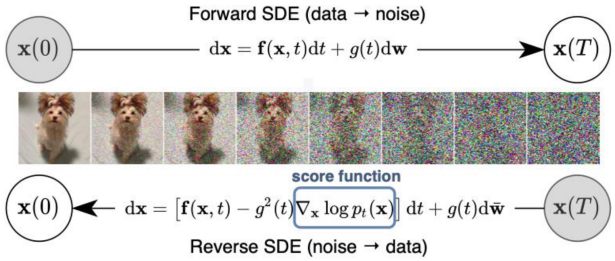
1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}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

连续时间极限（Continuous time limit）

DDPM 原始形式是离散时间下的马尔可夫链，通过预定义的时间步长(如 1000 步)，逐步向数据添加高斯噪声。这个过程是离散的、跳跃的，每一步都对应一个固定的噪声水平。

SDE（随机微分方程）形式（连续时间）将离散过程推广到连续域，使模型具有更严格的数学基础，可以研究扩散过程的连续极限性质，能够应用随机微分方程的丰富理论工具。

采样时可以灵活选择时间步长，可以设计自适应的采样策略，便于分析和改进模型性能。



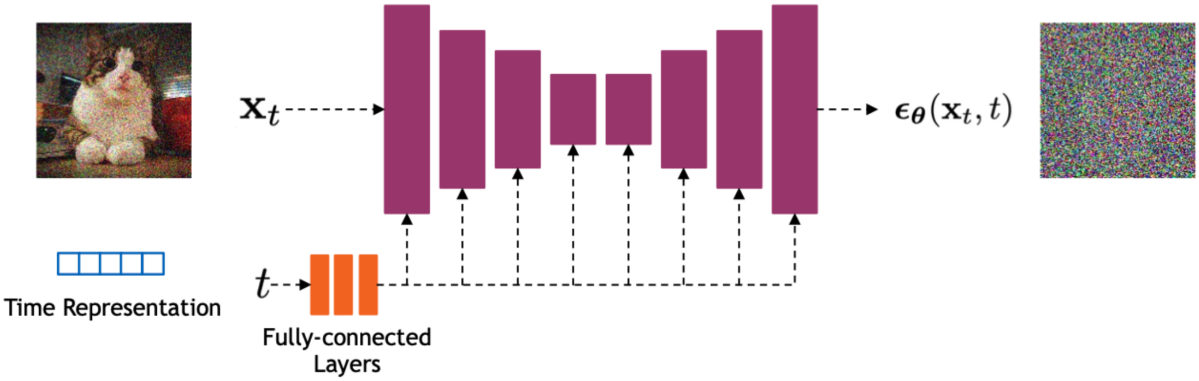
Forward Process：$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t \ + \ g(t)\mathrm{d}\mathbf{w}$
- $\mathbf{f}(\mathbf{x}, t)$：漂移项，控制数据的确定性变化
- $g(t)$：扩散系数，控制噪声的强度
- $\mathrm{d}\mathbf{w}$：维纳过程，提供随机性

Reverse Process：$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - \ g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \ ]\mathrm{d}t \ + \ g(t)\mathrm{d}\bar{\mathbf{w}}$
- $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$：评分函数，估计数据分布的对数梯度。评分函数学习数据分布的几何特征，在反向过程中指导去噪方向。

SDE 形式是 DDPM 的连续时间推广，当时间步长趋于无穷小时，离散 DDPM 会收敛到这个 SDE 形式。可视为 DDPM 的一个更一般的数学框架。

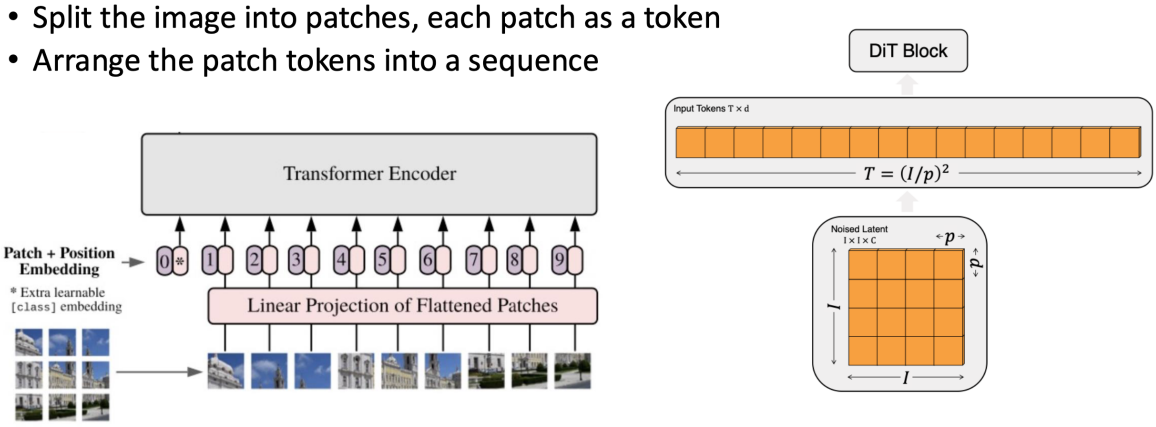
Architecture

U-Net



ViT

- Vision Transfomer (ViT) based approach:
  - Split the image into patches, each patch as a token
  - Arrange the patch tokens into a sequence



图像预处理:

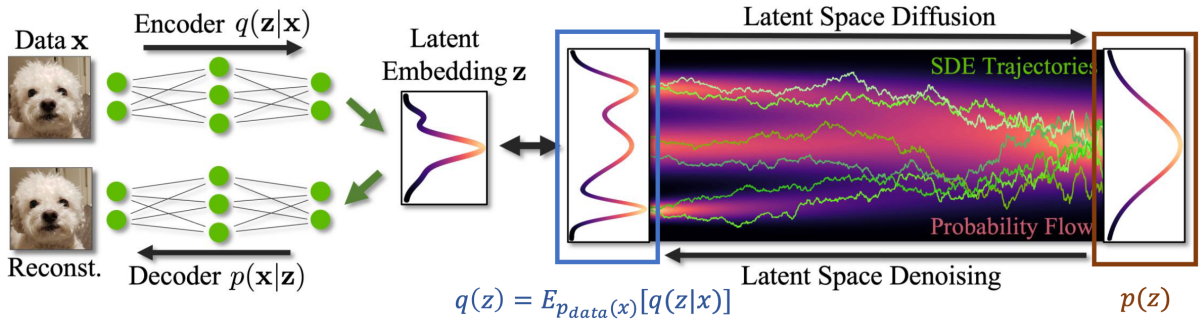
- 将输入图像分割成固定大小的补丁(patches)
- 每个补丁被展平(flatten)成一维向量
- 对每个补丁添加位置编码(position embedding)
- 添加一个特殊的[class]标记作为序列的开始

$$x: \begin{cases} \text{带噪声的图像补丁序列} \\ \text{时间步长 t 的编码} \end{cases} \quad \Longrightarrow \quad y: \text{每个补丁位置的噪声}$$

时间步长 t 的编码可以加入到 position embedding 中

潜在扩散模型 (Latent Diffusion Models, LDM)

U-Net 和 ViT 在处理高分辨率图像时计算成本很高，通过在低维潜在空间进行扩散可以显著降低计算复杂度。



Encoder $q(z \mid x)$	潜在空间扩散	Decoder $p(x \mid z)$
将高维图像压缩到低维潜在空间，减少需要处理的数据维度	在压缩后的潜在空间中进行扩散过程 使用 SDE(随机微分方程)轨迹描述噪声添加过程	将处理后的潜在表示重建回图像空间 恢复图像的细节和质量

Mathematics Behind Diffusion

$$\mathcal{N}(x;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

给定概率

$$p(x)=\mathcal{N}(x;\mu_x,\sigma_x^2)\qquad p(y\mid x)=\mathcal{N}(y;ax,\sigma_y^2)$$

则 Marginal Distribution $p(y)$ 和 Posterior Distribution $p(x\mid y)$ 也是正态分布

$$p(y)=\mathcal{N}(y;ax,a^2\sigma_x^2+\sigma_y^2)$$
$$p(\,x\mid y\,)=\mathcal{N}(x;\tilde{\mu},\tilde{\sigma}^2),\qquad \frac{1}{\tilde{\sigma}^2}=\frac{1}{\sigma_x^2}+\frac{a^2}{\sigma_y^2},\tilde{\mu}=\tilde{\sigma}^2(a\sigma_y^{-1}y+\sigma_x^{-1}\mu_x)$$

$$p(\,x\mid y\,)=\frac{p(y\mid x)p(x)}{p(y)}=\frac{\mathcal{N}(y;ax,\sigma_y^2)\mathcal{N}(x;\mu_x,\sigma_x^2)}{\mathcal{N}(y;ax,a^2\sigma_x^2+\sigma_y^2)}$$

考虑 $p(y)$ 是归一化系数，因此：

$$\begin{aligned}
p(\,x\mid y\,)&\equiv p(x)p(\,y\mid x\,)\\
&\propto\exp\left[-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right]\exp\left[-\frac{1}{2}\left(\frac{y-ax}{\sigma_y}\right)^2\right]\\
&=\exp\left[-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}\right]\exp\left[-\frac{1}{2}\frac{(y-ax)^2}{\sigma_y^2}\right]\\
&=\exp\left[-\frac{1}{2}\left\{\frac{(x-\mu_x)^2}{2\sigma_x^2}+\frac{(y-ax)^2}{2\sigma_y^2}\right\}\right]\\
&=\exp\left[-\frac{1}{2}\left\{\frac{x^2+\mu_x^2-2\mu_x x}{\sigma_x^2}+\frac{y^2+a^2x^2-2axy}{\sigma_y^2}\right\}\right]\\
&=\exp\left[-\frac{1}{2}\left\{x^2\left(\frac{1}{\sigma_x^2}+\frac{a^2}{\sigma_y^2}\right)-2x\left(\frac{\mu_x}{\sigma_x^2}+\frac{ay}{\sigma_x^2}\right)+\left(\frac{\mu_x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)\right\}\right]
\end{aligned}$$

对比系数可知

$$\left(\frac{x-\mu}{\sigma}\right)^2=\frac{x^2-2\mu x+\mu^2}{\sigma^2}=\frac{x^2}{\sigma^2}-\frac{2\mu x}{\sigma^2}+\frac{\mu^2}{\sigma^2}$$

$$\frac{1}{\sigma^2}=\frac{1}{\sigma_x^2}+\frac{a^2}{\sigma_y^2}\Longrightarrow\sigma^2=\frac{\sigma_x^2\sigma_y^2}{\sigma_y^2+a\sigma_x}$$

$$\mu=\underbrace{\left(\frac{\sigma_x^2\sigma_y^2}{\sigma_y^2+a\sigma_x}\right)}_{\sigma^2}\overbrace{\left(\frac{\mu_x}{\sigma_x^2}+\frac{ay}{\sigma_x^2}\right)}^{\mu\sigma^{-2}}_{a\sigma_y^{-1}y+\sigma_x^{-1}\mu_x}\Longrightarrow\mu=\frac{\sigma_y^2\mu_x+a\sigma_x^2y}{\sigma_y^2+a\sigma_x^2}$$

$$p(\,x\mid y\,)=\mathcal{N}\left(x;\frac{\sigma_x^2\sigma_y^2}{\sigma_y^2+a\sigma_x},\frac{\sigma_y^2\mu_x+a\sigma_x^2y}{\sigma_y^2+a\sigma_x^2}\right)$$

Probabilistic Graphical Models (PGM)

考虑 DAG $x_{t-1} \to x_t \to x_{t+1}$:

$$p(\, x_{t-1}, x_{t+1} \mid x_t \,) = \frac{p(x_{t-1}, x_t, x_{t+1})}{p(x_t)} = \frac{p(x_{t-1}, x_t) p(\, x_{t+1} \mid x_t, x_{t-1} \,)}{p(x_t)} = p(x_{t-1} \mid x_t) p(x_{t+1} \mid x_t)$$

Maximise Likelihood Estimate (MLE)

$$\begin{aligned}
\theta^* &= \operatorname*{argmax}_\theta \prod_{x \in \mathcal{X}} p_\theta(x) \quad = \operatorname*{argmax}_\theta \log \prod_{x \in \mathcal{X}} p_\theta(x) \\
&= \operatorname*{argmax}_\theta \sum_{x \in \mathcal{X}} \log p_\theta(x) \approx \operatorname*{argmax}_\theta \mathbb{E}_{x \sim p_{data}} [\log p_\theta(x)] \\
&= \operatorname*{argmax}_\theta \int_x p_{data}(x) \log p_\theta(x) \, dx \\
&= \operatorname*{argmax}_\theta \int_x p_{data}(x) \log p_\theta(x) \, dx - \int_x p_{data}(x) \log p_{data}(x) \, dx \\
&= \operatorname*{argmax}_\theta \int_x p_{data}(x) \log \frac{p_\theta(x)}{p_{data}(x)} dx \\
&= \operatorname*{argmax}_\theta - KL[p_{data}(x) \parallel p_\theta(x)] \\
&= \operatorname*{argmin}_\theta KL[p_{data}(x) \parallel p_\theta(x)]
\end{aligned}$$

VAE and Diffusion

VAE:

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

$$= \underbrace{\mathbb{E}_{q(z|x)} \left[\log \left(\frac{p_\theta(x, z)}{q(z \mid x)} \right) \right]}_{\mathcal{L}_{\text{VAE}}} + KL[q(\, z \mid x \,) \parallel p(z)]$$

$$\text{maxmise } \mathbb{E}_{q(z|x)} \left[\log \left(\frac{p_\theta(x, z)}{q(z \mid x)} \right) \right]$$

DDPM:

$$\log p_\theta(x_0) = \int_{x_1 : x_T} p(x_T) p_\theta(\, x_{T-1} \mid x_T \,) \cdots p_\theta(\, x_{t-1} \mid x_t \,) \cdots p_\theta(\, x_0 \mid x_1 \,) dx_1 : x_T$$

$$\text{maxmise } \mathbb{E}_{q(x_1 : x_T | x_0)} \left[\log \left(\frac{p(x_1 : x_T)}{q(x_1 : x_T \mid x_0)} \right) \right]$$

Down-Top: $q(\, x_1 : x_T \mid x_0 \,) = q(\, x_1 \mid x_0 \,) q(\, x_2 \mid x_1 \,) \cdots q(\, x_T \mid x_{T-1} \,) = \prod_{t=1}^{T} q(x_t \mid x_{t-1})$

如考虑 Markov 情况：

$$
\begin{aligned}
& q(\, x_1{:}x_T \mid x_0\,) \\
&= q(\, x_T \mid x_{T-1}, \dots, x_0\,) q(\, x_{T-1}, \dots, x_0 \mid x_0\,) \\
&= q(\, x_T \mid x_{T-1}, \dots, x_0\,) q(\, x_{T-1} \mid x_{T-2}, \dots, x_0\,) q(\, x_{T-2}, \dots, x_0 \mid x_0\,) \\
&= \cdots \\
&= q(\, x_T \mid x_{T-1}, \dots, x_0\,) q(\, x_{T-1} \mid x_{T-2}, \dots, x_0\,) q(\, x_{T-1} \mid x_{T-3}, \dots, x_0\,) \dots q(\, x_1 \mid x_0\,) \\
&= q(\, x_T \mid x_{T-1}\,) q(\, x_{T-1} \mid x_{T-2}\,) \dots q(\, x_1 \mid x_0\,)
\end{aligned}
$$

Top-Down: $q(\, x_1{:}x_T \mid x_0\,) = q(\, T \mid x_0\,) \prod_{t=2}^{T} q(x_{t-1} \mid x_t, 0)$

考虑

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

$$
\begin{aligned}
q(\, T \mid x_0\,) \prod_{t=2}^{T} q(x_{t-1} \mid x_t, 0) &= q(\, x_T \mid x_0\,) \prod_{t=2}^{T} \frac{q(\, x_t \mid x_{t-1}, x_0\,) q(x_{t-1} \mid x_0)}{q(\, x_t \mid x_0\,)} \\
&= q(\, x_T \mid x_0\,) \frac{q(x_1 \mid x_0) \prod_{t=2}^{T} q(\, x_t \mid x_{t-1}, x_0\,)}{q(\, x_T \mid x_0\,)} \\
&= q(x_1 \mid x_0) \prod_{t=2}^{T} q(\, x_t \mid x_{t-1}, x_0\,) \\
&= \prod_{t=1}^{T} q(\, x_t \mid x_{t-1}\,)
\end{aligned}
$$

进行 ELBO 分析：

$$
\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \\
&= \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} \\
&= \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] \\
&\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \prod_{t=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] \\
&= \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] \\
&= \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] \\
&= \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] \\
&= \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] \\
&= \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}} \\
&\qquad - \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}}
\end{aligned}
$$

上述优化是 suboptimal 的，引入新的概率：

$$
q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)} \tag{46}
$$

Armed with this new equation, we can retry the derivation resuming from the ELBO in Equation 37:

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}\text{maxmise } & \mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0 \mid x_1)] - KL[q(\, x_T \mid x_0\,) \parallel p(x_T)] \\ & - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[KL[q(x_{t-1} \mid x_t, x_0) \parallel p_\theta(x_{t-1} \mid x_t)]\right]\end{aligned}$$

注意到 $\Gamma_{PMT} = -KL[q(\, x_T \mid x_0\,) \parallel p(x_T)]$ 与参数 θ 无关（Noise 和 Noise 的 KL）。因此：

$$\begin{aligned}\text{maxmise } & \mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0 \mid x_1)] - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\left[KL[q(x_{t-1} \mid x_t, x_0) \parallel p_\theta(\, x_{t-1} \mid x_t\,)]\right] \\ & + \Gamma_{PMT}\end{aligned}$$

$q(x_{t-1} \mid x_t, x_0)$ 此项可以进行改写：

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$q(x_{t-1}\mid x_t,x_0)=\mathcal{N}\left(x_{t-1};\underbrace{\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t x_0+\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t}{1-\bar{\alpha}_t}}_{\mu_q(x_t,x_0)},\underbrace{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}_{\Sigma_q(t)}\mathbf{I}\right)$$

考虑 KL 的解析解：

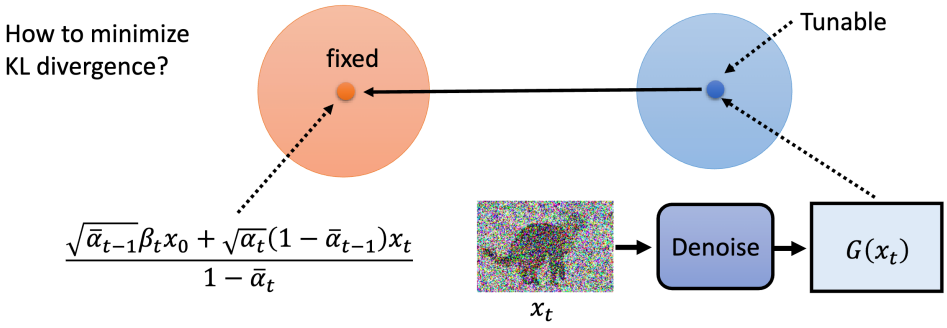
$$KL[\mathcal{N}(x;\mu_x,\Sigma_x)\parallel\mathcal{N}(y;\mu_y,\Sigma_y)]$$

$$=\frac{1}{2}\left[\log\frac{|\Sigma_y|}{|\Sigma_x|}-d+tr[\Sigma_y^{-1}\Sigma_x]+(\mu_y-\mu_x)^\top\Sigma_y^{-1}(\mu_y-\mu_x)\right]$$

$$KL[q(x_{t-1}\mid x_t,x_0)\parallel p_\theta(x_{t-1}\mid x_t)]$$

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

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



已知

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

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

$$
\begin{aligned}
\mu_q &= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t\dfrac{x_t-\sqrt{1-\bar{\alpha}_t}\varepsilon}{\sqrt{\bar{\alpha}_t}} + \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t}{1-\bar{\alpha}_t} \\
&= \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon\right)
\end{aligned}
$$

只有 ε 不知，因此 $\epsilon_\theta(x_t,t) = \varepsilon$

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** $\mathbf{x}_0$

最后这里的 $\sigma_t z$ 可能是类似于 NLP 中的 Sample etc. DDPM 本身是一个 Autoregressive 的模型，即可以一步到位。但是可能一步到位的效果不好，因此去选择分布采样。这就是为什么要加入 $\sigma_t z$。