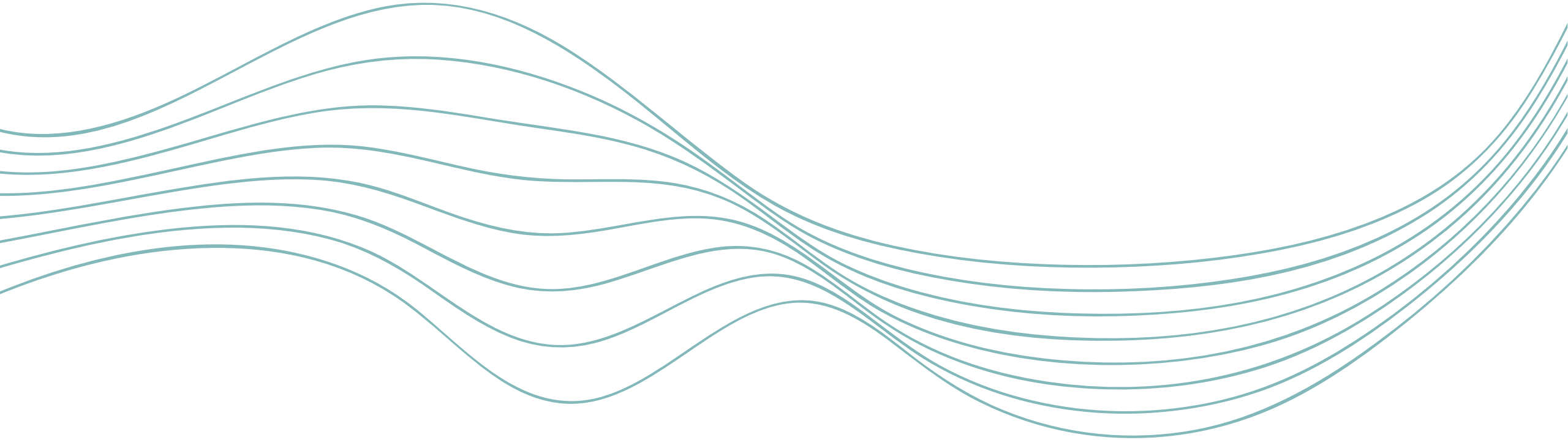


从DDPM、IDDPM到SD；
从VAE到DALL·E、DALL·E2

生成模型

定义

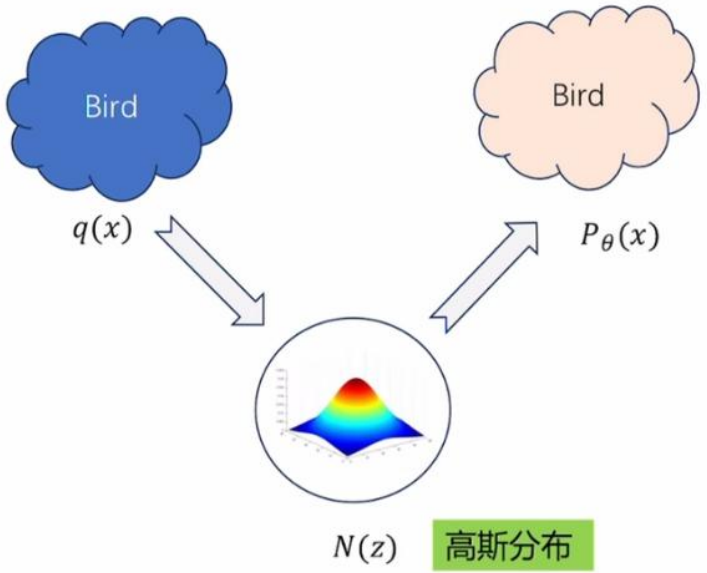
- 一个能随机生成与训练数据一致的模型

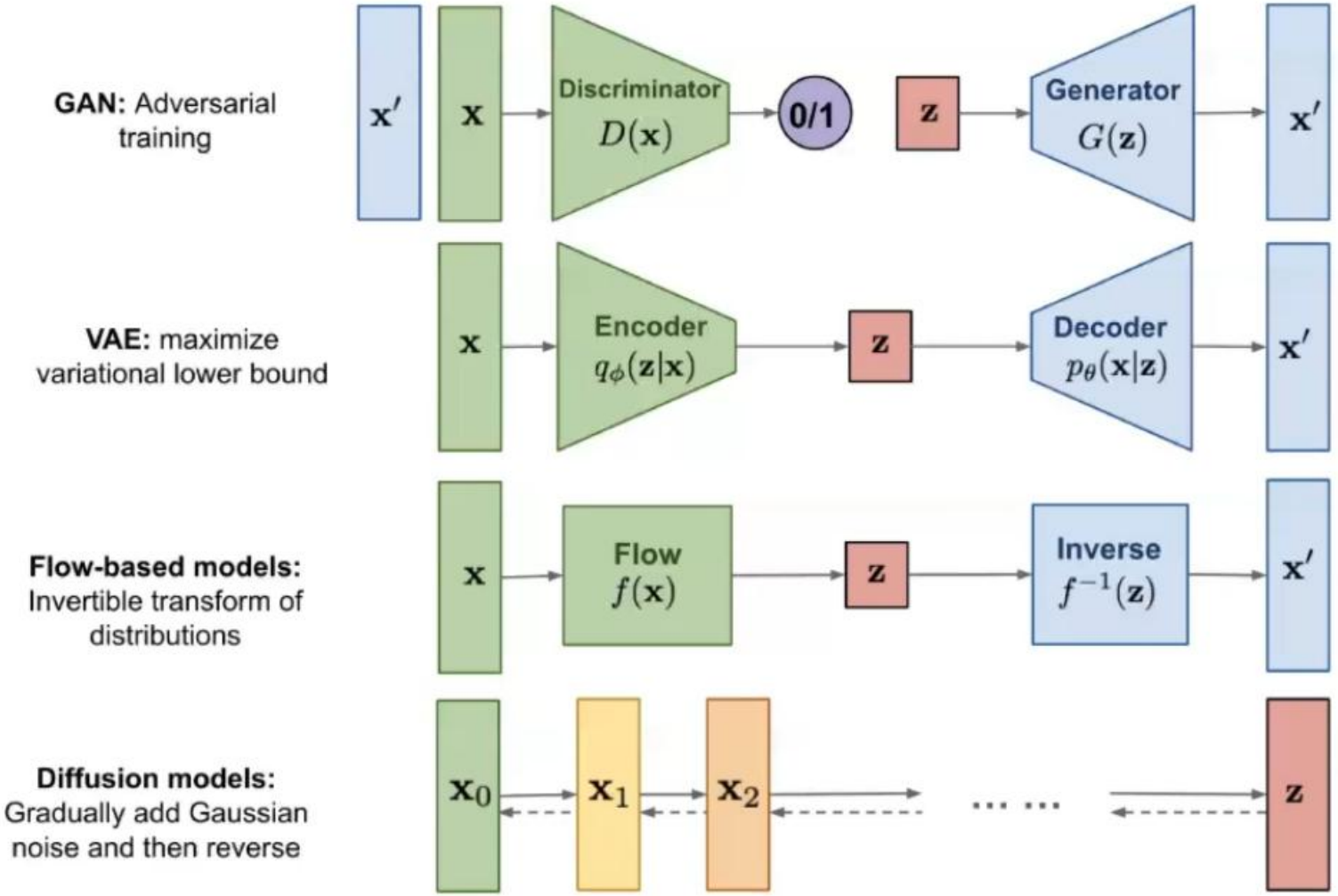
问题

- 如何对训练数据建模?
- 如何采样?

思路

- 从一个简单分布采样是容易的
- 从简单分布到观测数据分布是可以拟合的





生成模型的解题思路

- 将观测数据分布映射到简单分布 【Encoder】
- 从简单分布中映射到观测数据分布【Decoder】

DDPM

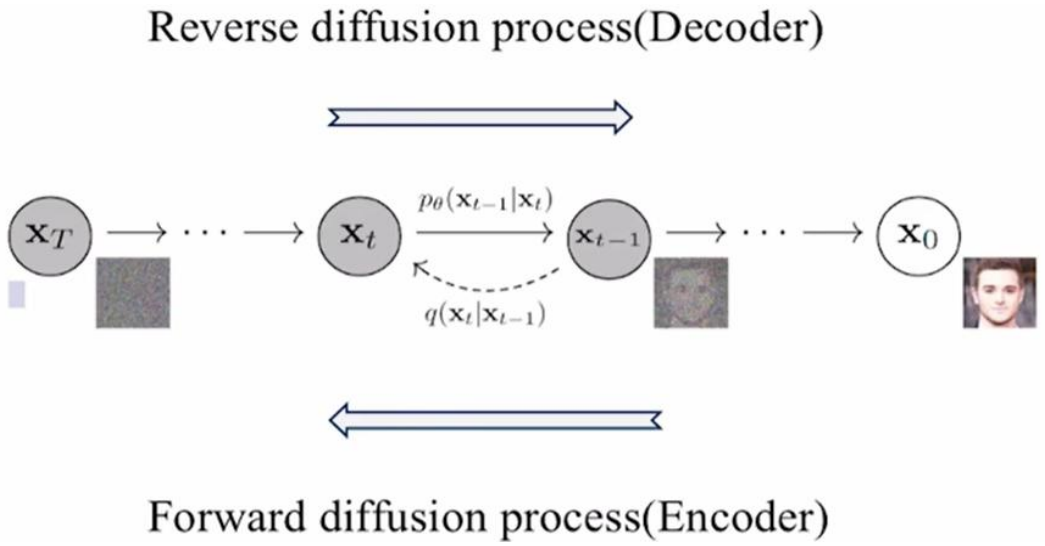
Diffusion Model作为生成模型的一类，同样包含encoder和decoder两个阶段

前向扩散过程

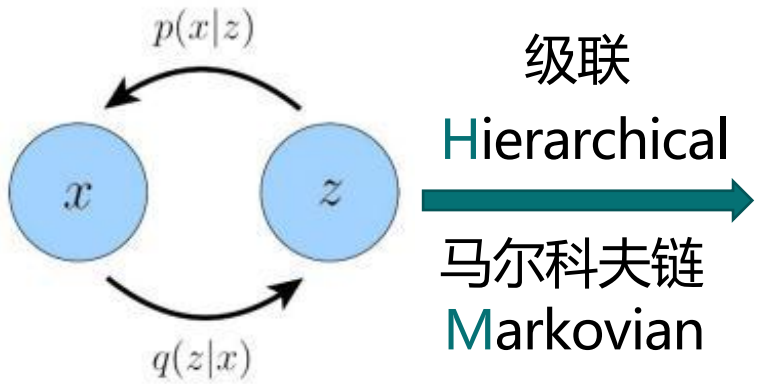
向观测数据中逐步加入噪声，直到观测数据变成高斯分布

反向生成过程

从一个高斯分布中采样，逐步消除噪声，知道变成清晰数据



VAE

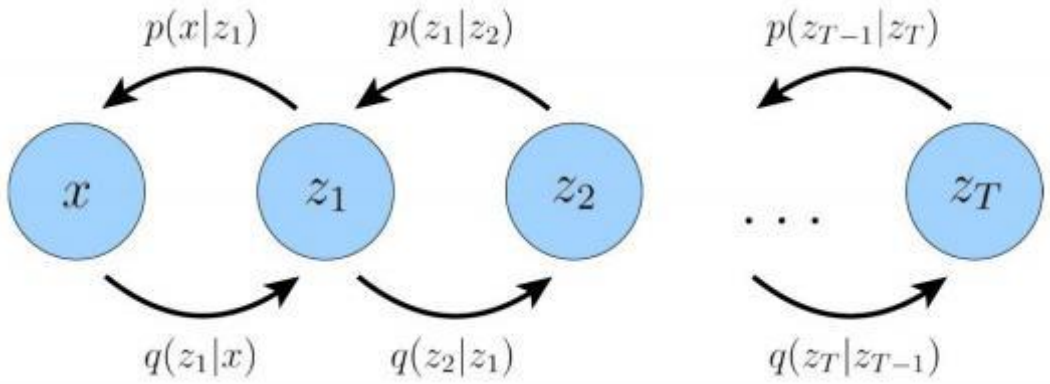


级联

Hierarchical

马尔科夫链

Markovian



MHVAE

Forward Diffusion Process

定义

一个真实分布的data，分布满足 $x_0 \sim q(x)$

一共加了T次噪声，得到$x_1, x_2, \ldots, x_T$

每次加噪声操作是对前一次加完噪声的结果操作(Markov chain)，

$$q(x_t|x_{t-1}) = N\left(x_t; \sqrt{1-\beta_t} * x_{t-1}, \beta_t I\right)$$

β_t是超参数，在DDPM中，β_t是随着t线性增长的（随着t的增大，x_t趋近于标准高斯分布）

加噪过程

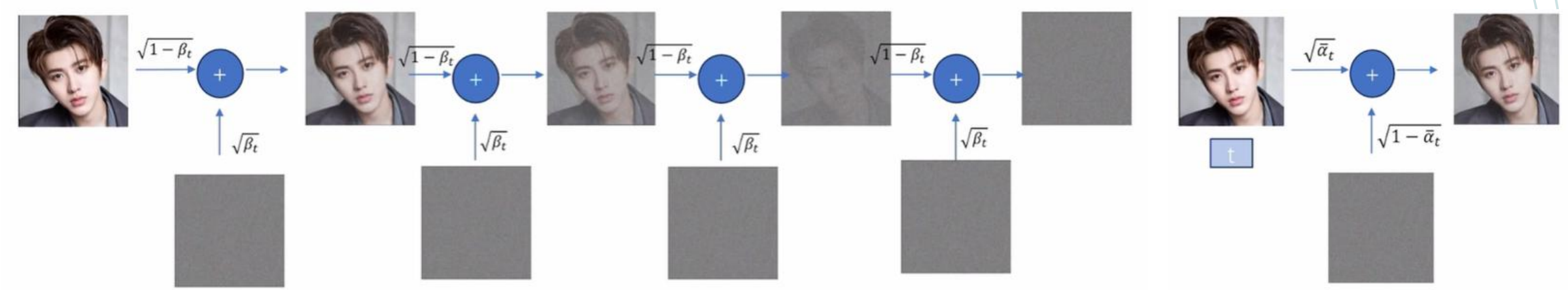
$$x_t = \sqrt{1-\beta_t} * x_{t-1} + \sqrt{\beta_t} * \varepsilon_{t-1}$$

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

其中 $\alpha_t = 1 - \beta_t$

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

给定图片和时间t，
前向扩散是一个确定的过程。



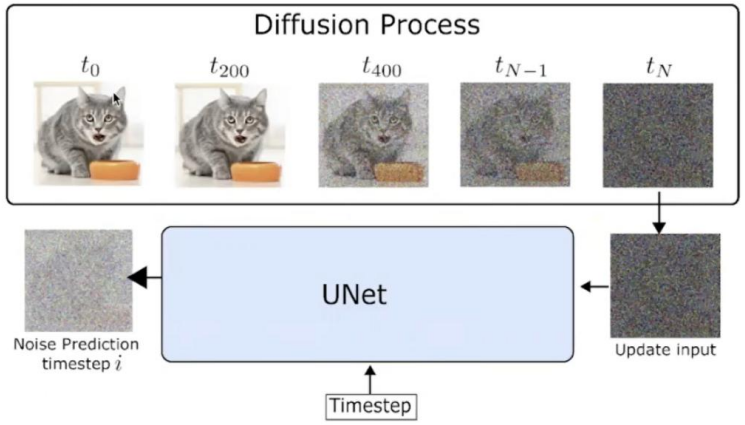
Reverse Diffusion Process

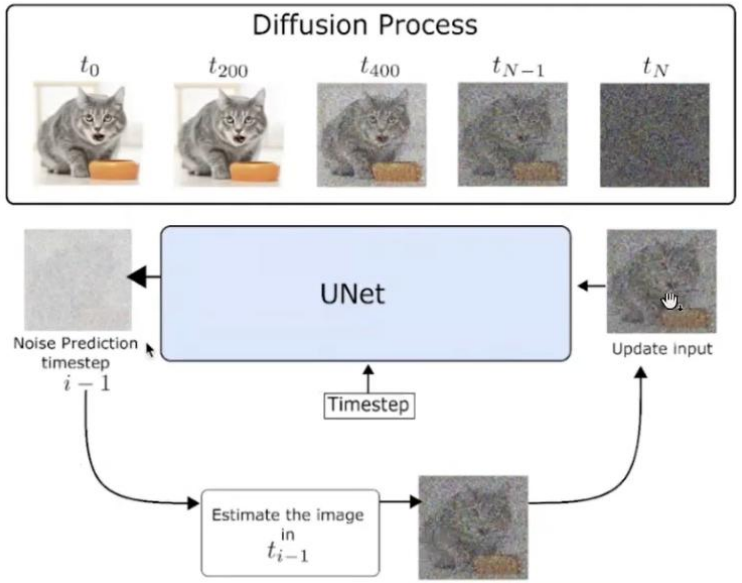
$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$$

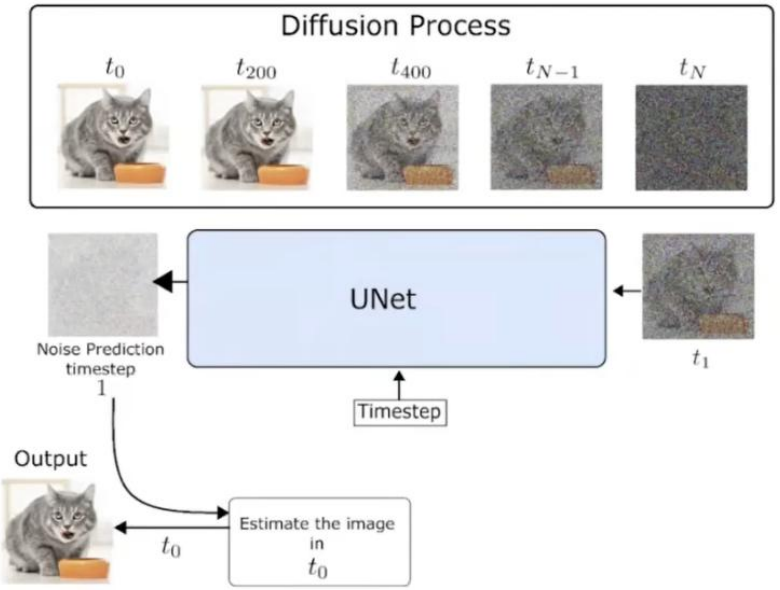
其中：

- $\mathbf{x}_t$：时间步t时的状态。
- $\mathbf{x}_{t-1}$：时间步$t-1$时的状态。
- α_t：与噪声添加过程相关的参数。
- $\bar{\alpha}_{t-1}$：累积噪声参数，表示从时间步0到时间步$t-1$的噪声水平。
- $\epsilon_\theta(\mathbf{x}_t, t)$：由神经网络$\theta$预测的噪声。
- σ_t：与时间步t相关的噪声标准差。
- $\mathbf{z}$：从标准正态分布中采样的噪声。

1. **去噪部分**：
   $\frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right)$
   - 这一部分用于从当前的状态$\mathbf{x}_t$中去除噪声。首先，将$\mathbf{x}_t$减去由神经网络预测的噪声项，随后乘以$\frac{1}{\sqrt{\alpha_t}}$，这一步可以视为去噪和重构数据。
2. **噪声添加部分**：
   $+\sigma_t \mathbf{z}$
   - 在去噪的基础上，添加一个新采样的噪声项$\sigma_t \mathbf{z}$，其中$\mathbf{z}$是从标准正态分布中采样的随机噪声。这一步确保生成过程的随机性，模拟真实数据的多样性。







DDPM

$$L_{t-1} = \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t,t)\|^2\right] + C$$

$$\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right]$$

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

三个限制：

- ✓ 数据x和所有隐变量z_t的维度相同。
- ✓ 所有的encoder $q(z_t|z_{t-1})$都不需要学习，而是预定义好的高斯分布模型，就是z_t状态为以z_{t-1}为均值的高斯分布。
- ✓ 最终T状态的分布z_T为标准高斯分布

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** $\mathbf{x}_0$

IDDPM

1. 学习反向扩散过程的方差

原始DDPM使用固定的方差（如β_t或$\tilde{\beta}_t$），而改进模型通过学习反向扩散过程的方差$\Sigma_\theta(xt,t)$，使用一个混合目标函数来优化方差。这一改进减少了采样步骤的数量，同时保持了高质量的样本生成。

2. 混合目标函数（Hybrid Objective）

IDDPM引入了一个混合目标函数（L_{hybrid}），结合了变分下界（VLB）和简化目标（L_simple）。这一目标函数不仅提高了对数似然，还减少了梯度噪声，提高了训练的稳定性。

$L_{hybrid} = L_{simple} + \lambda L_{vlb}$

其中，λ是一个小权重系数，用于平衡两个目标。

3. 重要性采样（Importance Sampling）

在优化变分下界时，使用了重要性采样技术来减少梯度噪声。通过动态调整每个时间步的采样概率，显著提高了对数似然的优化效果。

$L_{vlb} = E_{t\sim p_t}\left[\frac{L_t}{p_t}\right]$

其中，p_t与每个损失项的平方期望成正比。

4. 余弦噪声调度策略（Cosine Noise Schedule）

IDDPM使用余弦噪声调度策略来替代原始DDPM的线性噪声调度策略，使噪声在整个扩散过程中的增加更加平滑，避免了噪声水平的突然变化，从而提高了生成质量。

$\alpha_t = \cos\left(\frac{t/T+s}{1+s}\cdot\frac{\pi}{2}\right)^2$

其中，s是一个小偏移量，用于避免在过程开始时噪声过小。

5. 采样步骤的优化

通过引入学习到的方差，改进模型能够在更少的采样步骤下生成高质量样本。实验表明，IDDPM在减少采样步骤的情况下，仍然能够保持高质量的生成样本，这是通过参数化方差和改进的采样策略实现的。

6. 理论分析和扩展性

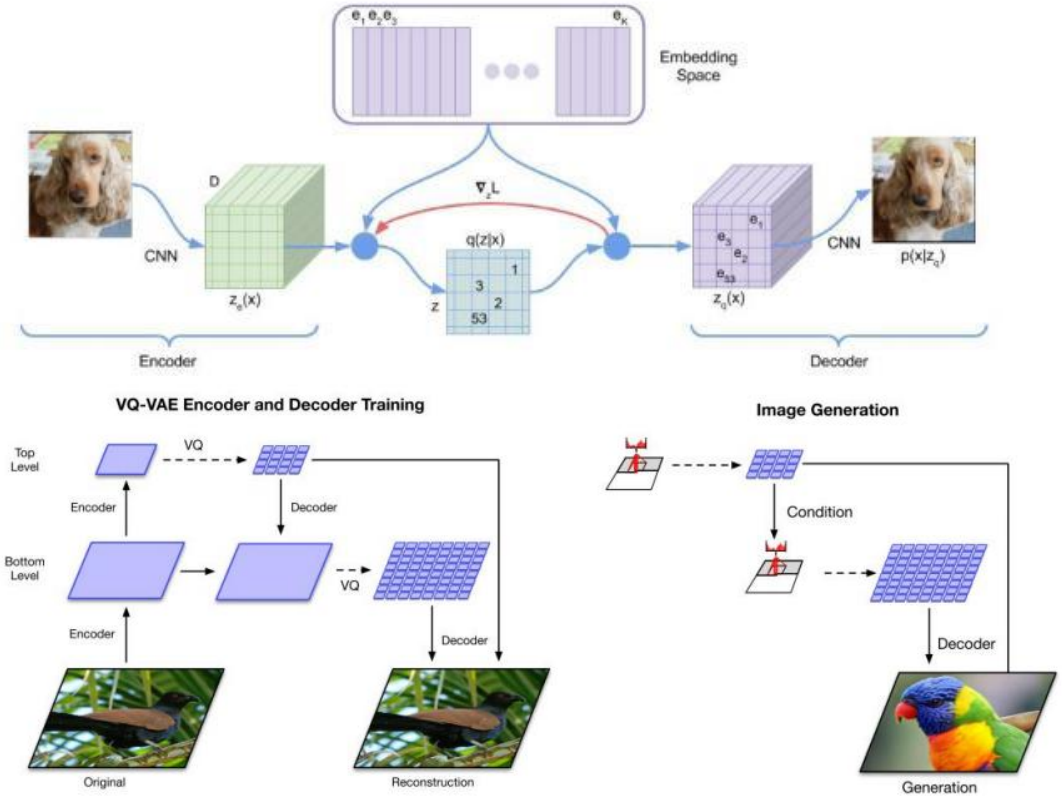
文章还通过理论分析展示了这些改进如何在不同数据集和模型规模上提升性能。随着模型容量和训练计算量的增加，IDDPM在对数似然和生成质量上表现出良好的扩展性，这表明这些改进具有广泛的应用潜力。

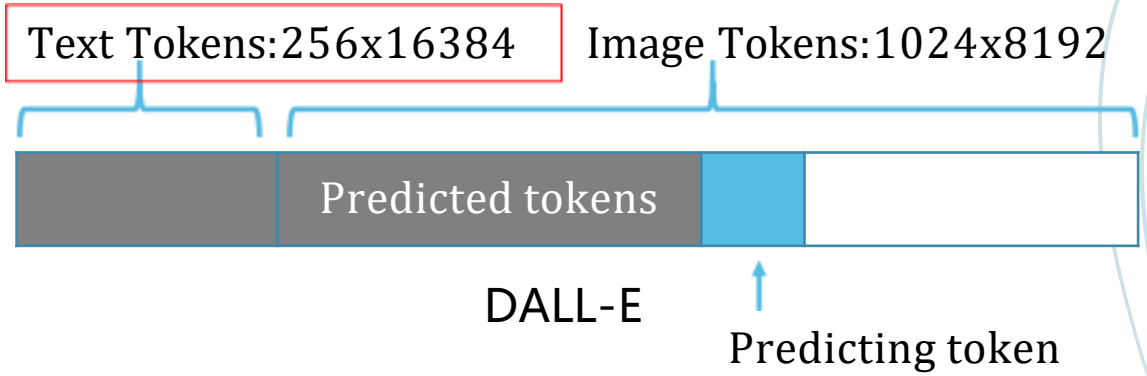
DALL-E

DALL-E沿用VQ-VAE两阶段训练模型范式。

Stage 1：训练Encoder， Decoder，得到图像codebook(size： 512 -> 8192)
Stage 2：训练Prior(PixelCNN -> 12B 基于transformer的自回归模型(GPT))

Prior改进： 图像单模态 - > 文本图像多模态



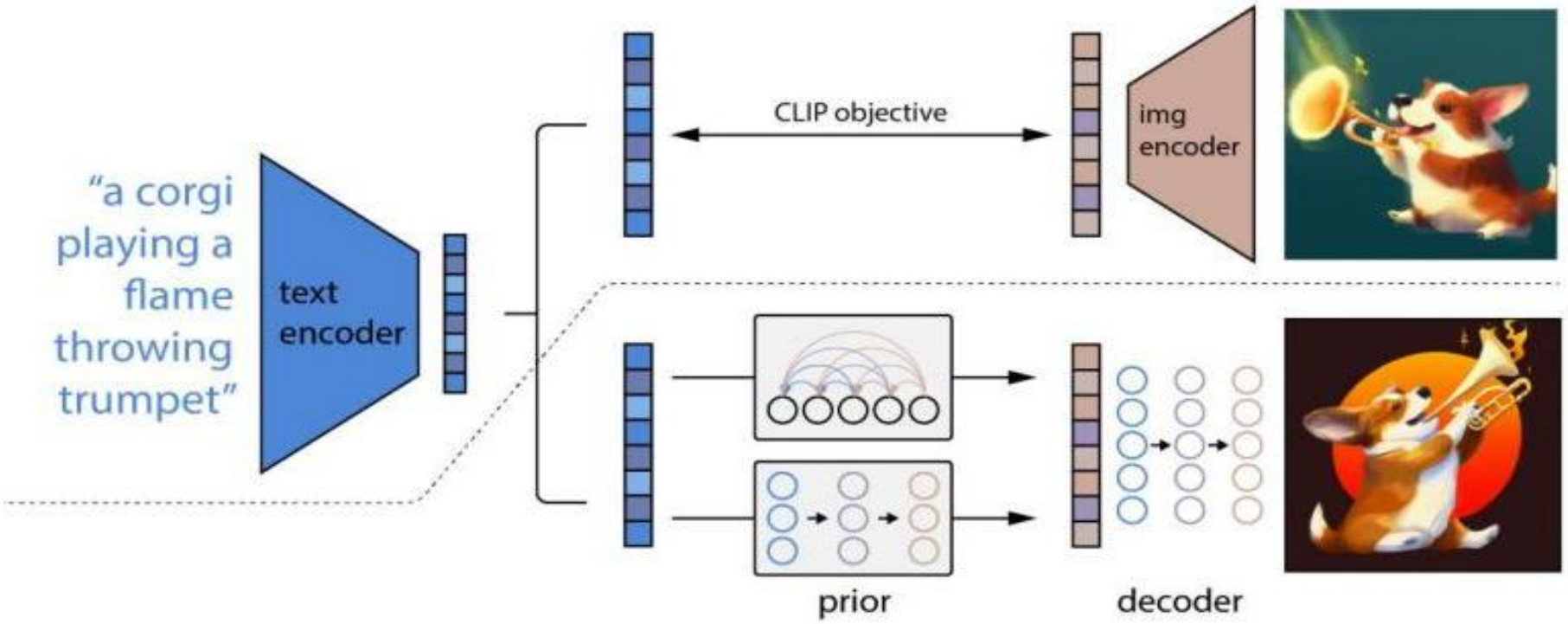


Reranking： 同1个text生成多张图片， 根据CLIP-score筛选好的生成结果。

Scale：2.5亿图片文本对，12B参数，使用16GB V100显卡
Mixed-Precision Training， Distributed Optimization

DALL-E 2

- 沿用DALL-E的两阶段prior+decoder架构
- Prior： 自回归模型 -> 扩散模型，从文本特征得到图像特征， $p(z_i|y)$
- Decoder：VQVAE的decoder -> 扩散模型，从图像特征得到图像$p(x|z_i, y)$ 借用CLIP模型作为监督信号训练prior



CLIP

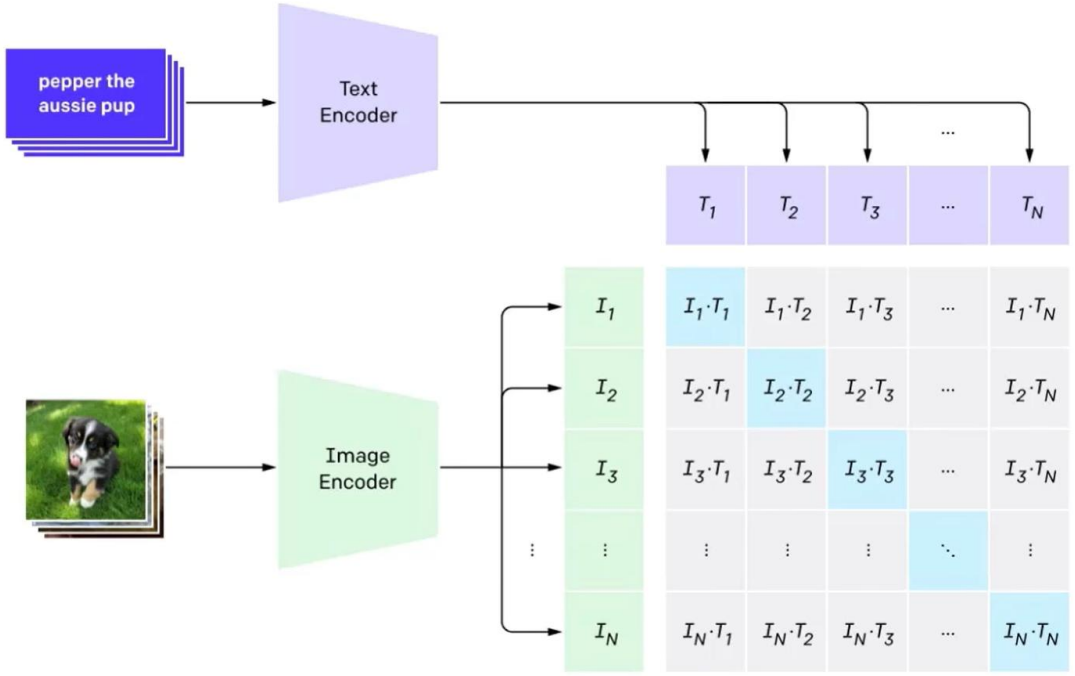
模型训练：

图像编码器->特征

文本编码器->特征

计算余弦相似度

对角线的是正样本



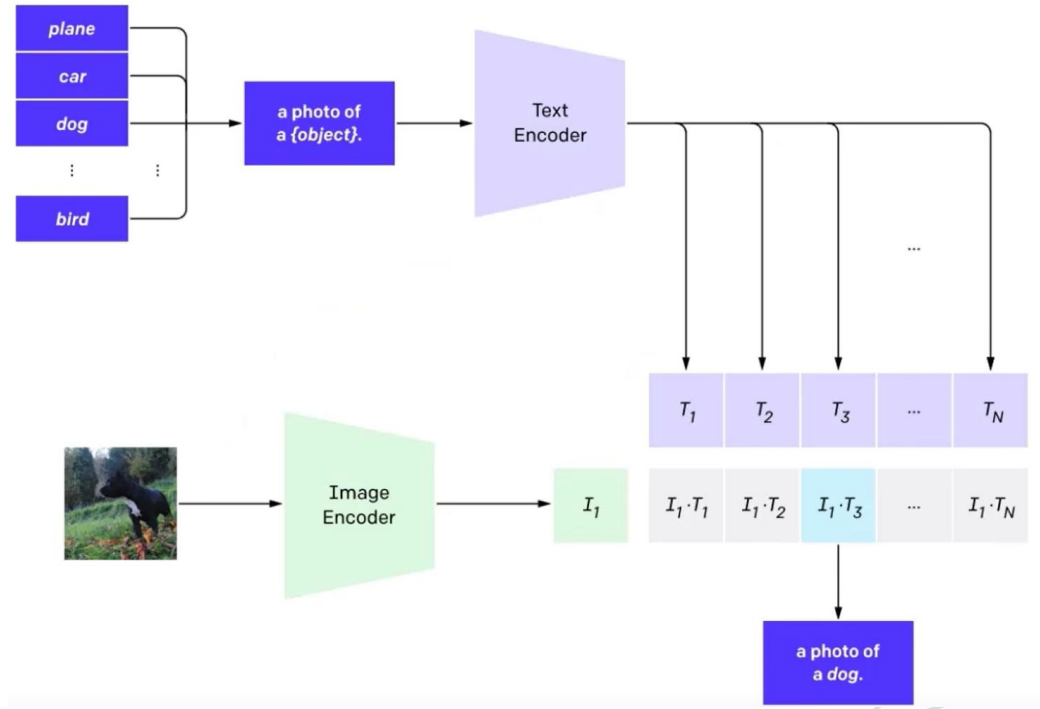
模型推理：

给一些提示文本(任意个)

提示要好好说话

然后每种提示算相似度

找到概率最高的即可



DALL-E 2

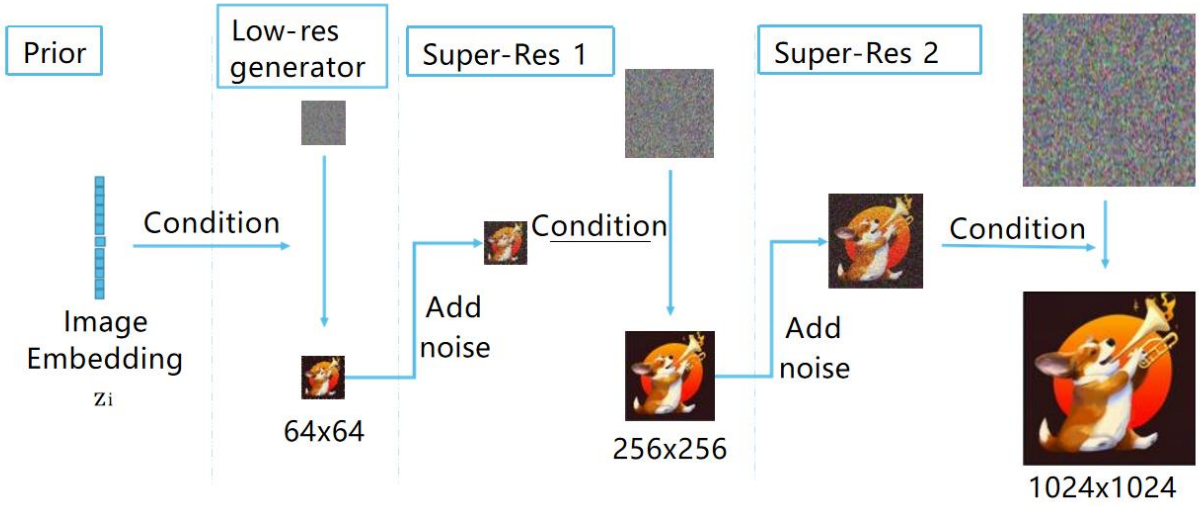
- Prior：
  - AR：文本embedding作为已知，逐个预测image embedding的每个位置的feature（同DALL-E）

  - Diffusion：通过一个transformer-based的模型预测未加噪的image embedding

$$L_{\text{prior}} = \mathbb{E}_{t\sim[1,T],z_i^{(t)}\sim q_t}\left[\|f_\theta(z_i^{(t)},t,y) - z_i\|^2\right.$$

  - Diffusion vs AR： Diffusion效果更优，且训练更高效

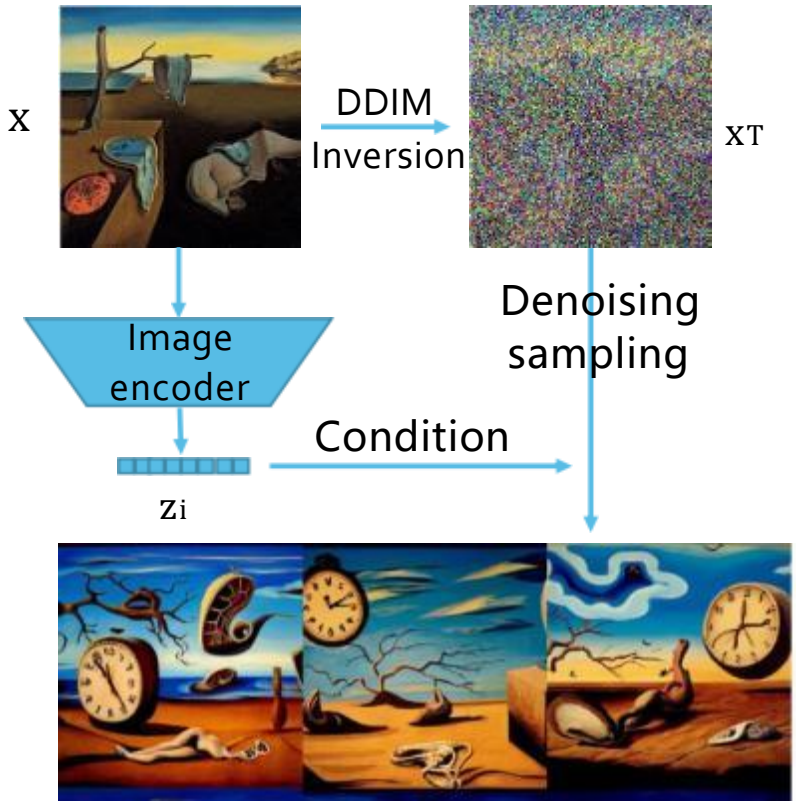
- Decoder：
  - 一个以prior输出的image embedding作为condition，生成64x64分辨率的普通扩散生成模型
  - 两个级联的升采样扩散模型，分别将分辨率升成256x256,1024x1024，每一级都以前一级的输出作为condition



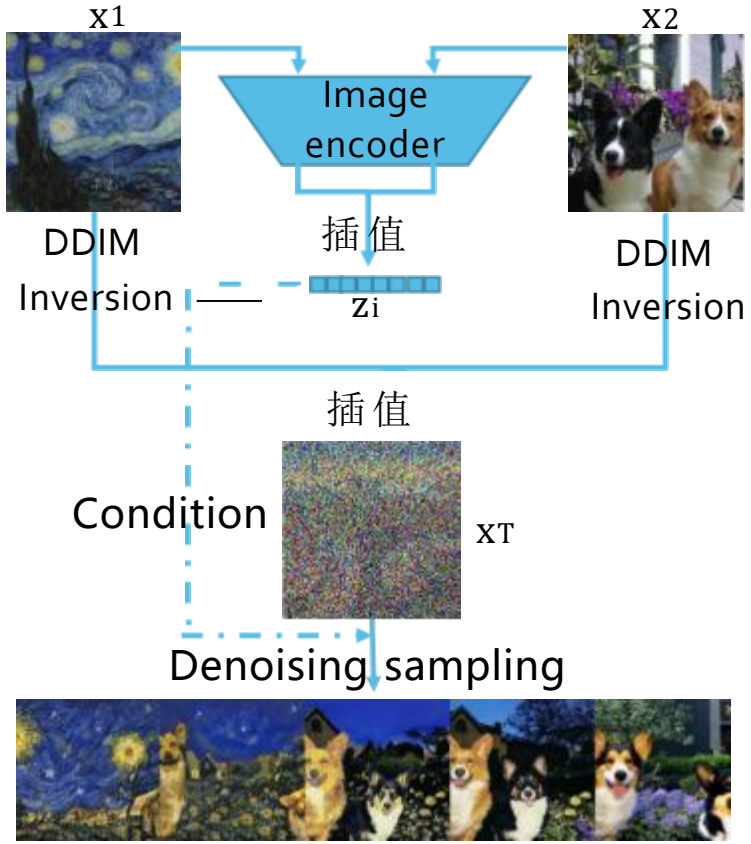
DALL-E 2

CLIP image encoder和text encoder的存在，他们将文本和图像映射到同一个隐空间z_t, z_i。我们拿到文本和图片之后，就可以对他们在隐空间上的特征进行多种操作。

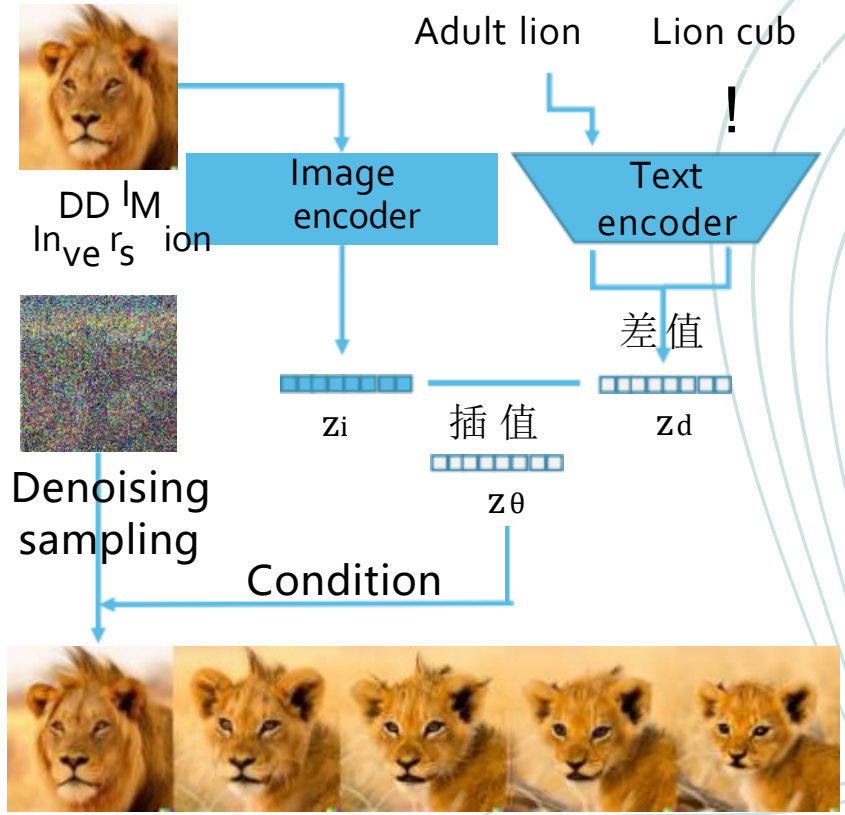
图片变体（Variations）



图片插值（Interpolations）



文本偏移（Text Diffs）



DALL-E 2

模型不足：

- 特征混淆
- 文字无法准确生成
- 复杂场景细节缺失严重

Figure 16: Samples from unCLIP for the prompt, "A sign that says deep learning."

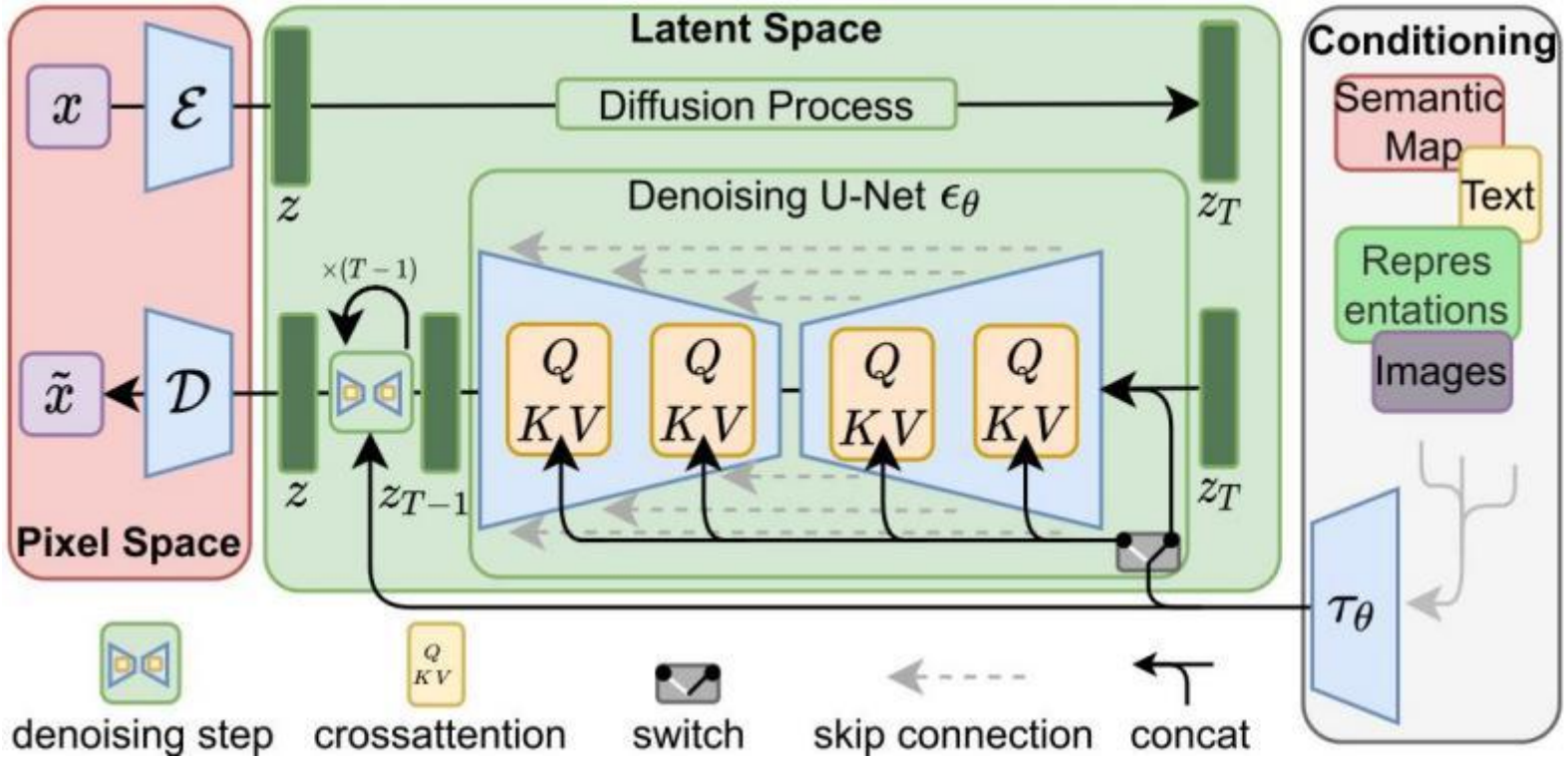
Figure 15: Reconstructions from the decoder for difficult binding problems. We find that the reconstructions mix up objects and attributes. In the first two examples, the model mixes up the color of two objects. In the rightmost example, the model does not reliably reconstruct the relative size of two objects.

(b) A high quality photo of Times Square.

LDM（Stable Diffusion）

和之前一些模型比，两个关键改动：

- 通过VAE将像素空间映射到隐空间，隐空间上进行扩散和去噪，实现高分辨率生成
- Cross attention实现不同类型的condition对模型的引导



2022年8月stability AI公司发布并开源Stable Diffusion模型。

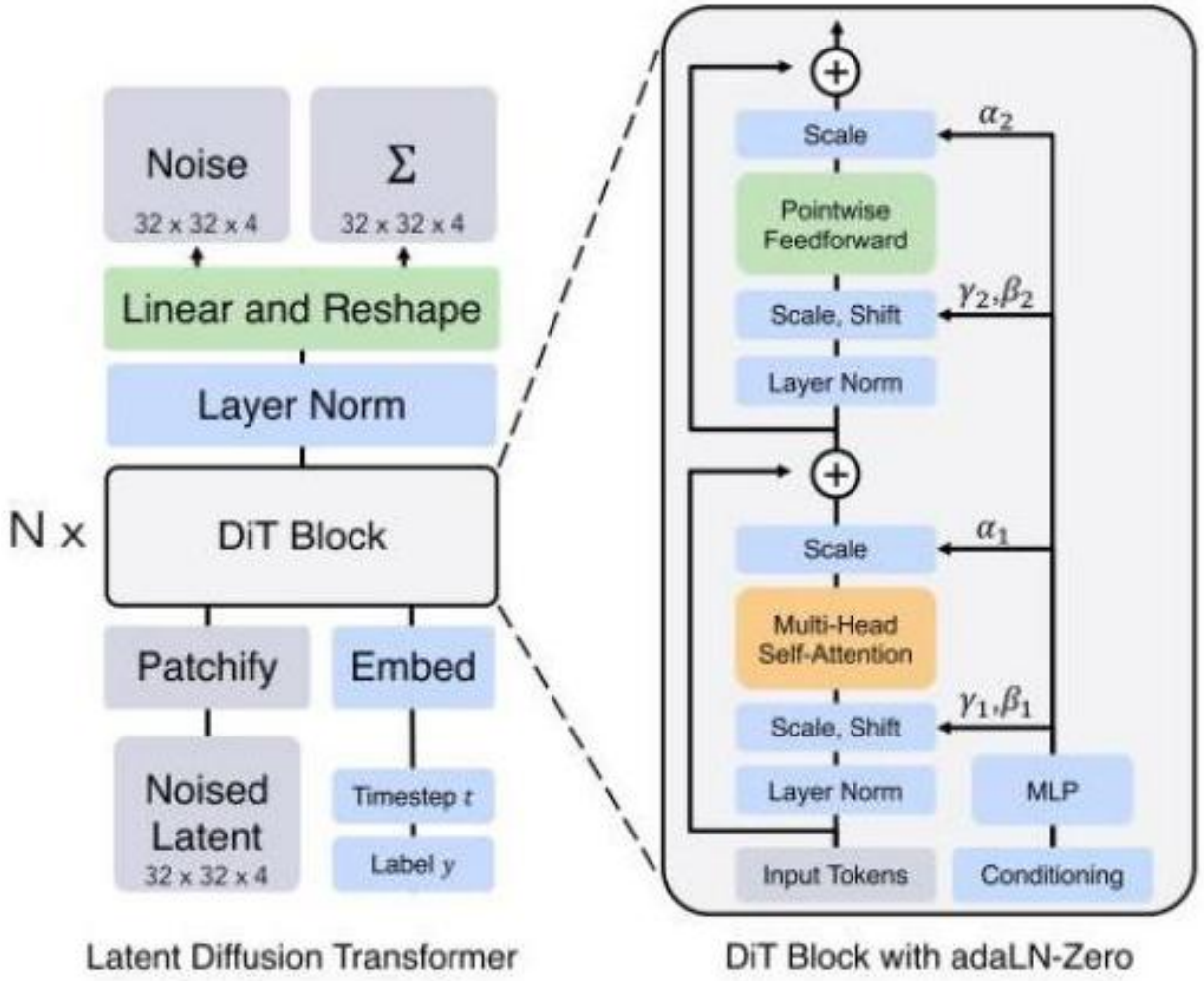
相关模型井喷式出现，例如controlnet，IP-adaptor，等等，反过来进一步提升了SD的流行程度。

DiTs（Diffusion Transformers）

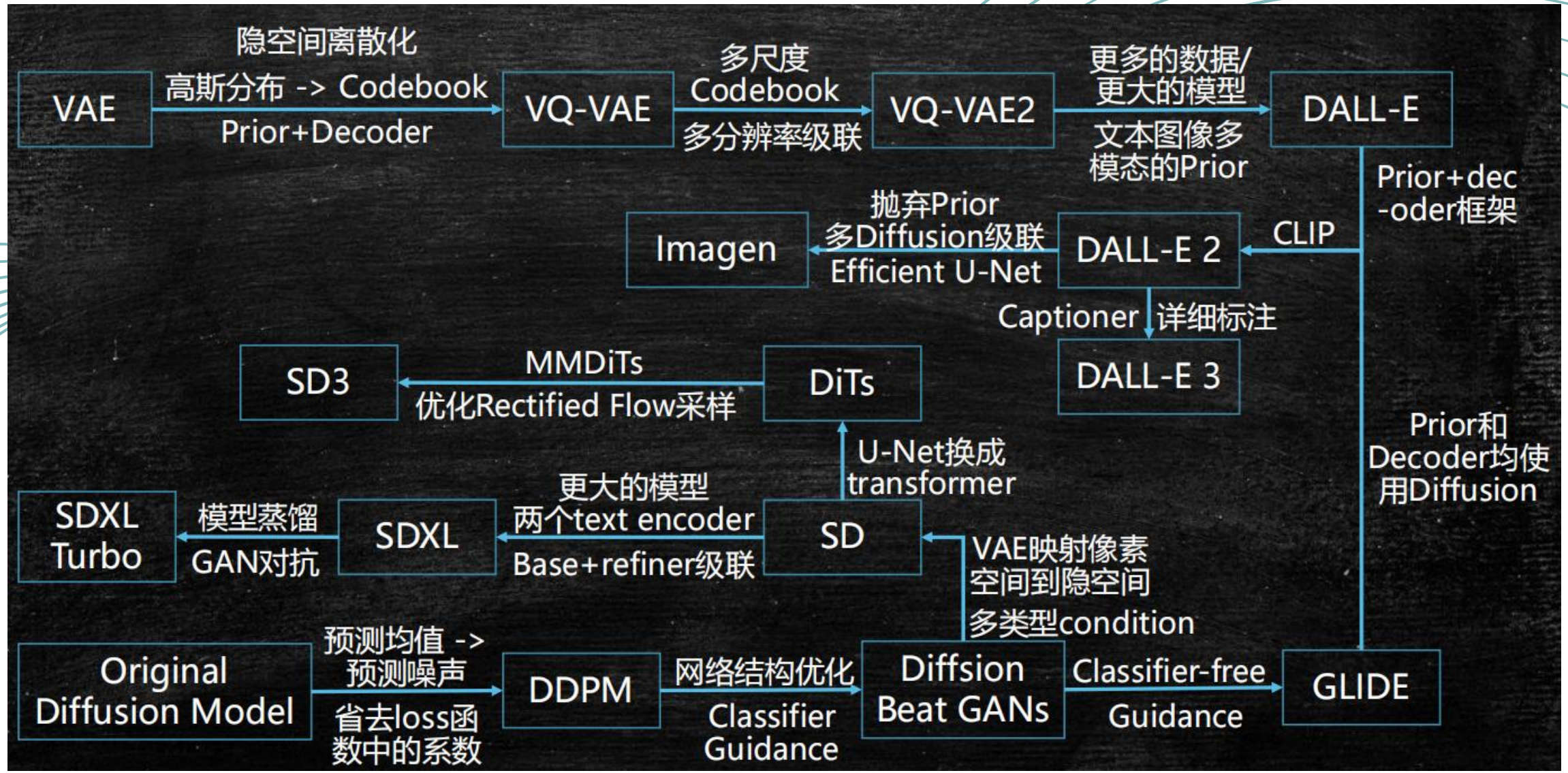
主要贡献：

- 将去噪模型从U-Net结构换成transformer结构。
- 实验证明了基于Transformer的去噪模型也有很好的可扩展性scalability。
- 基于DiTs的模型性能达到SOTA。

- DiTs逐渐取代U-net的趋势已显，例如SD3、Sora等。



整体梳理


VAE
隐空间离散化
高斯分布 -> Codebook
Prior+Decoder
VQ-VAE
多尺度
Codebook
多分辨率级联
VQ-VAE2
更多的数据/
更大的模型
文本图像多
模态的Prior
DALL-E
Prior+dec
-oder框架
Imagen
抛弃Prior
多Diffusion级联
Efficient U-Net
DALL-E 2
CLIP
Captioner
详细标注
DALL-E 3
SD3
MMDiTs
优化Rectified Flow采样
DiTs
U-Net换成
transformer
Prior和
Decoder均使
用Diffusion
SDXL
Turbo
模型蒸馏
GAN对抗
SDXL
更大的模型
两个text encoder
Base+refiner级联
SD
VAE映射像素
空间到隐空间
多类型condition
Original
Diffusion Model
预测均值 ->
预测噪声
省去loss函
数中的系数
DDPM
网络结构优化
Classifier
Guidance
Diffsion
Beat GANs
Classifier-free
Guidance
GLIDE

感谢观看

张心悦