思考：dalle2的prior和decoder应该怎么训练。

Prior输入的是text embedding, caption输出img embedding。

采用的是diffusion结构，实现上用的是transformer，探究transformer是怎么用到这里面的 以及中间到底发生了什么变化。

Transformer到底是用来干什么的：注意力和权重？

Diffusion的正向传播是一步步加噪声，把图片转换成正态分布的噪声。

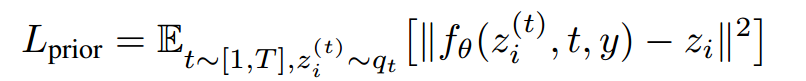
反向传播是给出带噪声的图片（还是纯噪声？）恢复出原来的图片。

训练的时候训练的是反向传播吗？

Loss:

For the diffusion prior, we train a decoder-only Transformer with **a causal attention mask** on a sequence consisting of, in order: **the encoded text**, **the CLIP text embedding**, an embedding for the **diffusion timestep**, the **noised CLIP image embedding**, and a final embedding whose output from the Transformer is used to predict the unnoised CLIP image embedding. we improve quality during sampling time by generating two samples of zi and selecting the one with a higher dot product with zt.

we find it better to train our model to predict the unnoised zi directly, and use a mean-squared error loss on this prediction:



是一个decoder-only transformer，输入是encoded text + CLIP text embedding + noised CLIP Image [embedding](https://www.zhihu.com/search?q=embedding&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra={"sourceType":"answer","sourceId":"2516184924"}" \t "https://www.zhihu.com/_blank) + 额外的token（类似class embedding），输出一个unnoised CLIP Image embedding。

embed和encode: