Generating Images from Text

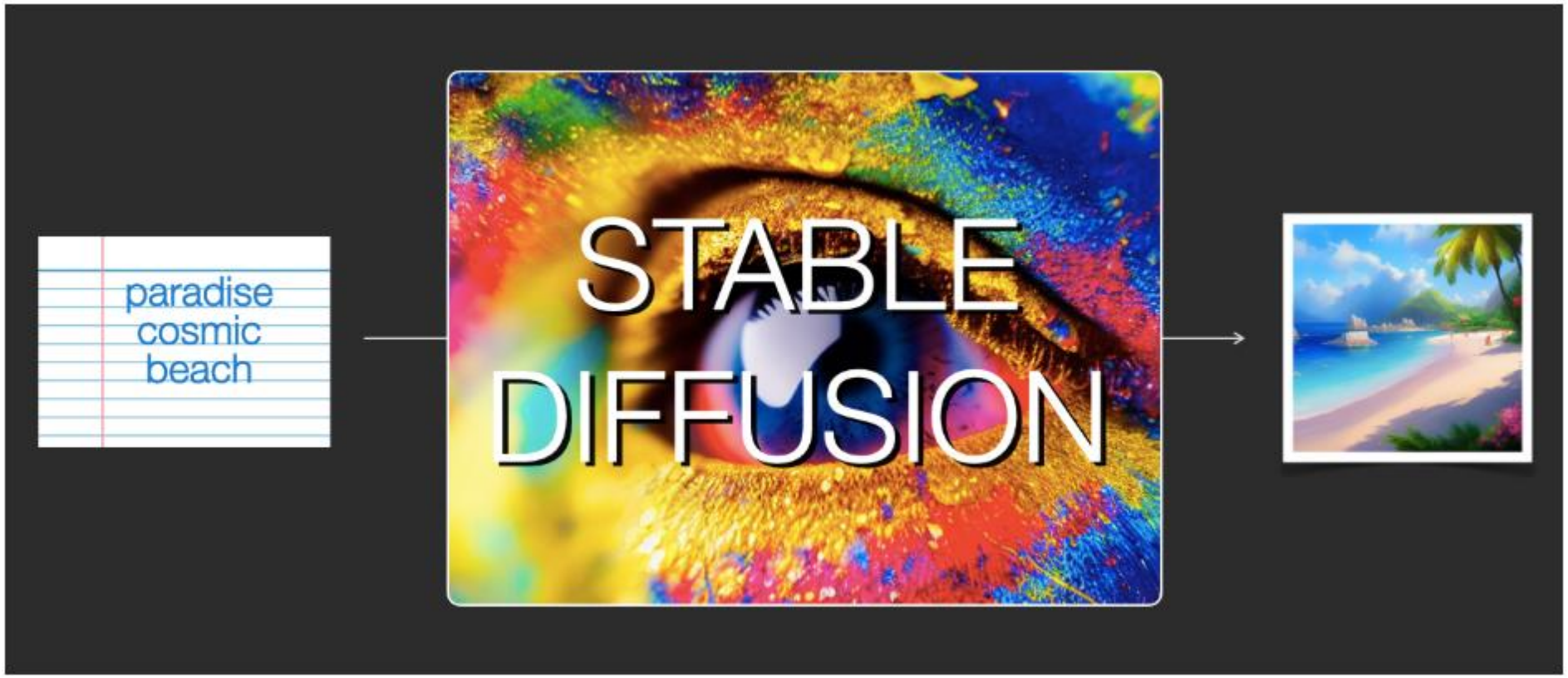
HOW DOES STABLE DIFFUSION WORK?

What is Stable Diffusion?

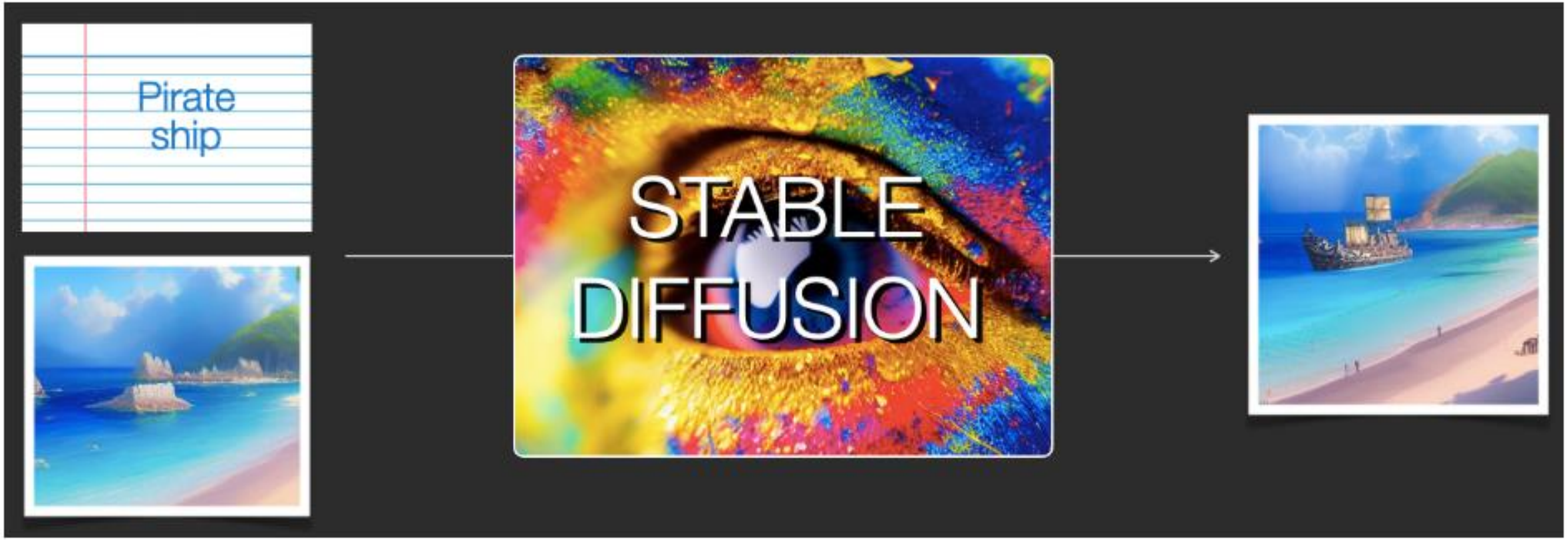
- Abilities:
  - Generates images from text
  - Modifies images based on text
  - Enhance low-resolution or low-detail images

- Deep-learning, text to image model
- Generate detailed images based on text descriptions
- Stable Diffusion v2.1 & Stable Diffusion v3.5

How Stable Diffusion Works



How Stable Diffusion Works



Components of Stable Diffusion

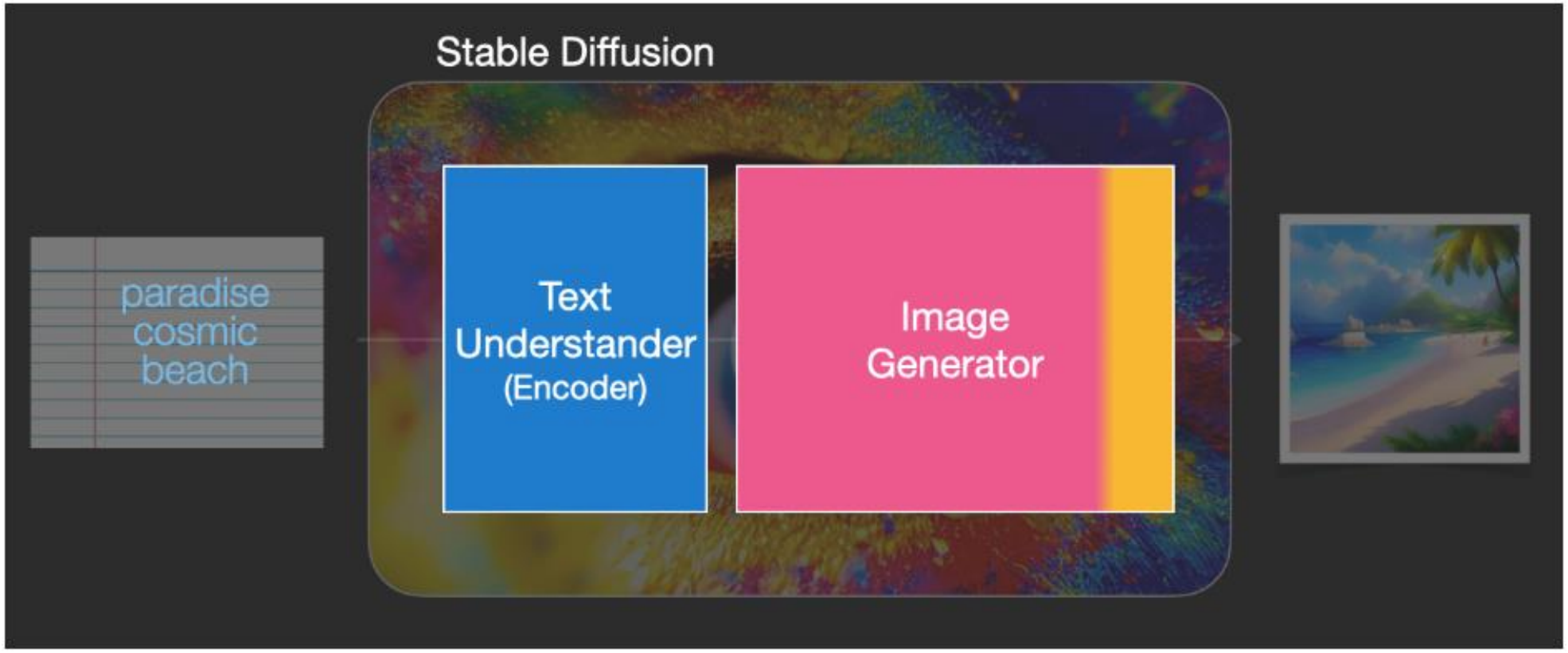


Image Generator

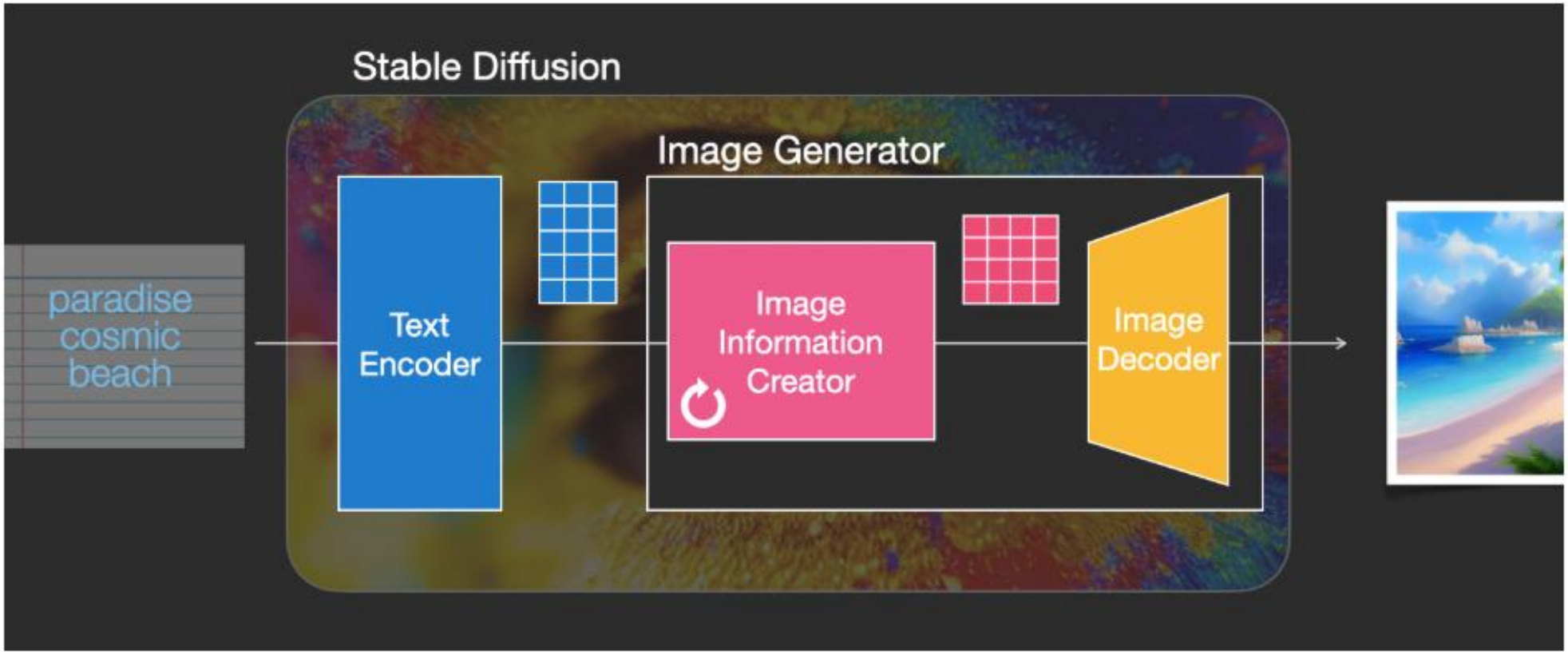
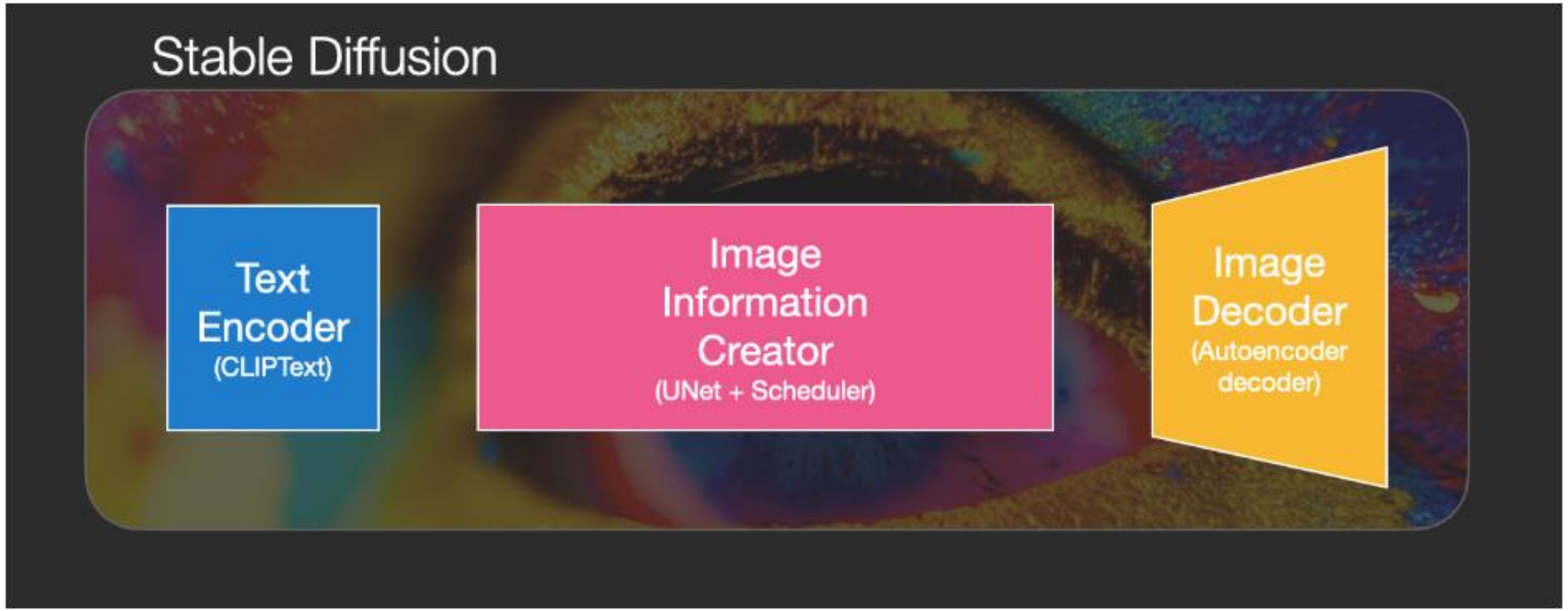
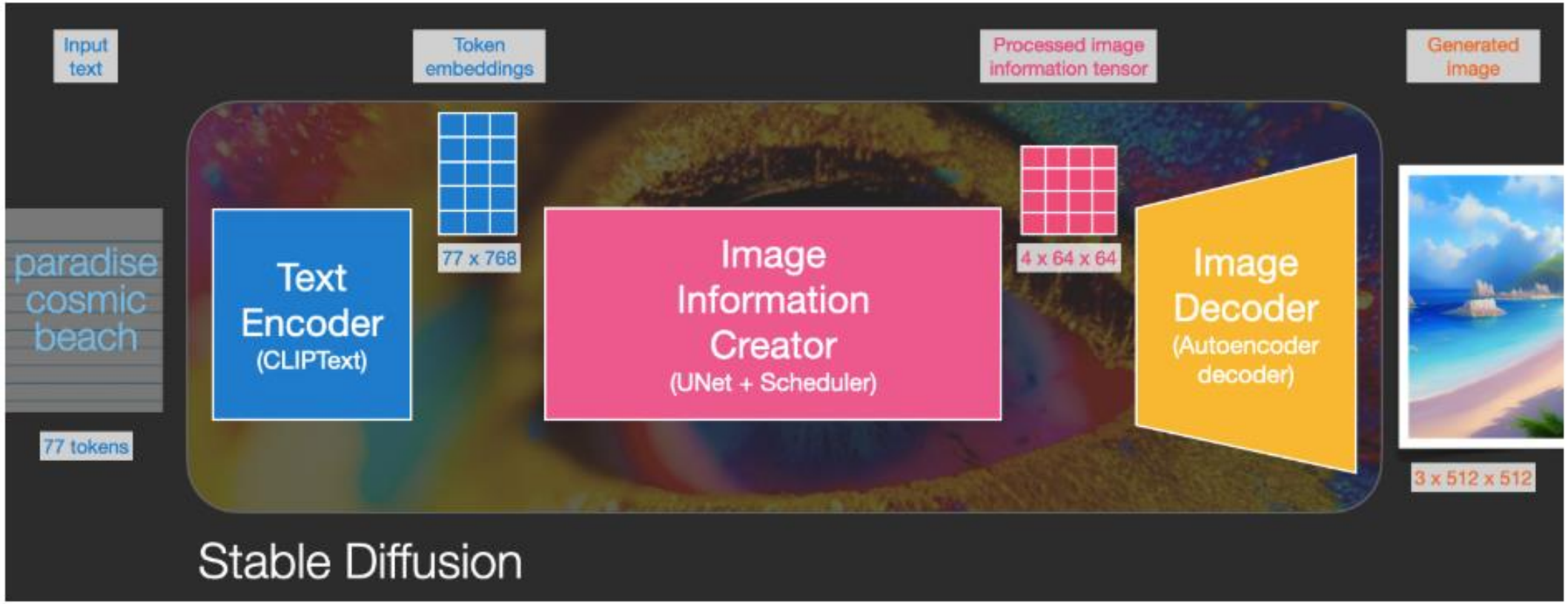


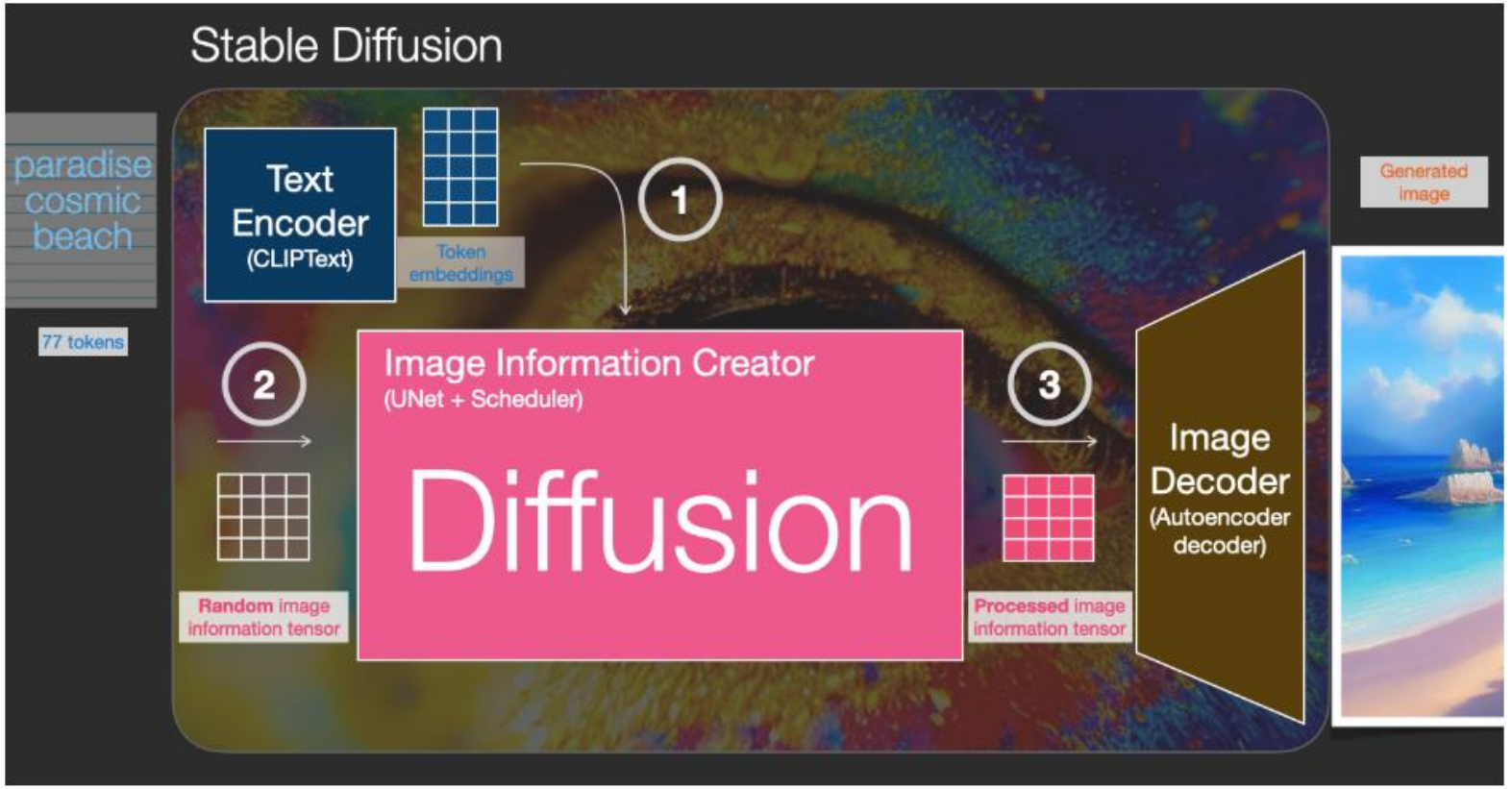
Image Decoder



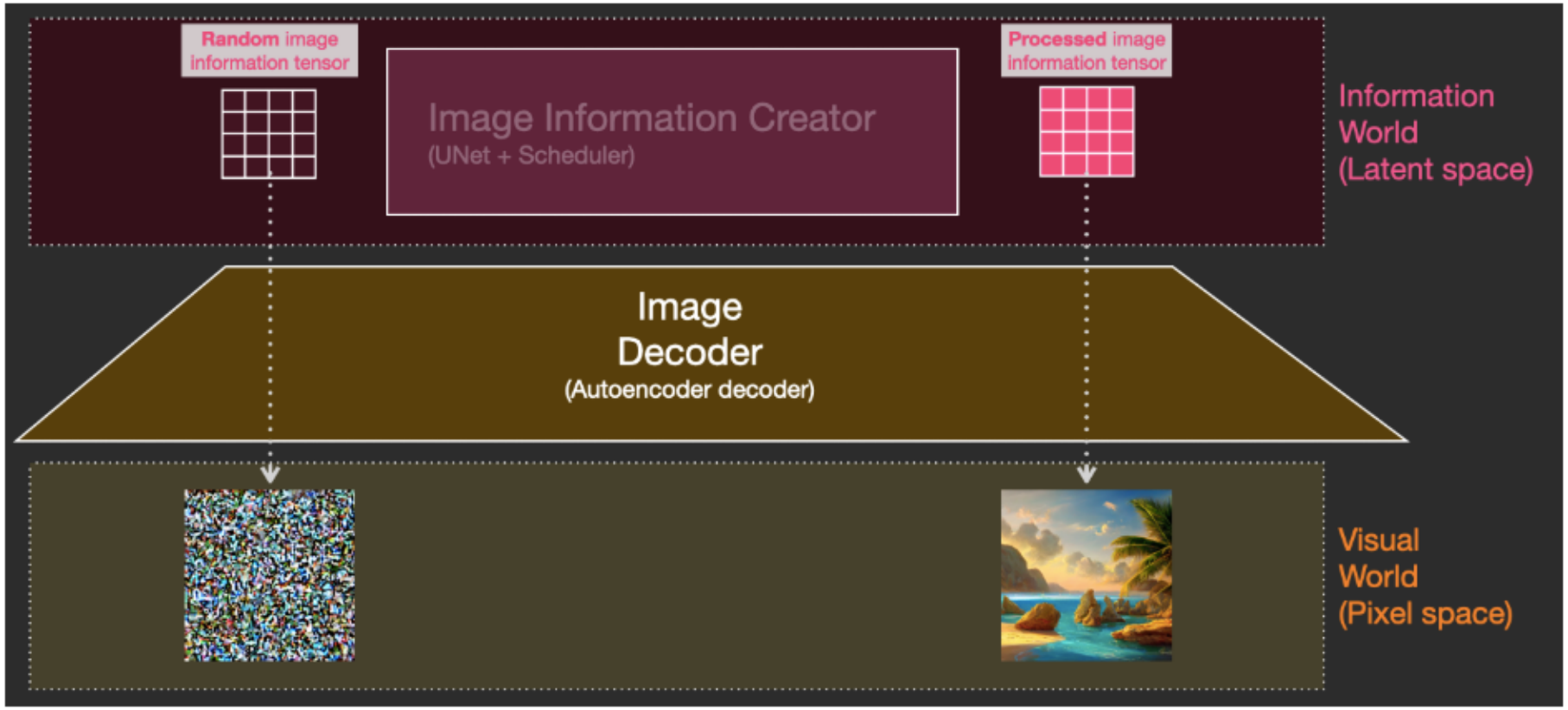
Overview of Main Components



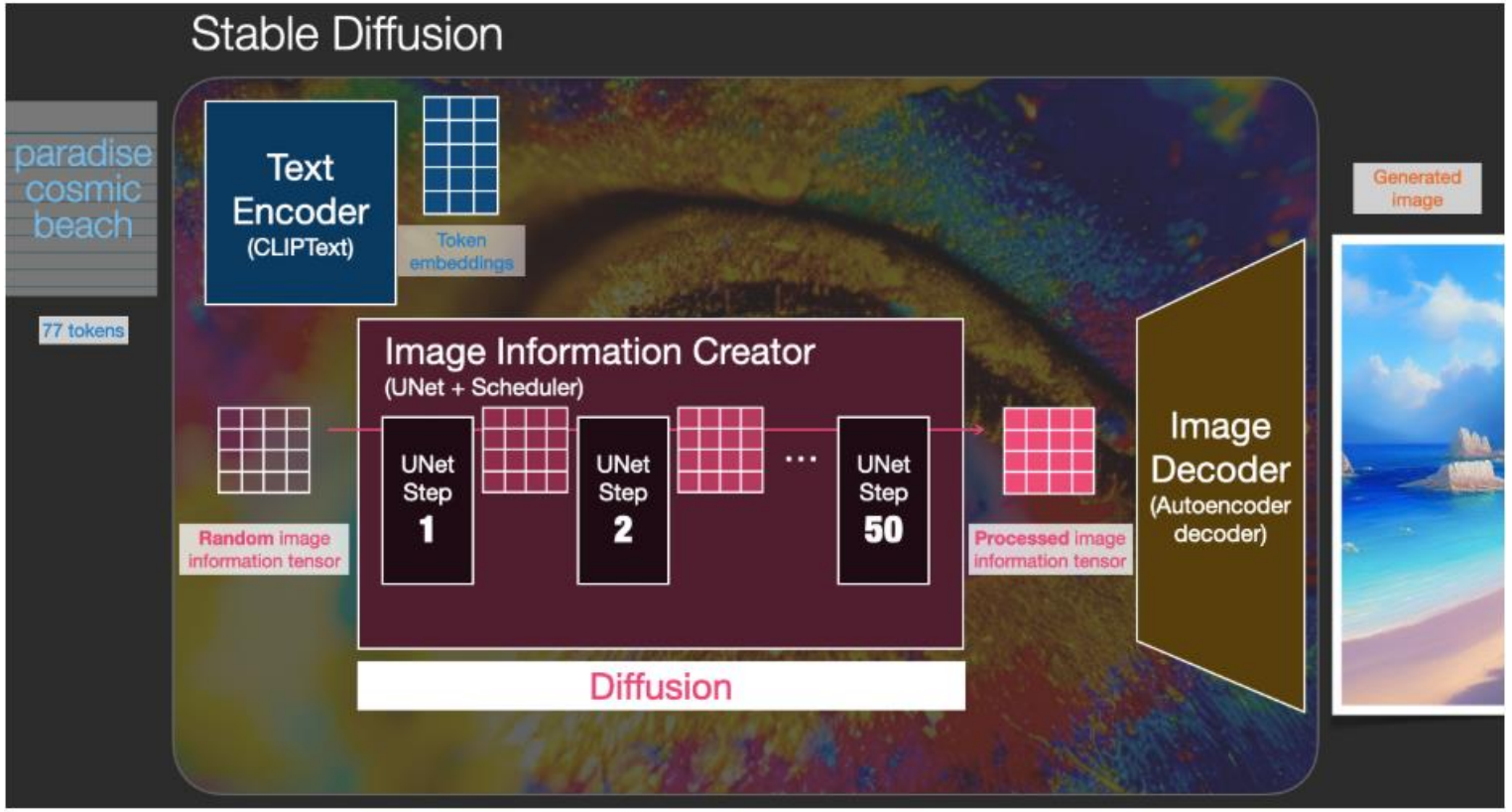
Diffusion Recap



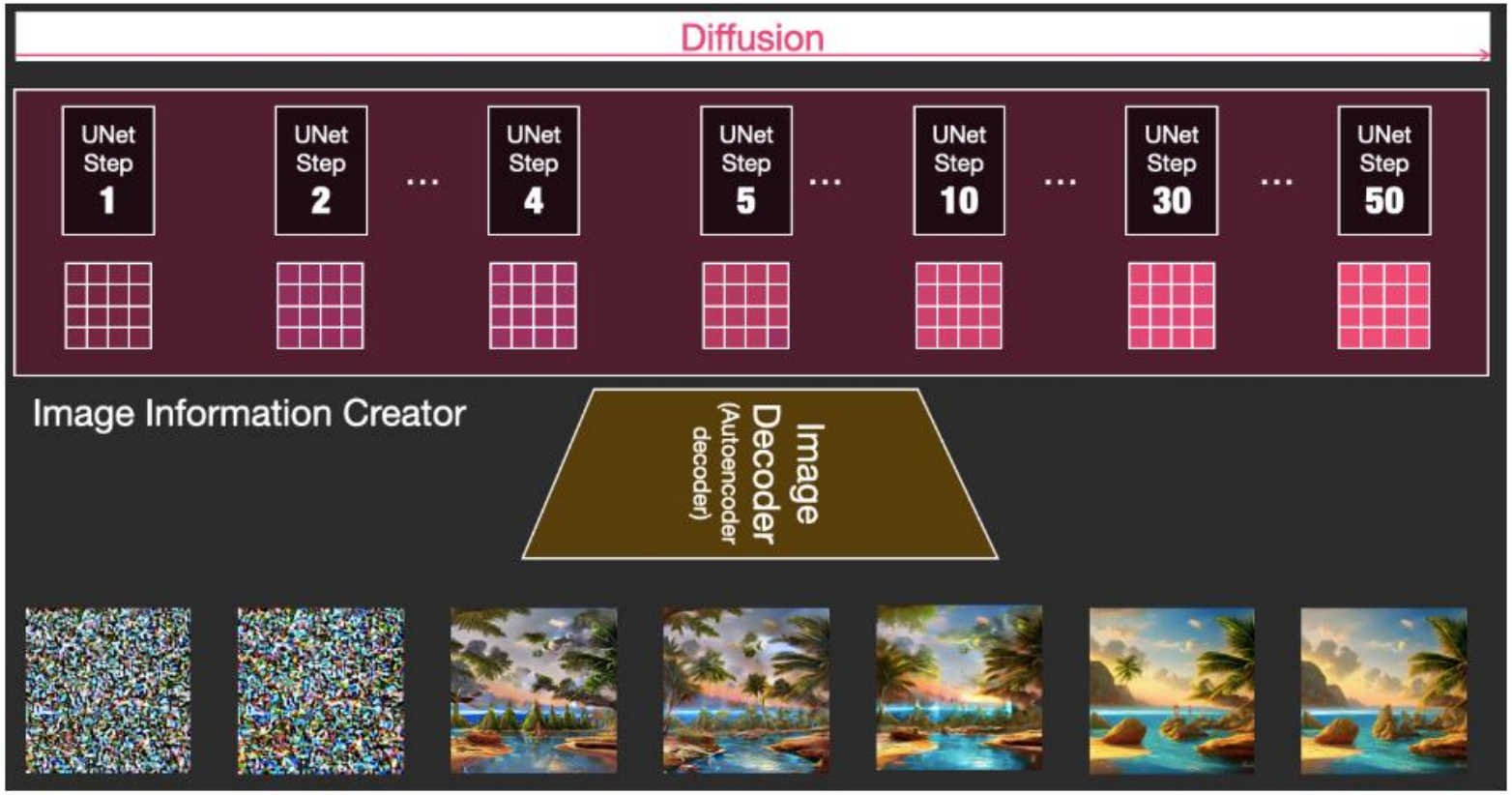
SD: Step-by-Step Process



Diffusion


Stable Diffusion
paradise cosmic beach
77 tokens
Text Encoder (CLIPText)
Token embeddings
Image Information Creator
(UNet + Scheduler)
UNet Step 1
UNet Step 2
UNet Step 50
Random image information tensor
Processed image information tensor
Image Decoder (Autoencoder decoder)
Generated image
Diffusion

Visualize Set of Latents



Diffusion Process



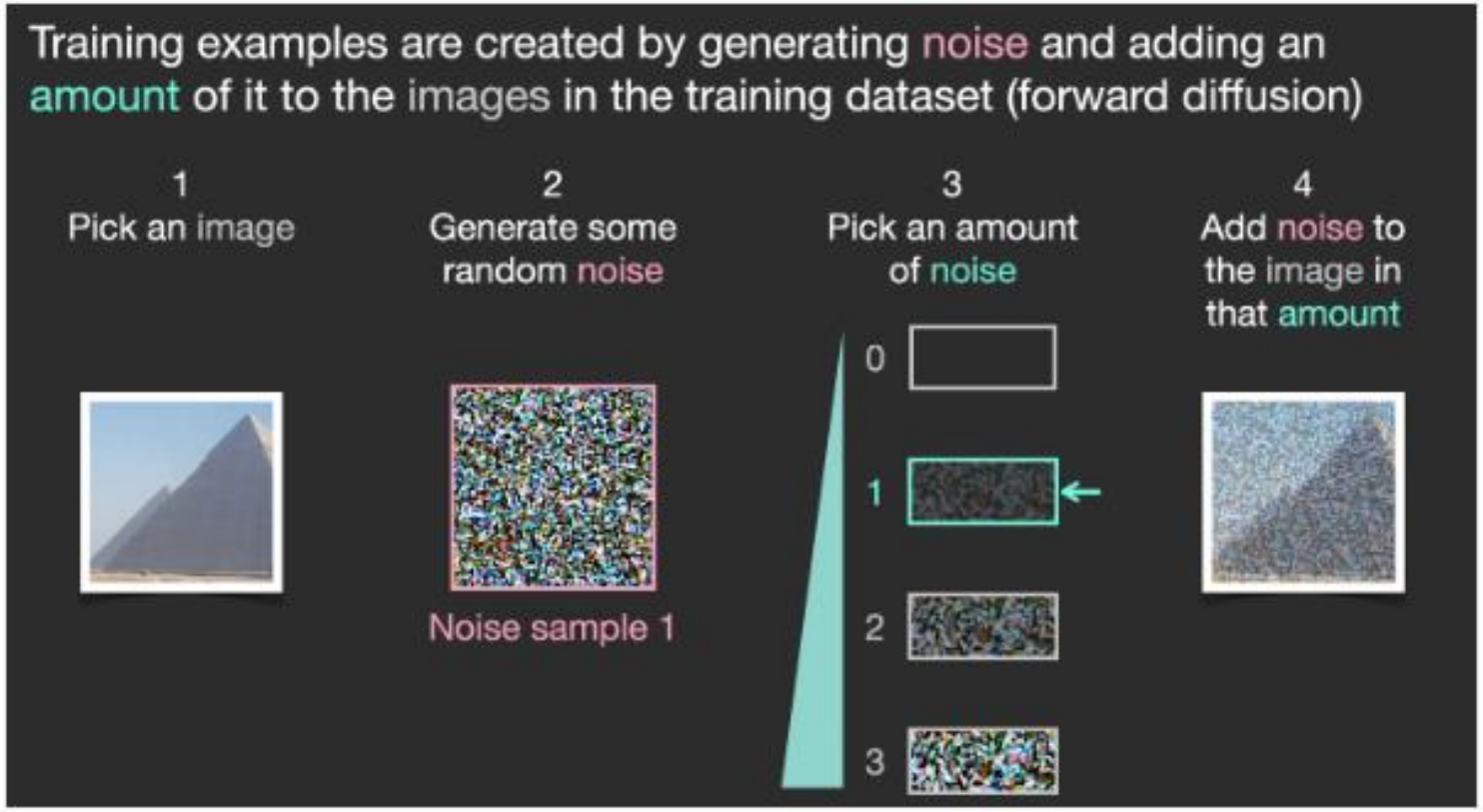
Steps 2 and 4



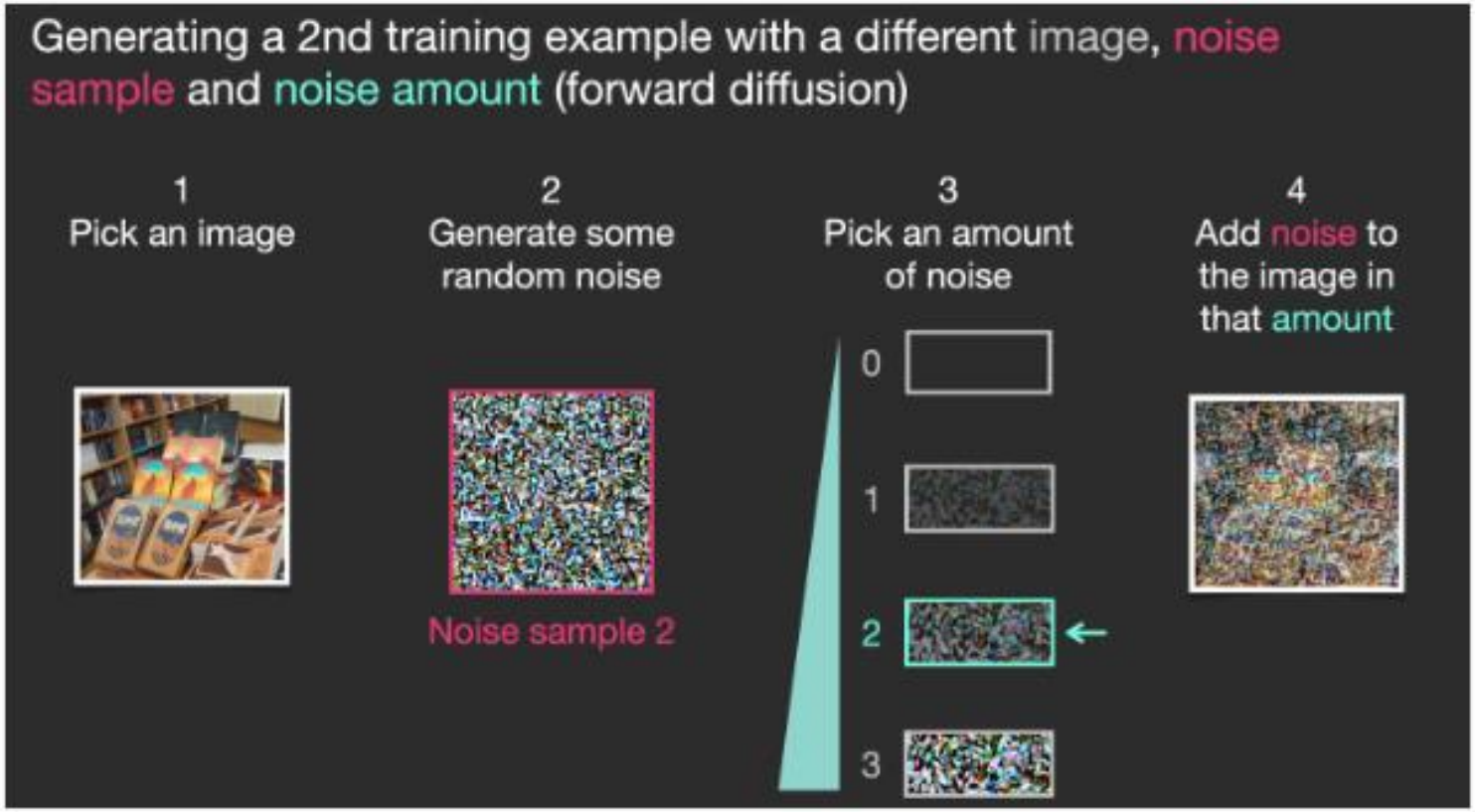
Another example of the diffusion process



Diffusion Process



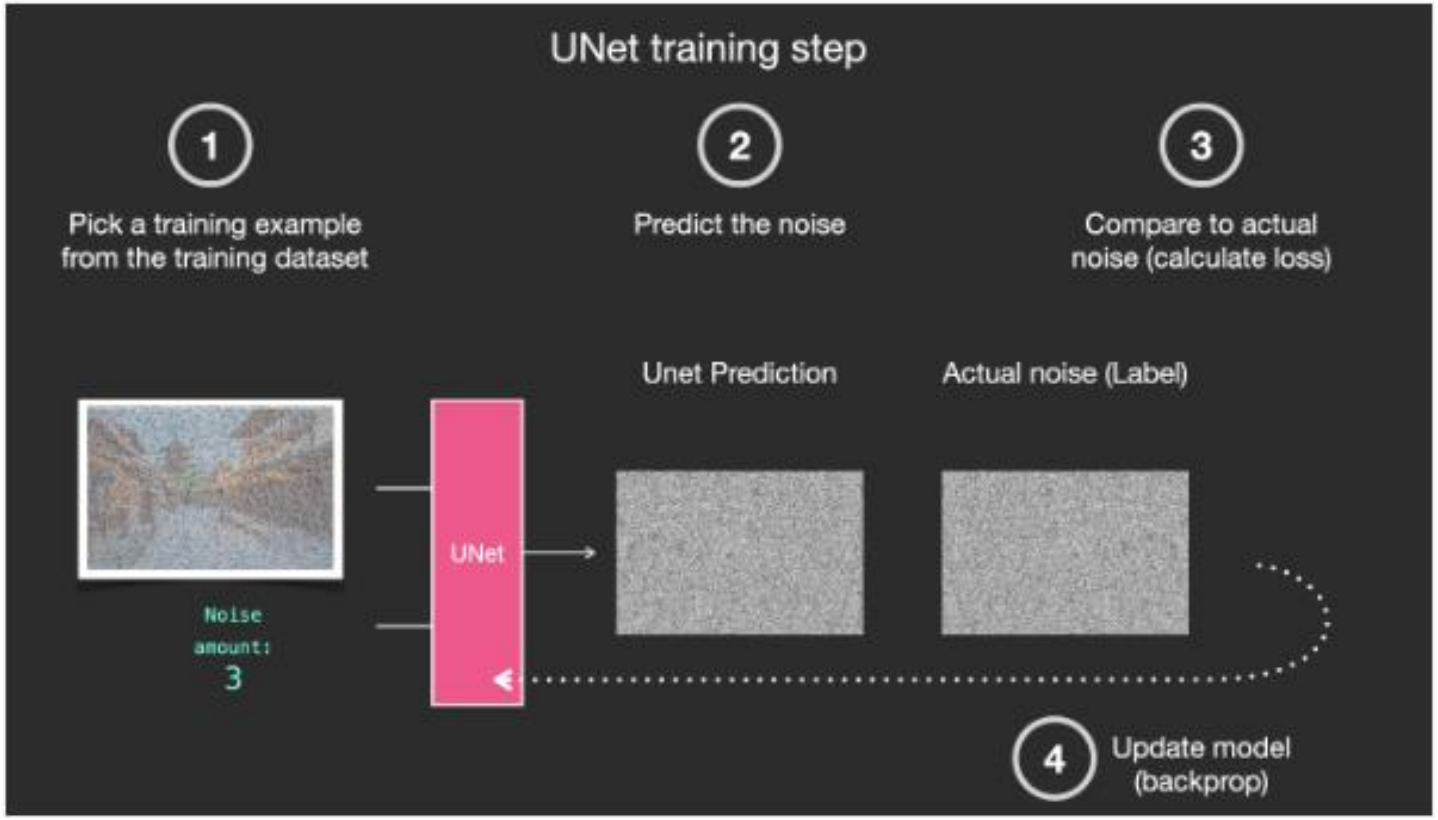
Training Example



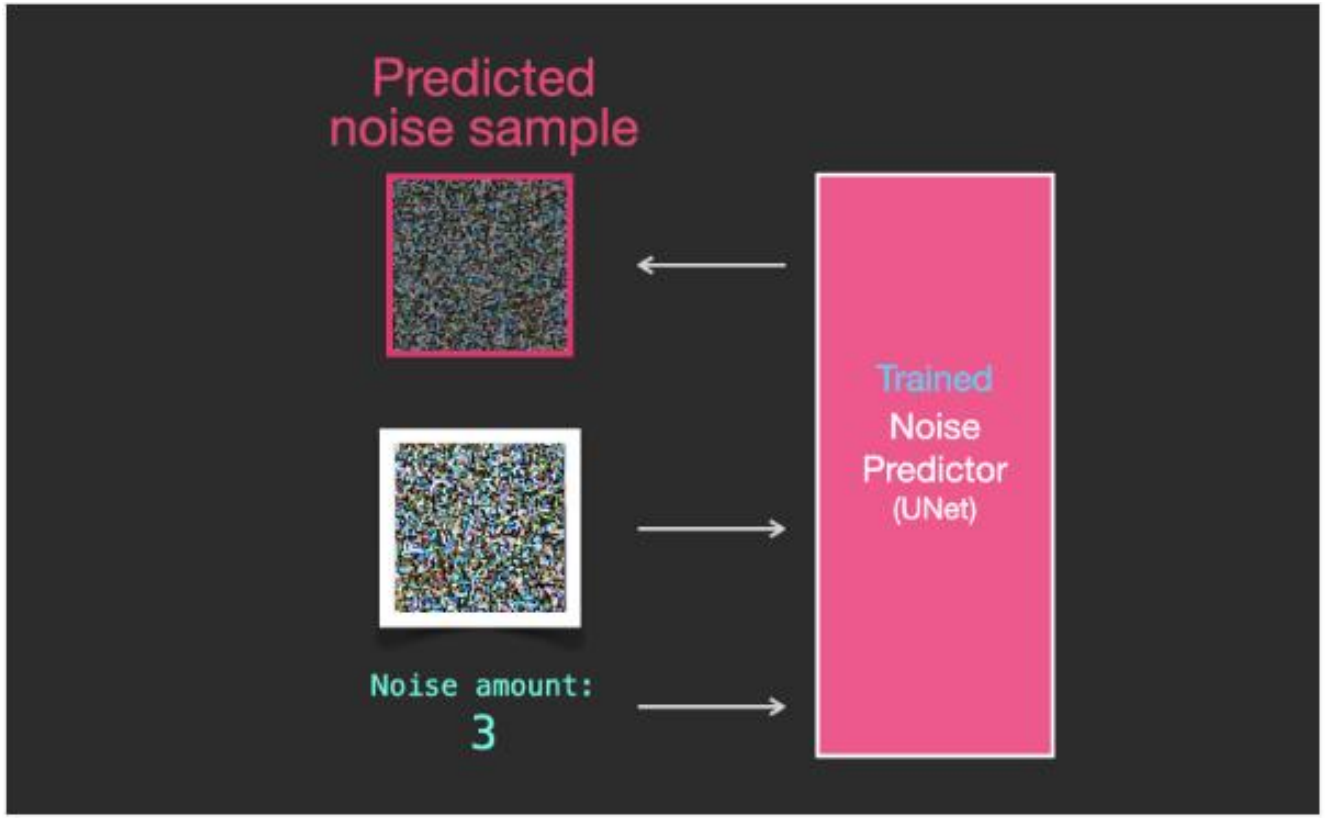
Generating Training Examples



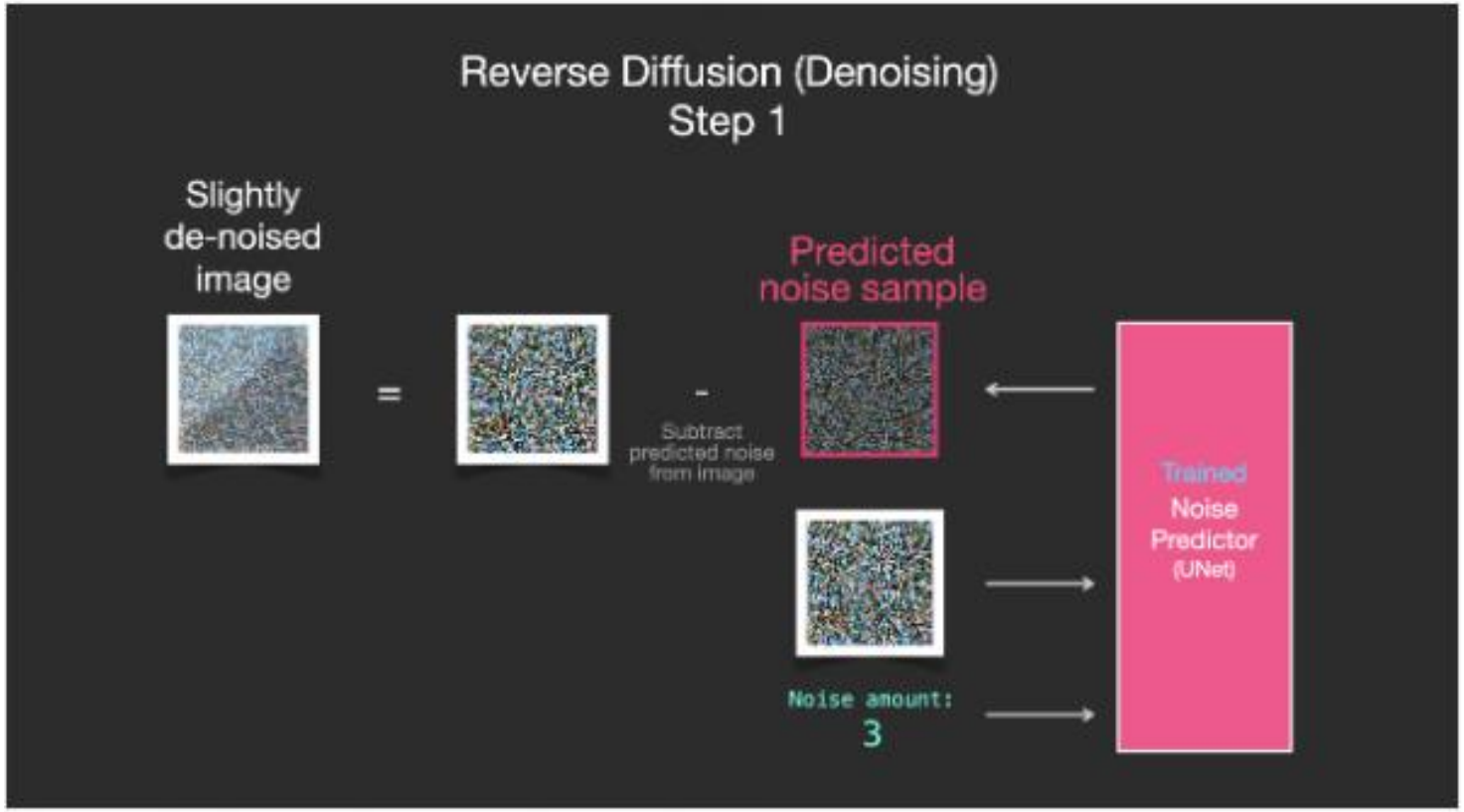
Noise Predictor



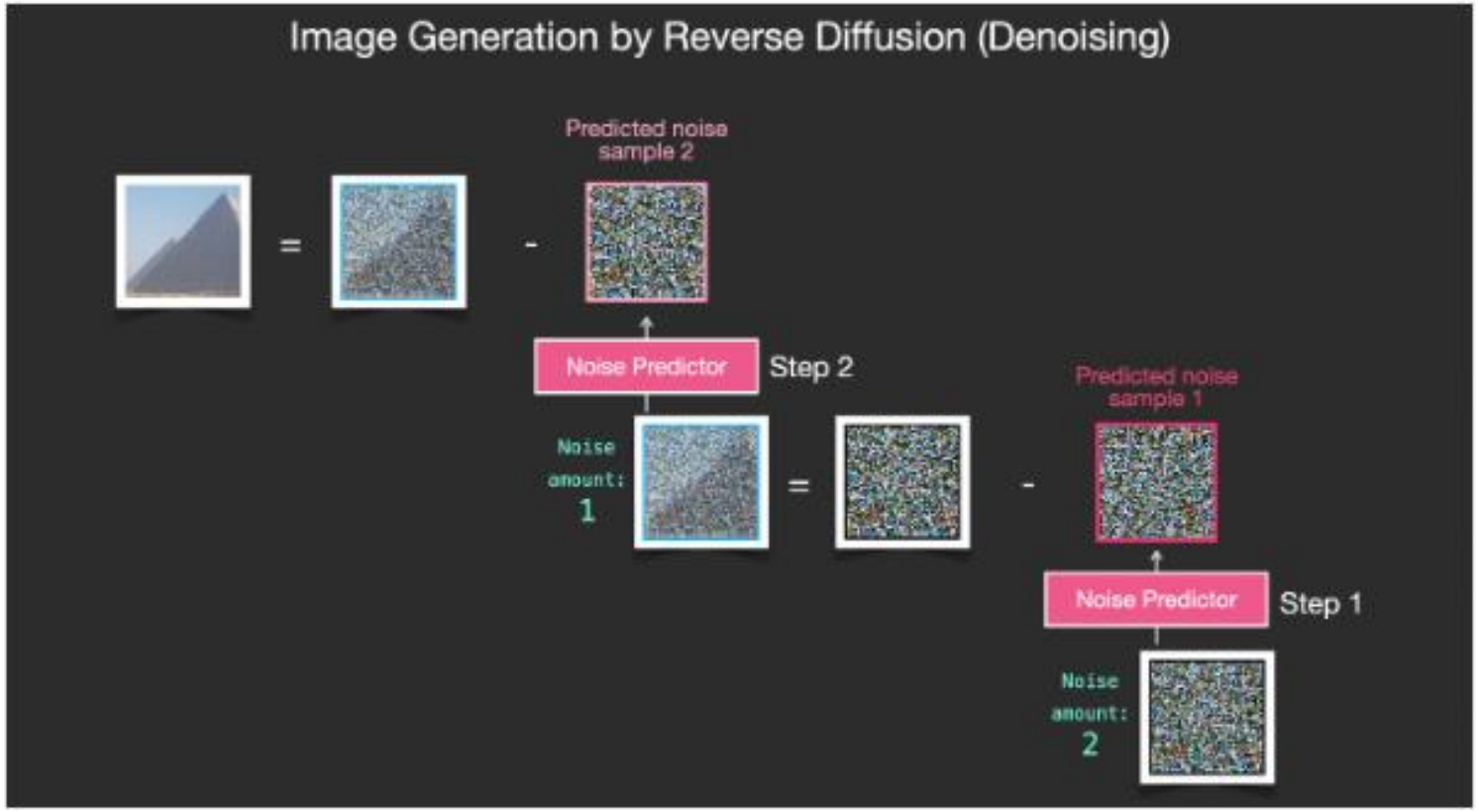
Removing noise to get an image



Denoising


Reverse Diffusion (Denoising)
Step 1
Slightly de-noised image
=
-
Subtract predicted noise from image
Predicted noise sample
Noise amount: 3
Trained Noise Predictor (UNet)

Continuing to Denoise



Denoising Diffusion Probabilistic Models

- Paper published in December 2020
- Process used in Stable Diffusion, Dall-E, Imagen
- Generates images without text data

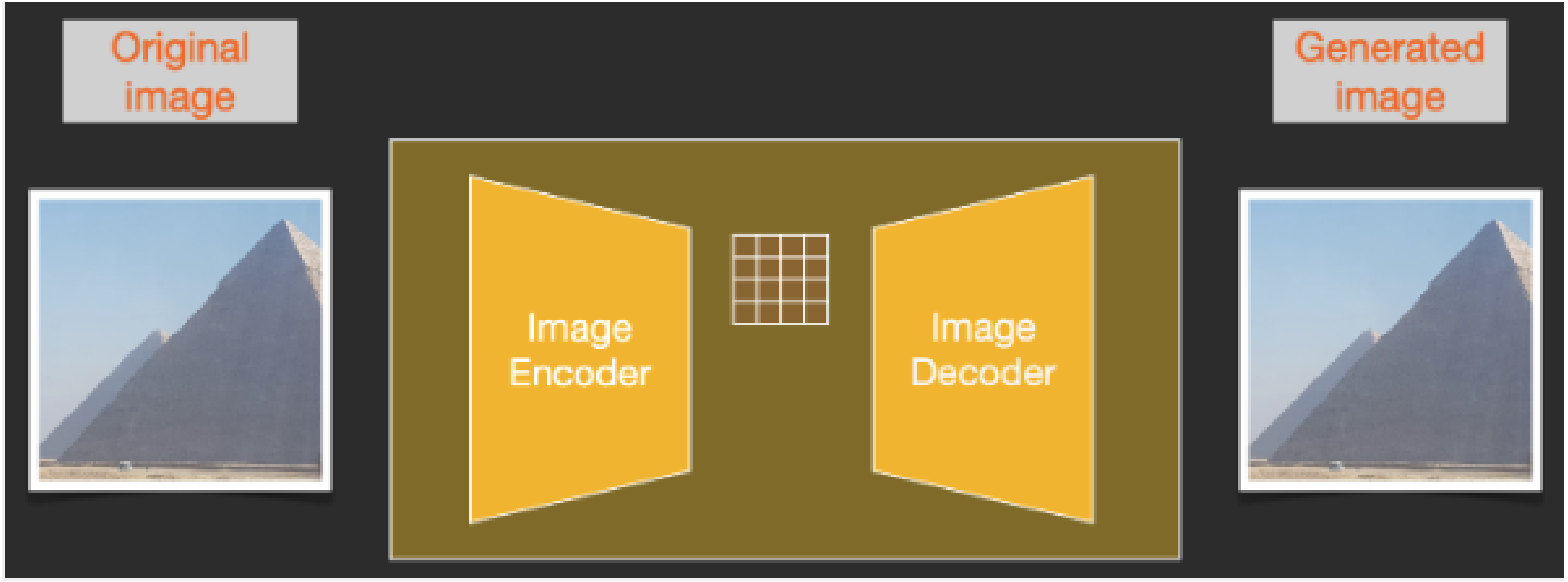
High-Resolution Image Synthesis with Laten Diffusion Models

- Paper published in 2021/2022 (update)

- Increase speed of image generation process

- Run diffusion on compressed version of image

Compression via Autoencoder



Forward Process

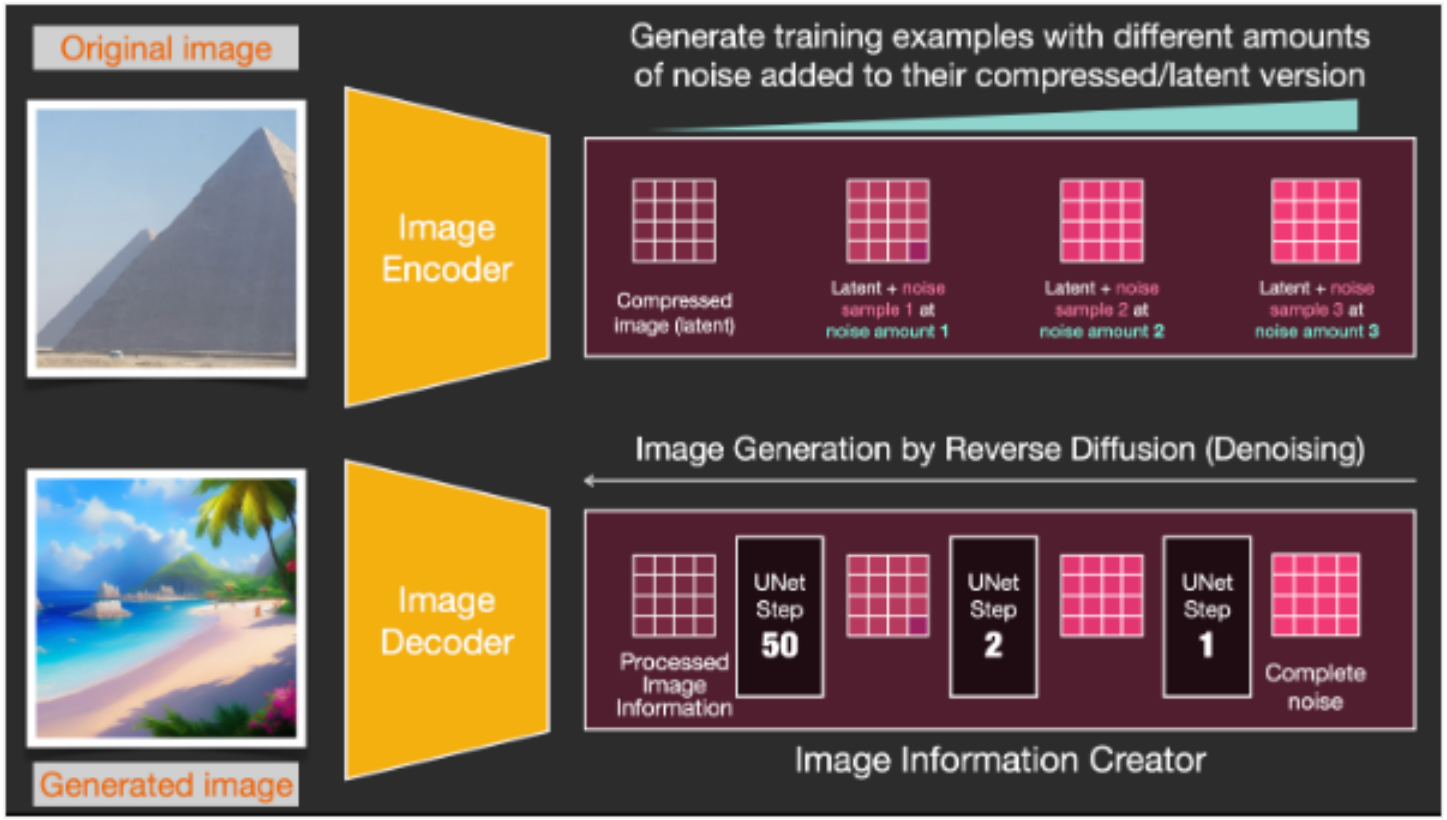
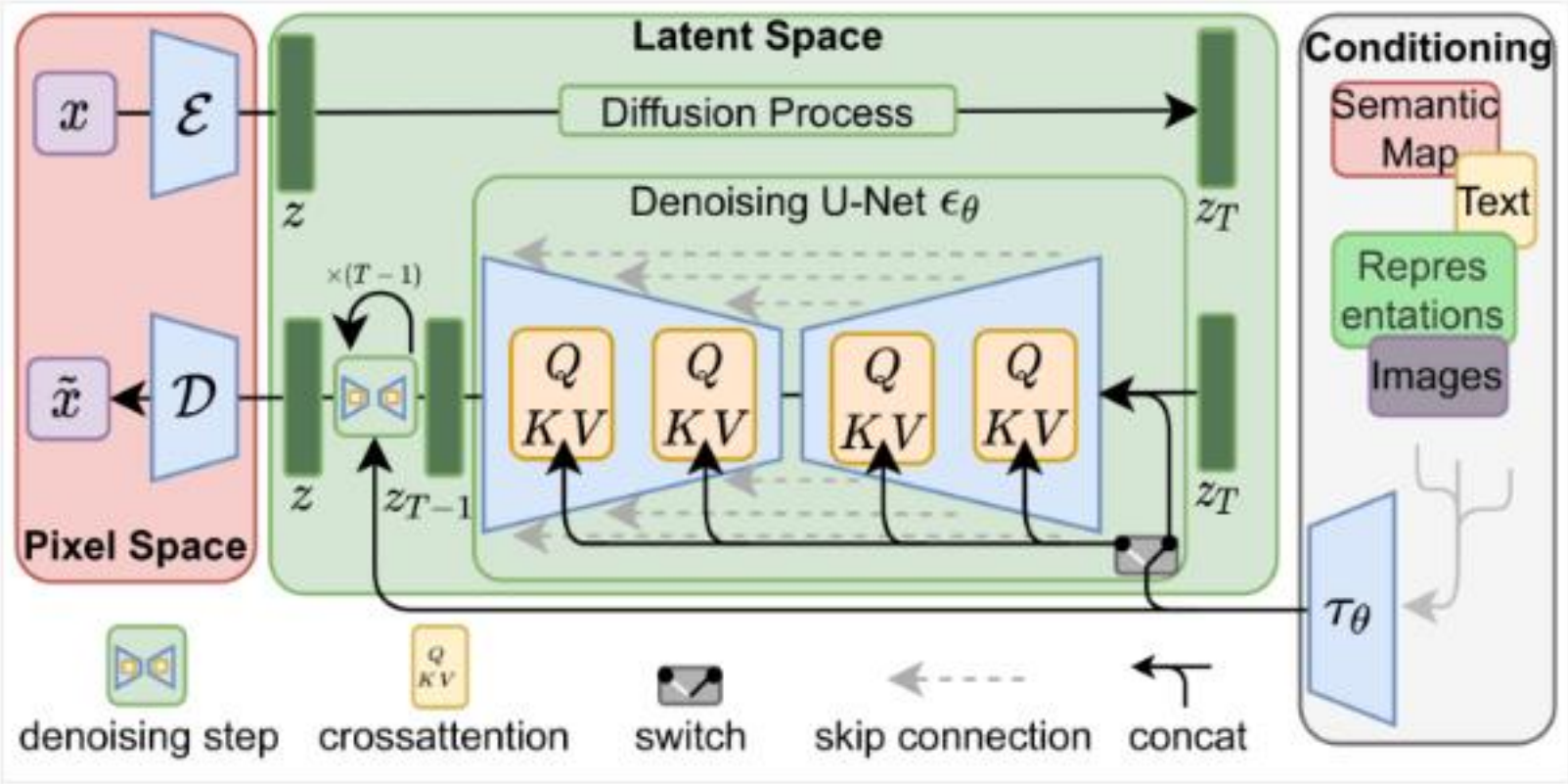


Figure 3. Latent Diffusion Models (LDMs)



Latent Variable Space

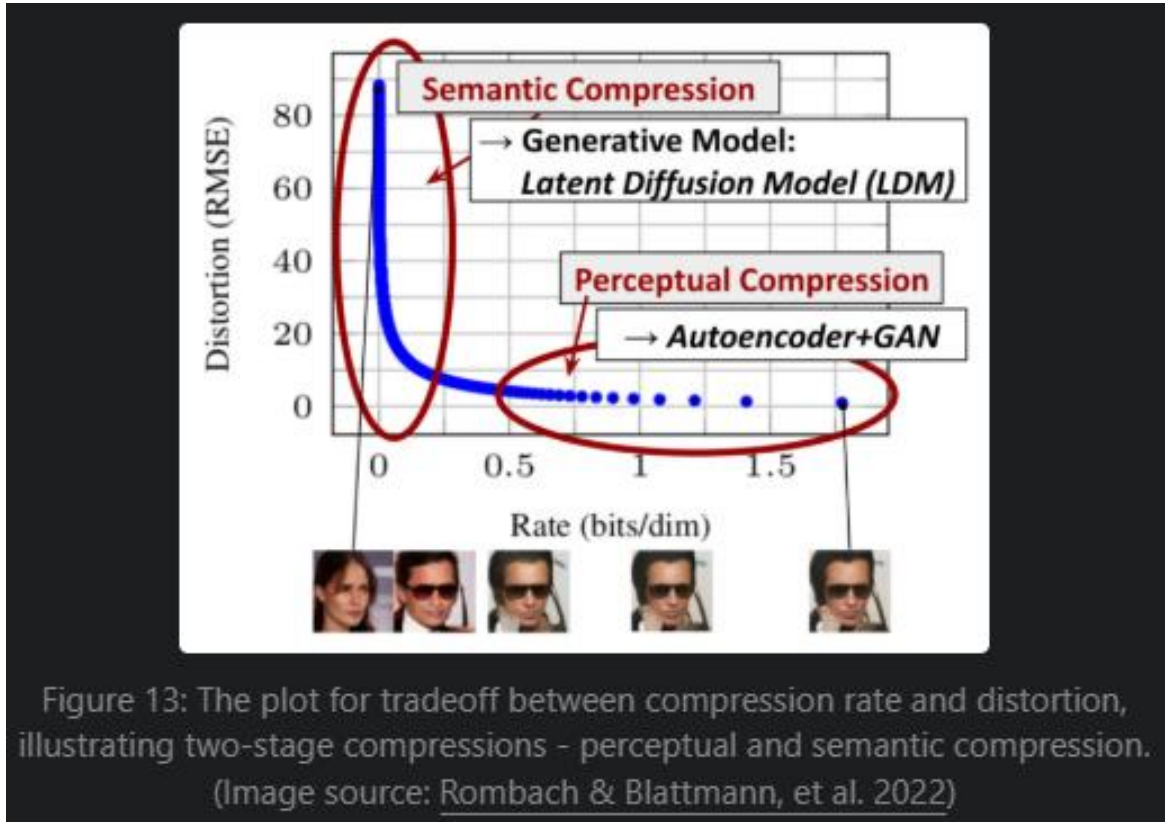
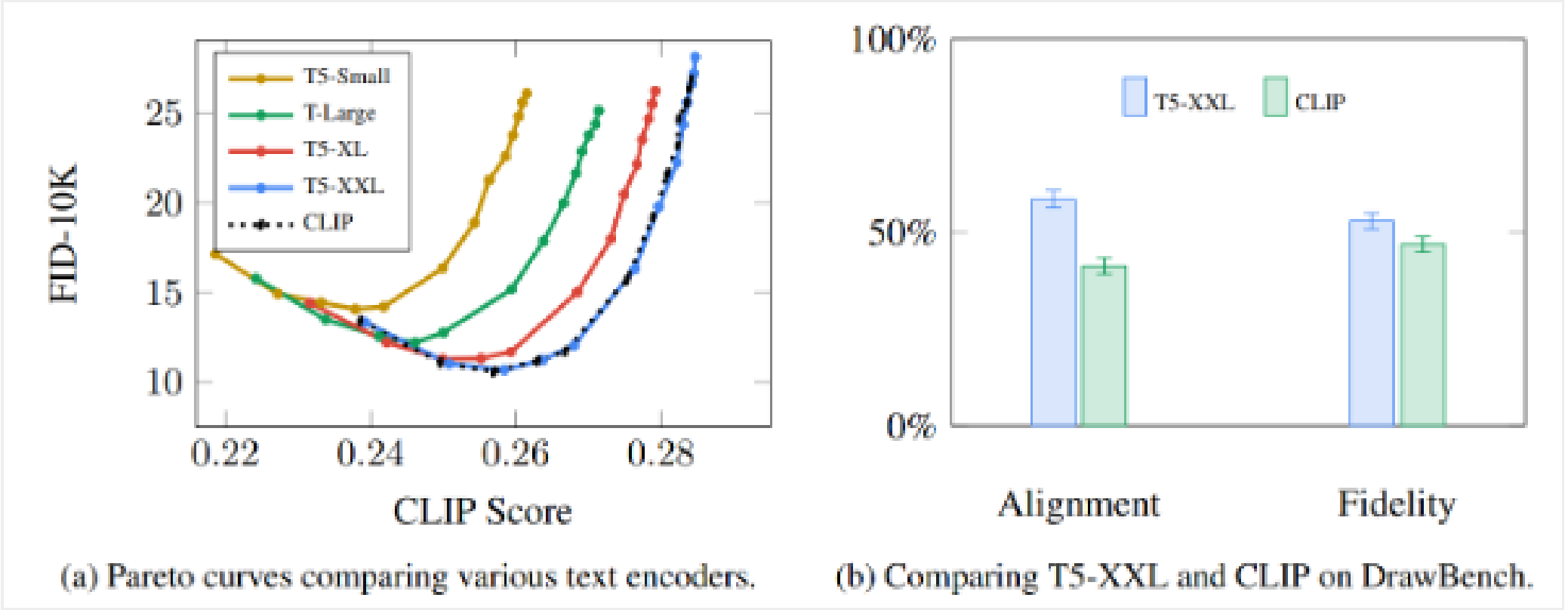


Figure 13: The plot for tradeoff between compression rate and distortion, illustrating two-stage compressions - perceptual and semantic compression. (Image source: Rombach & Blattmann, et al. 2022)

Text Encoder



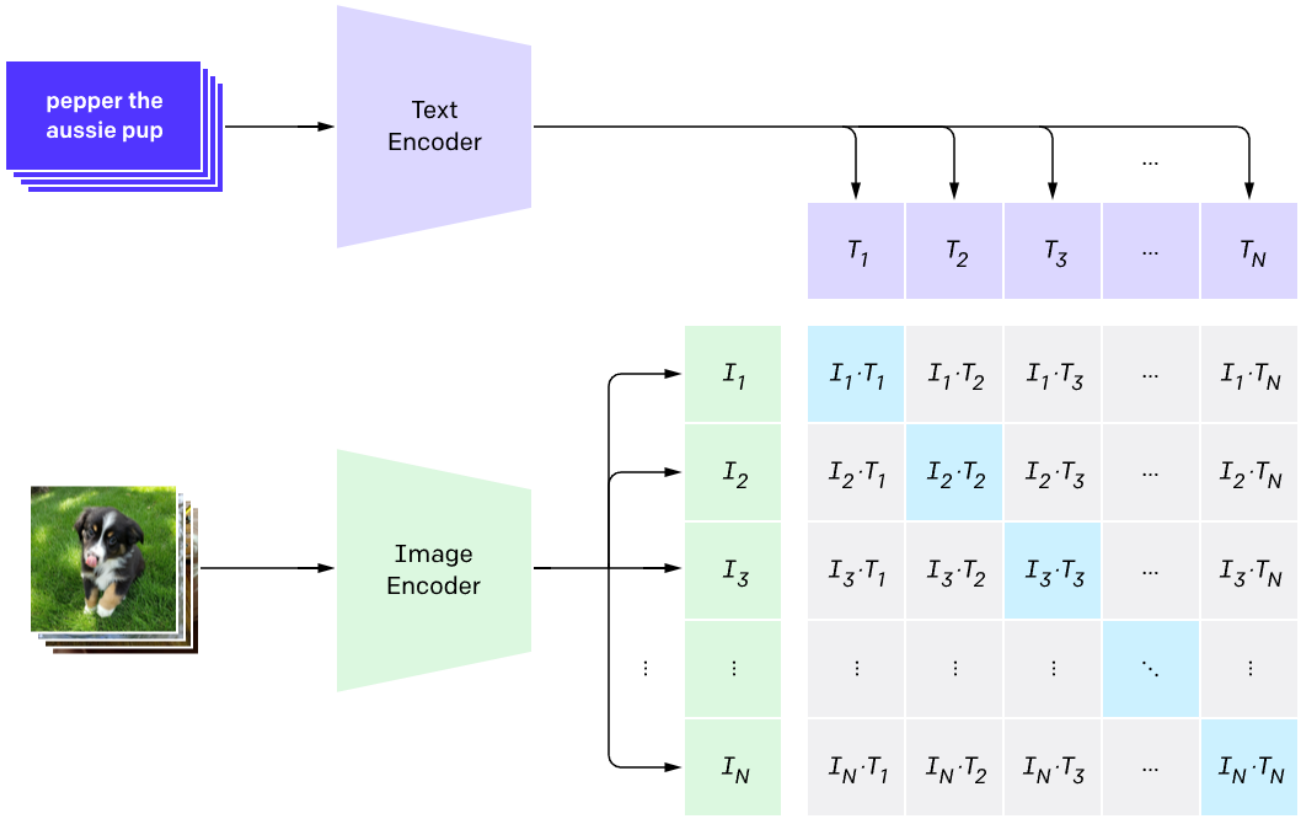
(a) Pareto curves comparing various text encoders.

(b) Comparing T5-XXL and CLIP on DrawBench.

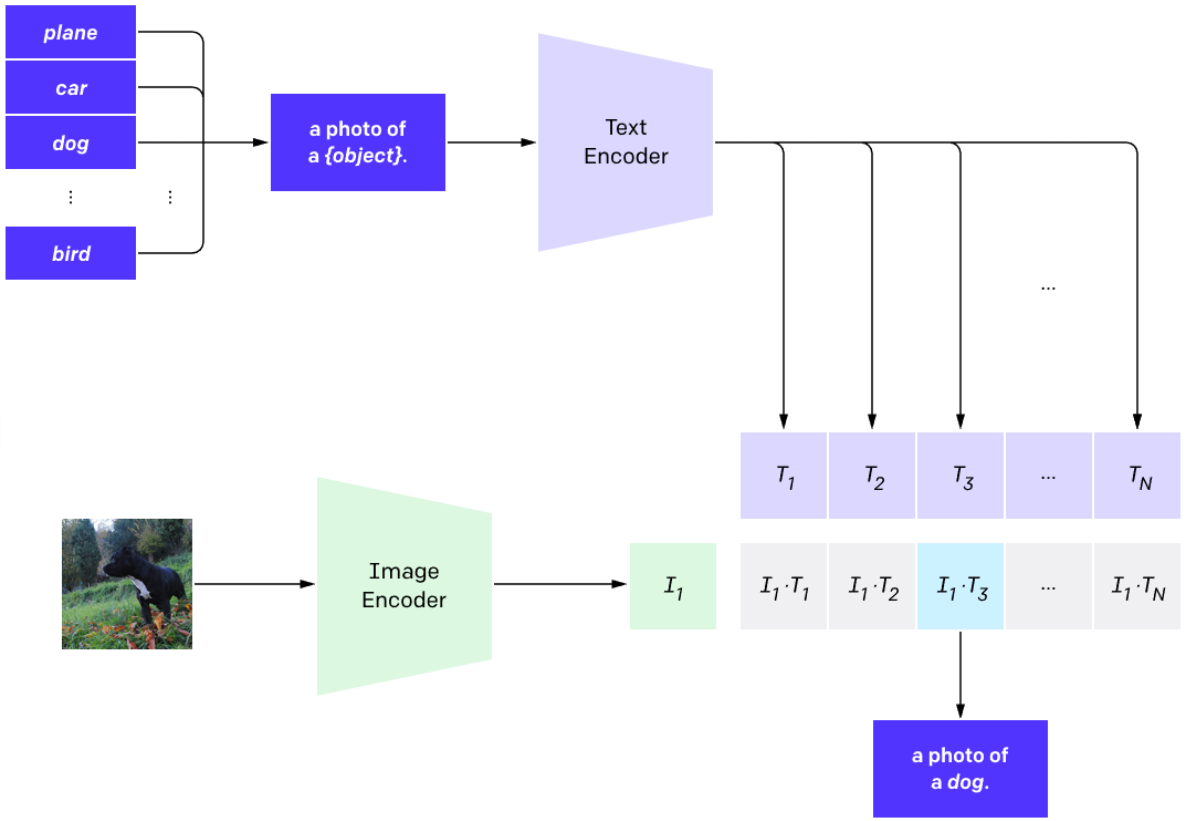
Training CLIP



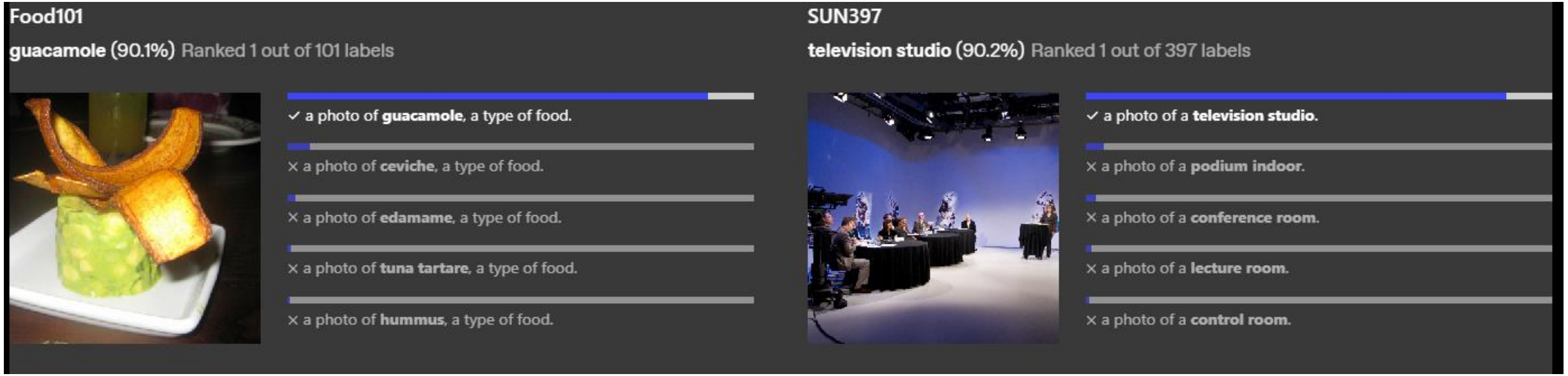
Contrastive pre-training



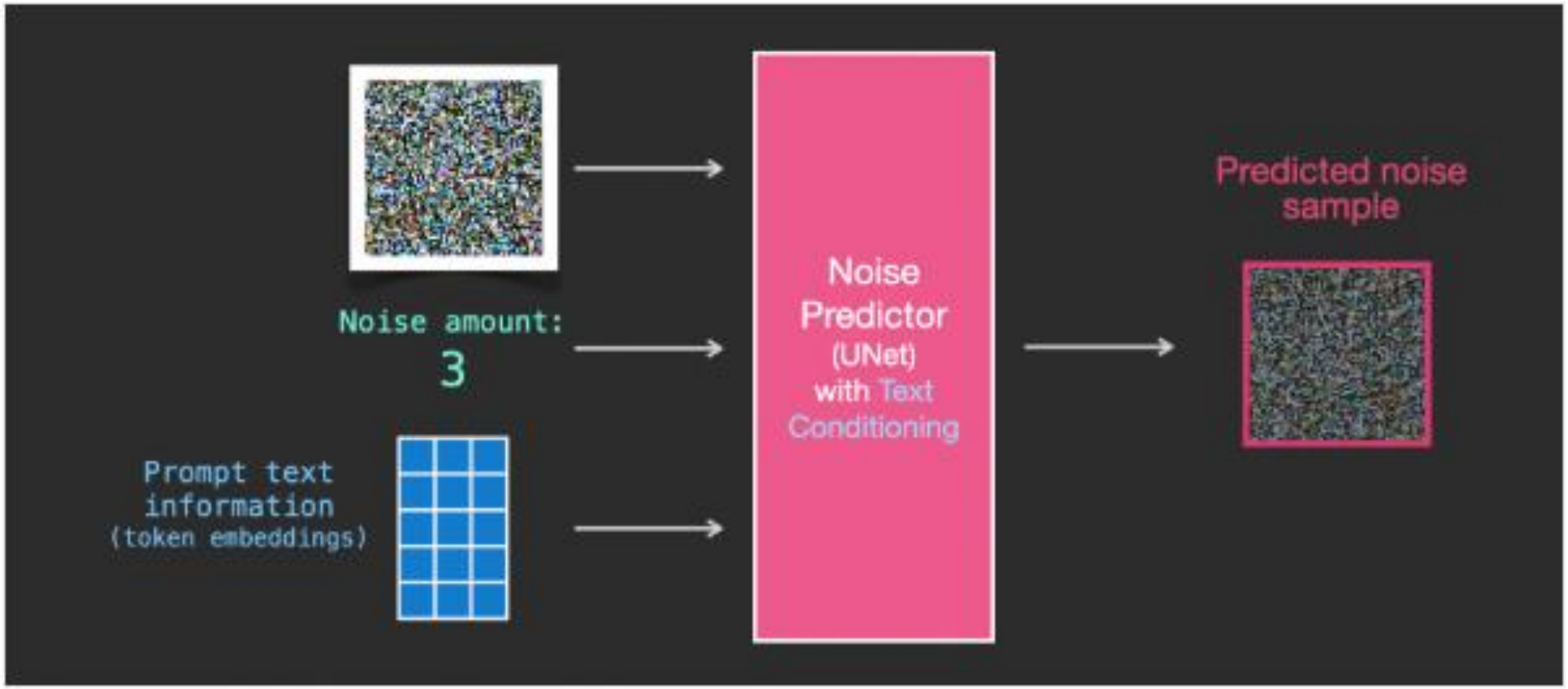
Create dataset classifier from label text



Examples of CLIP on various datasets



Feeding Text Information into Image Generation



Dataset now includes encoded text



U-Net Architecture

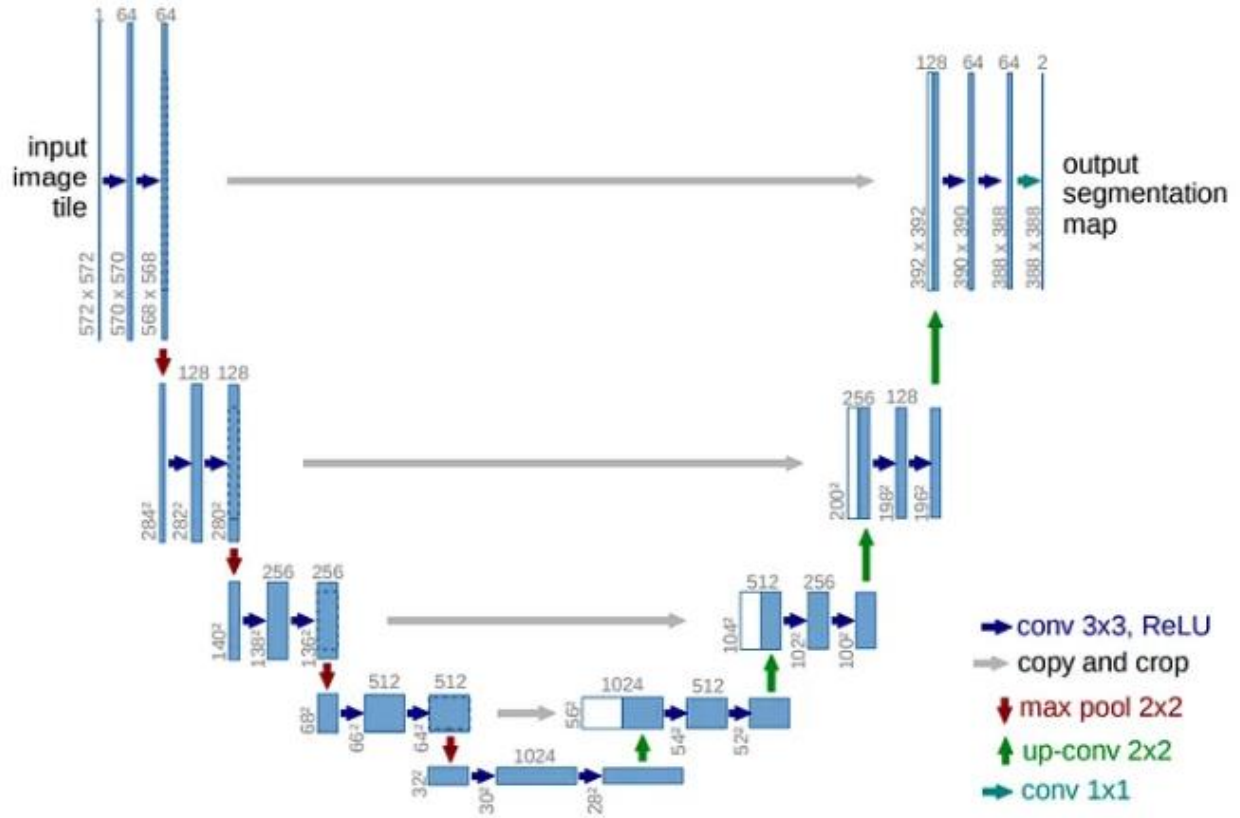
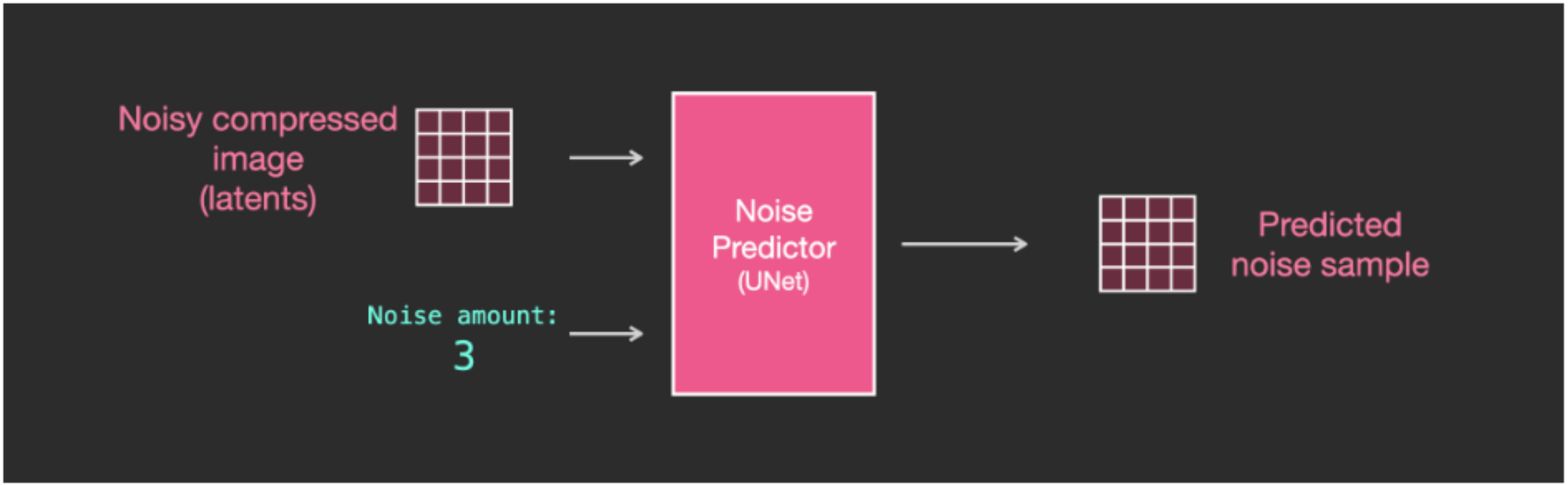


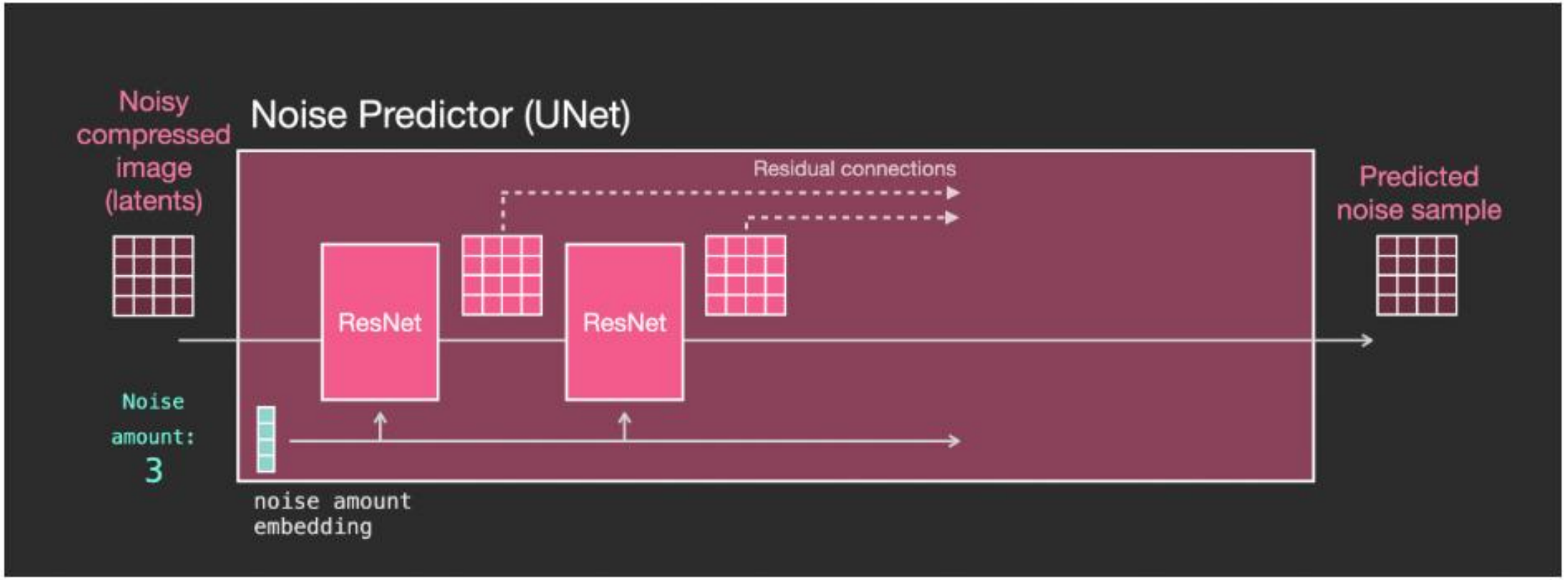
Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

Illustration of a Unet architecture from the original paper

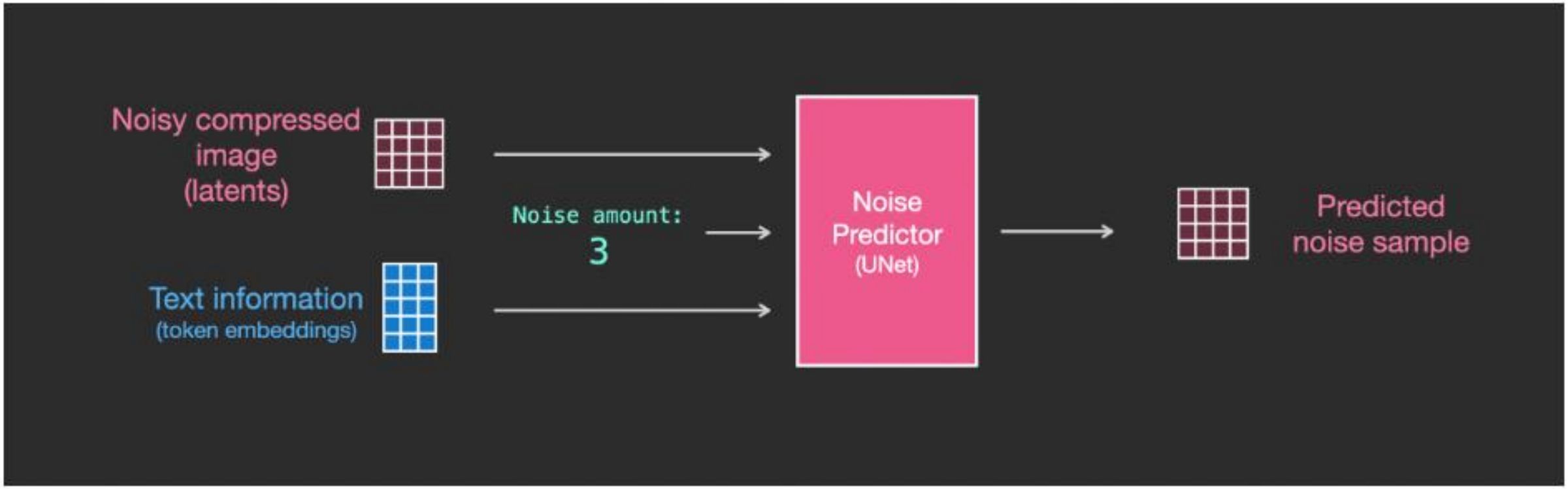
UNet Noise Predictor (no text)



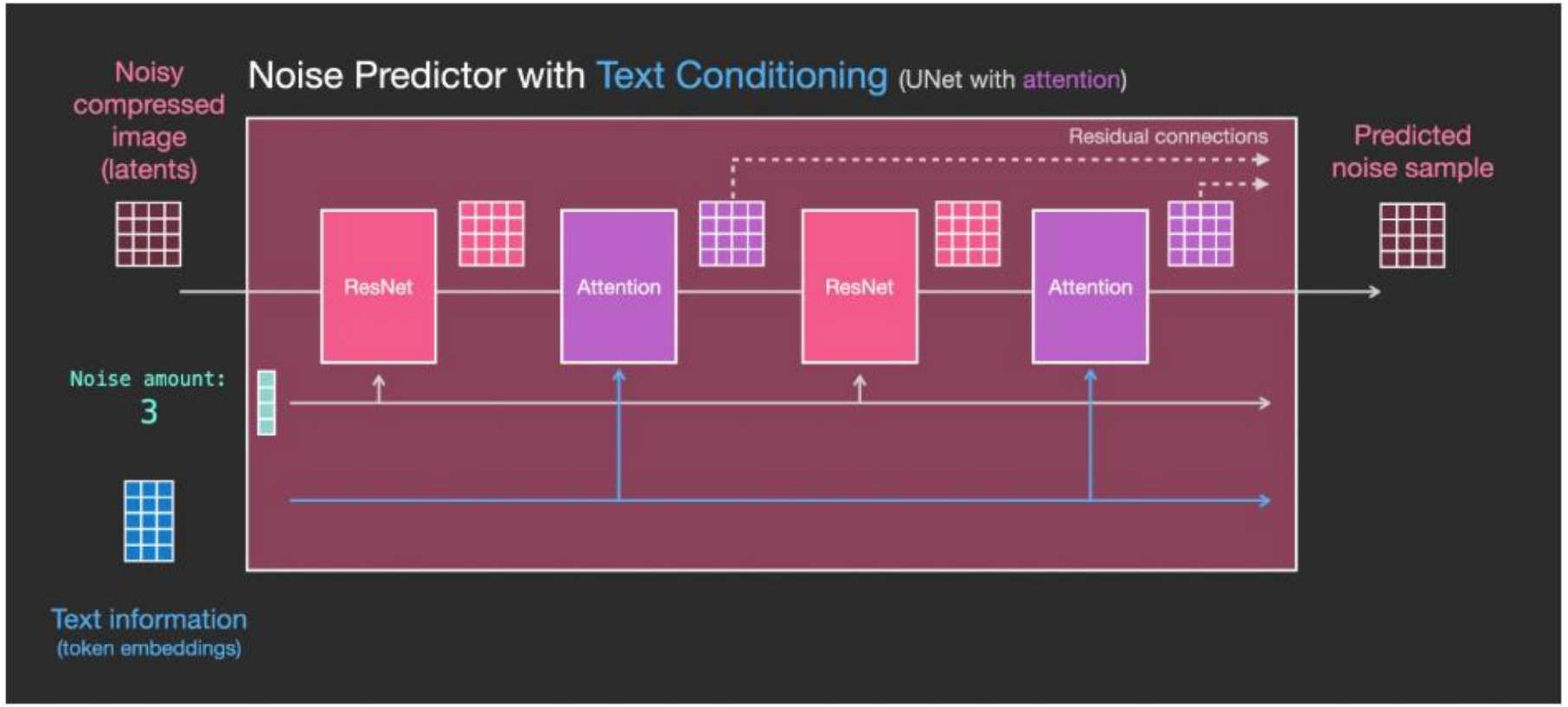
More Details



Adding Text to UNet Noise Predictor



Looking Inside



Current State of the Art

- Some of the major players:
  - Stability AI
  - DALL-E 3
  - Midjourney
  - Seedream AI

- Large Dueling Arena
  - Artificial Analysis Text to Image Leaderboard

Additional References

- The Illustrated Stable Diffusion – Jay Alammar
- How does Stable Diffusion work? – Andrew
- Stable Diffusion Public Release – stability.ai
- Latent text-to-image diffusion models – Hugging Face
- Stable Diffusion Example Jupyter Notebook – Google Colab
- Stable Diffusion 2.1 Demo – Hugging Face/StabilityAI
- Stable Diffusion with Diffusers – Patil, Cuenca, Lamber, and von Platen
- The Annotated Diffusion Model – Niels Rogge and Kashif Rasul
- Mikes-StableDiffusionNotes – Mike Brave