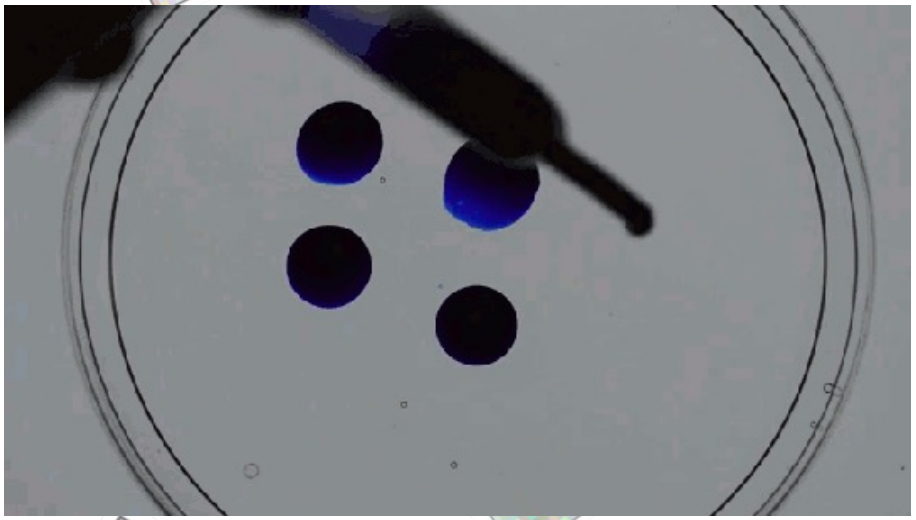
Stable Diffusion

Edmund Dable-Heath

Diffusion in the physical world

- Time-dependent random process of something moving from an area of high concentration to low concentration.
- How do we model this?
- If we know exactly how we got to a diffuse state, can we get back to a concentrated one?



$$P(X_{t+1} = x \mid X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t) = P(X_{t+1} = x \mid X_t = x_t)$$

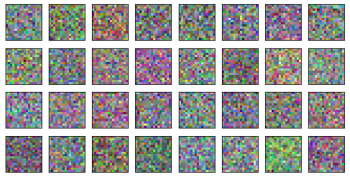
Diffusing Images

- Consider an image as a concentrated collection of coherent information within pixel space.

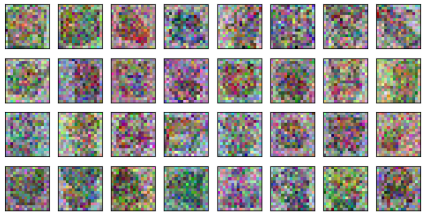
- Then we can think of adding noise – or blurring the image – as diffusion.

- Taking a trip through the image library of babel.

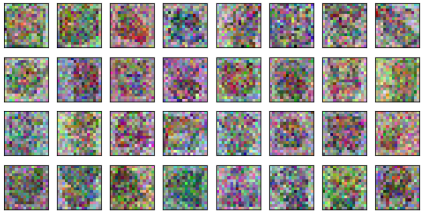
So what do these models look like?



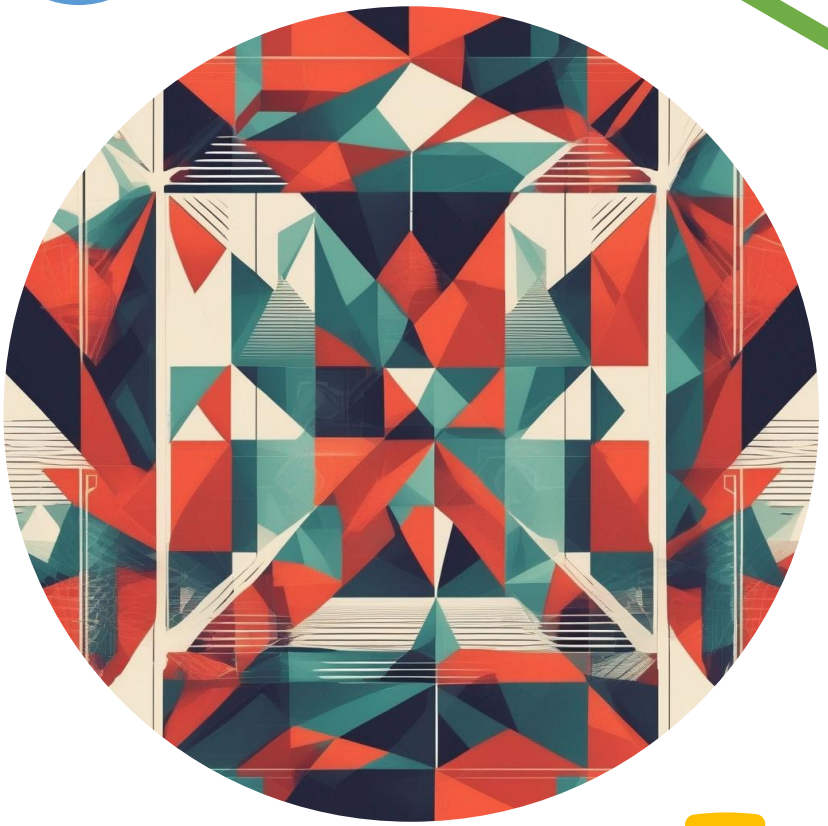
x_T → Neural Net → Remove Noise → x_{T-1}



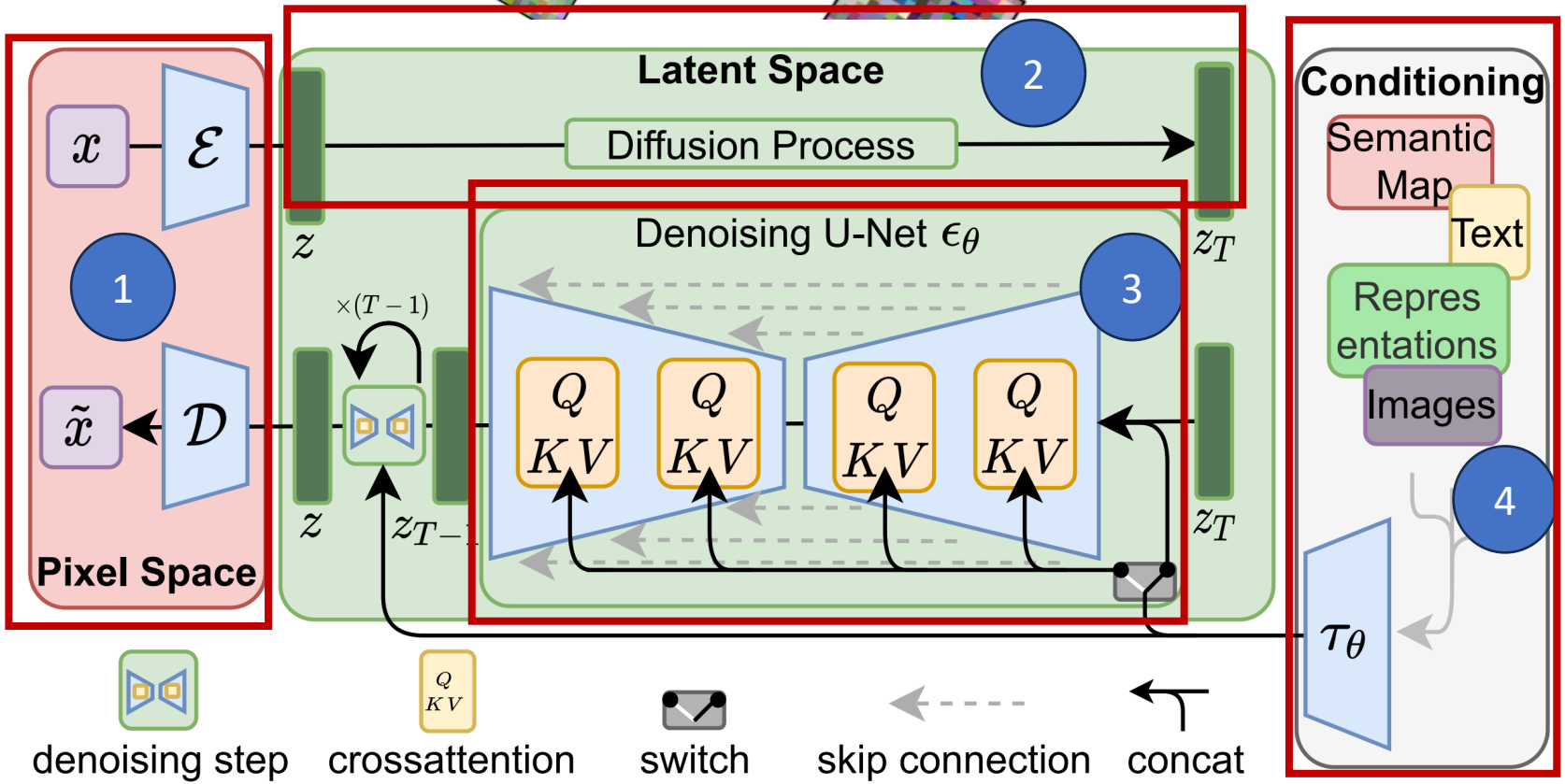
x_{T-1} → Neural Net → Remove Noise → x_{T-2}



Repeat → x_0



Architecture Overview


Latent Space
2
Diffusion Process
Conditioning
Semantic Map
Text
Repres entations
Images
4
Denoising U-Net ϵ_θ
3
$\times(T-1)$
x
$\mathcal{E}$
1
$\tilde{x}$
$\mathcal{D}$
Pixel Space
z
z_T
z_{T-1}
Q K V
τ_θ
denoising step
crossattention
switch
skip connection
concat

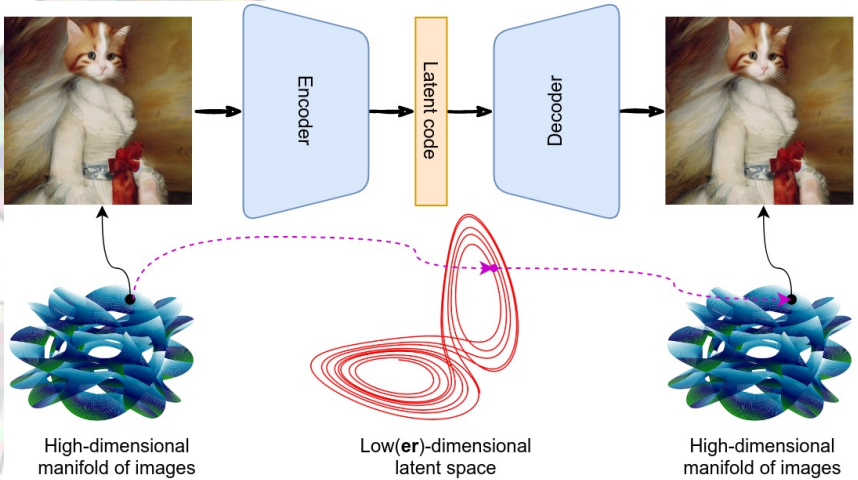
Pixel Space to Latent Space

Latent Space

- Also known as embedding space.

- A learned mathematical space of representations of the data.

- Typically unintuitive and hard to interpret, but effective if employed well.



Why Latent Space?

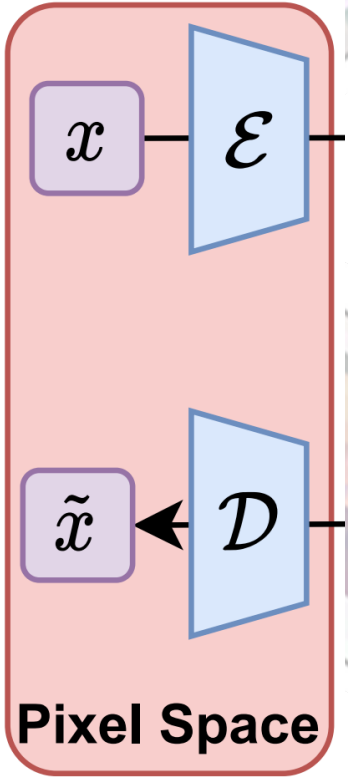
- Pixel-space is expensive:
    - Large memory constraints.
    - Costly function evaluations.

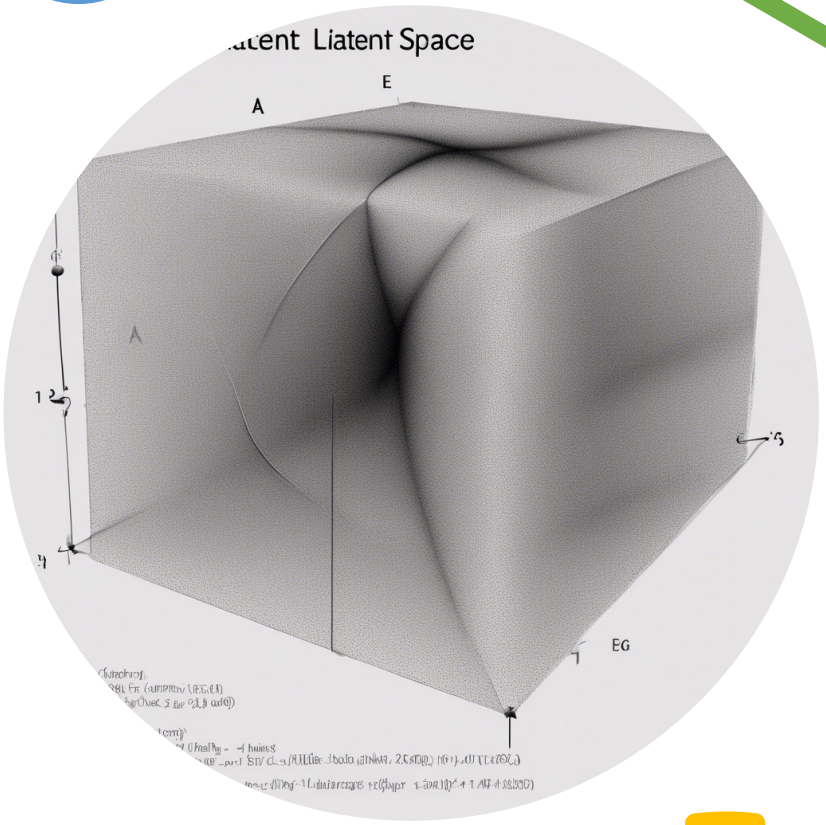
- Less need for quality-reducing compression of previous approaches.

- Once a compression model for a latent space has been learned, it and the latent space can be reused to train different generative models.

Perceptual Image Compression

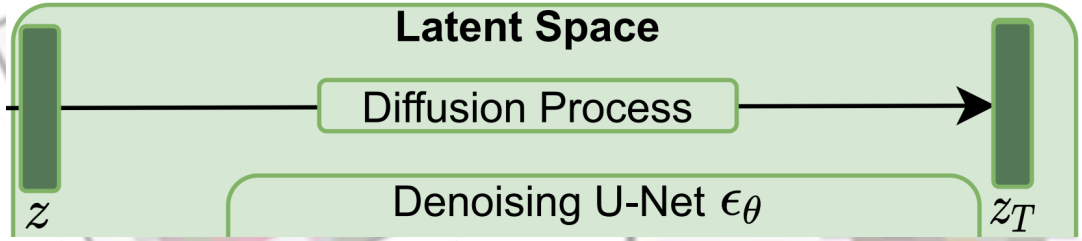
- Want a latent space that is perceptually equivalent to larger space.
- Autoencoder down samples, with the decoder reconstructing.
- Trained via perceptual loss and patch-based adversarial objective.
- Uses regularization (KL or VQ) to arbitrarily high-variance latent spaces.





Diffusion in a Latent Space

Diffusion in a Latent Space



- Take a randomly sampled ‘blur’ in the pixel space and encode it into the latent space.

- This then forms the starting point for our denoising process for training, having ‘blurred’ the images to a requisite state.

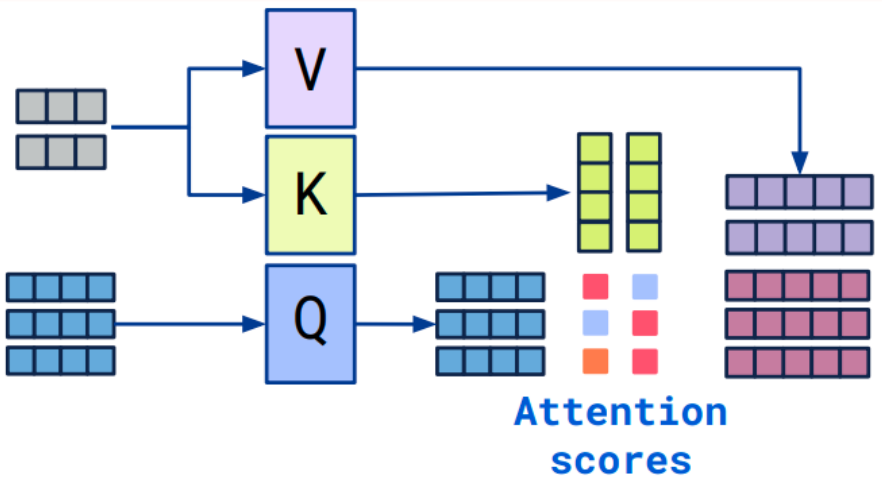
- At inference time we start with z_T if just generating images, and run the diffusion process if starting with a particular image as an input.

Denoising U-Net with Embedded Cross-Attention

Attention Recap

- Mechanism mimicking cognitive attention.

- Learns weights for relationship between each input token and each other token.

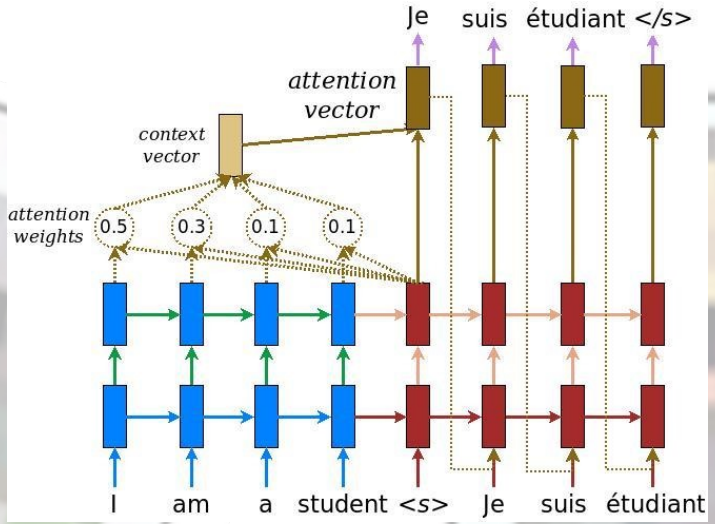


Why Attention?

- Effective at learning models with various input modalities.

- Interfaces well with pre-trained conditioning systems.

- Cross-attention allows for the comparison of two inputs.

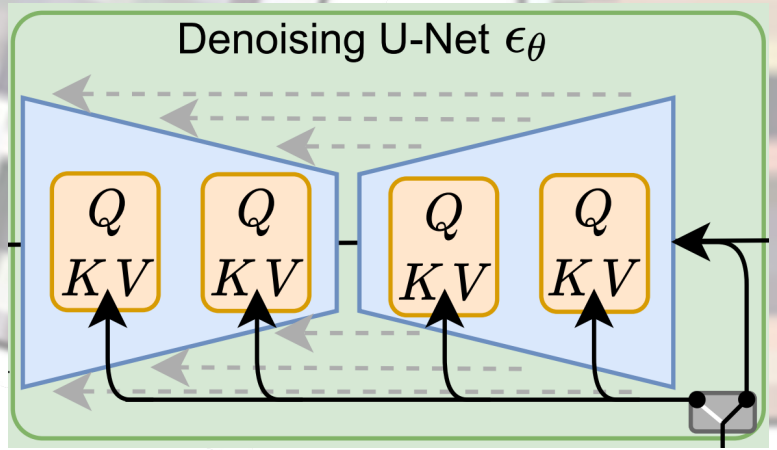


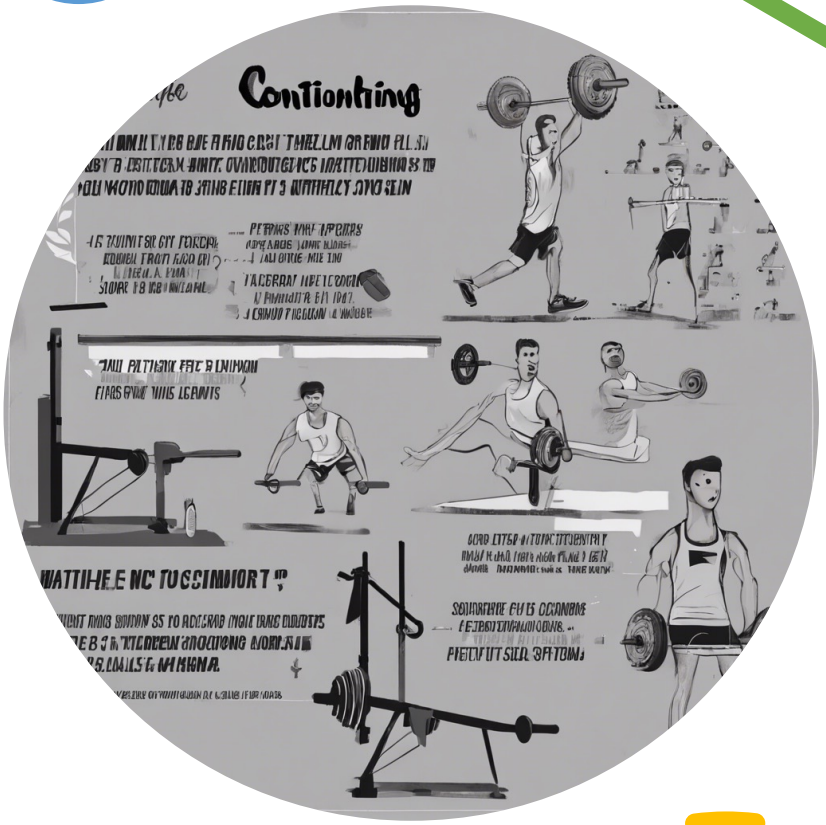
Embedded Attention

- At each layer of the U-Net a cross attention mechanism relates the context from an embedding to the latent representation.

- Each projection matrix is learnable.

- At training time the related embedding for a caption is given alongside the image that is being denoised.





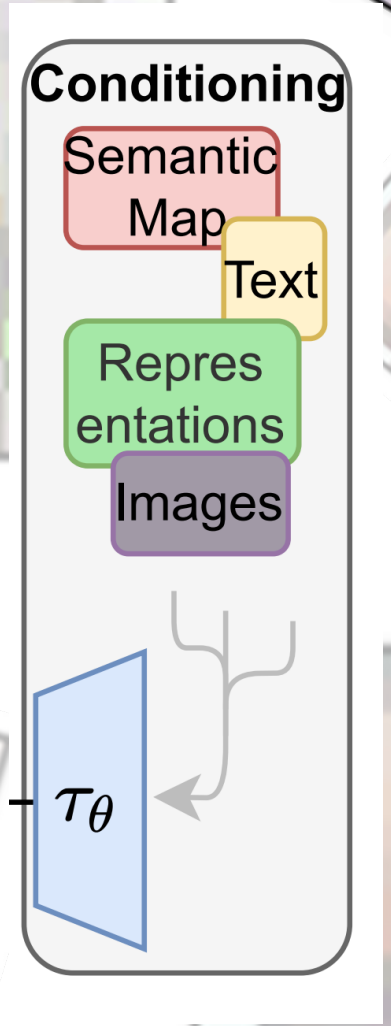
Conditioning

Conditioning Models

- In the paper BERT-tokeniser used to provide latent space representations of semantic input.

- This is then fed to the U-Net as guidance via cross-attention.

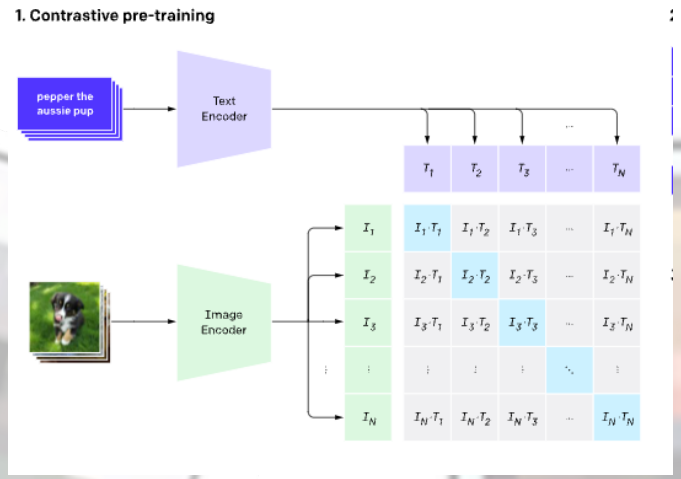
- Available version online switched to CLIP.

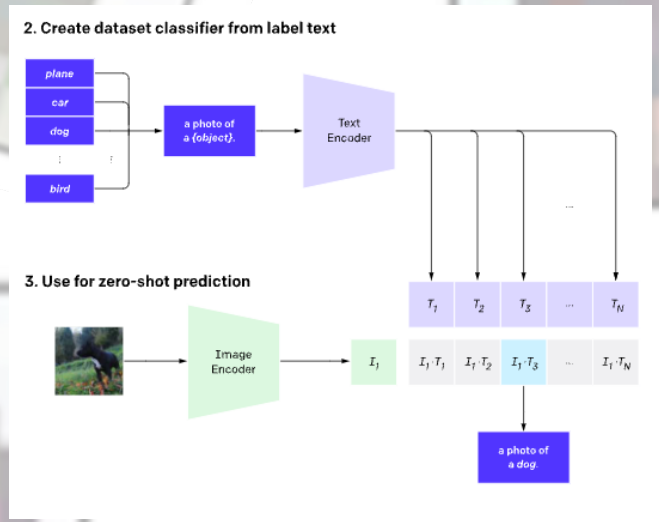


CLIP

- CLIP trained on captioned images, providing related embeddings for images and their captions.

- Negative examples of mismatched captions and images also used in training.

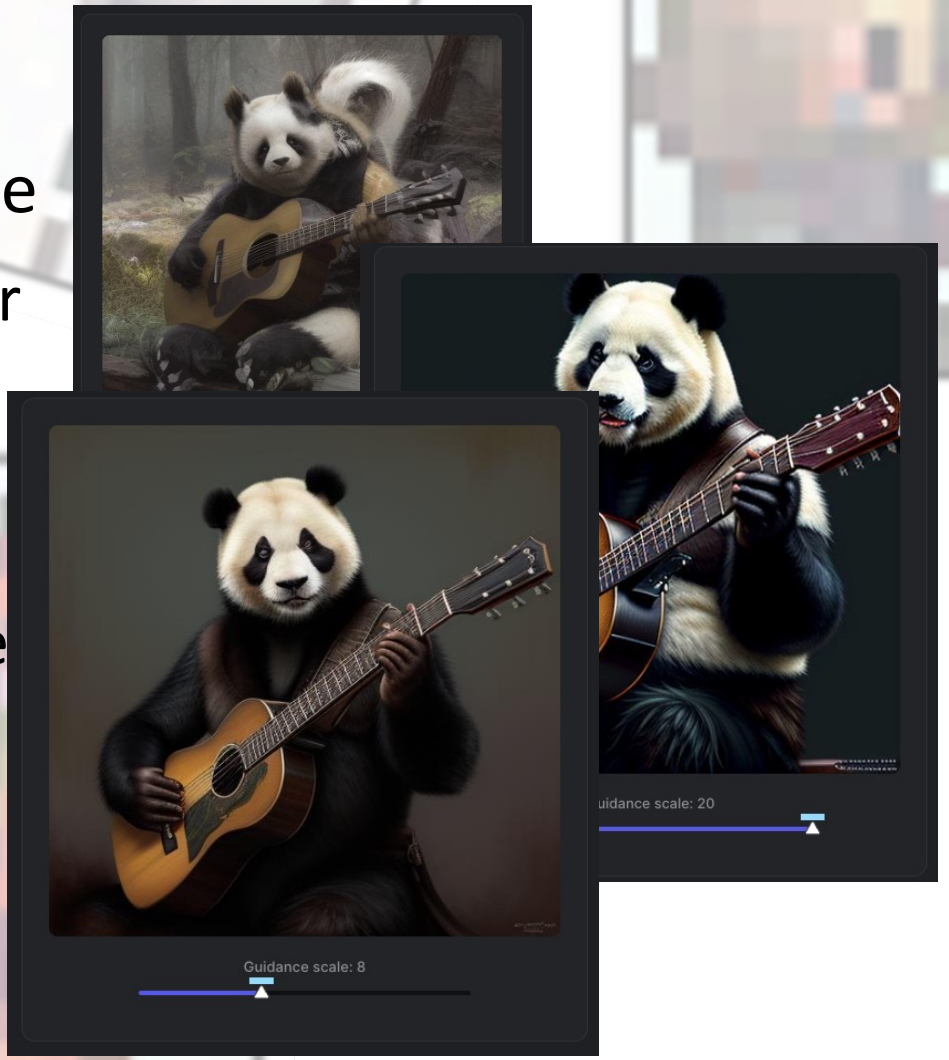


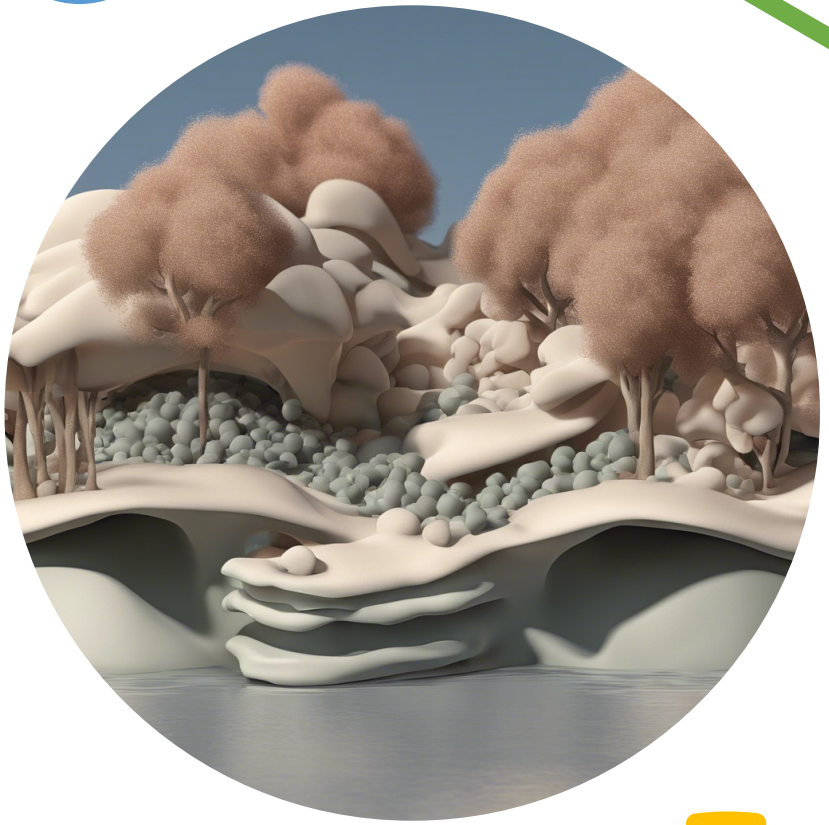


Classifier Free Guidance Scale

- A single value hyperparameter that controls how much ‘creativity’ the model has.

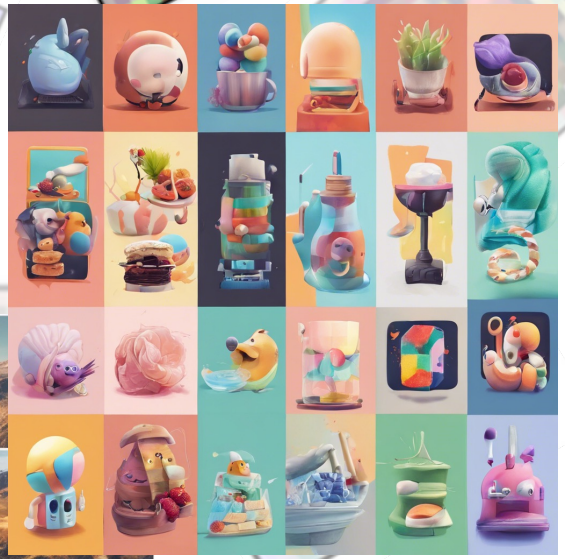
- Generally advised that a value of ‘around 7-13’ will give the best results, but this varies depending on task.

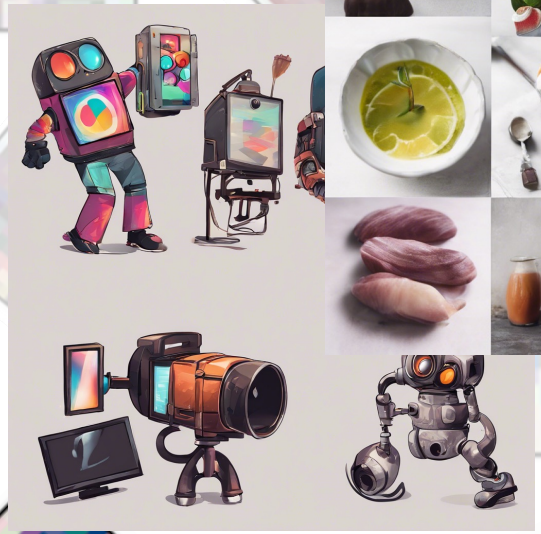


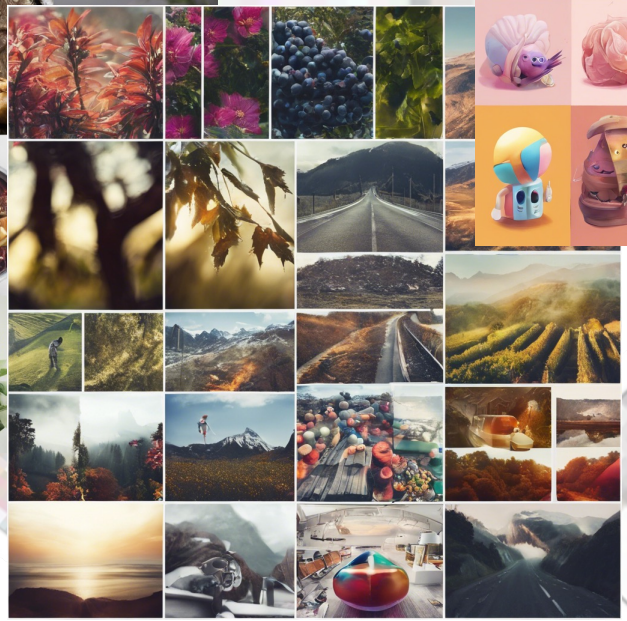


Applications of Stable Diffusion

Image Generation







Super Resolution

- The latent space method allows for up sampling from an initial resolution efficiently.

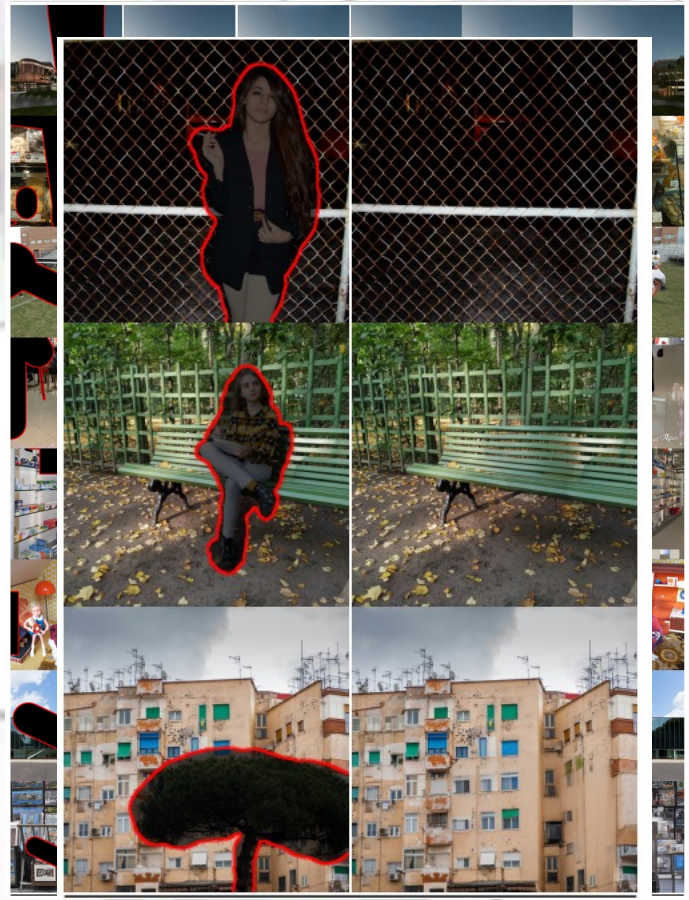
- Trained on down-sampled images as input, with high resolution images as final output.

In Painting

- The task is to remove an object from the frame.

- Training is done by recovering randomly removed segments from images.

- Inference input is then an image with the object to be removed cropped out.



Thanks for Listening!