# Title：

Hierarchical Text-Conditional Image Generation with CLIP Latents

使用CLIP训练好的特征，来做层级式的，依托于文本的图像生成。

64\*64->256\*256->1024\*1024

# Abstract：

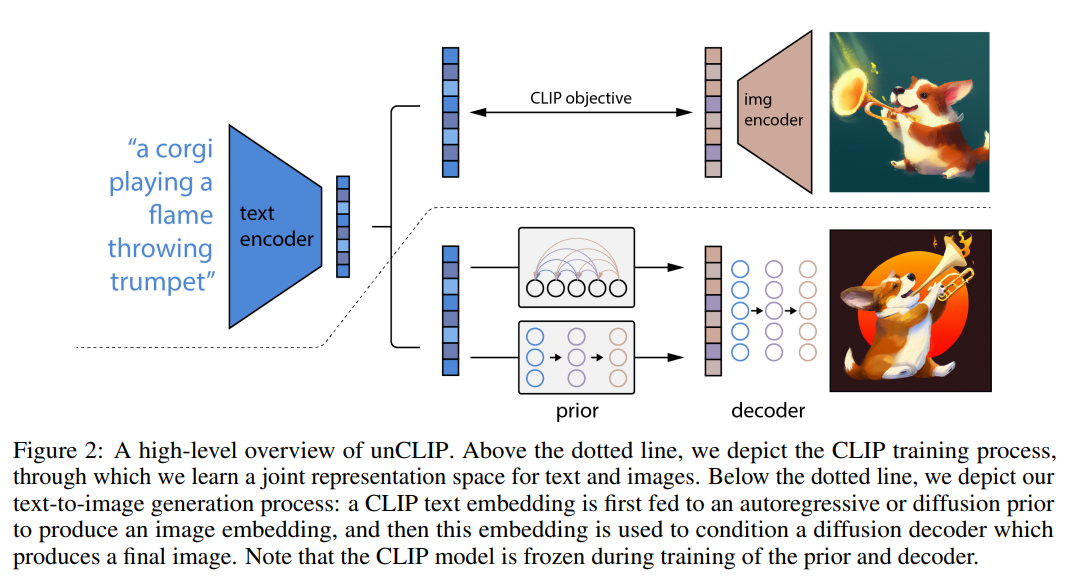
之前的对比学习的方法比如说CLIP模型，这种模型能学习到很稳健的图像特征，它既能抓住图像的语义信息也能抓住图像的风格信息，作者借助这些信息来做图像生成：提出了一个两阶段的模型，分别是prior和decoder。

prior是给定一个文本特征（CLIP生成），生成一个类似于CLIP的图像特征，decoder是给定特征后，可以根据特征生成图像（CLIP的图像特征作为ground truth）。

因为DALL·E 2是利用文本特征生成的图像，所以可以使用文本描述直接修改生成的图像。DALL·E 2 整个全是扩散模型。

# Introduction

CLIP通过最简单的对比学习方式，就能学习到很好的特征，比如他们有特别好的zero-shot能力，在CLIP上做fine-tune能在一系列任务上取得很好的成绩。Diffusion模型也是一个很好用的图像生成工具，但是保真度比不过GAN，但是最近有一系列技巧提升Diffusion的保真度，比如guidance technique，通过牺牲一部分多样性，提升扩散模型的保真度。

上半部分是CLIP，下半部分是DALL·E 2。

CLIP模型：文本通过文本编码器得到一个文本特征，图像通过一个图像编码器得到一个图像特征，它们俩组成了一个正样本对，其他的组合成为负样本。通过这种方式做对比学习最后把图像编码器和文本编码器都学的很好，而且文本和图像的特征都联系到一起，是一个合并的、多模态的特征空间。在DALL·E 2模型里，CLIP模型是锁住的，不会经过任何fine-tune。

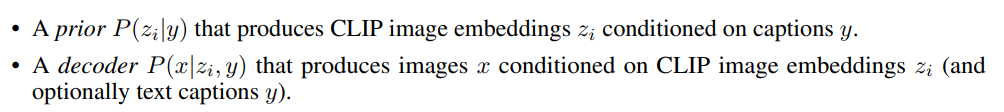
接下来就是两阶段的训练方式：

如果先从文本生成文本特征，再依据这个文本特征生成图像特征，然后再由这个图像特征生成图像，存在这一显式的生成方式效果好很多。

在训练时，先给定一个文本，通过给定锁住的文本编码器得到一个文本特征，然后想通过prior得到一个图像特征，通过CLIP，可以拿到正样本对的图像特征，可以作为监督（通过文本特征预测ground truth），这样就可以把prior模型训练出来。接下来就是使用常规的Diffusion模型作为解码器生成图像。

DALL·E 2把自己叫做unCLIP，是从文本到图像的反过程。

# Method

训练数据集里面是一种图像文本对（x，y），给定图片x，用zi和zt表示CLIP出来的图像和文本特征，整个DALL·E 2的网络分成了两部分，一个是prior，一个是decoder

prior表示根据文本y去生成图像zi，decoder的输入是zi（有时会带上文本y），把图像恢复成x，这样就完成文本到图像的过程。

为什么能用两阶段的方式来实现这个过程呢：

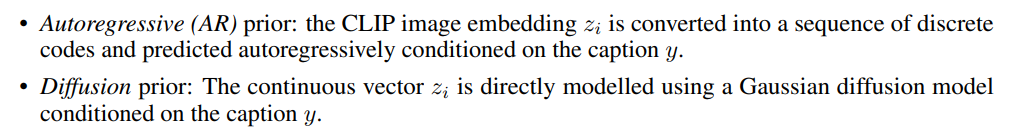
给定一个文本，恢复成x，可以写成以上概率的形式：给定一个文本，生成x和zi，zi和x是一对一的关系。然后依据chain rule，可以往后推导成上图形式

## Decoder

decoder是GLIDE模型的一个变体，首先使用CLIP模型的guidance，然后使用了classifier-free guidance，他们的guidance信号要么来自于CLIP要么来自于文本。

然后使用级联式的生成：图像分辨率64\*64->256\*256->1024\*1024，使用的Diffushion模型就是CNN，没有用Transformer，所以做推理的时候可以用在任何的尺寸上。

## Prior

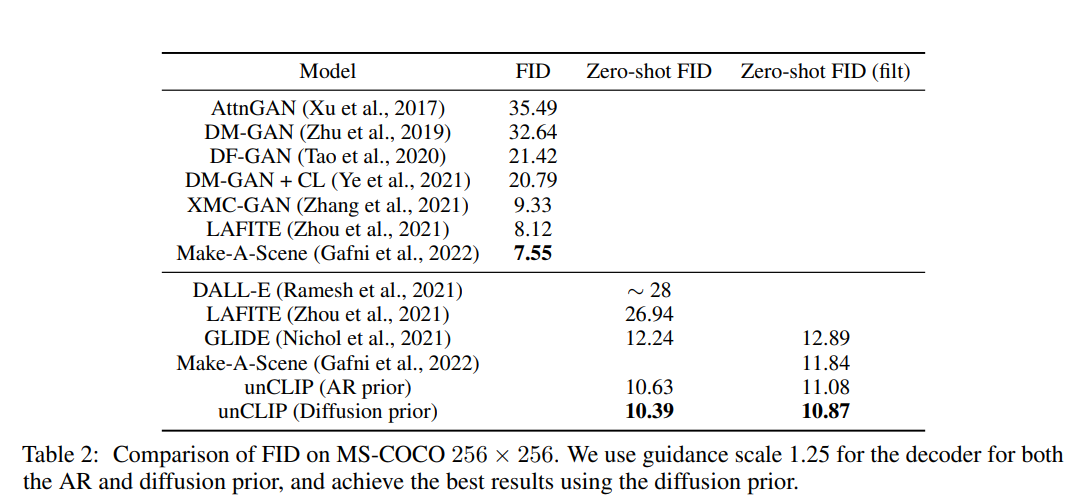
prior模型的作用是给定一个文本y，怎么去生成一个图像特征zi，作者尝试了两种方案，要么是自回归，要么是扩散模型

为了训练更快速，效果更好，作者使用了Diffusion模型。

无论是AR还是Diffusion，都使用了classifier-free guidance。

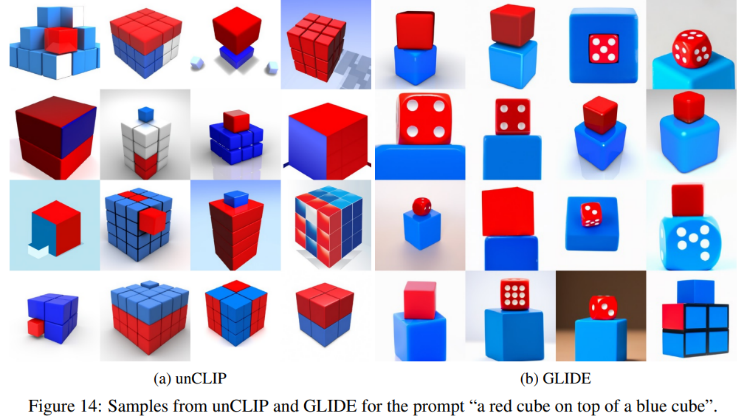
这里的diffusion，训练了一个Transformer的decoder（因为输入输出是embedding）。最终的embedding拿去预测没有加过噪声的CLIP的图像特征。这里的训练没有像DDPM那样去预测噪声（残差），而是直接预测图像特征，在此处效果比较好。

# Comparison on MS-COCO



# Limitations and Risks

1.DALL·E 2不能很好的把物体和它的属性结合起来



比如说这里红色方块在蓝色方块上面，GLIDE做的不错，但对于DALL·E 2比较差，作者说很可能是用了CLIP模型的原因（CLIP模型注重相似性，它不了解上下左右，是或者不是，所以用CLIP模型做下游任务的时候，就不能很好的区分物体和物体的属性）



2.DALL·E 2还是不能生成一些复杂的场景，一些细节生成不好：



没有完全体现出“湖”“玩”的要素

具体的细节是随机生成的一些颜色块，没有任何语义信息