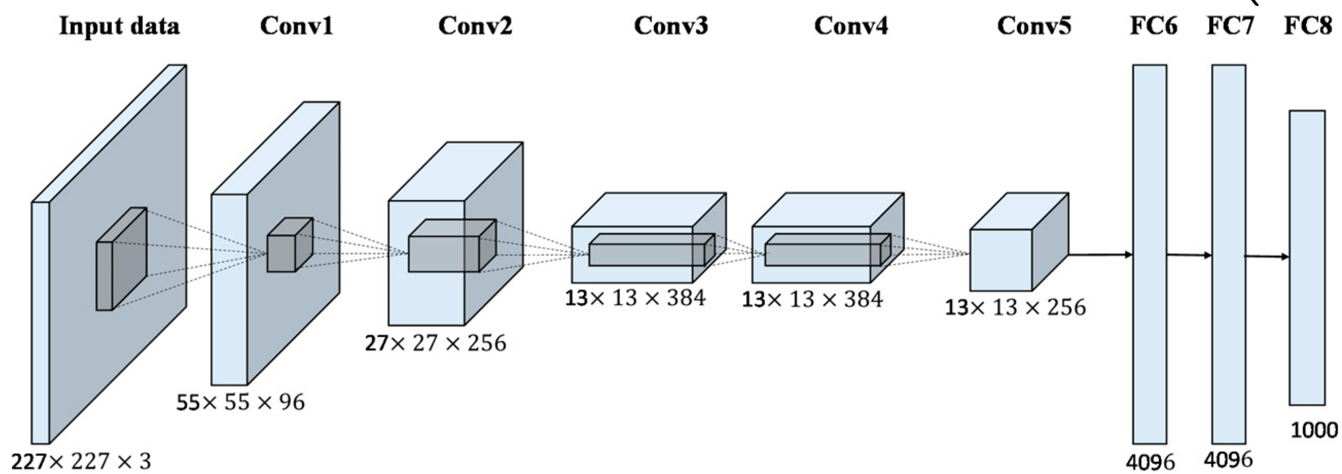
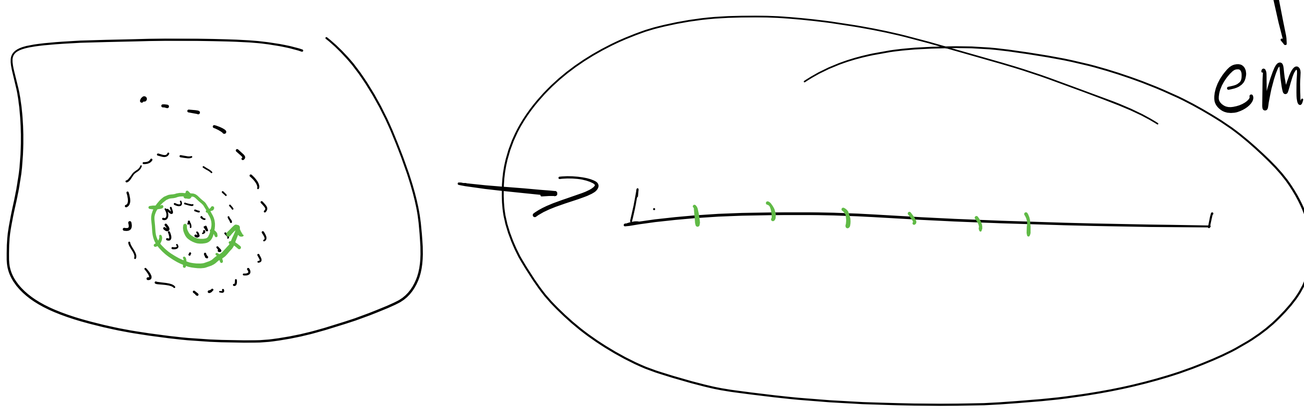
# (Sharp?) left turn:

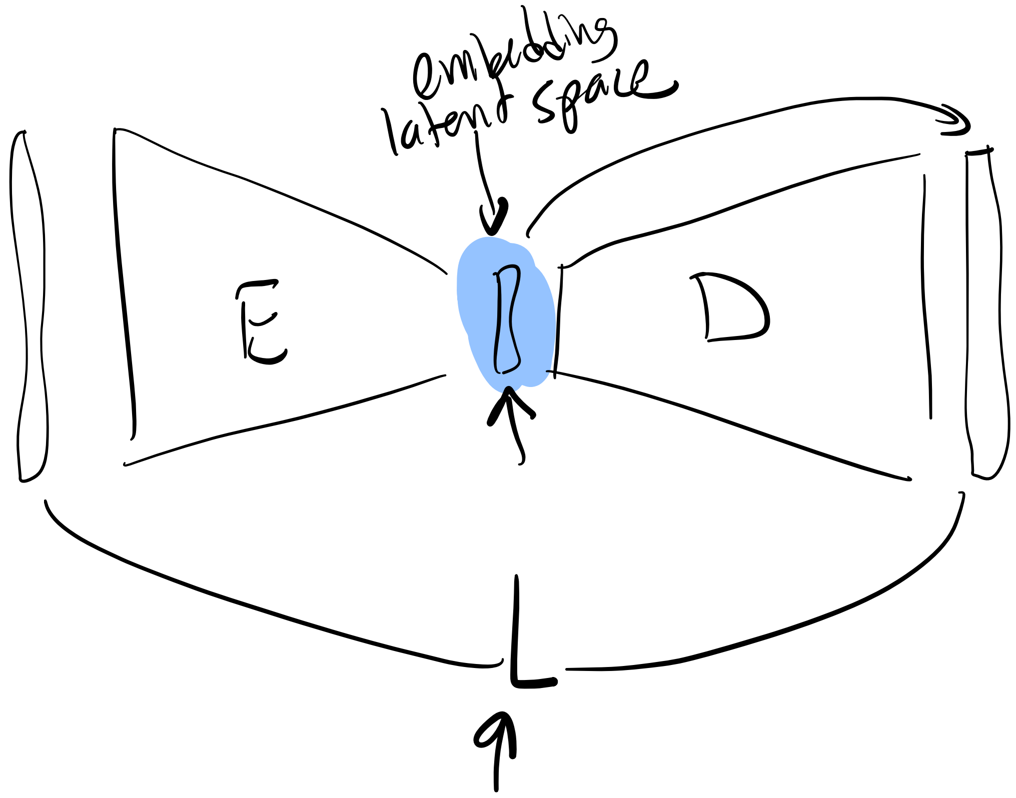
# Embeddings, Manifold Learning, and Autoencoders

embedding model




embedding





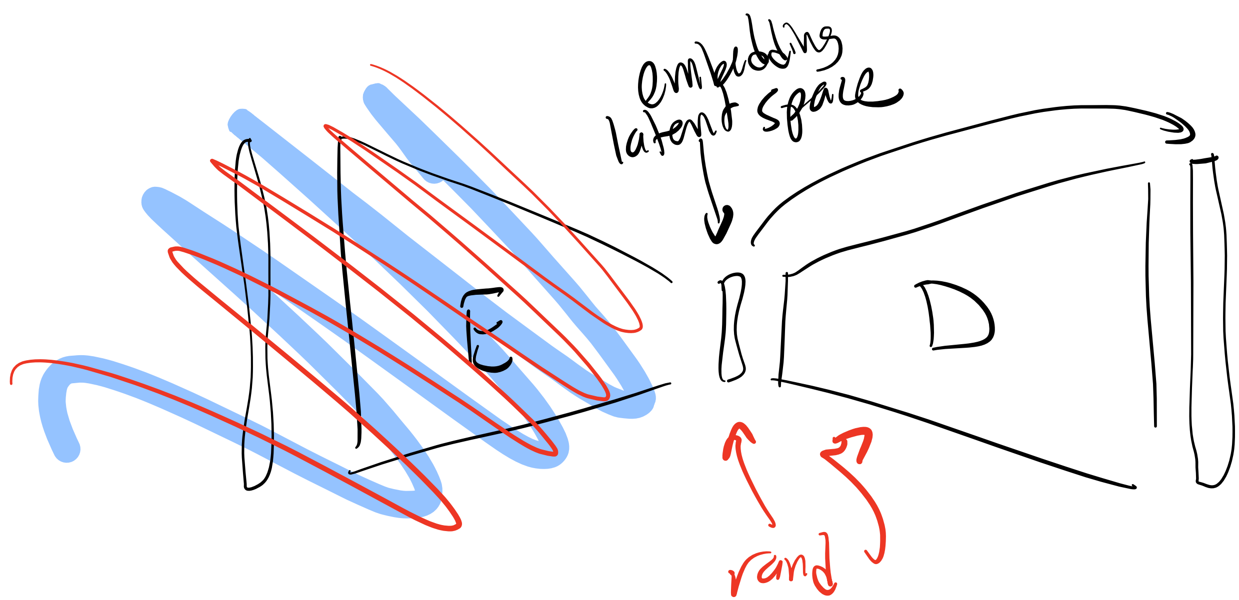
embedding
latent space
E
D
L

# Generative Modeling

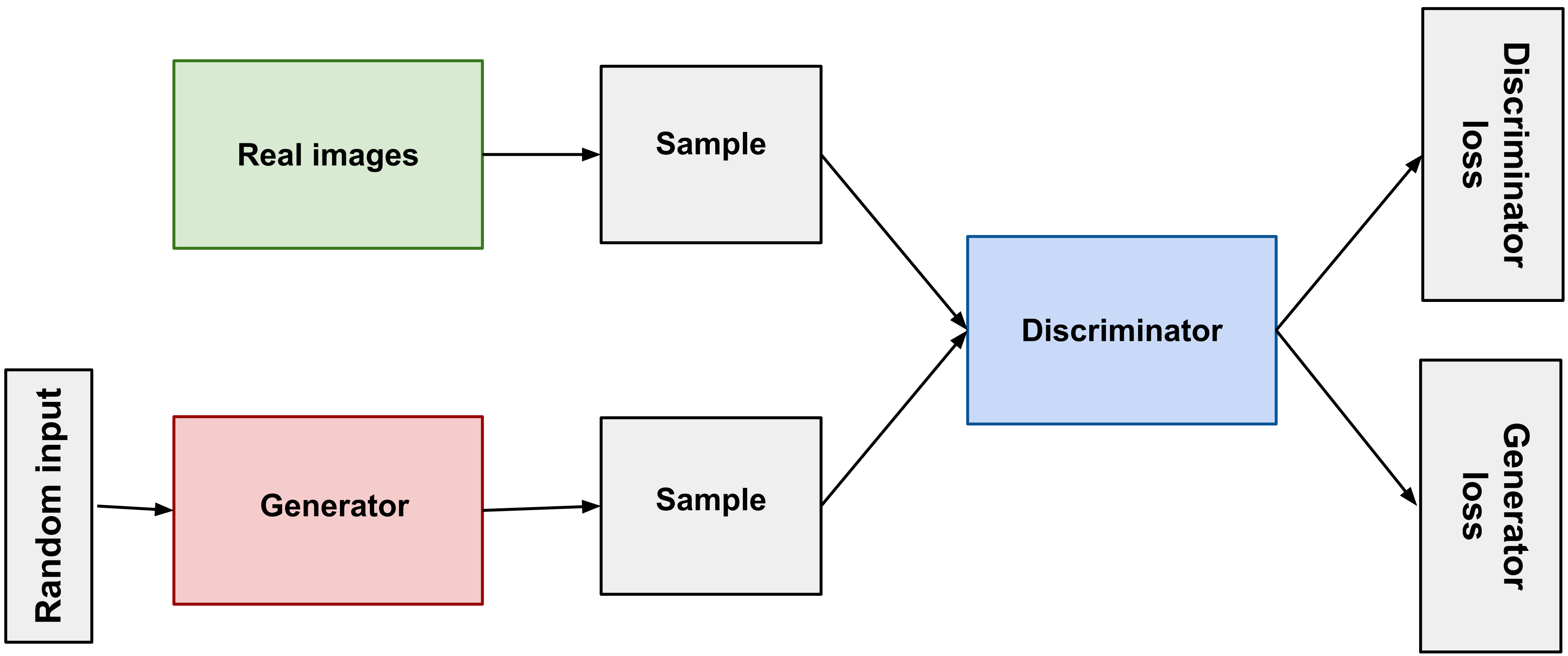
Disc: $p(y|x)$

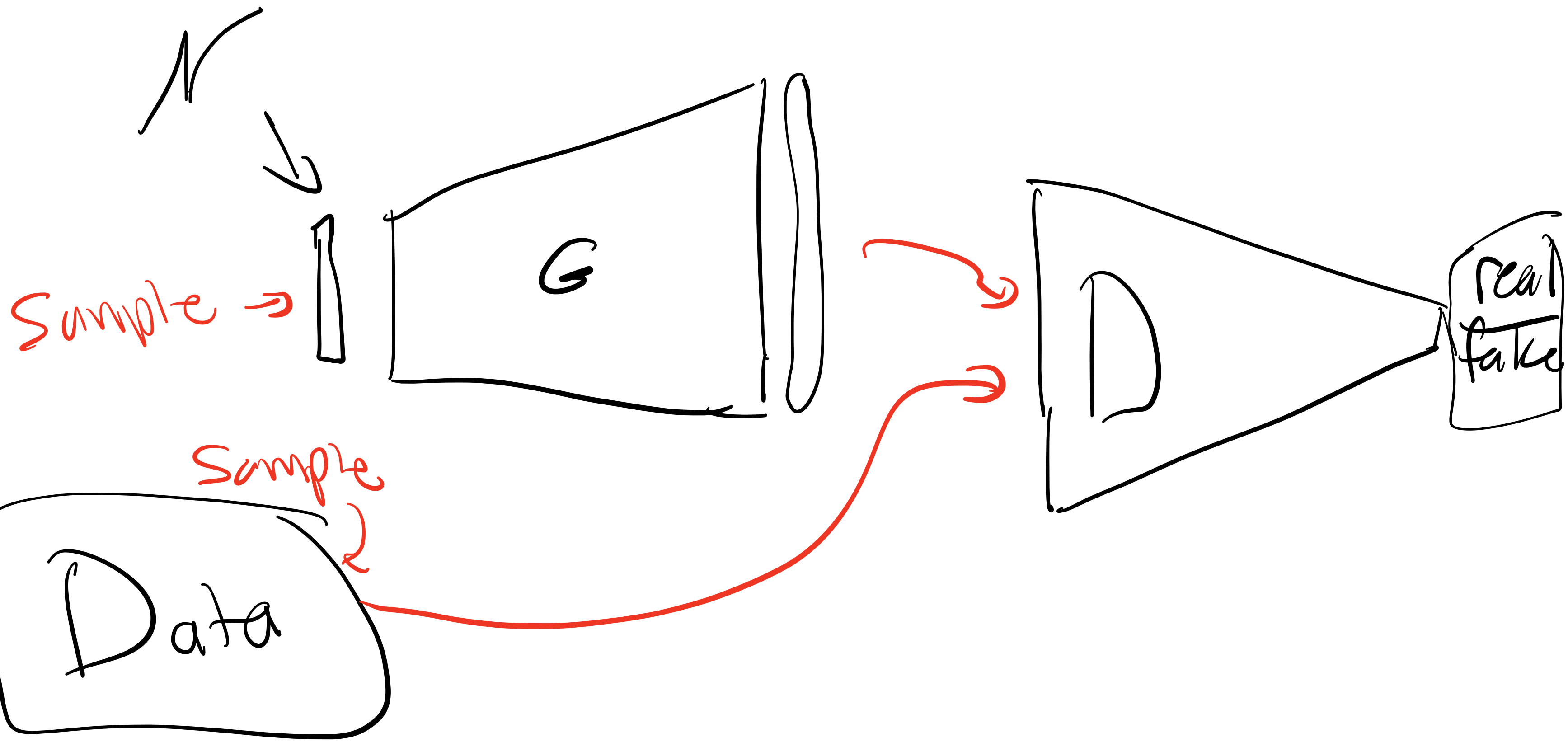
Gen: $p(x,y)$

sample



# Generative Adversarial Networks

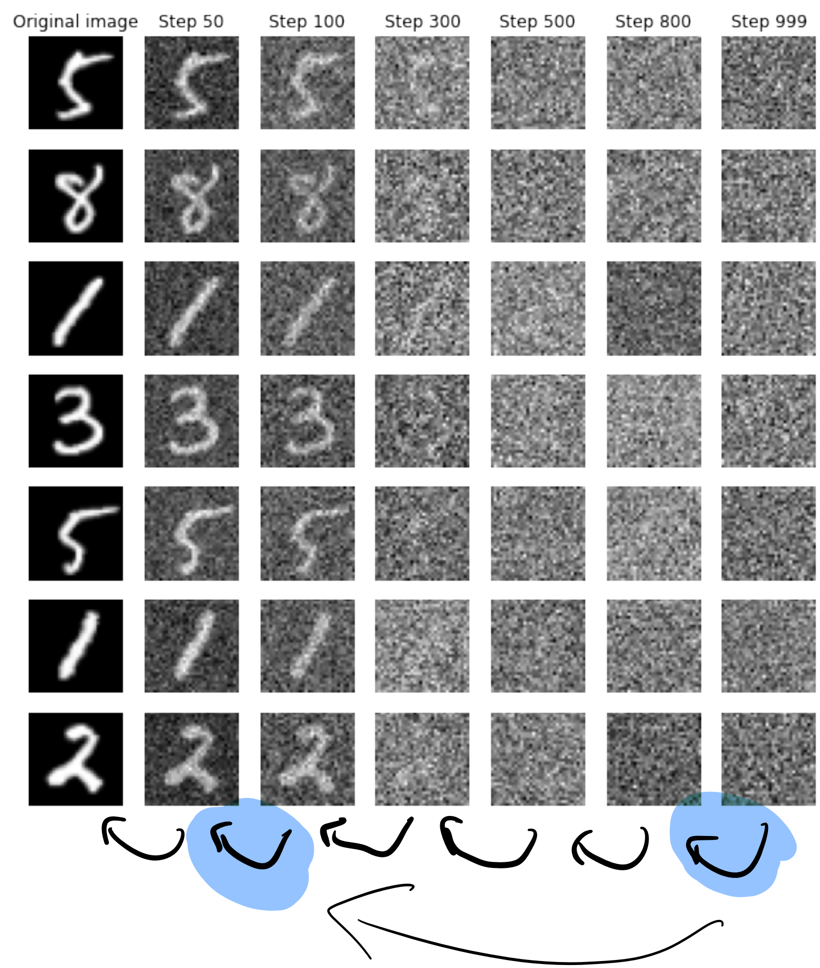




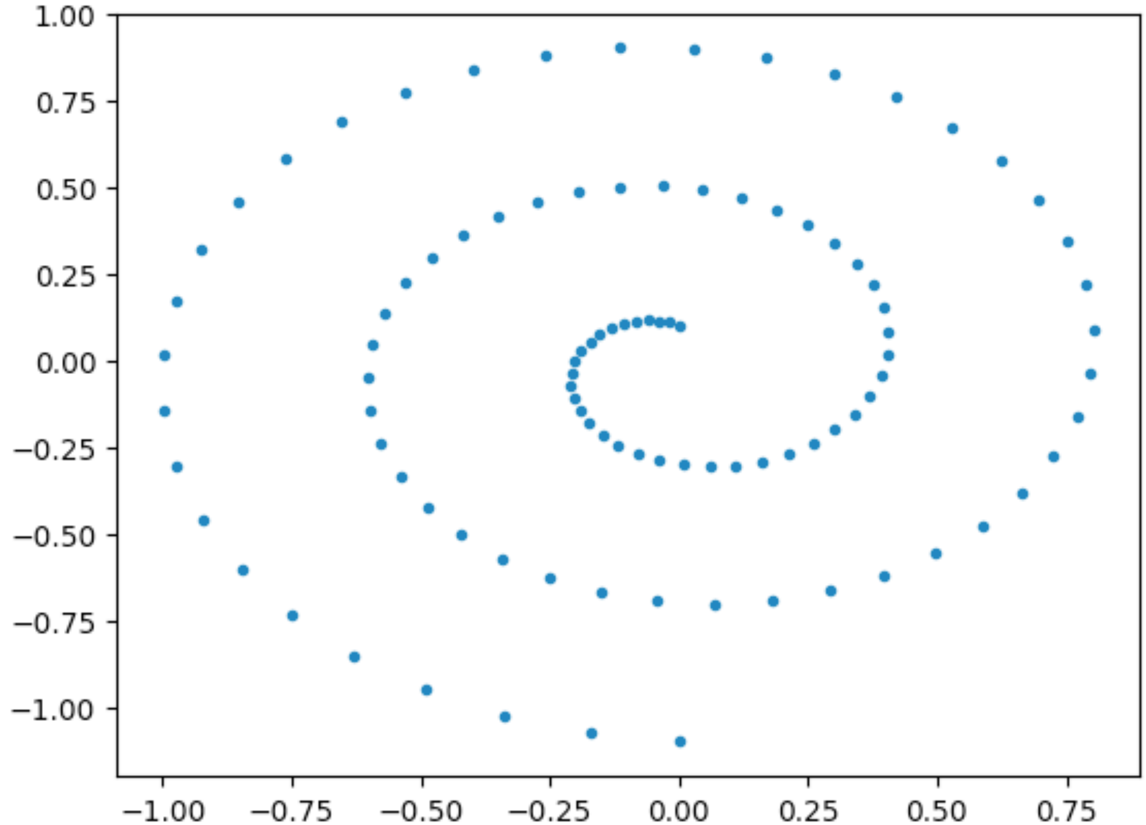
N
Sample
G
D
real
fake
Sample
Data


pixers

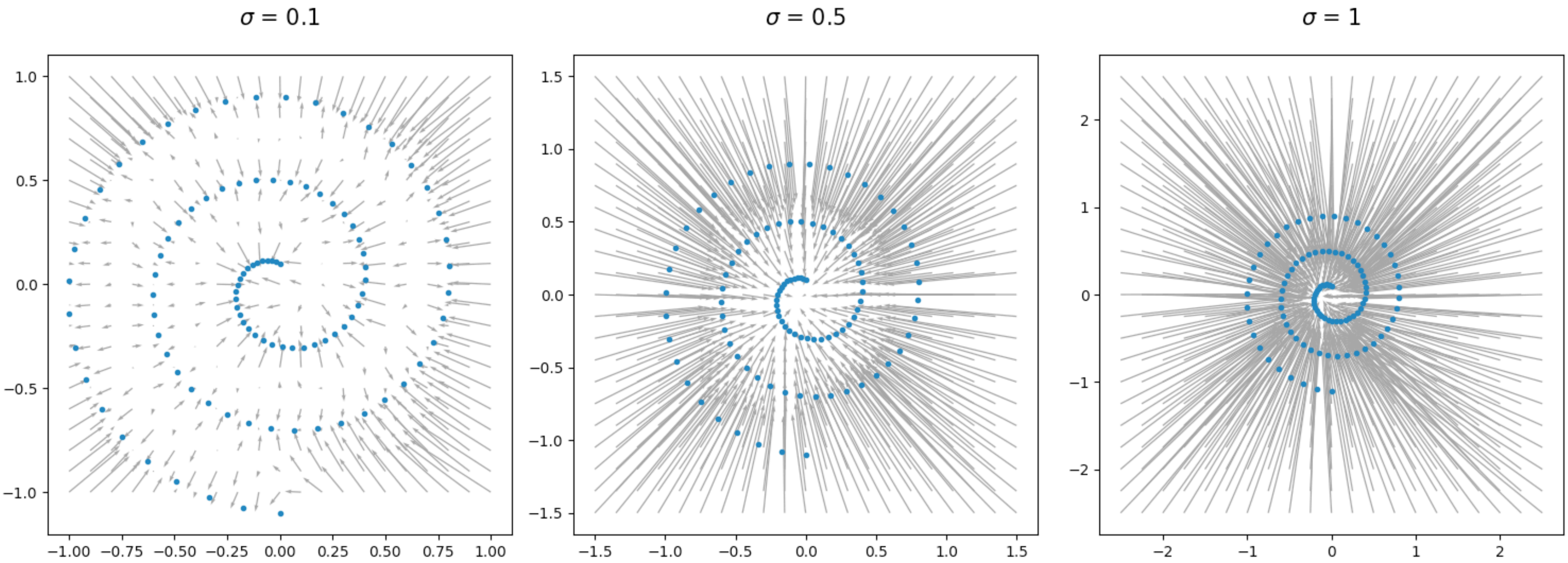
# Diffusion Models




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



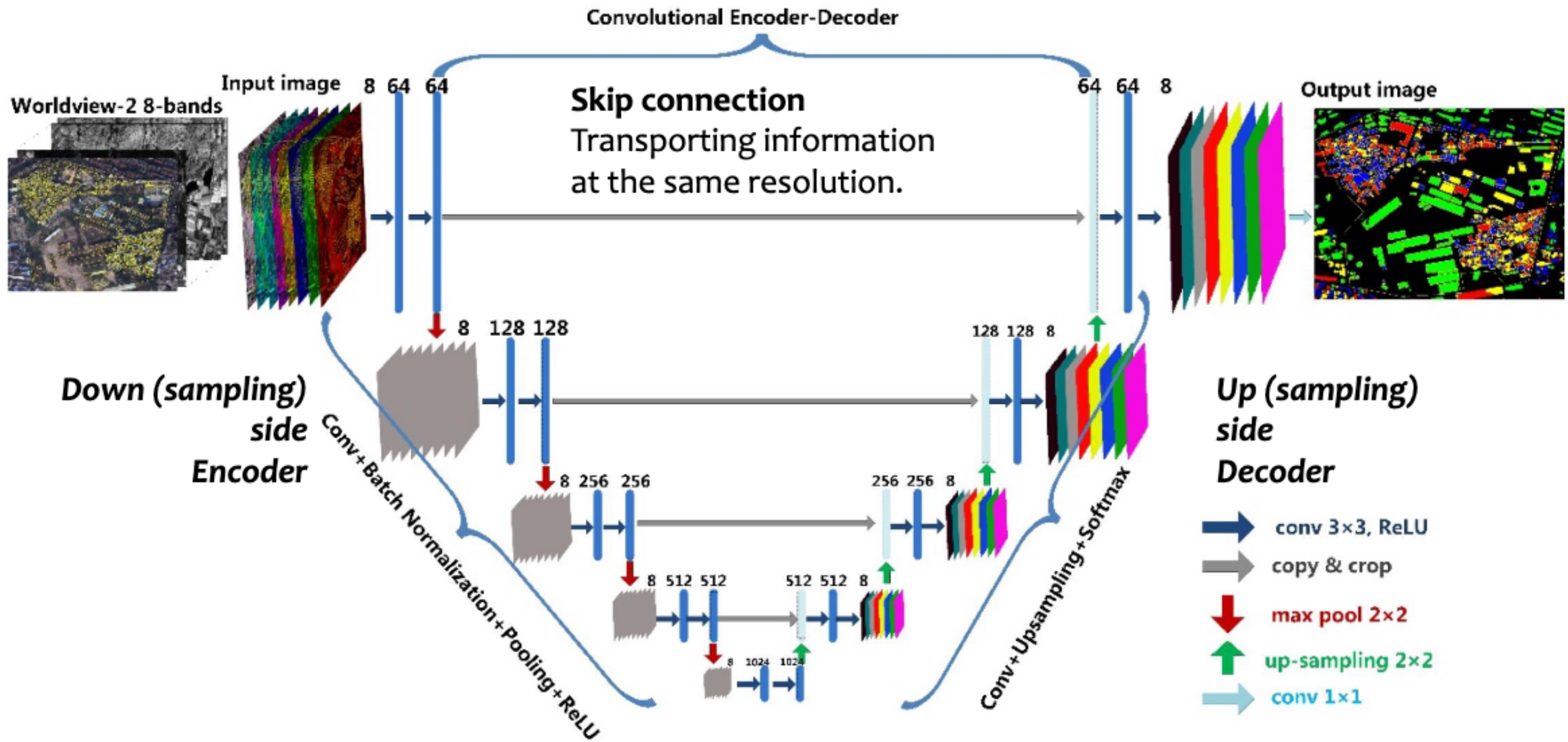
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75



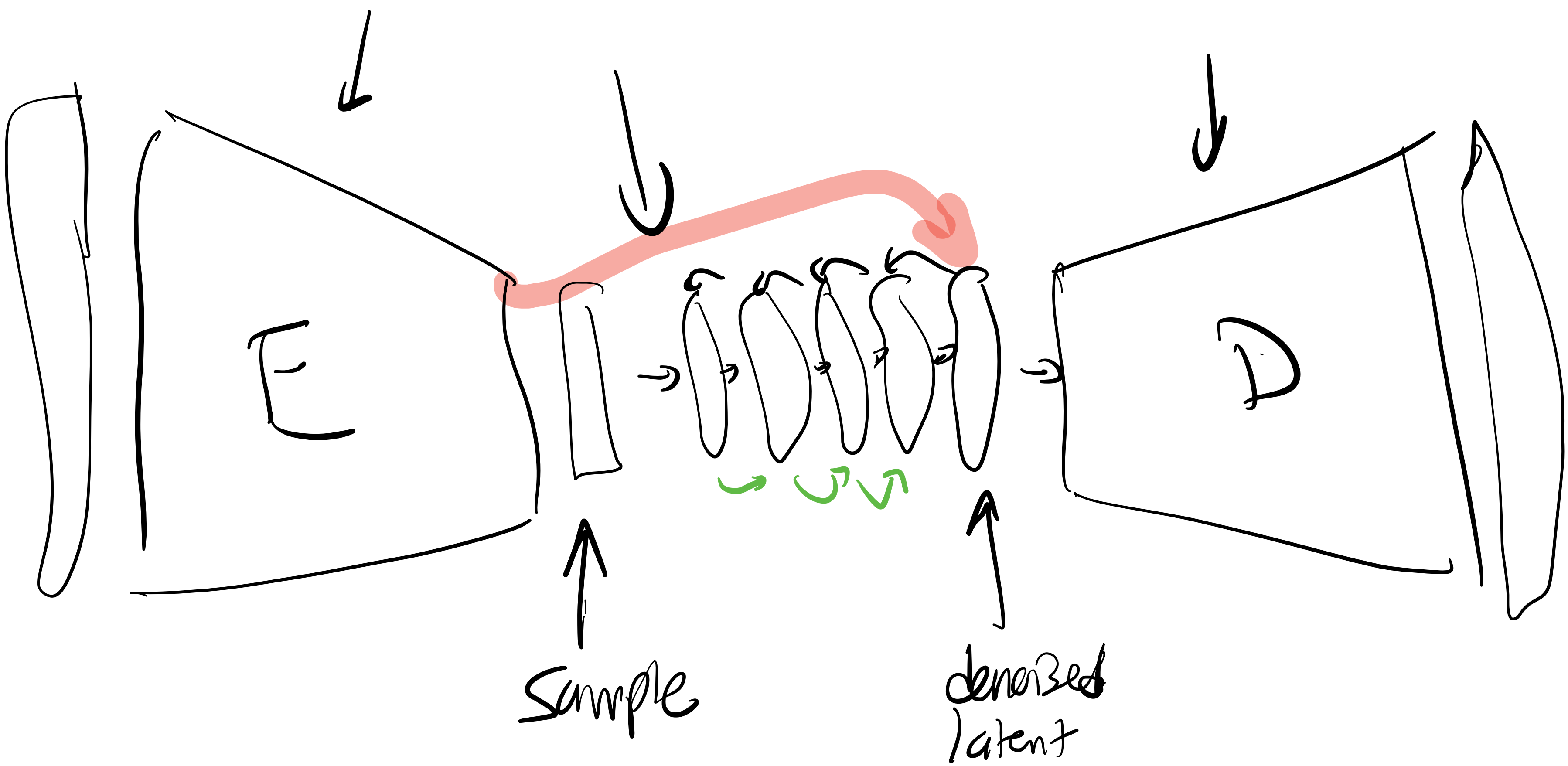
σ = 0.1
σ = 0.5
σ = 1

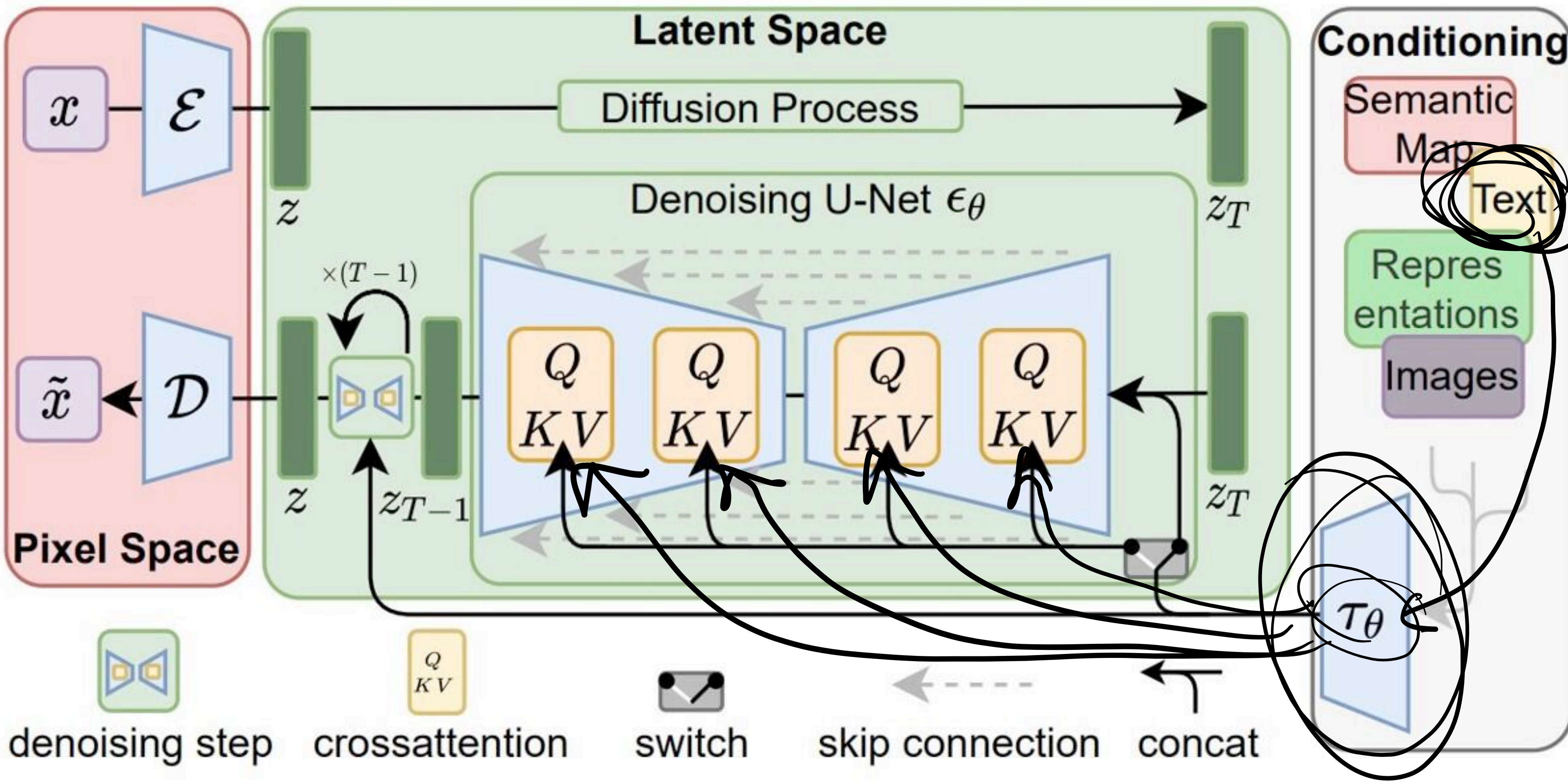

© JIAN

# UNet - a more detailed picture



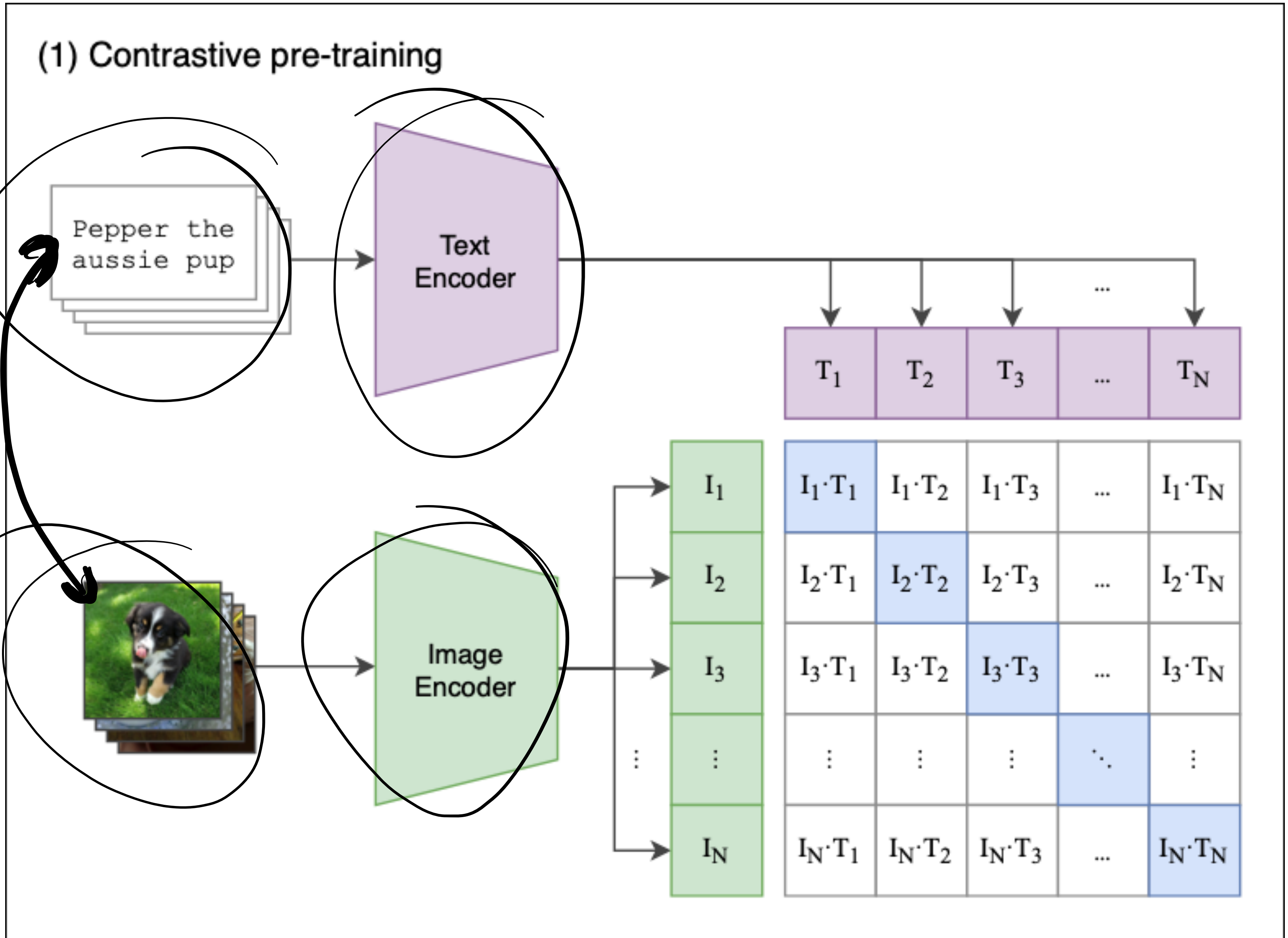
# Stable Diffusion (without the conditioning)




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

# Vision and Language

# Case study: CLIP



# "Attention"

The $t_1$

dog $t_2$

is $t_3$

cute $t_4$

$$t_i' = \sum w_i f(t_i)$$

# unCLIP aka DALL-E 2

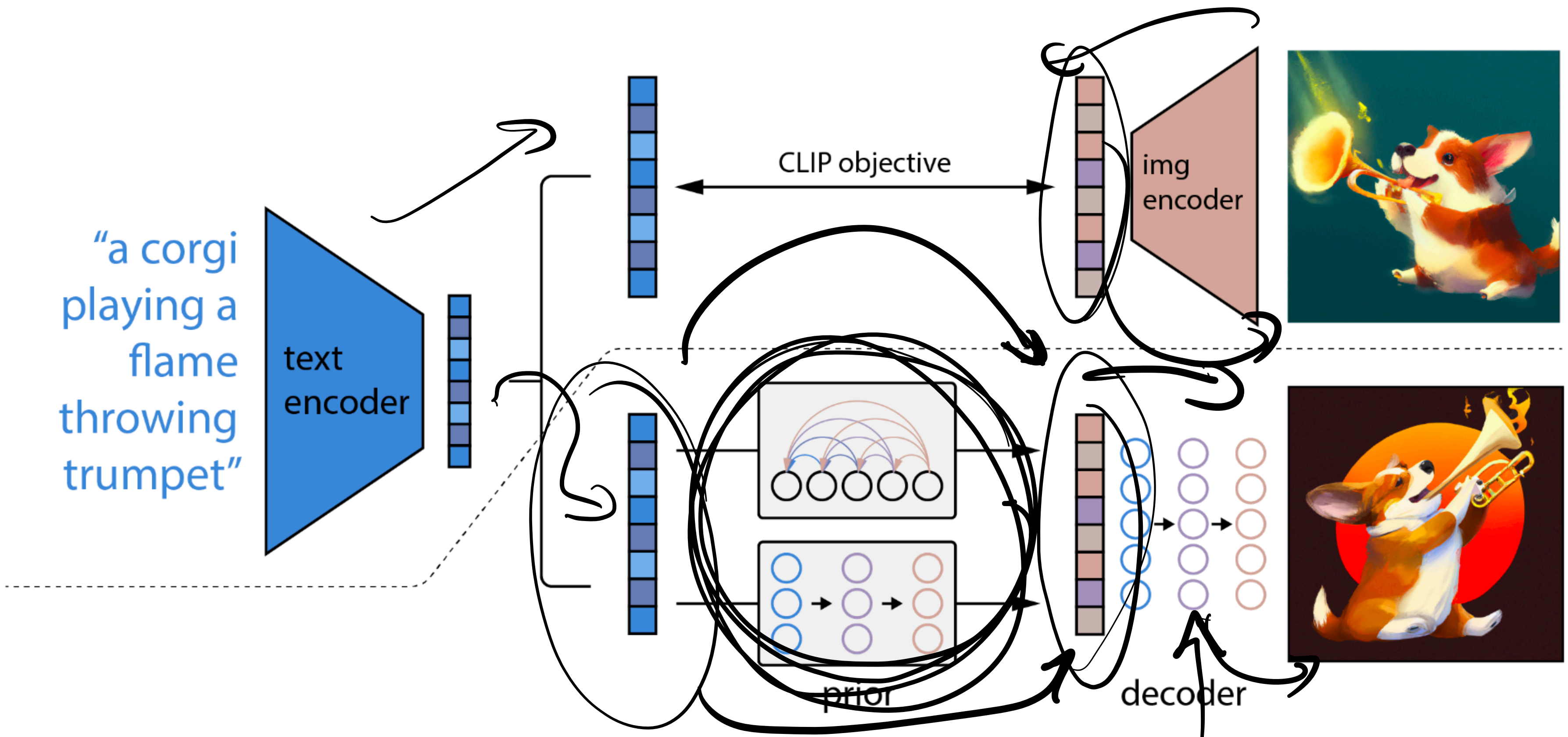



Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive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