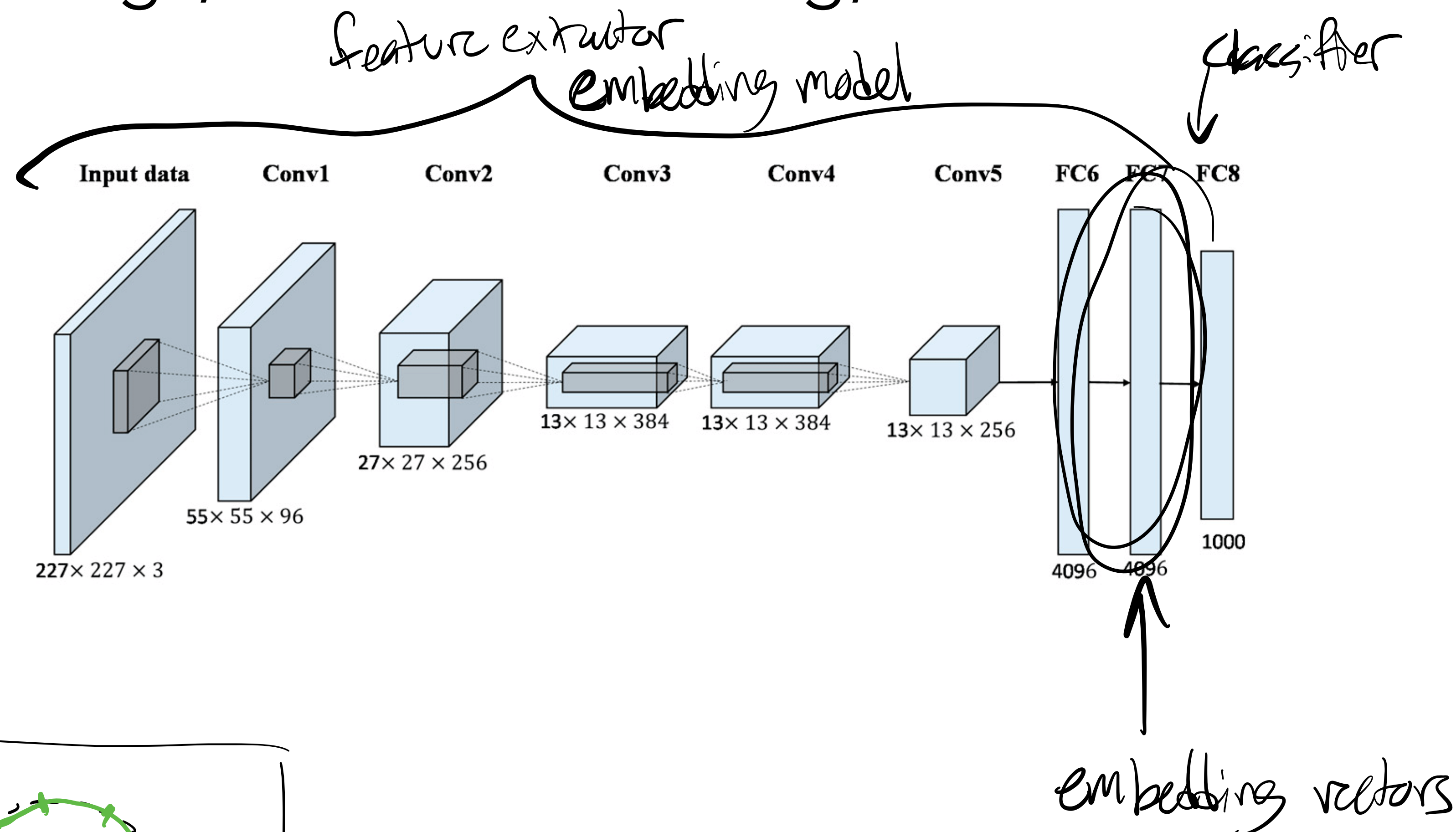
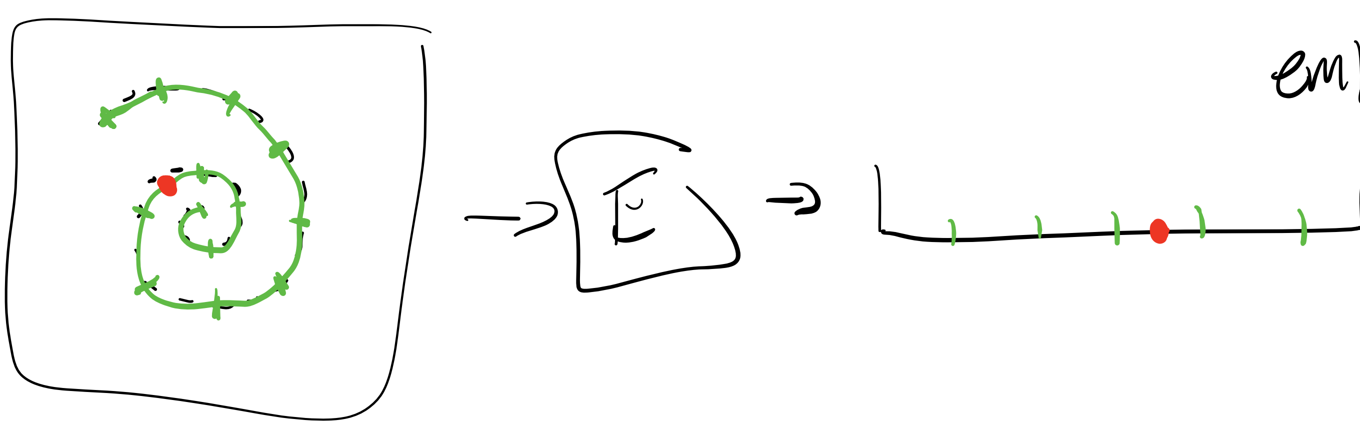
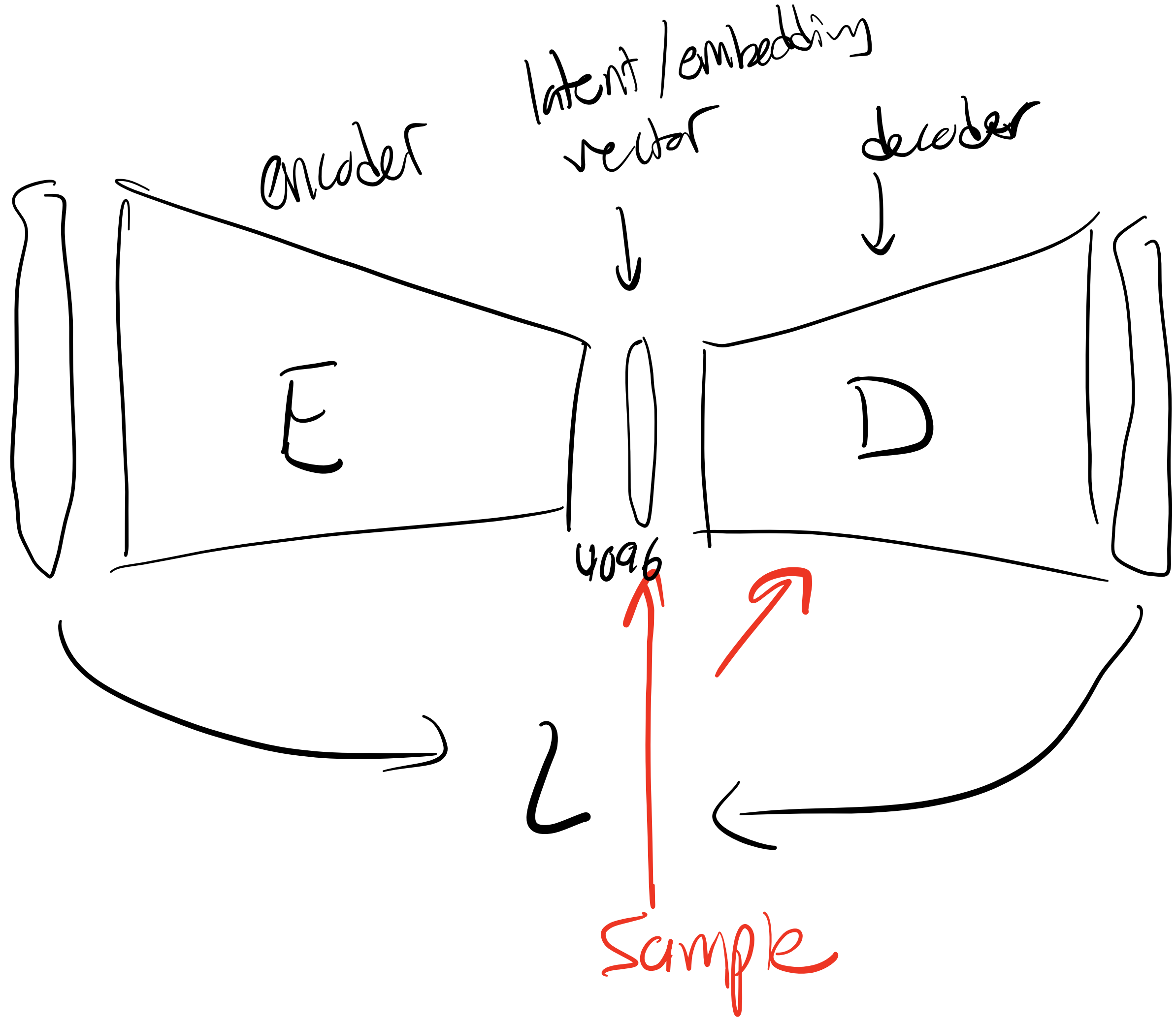
(Sharp?) left turn:
Embeddings, Manifold Learning, and Autoencoders



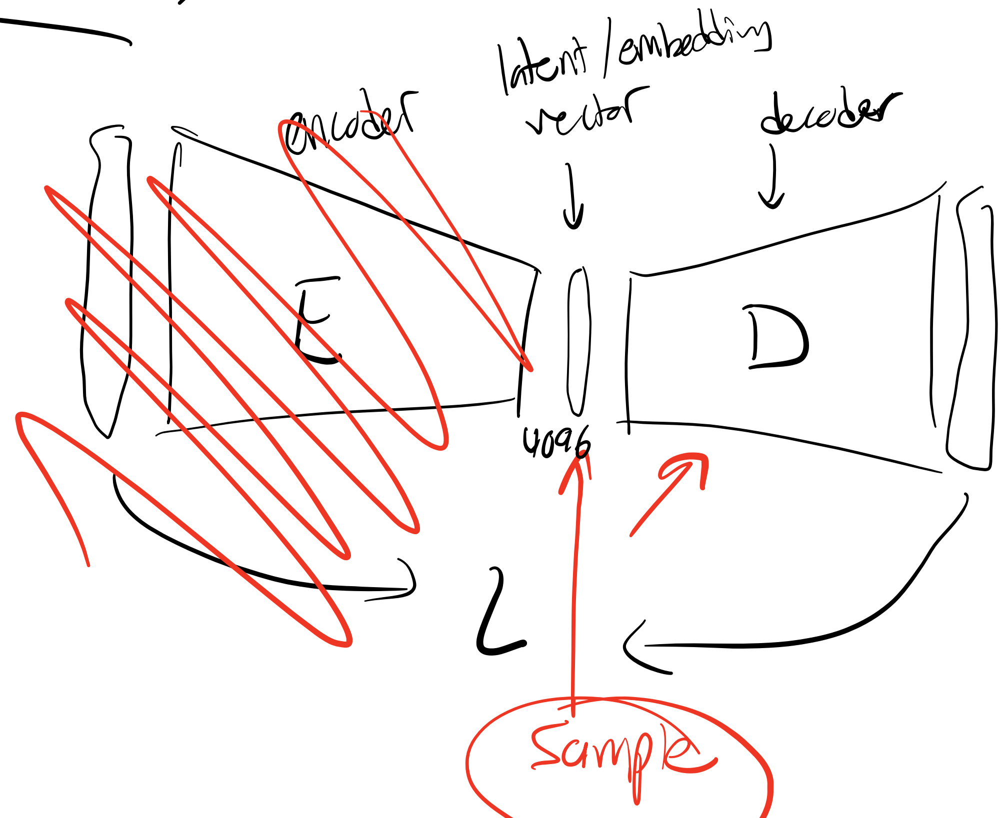



latent / embedding vector
encoder
decoder
E
D
4096
L
Sample

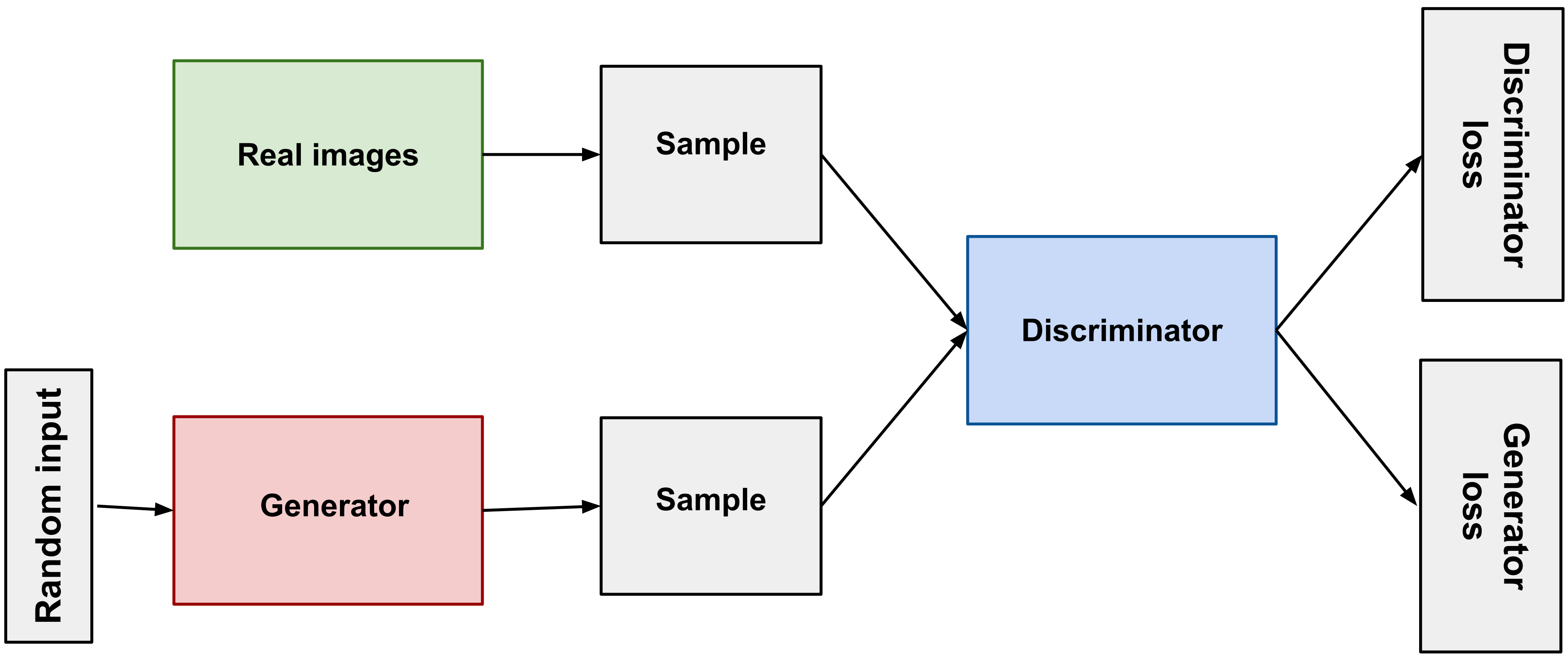
Generative Modeling

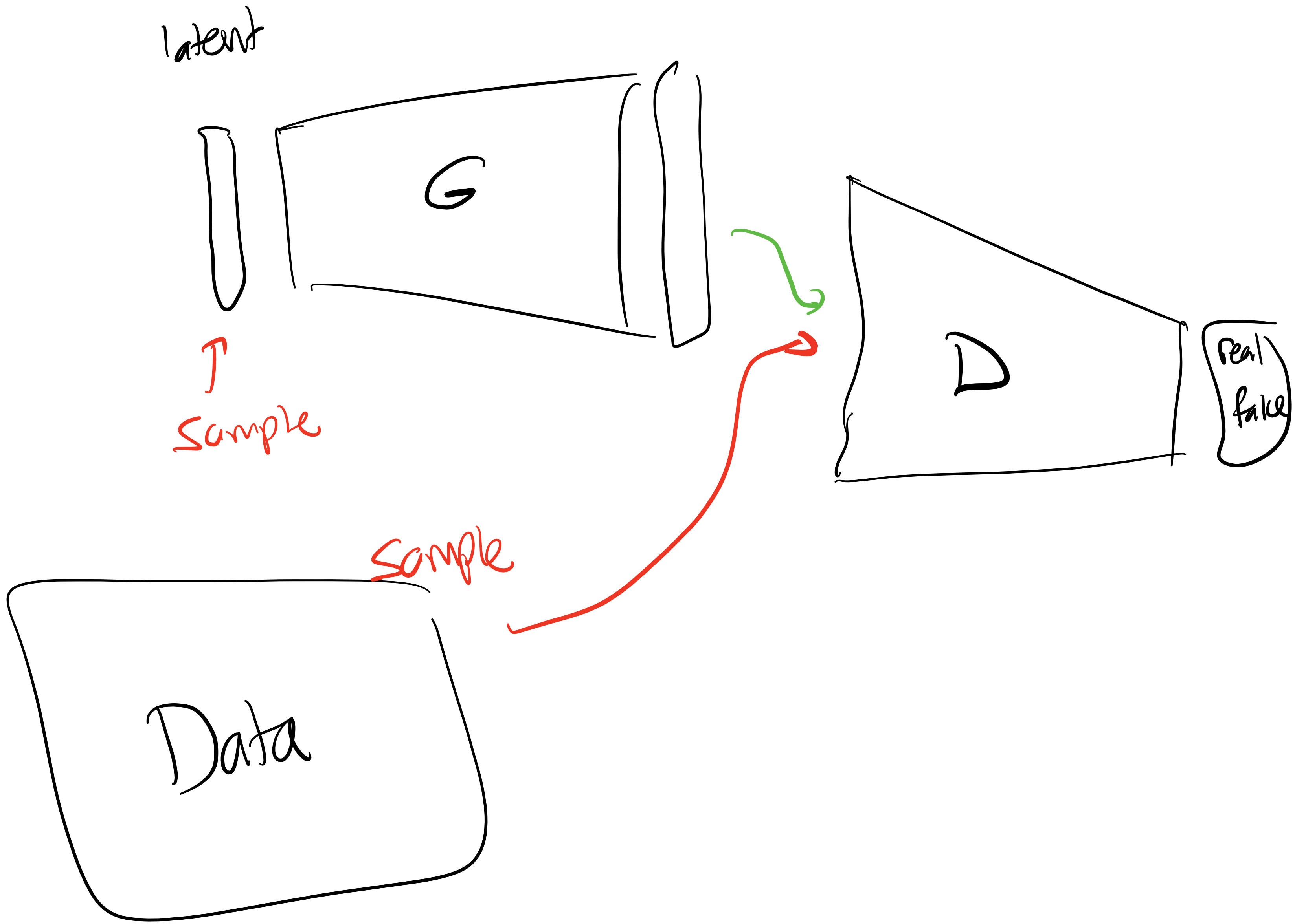
Dis: $p(y|x)$

Gen: $p(x, y)$



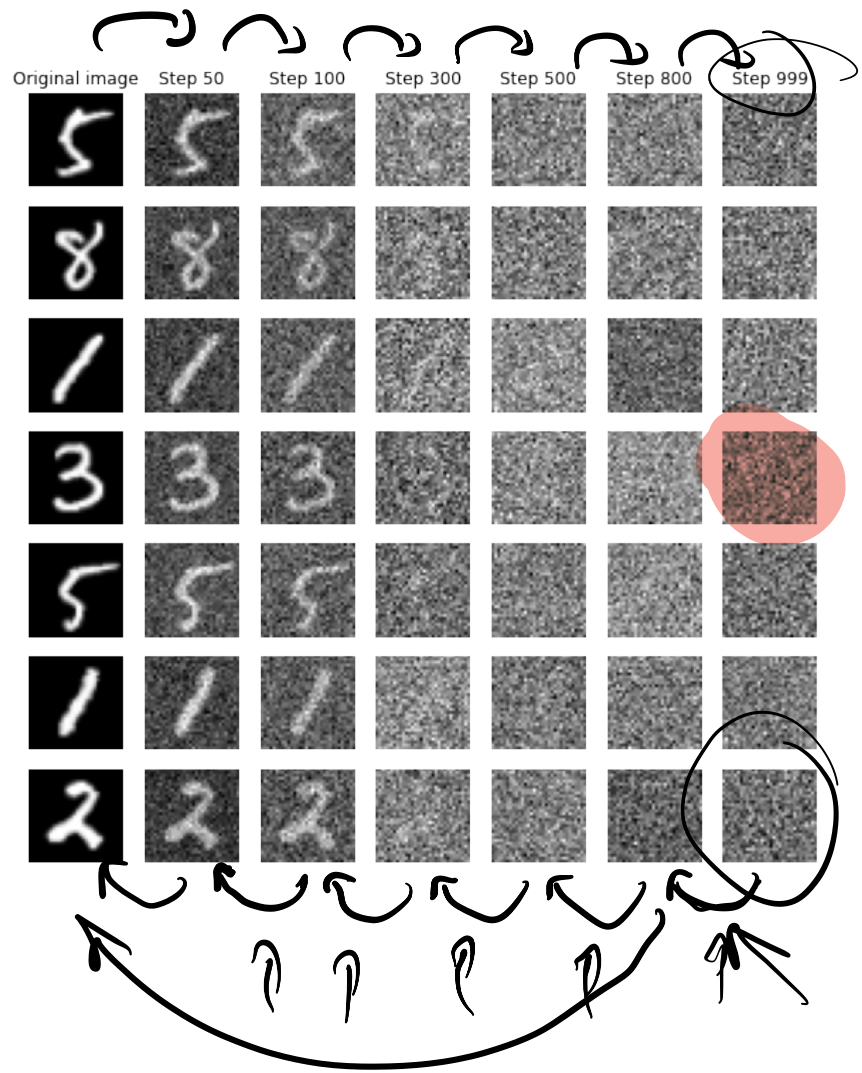
Generative Adversarial Networks



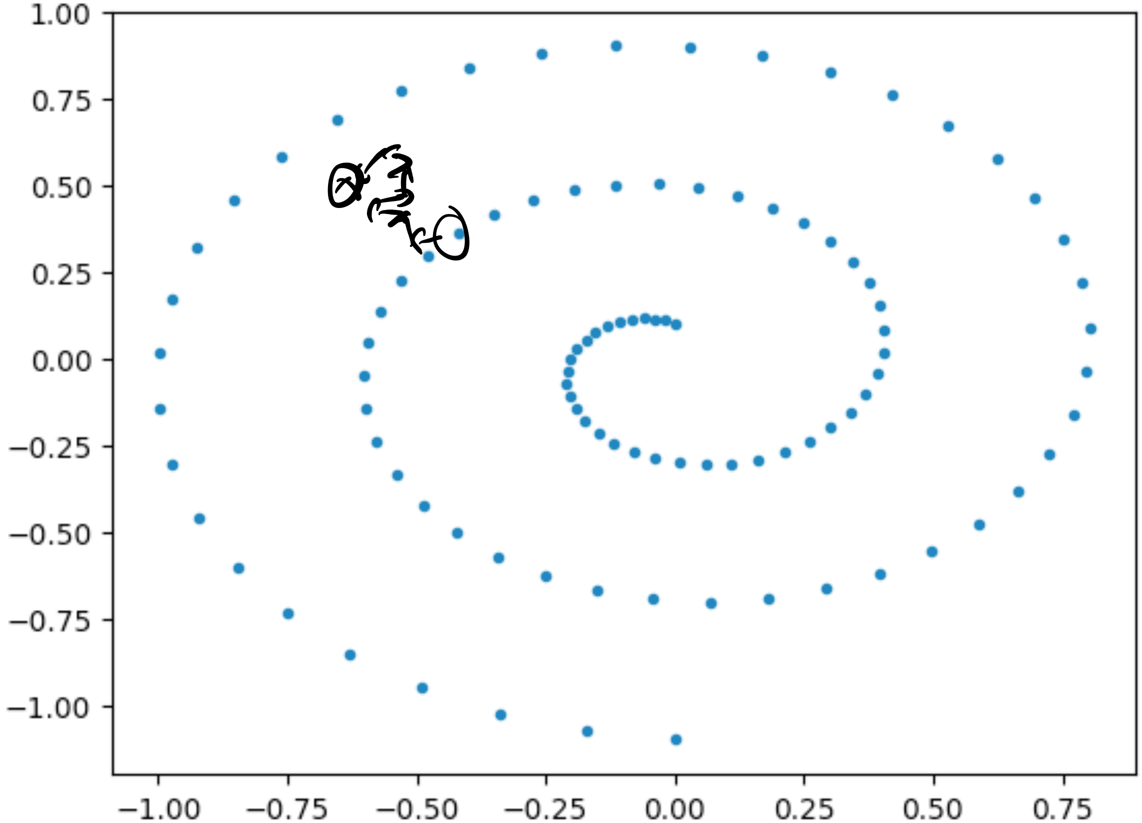

latent
G
sample
D
real
fake
sample
Data

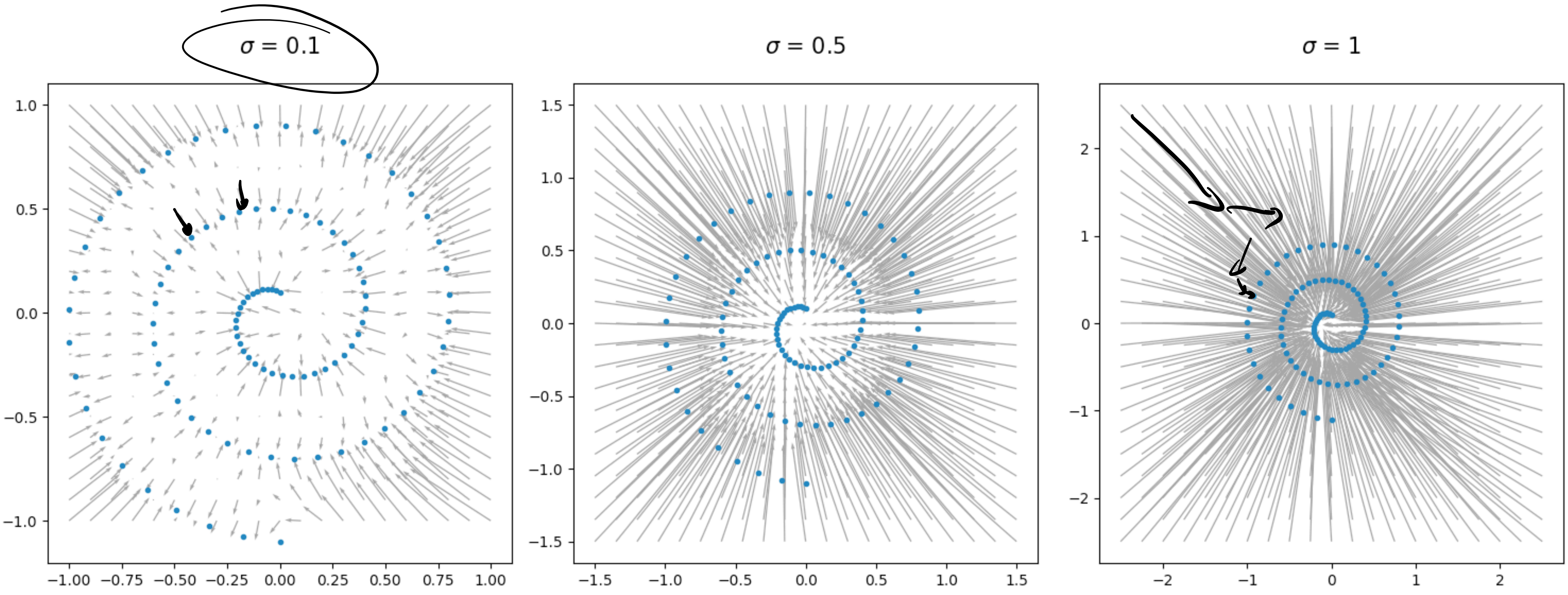

pixers

Diffusion Models



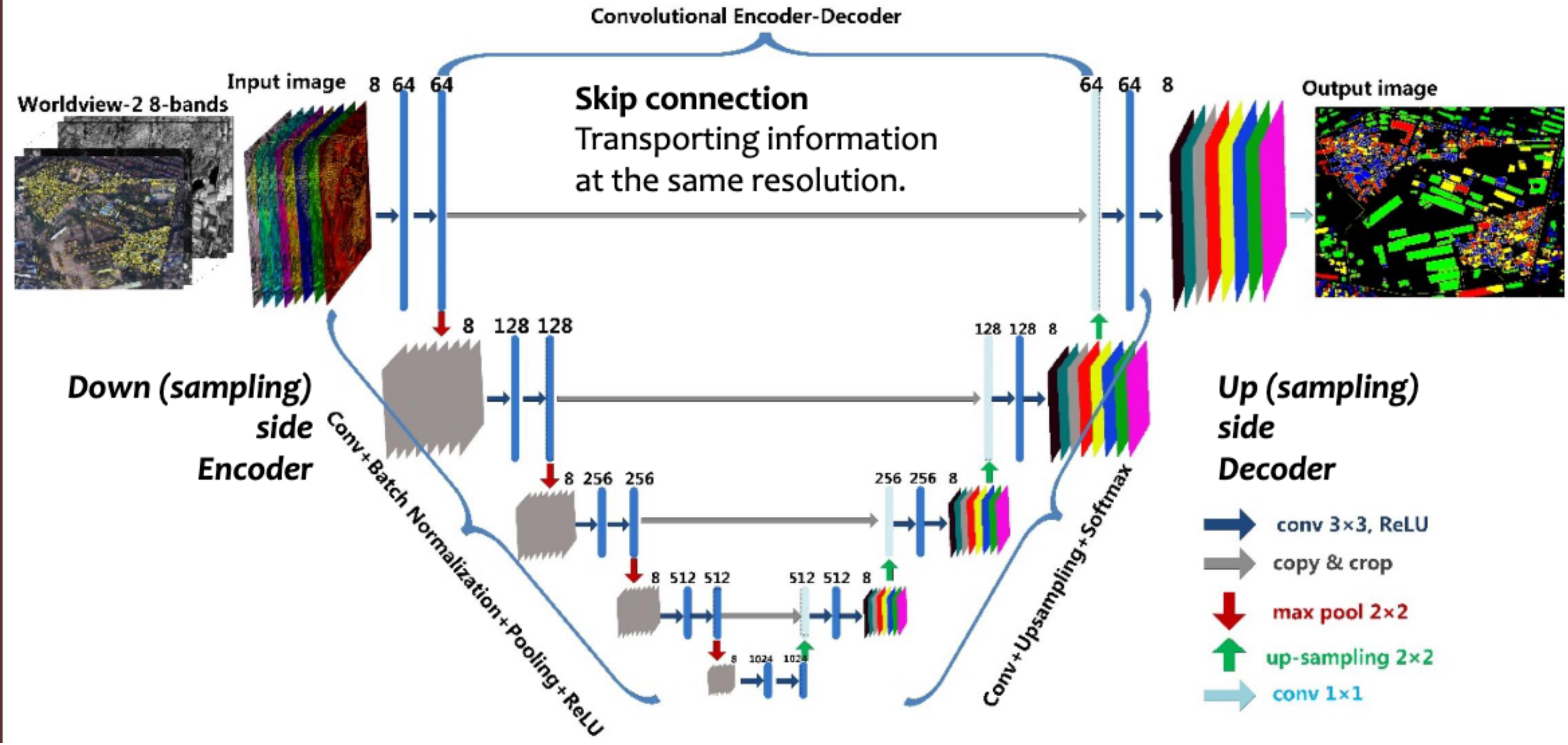
Some other good visuals: https://www.chenyang.co/diffusion.html


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75

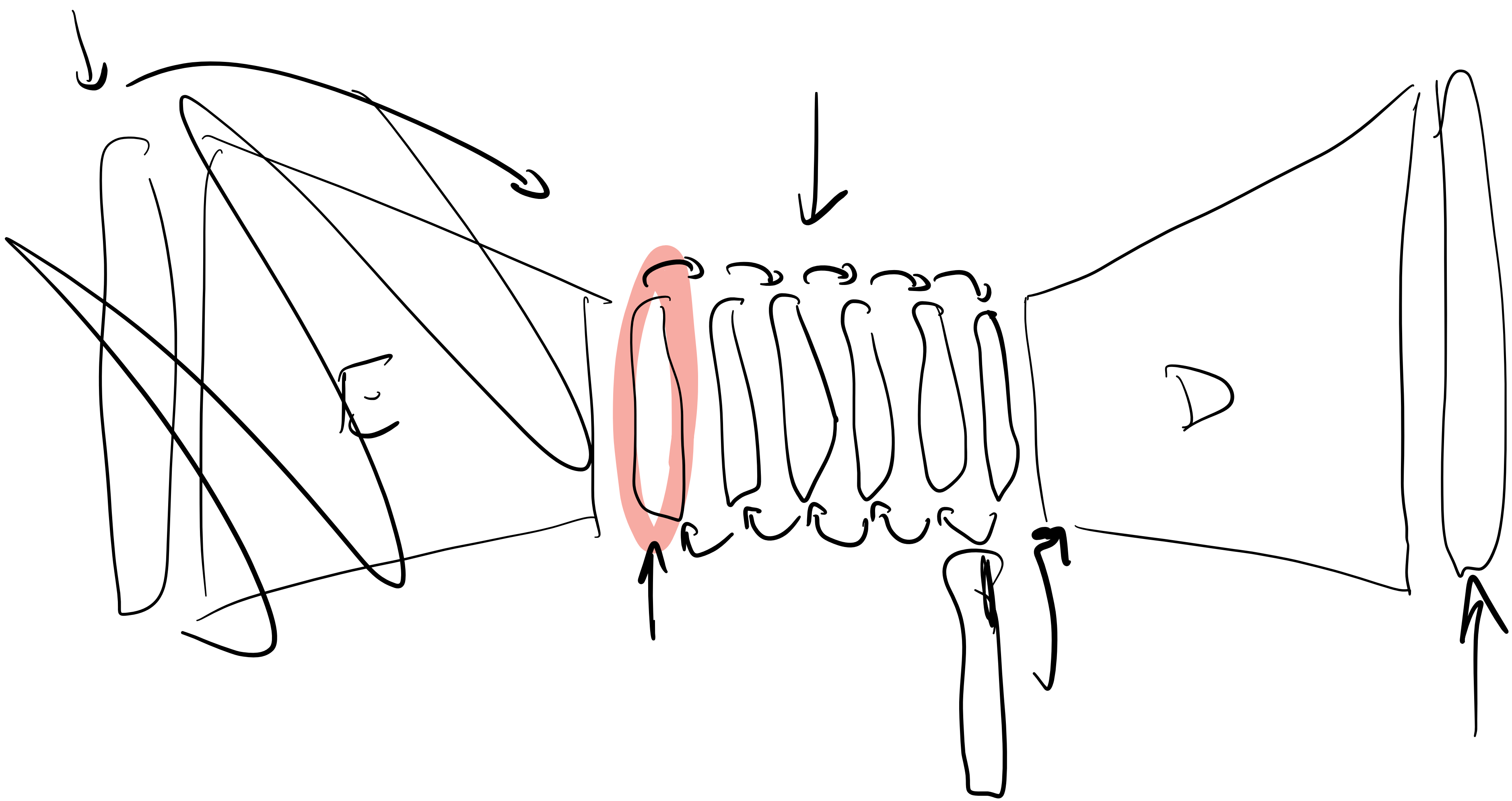

$\sigma = 0.1$
$\sigma = 0.5$
$\sigma = 1$
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
−1.0 −0.5 0.0 0.5 1.0
−1.5 −1.0 −0.5 0.0 0.5 1.0 1.5
−2 −1 0 1 2

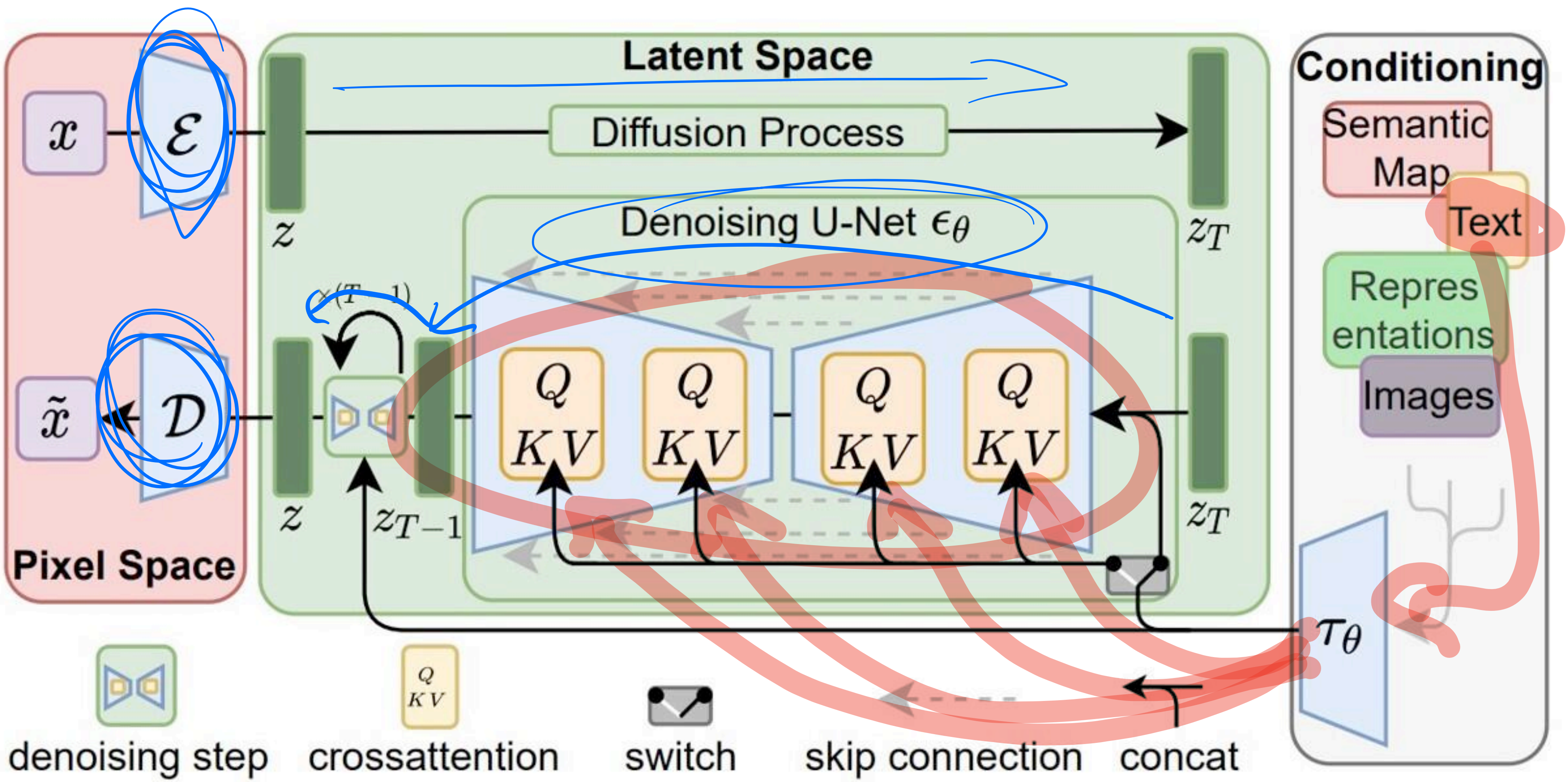

© JIAN

UNet - a more detailed picture



Stable Diffusion (without the conditioning)

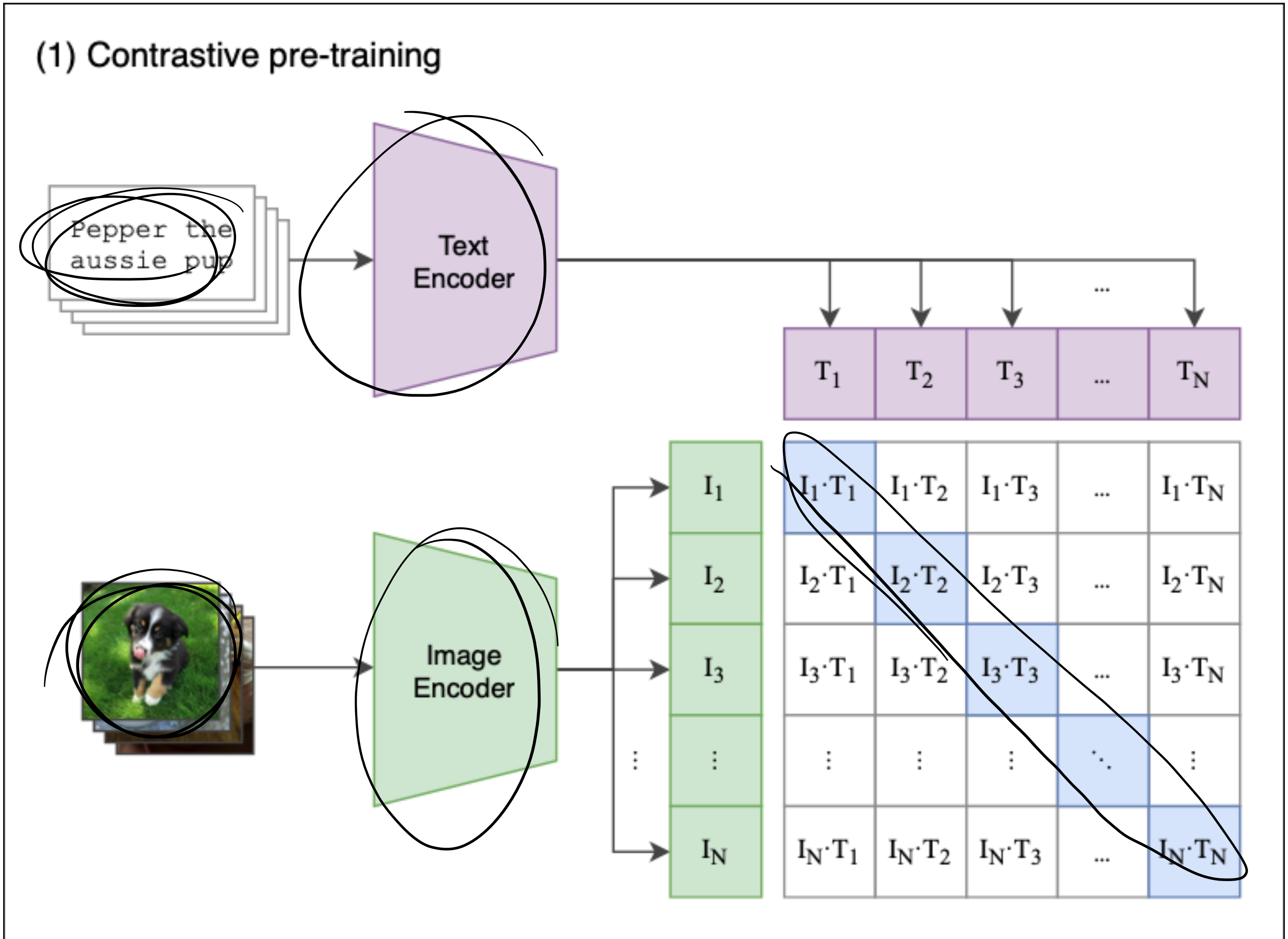




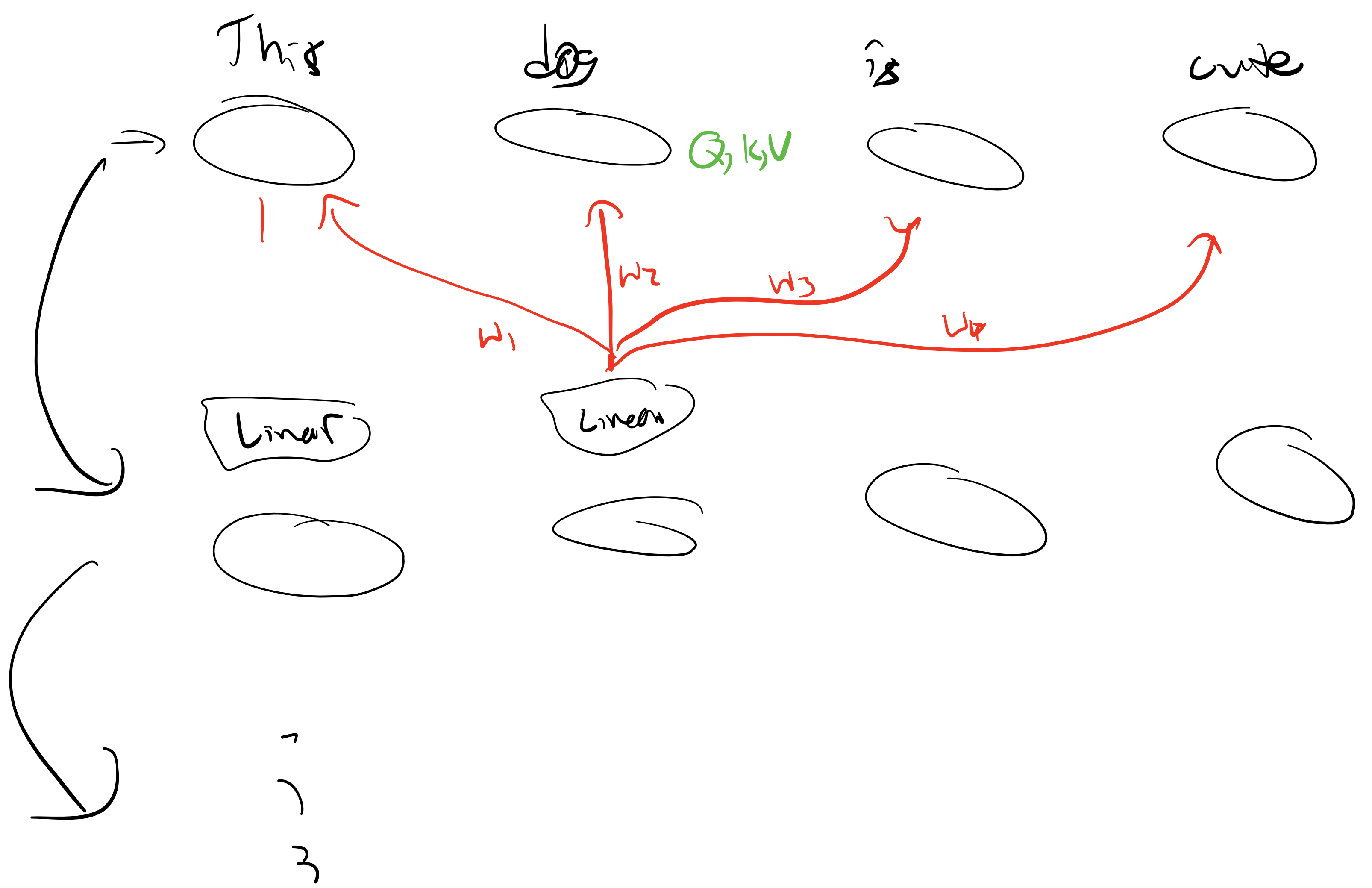
Latent Space
Conditioning
Diffusion Process
Semantic Map
Text
Repres entations
Images
Denoising U-Net ϵ_θ
x
$\mathcal{E}$
z
z_T
$\tilde{x}$
$\mathcal{D}$
z_{T-1}
$\times(T-1)$
Q
$K\ V$
τ_θ
Pixel Space
denoising step
crossattention
switch
skip connection
concat

Vision and Language

Case study: CLIP



"Attention"



unCLIP aka DALL-E 2

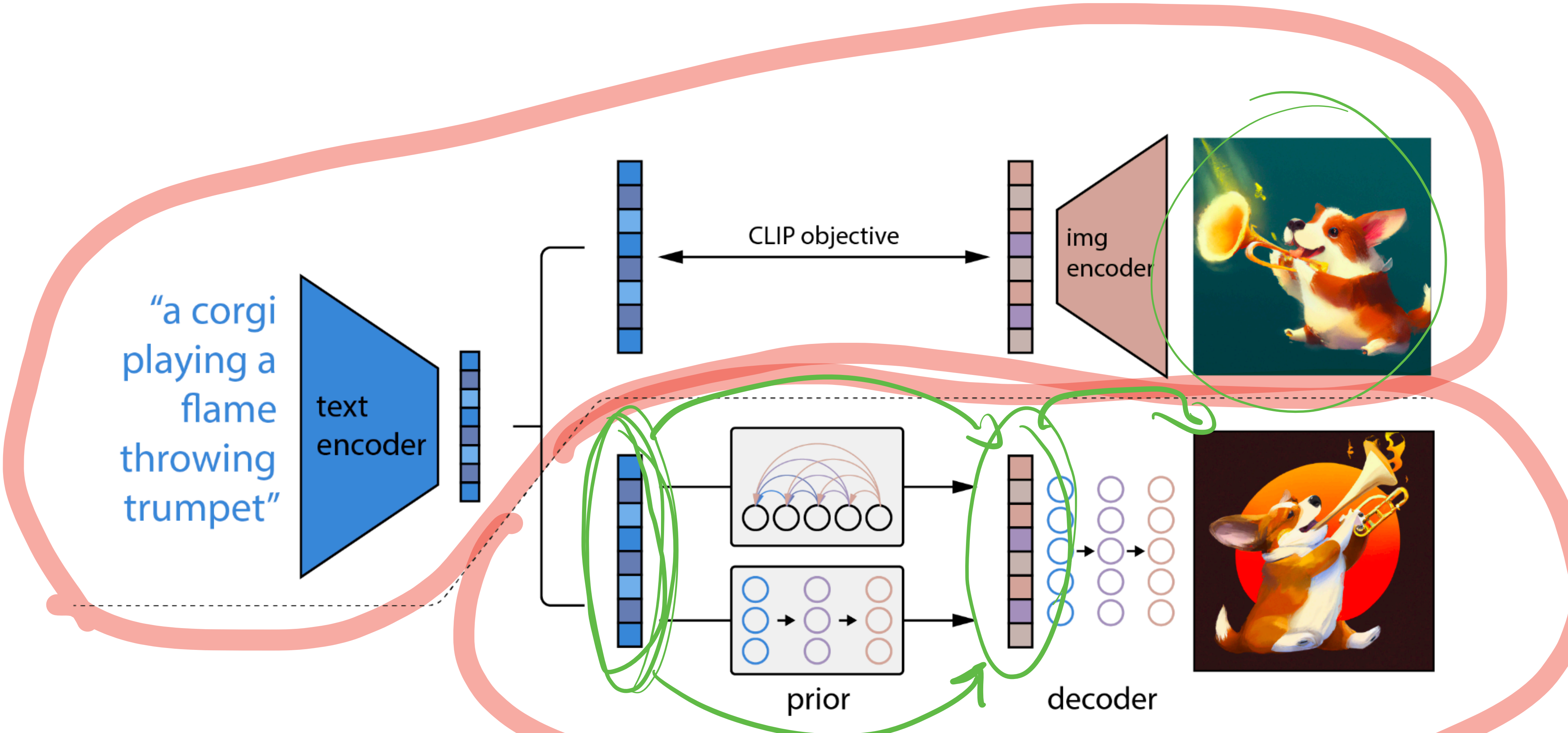


Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive 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.

Stable Diffusion
(with the conditioning)