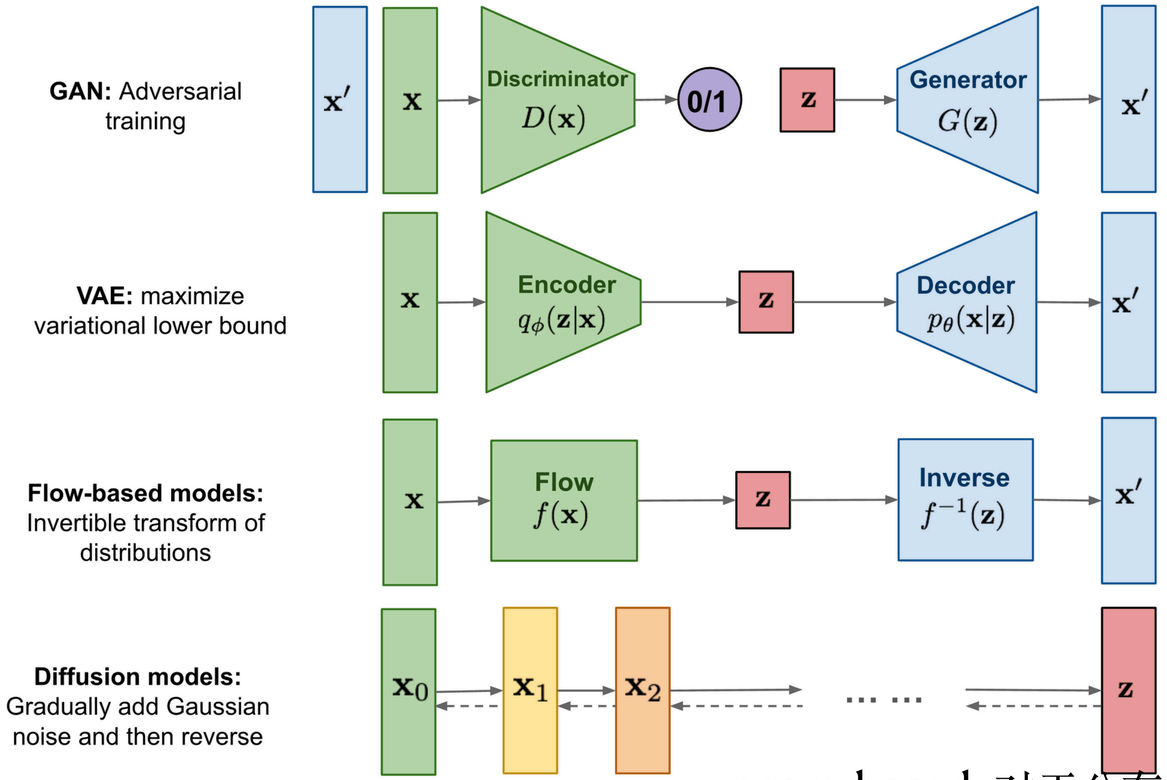
Diffusion Model

Introduction: Generative Models

生成模型的目标：给定某一分布的观测数据 $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$，建模估计其分布 $p_\theta(\mathbf{x})$



建模从简单分布到复杂分布的采样过程

likelihood-based: $\theta^* = \arg\max \prod p_\theta(\mathbf{x}_i)$

score-based: 对于分布 $p_\theta(\mathbf{x}) = \dfrac{e^{-f_\theta(\mathbf{x})}}{Z_\theta}$, $Z_\theta = \int e^{-f_\theta(\mathbf{x})} \mathrm{d}\mathbf{x}$

学习建模其 score function $\mathbf{s}_\theta(\mathbf{x}) = \nabla_\mathbf{x} \log p_\theta(\mathbf{x}) = -\nabla_\mathbf{x} f_\theta(\mathbf{x}) - \underbrace{\nabla_\mathbf{x} \log Z_\theta}_{=0} = -\nabla_\mathbf{x} f_\theta(\mathbf{x})$

Likelihood-base Model

给定先验 $p(\mathbf{z})$，建模 $p_\theta(\mathbf{x} \mid \mathbf{z})$，目标是最大化似然 $\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log p_\theta(\mathbf{x}_i)$

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z}) \mathrm{d}\mathbf{z}$$

积分 intractable

$$p_\theta(\mathbf{x}) = \frac{p_\theta(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z})}{p_\theta(\mathbf{z} \mid \mathbf{x})}$$

在给定参数 θ 的条件下，无法获得 $p_\theta(\mathbf{z} \mid \mathbf{x})$

$q_\phi(\mathbf{z} \mid \mathbf{x})$

Variational Inference: 寻找一个简单的分布 q 来近似推断问题中无法求解的后验概率密度 $p(\mathbf{z} \mid \mathbf{x})$

此时优化目标变成：1）最大化似然；2）最小化 p_θ 和 q_ϕ 的距离

ELBO

$$\log p_\theta(\mathbf{x}) = \log p_\theta(\mathbf{x}) \int q_\phi(\mathbf{z} \mid \mathbf{x}) \mathrm{d}\mathbf{z} \qquad \text{(Multiply by } 1 = \int q_\phi(\mathbf{z} \mid \mathbf{x}) \mathrm{d}\mathbf{z}\text{)}$$

$$= \int q_\phi(\mathbf{z} \mid \mathbf{x}) (\log p_\theta(\mathbf{x})) \mathrm{d}\mathbf{z} \qquad \text{(Bring evidence into integral)}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x})] \qquad \text{(Definition of Expectation)}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{z} \mid \mathbf{x})} \right] \qquad \text{(Chain Rule of Probability)}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z}) q_\phi(\mathbf{z} \mid \mathbf{x})}{p_\theta(\mathbf{z} \mid \mathbf{x}) q_\phi(\mathbf{z} \mid \mathbf{x})} \right] \qquad \text{(Multiply by } 1 = \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\text{)}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} \right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{p_\theta(\mathbf{z} \mid \mathbf{x})} \right] \qquad \text{(Split the Expectation)}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} \right] + D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z} \mid \mathbf{x})) \qquad \text{(Definition of KL Divergence)}$$

$$\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} \right] \qquad \text{(KL Divergence always } \geq 0\text{)}$$

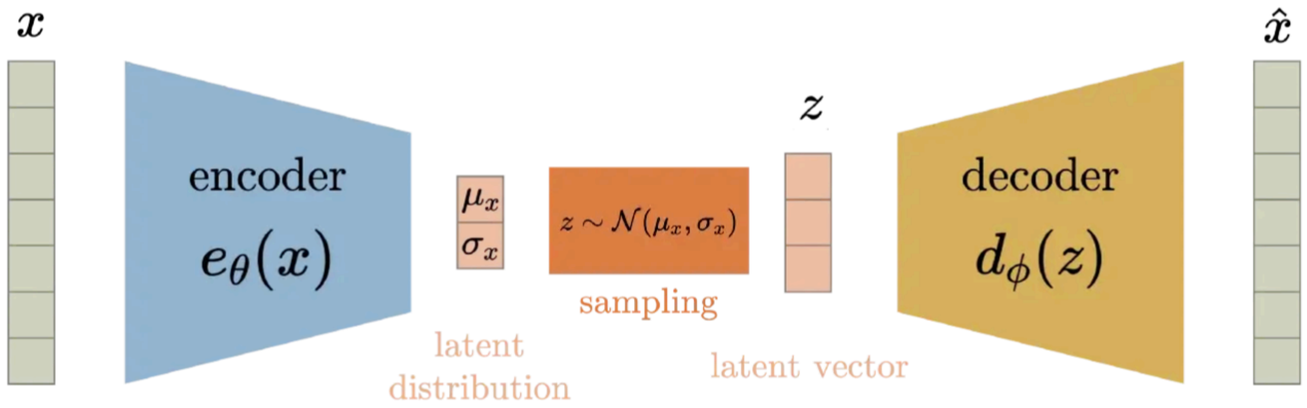
ELBO

$$-Loss = \sum_{i=1}^{N} \log p_\theta(\mathbf{x}_i) = \sum_{i=1}^{N} \left[\underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \left[\log \frac{p_\theta(\mathbf{x}_i, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x}_i)} \right]}_{\text{ELBO}} + D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}_i) \| p_\theta(\mathbf{z} \mid \mathbf{x}_i)) \right]$$

- $p_\theta(\mathbf{x})$ 对于参数 ϕ 来说是一个常量，优化 ϕ 使得 ELBO 越大的同时，会使得 D_{KL} 越小
- 当 θ 保持不变，ϕ 被优化到最佳的时候 $D_{\mathrm{KL}} = 0$，此时最大似然等于 ELBO，此时优化 θ 使得 ELBO 越大等价于使得最大似然越大
- 综上，联合优化 θ 和 ϕ 使得 ELBO 越大等价于最大化似然和最小化 D_{KL}，因此 ELBO 是一个完美的代理优化目标

VAE = VI + 看着像 AE 👀

$$
\begin{aligned}
-Loss = \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \left[\log \frac{p_\theta(\mathbf{x}_i, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x}_i)} \right] &= \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \left[\log \frac{p_\theta(\mathbf{x}_i \mid \mathbf{z}) p(\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x}_i)} \right] && \text{(Chain Rule of Probability)} \\
&= \sum_{i=1}^{N} \left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \left[\log p_\theta(\mathbf{x}_i \mid \mathbf{z}) \right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \left[\log \frac{p(\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x}_i)} \right] \right] && \text{(Split the Expectation)} \\
&= \sum_{i=1}^{N} \left[\underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \left[\log p_\theta(\mathbf{x}_i \mid \mathbf{z}) \right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}} \left(q_\phi(\mathbf{z} \mid \mathbf{x}_i) \| p(\mathbf{z}) \right)}_{\text{prior matching term}} \right] && \text{(Definition of KL Divergence)}
\end{aligned}
$$



理论上来说，对于数据分布 $p_\theta(\mathbf{x})$，我们对 $p(\mathbf{z})$、$p_\theta(\mathbf{x} \mid \mathbf{z})$ 和 $p_\theta(\mathbf{z} \mid \mathbf{x})$ 的分布不做任何要求，只要是满足 $p_\theta(\mathbf{x})p_\theta(\mathbf{z} \mid \mathbf{x}) = p(\mathbf{z})p_\theta(\mathbf{x} \mid \mathbf{z})$ 的一组解即可

VAE 中假设这三个都是高斯分布，后面讲会回顾其中存在的问题

VAE

$$-Loss = \sum_{i=1}^{N} \left[\underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \left[\log p_\theta(\mathbf{x}_i \mid \mathbf{z}) \right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}} \left(q_\phi(\mathbf{z} \mid \mathbf{x}_i) \| p(\mathbf{z}) \right)}_{\text{prior matching term}} \right]$$

$$\begin{aligned}
& D_{\mathrm{KL}} \left(q_\phi(\mathbf{z} \mid \mathbf{x}_i) \| p(\mathbf{z}) \right) \\
&= D_{\mathrm{KL}} \left(\mathcal{N} \left(\mu_\phi(\mathbf{x}_i), \sigma_\phi(\mathbf{x}_i) \right) \| \mathcal{N} \left(\mathbf{0}, \mathbf{I} \right) \right) \\
&= \| \mu_\phi(\mathbf{x}_i) \|_2^2 + \| \sigma_\phi(\mathbf{x}_i) \|_2^2 - 2 \sum_{d=1}^{D} \log \sigma_\phi^{(d)}(\mathbf{x}_i) + \text{constant} \\
&\propto \| \mu_\phi(\mathbf{x}_i) \|_2^2 + \| \sigma_\phi(\mathbf{x}_i) \|_2^2 - 2 \sum_{d=1}^{D} \log \sigma_\phi^{(d)}(\mathbf{x}_i)
\end{aligned}$$

VAE

假设 $p_\theta(\mathbf{x}_i \mid \mathbf{z})$ 是有固定方差 $\sigma_\mathbf{z}$ 的高斯分布

$$-Loss = \sum_{i=1}^{N} \left[\underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \left[\log p_\theta(\mathbf{x}_i \mid \mathbf{z}) \right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}} \left(q_\phi(\mathbf{z} \mid \mathbf{x}_i) \| p(\mathbf{z}) \right)}_{\text{prior matching term}} \right]$$

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \left[\log p_\theta(\mathbf{x}_i \mid \mathbf{z}) \right]$$

$$= \frac{1}{J} \sum_{j=1}^{J} \log p_\theta(\mathbf{x}_i \mid \mathbf{z}_j)$$

$$= \frac{-1}{2J} \sum_{j=1}^{J} \left[\sum_{k=1}^{K} \left[\frac{\left(\mathbf{x}_i^{(k)} - \mu_\theta^{(k)}(\mathbf{z}_j) \right)^2}{\sigma_\mathbf{z}^{(k)}} + \log 2\pi\sigma_\mathbf{z}^{(k)} \right] \right]$$

$$= \frac{-1}{2J} \sum_{j=1}^{J} \|\mathbf{x}_i - \mu_\theta(\mathbf{z}_j)\|_2^2 + \text{constant} \propto \|\mathbf{x}_i - \mu_\theta(\mathbf{z}')\|_2^2$$

上式即是通过从$q_\phi(z \mid x_i)$中采样m次z_j，来逼近$\mathbb{E}_{q_\phi}[\log p_\theta(x_i \mid z)]$。也许我们会好奇，之前两次我们都说积分太难求了，采样逼近代价太大了，所以不能采样逼近，为什么这里又可以采样逼近了呢？

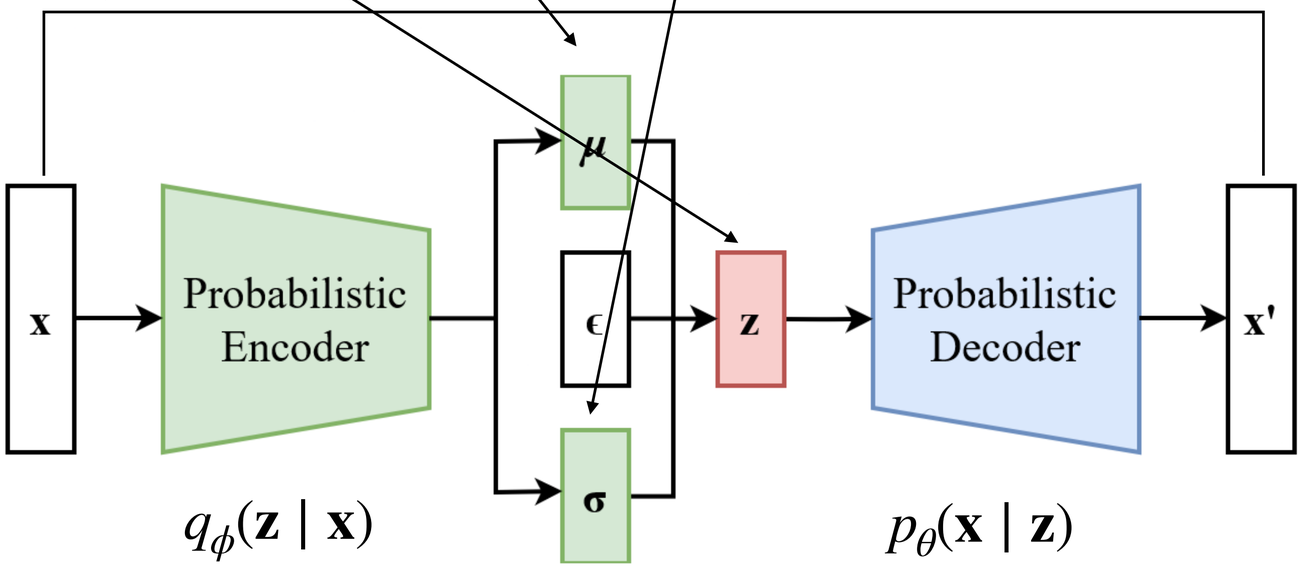
答案就是：之前我们都只能从$p(z)$中采样z_j，这样的话，采样到和x_i有关联的z_j的概率实在是很低，所以为了更好的逼近积分只能采样大量的z_j，这样的代价自然是极大的；然而，在上式中，我们其实是从$q_\phi(z \mid x_i)$中采样得到z_j。随着网络的训练，近似后验$q_\phi(z \mid x_i)$，很快就会比较接近真实的后验分布。这样一来，我们有很大可能能够在有限次数的采样中，采样到与x_i关联的z_j。

事实上，从经验来看，从$q_\phi(z \mid x_i)$中采样z_j估计$\mathbb{E}_{q_\phi}[\log p_\theta(x_i \mid z)]$是比较高效的。在实践中我们往往对一个$x_i$只采样一个$z_j$，即$m = 1$，就能达到可观的效果。所以我们可以将损失改写，并继续往下展开：

VAE

$$-Loss = \sum_{i=1}^{N} \left[\underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \left[\log p_\theta(\mathbf{x}_i \mid \mathbf{z}) \right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}} \left(q_\phi(\mathbf{z} \mid \mathbf{x}_i) \| p(\mathbf{z}) \right)}_{\text{prior matching term}} \right]$$

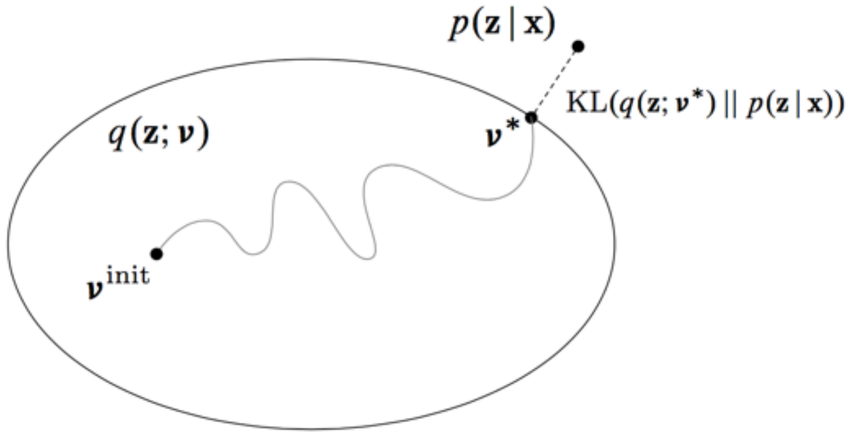
$$= \sum_{i=1}^{N} \left[\| \mathbf{x}_i - \mu_\theta(\mathbf{z}') \|_2^2 - \left(\| \mu_\phi(\mathbf{x}_i) \|_2^2 + \| \sigma_\phi(\mathbf{x}_i) \|_2^2 - 2 \sum_{d=1}^{D} \log \sigma_\phi^{(d)}(\mathbf{x}_i) \right) \right]$$



Hierarchical VAE

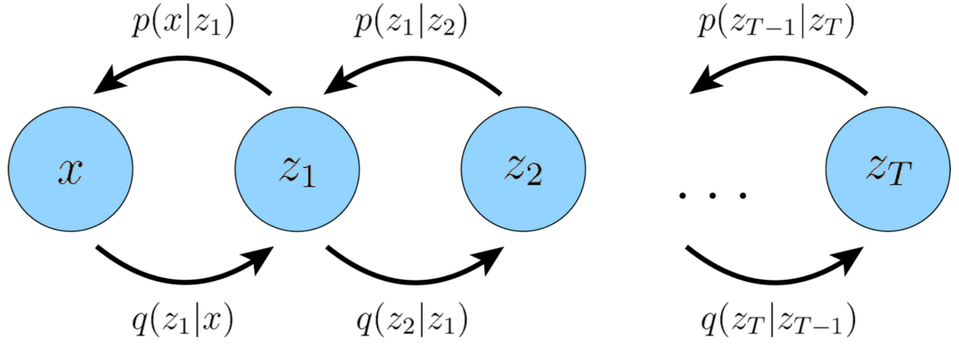
VAE 假设 $p(\mathbf{z})$、$p_\theta(\mathbf{x} \mid \mathbf{z})$ 和 $q_\phi(\mathbf{z} \mid \mathbf{x})$ 都是高斯分布

$$p_\theta(\mathbf{z} \mid \mathbf{x}) = \frac{p_\theta(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})}{p_\theta(\mathbf{x})} = \frac{\text{Gaussian} * \text{Gaussian}}{\text{a complex distribution}}$$



- VI turns **inference into optimization.**

$$\begin{aligned}
p_\theta(\mathbf{x}) &= \int p_\theta(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} && \text{GMM 理论上可以拟合任意复杂分布，像一个单层无限神经元的网络} \\
&= \int p_\theta(\mathbf{x} \mid \mathbf{z}_1)p(\mathbf{z}_1)\mathrm{d}\mathbf{z}_1 && \text{fat} \\
&= \int p_\theta(\mathbf{x} \mid \mathbf{z}_1)\left[p_\theta(\mathbf{z}_1 \mid \mathbf{z}_2)p(\mathbf{z}_2)\mathrm{d}\mathbf{z}_2\right]\mathrm{d}\mathbf{z}_1 && \text{deep + fat} \\
&\cdots \\
&= \int p_\theta(\mathbf{x} \mid \mathbf{z}_1)\prod_{t=2}^{T} p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t)p(\mathbf{z}_T)\mathrm{d}\mathbf{z}_{1:T} && \text{deeper is better}
\end{aligned}$$



即使假设 $q_\phi(\mathbf{z}_{t-1} \mid \mathbf{z}_t)$ 都是高斯分布，但多个高斯叠加在一起的复杂分布可以减小由于假设带来的和真实后验的误差

Hierarchical VAE

联合分布：$p_\theta(\mathbf{x}, \mathbf{z}_{1:T}) = p_\theta(\mathbf{x} \mid \mathbf{z}_1) \prod_{t=2}^{T} p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t) p(\mathbf{z}_T)$

变分后验：$q_\phi(\mathbf{z}_{1:T} \mid \mathbf{x}) = q_\phi(\mathbf{z}_1 \mid \mathbf{x}) \prod_{t=2}^{T} q_\phi(\mathbf{z}_t \mid \mathbf{z}_{t-1})$

$$
\begin{aligned}
-Loss = \text{ELBO} &= \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \left[\log \frac{p_\theta(\mathbf{x}_i, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x}_i)} \right] \\
&= \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_i)} \left[\log \frac{p_\theta(\mathbf{x}_i, \mathbf{z}_{1:T})}{q_\phi(\mathbf{z}_{1:T} \mid \mathbf{x}_i)} \right]
\end{aligned}
$$

Diffusion Model

- 数据 $\mathbf{x}$ 和隐变量 $\mathbf{z}_i$ 统一维度，因此每一个 p_θ 和 q_ϕ 的输入形式一致，可以复用网络
- 对于马尔可夫概率转移 $q_\phi(\mathbf{z}_t \mid \mathbf{z}_{t-1})$，是可以人工设计一条平稳马尔可夫链，使得终态为 $\mathcal{N}(\mathbf{0}, \mathbf{I})$
- 因此目标从联合训练(ϕ, θ)，从解空间找到一组解 → 人工给定一个 ϕ，找到对应的解 θ

仅定义 $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I})$

推导可得 $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right)$, where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$

设计 $\lim_{t\to\infty} \bar{\alpha}_t = 0$，可得 $\lim_{t\to\infty} q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{0}, \mathbf{I})$

Diffusion Model

$$
\begin{aligned}
-Loss = \text{ELBO} &= \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0^{(i)})} \left[\log \frac{p_\theta(\mathbf{x}_0^{(i)}, \mathbf{x}_{1:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0^{(i)})} \right] \\
&= \sum_{i=1}^{N} \left\{ \underbrace{\mathbb{E}_{q\left(\mathbf{x}_1 \mid \mathbf{x}_0^{(i)}\right)} \left[\log p_\theta \left(\mathbf{x}_0^{(i)} \mid \mathbf{x}_1 \right) \right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}} \left(q \left(\mathbf{x}_T \mid \mathbf{x}_0^{(i)} \right) \| p \left(\mathbf{x}_T \right) \right)}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q\left(\mathbf{x}_t \mid \mathbf{x}_0^{(i)}\right)} \left[D_{\mathrm{KL}} \left(q \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0^{(i)} \right) \| p_\theta \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t \right) \right) \right]}_{\text{denoising matching term}} \right\} \\
&= \sum_{i=1}^{N} \left\{ \mathbb{E}_{q\left(\mathbf{x}_1 \mid \mathbf{x}_0^{(i)}\right)} \left[\log p_\theta \left(\mathbf{x}_0^{(i)} \mid \mathbf{x}_1 \right) \right] - \sum_{t=2}^{T} \mathbb{E}_{q\left(\mathbf{x}_t \mid \mathbf{x}_0^{(i)}\right)} \left[D_{\mathrm{KL}} \left(q \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0^{(i)} \right) \| p_\theta \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t \right) \right) \right] \right\} \\
&= - \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}_{q\left(\mathbf{x}_t \mid \mathbf{x}_0^{(i)}\right)} \left[D_{\mathrm{KL}} \left(q \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0^{(i)} \right) \| p_\theta \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t \right) \right) \right] \\
&= - \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \mathbf{x}_t \sim q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)} \left[D_{\mathrm{KL}} \left(q \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0 \right) \| p_\theta \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t \right) \right) \right]
\end{aligned}
$$

Diffusion Model

$$Loss = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0)} \left[D_{\mathrm{KL}} \left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) \right) \right]$$

$$q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) = \frac{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_0\right) q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)}$$

$$= \mathcal{N}\left(\mathbf{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}\left(1-\alpha_t\right)\mathbf{x}_0}{1-\bar{\alpha}_t}}_{\mu_q\left(\mathbf{x}_t, \mathbf{x}_0\right)}, \underbrace{\frac{\left(1-\alpha_t\right)\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)=\sigma_q^2\mathbf{I}} \right)$$

为了让 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ 尽可能接近真值 $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$，因此 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ 建模成高斯分布

即 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_q(t)\right)$，也就是让均值 μ_q 和 μ_θ 越接近越好

Diffusion Model

$$\mu_q(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}\left(1-\alpha_t\right)\mathbf{x}_0}{1-\bar{\alpha}_t}$$

$$\mu_\theta\left(\mathbf{x}_t, t\right) = \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}\left(1-\alpha_t\right)\hat{\mathbf{x}}_\theta\left(\mathbf{x}_t, t\right)}{1-\bar{\alpha}_t}$$

$$Loss = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \mathbf{x}_t \sim q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)}\left[D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right)\right]$$

$$= \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \mathbf{x}_t \sim q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)}\left[\underbrace{\frac{1}{2\sigma_q^2(t)}\frac{\bar{\alpha}_{t-1}\left(1-\alpha_t\right)^2}{\left(1-\bar{\alpha}_t\right)^2}}_{\lambda(t)}\left\| \hat{\mathbf{x}}_\theta\left(\mathbf{x}_t, t\right) - \mathbf{x}_0 \right\|_2^2\right]$$

Diffusion Model

有 $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right)$，可得 $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$，因此有 $\mathbf{x}_0 = \dfrac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon}{\sqrt{\bar{\alpha}_t}}$

$$\mu_q(\mathbf{x}_t, \mathbf{x}_0) = \frac{1}{\sqrt{\alpha_t}}\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon$$

$$\mu_\theta\left(\mathbf{x}_t, t\right) = \frac{1}{\sqrt{\alpha_t}}\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\epsilon}_\theta\left(\mathbf{x}_t, t\right)$$

$$Loss = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})}\left[\frac{1}{2\sigma_q^2(t)}\frac{\left(1-\alpha_t\right)^2}{\left(1-\bar{\alpha}_t\right)\alpha_t}\left\|\epsilon - \hat{\epsilon}_\theta\left(\underbrace{\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon}_{\mathbf{x}_t}, t\right)\right\|_2^2\right]$$

Diffusion Model

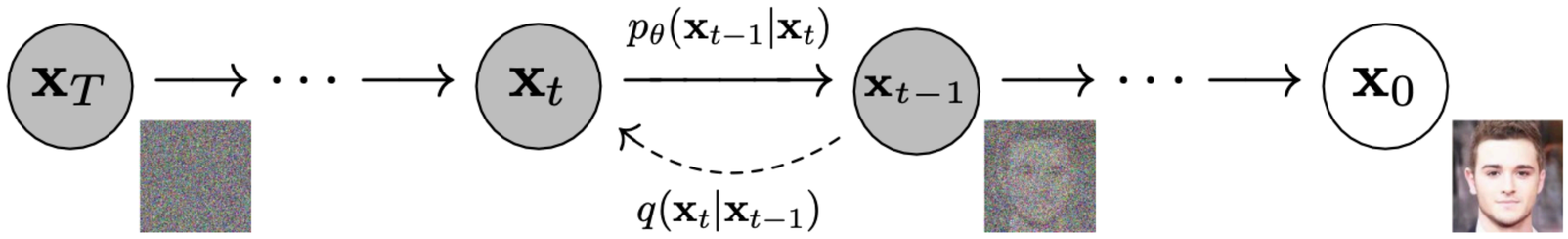
$$Loss = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})} \left[\frac{1}{2\sigma_q^2(t)} \frac{\left(1-\alpha_t\right)^2}{\left(1-\bar{\alpha}_t\right)\alpha_t} \left\| \epsilon - \hat{\epsilon}_\theta \left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t \right) \right\|_2^2 \right]$$

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** $\mathbf{x}_0$



DDIM

训练 采样

$$Loss = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0)} \left[D_{\mathrm{KL}} \left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) \right) \right]$$

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** $\mathbf{x}_0$

虽然我们是从假设状态转移方程 $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}), (1-\alpha_t)\mathbf{I})$ 开始推导

但训练和采样过程从未涉及 $q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$

DDIM

$$Loss = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0)} \left[D_{\mathrm{KL}} \left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) \right) \right]$$

仅定义 $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right)$, where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$

$$q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) = \frac{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_0\right) q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)}$$

不再假设马尔科夫性，也不再假设其分布

此时 $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)$ 仅需满足方程 $\int q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) \mathrm{d}\mathbf{x}_t = q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)$ 即可

DDIM

仅定义 $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right)$, where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$

此时 $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)$ 仅需满足方程 $\int q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) \mathrm{d}\mathbf{x}_t = q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)$ 即可

待定系数 $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; k_t\mathbf{x}_t + \lambda_t\mathbf{x}_0, \sigma_t^2\mathbf{I}\right)$

$$\int q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) \mathrm{d}\mathbf{x}_t$$

$$= k_t\mathbf{x}_t + \lambda_t\mathbf{x}_0 + \sigma_t\epsilon_1$$

$$= k_t(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_2) + \lambda_t\mathbf{x}_0 + \sigma_t\epsilon_1$$

$$= (k_t\sqrt{\bar{\alpha}_t} + \lambda_t)\mathbf{x}_0 + \sqrt{k_t^2(1-\bar{\alpha}_t) + \sigma_t^2}\epsilon = q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right) = \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon$$

DDIM

$$\begin{cases} k_t\sqrt{\bar{\alpha}_t} + \lambda_t = \sqrt{\bar{\alpha}_{t-1}} \\ \sqrt{k_t^2(1-\bar{\alpha}_t) + \sigma_t^2} = \sqrt{1-\bar{\alpha}_{t-1}} \end{cases} \rightarrow \begin{cases} k_t = \dfrac{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}} \\ \lambda_t = \sqrt{\bar{\alpha}_{t-1}} - \dfrac{\sqrt{\bar{\alpha}_t}\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}} \end{cases}$$

$$q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; k_t\mathbf{x}_t + \lambda_t\mathbf{x}_0, \sigma_t^2\mathbf{I}\right)$$

$$= \mathcal{N}\left(\mathbf{x}_{t-1}; \frac{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}}\mathbf{x}_t + (\sqrt{\bar{\alpha}_{t-1}} - \frac{\sqrt{\bar{\alpha}_t}\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}})\mathbf{x}_0, \sigma_t^2\mathbf{I}\right)$$

DDIM

$$q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)=\frac{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_0\right) q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)}$$

$$=\mathcal{N}\left(\mathbf{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right) \mathbf{x}_t+\sqrt{\bar{\alpha}_{t-1}}\left(1-\alpha_t\right) \mathbf{x}_0}{1-\bar{\alpha}_t}}_{\mu_q\left(\mathbf{x}_t, \mathbf{x}_0\right)}, \underbrace{\frac{\left(1-\alpha_t\right)\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} \mathbf{I}}_{\mathbf{\Sigma}_q(t)=\sigma_q^2 \mathbf{I}}\right)$$

$$q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)=\mathcal{N}\left(\mathbf{x}_{t-1} ; k_t \mathbf{x}_t+\lambda_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}\right)$$

$$=\mathcal{N}\left(\mathbf{x}_{t-1} ; \frac{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}} \mathbf{x}_t+(\sqrt{\bar{\alpha}_{t-1}}-\frac{\sqrt{\bar{\alpha}_t} \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}}) \mathbf{x}_0, \sigma_t^2 \mathbf{I}\right)$$

DDPM 就是 $\sigma_t^2=\frac{\left(1-\alpha_t\right)\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}$ 的一个特例，还恰好满足了马尔科夫性

DDIM

$$Loss = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})} \left[\lambda_t(\alpha_{1:t}) \left\| \epsilon - \hat{\epsilon}_\theta \left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t \right) \right\|_2^2 \right]$$

$$Loss = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})} \left[\lambda_t(\alpha_{1:t}, \sigma_{1:t}) \left\| \epsilon - \hat{\epsilon}_\theta \left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t \right) \right\|_2^2 \right]$$

$$Loss_{simple} = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})} \left[\left\| \epsilon - \hat{\epsilon}_\theta \left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t \right) \right\|_2^2 \right]$$

DDPM 使用 $Loss_{simple}$ 训练，效果更好

忽略 λ_t，此时和 DDPM 训练目标一致，只有采样过程和 DDPM 不一样，其依赖于 σ_t

DDIM

Algorithm 2 Sampling

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

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \frac{1}{\sqrt{\alpha}_t} \left(\mathbf{x}_t - \left(\sqrt{1-\bar{\alpha}_t} - \sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} \right) \epsilon_\theta(\mathbf{x}_t, t) \right), \sigma_t^2 \mathbf{I} \right)$$

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha}_t} \left(\mathbf{x}_t - \left(\sqrt{1-\bar{\alpha}_t} - \sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} \right) \epsilon_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$$

$$= \frac{1}{\sqrt{\bar{\alpha}_{t-1}}} \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\, \epsilon_\theta(\mathbf{x}_t, t) + \sigma_t \mathbf{z}$$

DDIM 即 $\sigma_t = 0$ 的情况，此时采样的过程就是一个确定性的过程

DDIM

$$Loss_{simple} = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})} \left[\left\| \epsilon - \hat{\epsilon}_\theta \left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t \right) \right\|_2^2 \right]$$

假设 $[\tau_1, \tau_2, \ldots, \tau_s]$ 是 $[1,2,\ldots,T]$ 的子序列

$Loss_{simple} = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, \tau \sim [\tau_1,\tau_2,\ldots,\tau_s],\, \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})}$ 是 $Loss_{simple} = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})}$ 的子集

在模型拟合能力足够好的情况下，它其实包含了任意子序列参数的训练结果

生成过程也可以从 $[1,2,\ldots,T] \rightarrow [\tau_1, \tau_2, \ldots, \tau_s]$

Table 1: CIFAR10 and CelebA image generation measured in FID. $\eta = 1.0$ and $\hat{\sigma}$ are cases of DDPM (although Ho et al. (2020) only considered $T = 1000$ steps, and $S < T$ can be seen as simulating DDPMs trained with S steps), and $\eta = 0.0$ indicates DDIM.

	S	CIFAR10 (32 × 32) 10	20	50	100	1000	CelebA (64 × 64) 10	20	50	100	1000
η	0.0	**13.36**	**6.84**	**4.67**	**4.16**	4.04	**17.33**	**13.73**	**9.17**	**6.53**	3.51
	0.2	14.04	7.11	4.77	4.25	4.09	17.66	14.11	9.51	6.79	3.64
	0.5	16.66	8.35	5.25	4.46	4.29	19.86	16.06	11.01	8.09	4.28
	1.0	41.07	18.36	8.01	5.78	4.73	33.12	26.03	18.48	13.93	5.98
$\hat{\sigma}$		367.43	133.37	32.72	9.99	**3.17**	299.71	183.83	71.71	45.20	**3.26**

$$\sigma_t^2 = \eta \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}$$

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; k_t\mathbf{x}_t + \lambda_t\mathbf{x}_0, \sigma_t^2\mathbf{I})$$

$$= \mathcal{N}\left(\mathbf{x}_{t-1}; \frac{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}}\mathbf{x}_t + (\sqrt{\bar{\alpha}_{t-1}} - \frac{\sqrt{\bar{\alpha}_t}\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}})\mathbf{x}_0, \sigma_t^2\mathbf{I}\right)$$

σ_t 越小，红框部分越小，对于估计部分的 $\hat{\mathbf{x}}_0$ 依赖就越少

Condition

$$
\begin{aligned}
-Loss = \text{ELBO} &= \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0^{(i)})} \left[\log \frac{p_\theta(\mathbf{x}_0^{(i)}, \mathbf{x}_{1:T} \mid \mathbf{y})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0^{(i)})} \right] \\
&= \sum_{i=1}^{N} \left\{ \underbrace{\mathbb{E}_{q(\mathbf{x}_1 \mid \mathbf{x}_0^{(i)})} \left[\log p_\theta \left(\mathbf{x}_0^{(i)} \mid \mathbf{x}_1, \mathbf{y} \right) \right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}} \left(q \left(\mathbf{x}_T \mid \mathbf{x}_0^{(i)} \right) \| p \left(\mathbf{x}_T \mid \mathbf{y} \right) \right)}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\mathbf{x}_t \mid \mathbf{x}_0^{(i)})} \left[D_{\mathrm{KL}} \left(q \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0^{(i)} \right) \| p_\theta \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y} \right) \right) \right]}_{\text{denoising matching term}} \right\} \\
&= \sum_{i=1}^{N} \left\{ \mathbb{E}_{q(\mathbf{x}_1 \mid \mathbf{x}_0^{(i)})} \left[\log p_\theta \left(\mathbf{x}_0^{(i)} \mid \mathbf{x}_1, \mathbf{y} \right) \right] - \sum_{t=2}^{T} \mathbb{E}_{q(\mathbf{x}_t \mid \mathbf{x}_0^{(i)})} \left[D_{\mathrm{KL}} \left(q \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0^{(i)} \right) \| p_\theta \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y} \right) \right) \right] \right\} \\
&= - \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{x}_t \mid \mathbf{x}_0^{(i)})} \left[D_{\mathrm{KL}} \left(q \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0^{(i)} \right) \| p_\theta \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y} \right) \right) \right] \\
&= - \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0),\, t \sim \mathcal{U}(1,T),\, \mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0)} \left[D_{\mathrm{KL}} \left(q \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0 \right) \| p_\theta \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y} \right) \right) \right]
\end{aligned}
$$

Classifier Guidance

Algorithm 1 Classifier guided diffusion sampling, given a diffusion model $(\mu_\theta(x_t), \Sigma_\theta(x_t))$, classifier $p_\phi(y|x_t)$, and gradient scale s.

Input: class label y, gradient scale s
$x_T \leftarrow$ sample from $\mathcal{N}(0, \mathbf{I})$
for all t from T to 1 **do**
  $\mu, \Sigma \leftarrow \mu_\theta(x_t), \Sigma_\theta(x_t)$
  $x_{t-1} \leftarrow$ sample from $\mathcal{N}(\mu + s\Sigma \nabla_{x_t} \log p_\phi(y|x_t), \Sigma)$
end for
return x_0

$$p(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \rightarrow p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y})$$

$$p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{y}) = \frac{p(\mathbf{x_{t-1}} \mid \mathbf{x}_t) p(\mathbf{y} \mid \mathbf{x}_{t-1})}{p(\mathbf{y} \mid \mathbf{x}_t)}$$

$$= \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t), \sigma_t^2 \mathbf{I}) \exp\left(\log p(\mathbf{y} \mid \mathbf{x}_{t-1}) - \log p(\mathbf{y} \mid \mathbf{x}_t)\right)$$

$$\approx \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t), \sigma_t^2 \mathbf{I}) \exp\left(\log p(\mathbf{y} \mid \mathbf{x}_{t-1}) - \log p(\mathbf{y} \mid \mu_\theta(\mathbf{x}_t))\right)$$

$$\approx \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t), \sigma_t^2 \mathbf{I}) \exp\left(\left(\mathbf{x}_{t-1} - \mu_\theta(\mathbf{x}_t)\right) \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x}) \mid_{\mathbf{x}=\mu_\theta(\mathbf{x}_t)}\right)$$

$$\propto \exp\left(-\frac{\|\mathbf{x}_{t-1} - \mu_\theta(\mathbf{x}_t)\|^2}{2\sigma_t^2} + \left(\mathbf{x}_{t-1} - \mu_\theta(\mathbf{x}_t)\right) \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x}) \mid_{\mathbf{x}=\mu_\theta(\mathbf{x}_t)}\right)$$

$$\propto \exp\left(-\frac{\|\mathbf{x}_{t-1} - \mu_\theta(\mathbf{x}_t) - \sigma_t^2 \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x}) \mid_{\mathbf{x}=\mu_\theta(\mathbf{x}_t)} \|^2}{2\sigma_t^2}\right)$$

$$\approx \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t) + \sigma_t^2 \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x}) \mid_{\mathbf{x}=\mu_\theta(\mathbf{x}_t)}, \sigma_t^2 \mathbf{I})$$

Classifier Guidance

Tweedie's Formula: Suppose $\mathbf{x} \mid \mu \sim \mathcal{N}\left(\mathbf{x}; \mu, \sigma^2 \mathbf{I}\right)$, where $\mu \sim p(\mu)$.

To estimate the expectation $\mathbb{E}[\mu \mid \mathbf{x}]$, Tweedie says $p(\mathbf{x})$ is all you need!

$$\mathbb{E}[\mu \mid \mathbf{x}] = \mathbf{x} + \sigma^2 \nabla_{\mathbf{x}} \log p(\mathbf{x})$$

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

Given sample $\mathbf{x}_t$, we can find the posterior expectation of the mean as:

$$\sqrt{\bar{\alpha}_t}\mathbf{x}_0 = \mathbb{E}\left[\mu \mid \mathbf{x}_t\right] = \mathbf{x}_t + (1 - \bar{\alpha}_t)\nabla \log p\left(\mathbf{x}_t\right)$$

$$\begin{cases} \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 - (1 - \bar{\alpha}_t)\nabla \log p\left(\mathbf{x}_t\right) \\ \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon \end{cases} \rightarrow \nabla \log p\left(\mathbf{x}_t\right) = -\frac{1}{\sqrt{1 - \bar{\alpha}_t}}\epsilon$$

Classifier Guidance

$$\nabla \log p(\mathbf{x}_t) \rightarrow \nabla \log p(\mathbf{x}_t \mid \mathbf{y})$$

$$\begin{aligned}
\nabla \log p\left(\mathbf{x}_t \mid \mathbf{y}\right) &= \nabla \log \left(\frac{p\left(\mathbf{x}_t\right) p\left(\mathbf{y} \mid \mathbf{x}_t\right)}{p(\mathbf{y})} \right) \\
&= \nabla \log p\left(\mathbf{x}_t\right) + \nabla \log p\left(\mathbf{y} \mid \mathbf{x}_t\right) - \nabla \log p(\mathbf{y}) \\
&= \nabla \log p\left(\mathbf{x}_t\right) + \nabla \log p\left(\mathbf{y} \mid \mathbf{x}_t\right) \\
&= -\frac{1}{\sqrt{1-\bar{\alpha}_t}} \epsilon + \nabla \log p\left(\mathbf{y} \mid \mathbf{x}_t\right)
\end{aligned}$$

$$\hat{\epsilon}\left(\mathbf{x}_t\right) = \epsilon_\theta\left(\mathbf{x}_t\right) - \sqrt{1-\bar{\alpha}_t} \nabla \log p\left(\mathbf{y} \mid \mathbf{x}_t\right)$$

Algorithm 2 Classifier guided DDIM sampling, given a diffusion model $\epsilon_\theta(x_t)$, classifier $p_\phi(y|x_t)$, and gradient scale s.

Input: class label y, gradient scale s
$x_T \leftarrow$ sample from $\mathcal{N}(0, \mathbf{I})$
for all t from T to 1 **do**
  $\hat{\epsilon} \leftarrow \epsilon_\theta(x_t) - \sqrt{1-\bar{\alpha}_t}\, \nabla_{x_t} \log p_\phi(y|x_t)$
  $x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon}$
end for
return x_0

Classifier-Free Guidance

$$
\begin{aligned}
\nabla \log p\left(\mathbf{x}_t \mid \mathbf{y}\right) &= \nabla \log p\left(\mathbf{x}_t\right) + \gamma\left(\nabla \log p\left(\mathbf{x}_t \mid \mathbf{y}\right) - \nabla \log p\left(\mathbf{x}_t\right)\right) \\
&= \nabla \log p\left(\mathbf{x}_t\right) + \gamma \nabla \log p\left(\mathbf{x}_t \mid \mathbf{y}\right) - \gamma \nabla \log p\left(\mathbf{x}_t\right) \\
&= \gamma \nabla \log p\left(\mathbf{x}_t \mid \mathbf{y}\right) + (1-\gamma) \nabla \log p\left(\mathbf{x}_t\right) \\
&= \gamma \nabla \log p\left(\mathbf{x}_t \mid \mathbf{y}\right) + (1-\gamma) \nabla \log p\left(\mathbf{x}_t \mid \Phi\right)
\end{aligned}
$$

Algorithm 1 Joint training a diffusion model with classifier-free guidance

Require: p_{uncond}: probability of unconditional training

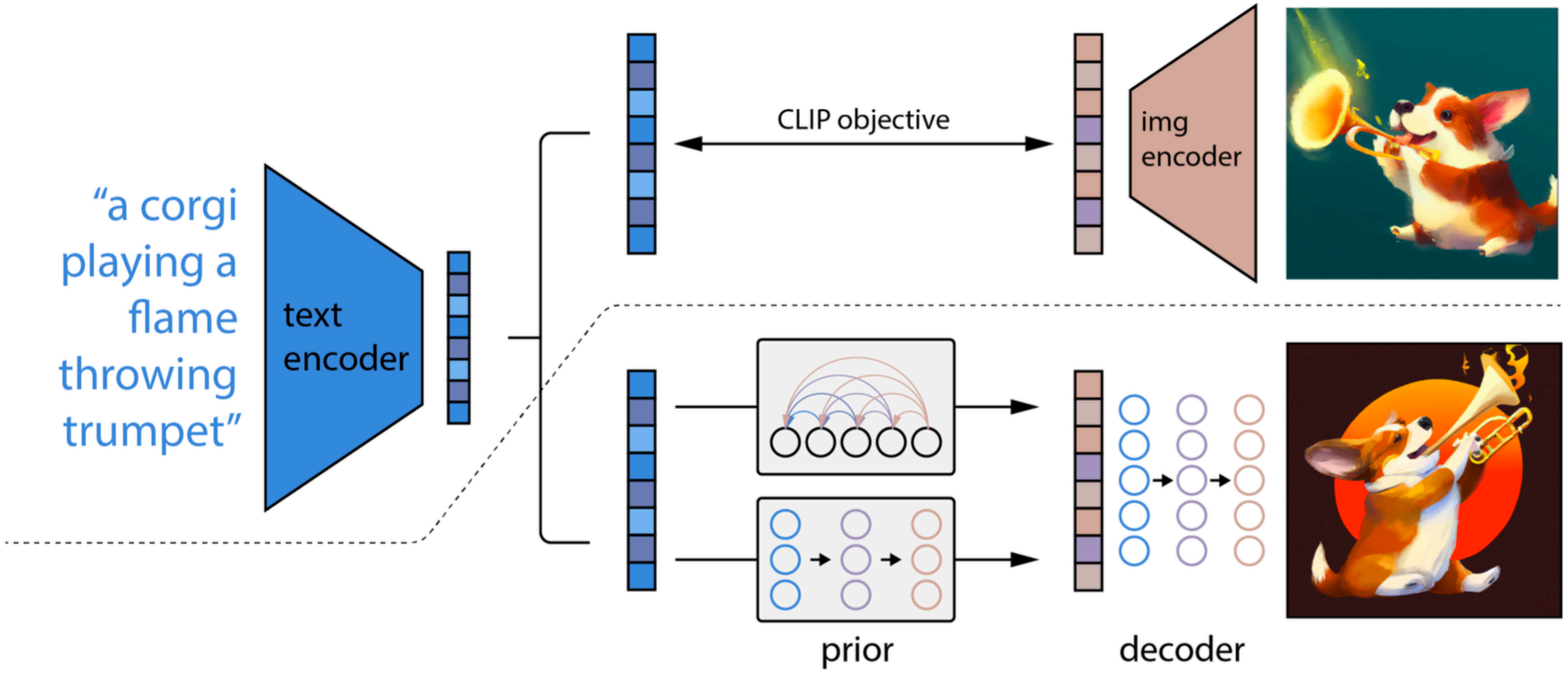
1: **repeat**
2: $(\mathbf{x}, \mathbf{c}) \sim p(\mathbf{x}, \mathbf{c})$ ▷ Sample data with conditioning from the dataset
3: $\mathbf{c} \leftarrow \varnothing$ with probability p_{uncond} ▷ Randomly discard conditioning to train unconditionally
4: $\lambda \sim p(\lambda)$ ▷ Sample log SNR value
5: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
6: $\mathbf{z}_\lambda = \alpha_\lambda \mathbf{x} + \sigma_\lambda \boldsymbol{\epsilon}$ ▷ Corrupt data to the sampled log SNR value
7: Take gradient step on $\nabla_\theta \left\|\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) - \boldsymbol{\epsilon}\right\|^2$ ▷ Optimization of denoising model
8: **until** converged

Algorithm 2 Conditional sampling with classifier-free guidance

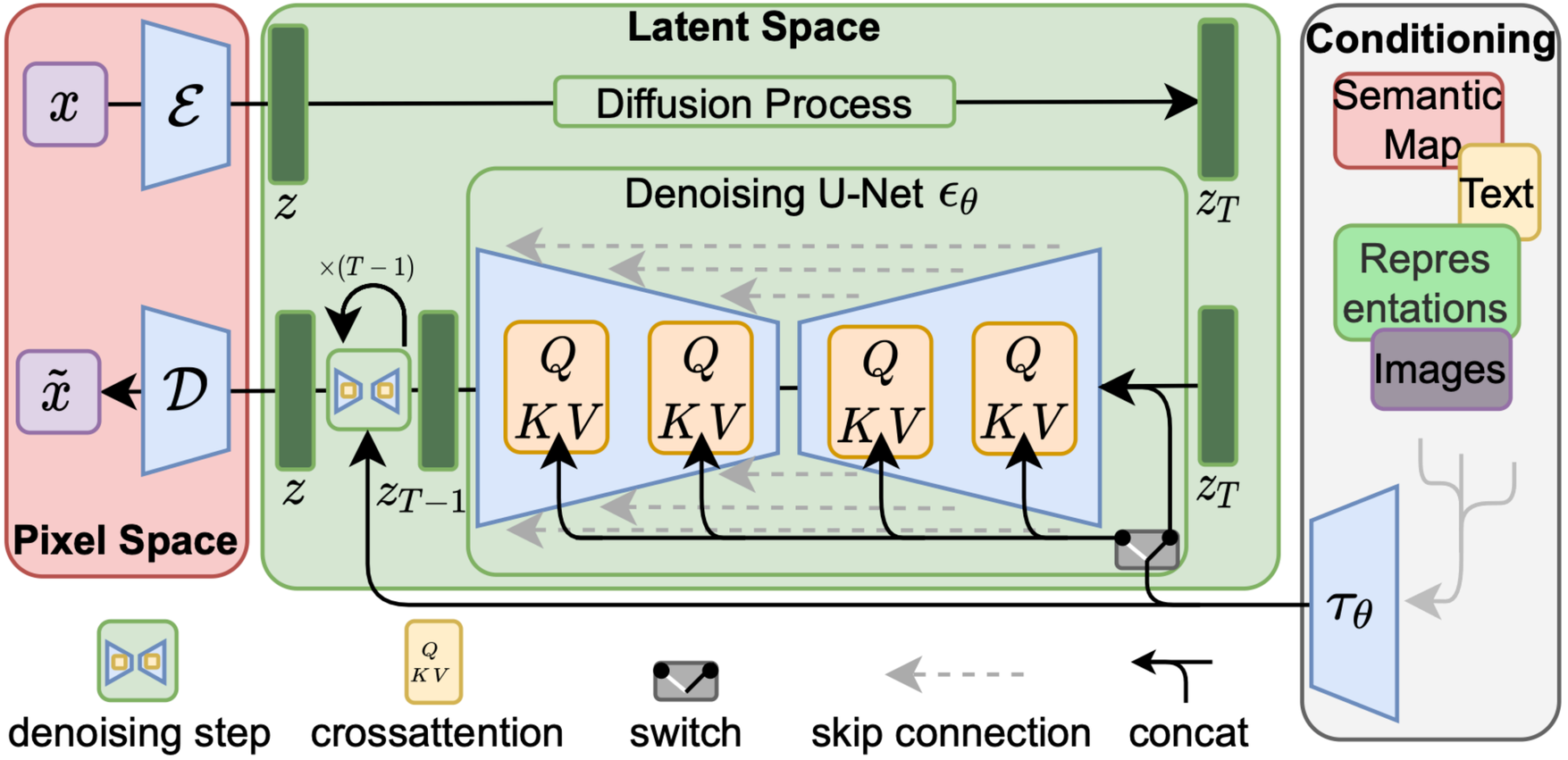
Require: w: guidance strength
Require: $\mathbf{c}$: conditioning information for conditional sampling
Require: $\lambda_1, \ldots, \lambda_T$: increasing log SNR sequence with $\lambda_1 = \lambda_{\min}$, $\lambda_T = \lambda_{\max}$

1: $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = 1, \ldots, T$ **do**
▷ Form the classifier-free guided score at log SNR λ_t
3: $\tilde{\boldsymbol{\epsilon}}_t = (1+w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{z}_t)$
▷ Sampling step (could be replaced by another sampler, e.g. DDIM)
4: $\tilde{\mathbf{x}}_t = (\mathbf{z}_t - \sigma_{\lambda_t}\tilde{\boldsymbol{\epsilon}}_t)/\alpha_{\lambda_t}$
5: $\mathbf{z}_{t+1} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}_{\lambda_{t+1}|\lambda_t}(\mathbf{z}_t, \tilde{\mathbf{x}}_t), (\tilde{\sigma}^2_{\lambda_{t+1}|\lambda_t})^{1-v}(\sigma^2_{\lambda_t|\lambda_{t+1}})^v)$ if $t < T$ else $\mathbf{z}_{t+1} =$
6: **end for**
7: **return** $\mathbf{z}_{T+1}$

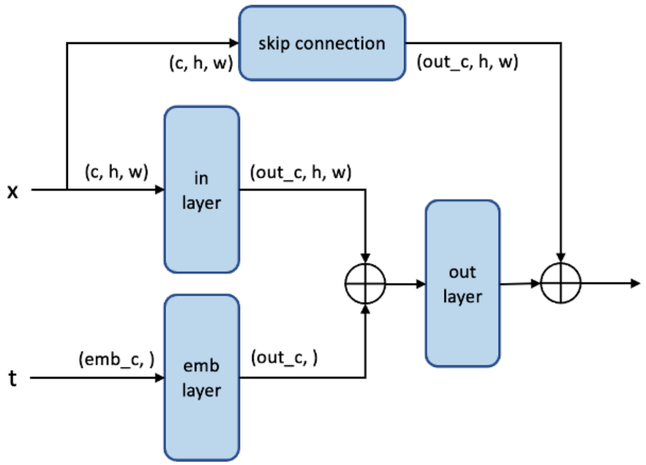
DALL·E 2

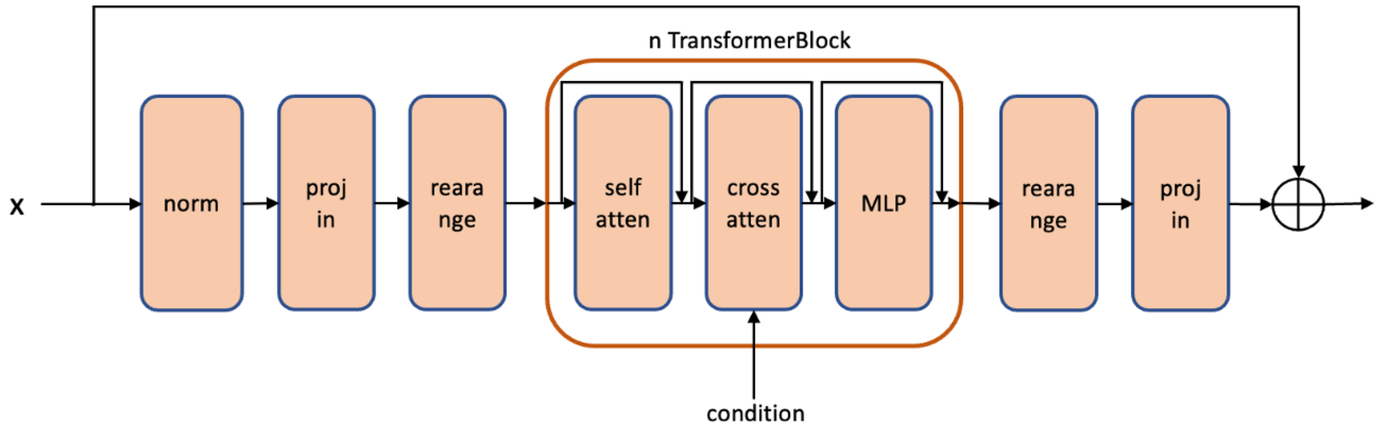

"a corgi playing a flame throwing trumpet"
text encoder
CLIP objective
img encoder
prior
decoder

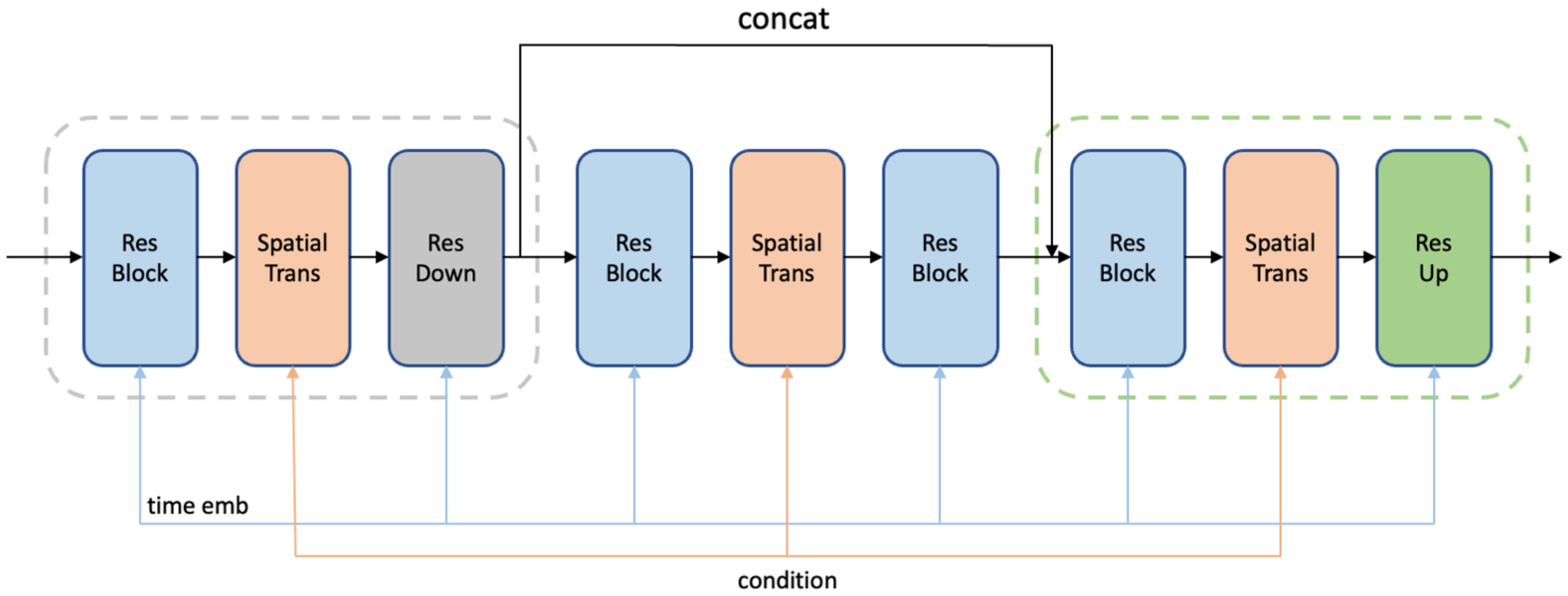
Stable Diffusion



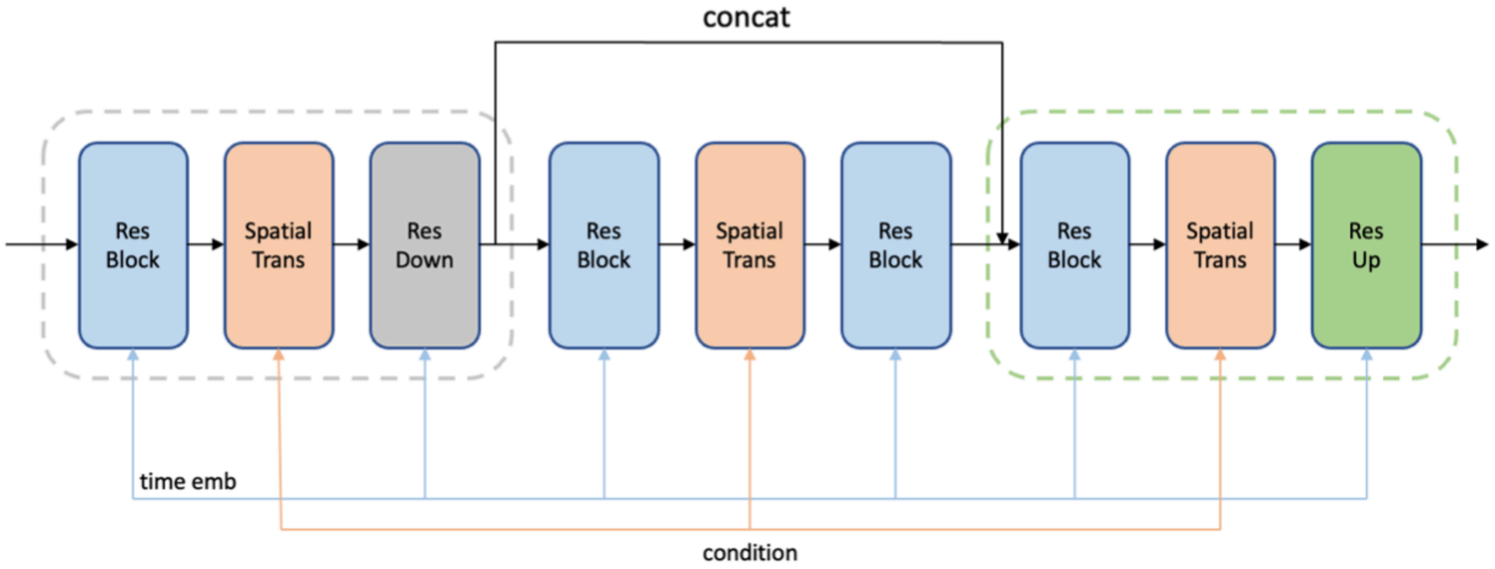
Unet architecture







U-ViT architecture



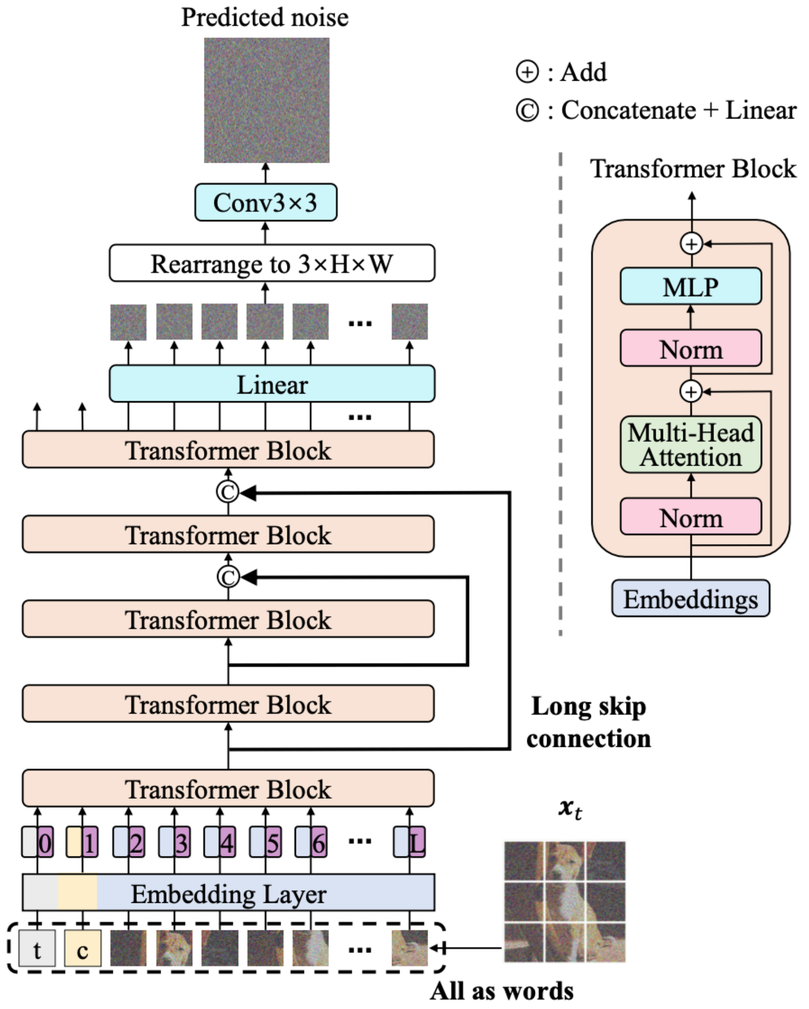


Figure 1. The **U-ViT** architecture for diffusion models, which is characterized by treating **all** inputs including the time, condition and noisy image patches **as tokens** and employing **long skip connections** between shallow and deep layers.

DiT

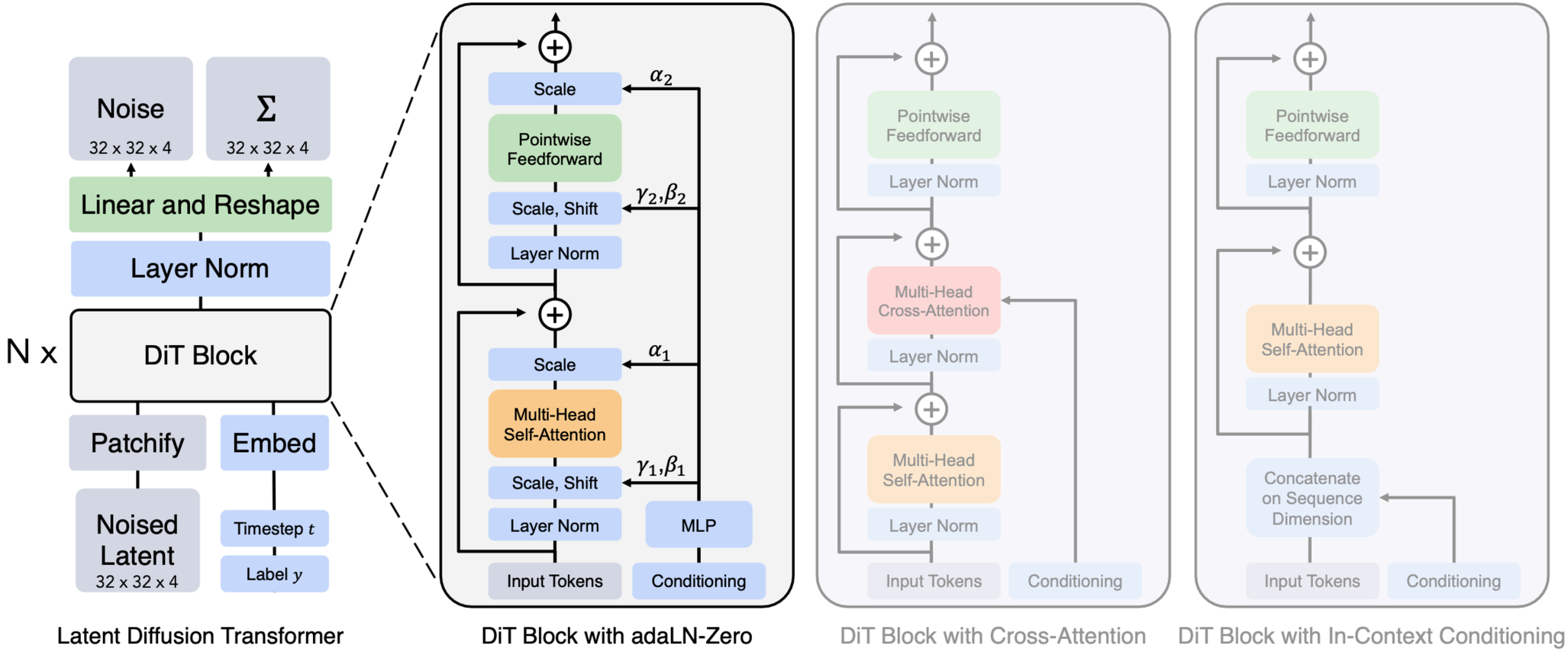
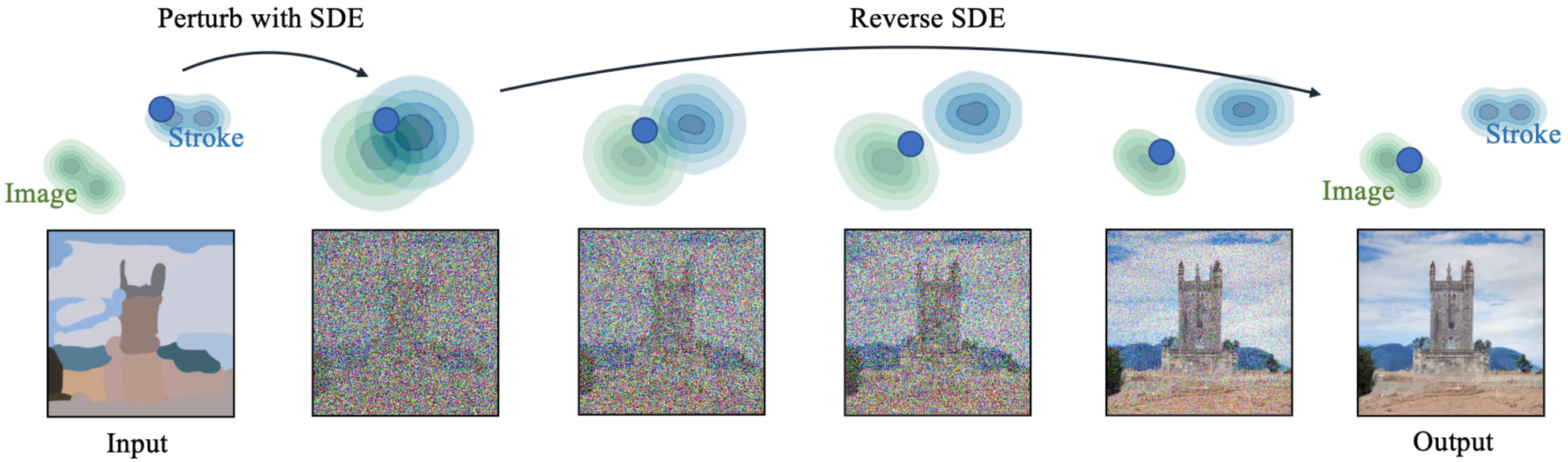
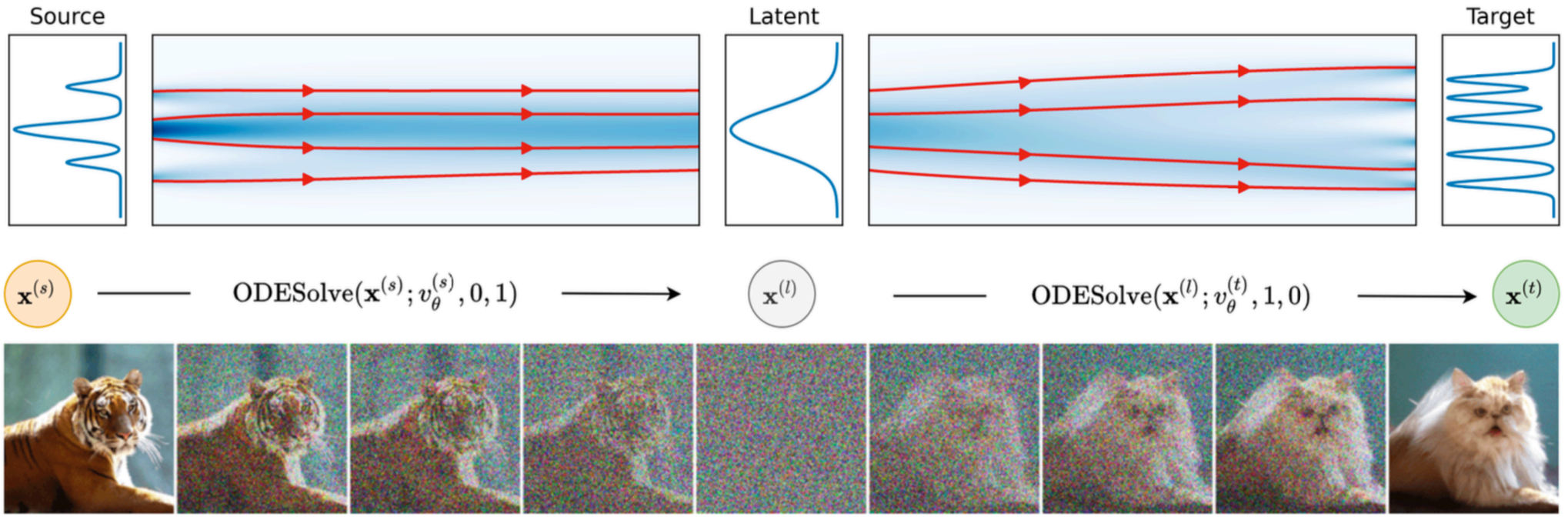


Figure 3. **The Diffusion Transformer (DiT) architecture.** *Left:* We train conditional latent DiT models. The input latent is decomposed into patches and processed by several DiT blocks. *Right:* Details of our DiT blocks. We experiment with variants of standard transformer blocks that incorporate conditioning via adaptive layer norm, cross-attention and extra input tokens. Adaptive layer norm works best.

SDEdit


Perturb with SDE
Reverse SDE
Stroke
Image
Image
Stroke
Input
Output

Zero shot Image-to-Image Translation



DiffEdit

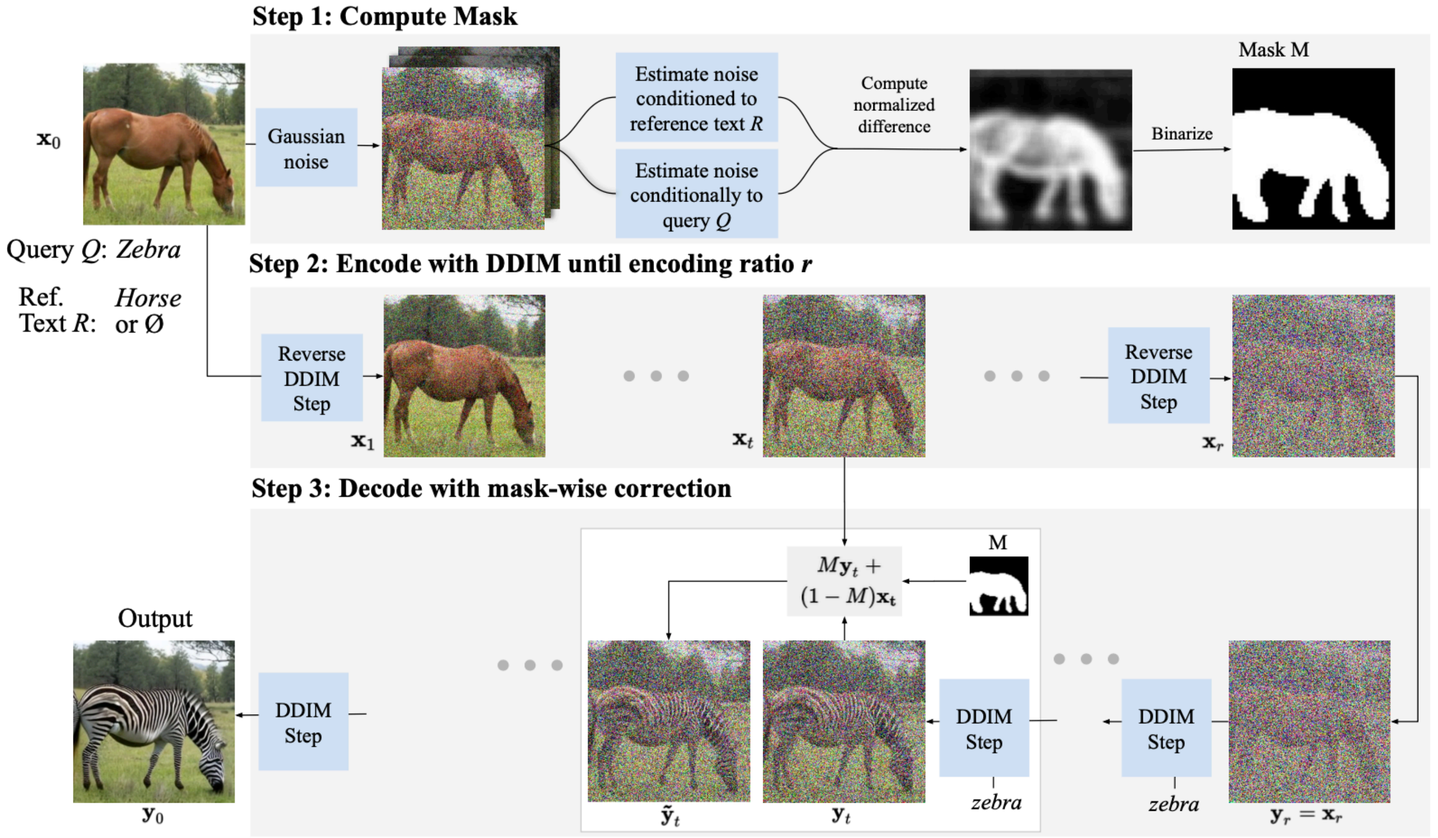


Figure 2: The three steps of DIFFEDIT. **Step 1:** we add noise to the input image, and denoise it: once conditioned on the query text, and once conditioned on a reference text (or unconditionally). We derive a mask based on the difference in the denoising results. **Step 2:** we encode the input image with DDIM, to estimate the latents corresponding to the input image. **Step 3:** we perform DDIM decoding conditioned on the text query, using the inferred mask to replace the background with pixel values coming from the encoding process at the corresponding timestep.

DiffusionDet