Method:
CLIP feature+ 2 stage (prior, decoder)

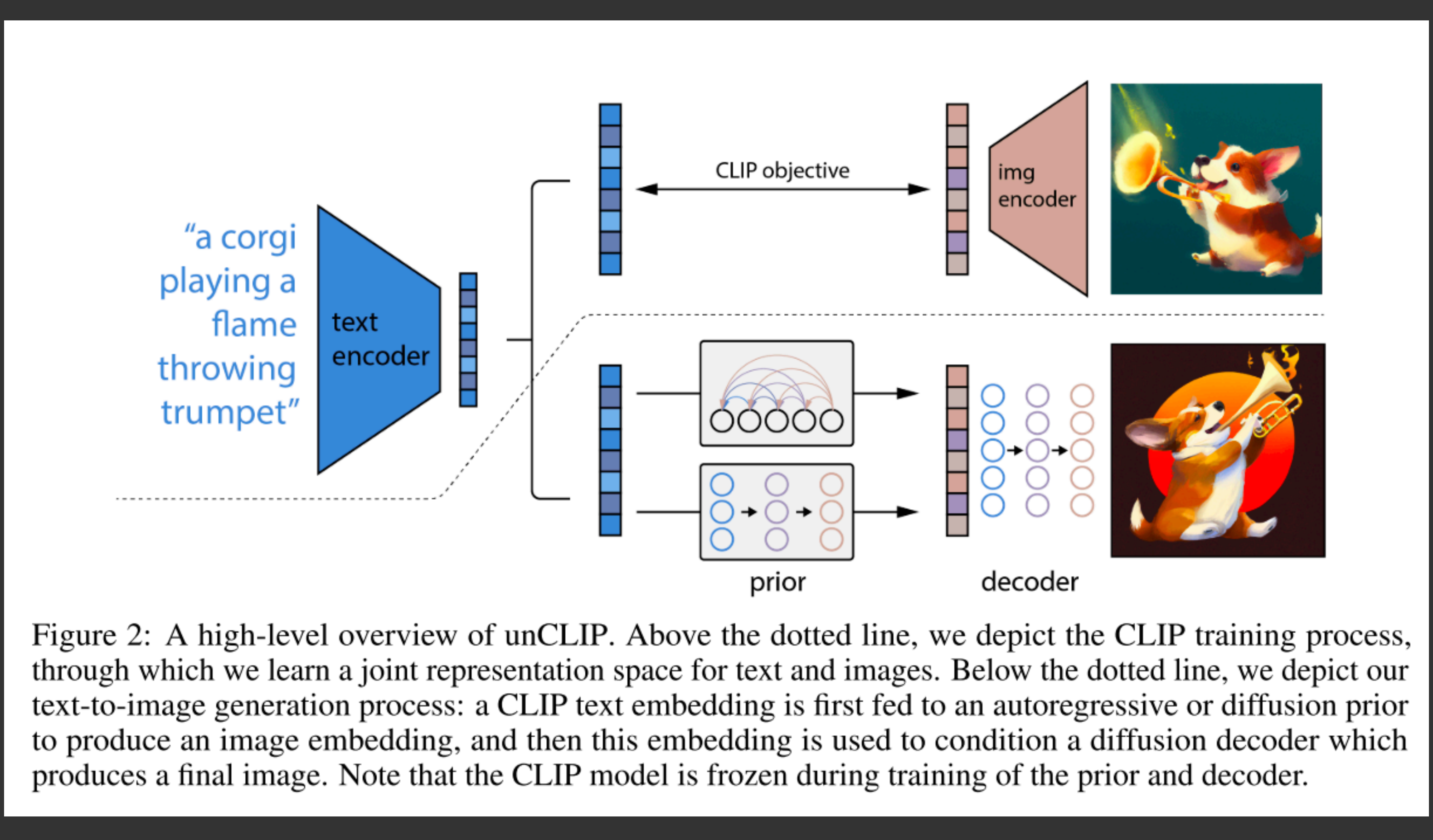
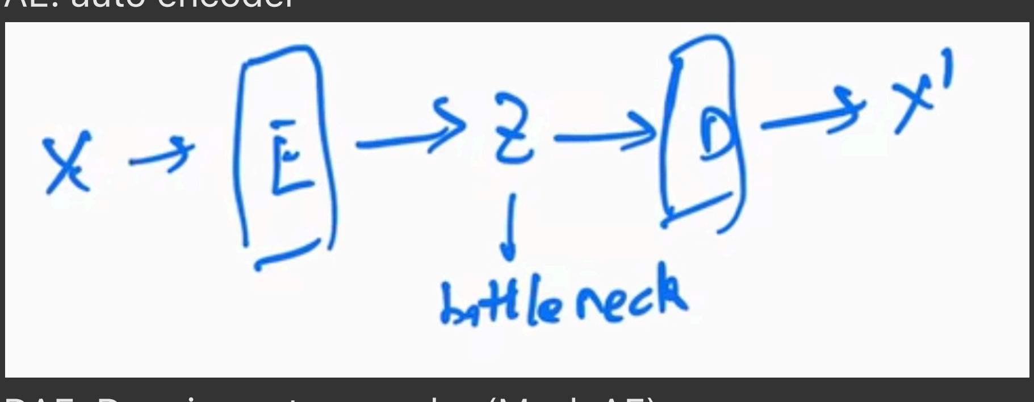


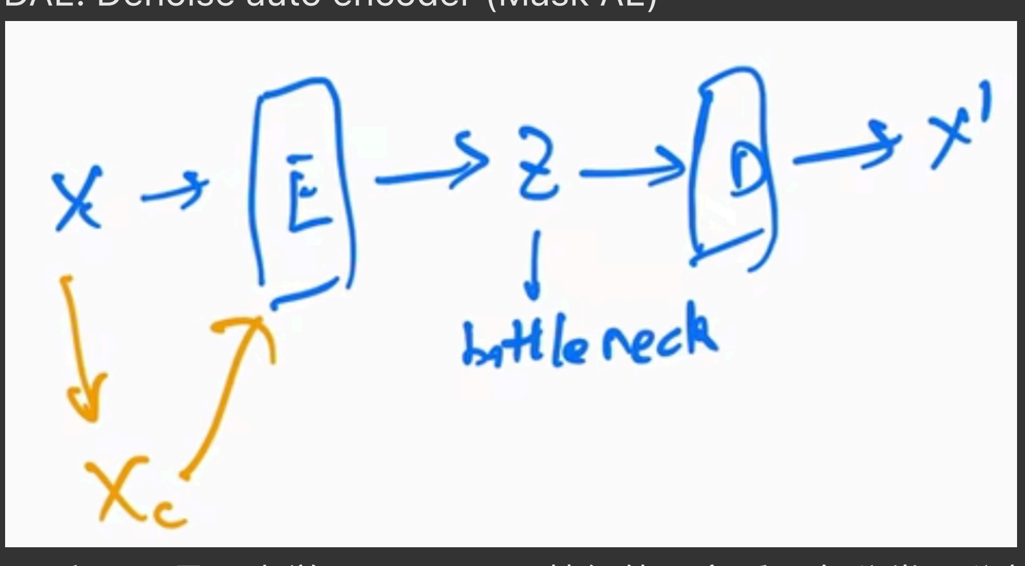
Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

CLIP + GLIDE + many useful training skills

AE: auto encoder

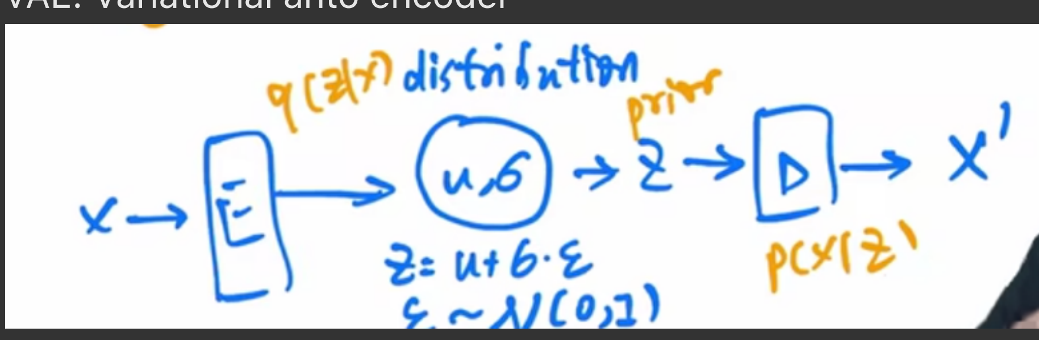


DAE: Denoise auto encoder (Mask AE)



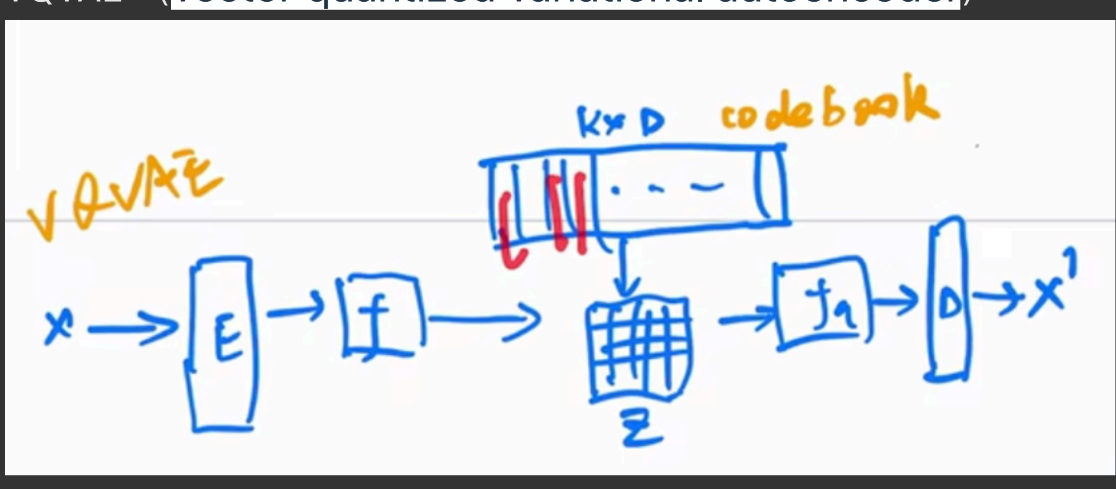
AE 和DAE是用来学习bottleneck特征的，之后服务分类、分割、检测等等，他学到的不是分布，无法进行采样。主要是用于重建的。

VAE: Variational anto encoder



VAE预测的是分布。
q（z/x）是后验概率
学习的distribution是先验分布。
p（x/z）是likelyhood，整个过程是最大化似然函数。

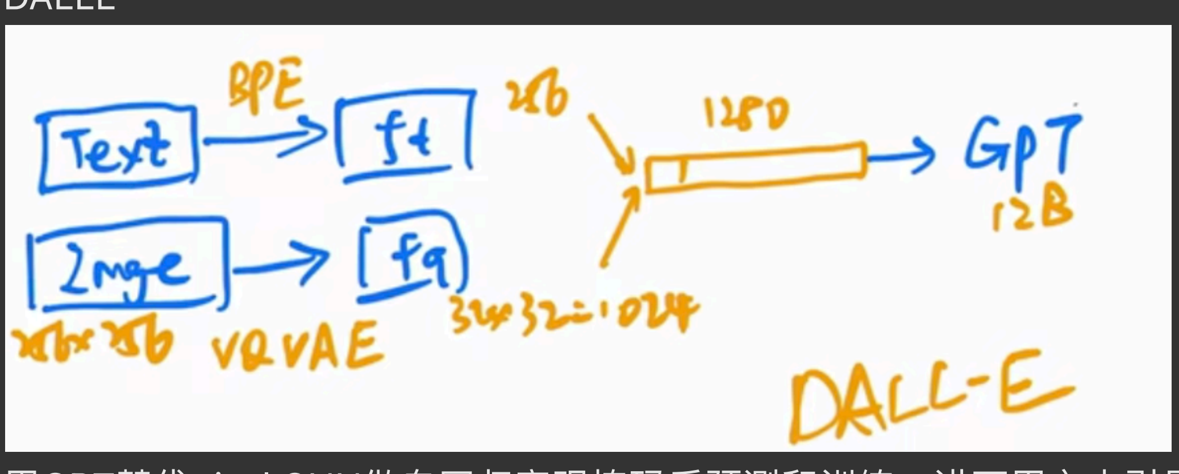
VQVAE （vector quantized variational autoencoder）



VAE不容易把模型做大（图像重建尺度做大，分布不好学）用codebook来代替。
K一般是8192，D是512或者768
因此经过编码器后的特征可以直接从码本里面找最近邻进行量化表示，整个待学习空间十分可控。
codebook和fq是用来做highlevel的任务的（分类、检测），不具备随机生成属性。

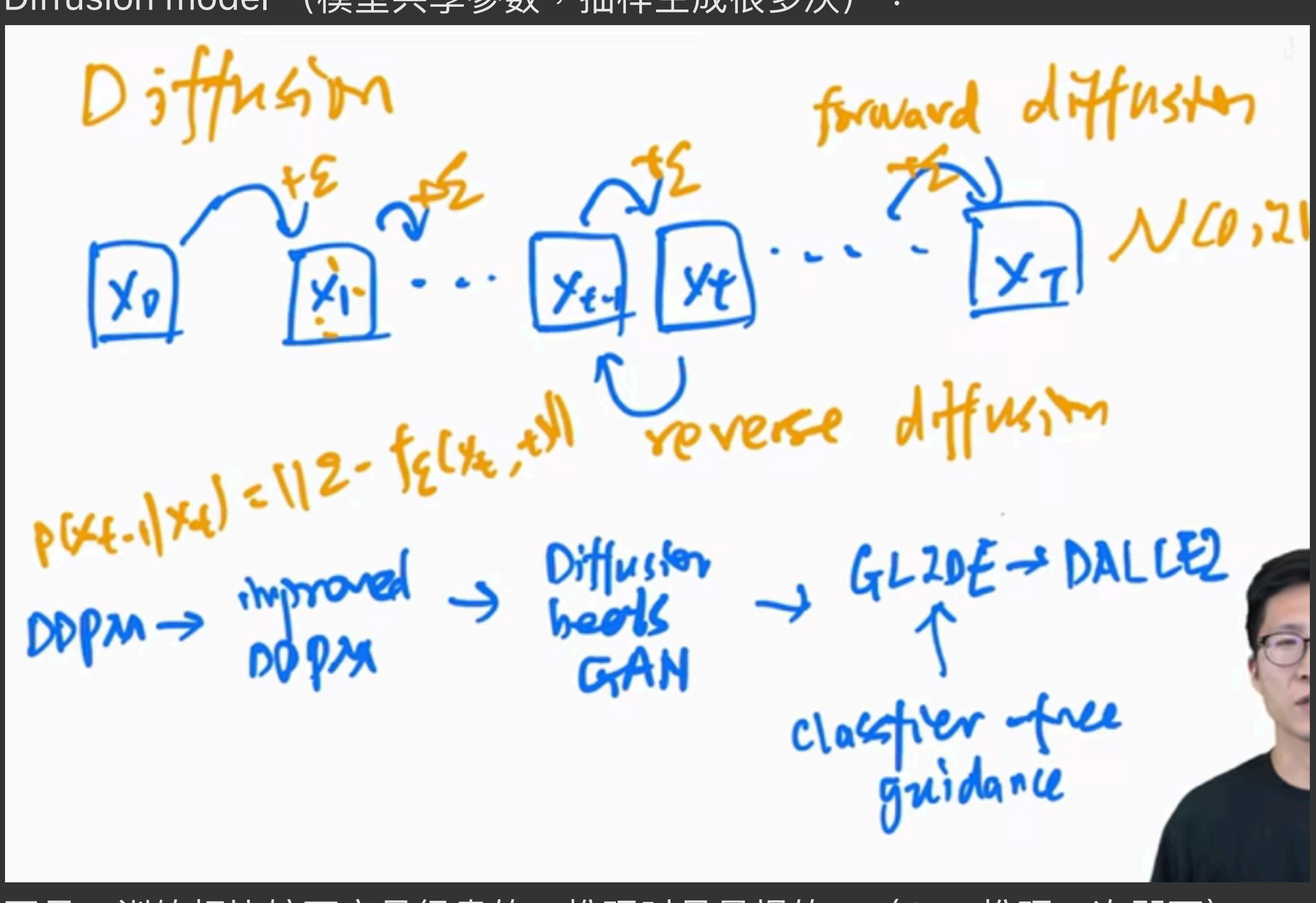
VQVAE2：
做成层级式的，不仅有局部的建模，还有全局的建模，+attention 模型表达能力变强了。根据codebook学习了一个prior（pixel CNN， 一个auto regressive模型）。从而实现了可以生成的能力。

DALLE



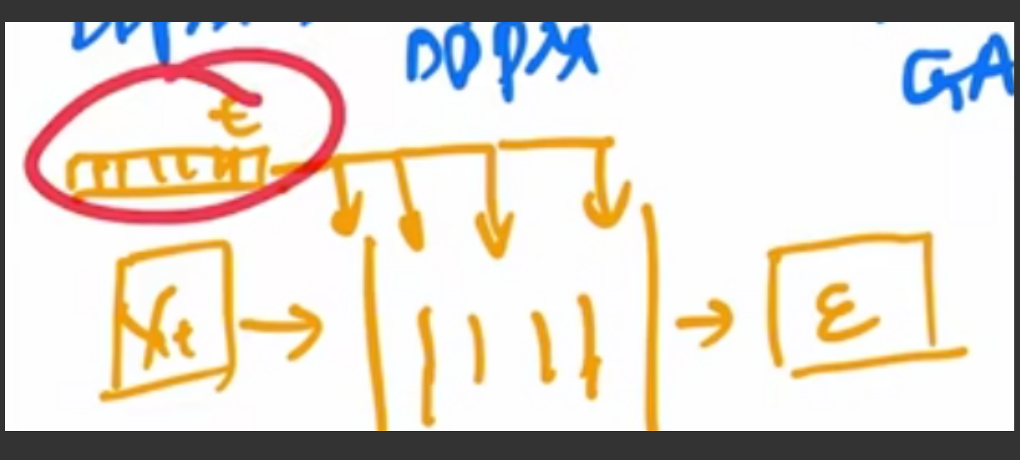
用GPT替代pixel CNN做自回归实现掩码后预测和训练，进而用文本引导图像生成。
BPE：byte pair encoder
VQVAE部分即用上述方法生成的codebook，直接用。
将文本+图像特征concate起来，给GPT，用掩码遮住做自回归，从而训练文生图。
会生成很多图像，用CLIP特征进行排位，排位最靠前的选做结果。

Diffusion model （模型共享参数，抽样生成很多次）：



不足：训练相比较而言是很贵的，推理时是最慢的。（GAN推理一次即可）

DDPM贡献
1. 预测每步叠加的噪声值（类似resnet ，预测噪声而不是xt-1。）
2. 加入了time embedding（类似 positional encoding，最开始反向传播中学习轮廓，最后再学习细节，时间编码用于生成和采样控制）
3. 只需要预测噪声均值即可。

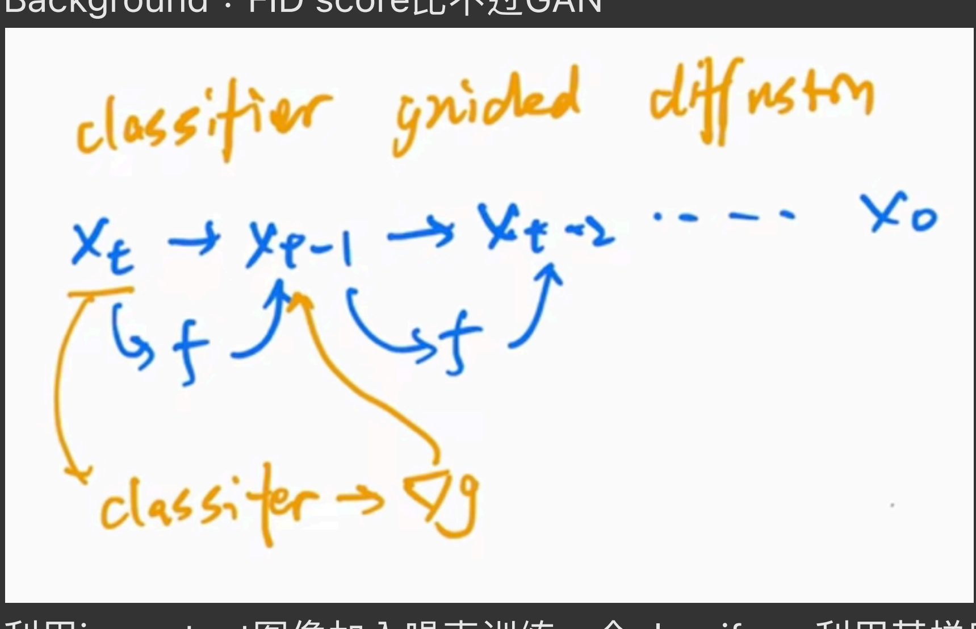


DDPM与VAE模型的不同之处：
1. 虽然都是编码器+解码器过程。但DDPM编码过程固定，每次加入固定噪声。VAE的编码器是学习获得的。
2. DDPM 刚开始和中间过程的维度都是相同大小。而AE和VAE一般情况下都是不同的。
3. Diffusion有步数的概念，要经过很多很多步才能图像生成。有time step和time embedding这些概念。所有timestep所有模型都是共享参数的。VAE没有这些。

Improved DDPM 相对于DDPM的改动：
1. 学习方差。
2. 噪声生成从线性schedule变成余弦schedule
3. 大模型，scale很好。

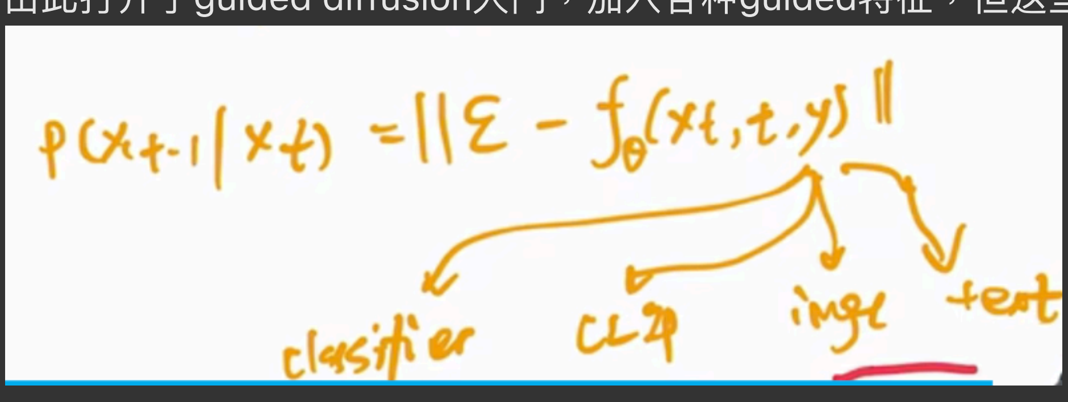
Diffusion beats GAN：
1. 大模型（加大、加宽，增加attention head数量）single scale attention head 变成multi scale attention head 又大又复≣
2. adaptive group normalization. (用步数做自适应归一化)
3. Classifier Guidance 引导模型采样和生成 25次采样即可。

关于Classifier Guidance diffusion：
Background：FID score比不过GAN

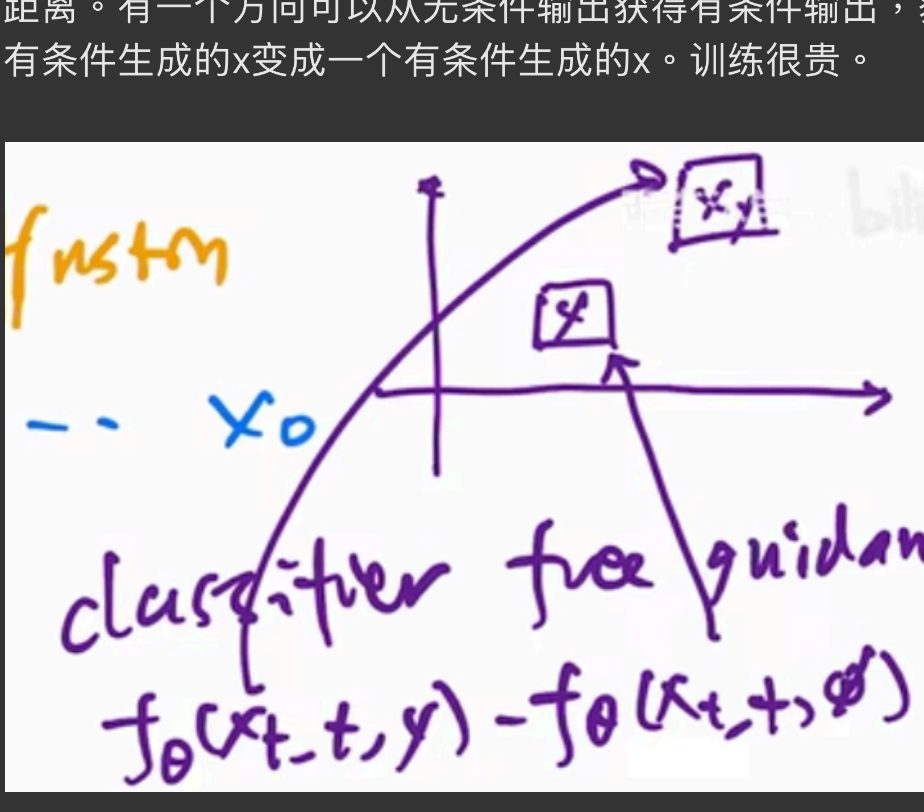


利用imagetnet图像加入噪声训练一个classifer，利用其梯度去指导xt-1的生成。使得效果更逼真。（超过BIGGAN）
牺牲多样性获得逼真感。

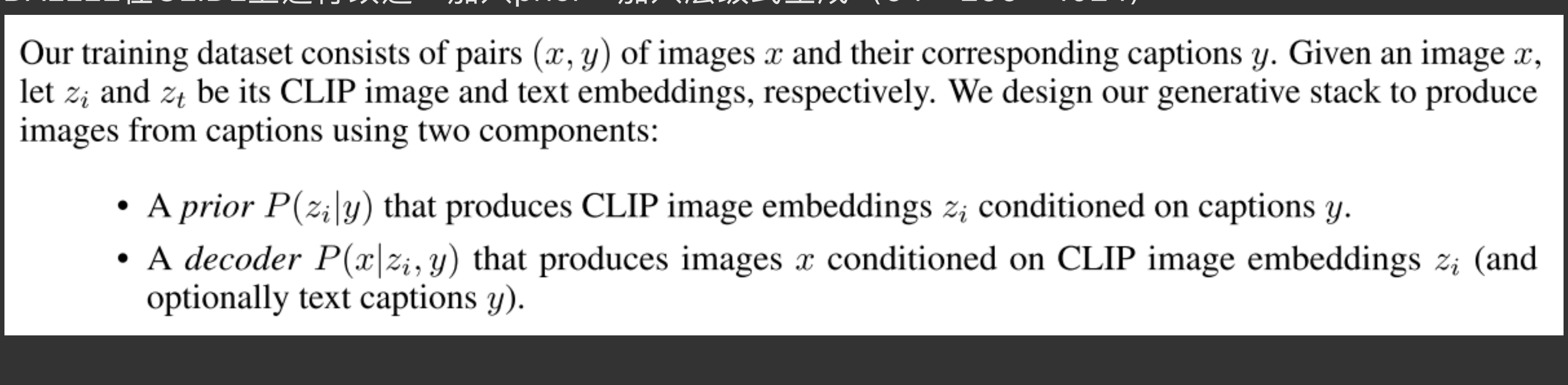
由此打开了guided diffusion大门，加入各种guided特征，但这些方法也都引入了一个另外的模型引导。



改进版本：**classifier free guidance**！！（GLIDE模型，3.5B参数效果直逼DALLE12B）训练模型的时候生成了两个输出，一个有条件的输出，一个没条件的输出。
例如用图像文本对做训练：1个是加入文本的，1个是没用用文本的。生成了两个图像，训练获得他们之间的距离。有一个方向可以从无条件输出获得有条件输出，获得了差距。这样再做生成的时候，可以获得用没有条件生成的x变成一个有条件生成的x。训练很贵。



DALLE2在GLIDE上进行改进。加入prior，加入层级式生成（64->256->1024）。

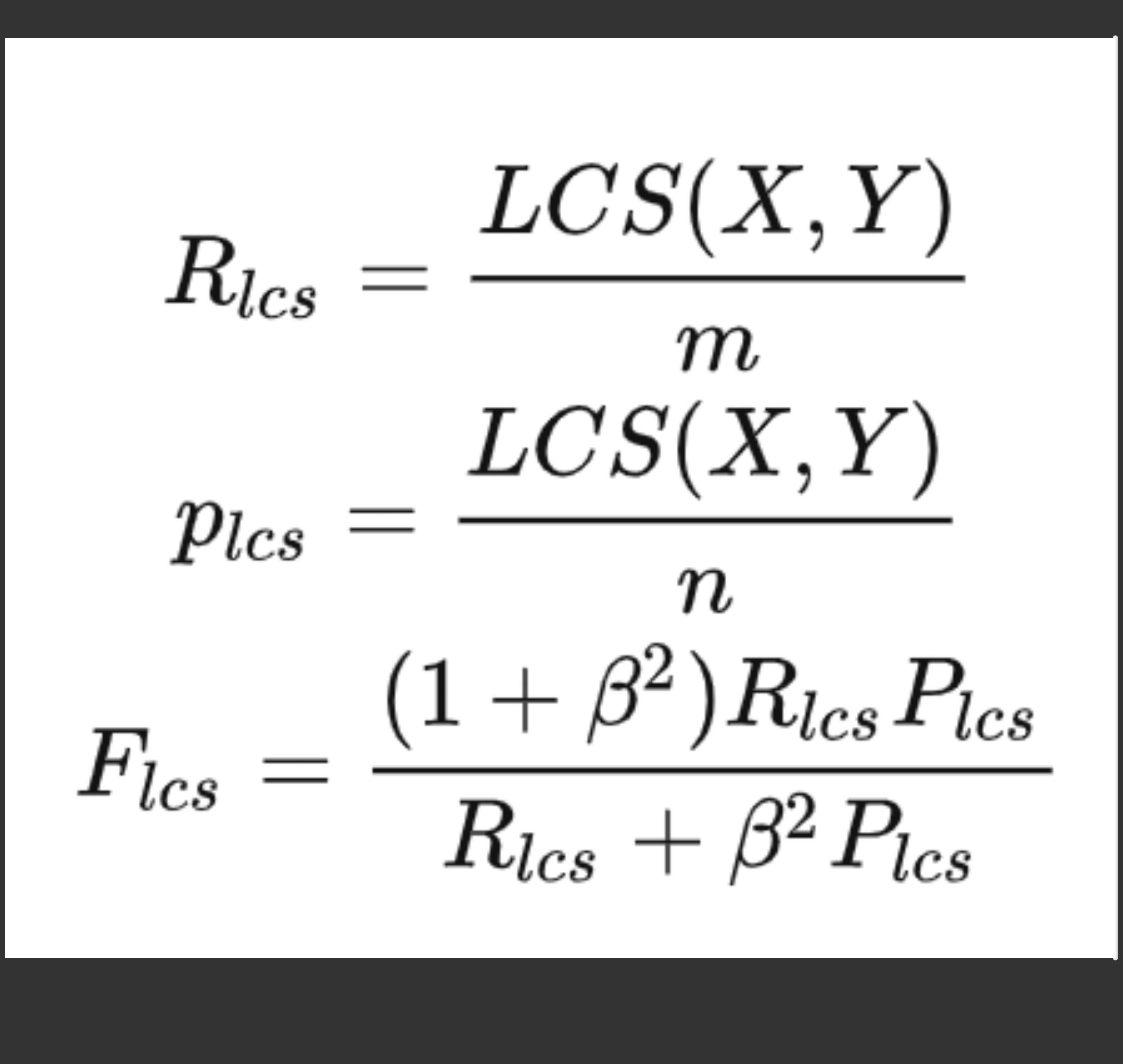


Our training dataset consists of pairs (x, y) of images x and their corresponding captions y. Given an image x, let z_i and z_t be its CLIP image and text embeddings, respectively. We design our generative stack to produce images from captions using two components:

- A *prior* $P(z_i|y)$ that produces CLIP image embeddings z_i conditioned on captions y.
- A *decoder* $P(x|z_i, y)$ that produces images x conditioned on CLIP image embeddings z_i (and optionally text captions y).

Decoder:
1. CLIP guidance + **classifier free guidance** (CLIP图像特征或者文本特征)
2. upsample model， UNET

Prior：
训练快速：
方法一：AR自回归，太慢。
方法二：diffusion model。 decoder-only Transformer（输入输出都是序列）。对于特征重建，预测特征比预测噪声效果要好。



$$R_{lcs} = \frac{LCS(X,Y)}{m}$$

$$p_{lcs} = \frac{LCS(X,Y)}{n}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$