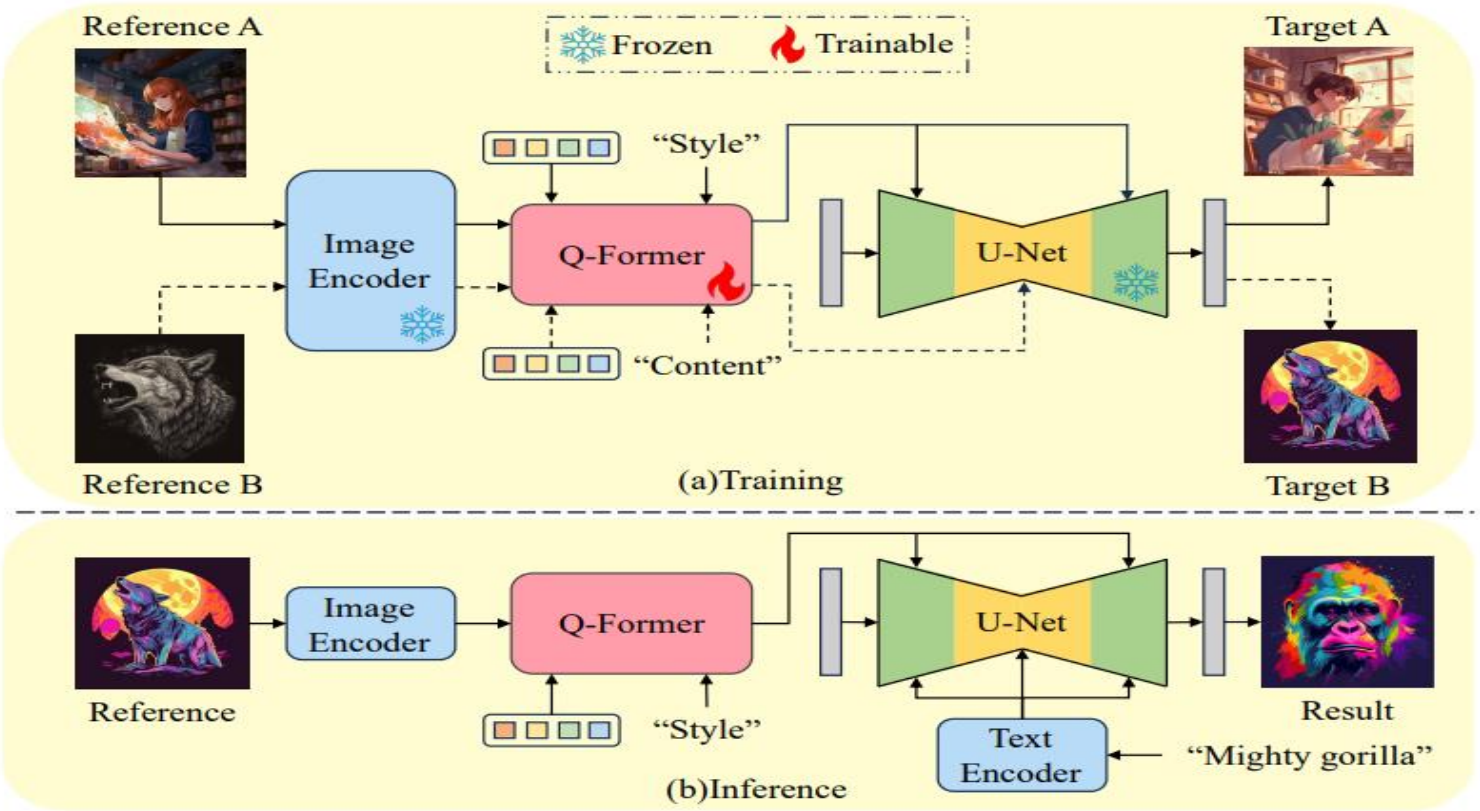
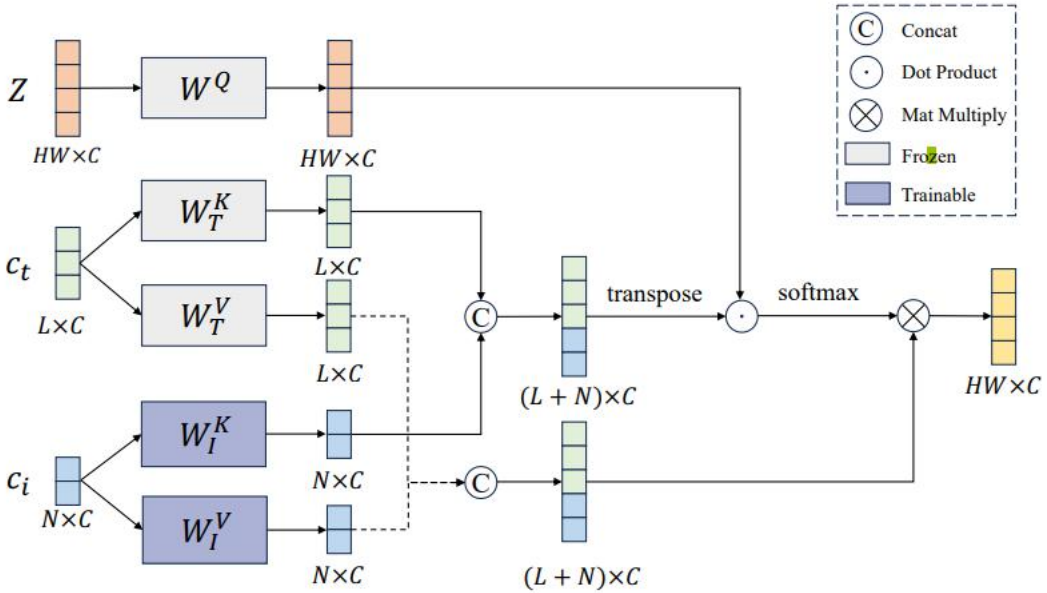
DEADiffusion

- 解决的问题：现有SD模型语义理解能力差，难以理解较为复杂的prompt。
- 主要创新点：
  - 1.双解耦特征提取：image_after_encode+Q-former’s learnable query tokens+text(style or content)



- 2.解纠缠的条件反射机制:跨注意层的不同部分分别负责图像样式/语义表示的注入
  - 原理：解纠缠调节机制(Disentangled Conditioning Mechanism，DCM)
  - DCM对空间分辨率较低的粗层进行语义约束，对空间分辨率较高的细层进行风格约束。如图所示，只将带有 “style”条件的Q-Former的输出查询注入到精细层
  - 图像-文本交叉注意力层：



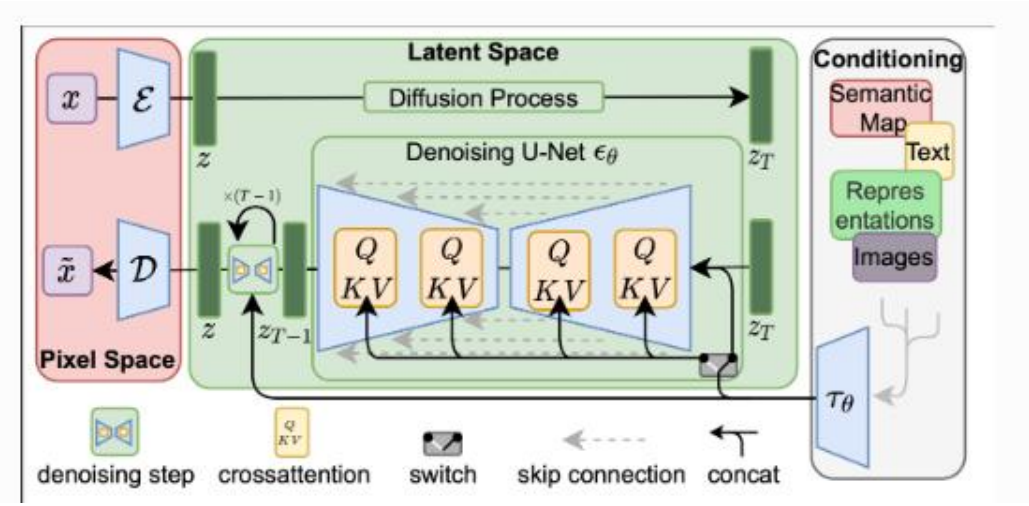
$$Q = ZW^Q, \tag{2}$$

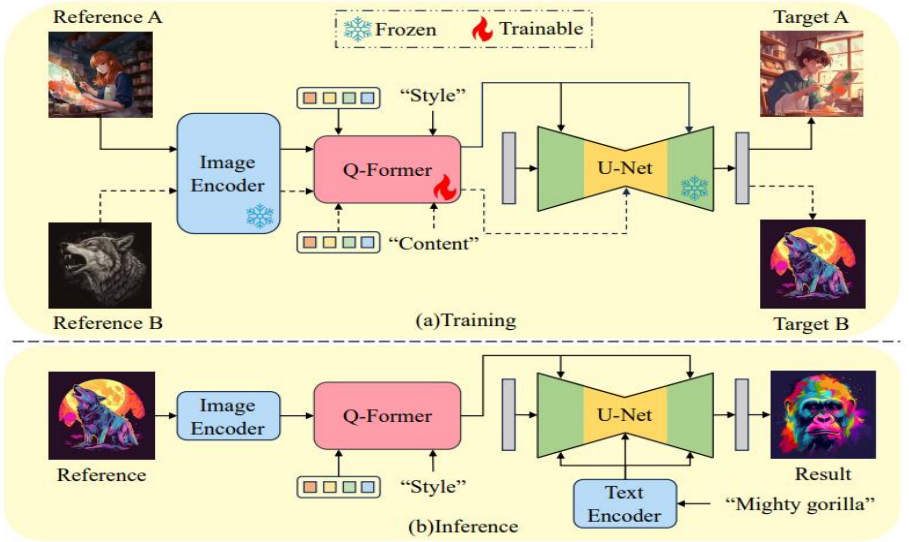
$$K = Concat(c_t W_T^K, c_i W_I^K), \tag{3}$$

$$V = Concat(c_t W_T^V, c_i W_I^V), \tag{4}$$

$$Z^{new} = Softmax(\frac{QK^T}{\sqrt{d}})V. \tag{5}$$

- 3.使用非重构训练范式：构建了两个成对的数据集





文章作者认为：传统的SD模型，模型会倾向于还原风格而忽略语义信息。所以采用了成对数据集，而且reference和target只保持语义或风格的相似性

现有问题和挑战：

1.训练代码未开源：需要我们复现训练内容（难）

但是或许可以参考BCLIP-Diffusion

2.源代码基于stablediffusion v1.5

3.数据集的搭建

FCDiffusion（Frequency-Controlled Diffusion）

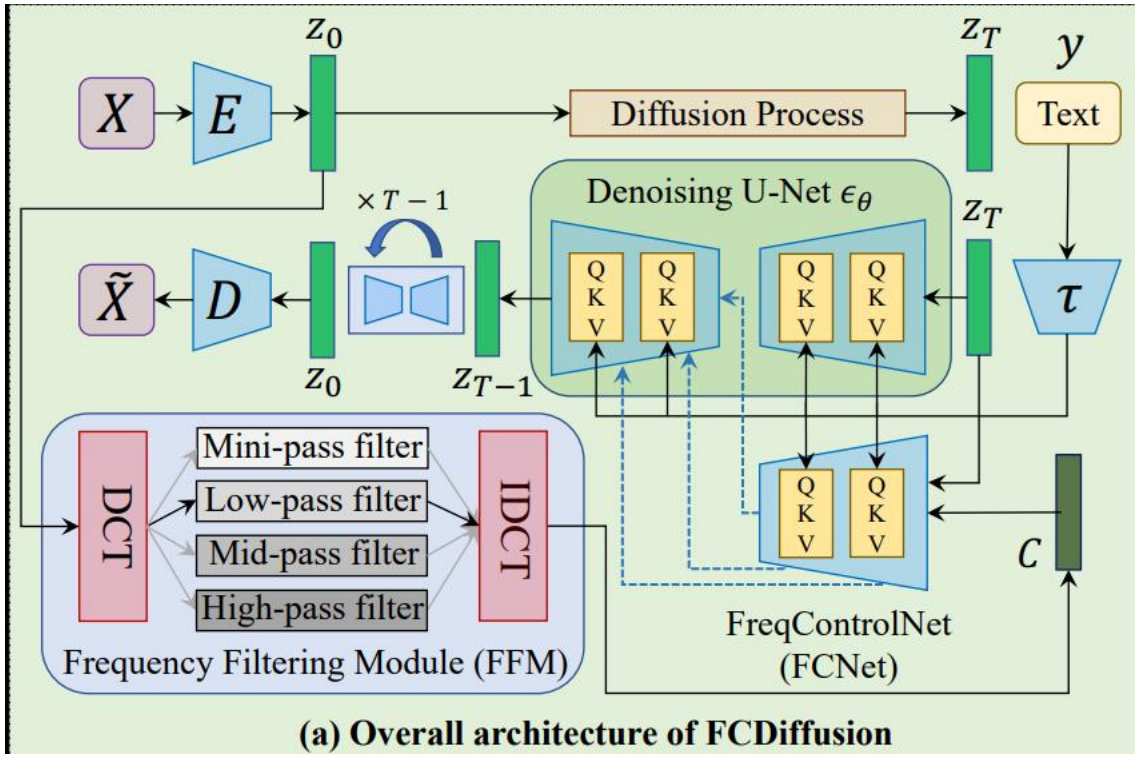
构成：Latent diffusion+FFM+FCNet

LDM：Stable Diffusion v2-1-baseFFM

FFM：核心是DCT（离散余弦变换）

应用不同的DCT滤波器构建相应的控制信号

FCNet：以当前的去噪结果、控制信号和文本嵌入作为输入，并输出多尺度特征图，引导预训练的 LDM 重建。借鉴 ControlNet (Zhang and Agrawala 2023)



- FFM：
  - 1、频域：以频率为横坐标，振幅为纵坐标的坐标轴
  - 2、离散余弦变换（DCT）：是对离散信号列做fourier变换，其变换核为余弦函数，它的变换阵的基向量能很好地描述人类语音信号和图像信号的相关特征。因此，在对语音信号、图像信号的变换中，DCT变换被认为是一种准最佳变换。

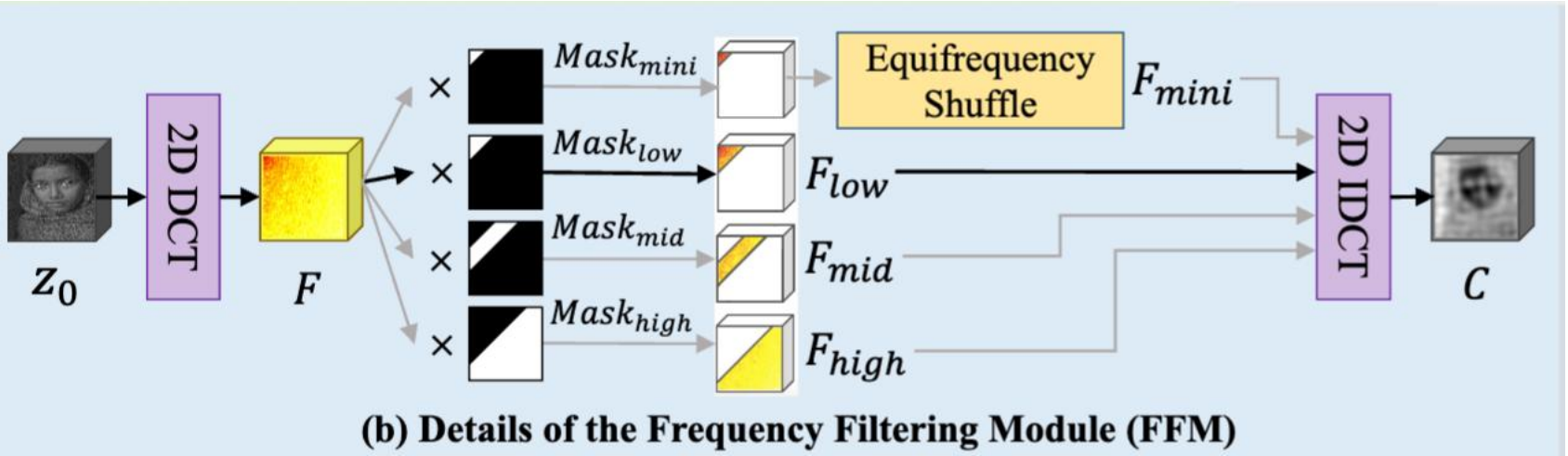
$$g(x,u)=C(u)\sqrt{\frac{2}{N}}\cos\frac{(2x+1)u\pi}{2N} \qquad C(u)=\begin{cases}\frac{1}{\sqrt{2}} & u=0\\ 1 & \text{其他}\end{cases}$$

$$F(u)=C(u)\sqrt{\frac{2}{N}}\sum_{x=0}^{N-1}f(x)\cos\frac{(2x+1)u\pi}{2N}\quad (u,\ x=0,\ 1,\ 2,\ \ldots,\ N-1)$$

$$g(x,y,u,v)=\frac{2}{\sqrt{MN}}C(u)C(v)\cos\frac{(2x+1)u\pi}{2M}\cos\frac{(2y+1)v\pi}{2N}$$

$$F(u,v)=\frac{2}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)C(u)C(v)\cos\frac{(2x+1)u\pi}{2M}\cos\frac{(2y+1)v\pi}{2N}$$

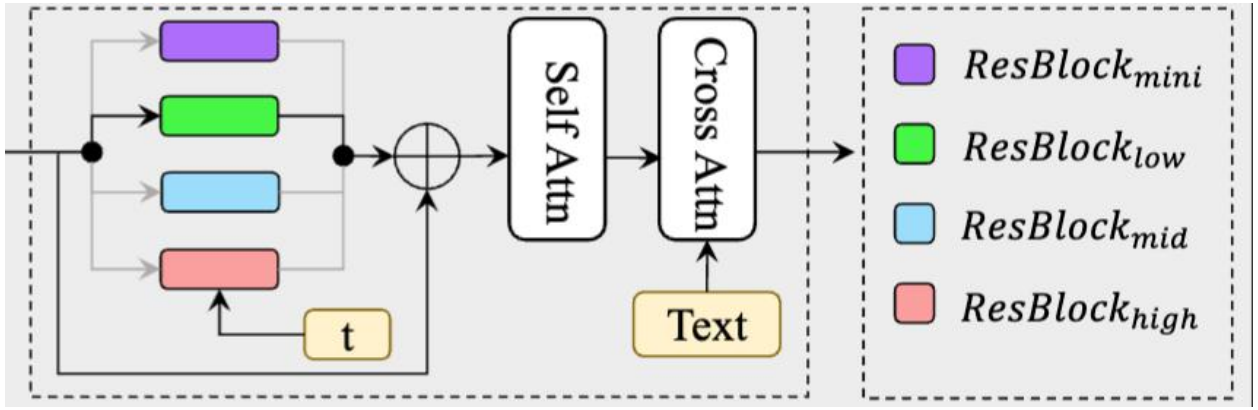
- 图片经过DCT变换之后，信息主要集中在频域图的左上角，即低频部分

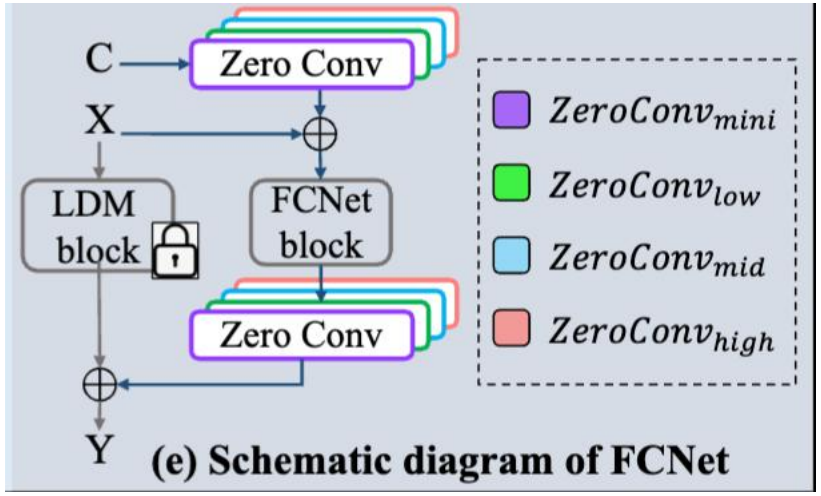


设置了不同的Mask以获取不同频率段的信息。包括微通频、低通频、中通频和高通频。
然后通过Fourier反变换（Inverse DCT）将图片还原回去。

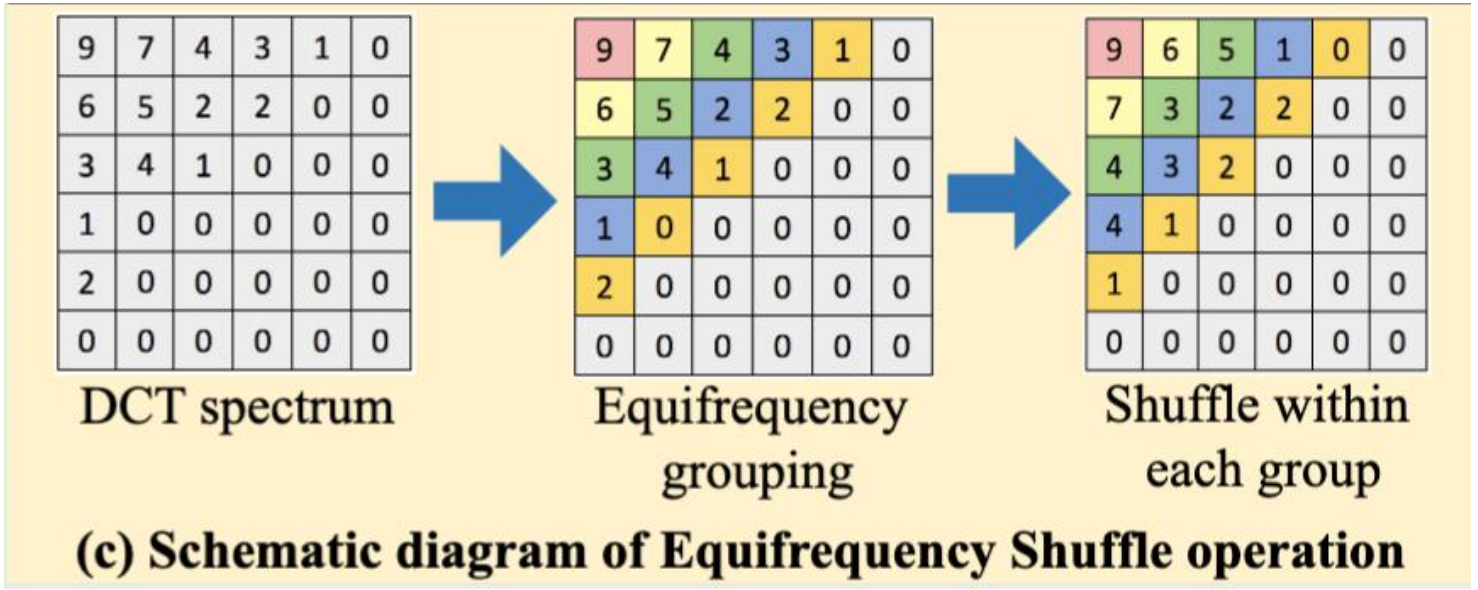
FCNet

- FCNet 是 LDM U-Net 编码器的一个可训练副本。FCNet 中的每个 U-Net 块包含一个结合了时间嵌入的 ResBlock、一个自注意力块和一个结合了文本嵌入的交叉注意力块。FCNet 中的每个 ResBlock 有四个并行副本，分别对应于 FFM 中的四个 DCT 过滤分支。使用了 ControlNet 中提出的零卷积，用于平滑地将特征注入预训练的 LDM。同样，每个零卷积也有四个并行副本，分别对应于四个 DCT 过滤分支。





为了解决微通频（mini）结构性特征无法消除，引入了等频扰动。等频扰动在不改变总体能量分布的情况下随机扰动 DCT 频谱，有助于消除源图像对生成图像的空间结构影响，同时保持 I2I 风格关联。



- 在微频率控制下，转换图像仅保留原始风格信息，没有源图像的结构约束，从而实现了基于风格的内容创建应用，即根据文本提示重新创建任何图像内容而不改变图像风格。
- 在低频率控制下，生成图像保留了源图像的风格和空间结构，适合进行源图像的小规模编辑，即图像语义操作。
- 在高频率控制下，转换图像在物体轮廓上与源图像保持一致，对风格外观的约束较少，允许根据文本提示操控图像风格，即图像风格转换。此外，还实现了图像场景转换的应用，
- 使用中频率控制来绕过源图像在低频率风格和高频率轮廓方面的约束。

- 问题：
- 1、数据集的构建（正在下载数据集……）
- 2、jittor框架的改写