AI LAB PPT

STABLE DIFFUSION

PRESENTATION

N_Koda

2025

STUDIO SHODWE

brief description

- A Generative AI Model specifically designed for image generation.
- It is based on the Latent Diffusion Model (LDM) architecture.
- An incredibly popular and widely adopted Open Source system, trained on vast datasets of paired images and texts.

STUDIO SHODWE

Model Uses

The primary function of the Stable Diffusion model is Text-to-Image Generation.
It is used to generate art, concept designs, photorealistic images, and to manipulate existing images based on the provided text descriptions.

It is used by

VIRTUAL REALITY

Mage.space is an AI image generation platform often recognized for providing a frequently free and unlimited service, and Stable Diffusion serves as its core engine. Mage.space leverages the open-source nature of Stable Diffusion to enable users to generate digital art via text prompts. The platform also offers access to a wide array of trained and custom models (such as SDXL) built upon the foundational Stable Diffusion architecture. This diversity allows users to select from various artistic styles, making Mage.space an accessible and straightforward destination for experiencing the power of Stable Diffusion without the need for complex local software installation.

AUGMENTED REALITY

Leonardo AI is a specialized platform for creating concept art and visual assets, relying heavily on modified and optimized Stable Diffusion models to deliver high-quality, detailed outputs. While the site offers its own powerful core model, the essence of the platform lies in its vast library of community-created models that have been "finetuned" on the Stable Diffusion base. This allows users to select models specifically trained for niche styles, such as game assets, anime, or hyper-realistic photography, establishing it as a powerful tool for professionals and hobbyists looking to exploit the maximum flexibility and capabilities of Stable Diffusion.

STUDIO SHODWE

The Architecture

Text Encoder

This component's role is to understand the user's prompt.

- It typically uses a pre-trained language model, like CLIP (Contrastive Language-Image Pre-training).
- Function: It converts the input text (your prompt) into numerical representations called embeddings. These embeddings serve as the "conditions" or instructions that guide the image generation process.

U-Net

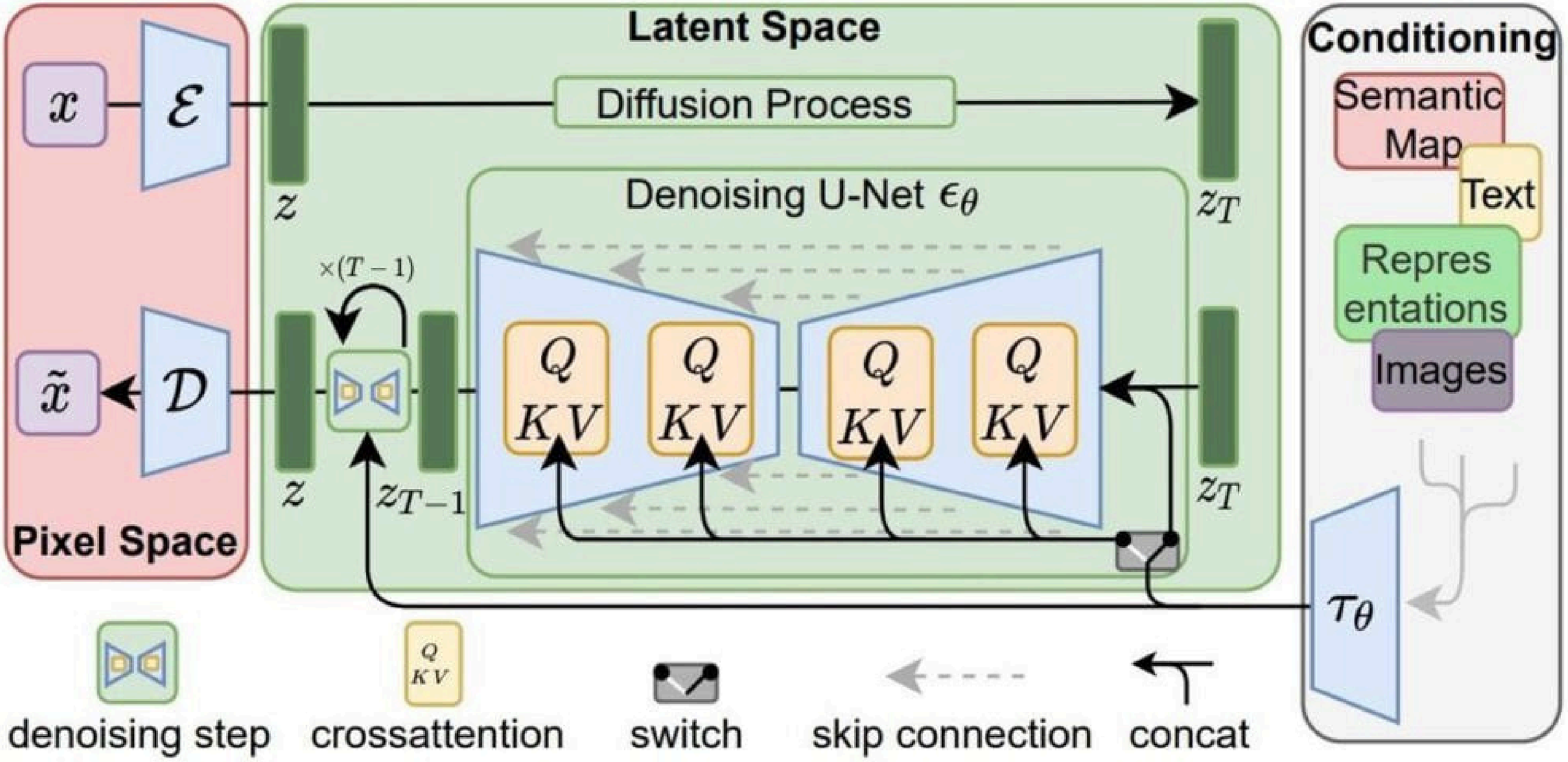
The U-Net is the engine of the model where the actual generation occurs.

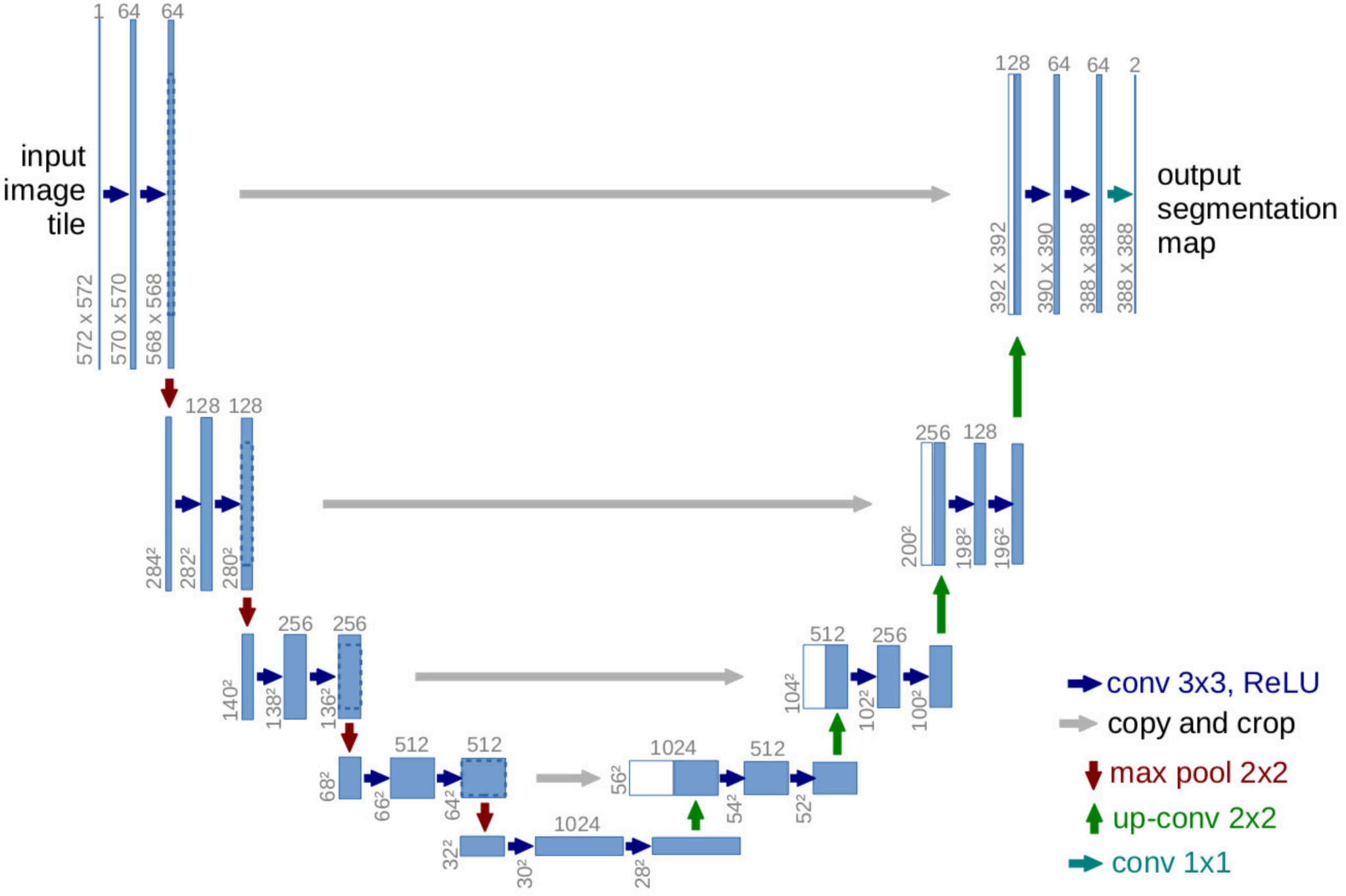
- Function: It takes random noise (in the Latent Space) and the text embeddings, and performs a multi-step Denoising process.
- In each step, the U-Net predicts and removes the noise, gradually transforming the random input into a coherent, meaningful image representation, conditioned by the text instructions.

(VAE) Autoencoder

The Variational Autoencoder (VAE) manages the high-dimensional data efficiently.

- Encoder: Compresses the original image data (if training or fine-tuning) from the pixel space into the smaller, optimized Latent Space. This is what makes the process fast.
- Decoder: Takes the final, denoised output from the U-Net (which is still in the Latent Space) and decompresses it back into a viewable, high-resolution Pixel Image for the user.


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


input image tile
572 x 572
570 x 570
568 x 568
1
64
64
128
128
284²
282²
280²
256
256
140²
138²
136²
512
512
68²
66²
64²
1024
32²
30²
28²
1024
512
56²
54²
52²
512
256
104²
102²
100²
256
128
200²
198²
196²
128
64
64
2
392 x 392
390 x 390
388 x 388
388 x 388
output segmentation map
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1

WINK WINK

THANK YOU!

N-KODA