

Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

Advanced Deep Learning

Dr. Rastgoo







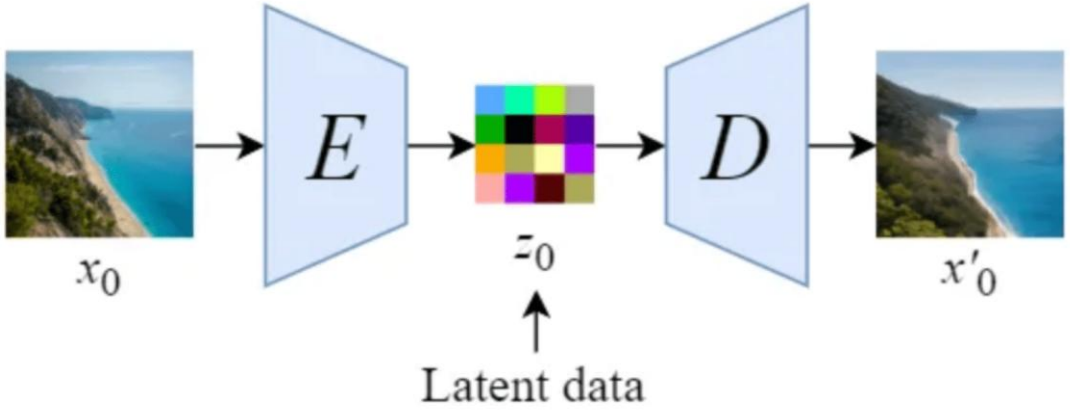
Stable Diffusion

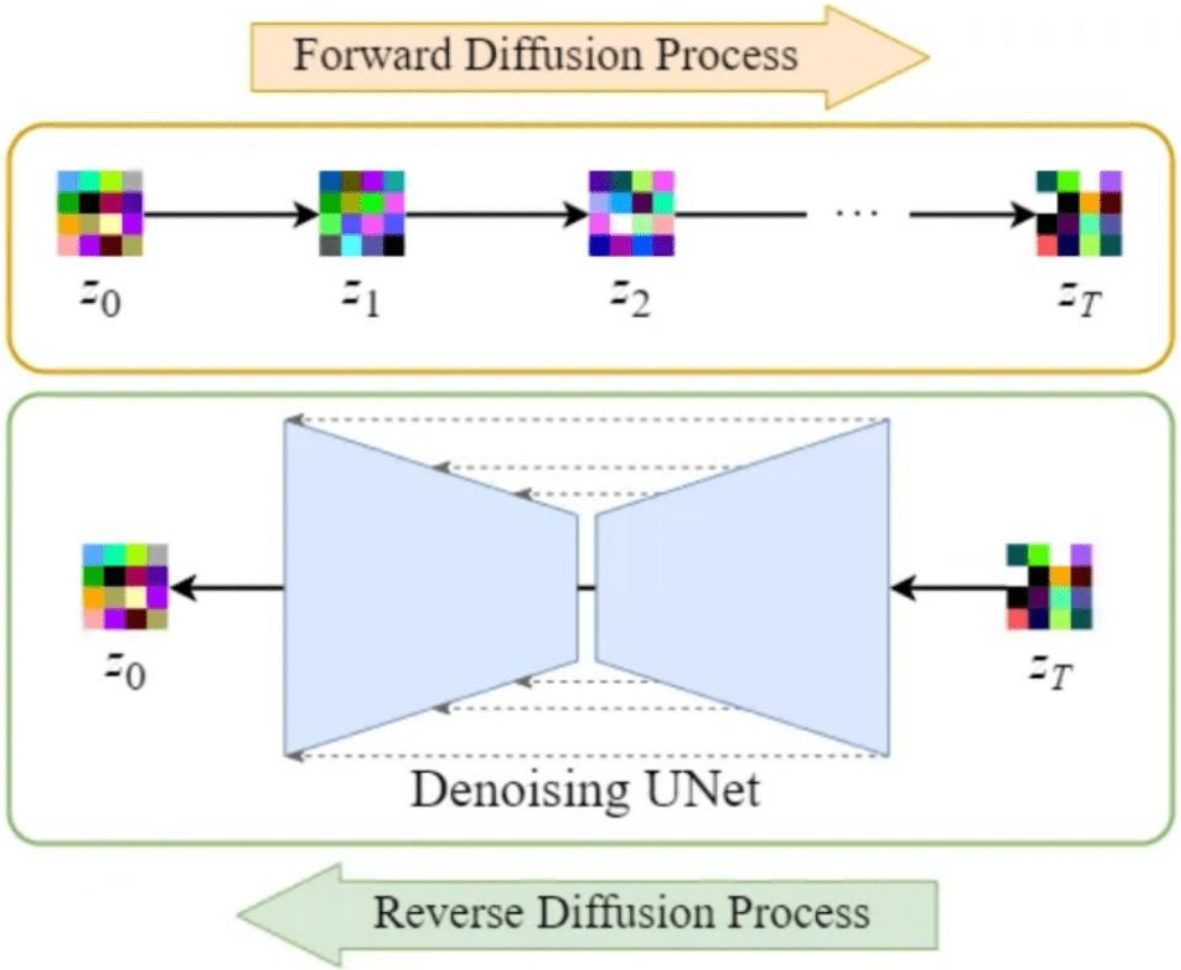
- The reverse diffusion process in conventional diffusion models involves iteratively passing a full-sized image through the U-Net architecture in order to obtain the final denoised result.
- However, this iterative nature presents challenges in terms of computational efficiency.
- This is emphasized when dealing with large image sizes and a high number of diffusion steps (T).
- The time required for denoising the image from Gaussian noise during sampling can become prohibitively long.
- To address this issue, a group of researchers proposed a novel approach called Stable Diffusion, originally known as Latent Diffusion Model (LDM) [15].

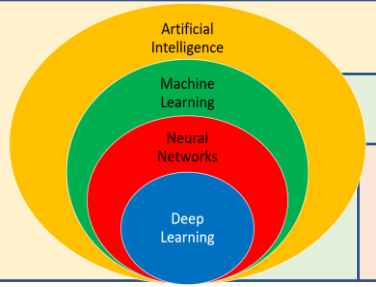
Latent Diffusion Models

- Stable Diffusion introduces a key modification by performing the diffusion process in the latent space.
- This works by using a trained Encoder E for encoding a full-size image to a lower dimension representation (latent space).
- Then making the forward diffusion process and the reverse diffusion process within the latent space.
- Later on, with a trained Decoder D, we can decode the image from its latent representation back to the pixel-space.
- For constructing the encoder and decoder, we can train some variant of a Variational AutoEncoder (VAE). This network is then decoupled for using both components separately.

Latent Diffusion Models







Artificial Intelligence (AI)

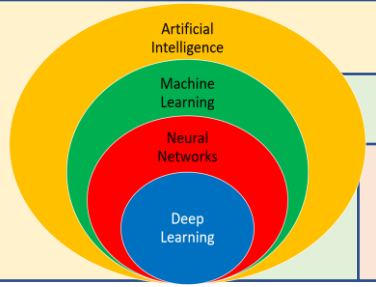
Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

Latent Diffusion Models

- Shifting diffusion operations to the latent space in Stable Diffusion enhances speed and reduces costs.
- This advancement accelerates denoising and sampling processes, making it an efficient solution for high-quality image generation and stable training.
- By leveraging the latent space, Stable Diffusion eases the computational burden in the reverse diffusion process.
- This enables quicker denoising of images, enhancing both speed and overall model stability and robustness.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

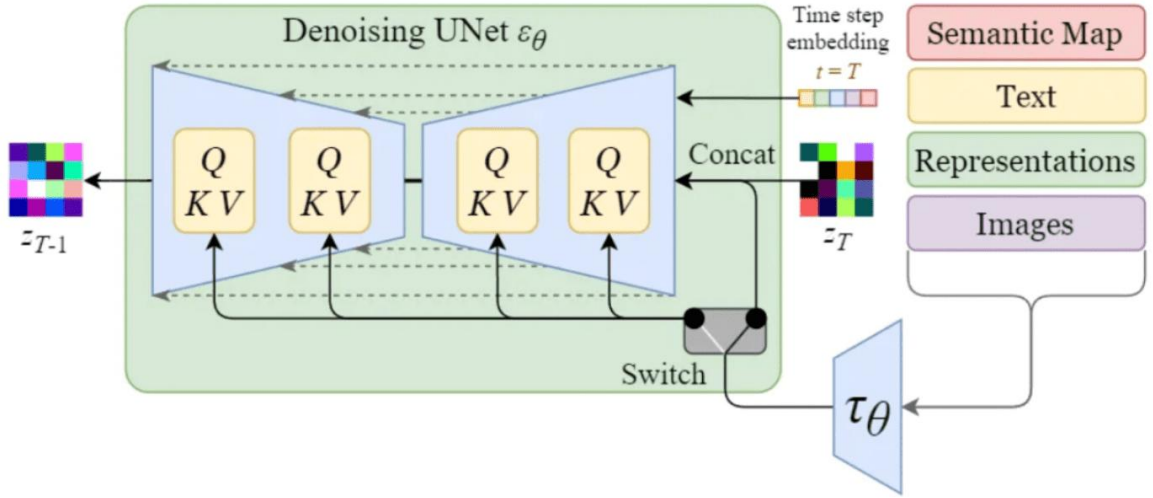
Deep Learning (DL)

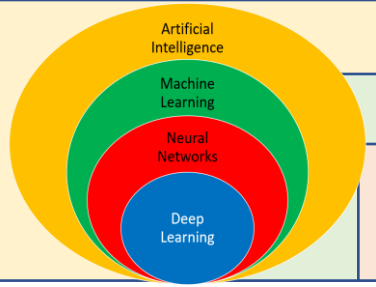
Conditioning

- Until then, generating images of a specific class was possible mainly through the addition of the class label in the input. Commonly known as Classifier Guidance.
- However, one of the standout features of the Stable Diffusion model, is its ability to generate images based on specific text prompts or other conditioning inputs.
- This is achieved by introducing conditioning mechanisms into the inner diffusion model.
- To enable conditioning, the denoising U-Net of the inner diffusion model makes use of a cross-attention mechanism.
- This allows the model to effectively incorporate conditioning information during the image generation (denoising) process.

Conditioning

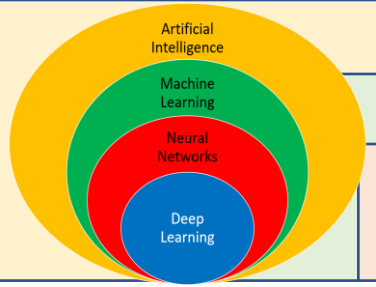
- The conditioning inputs can take various forms depending on the desired output:
- Text inputs are first transformed into embeddings through language models like BERT or CLIP. In the conditioning, we map these embeddings into the U-Net using a Multi-Head Attention layer, represented as Q, K, and V in the diagram.
- Other conditioning inputs such as spatially aligned data such as semantic maps, images, or inpainting act similarly.
- However, the integration of these conditioning mechanisms is usually achieved through concatenation.





Conditioning

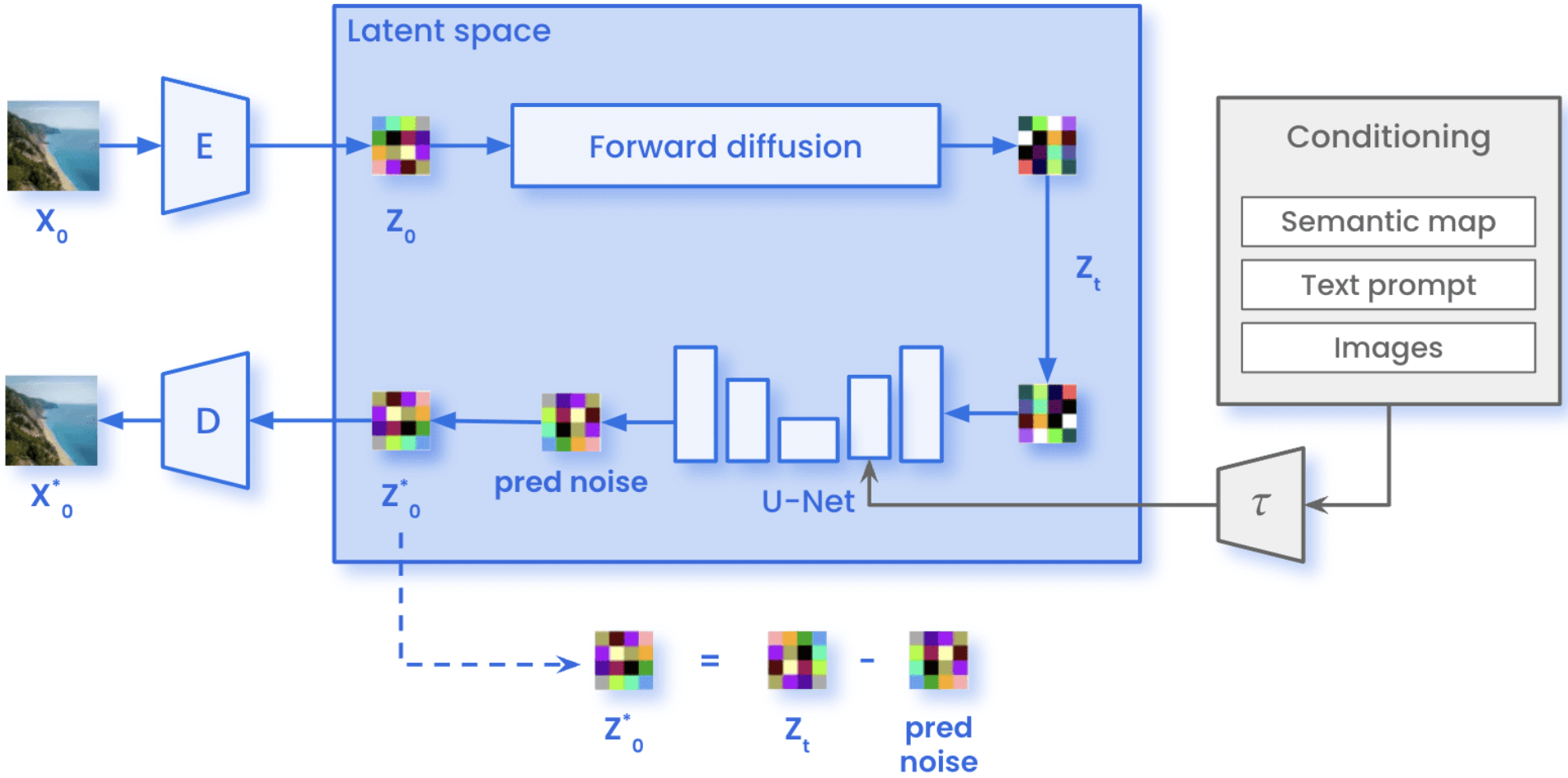
- By incorporating conditioning mechanisms, the Stable Diffusion model expands its capabilities to generate images based on specific additional inputs.
- Text prompts, semantic maps, or additional images, enable more versatile and controlled image synthesis.
- By using prompt engineering, it's possible to create even more compelling images.

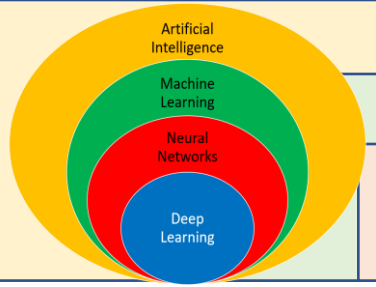


Architecture

- During training, the images (x_0) are encoded through the *Encoder E*, reaching the latent representation of the image (z_0).
- In the forward diffusion process, the image undergoes the addition of Gaussian noise, obtaining a noisy image (z_T).
- The image then passed through the U-Net, in order to predict the noise present in z_T.
- This comparison between the actual noise added in the forward diffusion and the prediction allows the calculation of the loss previously mentioned.
- With the calculated loss, we update the parameters of the U-Net through backpropagation.

Architecture

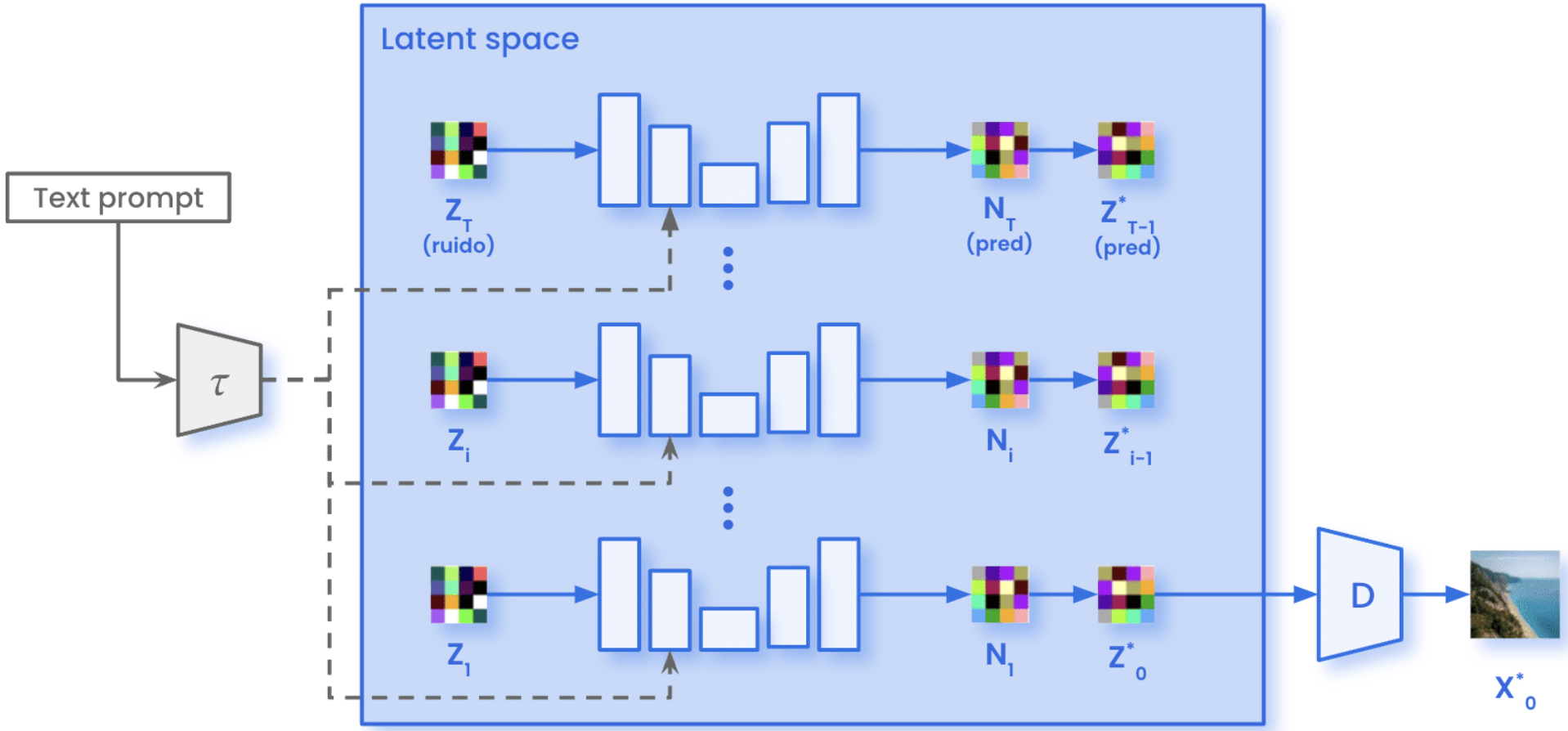


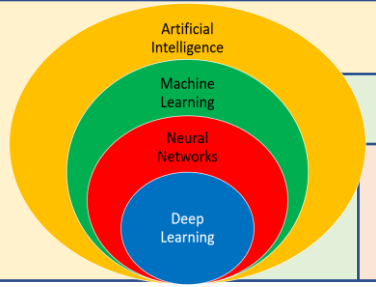


Architecture

- On the other hand, the forward diffusion process does not occur during sampling.
- We just sample Gaussian noise with the same dimensions present in the latent space (z_T).
- This noise passes through the U-Net for the specified number of inference steps T.
- At each step *t*, the U-Net predicts the whole noise present in the image.
- The model removes just a fraction of the predicted noise to obtain the representation of the image at timestep *t-1*.
- After all the T inference steps are iteratively, we obtain the representation within the latent space of the generated image ($\hat{z}_0$). Using the Decoder D, we can then transform that image from the latent space to the pixel-space (X_0).

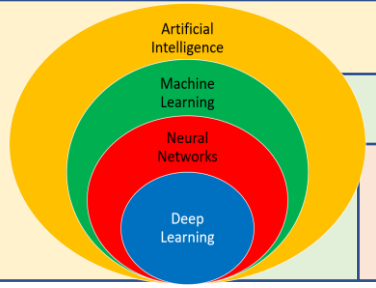
Architecture





Architecture

- On the other hand, the forward diffusion process does not occur during sampling.
- We just sample Gaussian noise with the same dimensions present in the latent space (z_T).
- This noise passes through the U-Net for the specified number of inference steps T.
- At each step *t*, the U-Net predicts the whole noise present in the image.
- The model removes just a fraction of the predicted noise to obtain the representation of the image at timestep *t-1*.
- After all the T inference steps are iteratively, we obtain the representation within the latent space of the generated image ($\hat{z}_0$). Using the Decoder D, we can then transform that image from the latent space to the pixel-space (X_0).



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

