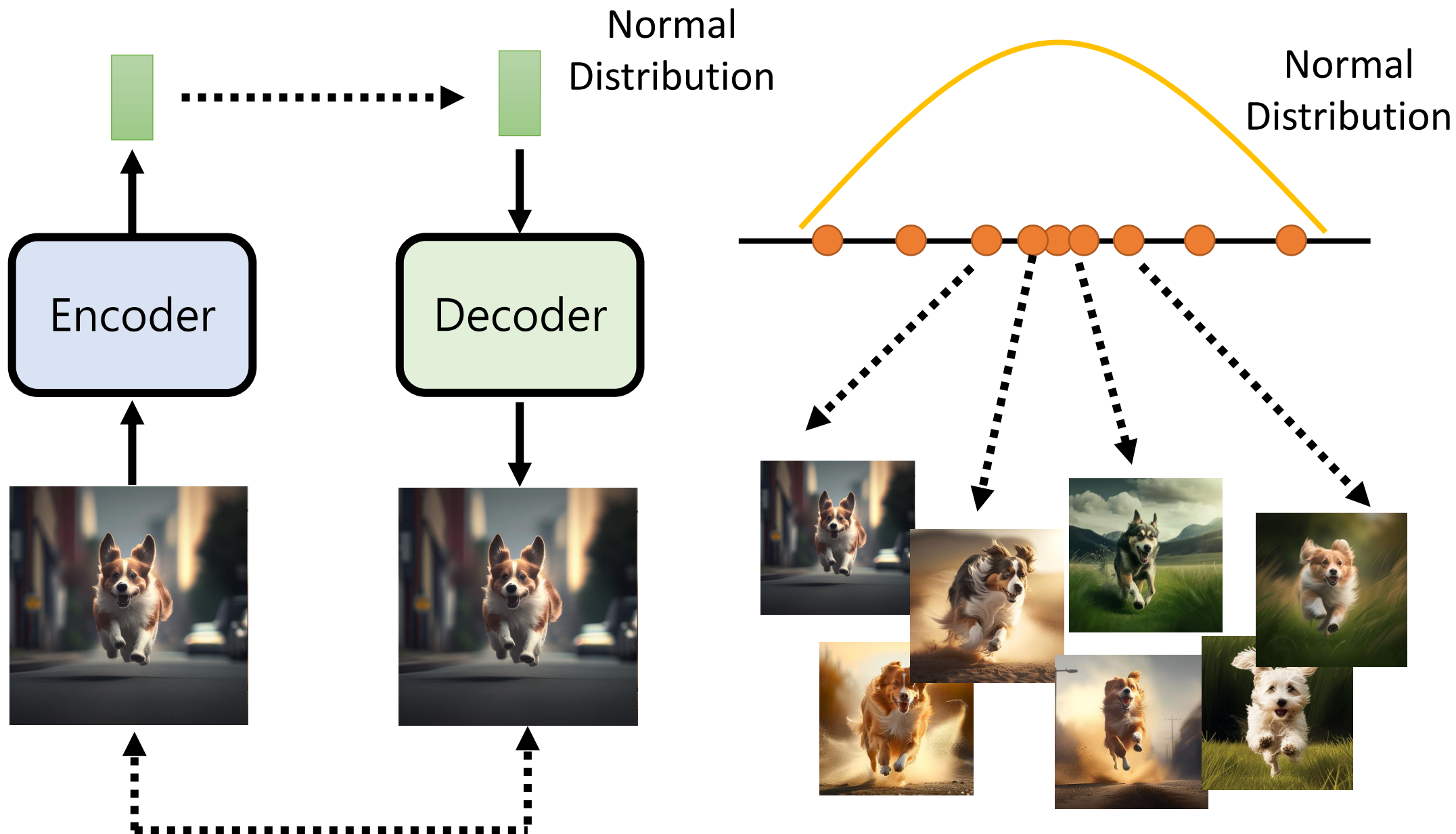
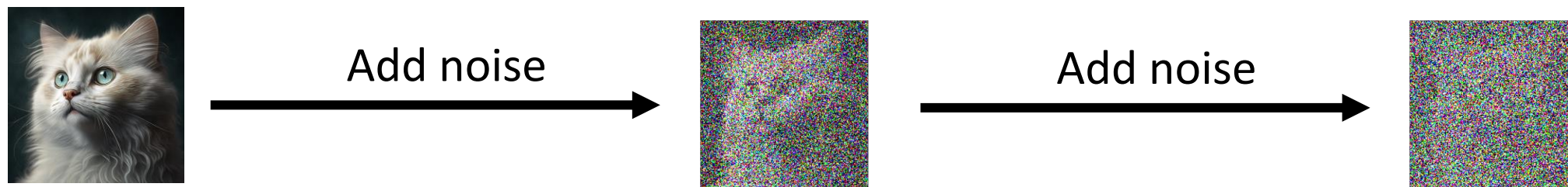


VAE

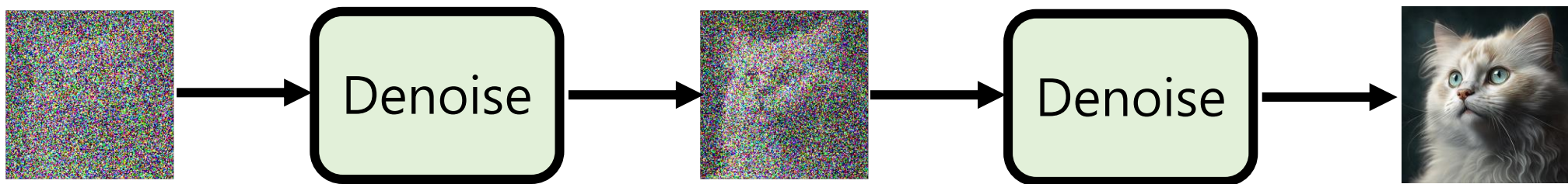


基本概念

Forward Process

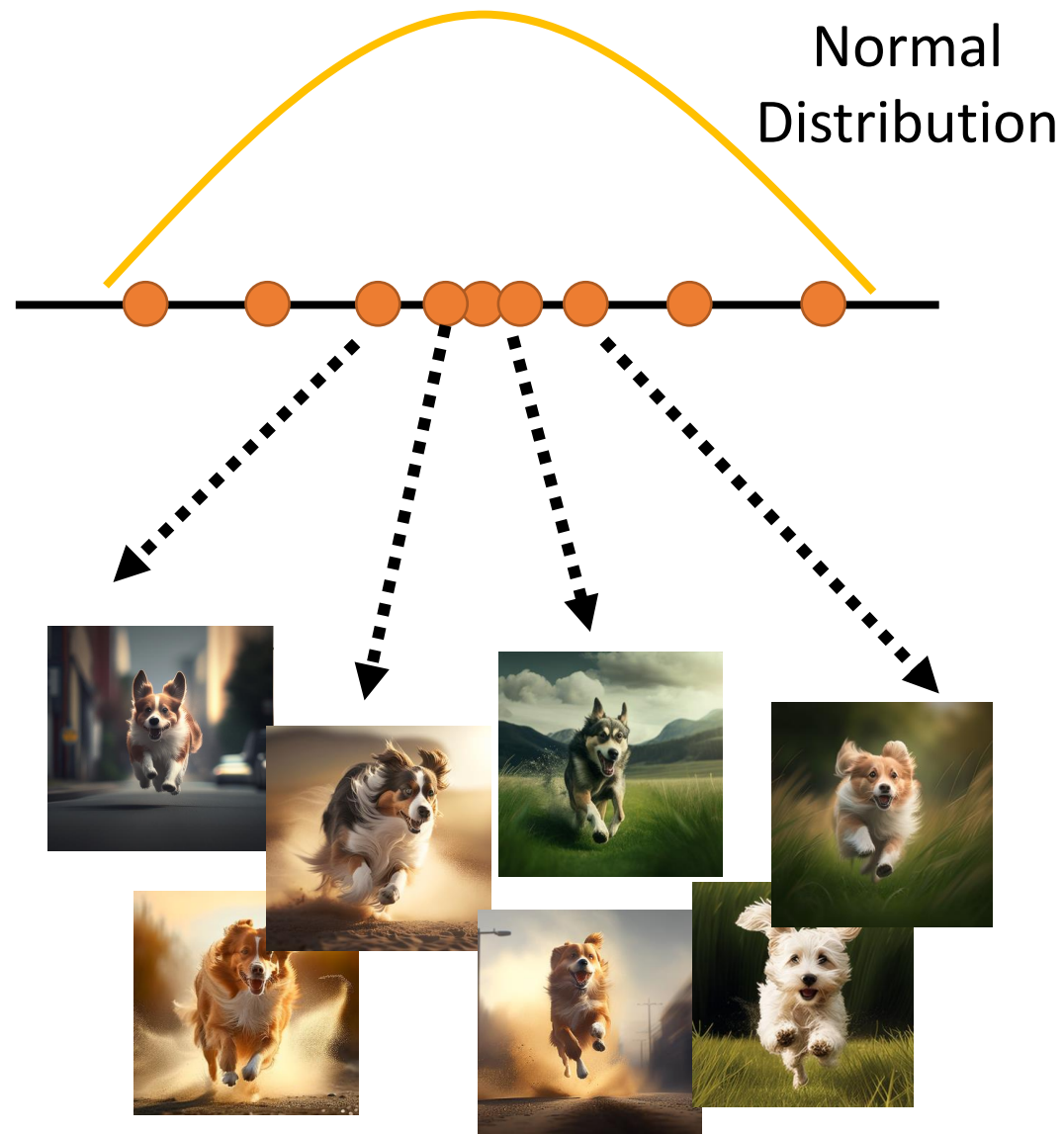
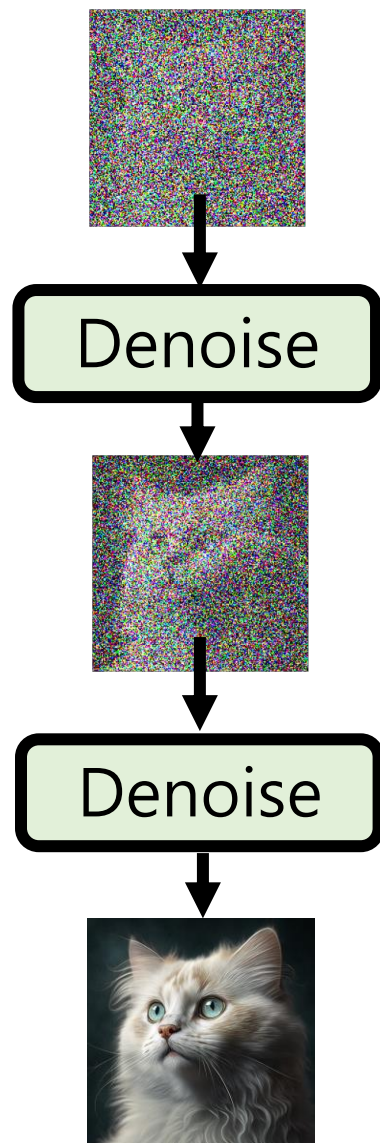
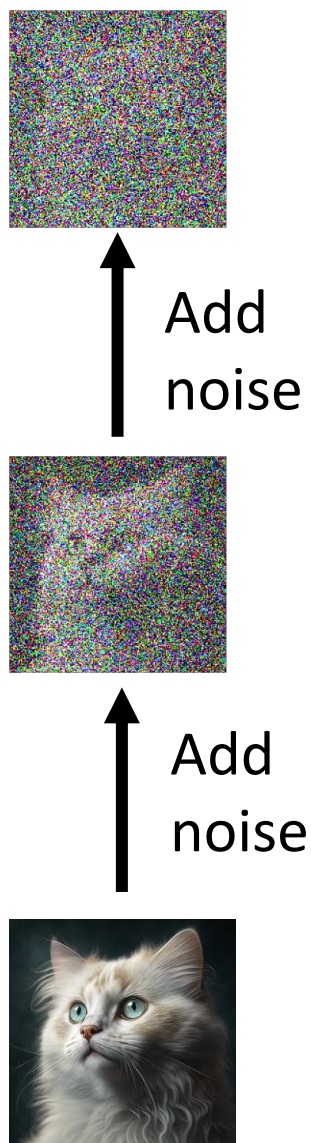


Reverse Process



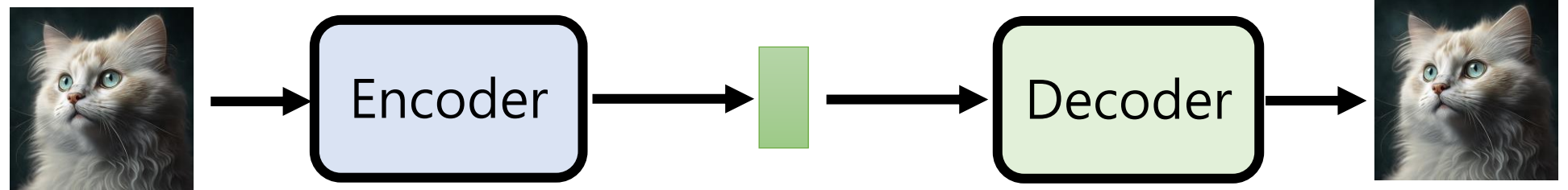
Diffusion

下一段課程會再細講

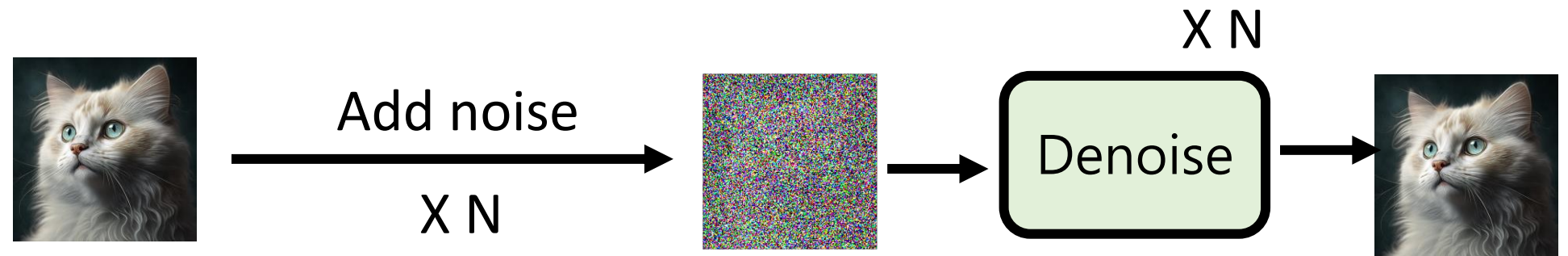


VAE vs. Diffusion Model

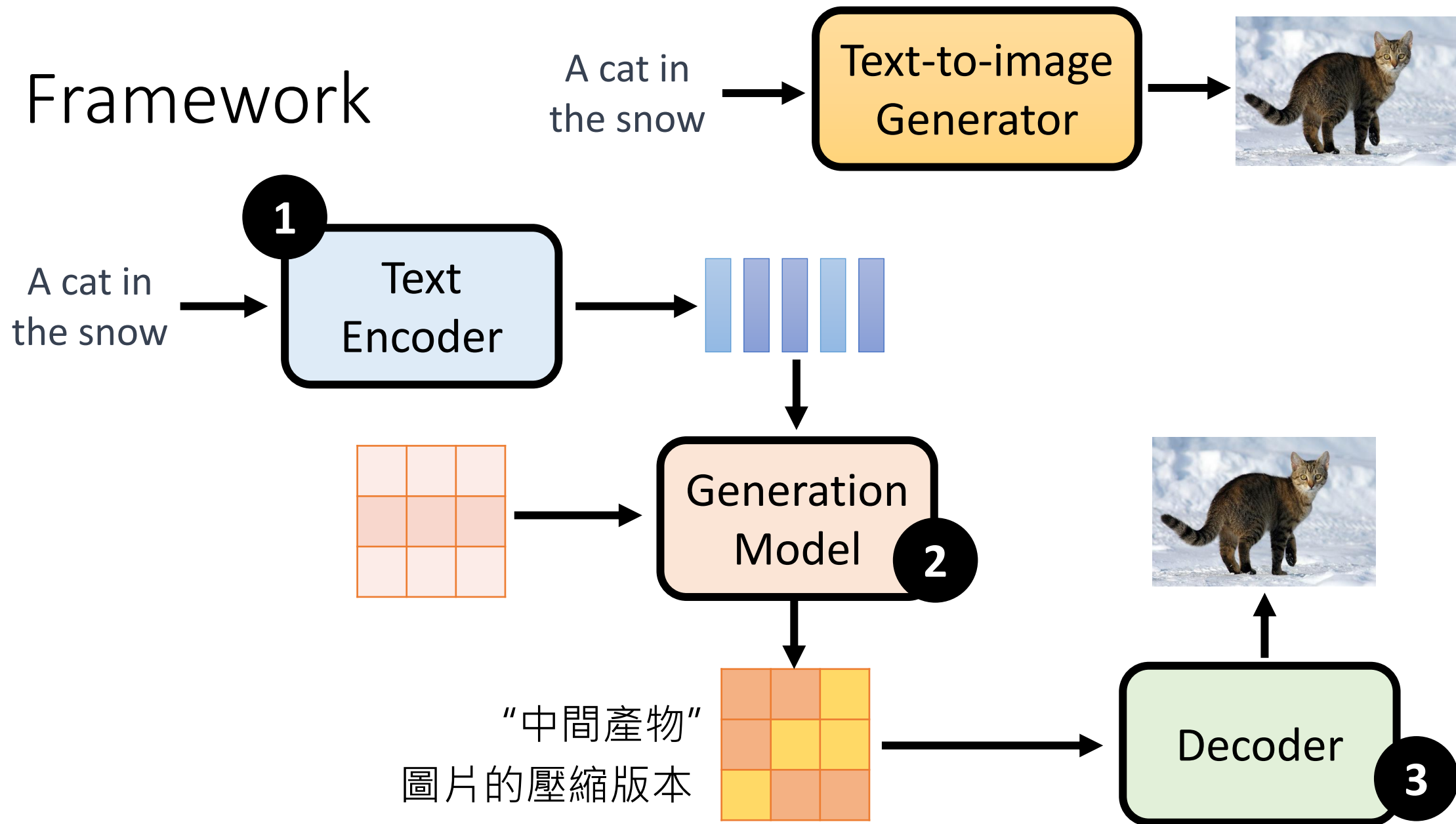
VAE



Diffusion

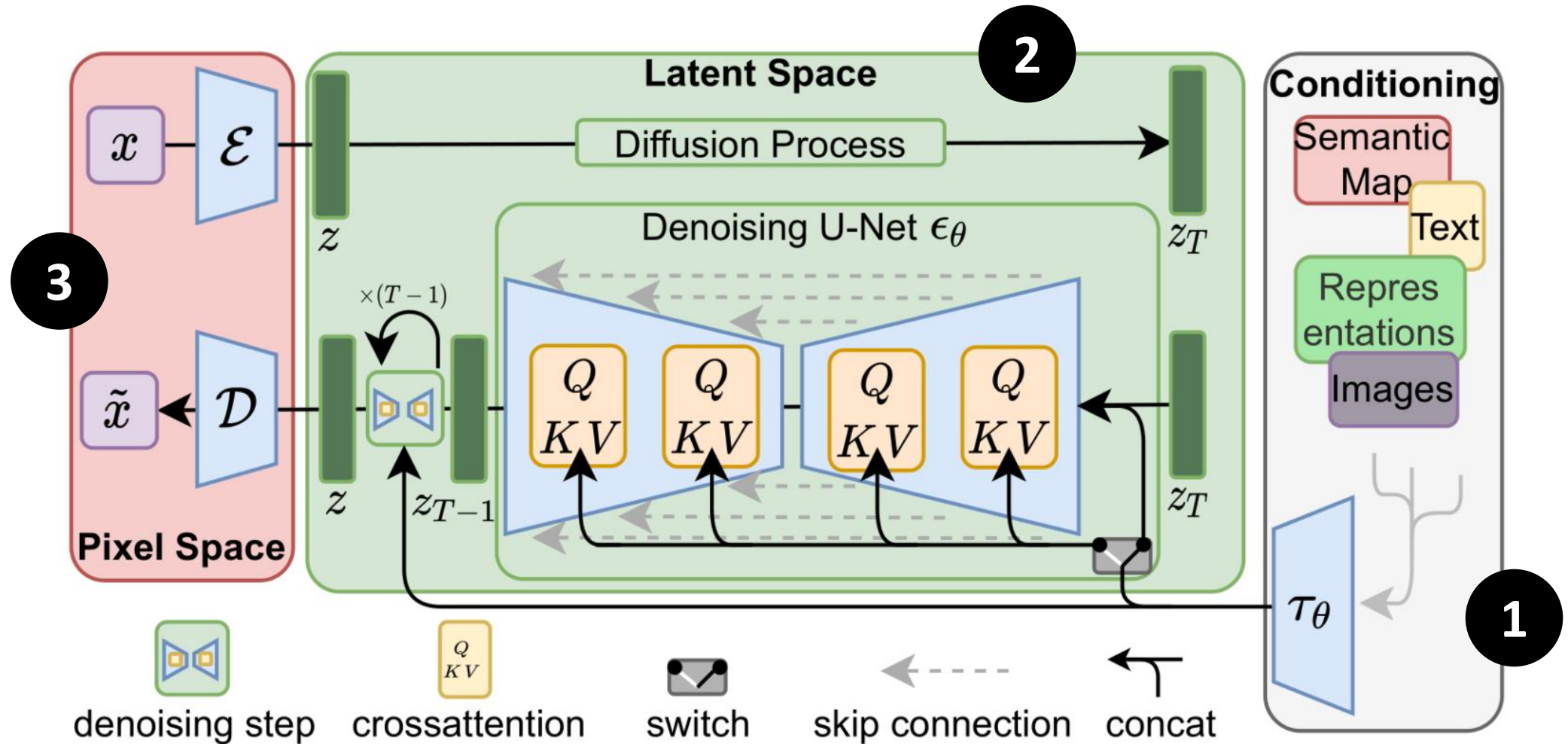


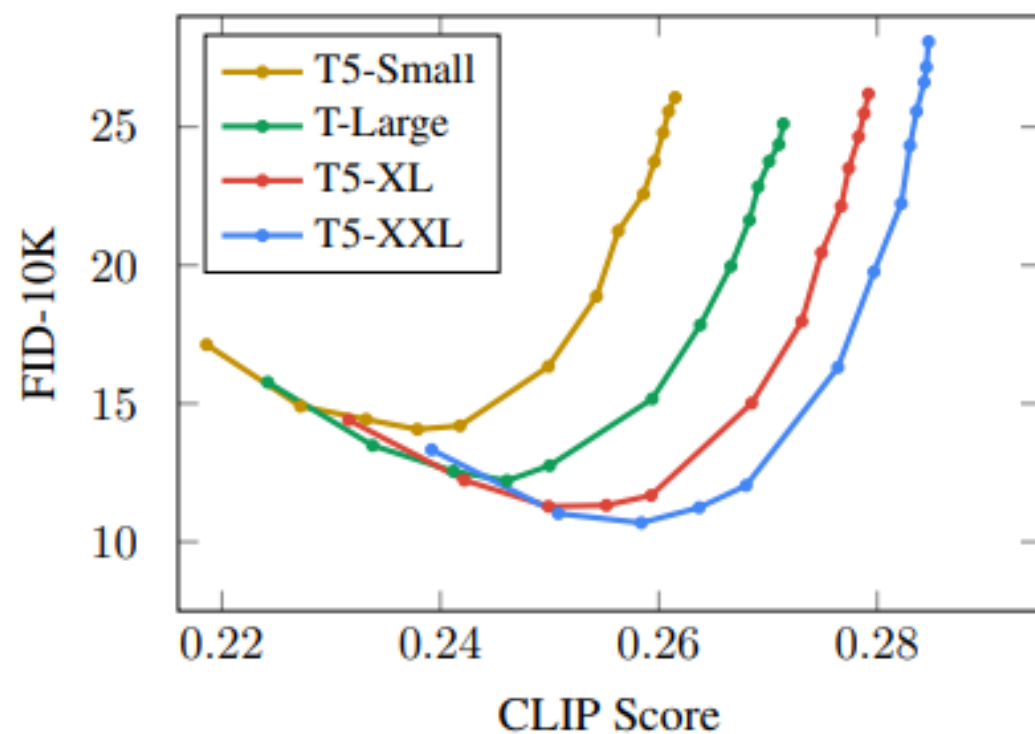
Framework



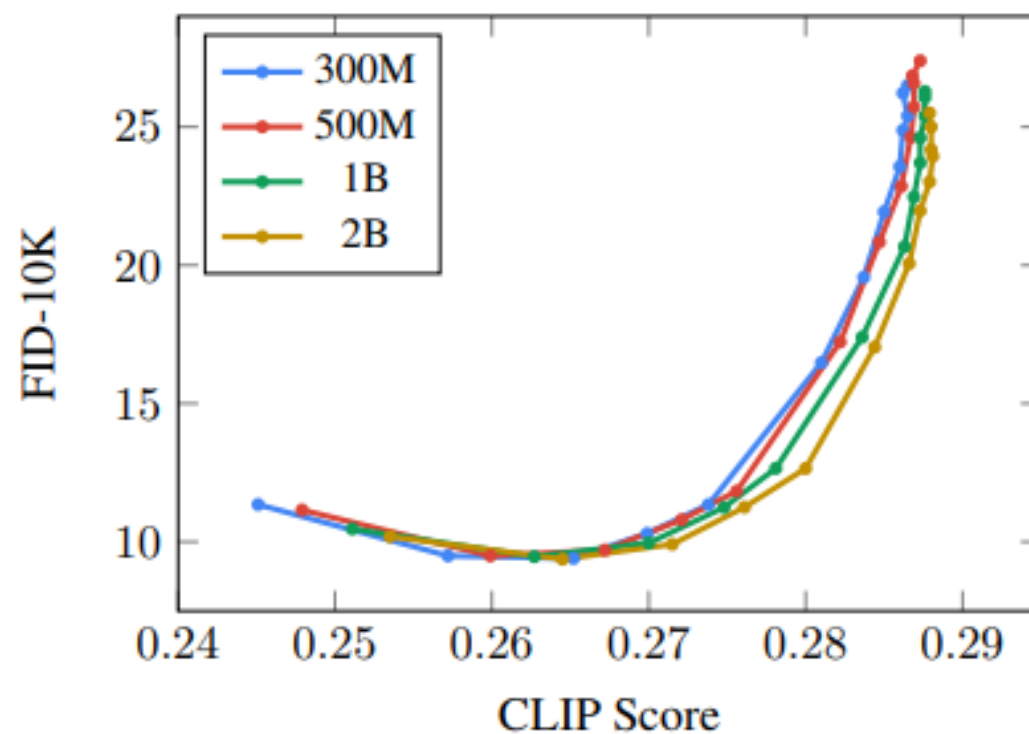
Stable Diffusion

<https://arxiv.org/abs/2112.10752>





(a) Impact of encoder size.

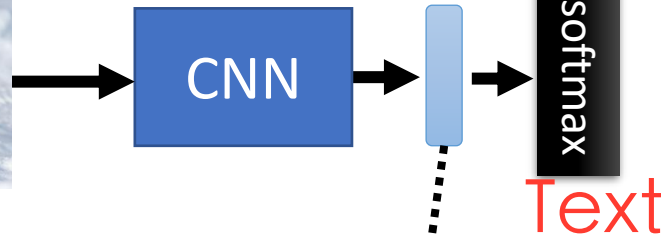


(b) Impact of U-Net size.

<https://arxiv.org/abs/2205.11487>

Fréchet Inception Distance (FID)

<https://arxiv.org/abs/1706.08500>



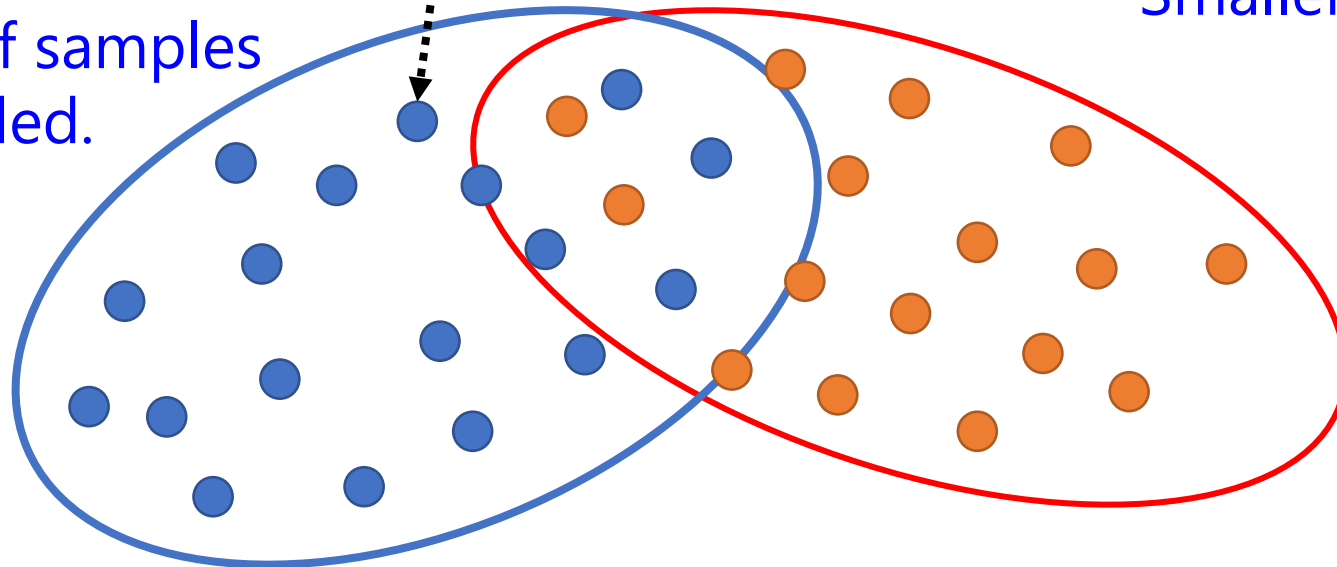
red points: real images

blue points: generated images

FID = Fréchet distance
between the two **Gaussians** ???

Smaller is better

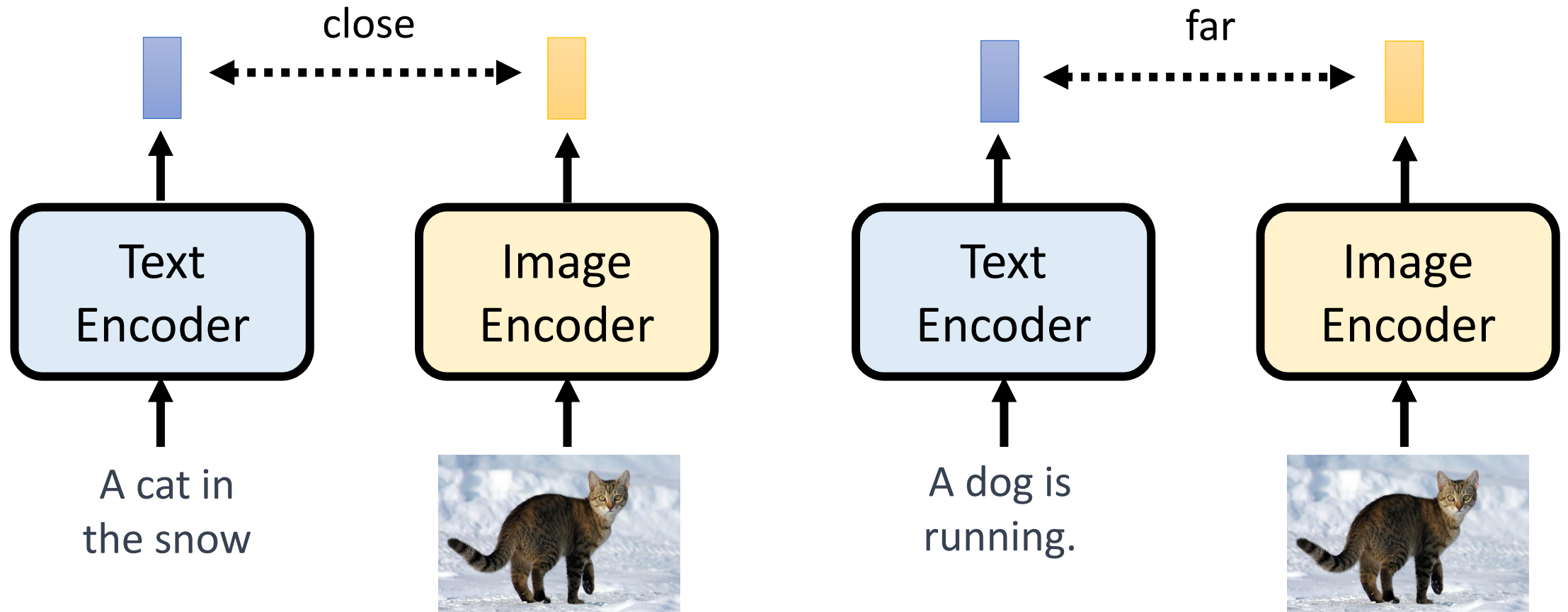
A lot of samples
is needed.



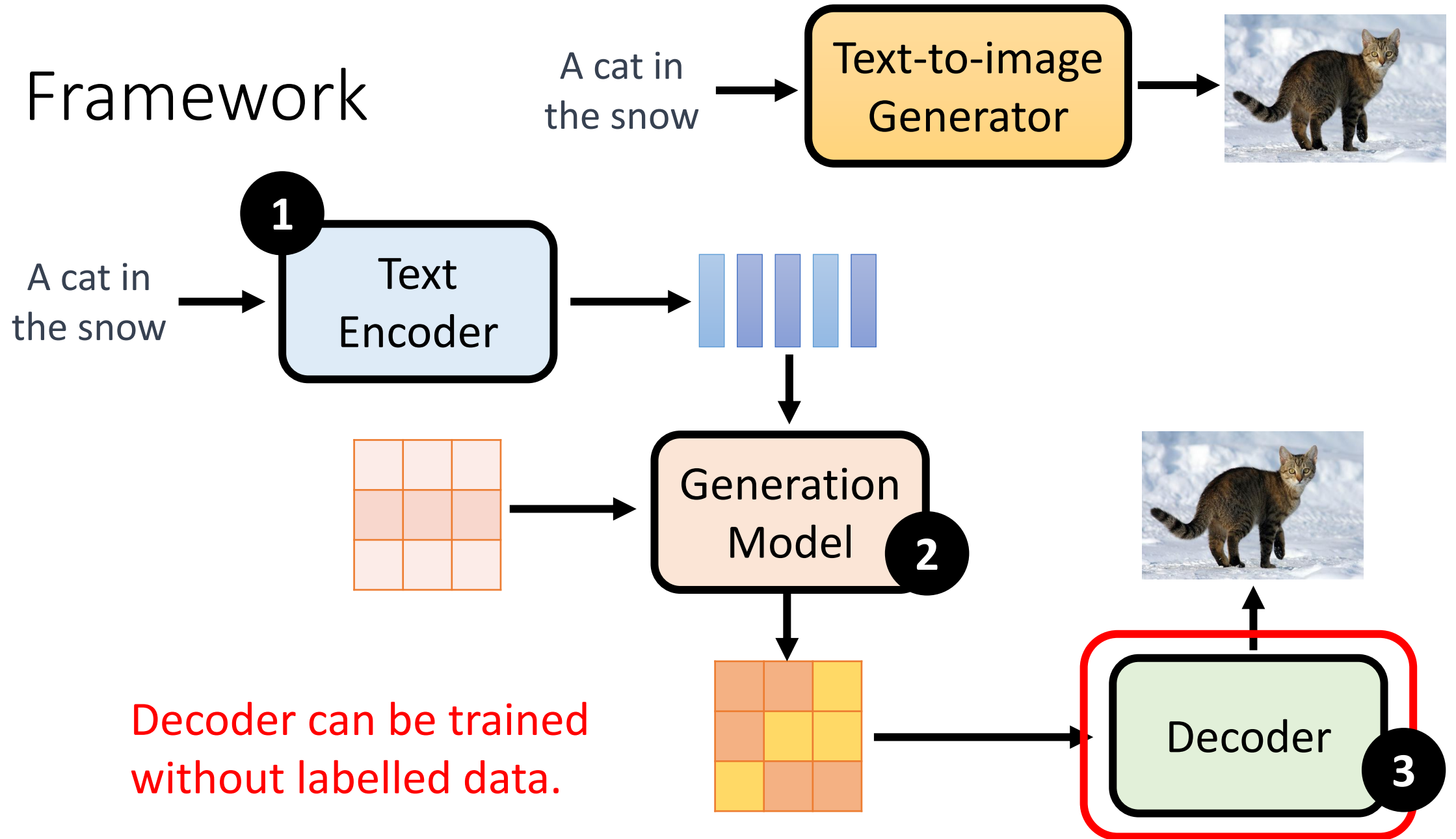
Contrastive Language-Image Pre-Training (CLIP)

<https://arxiv.org/abs/2103.00020>

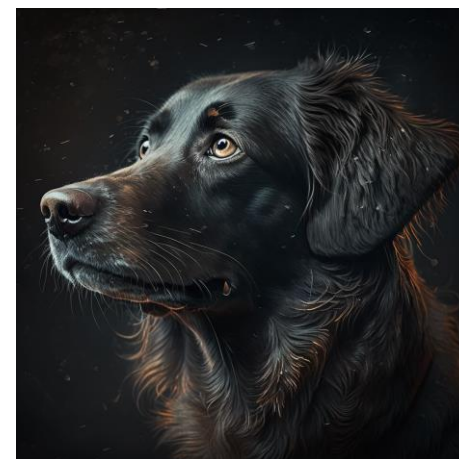
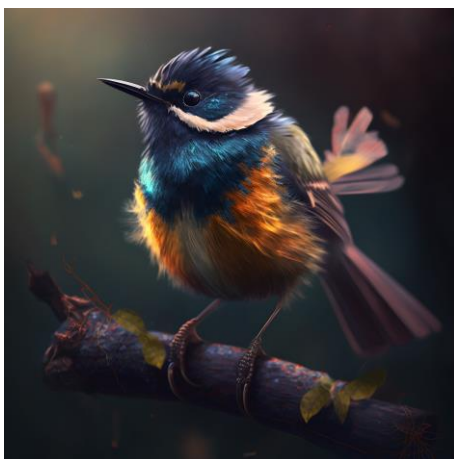
400 million image-text pairs



Framework



「中間產物」為小圖

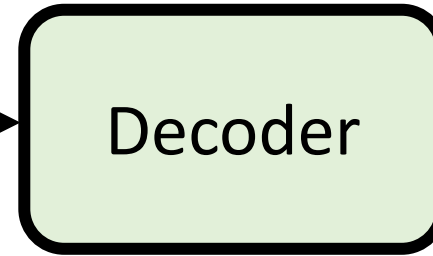
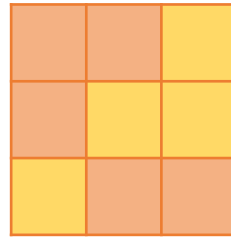


(Images are generated
by Midjourney)

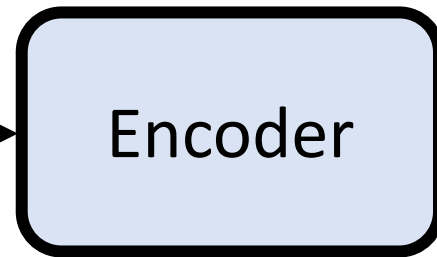
「中間產物」為「Latent Representation」

Auto-encoder

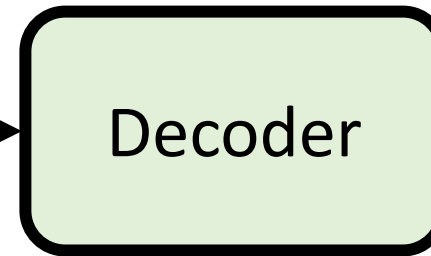
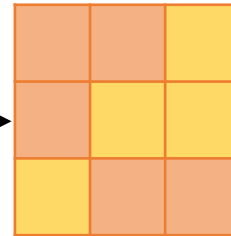
Latent
Representation



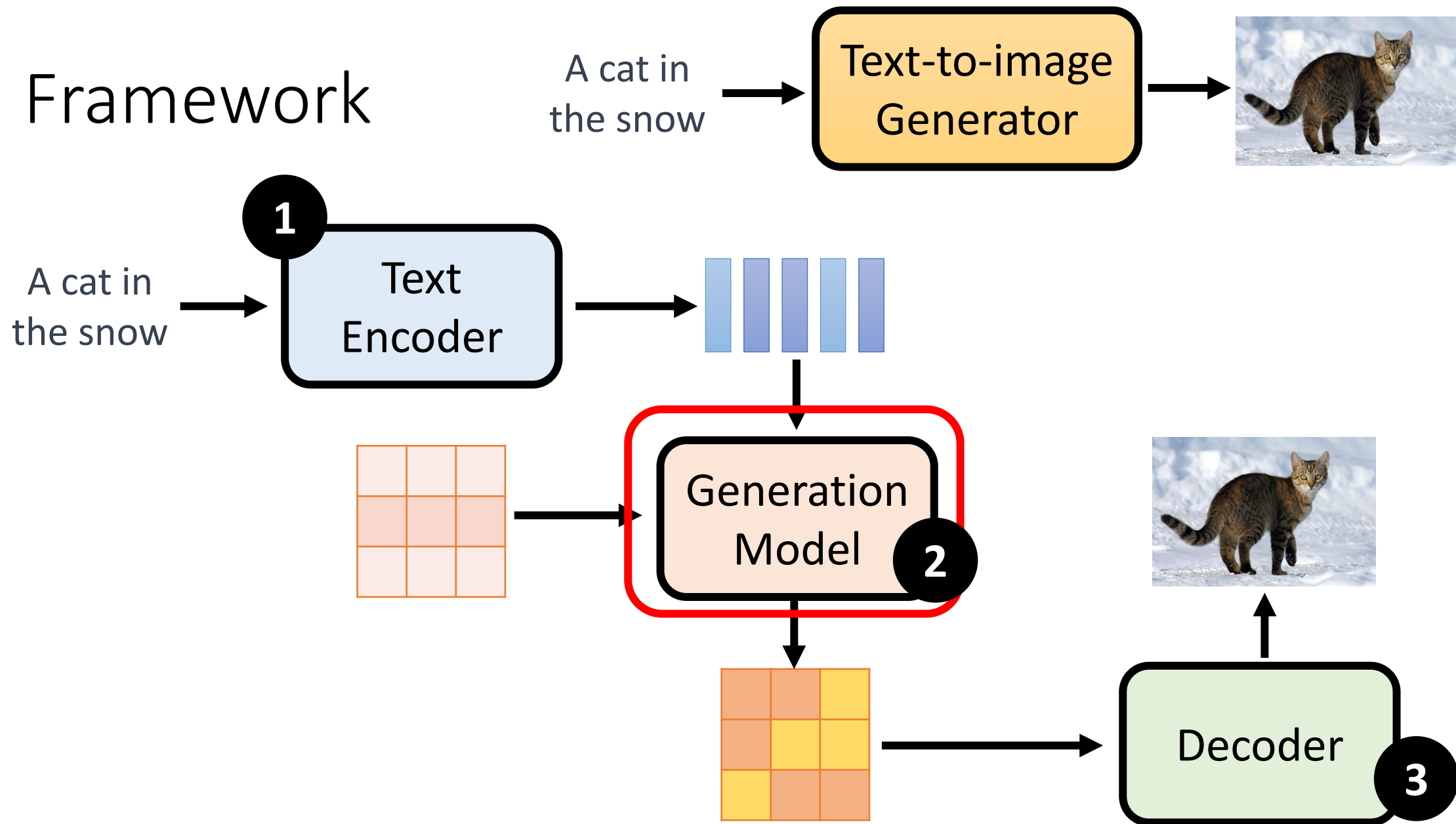
$H \times W \times 3$



$h \times w \times c$



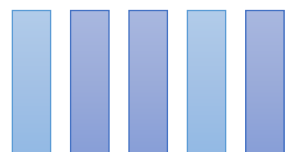
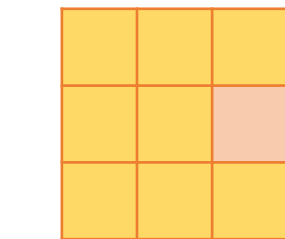
Framework



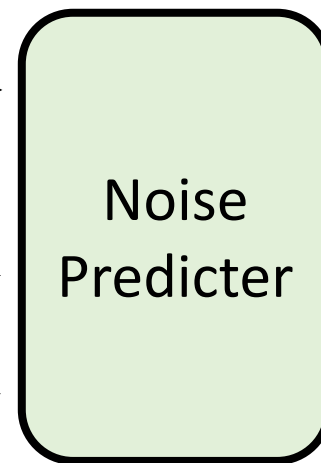
A cat in the snow



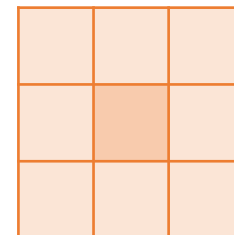
input



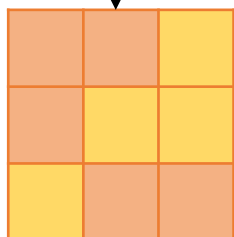
2



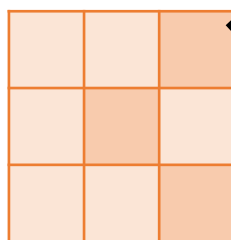
?????



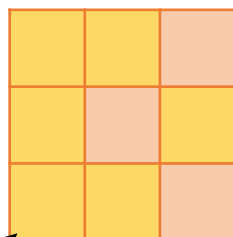
Encoder



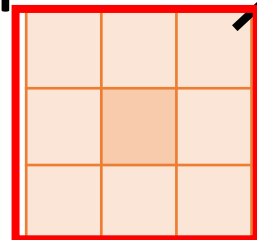
+



Step 1



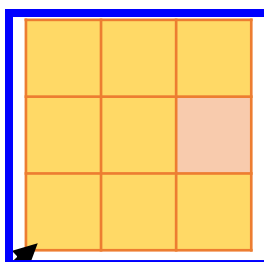
+



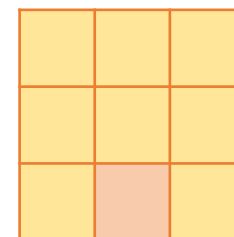
Step 2

ground truth

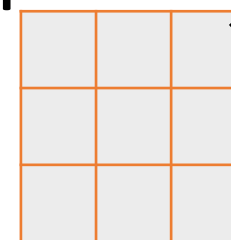
input



.....



+



Step 1000

