Stable Diffusion

Give it a text prompt. It will return an image matching the text.

- Kedar Kulkarni
SY AI promoter

Contents

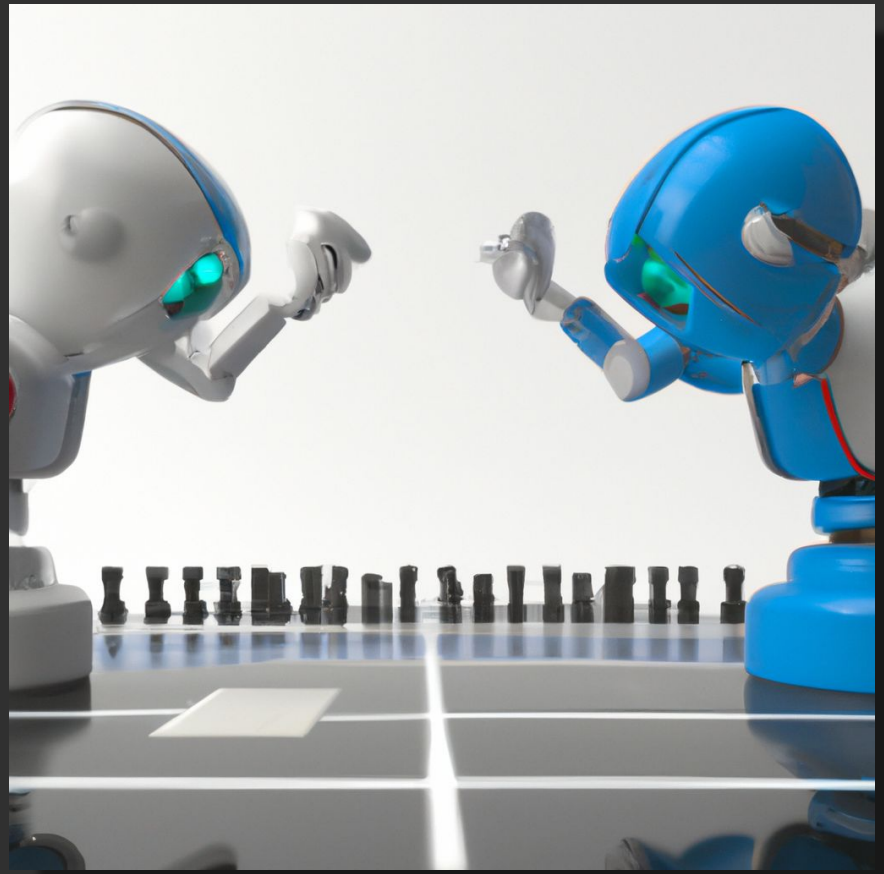
Introduction

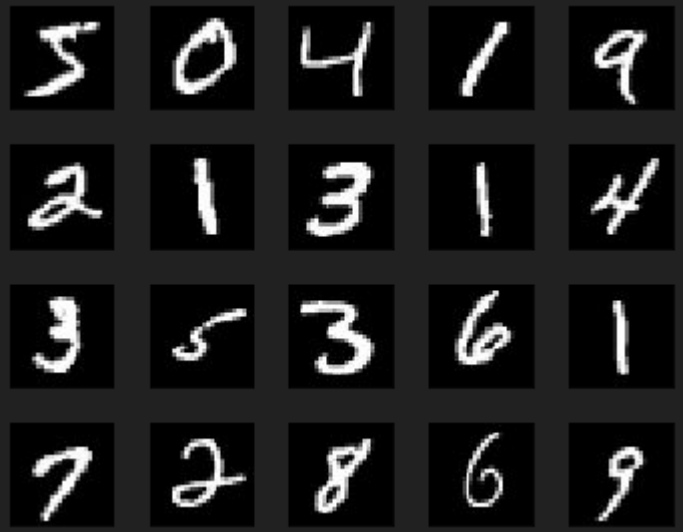
- Image generation is the creation of artificially generated images that look as realistic as real images.

- These images can be created by Generation Adversarial Networks(GAN) or with Variational Autoencoders, and more recently, Vector Quantized Variational Autoencoders (VQ-VAE), which create a discrete latent representation and create more variety of images and is easier to train compared to GANs.

GANs

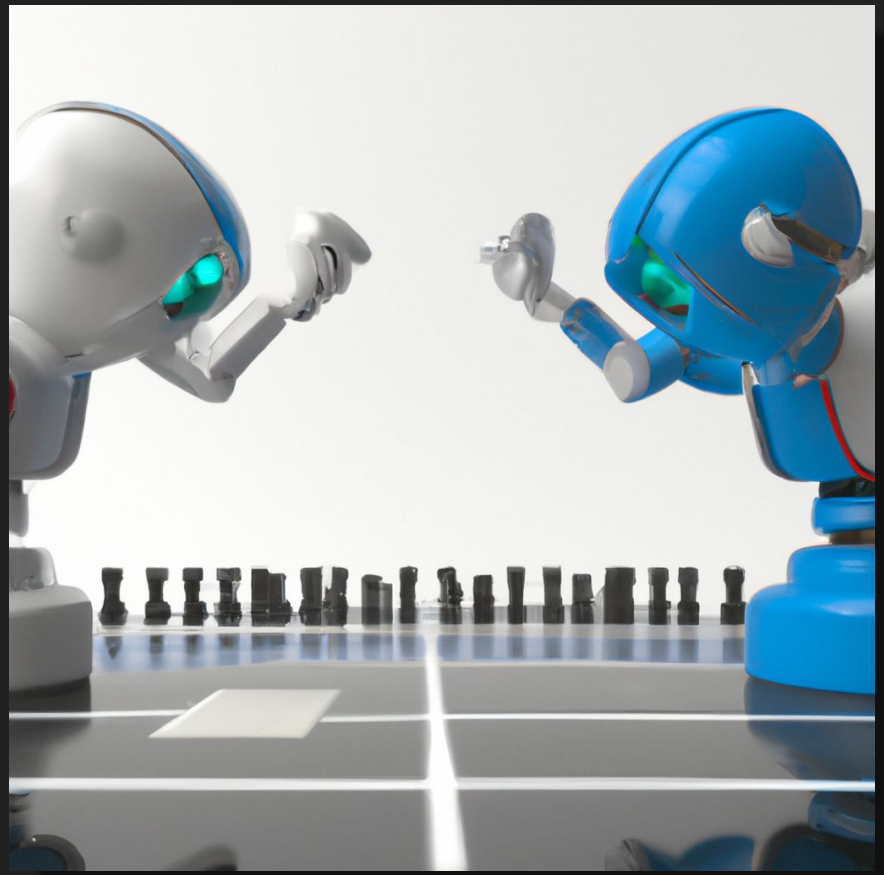
Generative Adversarial Networks



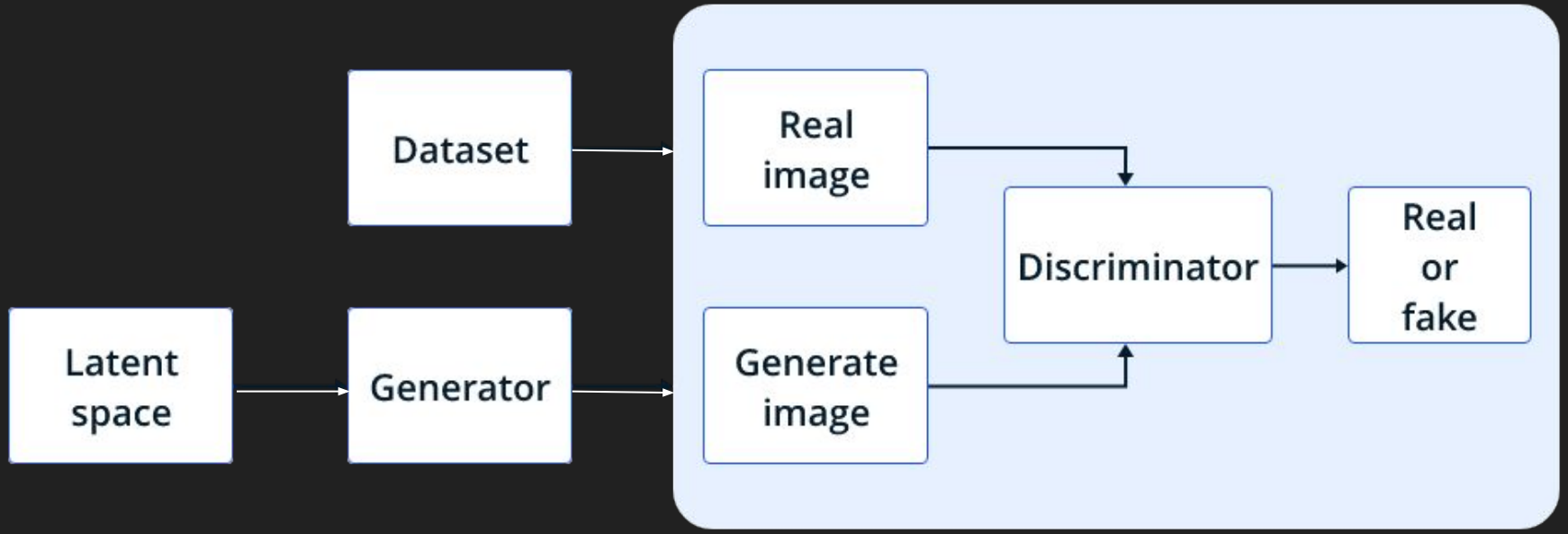


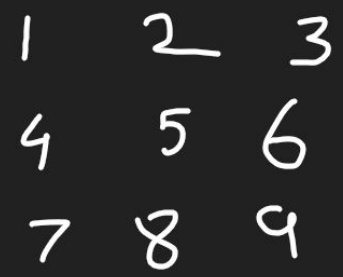
Dataset of labelled hand-drawn numbers

1 ⟷ 7

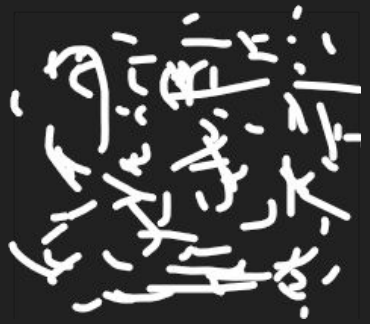


GAN Architecture





[Real Dataset]



[Noise]

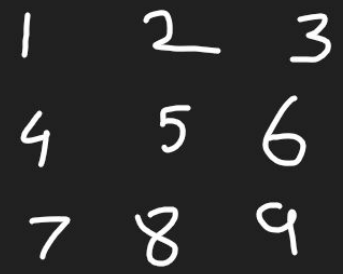
Generator

[Generator Output]

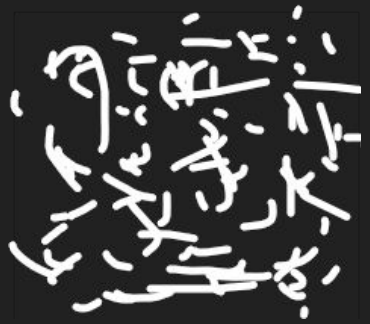
Discriminator

1
[REAL]

0
[FAKE]



[Real Dataset]



[Noise]

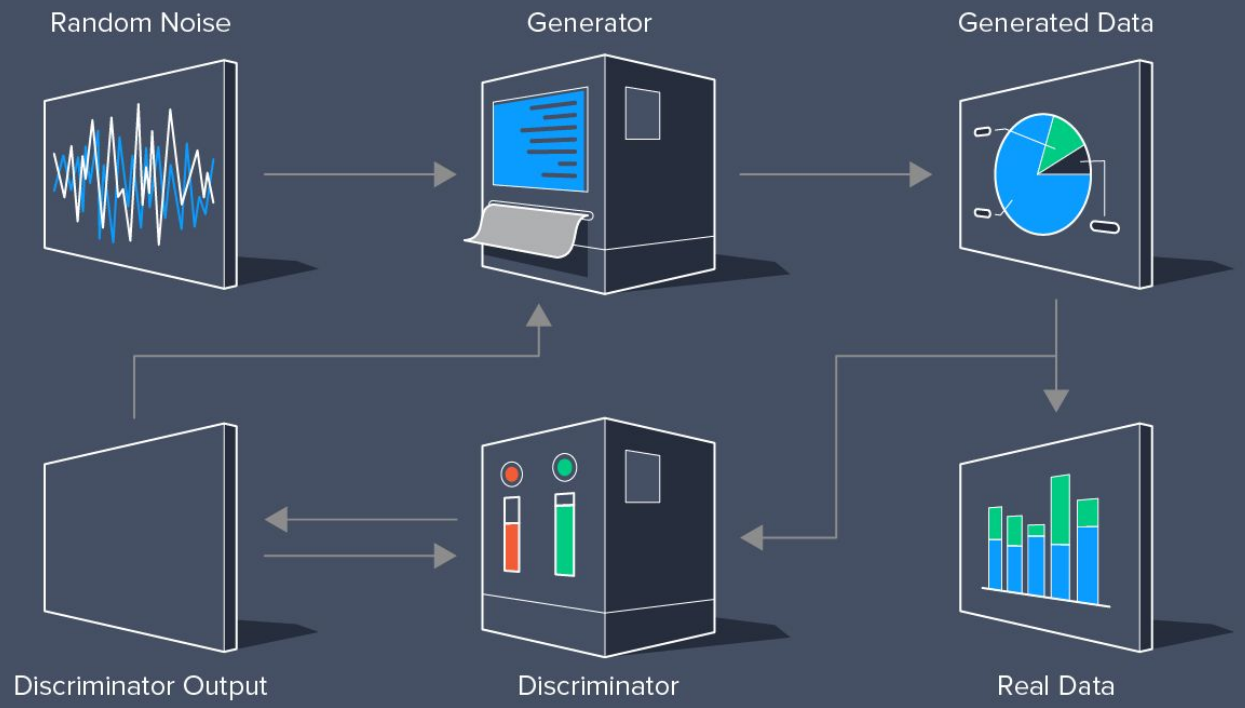
Generator

[Generator Output]

Discriminator

1
[REAL]

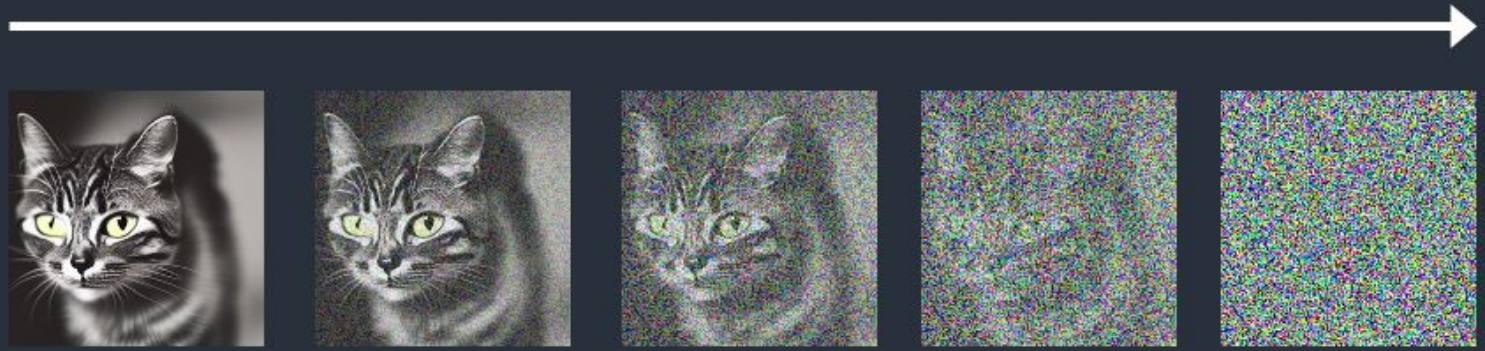
0.9
[FAKE]


Random Noise
Generator
Generated Data
Discriminator Output
Discriminator
Real Data

How Diffusion Models Work ?

Forward and Reverse Diffusion

Forward diffusion



Step

1 2 3 4

Noise 1 Noise 2 Noise 3 Noise 4

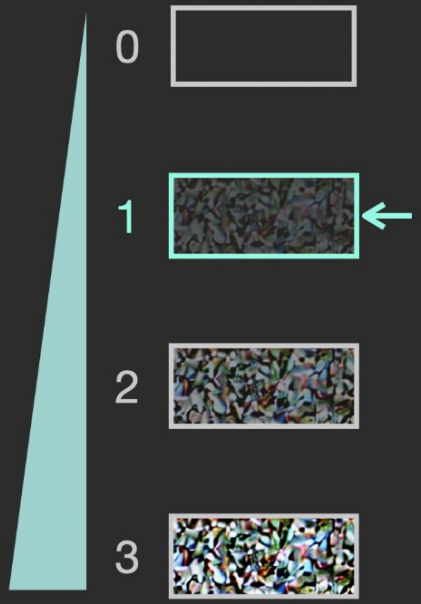
Training examples are created by generating noise and adding an amount of it to the images in the training dataset (forward diffusion)

1
Pick an image

2
Generate some random noise

Noise sample 1

3
Pick an amount of noise



4
Add noise to the image in that amount

DATASET

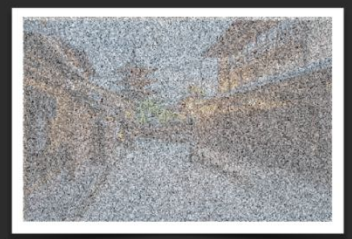
MODEL

INPUT		OUTPUT / LABEL
Noise Amount	Noisy Image	Noise sample
3		
14		
7		
42		
2		
21		

Noise Predictor
(UNet)

UNet training step

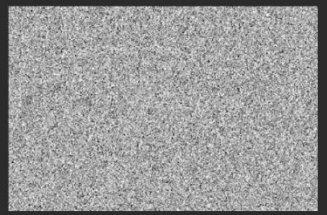
1. Pick a training example from the training dataset
2. Predict the noise
3. Compare to actual noise (calculate loss)



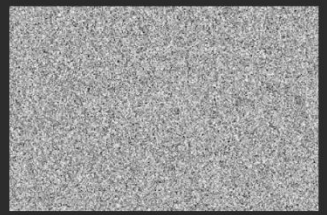
Noise amount: 3

UNet

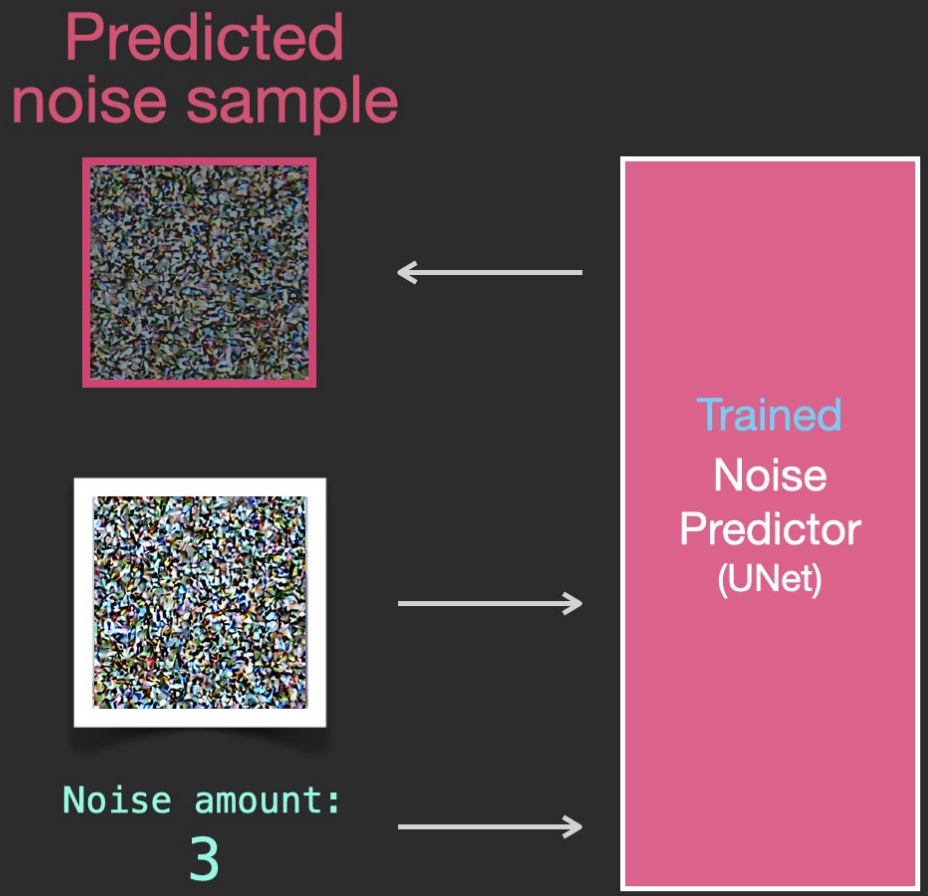
Unet Prediction



Actual noise (Label)



4. Update model (backprop)


Predicted
noise sample
Trained
Noise
Predictor
(UNet)
Noise amount:
3

Reverse Diffusion (Denoising)
Step 1

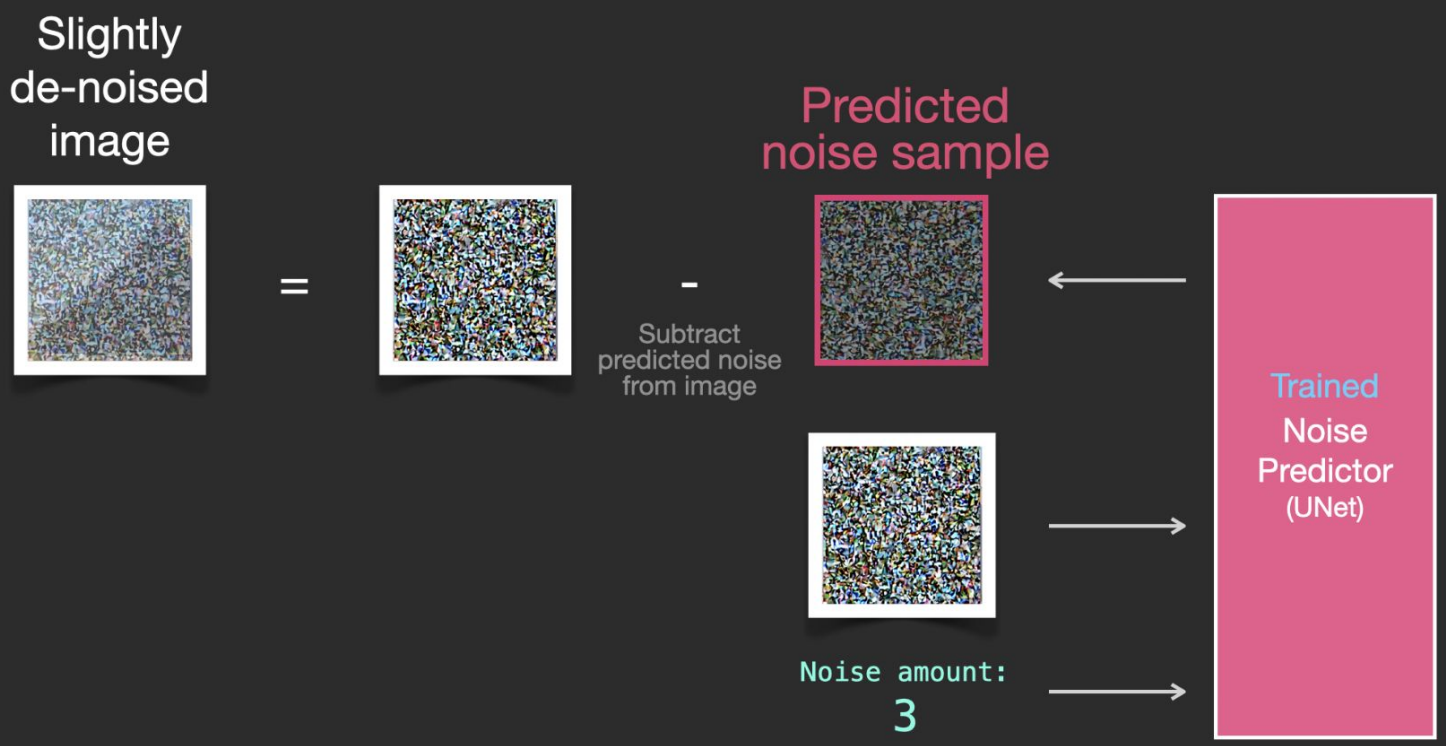
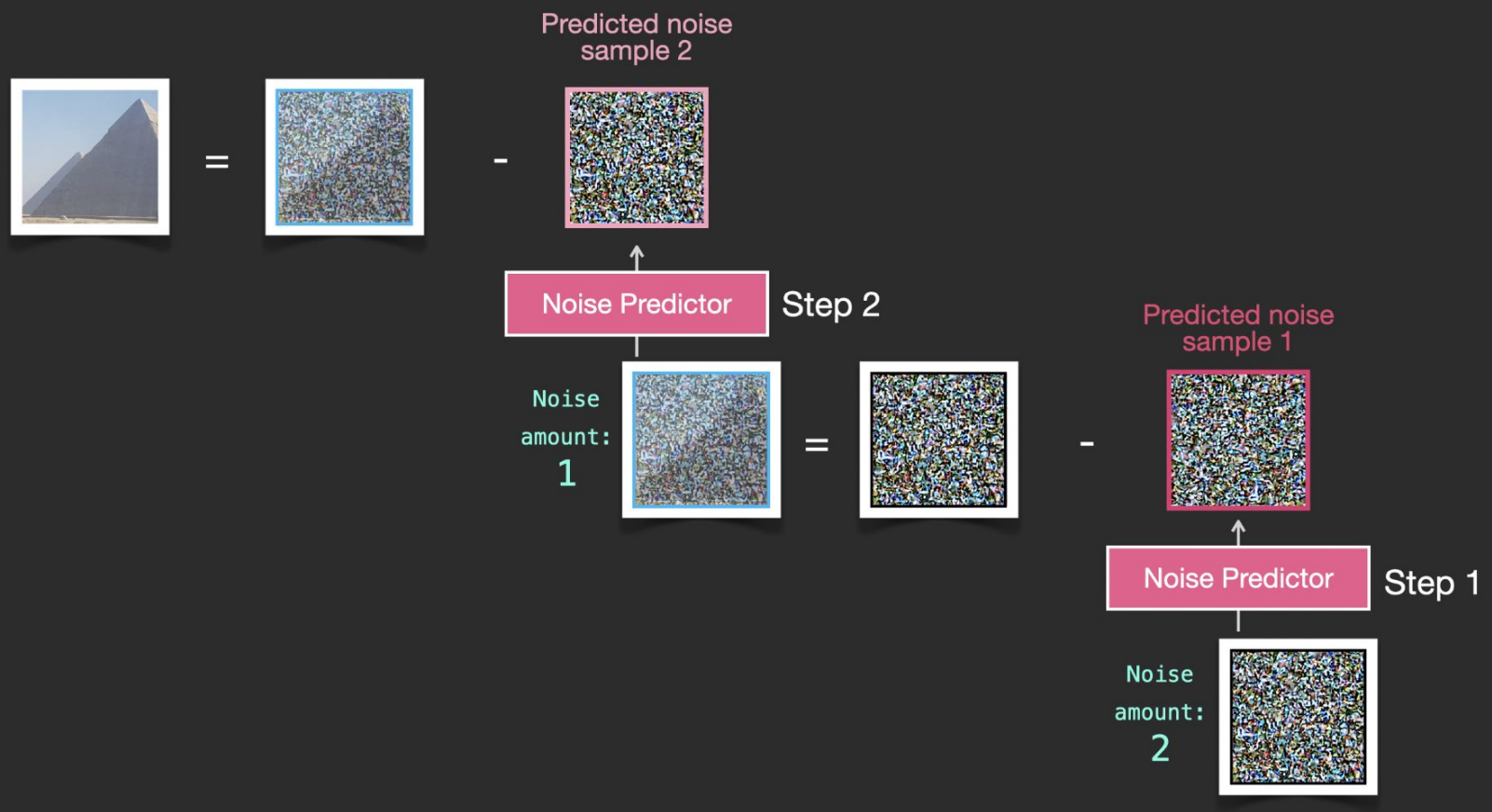
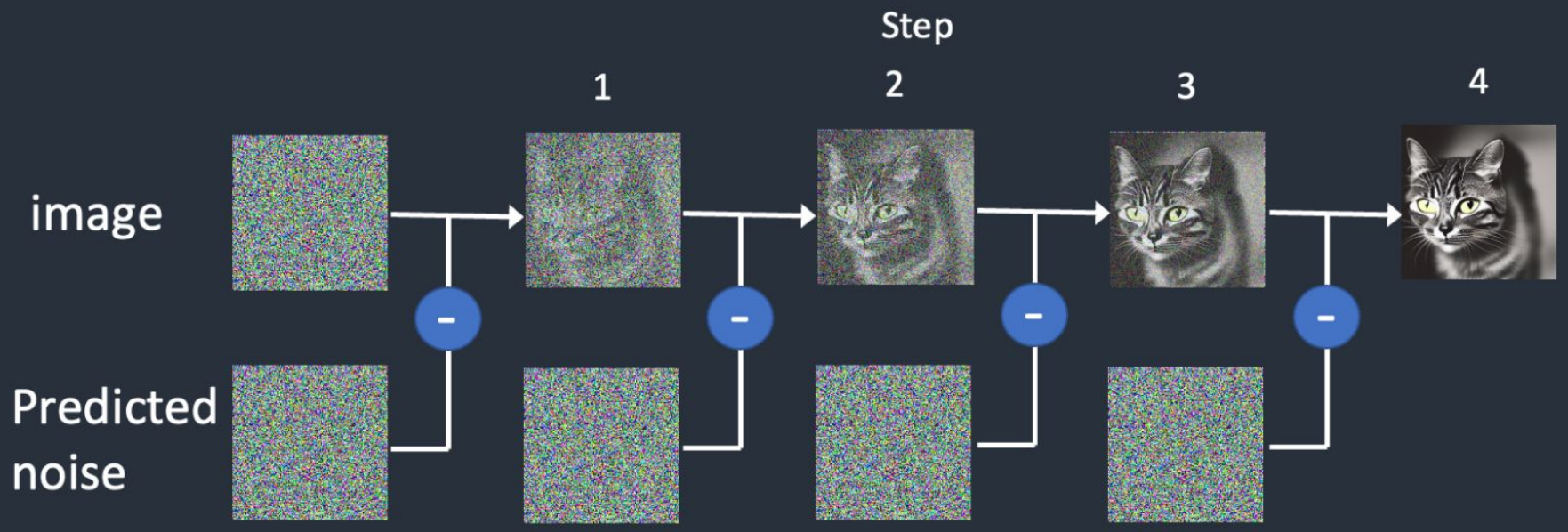


Image Generation by Reverse Diffusion (Denoising)

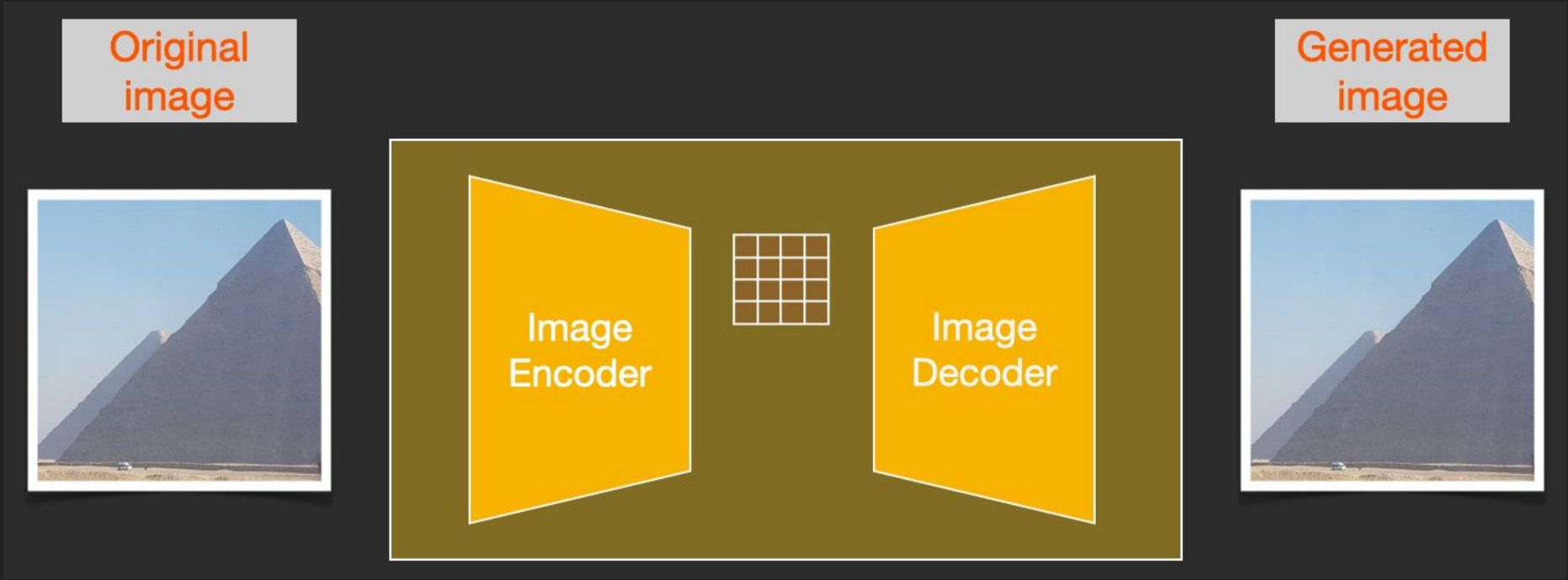


Reverse Diffusion



Reverse diffusion works by subtracting the predicted noise from the image successively.

Departure to Latent Space



Latent Space

512 x 512 x 3 = 786,432 !

[Pixel Space]

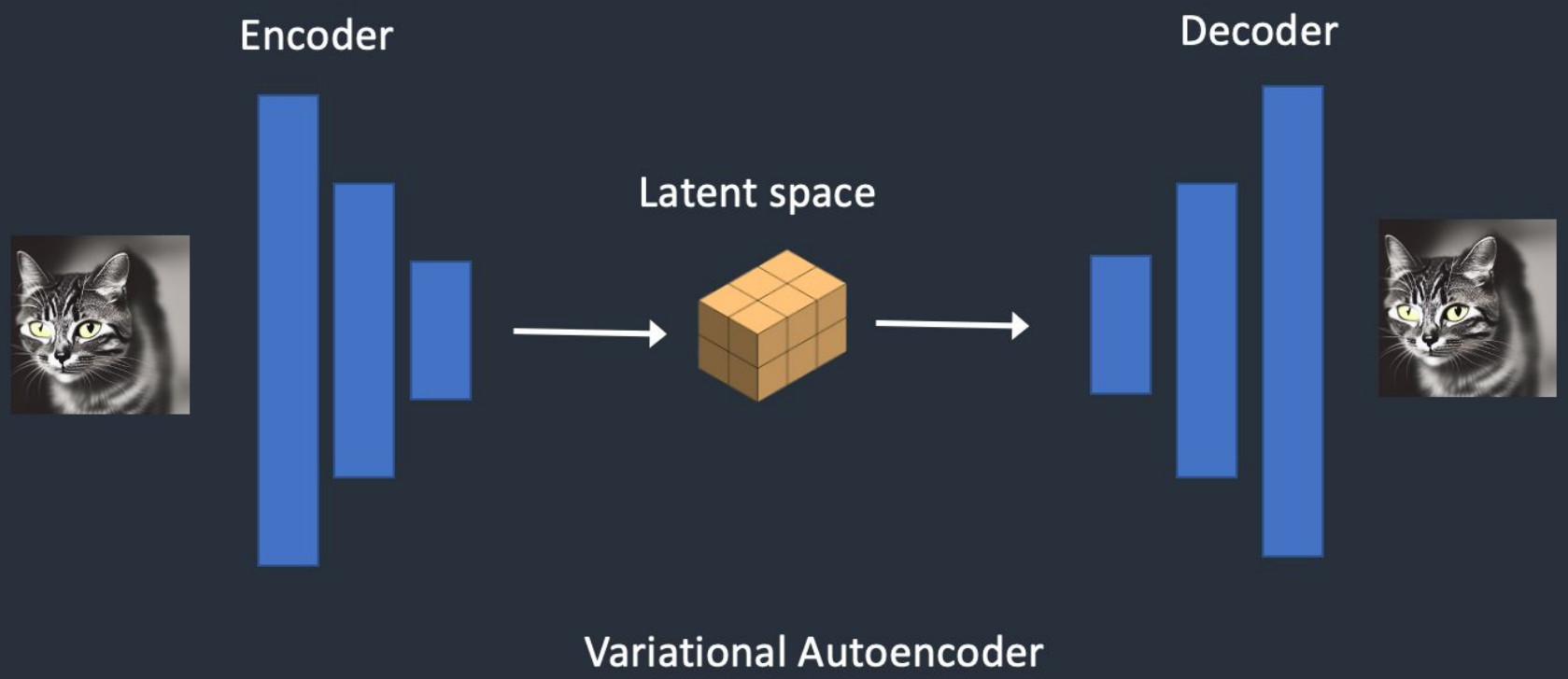
VAE encoder

48 times smaller !

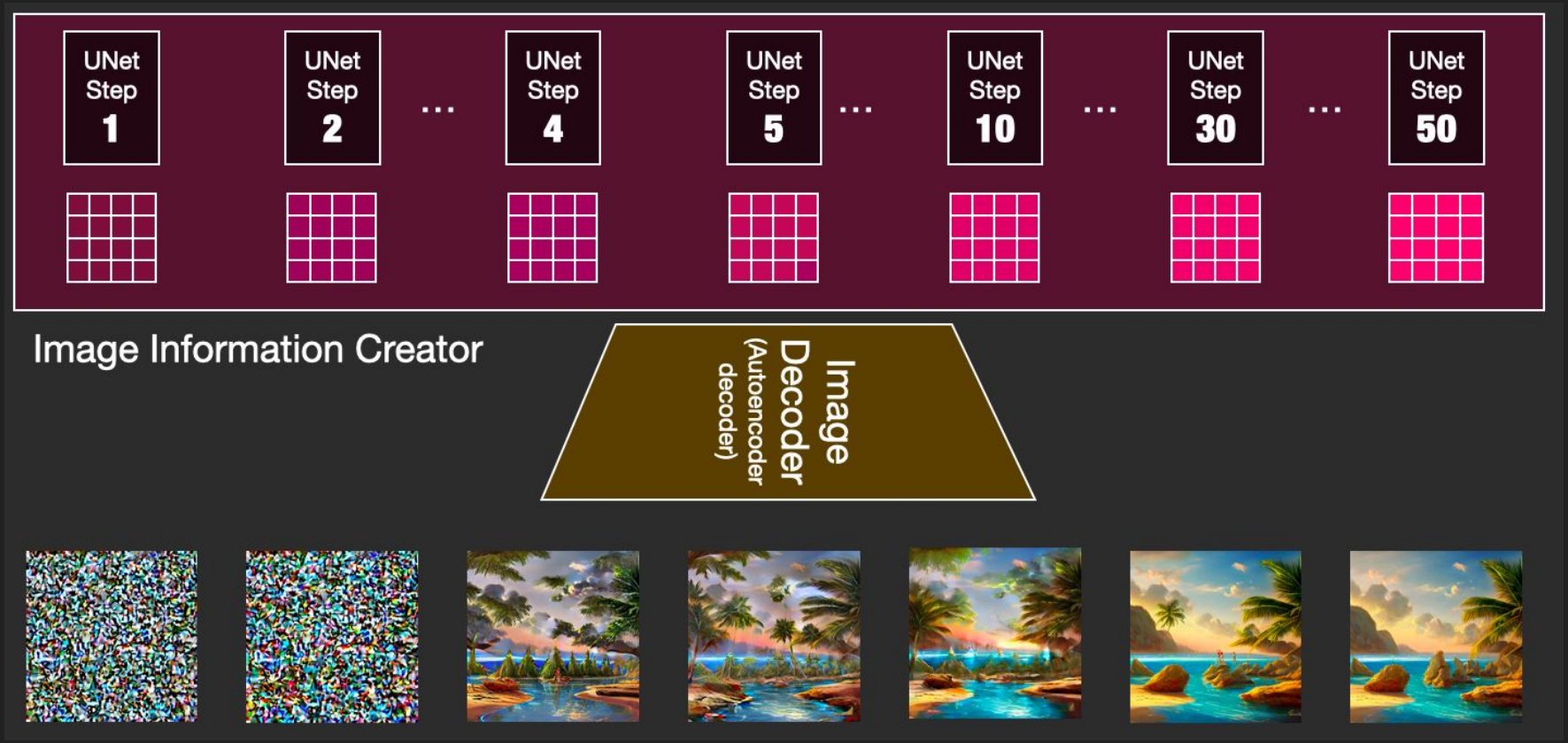
64 x 64 x 4 = 16,384

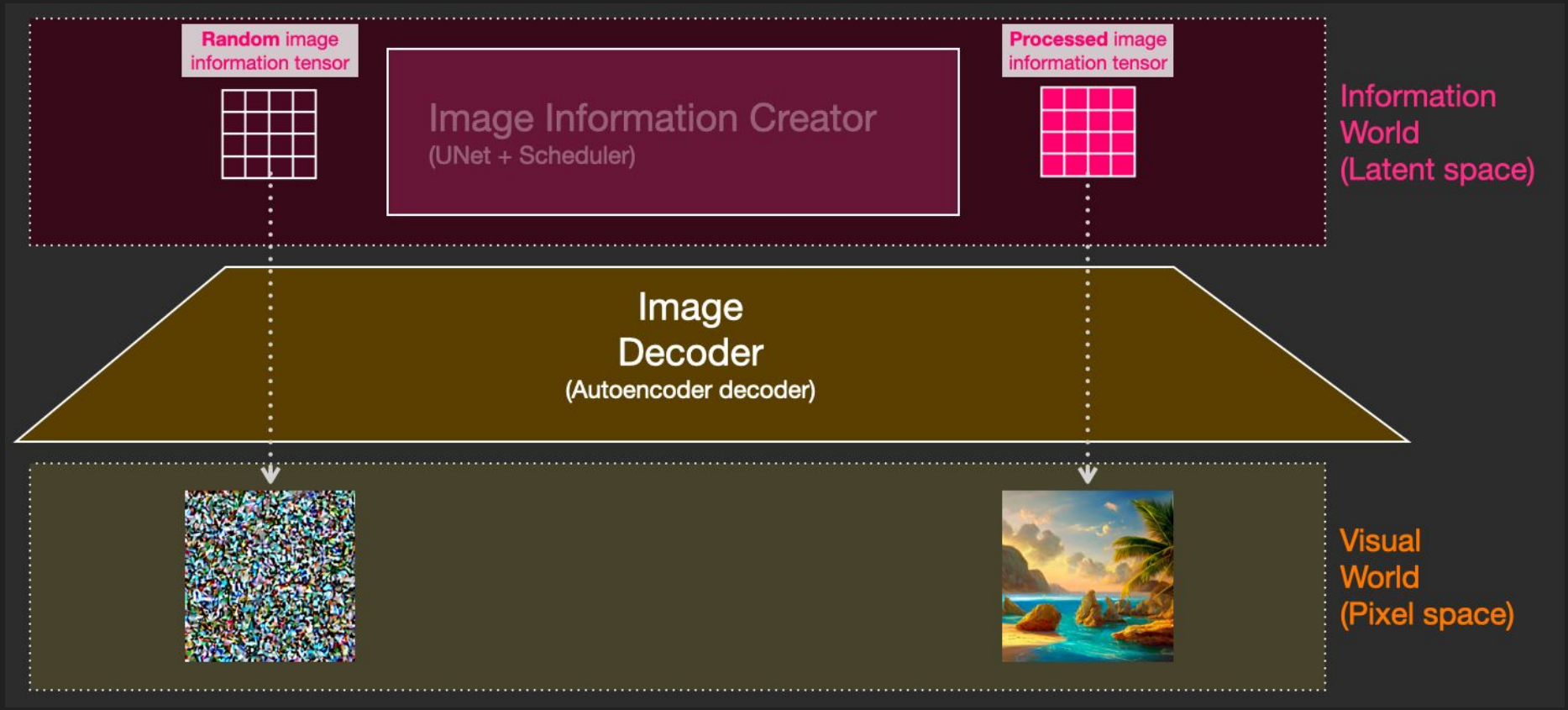
[Latent Space]

Variational AutoEncoder (VAE)

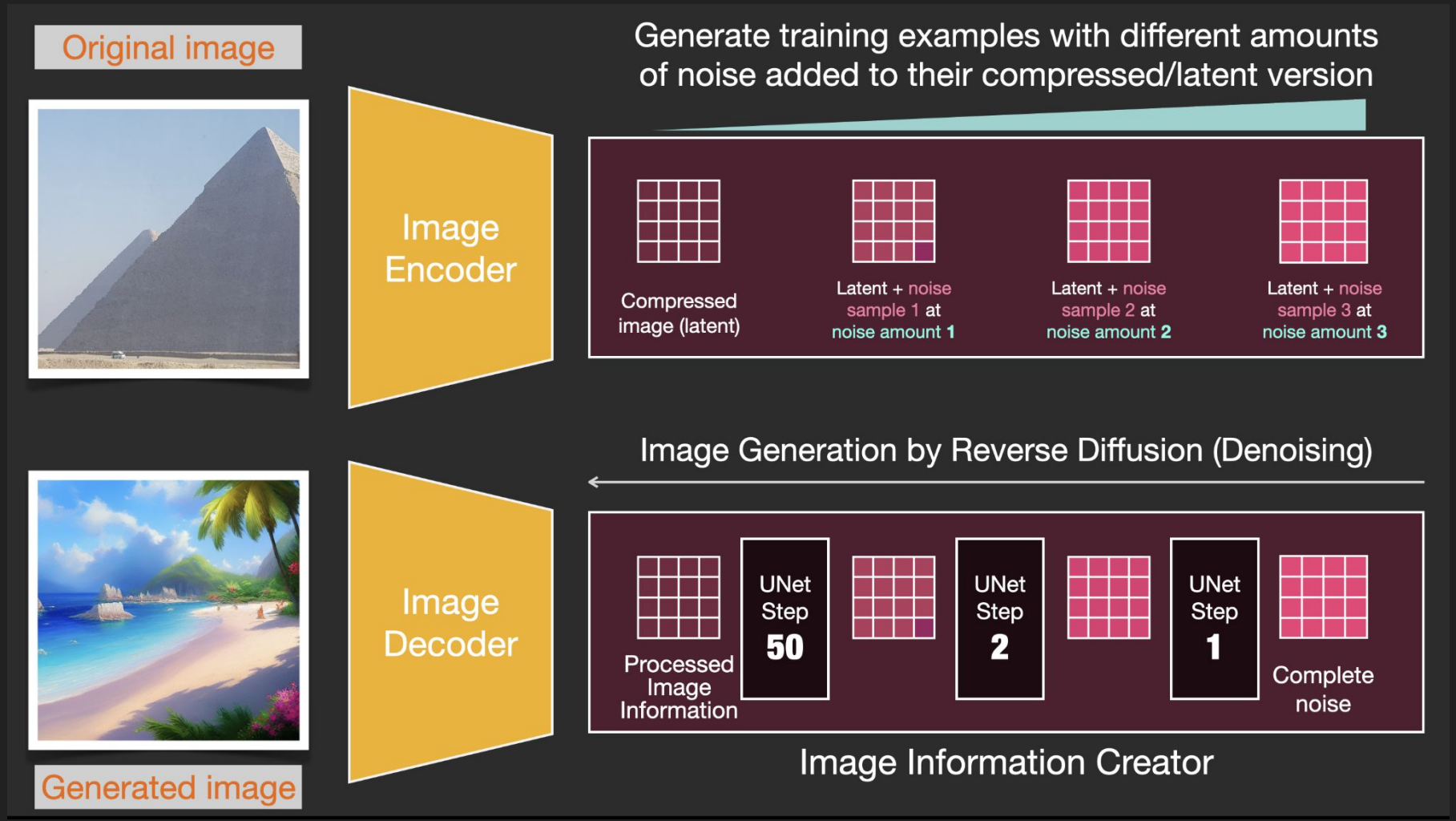


Latent Space




Random image information tensor
Image Information Creator
(UNet + Scheduler)
Processed image information tensor
Information World (Latent space)
Image Decoder
(Autoencoder decoder)
Visual World (Pixel space)

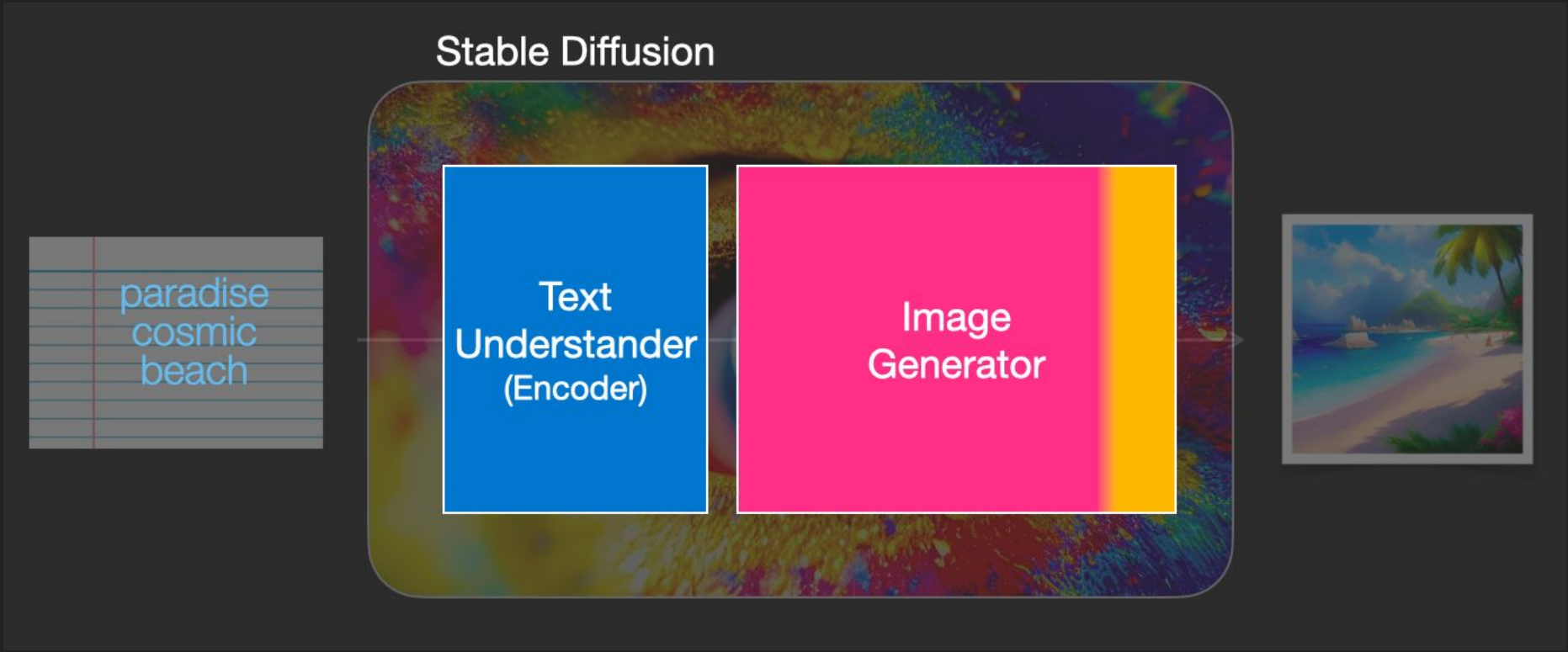
Quick Recap


Original image
Generate training examples with different amounts of noise added to their compressed/latent version
Image Encoder
Compressed image (latent)
Latent + noise sample 1 at noise amount 1
Latent + noise sample 2 at noise amount 2
Latent + noise sample 3 at noise amount 3
Image Generation by Reverse Diffusion (Denoising)
Image Decoder
Processed Image Information
UNet Step 50
UNet Step 2
UNet Step 1
Complete noise
Image Information Creator
Generated image

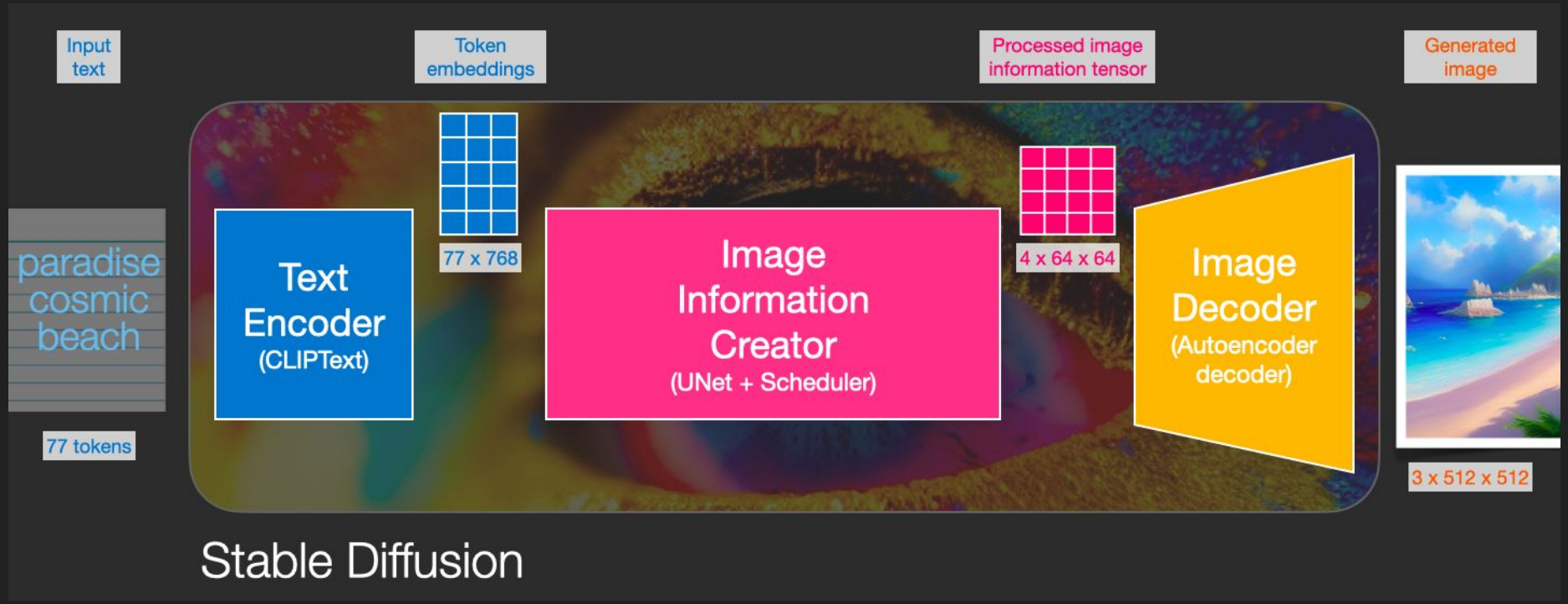
OK but when does the text prompt come into the picture ?

[Pun Intended]

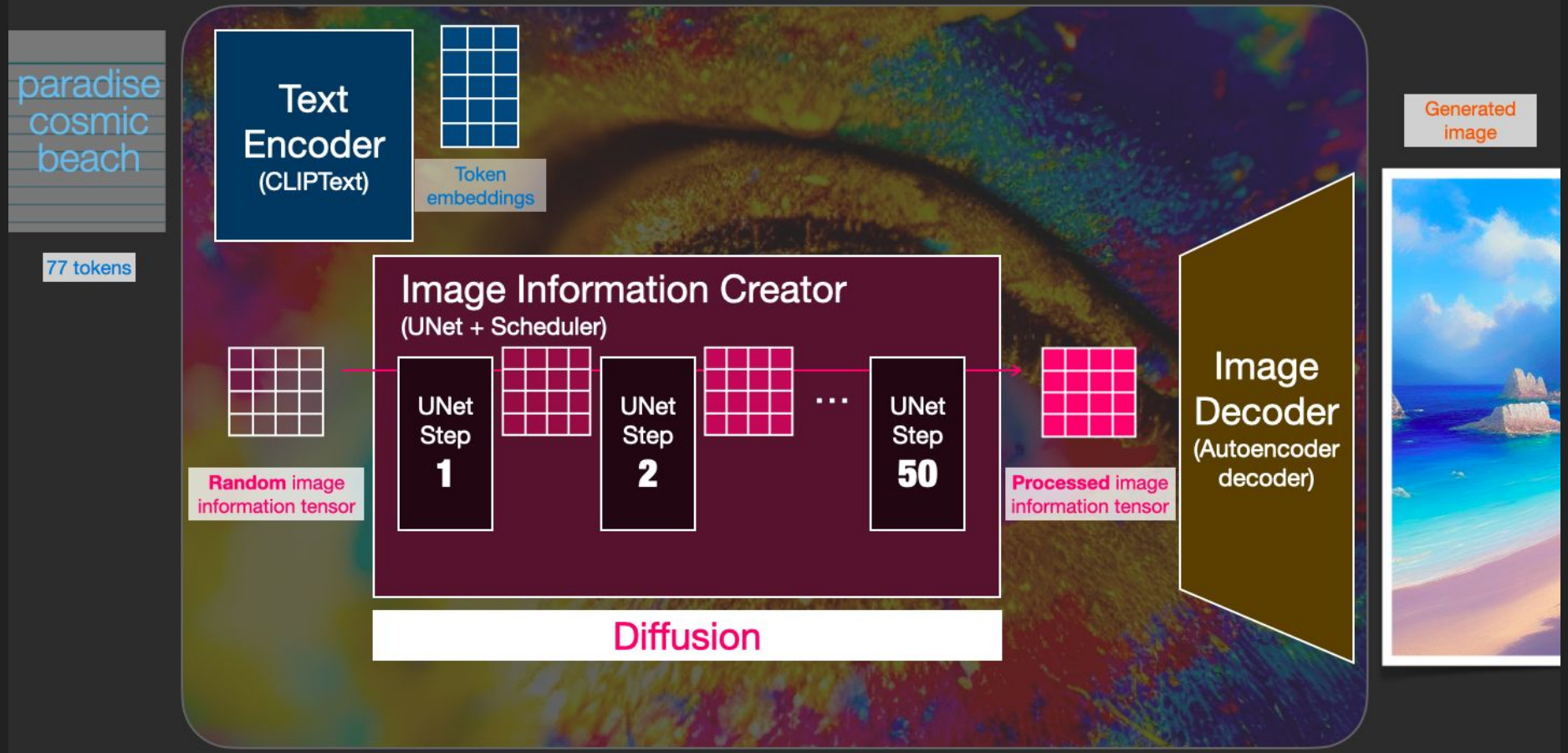
Conditioning

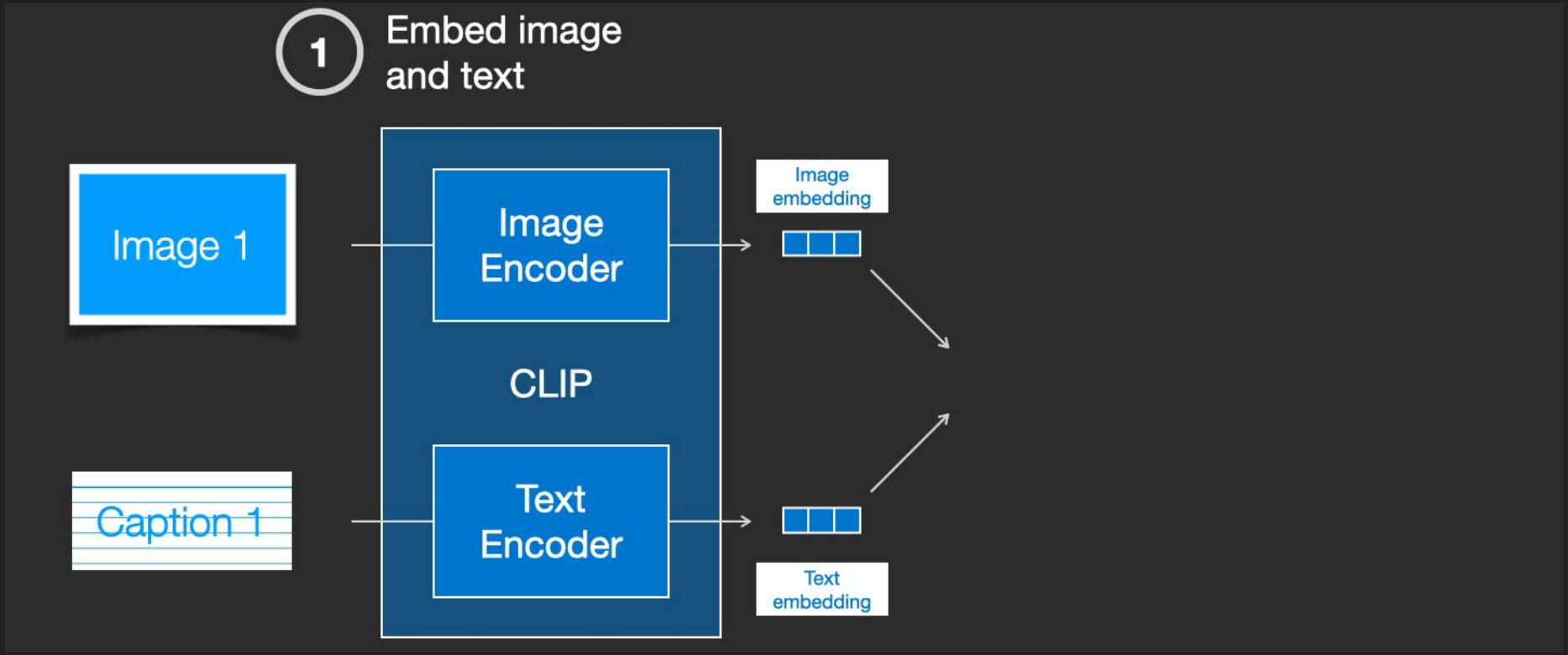


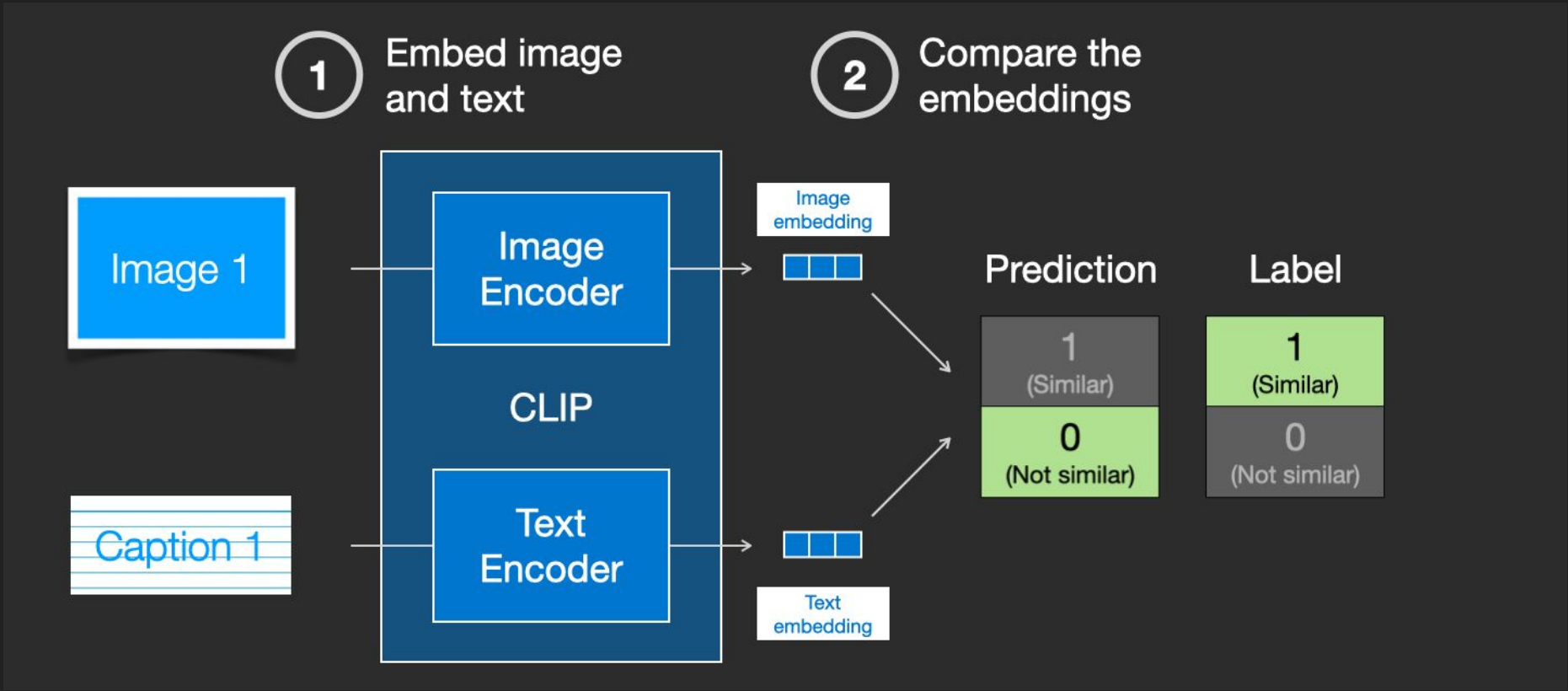
Stable Diffusion

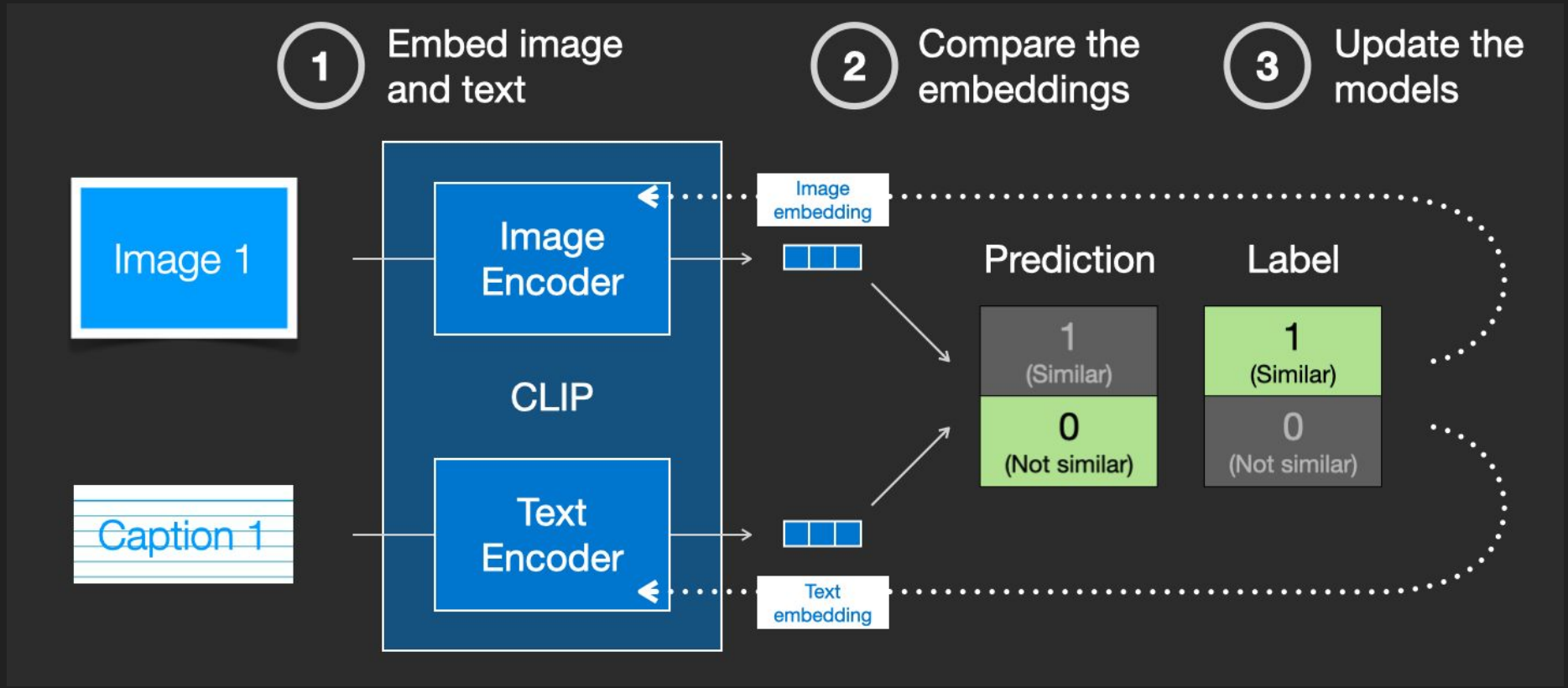

Input text
Token embeddings
Processed image information tensor
Generated image
paradise cosmic beach
77 tokens
Text Encoder
(CLIPText)
77 x 768
Image Information Creator
(UNet + Scheduler)
4 x 64 x 64
Image Decoder
(Autoencoder decoder)
3 x 512 x 512
Stable Diffusion

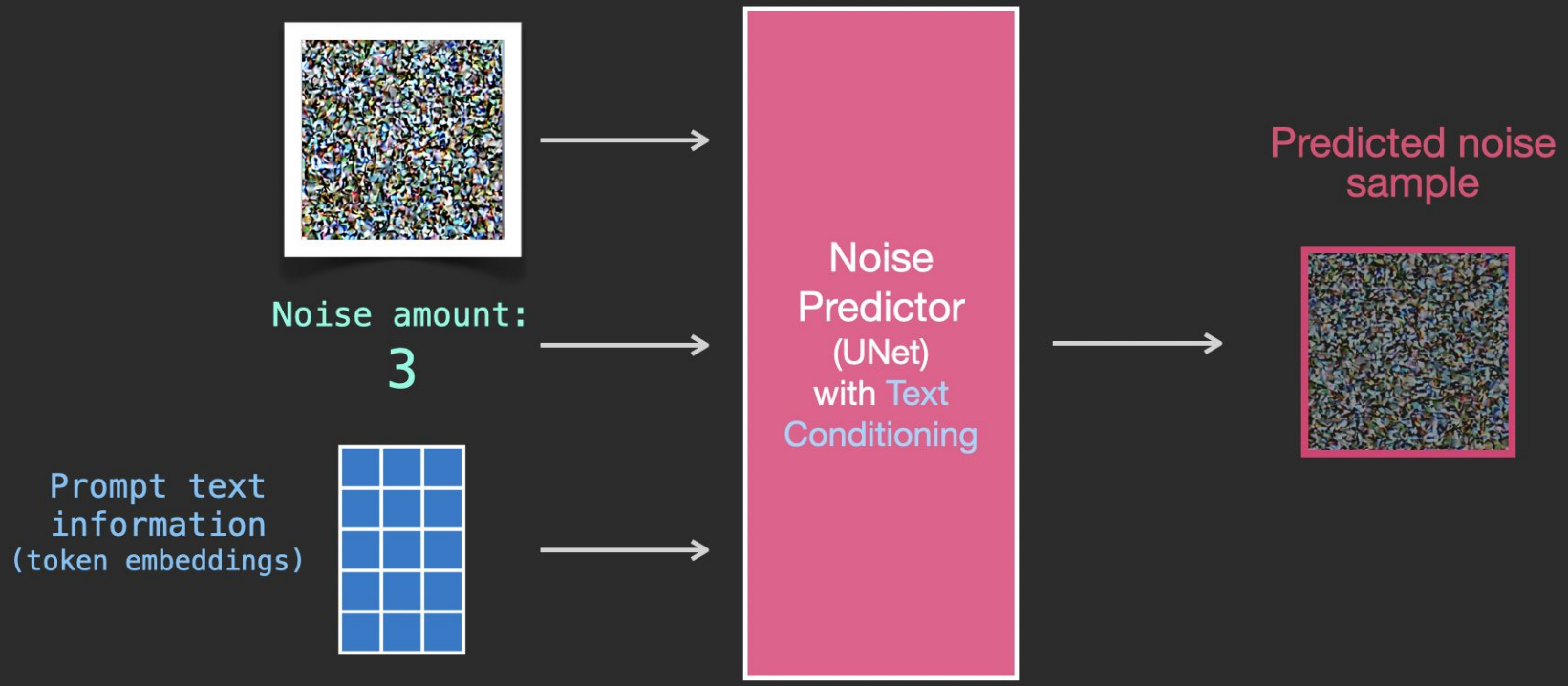
Stable Diffusion




1
Embed image and text
Image 1
Caption 1
Image Encoder
CLIP
Text Encoder
Image embedding
Text embedding


1
Embed image and text
2
Compare the embeddings
Image 1
Caption 1
Image Encoder
CLIP
Text Encoder
Image embedding
Text embedding
Prediction
1 (Similar)
0 (Not similar)
Label
1 (Similar)
0 (Not similar)

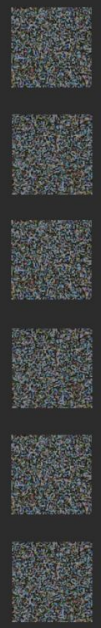

1
Embed image and text
2
Compare the embeddings
3
Update the models
Image 1
Caption 1
Image Encoder
CLIP
Text Encoder
Image embedding
Text embedding
Prediction
Label
1 (Similar)
0 (Not similar)
1 (Similar)
0 (Not similar)


Noise amount:
3
Prompt text information
(token embeddings)
Noise Predictor
(UNet)
with Text Conditioning
Predicted noise sample

DATASET

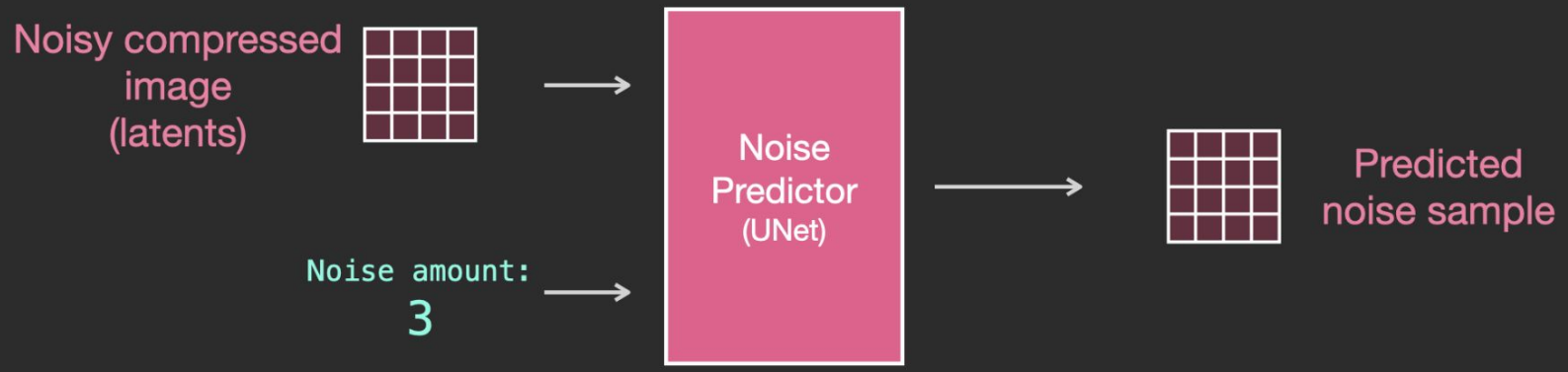
MODEL

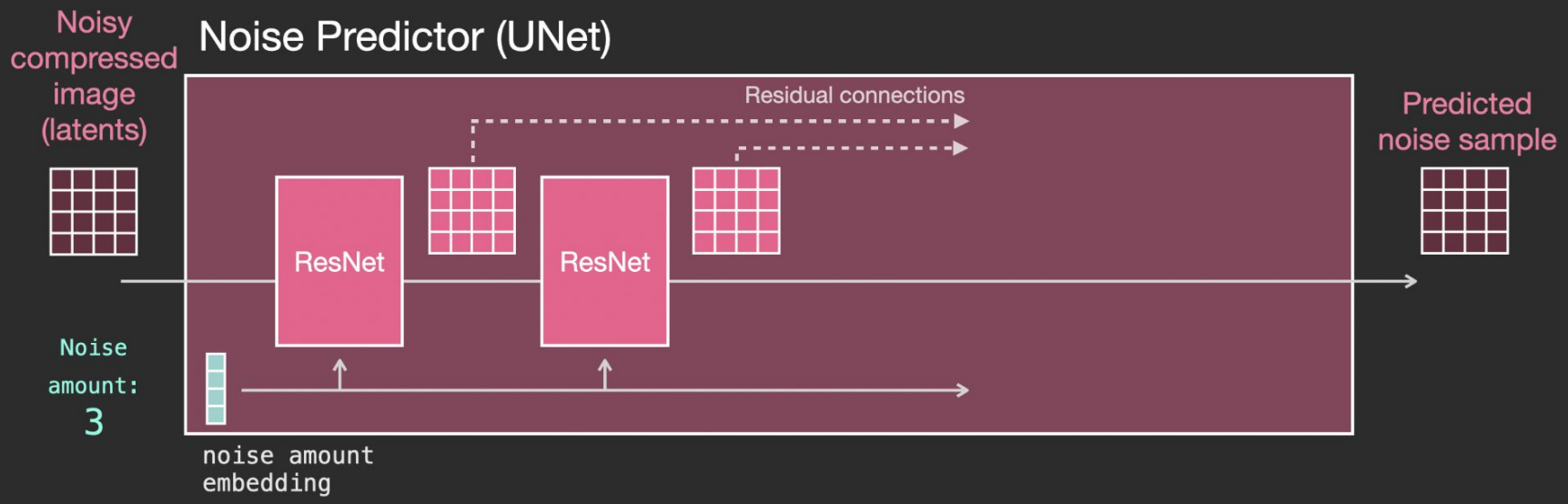
	INPUT		OUTPUT / LABEL
Step	Image	Text	noise sample
3			
14			
7			
42			
2			
21			



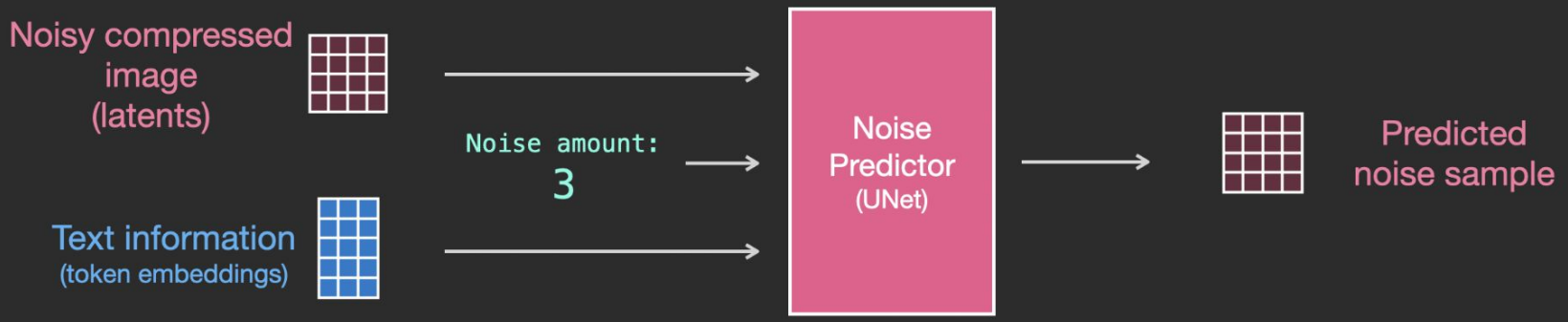
Noise Predictor
(UNet)
with Text
Conditioning

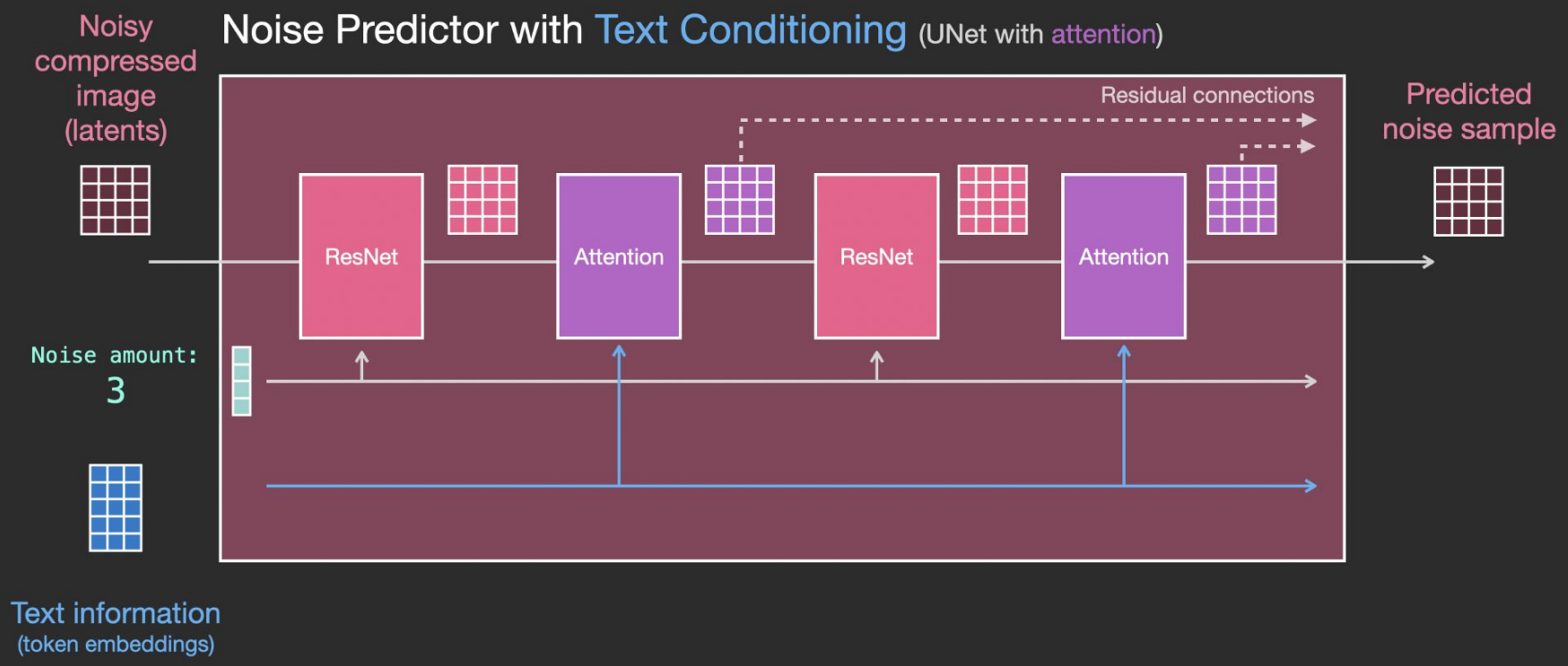
Without Text


Noisy compressed image (latents)
Noise amount:
3
Noise Predictor
(UNet)
Predicted noise sample


Noise Predictor (UNet)
Noisy compressed image (latents)
Residual connections
ResNet
ResNet
Predicted noise sample
Noise amount: 3
noise amount embedding

With Text


Noisy compressed image (latents)
Text information (token embeddings)
Noise amount: 3
Noise Predictor (UNet)
Predicted noise sample


Noise Predictor with Text Conditioning (UNet with attention)
Noisy compressed image (latents)
Residual connections
Predicted noise sample
ResNet
Attention
ResNet
Attention
Noise amount:
3
Text information
(token embeddings)

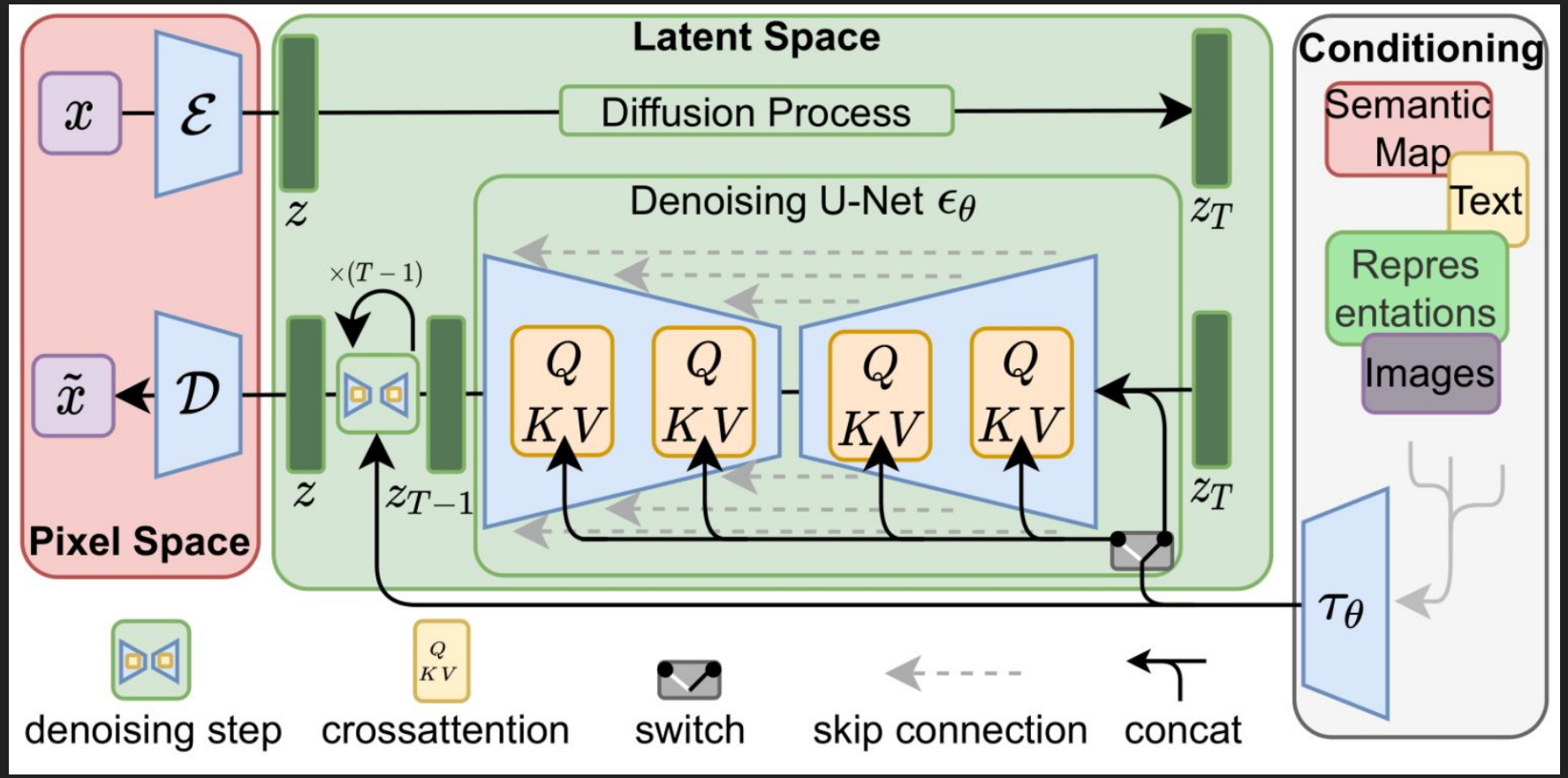
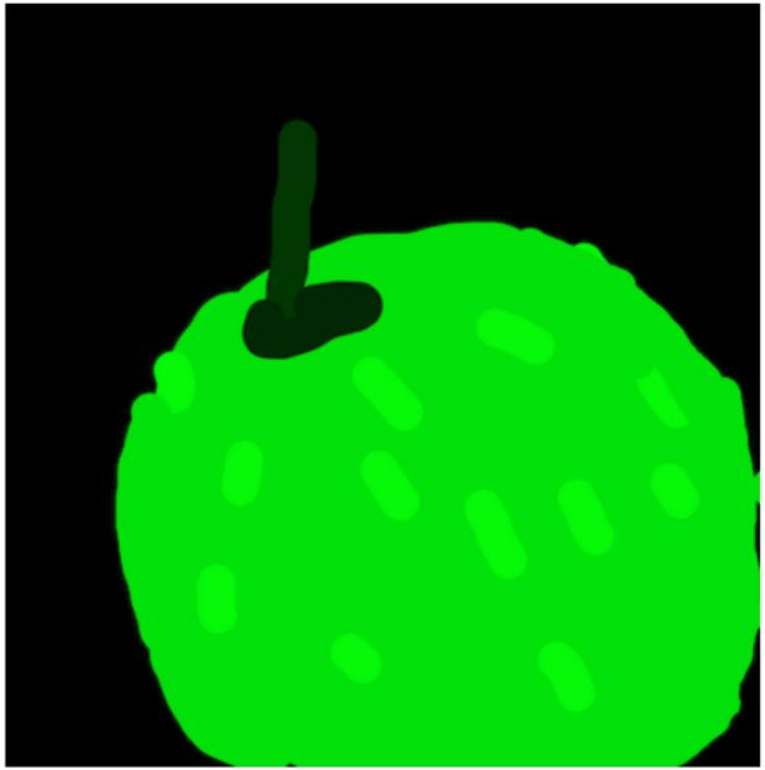
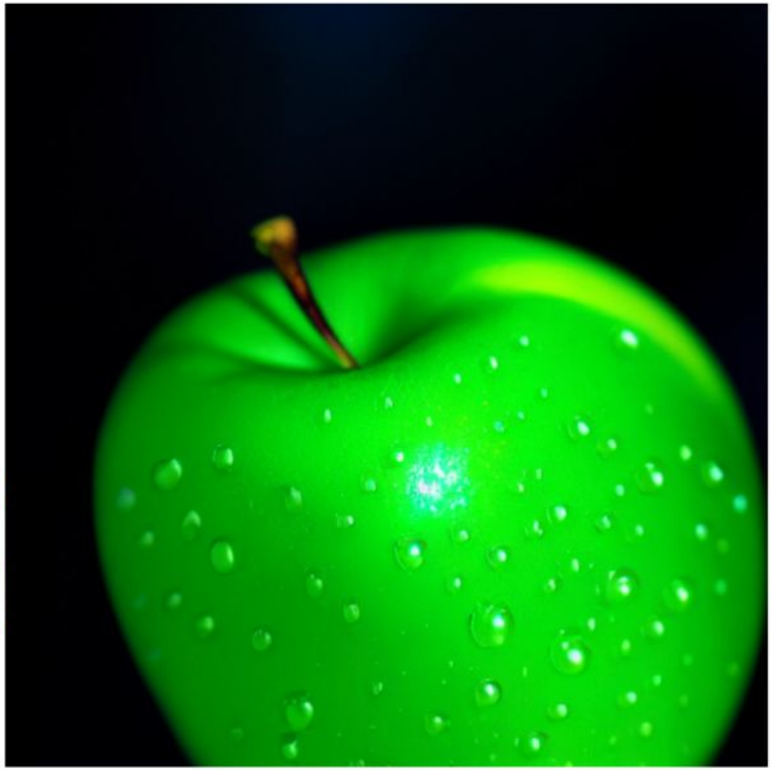
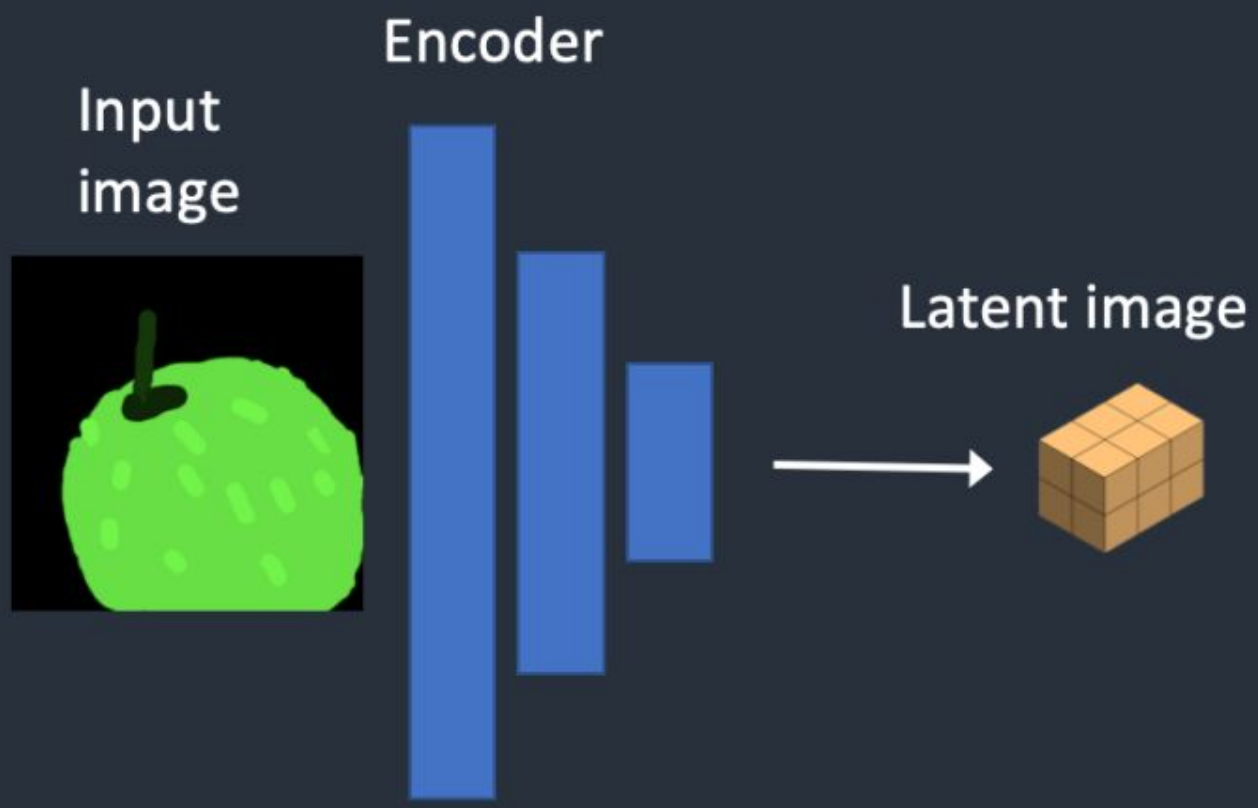

Latent Space
Diffusion Process
Denoising U-Net ϵ_θ
$\times(T-1)$
x
$\mathcal{E}$
$\tilde{x}$
$\mathcal{D}$
z
z_{T-1}
z_T
Q
K V
Pixel Space
Conditioning
Semantic Map
Text
Repres entations
Images
τ_θ
denoising step
crossattention
switch
skip connection
concat

Image-to-Image






Encoder
Input image
Latent image

REFERENCES

- https://huggingface.co/blog/annotated-diffusion
- https://jalammar.github.io/illustrated-stable-diffusion/
- https://stable-diffusion-art.com/how-stable-diffusion-work/