扩散模型 (Diffusion Model)

作者：Calvin
QQ: 179209347
Mail: 179209347@qq.com

介绍

笔记简介：

- 面向对象：深度学习初学者
- 依赖课程：**线性代数，统计概率**，优化理论，图论，离散数学，微积分，信息论

知乎专栏：

https://zhuanlan.zhihu.com/p/693738275

Github & Gitee 地址：

https://github.com/mymagicpower/AIAS/tree/main/deep_learning

https://gitee.com/mymagicpower/AIAS/tree/main/deep_learning

*** 版权声明：**

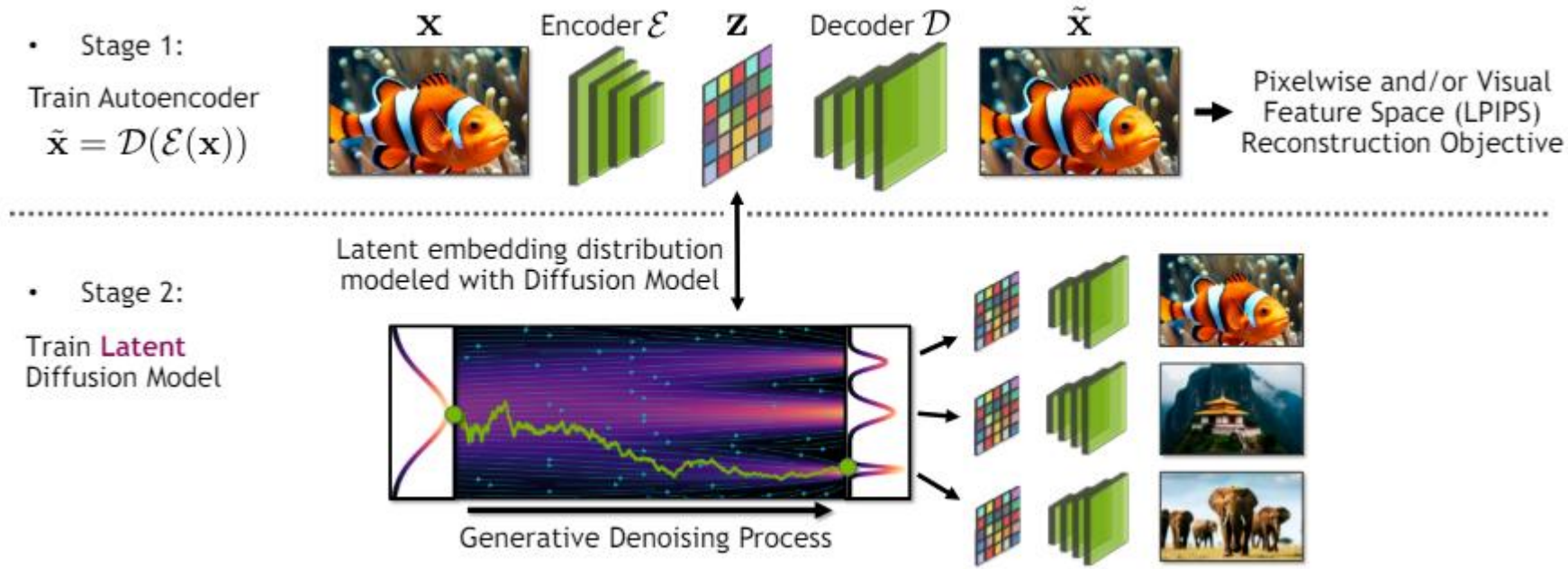
- 仅限用于个人学习
- 禁止用于任何商业用途

潜在扩散模型（LDM - Latent Diffusion Model）

潜在扩散模型（Latent Diffusion Model, LDM）是一种用于生成图像的深度学习模型，其核心思想是通过**在潜在空间中进行扩散过程**来生成图像。LDM将生成任务分解为一个通过噪声到数据的转换过程，使得模型能够高效地生成高质量的图像。

优势：

- **压缩潜在空间：**在低分辨率的潜在空间中训练扩散模型计算效率更高
- **规整的平滑/压缩潜在空间：**扩散模型任务更容易，采样速度更快
- **灵活性：**自动编码器可以根据数据（图像、视频、文本、图形、3D点云、网格等）进行定制

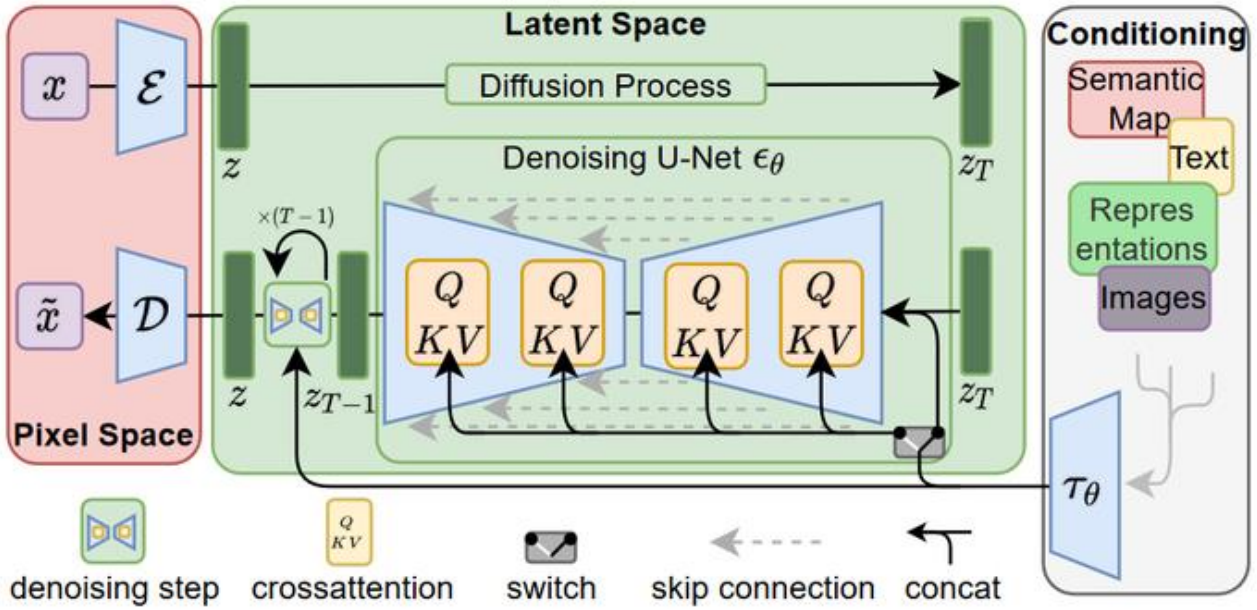


https://neurips2023-ldm-tutorial.github.io/

潜在扩散模型（LDM）

Latent Diffusion Model （LDM)

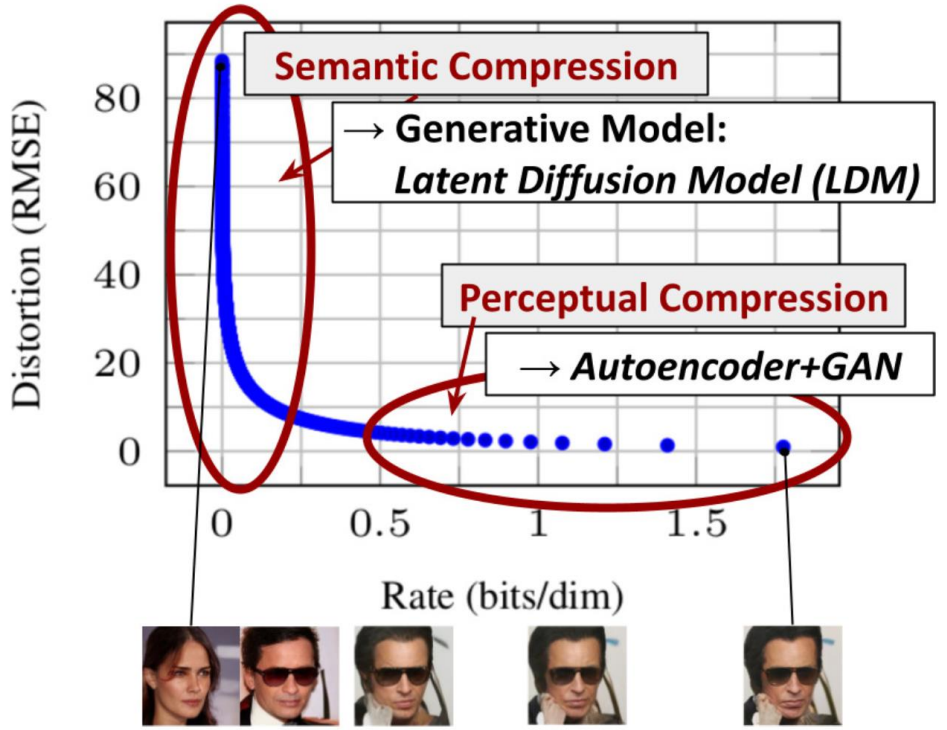
- 感知图像压缩（Perceptual Image Compression）：最左侧红框部分是一个VQ-VAE，用于将输入图像 x 编码为一个离散特征 z 。
- LDM：图中间绿色部分是在潜变量空间的扩散模型，其中上半部分是加噪过程，用于将特征 z 加噪为 z_T 。下半部分是去噪过程，去噪的核心结构是一个由交叉注意力（Cross Attention）组成的U-Net，用于将 z_T 还原为 z 。
- 条件机制（Conditioning Mechanisms）：右侧是一个条件编码器，用于将图像，文本等前置条件编码成一个特征向量 τ_θ ，并将其送入到扩散模型的去噪过程中。



- x 是输入图像
- $\tilde{x}$ 表示生成的圈像
- $\mathcal{E}$ 是编码器
- $\mathcal{D}$ 是解码器
- z 是潜在向量
- z_T 是增加噪声后的潜在向量
- τ_θ 是文本/图像的编码器（如：CLIP)，实现了语义压缩。

潜在扩散模型（LDM）

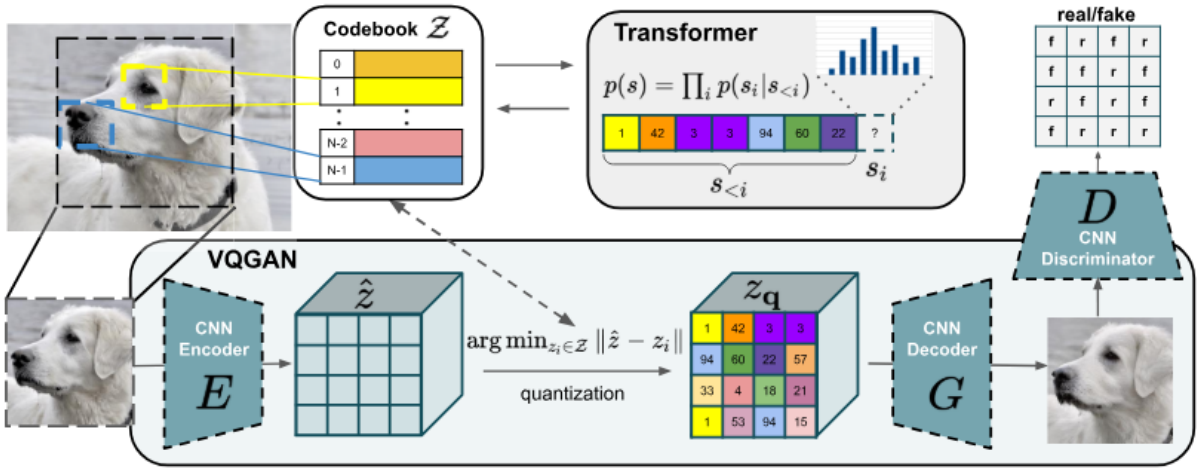
潜在扩散模型（LDM）在潜在空间而非像素空间中运行扩散过程，使训练成本更低，推断速度更快。其灵感来自观察到图像的大多数位对感知细节有贡献，并且在进行激进压缩后，语义和概念构成仍然存在。LDM通过首先利用自动编码器削减像素级冗余，然后在学习到的潜在空间上通过扩散过程操纵/生成语义概念，松散地分解了感知压缩和语义压缩。

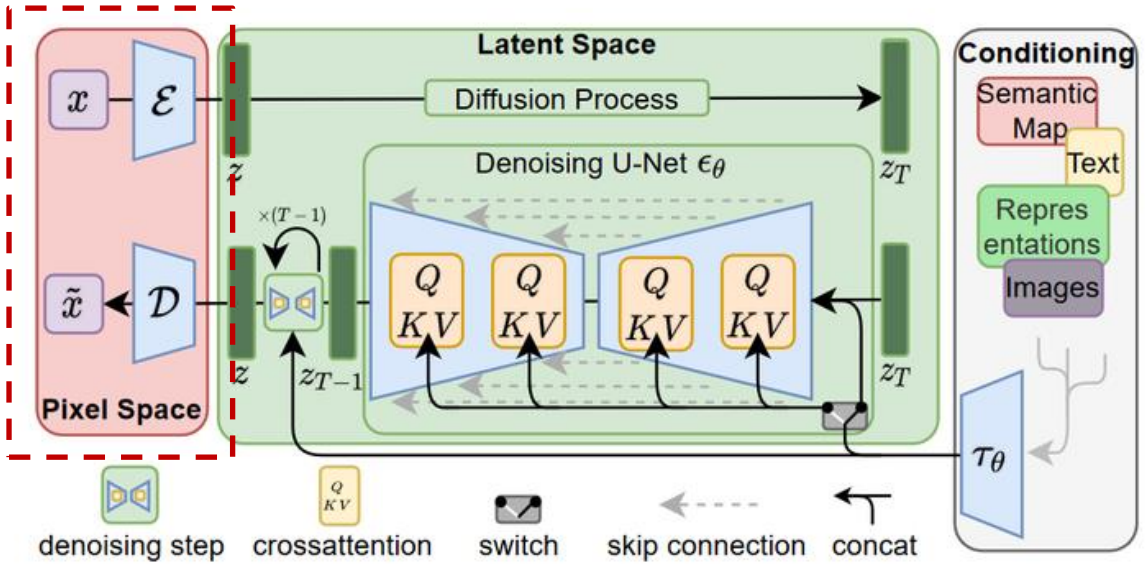


潜在扩散模型（LDM ）- 感知图像压缩

第一块模型选择用预训练好的 VQGAN 或者 VQ-VAE 来把图像降维。大部分LDM都选择的是VQGAN。
在感知图像压缩使用的训练好的VQ-GAN，它包括：

- 一个编码器 $\mathcal{E}$：编码器 $\mathcal{E}$ 用于将一个RGB彩色图像 $x \in \mathbb{R}^{H\times W\times 3}$ 压缩到一个特征向量 $z \in \mathbb{R}^{h\times w\times c}$。
- 一个解码器 $\mathcal{D}$：解码器 $\mathcal{D}$ 则用于将这个潜变量 z 还原为输入图像 $\tilde{x}$。



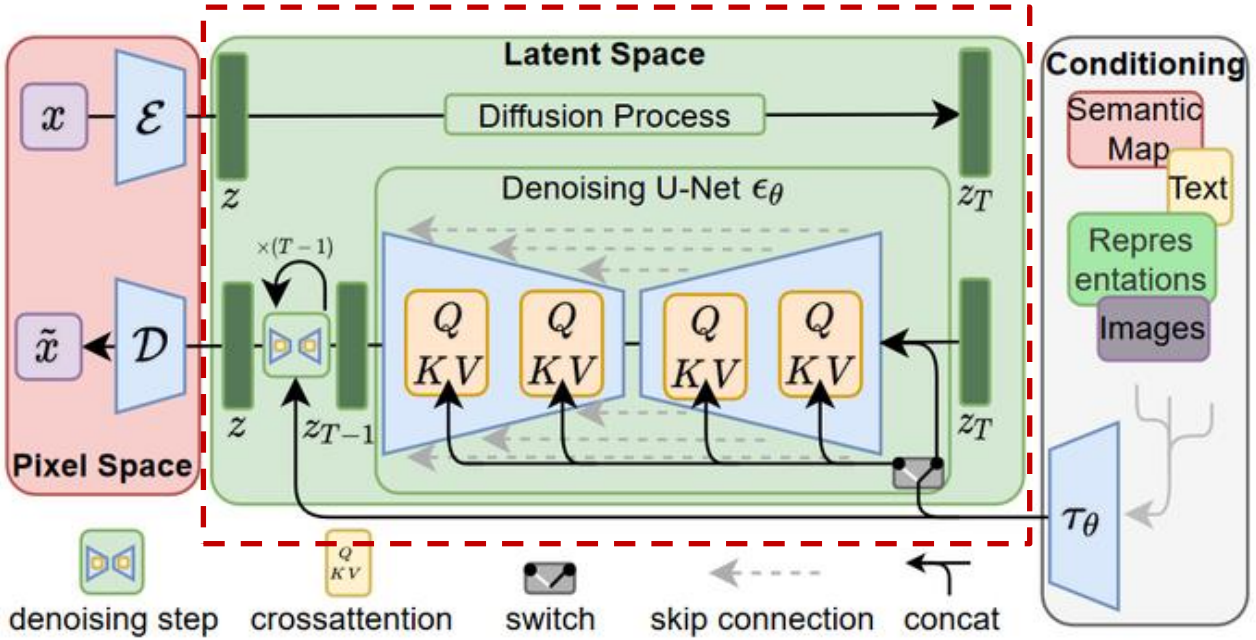


潜在扩散模型（LDM ）- LDM

第二块模型LDM：有了压缩后的图像潜变量，接下来对其进行扩散过程的加噪和去噪了。
因为LDM是作用在潜空间：

- 特征的大小要比图像空间小很多，推理速度是要快很多的。
- 更侧重于生成图像的语义信息而非图像的纹理细节。

U-Net将256*256*3的图像编码到潜空间中尺寸为32*32*4，并且在分辨率为32，16，8，4的层加入了self-attention layers和text-conditioned cross-attention layers。



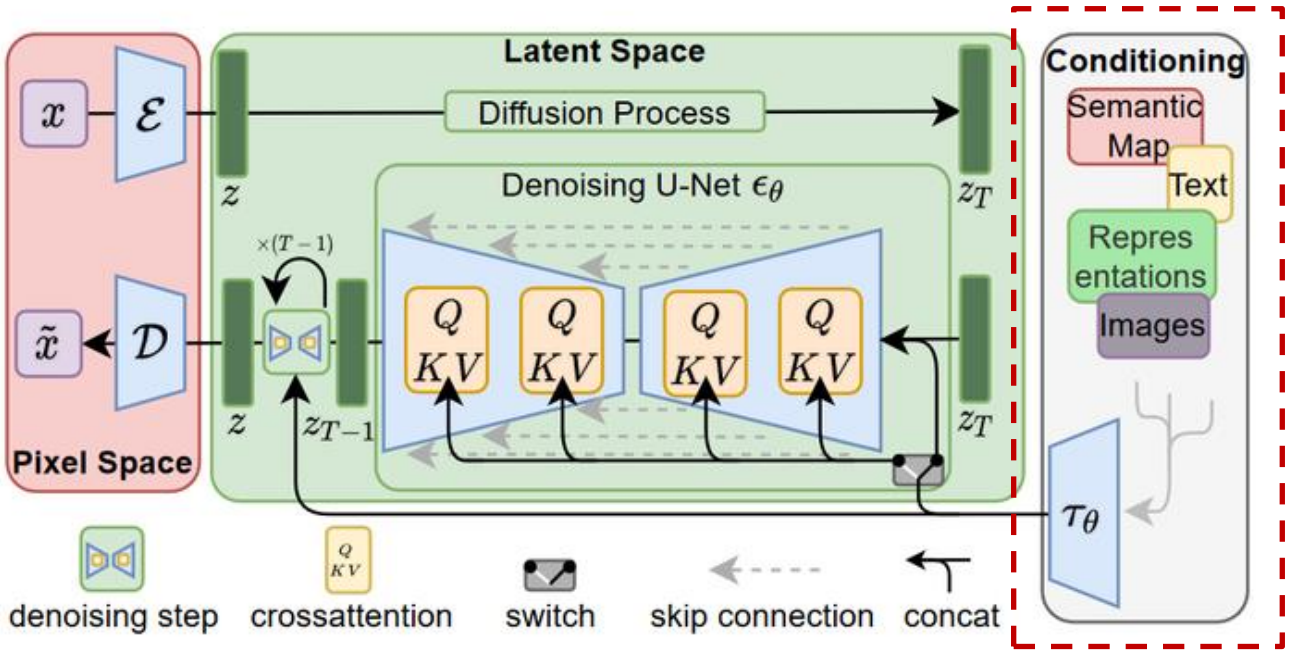
在训练时利用编码器得到 z_t ，在潜在表示空间中学习，相应的目标函数可以写成如下形式：

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\| \epsilon - \epsilon_\theta(z_t, t) \|_2^2\right]$$

潜在扩散模型（LDM ）- 条件机制

第三块模型是条件机制（Conditioning Mechanisms）：除了无条件图片生成外，也可以进行条件图片生成，主要是通过条件时序去噪自编码器（conditional denoising autoencoder ） $\epsilon_\theta(z_t, t, y)$ 来实现。

- $\epsilon_\theta(z_t, t, y)$ 通过在UNet主干网络上增加cross-attention，从而可以通过 y 来控制图片合成的过程。
- 论文引入了一个领域专用编码器（domain specific encoder）τ_θ ，将 y 映射为一个中间表示 $\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$，这样就可以很方便的引入各种形态的条件（文本、类别、layout等等）。
- 使用的是 32 层 transformer 与 U-Net 一起从头开始训练。



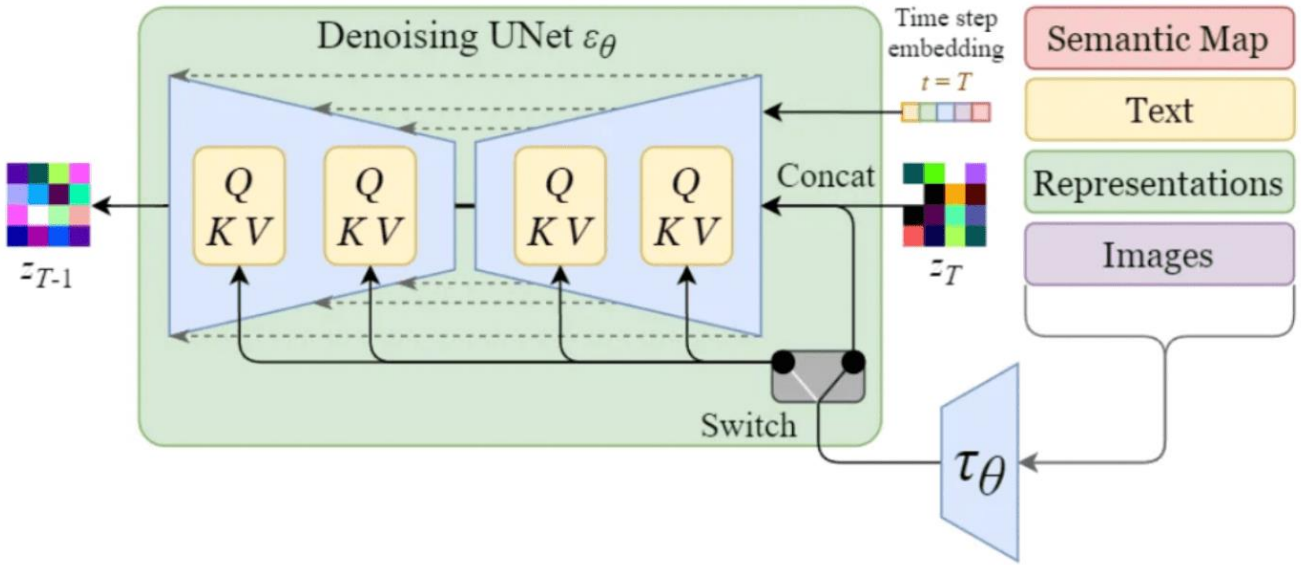
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \cdot \mathbf{V}$$

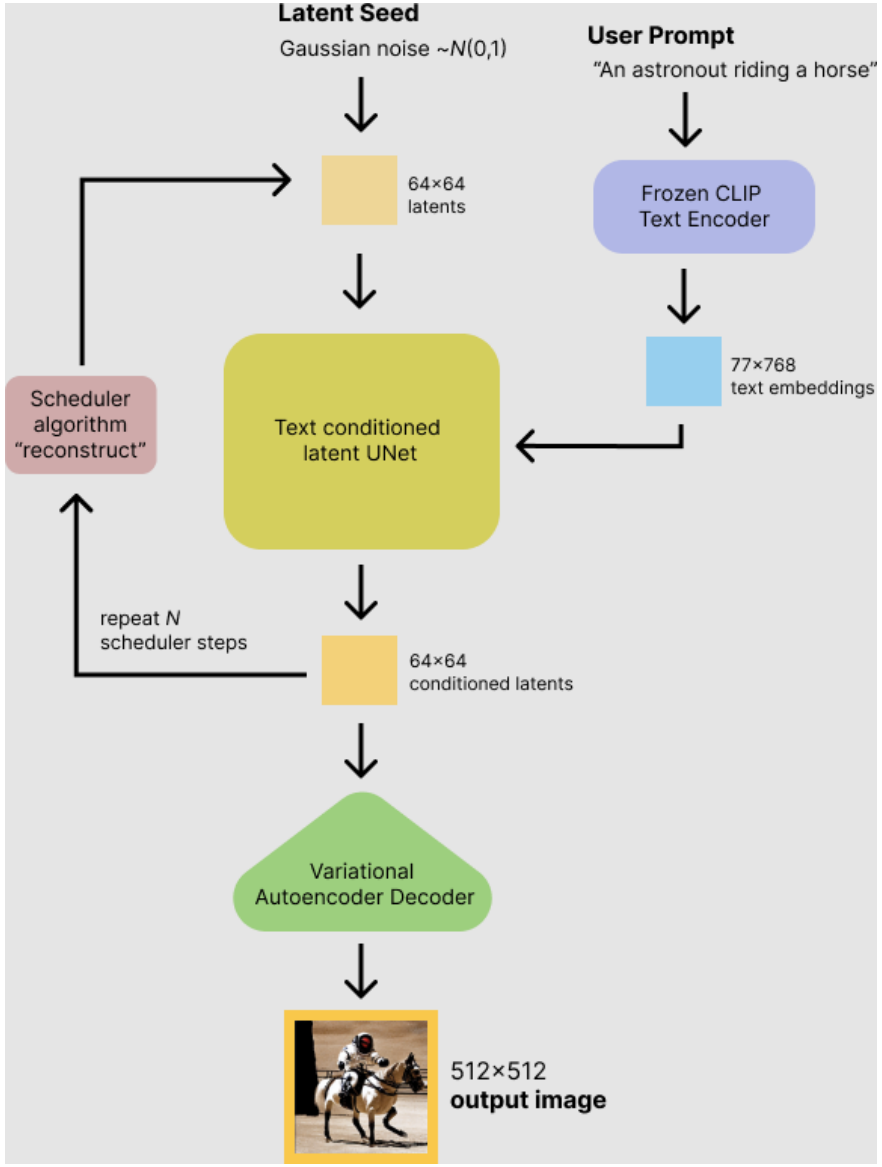
$$\text{where } \mathbf{Q} = \mathbf{W}_Q^{(i)} \cdot \varphi_i(\mathbf{z}_i),\ \mathbf{K} = \mathbf{W}_K^{(i)} \cdot \tau_\theta(y),\ \mathbf{V} = \mathbf{W}_V^{(i)} \cdot \tau_\theta(y)$$

$$\text{and } \mathbf{W}_Q^{(i)} \in \mathbb{R}^{d \times d_\epsilon^i},\ \mathbf{W}_K^{(i)}, \mathbf{W}_V^{(i)} \in \mathbb{R}^{d \times d_\tau},\ \varphi_i(\mathbf{z}_i) \in \mathbb{R}^{N \times d_\epsilon^i},\ \tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$$

稳定扩散模型 (Stable Diffusion Model)

稳定扩散是一种潜在的文本到图像扩散模型。由于Stability AI捐赠的计算资源以及LAION的支持，得以在LAION-5B数据库的子集上训练了一个Latent Diffusion模型，用于处理512x512的图像。类似于谷歌的Imagen，该模型使用了一个冻结的CLIP ViT-L/14文本编码器来根据文本提示对模型进行条件化。该模型具有8.6亿个UNet和1.23亿个文本编码器。





Latent Diffusion Model 和 Stable Diffusion Model 的区别联系

Latent Diffusion Model（LDM）

- 第一块模型选择用预训练好的VQGAN 或者VAE来把图像降维。官方大部分LDM都选择的是VQGAN。
- 第二块模型的U-Net将 256*256*3 的图像编码到潜空间中尺寸为32*32*4，并且在分辨率为32，16，8，4的层加入了self-attention layers和text-conditioned cross-attention layers。
- 第三块模型是条件机制（Conditioning Mechanisms），主要是text condition。

Stable Diffusion Model（LDM的升级版,使用更高分辨率的图像和更多的数据训练）

- 第一块模型选择用预训练好的VAE来把图像降维。
- 第二块模型的U-Net结构将 512*512*3 的图像编码到潜空间中尺寸为64*64*4，并且在分辨率为64，32，16，8的层加入了self-attention layers和text-conditioned cross-attention layers。以文本嵌入为条件迭代地对随机潜在图像表示进行去噪。
- 第三块模型的text encoder使用的是预训练好的CLIP text encoder。文本提示通过 CLIP 文本编码器转换为 77×768 的文本嵌入。

Stable Diffusion Model 应用：文生图与图生图

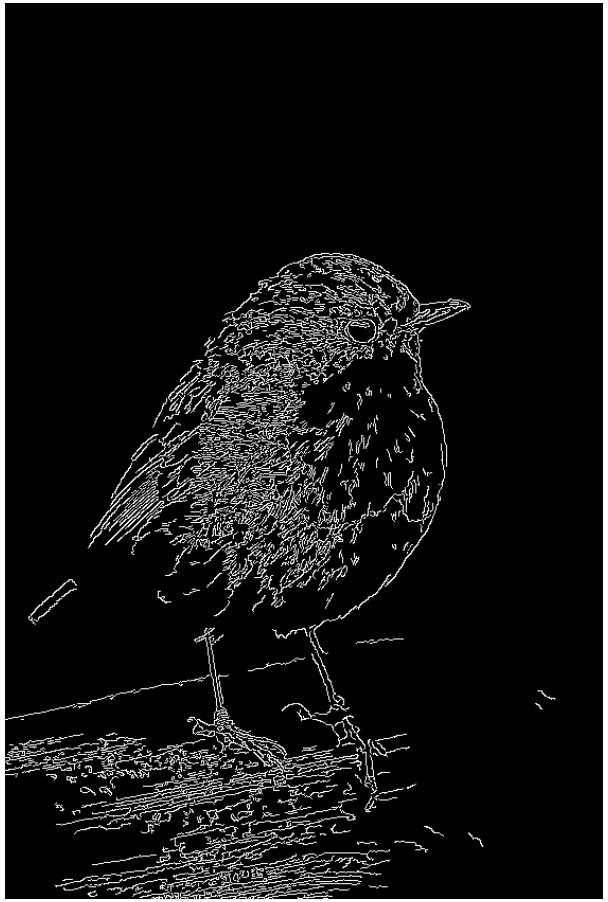
文生图测试

- 提示词
  a photo of an astronaut riding a horse on mars

- 生成图片效果如右图所示：

Stable Diffusion Model 应用： Canny 边缘检测图像生成

Canny 识别出图像内各对象的边缘轮廓，根据提示词生成新的图片。
对应ControlNet模型： control_canny

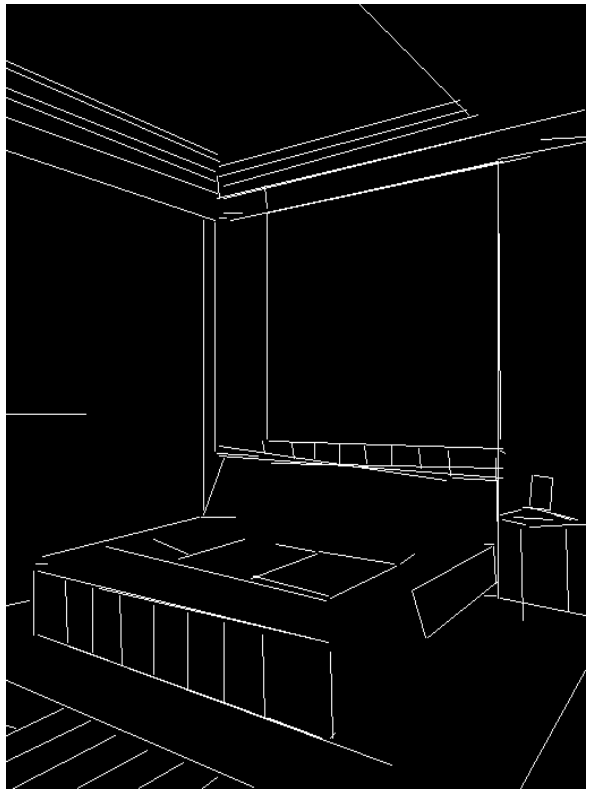


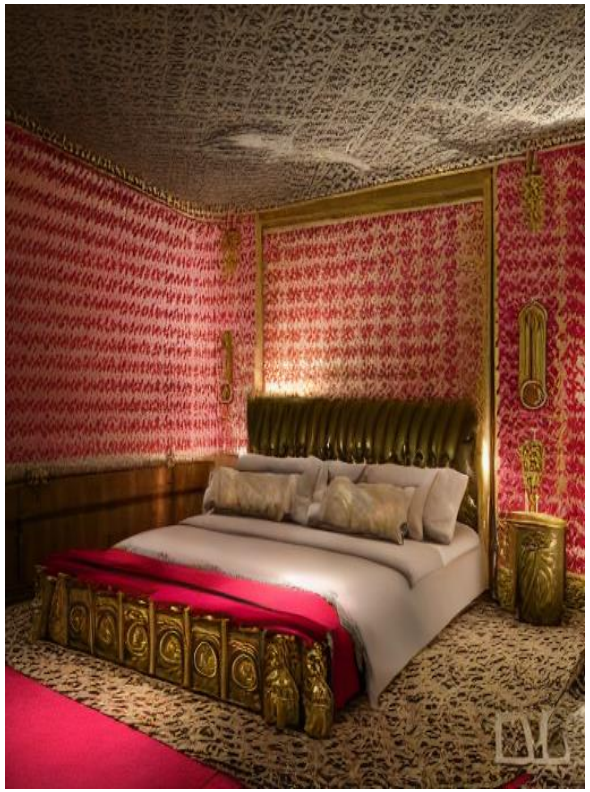
prompt = "masterpiece, best quality, ultra detailed, extremely detailed CG unity 8k wallpaper, best illumination, best shadow, an extremely delicate and beautiful, dynamic angle, finely detail, depth of field, bloom, shine, glinting stars, classic, illustration, painting, highres, original, perfect lighting";

Stable Diffusion Model 应用： MLSD 线条检测图像生成

MLSD 线条检测用于生成房间、直线条的建筑场景效果比较好。
对应ControlNet模型： control_mlsd

提示词："royal chamber with fancy bed";

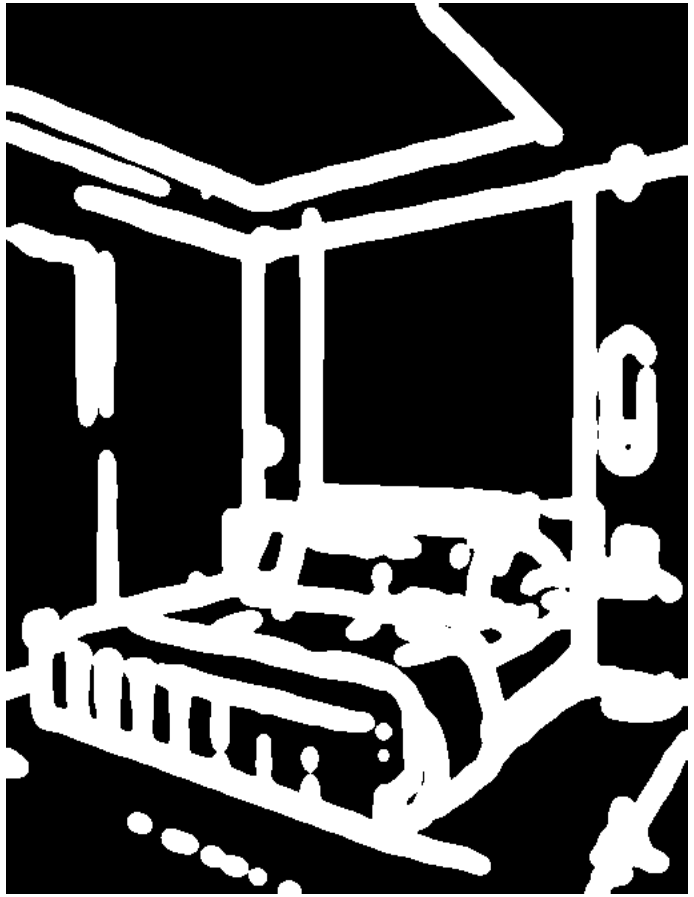




Stable Diffusion Model 应用： Scribble 涂鸦图像生成

根据涂鸦效果的草图线条生成新的图片。
对应ControlNet模型： control_scribble。

提示词： "royal chamber with fancy bed";



Stable Diffusion Model 应用： SoftEdge图像生成

SoftEdge 边缘检测可保留更多柔和的边缘细节，类似手绘效果。

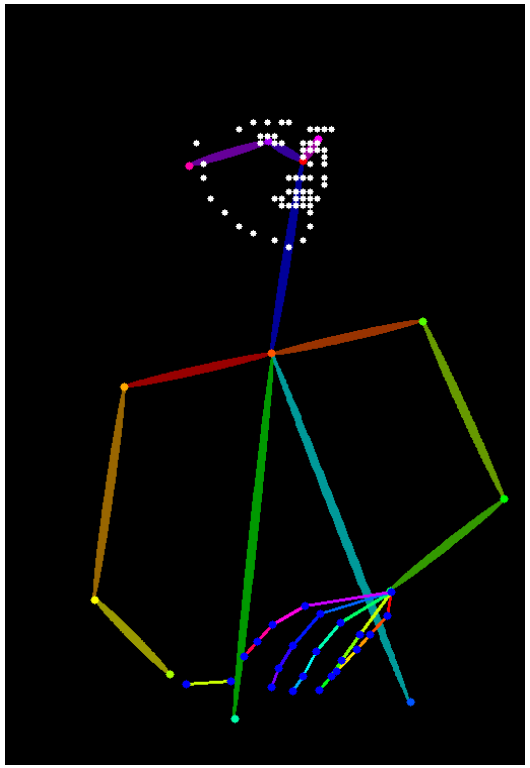
对应ControlNet模型： control_softedge。

提示词： "royal chamber with fancy bed";

Stable Diffusion Model 应用： OpenPose 姿态检测图像生成

OpenPose 姿态检测可生成图像中角色动作姿态的骨架图(含脸部特征以及手部骨架检测)，这个骨架图可用于控制生成角色的姿态动作。对应ControlNet模型：
control_openpose。

提示词： "chef in the kitchen";


Stable Diffusion Model 应用：语义分割图像生成

语义分割可多通道应用，原理是用颜色把不同类型的对象分割开，让AI能正确识别对象类型和需求生成的区界。

提示词： "old house in stormy weather with rain and wind";

对应ControlNet模型： control_seg。

Stable Diffusion Model 应用： Depth 深度检测图像生成

通过提取原始图片中的深度信息，生成具有原图同样深度结构的深度图，越白的越靠前，越黑的越靠后。

对应ControlNet模型： control_depth。

提示词： "Stormtrooper's lecture in beautiful lecture hall";

- Midas
- DPT

Stable Diffusion Model 应用： Normal Map法线贴图图像生成

根据图片生成法线贴图，便于AI给图片内容进行更好的光影处理，它比深度模型对于细节的保留更加的精确。

对应ControlNet模型： control_normal。

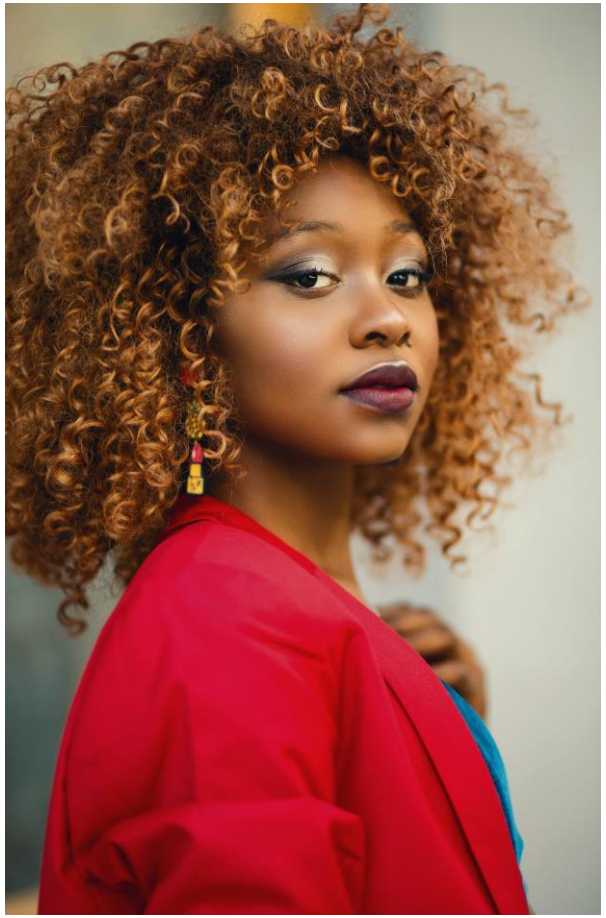
提示词： "A head full of roses";



Stable Diffusion Model 应用： Lineart图像生成

Lineart 边缘检测预处理器可很好识别出图像内各对象的边缘轮廓，用于生成线稿。
对应ControlNet模型： control_lineart。

提示词： "michael jackson concert";



Stable Diffusion Model 应用： Content Shuffle 图像生成

Content Shuffle 图片内容变换位置，打乱次序，配合模型 control_v11e_sd15_shuffle 使用。

对应ControlNet模型： control_shuffle。

提示词： "New York"


Thank
You