# Stable Diffusion.

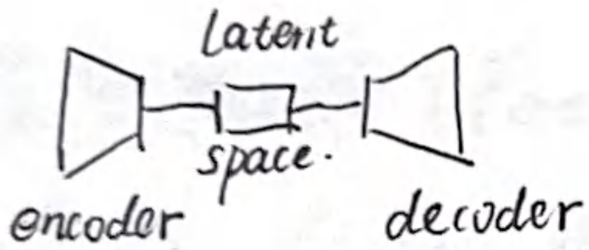
1. 使用了 CLIP (Contrastive Language-Image Pre-training) 中的 Text Encoder

2. Variational AutoEncoder.

(1) 用于压缩前/逆向过程的维度以提高计算效率
↳ latent diffusion.

→ 将学习 data distribution ⇒ latent representation of the data.

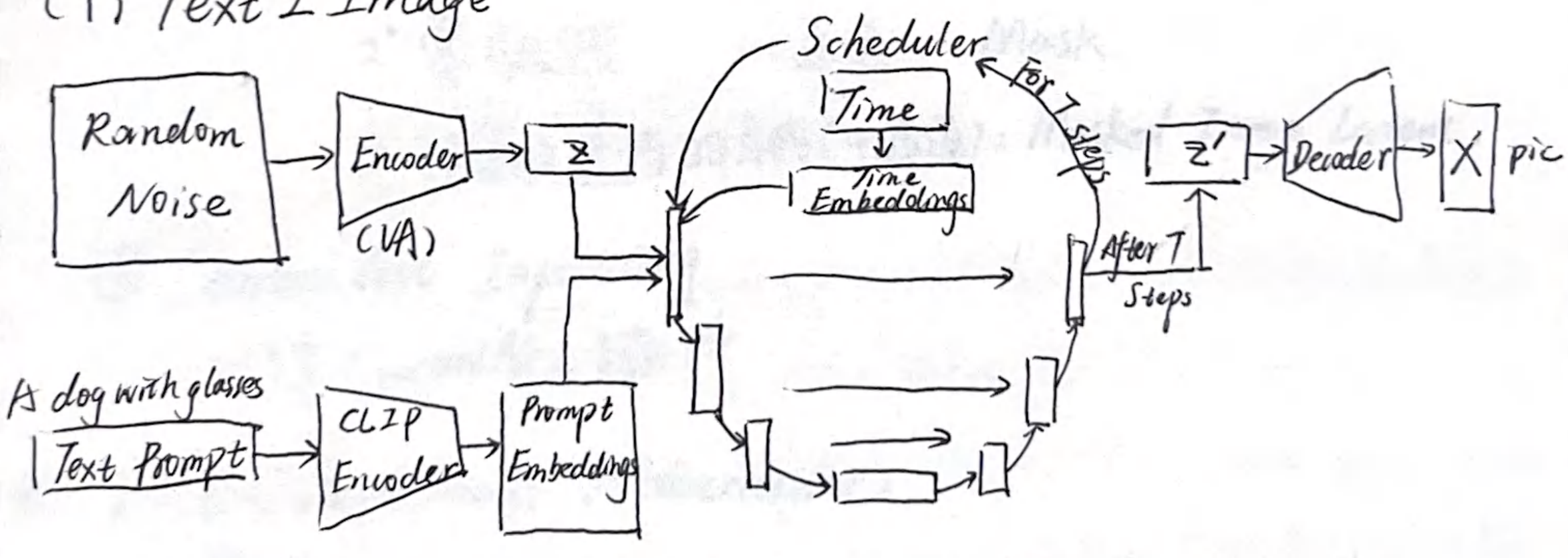
(2) 不压缩，同时也学习了一个 latent space.




↖ represents the parameters of a multivariate distribution (多元的)
↑
(Gaussion) 学的是 μ 和 σ.

3. Architecture.

(1) Text 2 Image




(2) Image 2 Image

将 Random Noise 换成真实图片：encoder 得到的z是后，Add noise to latent.

- 添加的噪声少，生成的自由度少
- 少的噪声 means 不能从根本上 radically 改变图片.
- 噪声多少 ⇔ 对原图有多大注意力

(3) In-Painting. (fool the model) ⟹ 引入约束

① 推理时干预 (Inference-time Strategy)

• 代表作: RePaint

• 思想: 不重新训练模型, 在去噪过程作弊.

$$X_{t-1}^{final} = Mask \cdot X_{t-1}^{pred} + (1-Mask) \cdot X_{t-1}^{Known}$$

$X_{t-1}^{pred}$: 被遮住的部分, 使用模型推理结果

$X_{t-1}^{Known}$: 保留区域: 不是直接将原图 $X_0$ 贴上, 而是前向到 $t-1$ 强度的噪声.

• 问题: 边缘不连贯.

⇒ 生成 $X_{t-1}$ 后, 强行加噪回 $X_t$, 反复横跳. (计算量大)

② 模型微调 (Fine-tuning Strategy)

• 代表作: Stable Diffusion Inpainting Model.

• 思想: 将输入通道 4 → 9 ⇒ UNet.

1° 当前噪声图 (4通道): $Z_t$

2° 蒙版图 (1通道): Mask

3° 被蒙版遮挡的原图 (4通道): Masked Image Latent.

③ ControlNet Inpainting

训练 ControlNet 插件.

4. VAE (Variational Autoencoder)

(1) • 原理: Not learning how to compress the data but learning a latent space which are the parameters of a multivariate Gaussian distribution.

→ Actually, the VAE is trained to learning the $\mu$ and $\sigma$. (⇓ $\log\sigma$.)

→ Then we will sample the distribution.

## 5. 核心参数

① Strength：重绘程度

- 决定了生成图和原图有多像　图片还未完全变成噪声
- 代码逻辑：起始噪声水平（从中间步开始降噪）

② do_cfg = True：启用分类器自由引导.

- 是否听懂人话
- 代码逻辑：如果为True，模型的每步预测会同时计算2个结果：
  1° 有提示词预测（output_cond）
  2° 无 ～　（output_uncond）：完全盲猜，根据负向prompt或空内容去预测.
  3° 将2种output结合
- 如果不开启，生图虽然自然，但不会理会prompt.

③ cfg_scale：引导系数

- 代码逻辑：最终结果 = 无引导结果 + 系数（有引导 - 无引导）
- < 2：不理会prompt
- > 15：强行匹配每一个词.

## 6. 代码小思考

(1) 调度策略：原DDPM用的Linear Schedule，SD中用的先开方 呈做插值再平方回去的 Scaled Linear Schedule. why?

Ans. DDPM是在像素空间训练的，SD是在Latent space上训练的，对于方差和分布更敏感，采用SLS，让β在初期增大得慢一些，适合Latent Space的信噪比变化规律.

(2) 项目中采用的是Pespaced DDPM，而非DDIM？

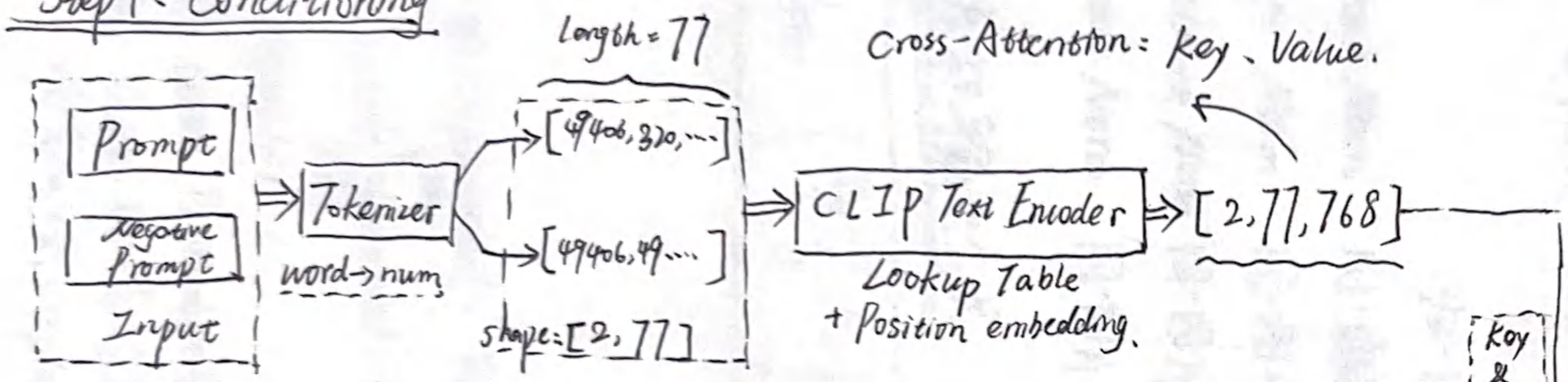
Ans. 因为每步噪声采样的variance ≠ 0，如果为0的话才是DDIM.

(3) VAE中添加的噪声和 diffusion 前向加的噪声的区别？

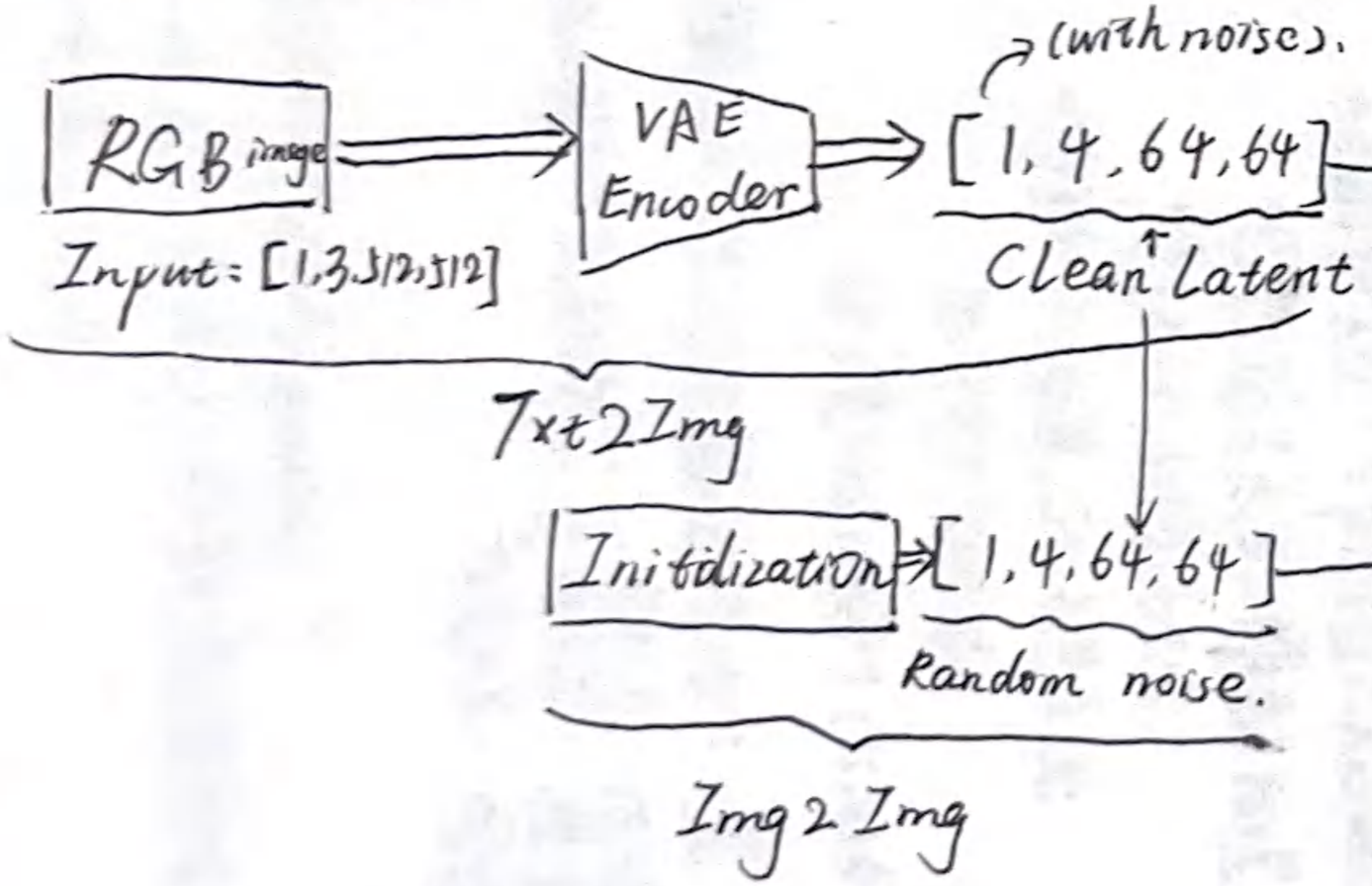
Ans. VAE把图片压缩后，得到的是一个概率分布 $(\mu, \sigma)$，

需要从中取出一个具体的Latent向量z为后面的流程用，
因此需要重参数化采样：Latent = Mean + Variance × Noise
没有这个噪音，VAE无法训练退化为普通的编码器。

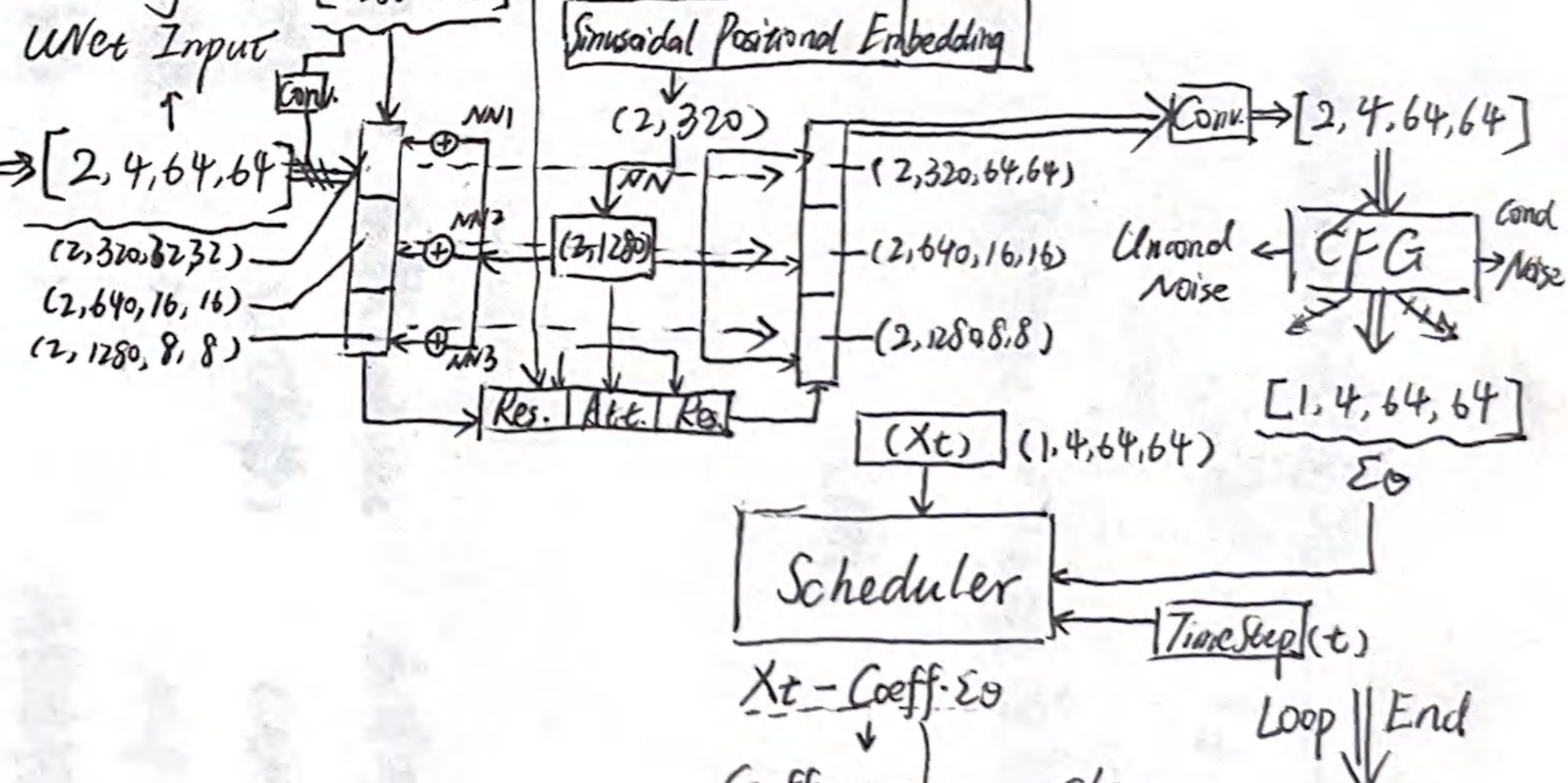
## Step 1: Conditioning



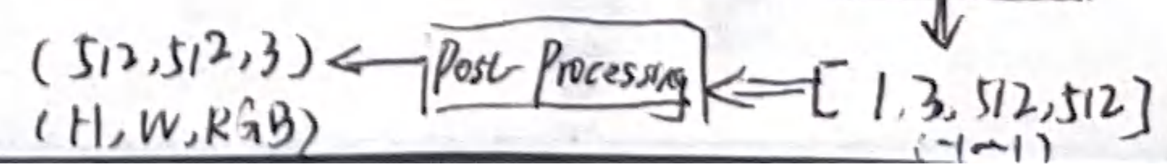

## Step 2: Latent Preparation




## Step 3: Denoising

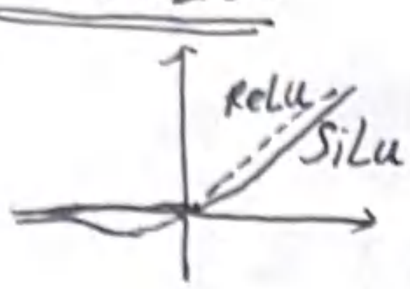



## Step 4: Decoding



# 代码知识：

## 1. SiLu




$SiLU(x) = X \cdot \sigma(x) = \frac{X}{1+e^{-x}}$

优势：①0点可导

②负值区梯度有传递，防止神经元死亡。

③隐式正则化

## 2. Normalization

$$y = \frac{x - E[x]}{\sqrt{Var[x]+\varepsilon}} * \gamma + \beta$$ （不改变形状）

标准化 → 仿射变换（$\gamma$、$\beta$可学习参数）

[图]

| | 归一化范围 | 依赖batch_size | 应用场景 | 比喻 |
|---|---|---|---|---|
| Batch Norm | 同一通道，所有照片 | √（劣势） | CNN | 全校单科排名 |
| Layer Norm | 同一图片，所有通道 | × | NLP | 个人综合总分 |
| Instance Norm | 同一图片，单一通道 | × | 风格迁移GAN | 个人单科 |
| Group Norm | 同一图片，一组通道 | × | VAE, Stable Diffusion | 个人理/文科 |

## 3. 卷积输出公式

$$Hout = \left[\frac{Hin + P - K}{S}\right] + 1$$

## "1×1卷积"

- 物理含义：作用在Channel维度上的全连接层。把每个像素点的特征向量做一次矩阵乘法，让特征间相互融合。
- 作用：①通道间信息交流
  ②升维与降维
  ③增加非线性（后面加激活函数）。

## 4. nn.Sequential：只能一种输入x，然后把输出传给下一层。

↓继承

- SwitchSequential：处理UNet每层不同的需求 ← 卷积层：图像x；残差层：图像x+时间t；注意力层：图像x+文本context。

↓包装

- ModuleList ：文件夹

▷ Why UNet用ModuleList包裹许多SwitchSequential？

Ans. UNet的编码器（Downloder）每经过一层都需保存当前特图，以便后面解码器（Uploder）的每一层concate。