Module 9 - Exercise

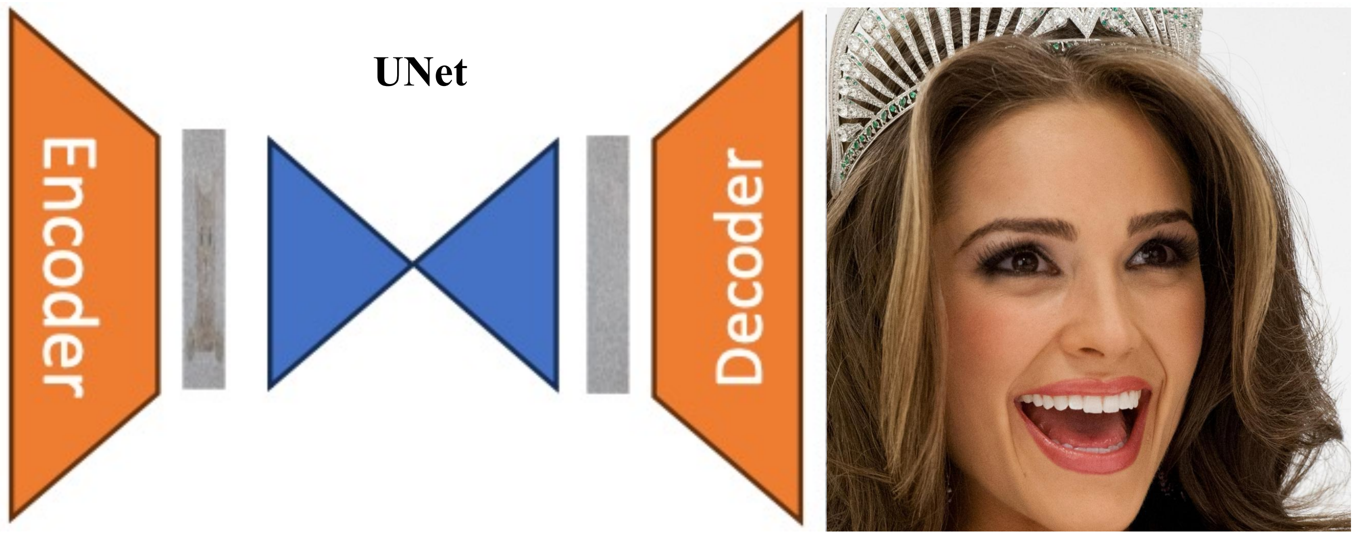
Text to Image Generation Using Stable Diffusion Model

AI VIET NAM
Nguyen Quoc Thai

Objectives

Text-to-Image using Stable Diffusion Model

The person has high cheekbones, and pointy nose. She is wearing lipstick.



Outline

- Introduction
- Stable Diffusion Model
- Text-to-Image Generation using SDM

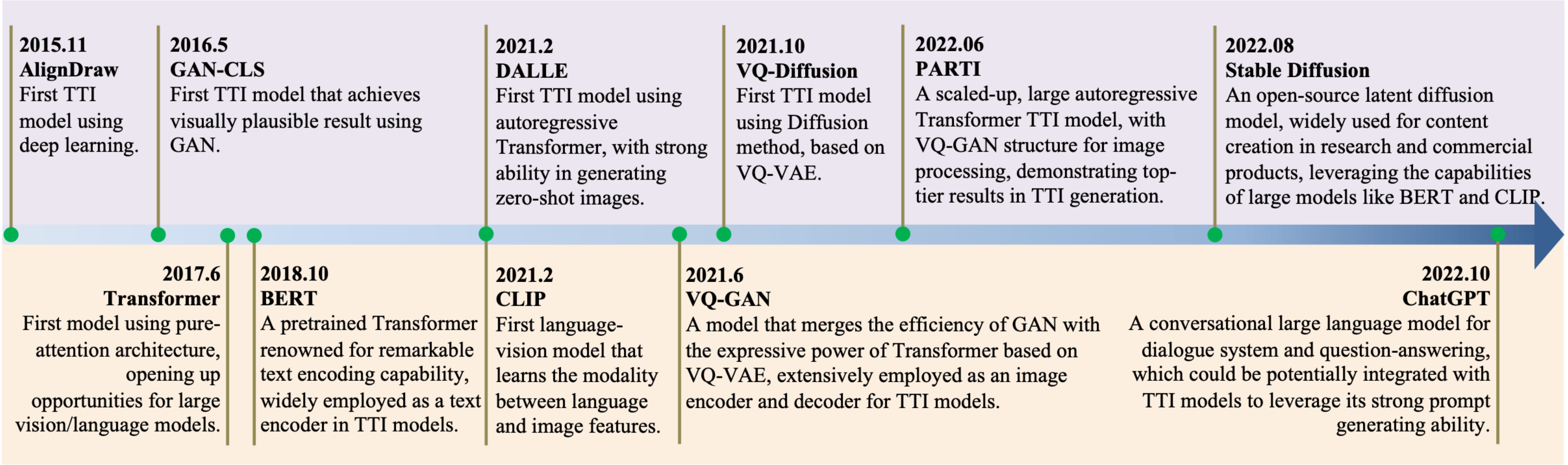
Introduction

Text to Image Generation

- The person has high cheekbones, and pointy nose. She is wearing lipstick.

Introduction

The milestones of text-to-image models and large models

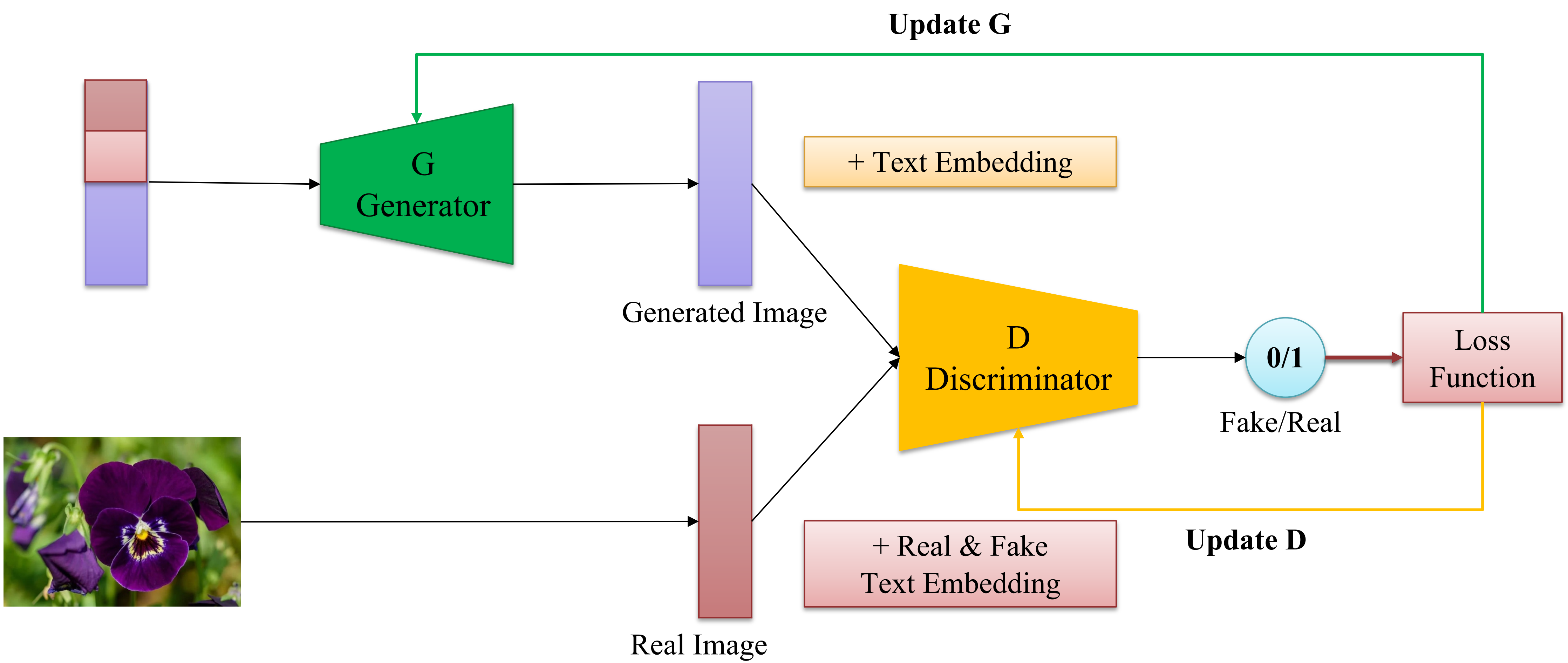


Source

AI VIET NAM
@aivietnam.edu.vn

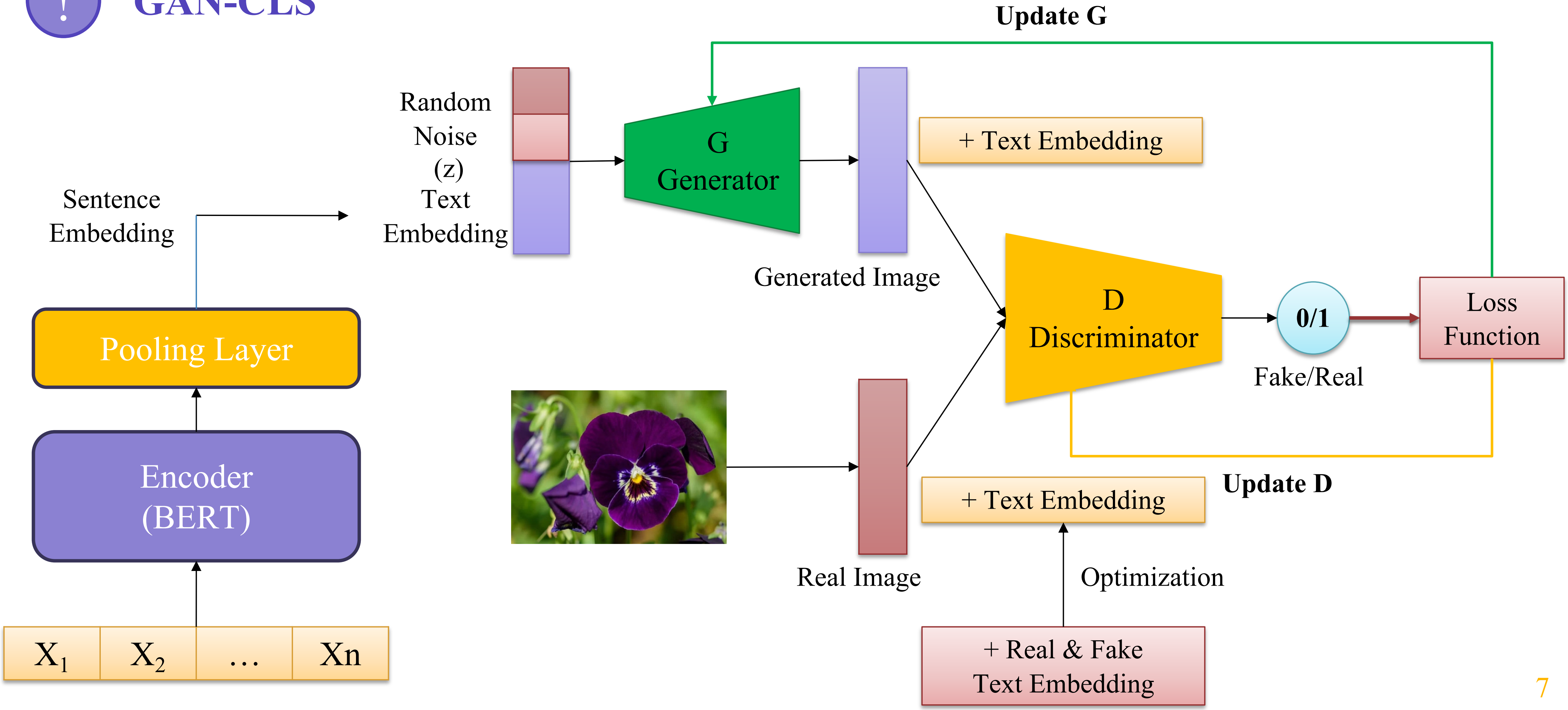
Introduction

GAN-CLS



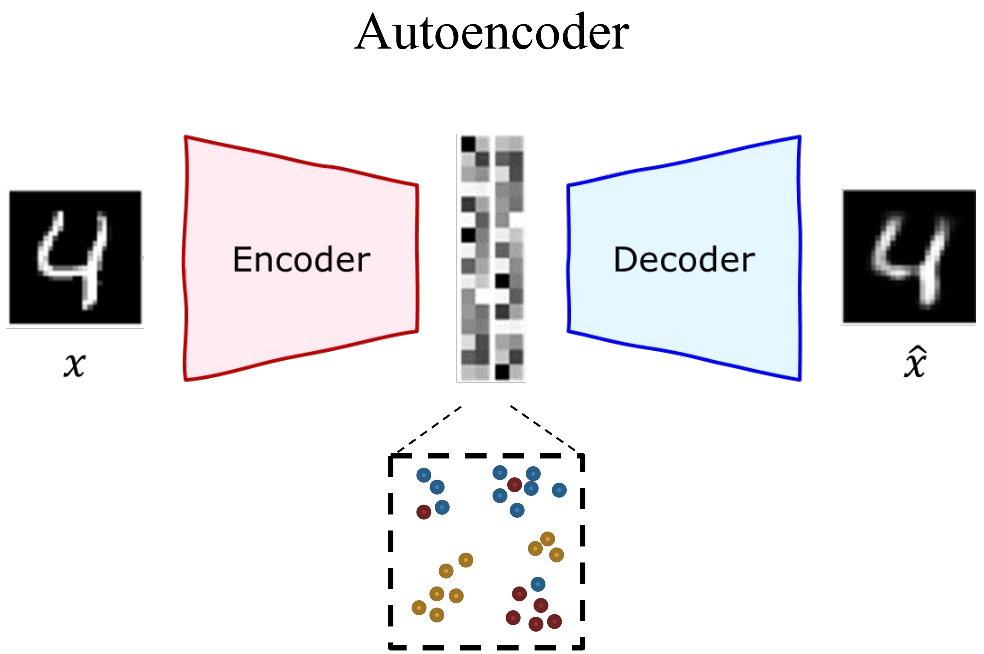
Introduction

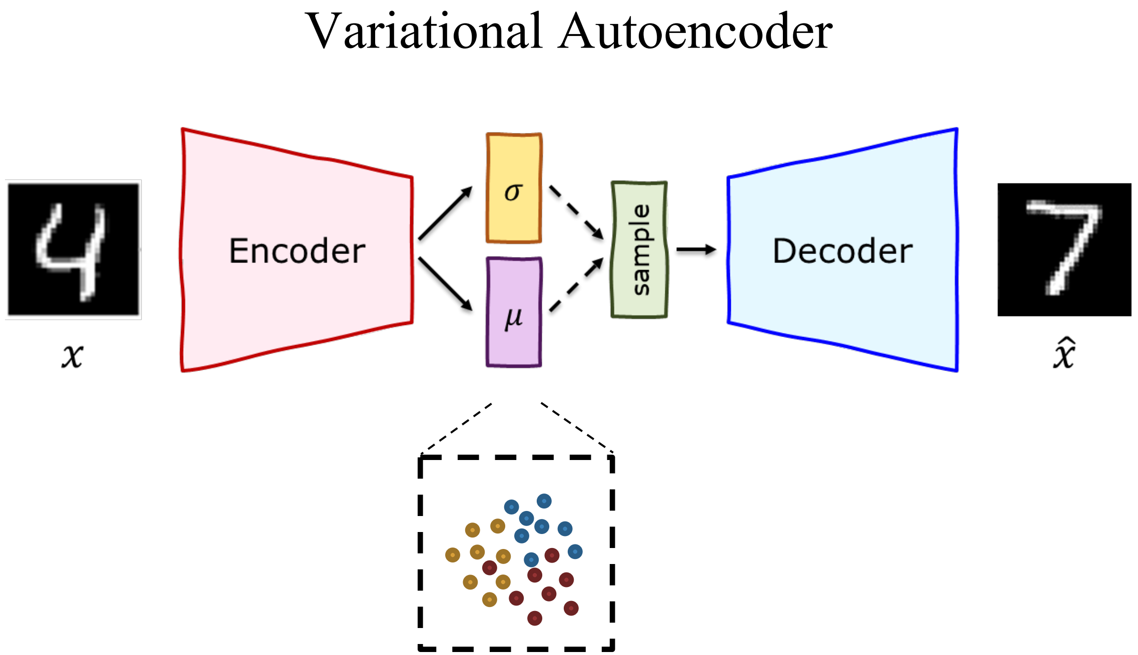
GAN-CLS



Introduction

Variatinal AutoEncoder

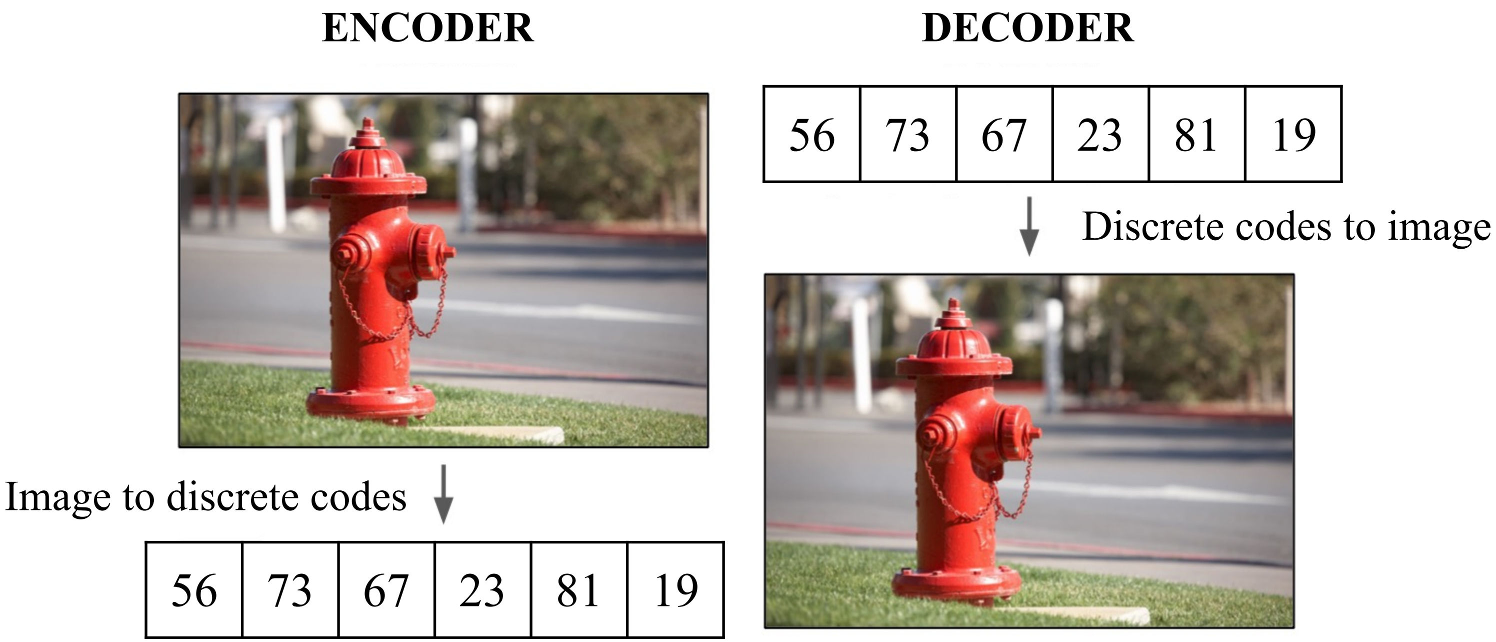




Introduction

Vector Quantized Variatinal AutoEncoder (VQ-VAE)

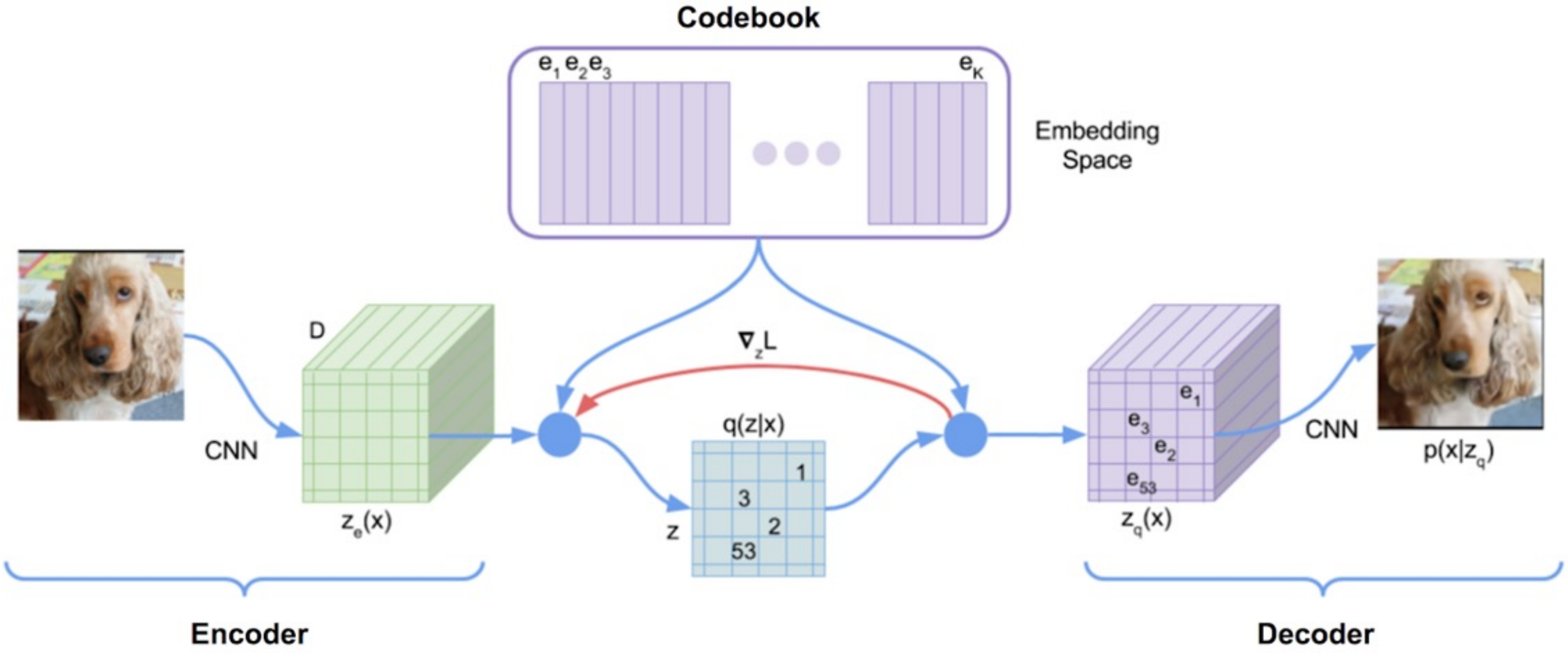
- VAE learns a discrete latent representation, instead continuous
- Latents do not necessarily need to be continuous vectors
- It just needs to be some form of numerical representation of the data



Introduction

Visual Vocabulary

- Introduces a Discrete Latent Codebook to store a finite set of possible latent vectors
- Describe an image as a sequence of symbols (language tokens)

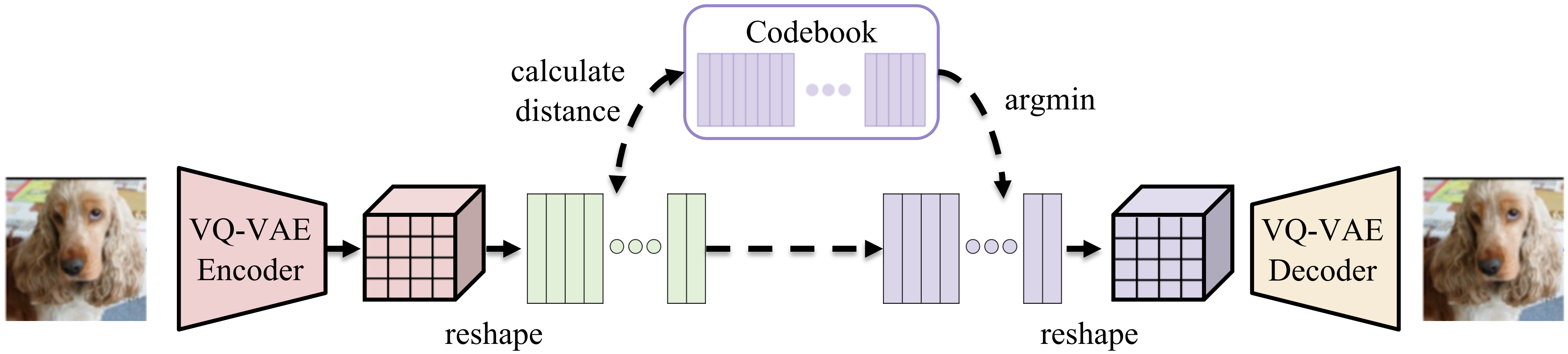


Introduction

Quantization Layer

During Training

- Output of encoder is compared to all the vectors in the codebook
- The codebook vector closet in Euclidean distance is fed to decoder

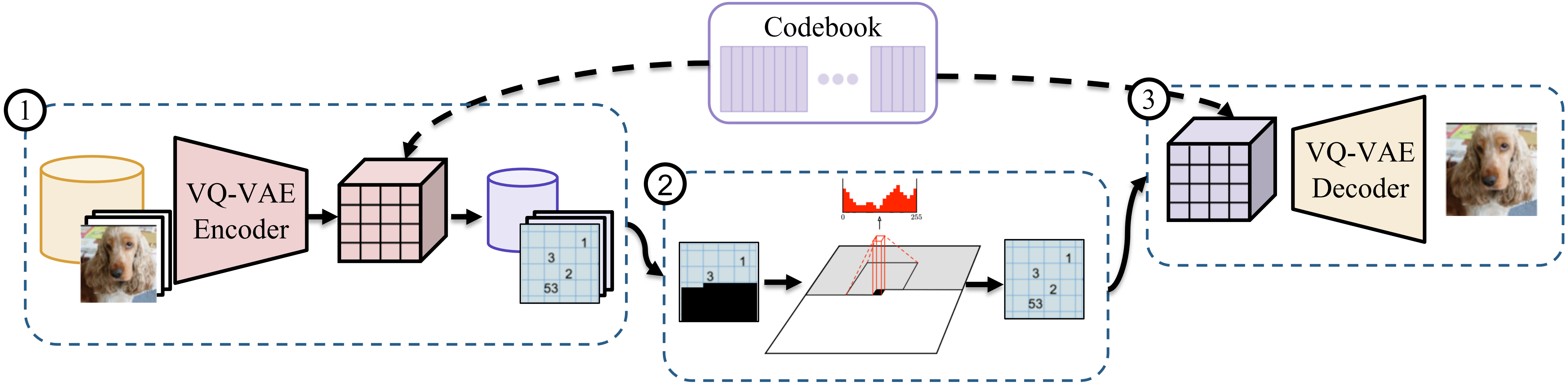


Introduction

Generate Image from Codebook

Train a PixelCNN as prior on the discretized 32x32 latent space

- Use VQ-VAE Encoder to extract latent space (codebook indicates) from dataset
- Train PixelCNN to auto-regressively complete the latent codebook
- Use VQ-VAE Decoder to generate image from the completed latent codebook



Introduction

DALL-E

Text-to-Image Generator model using a transformer that autoregressively models the text and image tokens as a single stream of data

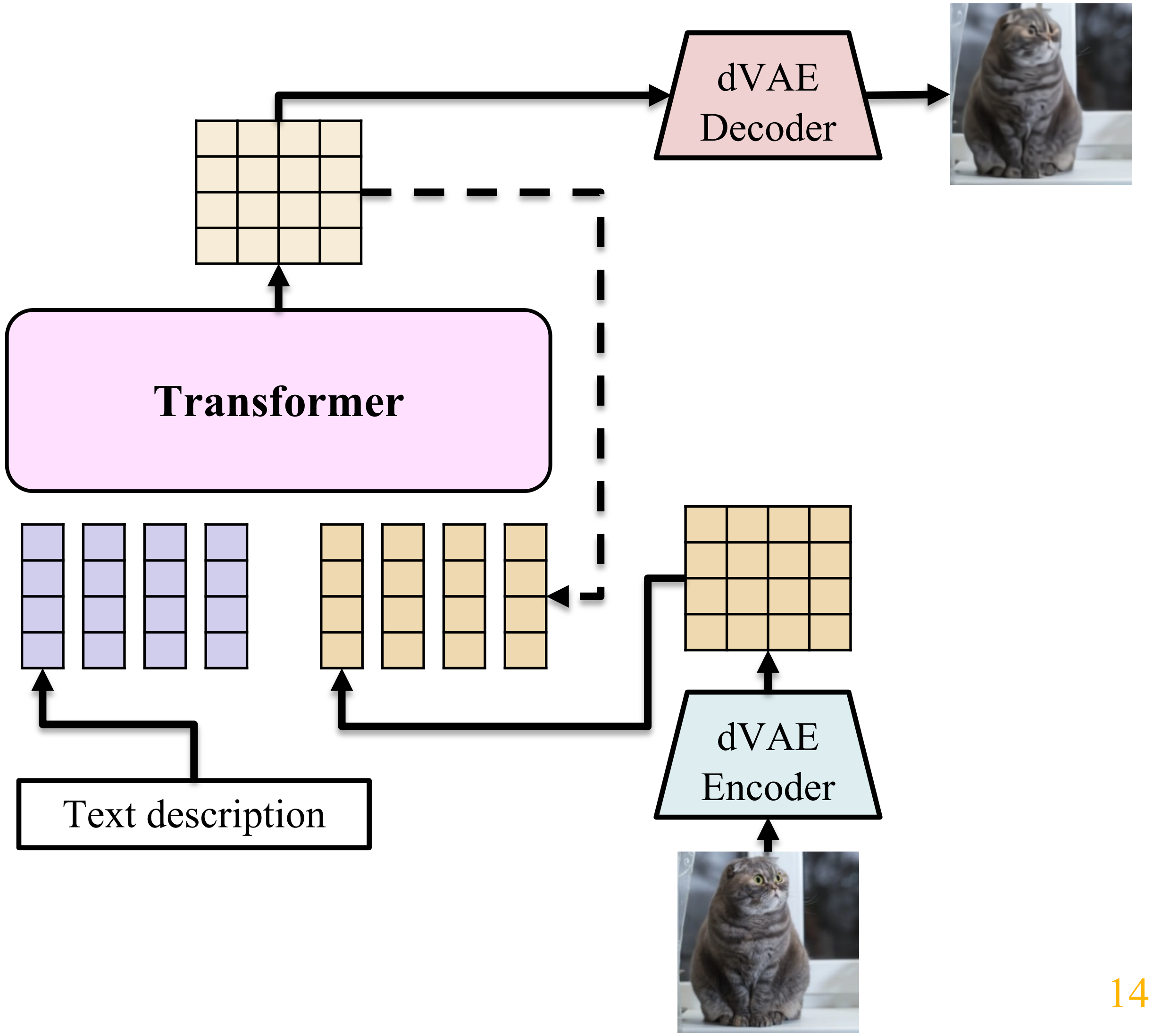
- Uses Discrete VAE
- Switch PixelCNN with a 12-billion parameter GPT-3
- Trained on 250 million image-text pairs

An armchair in the shape of an avocado

Introduction

DALL-E Parts

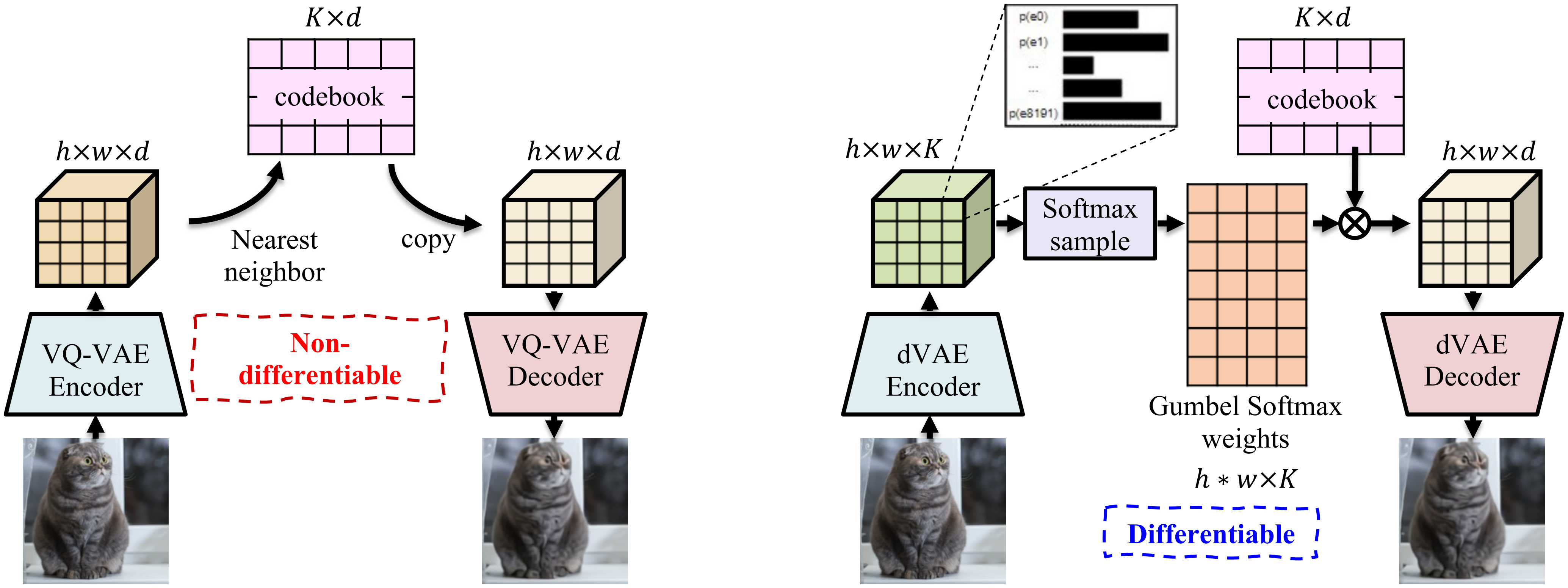
- Discrete VAE encoder and decoder
  - Inspired by VQ-VAE-2
  - Compress 256x256 RGB images into a 32x32 grid of image tokens
  - With 8192 possible codebook tokens
- Transformer Decoder
  - Concatenate text tokens with image tokens into single array
  - Train to predict text image token from the preceding tokens



Introduction

Gumbel Softmax Relaxation

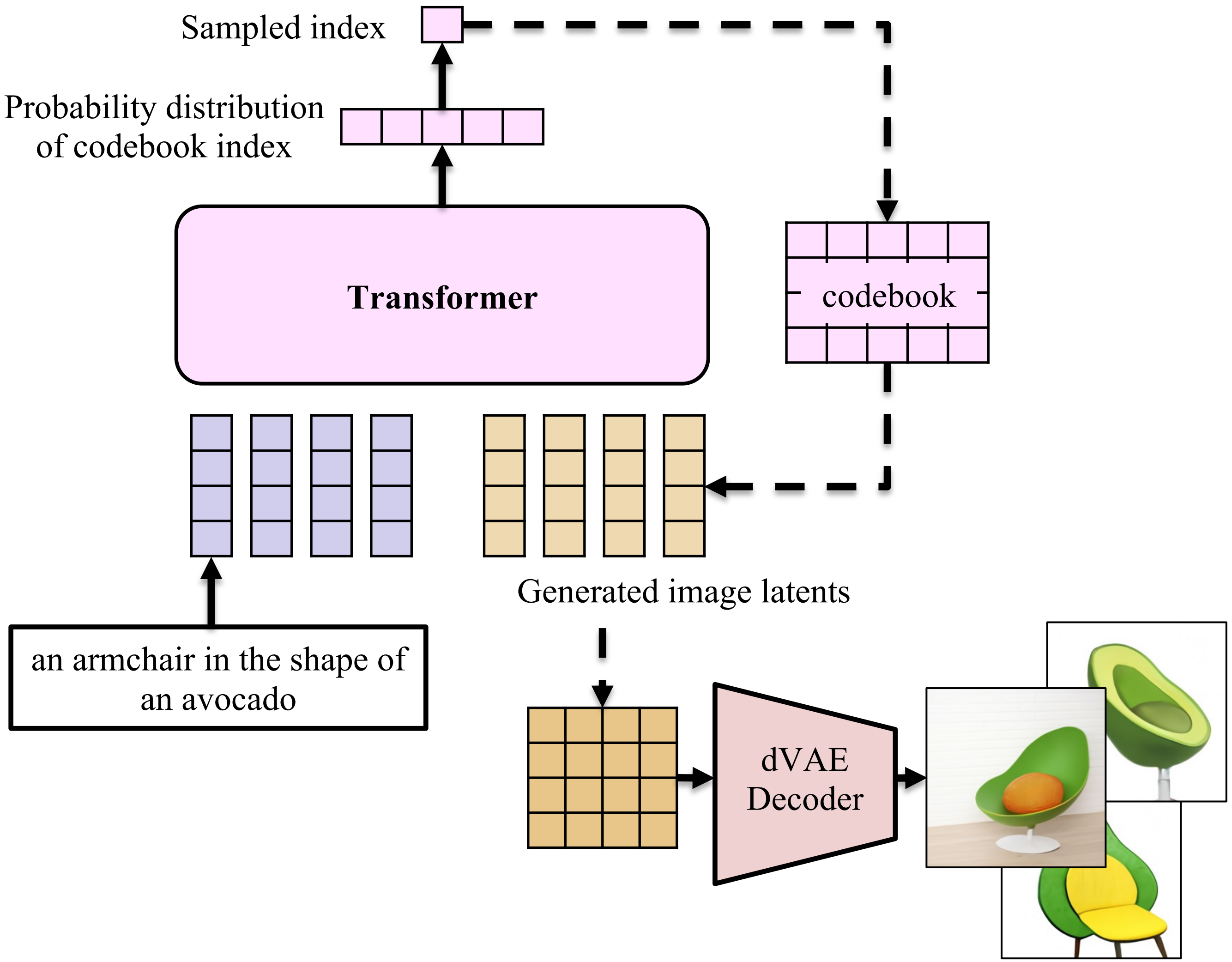
➢ Outputs a distribution over codebook vectors for each latent code instead of mapping deterministically to a single codebook vector.



Introduction

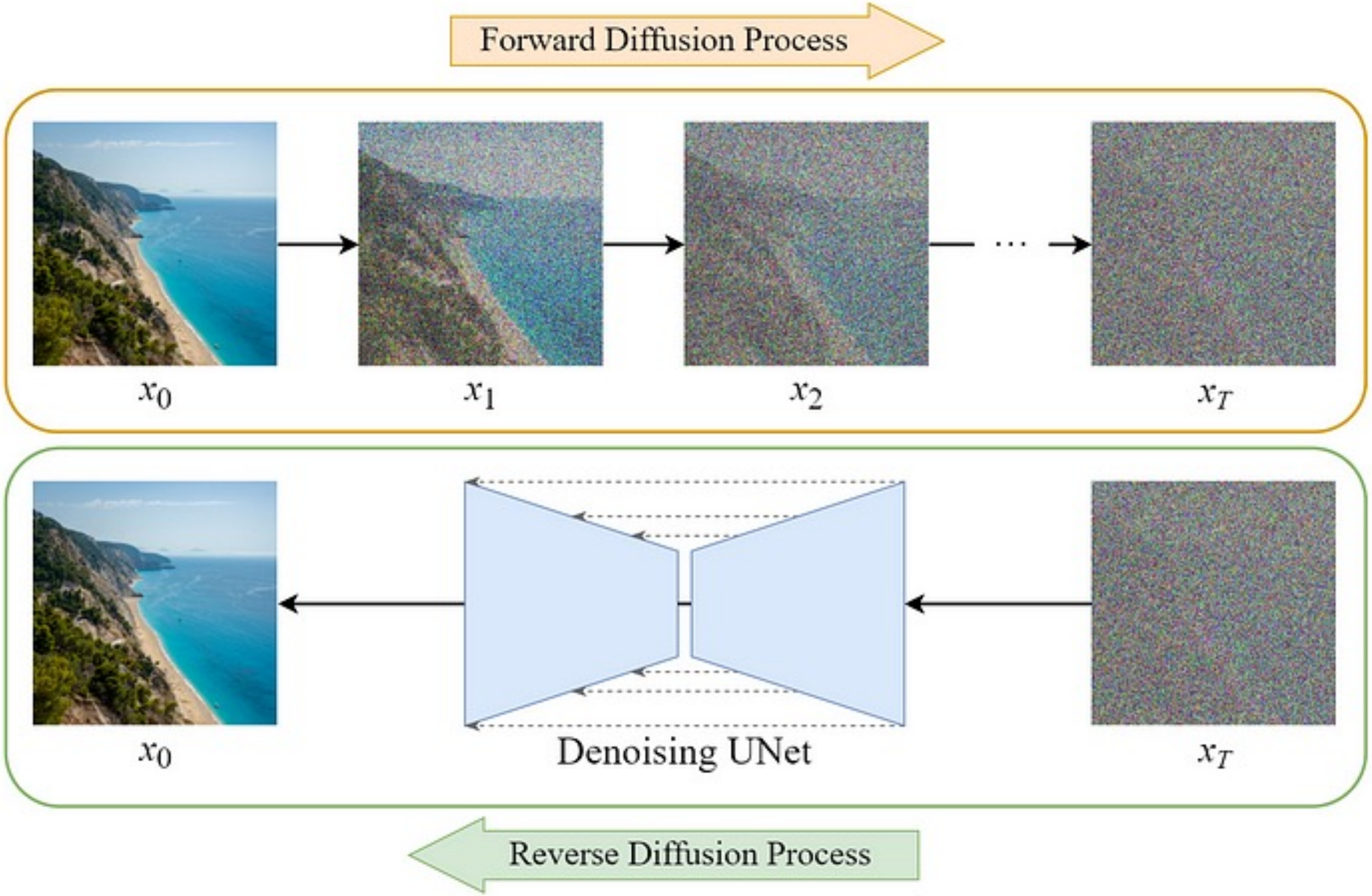
Dall-E Inference

- Not directly predict (choose) the next latent index, but predict the distribution
- Then sample the index from that distribution



Introduction

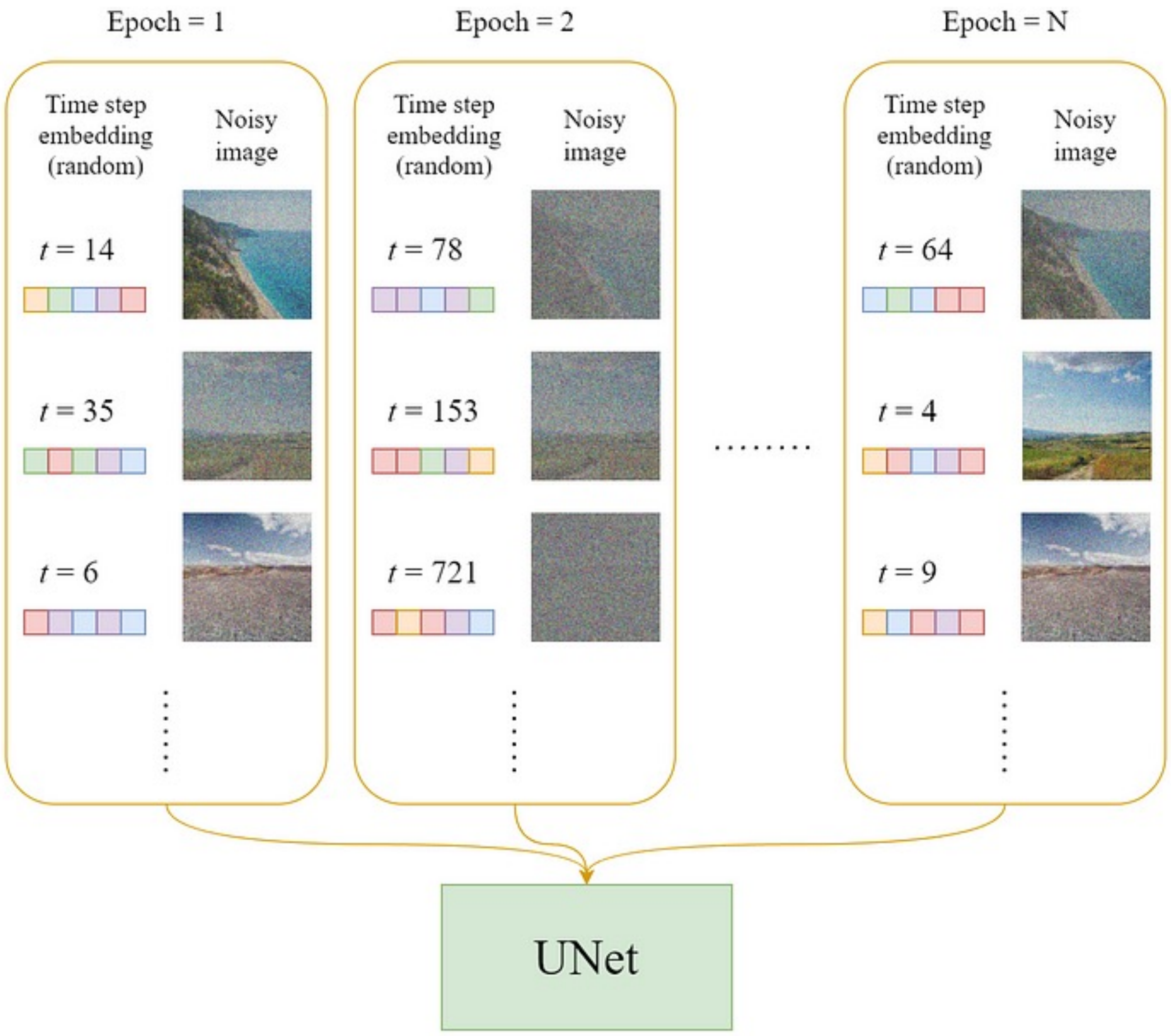
Diffusion Model



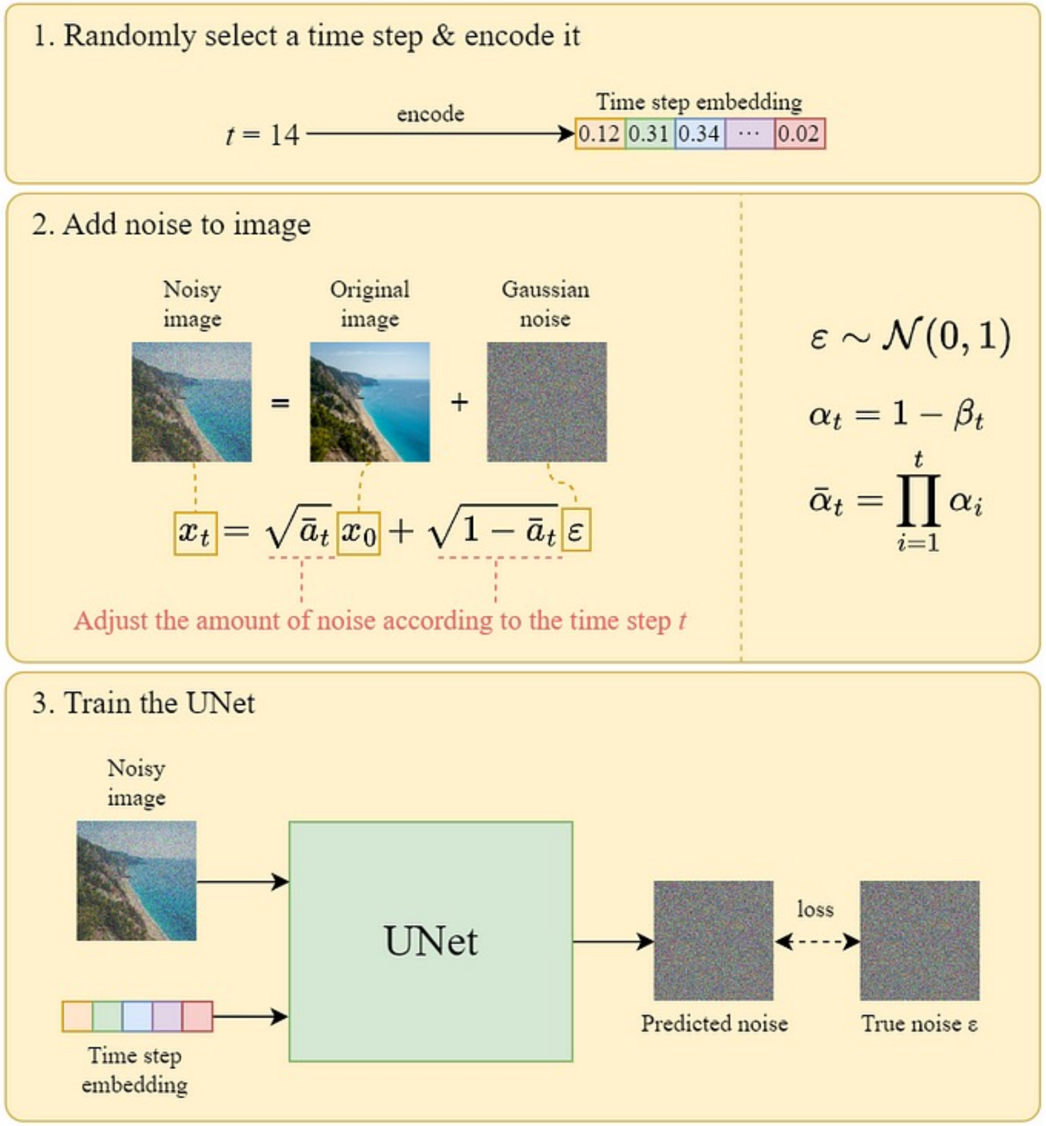
AI VIET NAM
@aivietnam.edu.vn

Introduction

Diffusion Model - Training

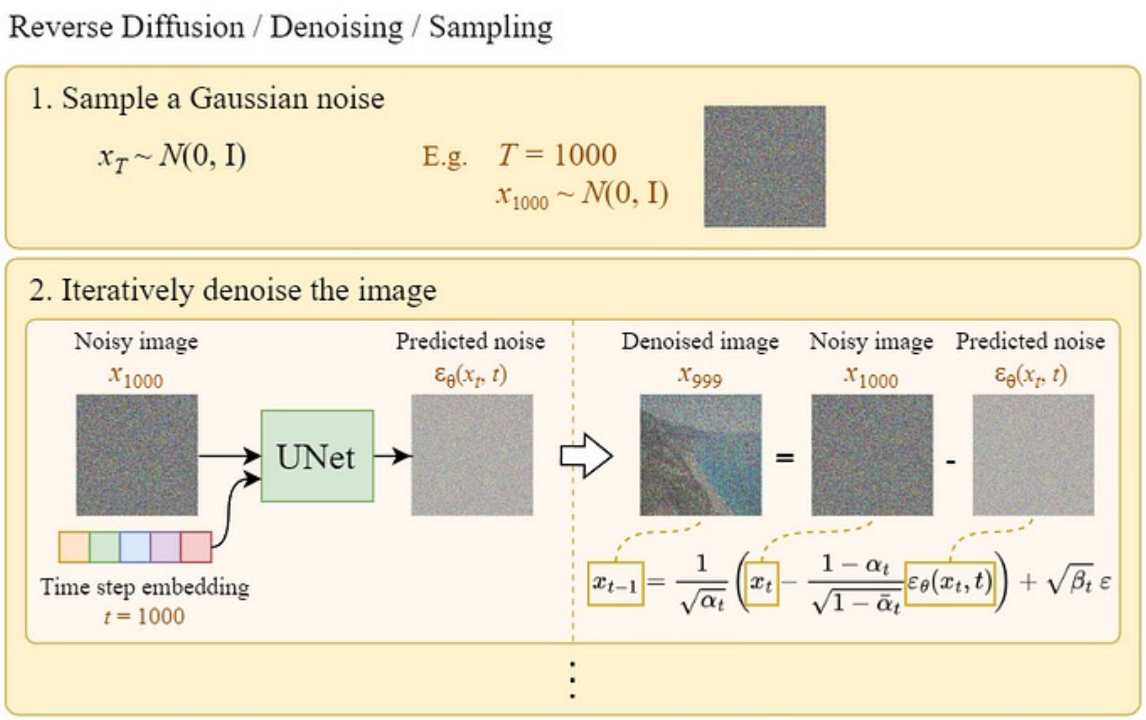


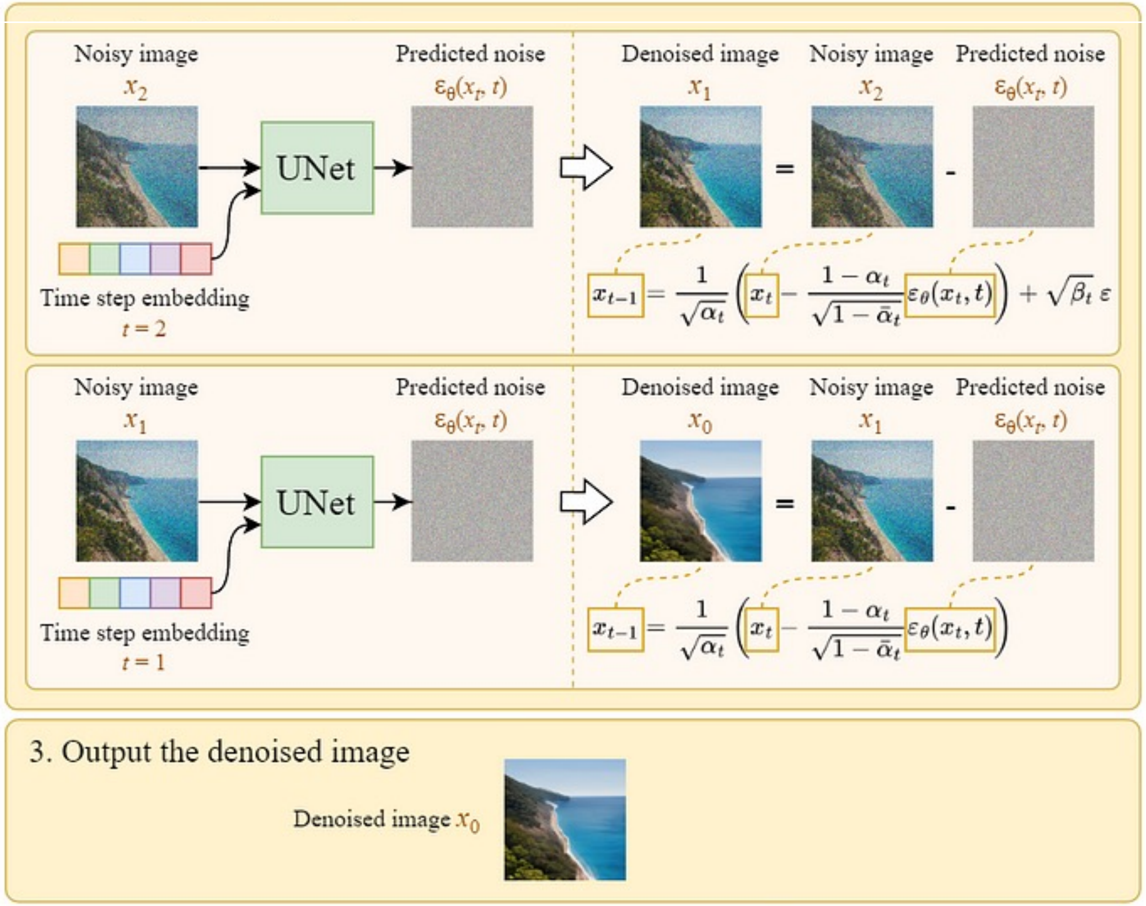
For each training step:



Introduction

Diffusion Model - Inference

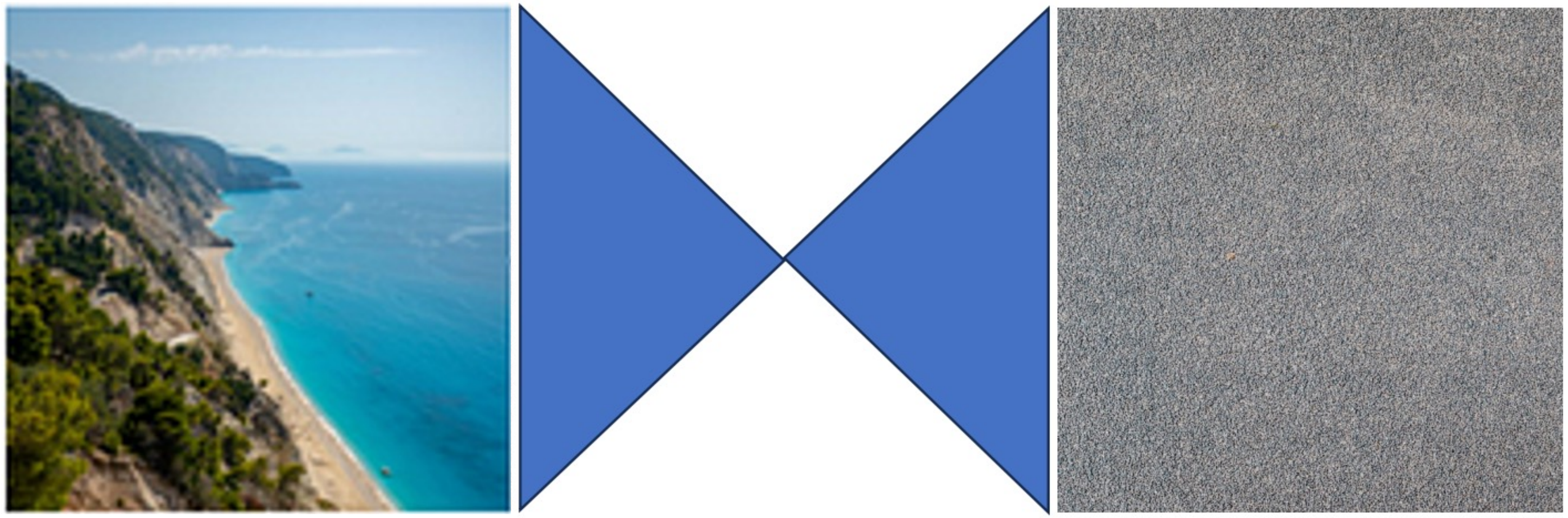




Introduction

Diffusion Model - Problem

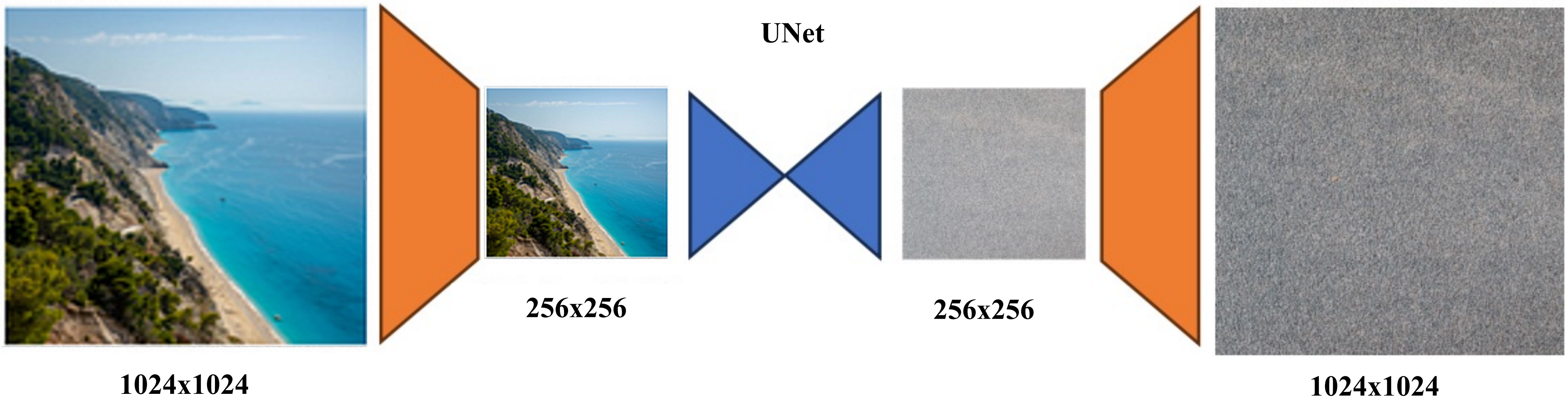
- Operating in the input space is very computationally expensive



1024x1024 1024x1024

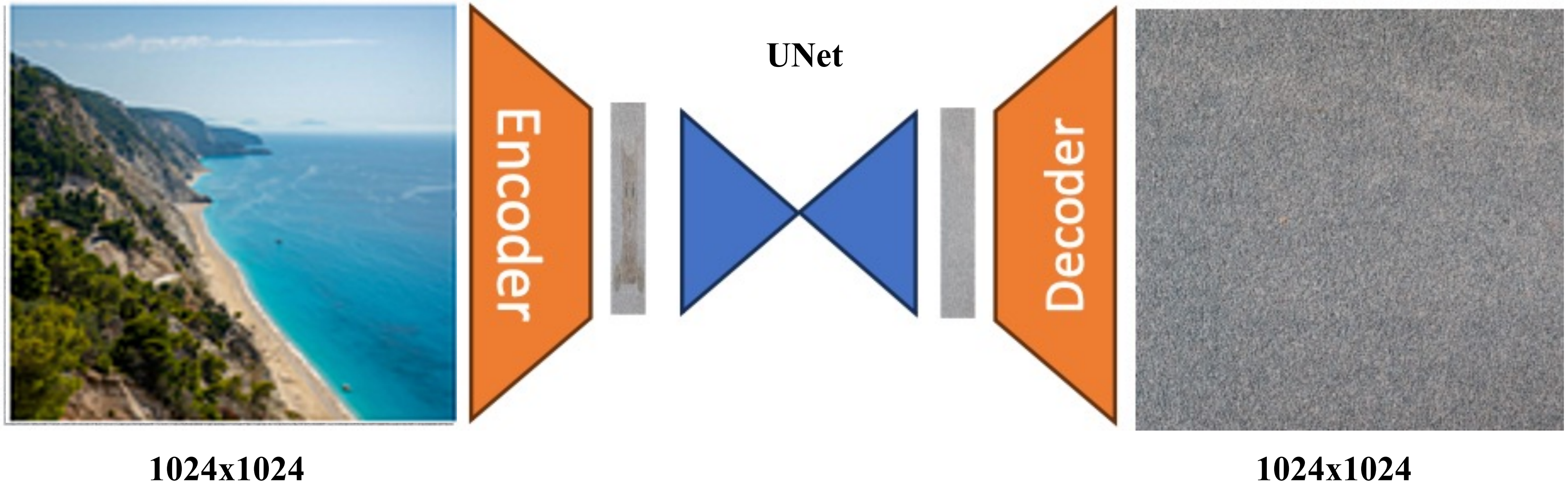
Introduction

Diffusion Model – Generate Low-Resolution + Upsample



Introduction

Diffusion Models – Generate in Latent Space



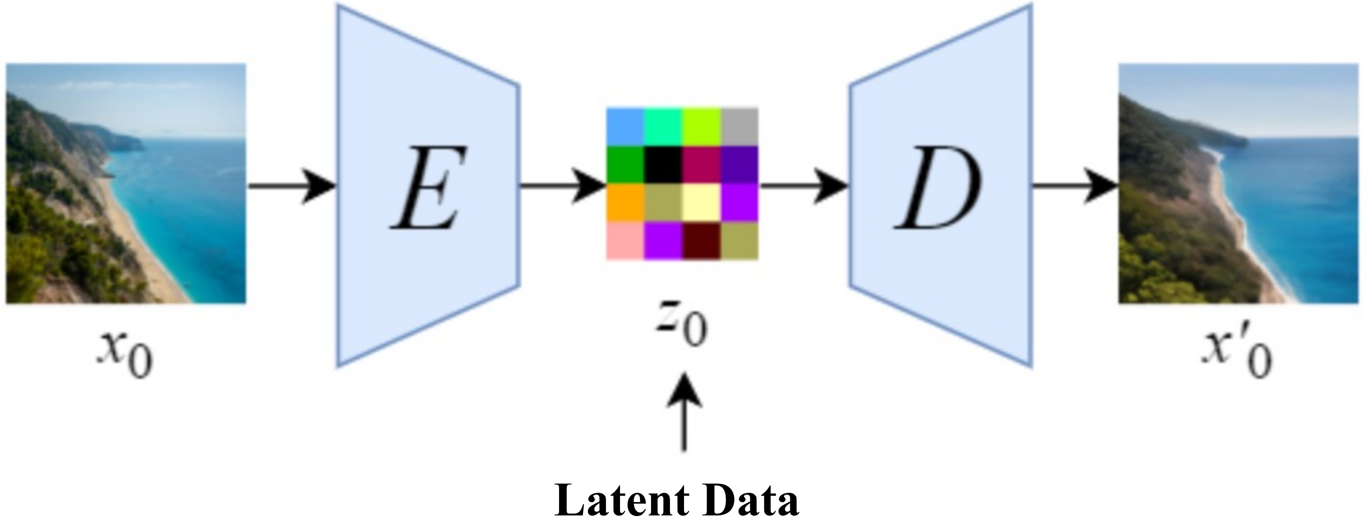
Outline

- Introduction
- Stable Diffusion Model
- Text-to-Image Generation using SDM

Stable Diffusion Model

Stable Diffusion Model (Latent Diffusion Model)

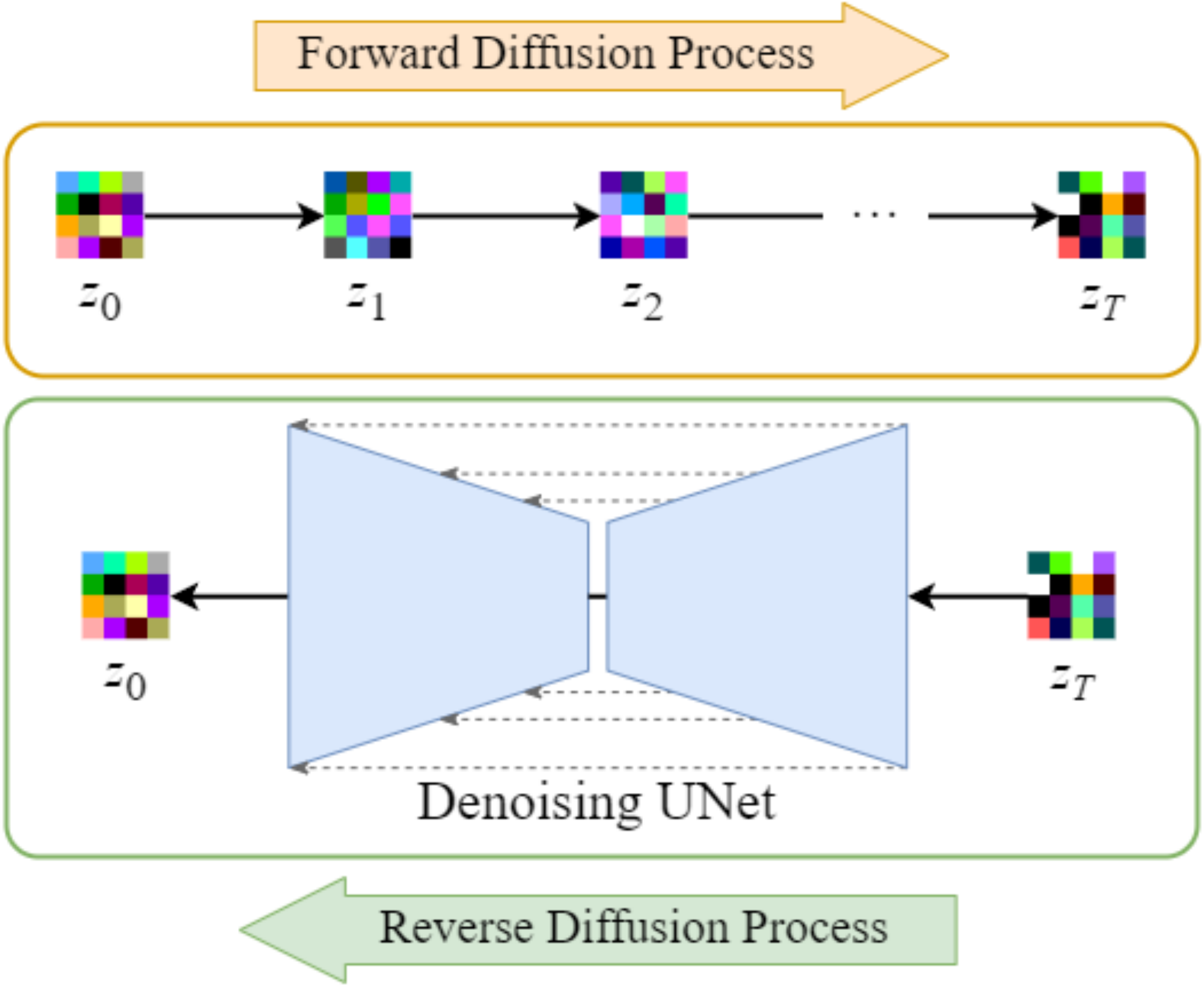
- The Diffusion process happens in the latent space
- First, train an autoencoder to learn to compress the image data into low-dimensional representation



Stable Diffusion Model

Stable Diffusion Model (Latent Diffusion Model)

- After encoding the images into latent data, the forward and reverse diffusion processes will be done in the latent space.

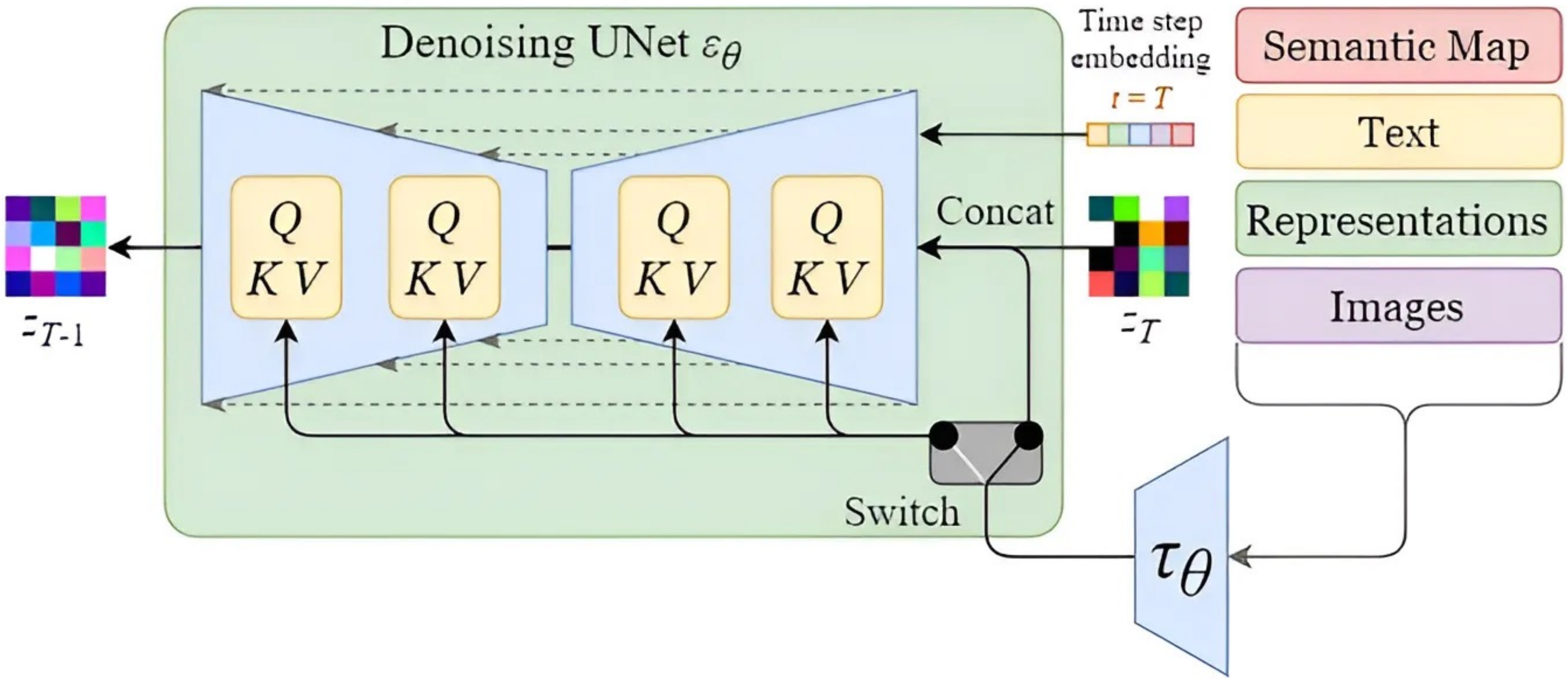


AI VIET NAM
@aivietnam.edu.vn

Stable Diffusion Model

Conditional Generation

- Condition denoising on text, images, etc,…



Stable Diffusion Model

Training

- The training objective (loss function) is pretty similar to the one in the pure diffusion model. The only changes are:
  - Input latent data z instead of the image x
  - Added conditioning input $\tau_\theta(y)$ to the UNet

$$\boxed{z_0 = E(x_0)}$$

$$\boxed{z_t = \sqrt{\bar{\alpha}_t}\,\boxed{z_0} + \sqrt{1 - \bar{\alpha}_t}\,\varepsilon}$$

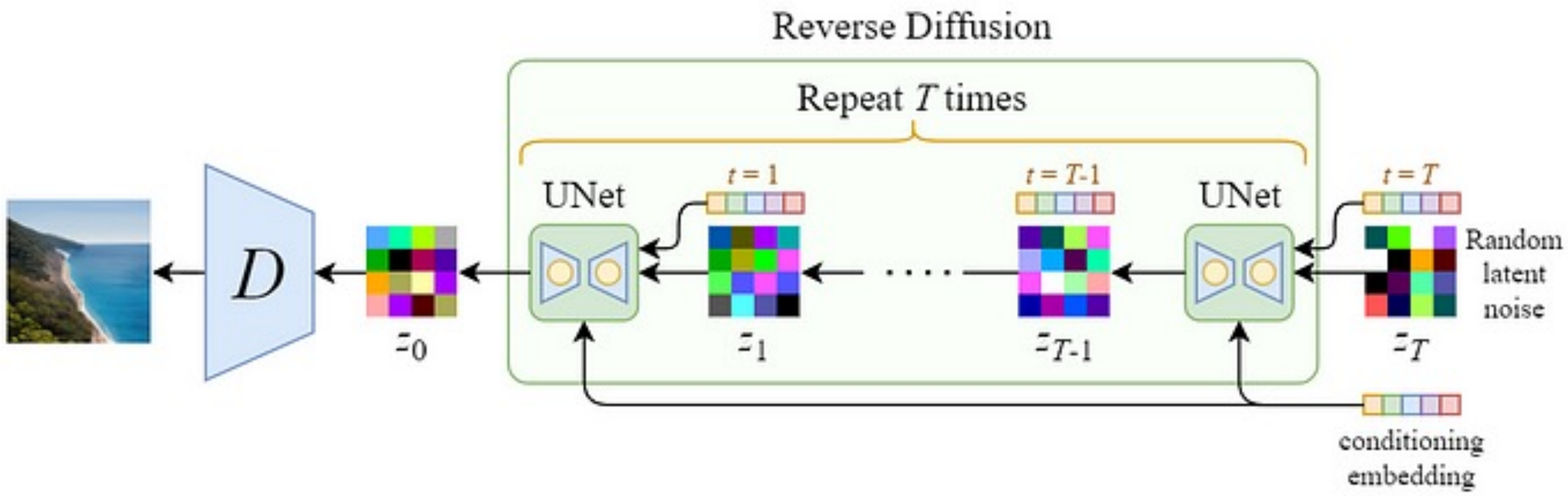
$$L_{\mathrm{LDM}} = \mathbb{E}_{t, z_0, \varepsilon, y}\left[\left\| \varepsilon - \varepsilon_\theta(\boxed{z_t}, t, \boxed{\tau_\theta(y)}) \right\|^2 \right]$$

Conditioning

Stable Diffusion Model

Sampling

- Stable Diffusion sampling process use the latent data
- The size of the latent data is much smaller than the original images, the denoising process will be much faster



Outline

- Introduction
- Stable Diffusion Model
- Text-to-Image Generation using SDM

Text-to-Image using SDM

Celeb-HQ Dataset

Text-to-Image using SDM

Celeb-HQ Dataset

- The person has high cheekbones, and pointy nose. She is wearing lipstick.
- The person has high cheekbones, and pointy nose. She is wearing lipstick.
- She is wearing lipstick. She is young, and smiling and has big lips, mouth slightly open, pointy nose, and high cheekbones.
- This attractive woman has high cheekbones, pointy nose, bushy eyebrows, mouth slightly open, wavy hair, arched eyebrows, and bags under eyes.
- …

AI VIET NAM
@aivietnam.edu.vn

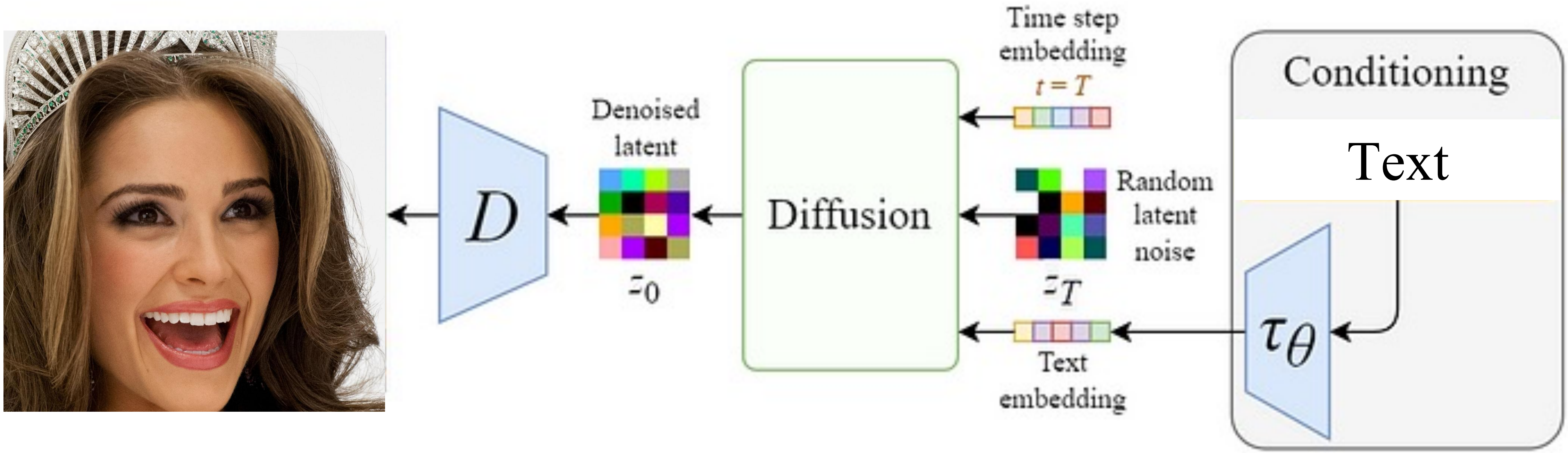
Text-to-Image using SDM

Celeb-HQ Dataset

- The person has high cheekbones, and pointy nose. She is wearing lipstick.

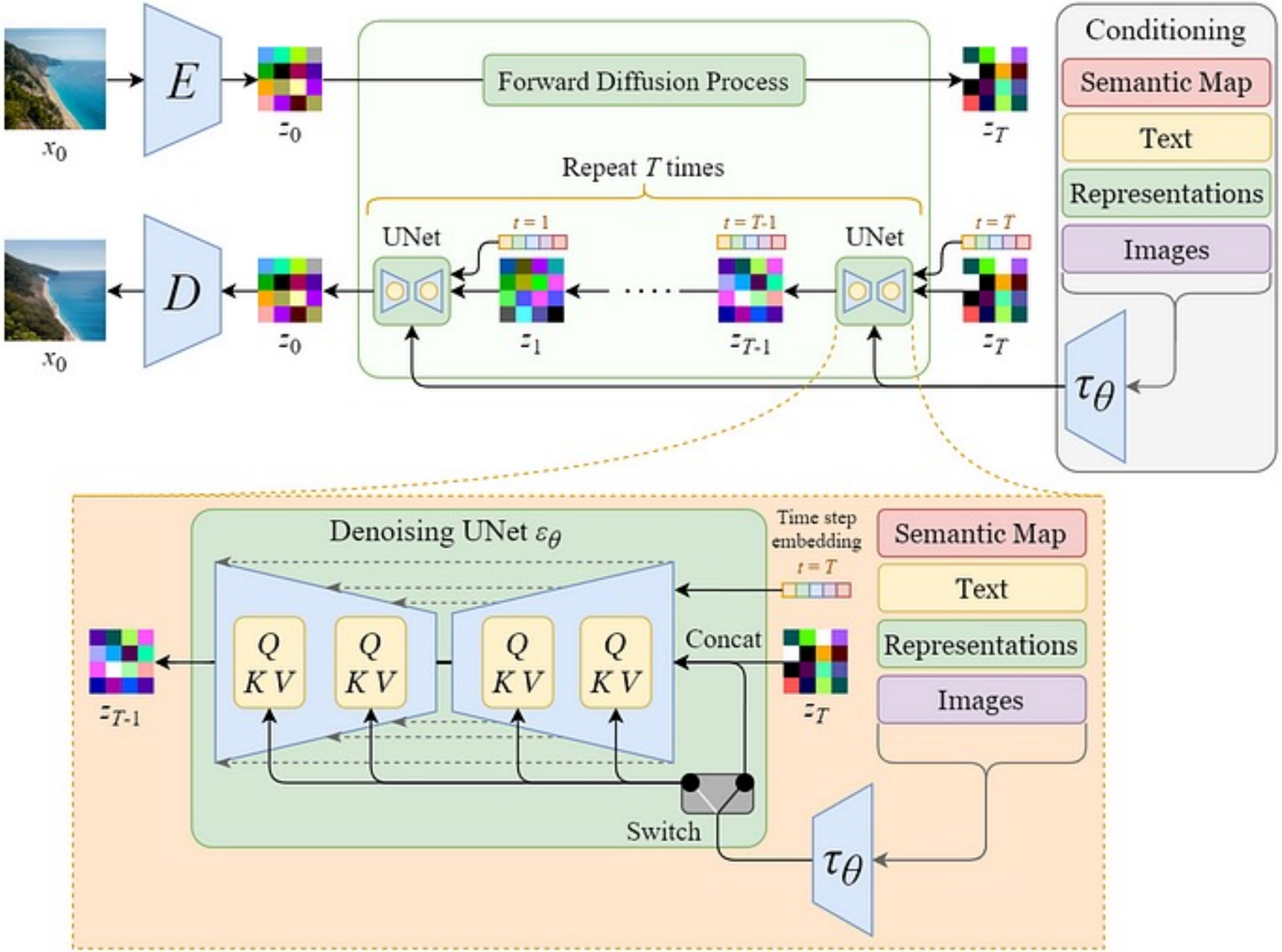
Text-to-Image using SDM

Stable Diffusion Model – Text Condition



BERT
CLIP

Summary


x_0
E
z_0
Forward Diffusion Process
z_T
Repeat T times
UNet
$t=1$
$t=T\text{-}1$
UNet
$t=T$
x_0
D
z_0
z_1
$z_{T\text{-}1}$
z_T
Conditioning
Semantic Map
Text
Representations
Images
τ_θ
Denoising UNet ε_θ
Time step embedding
$t=T$
Q K V
Q K V
Q K V
Q K V
Concat
$z_{T\text{-}1}$
z_T
Switch
Semantic Map
Text
Representations
Images
τ_θ

Thanks!

Any questions?