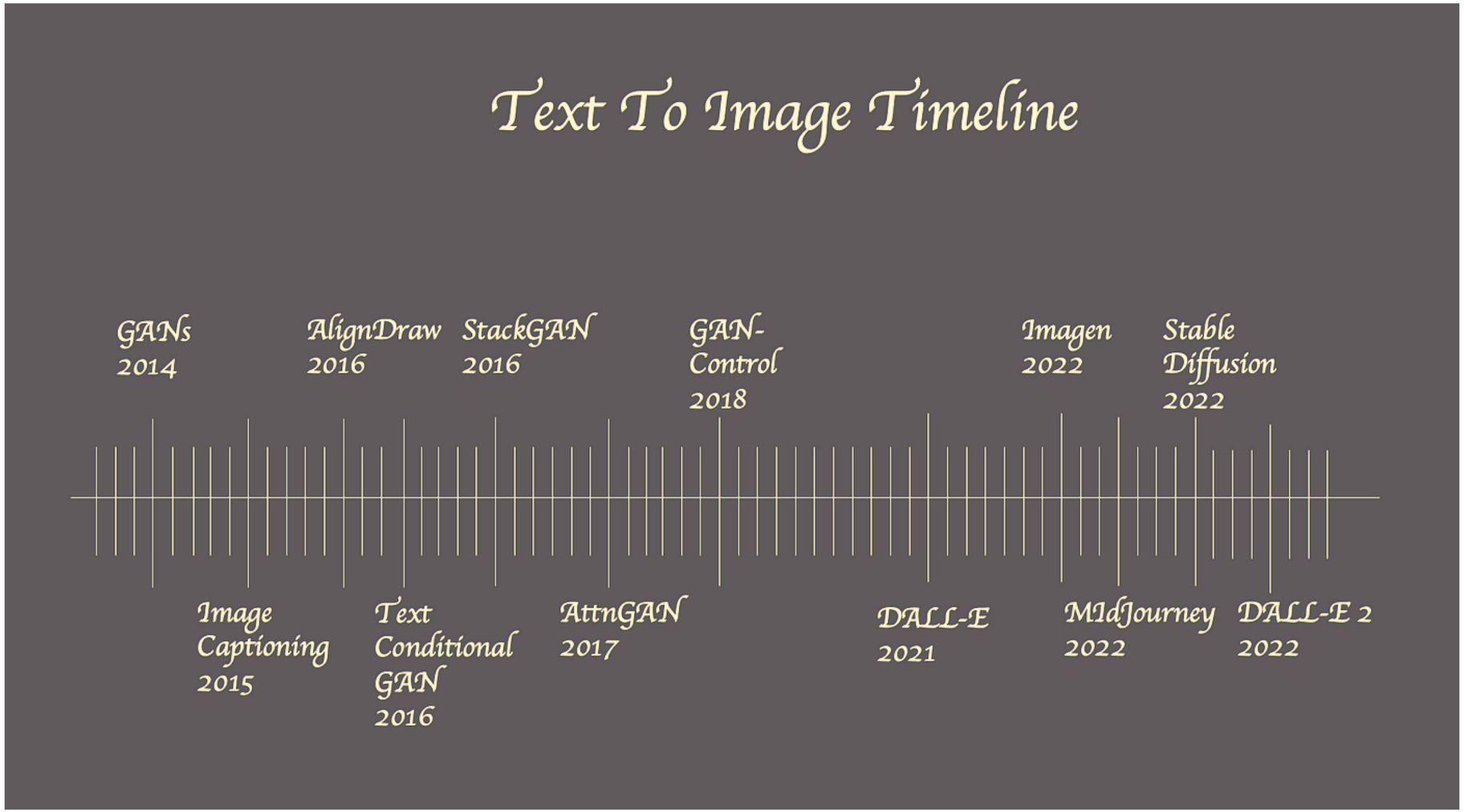
Generative AI

Timeline


Text To Image Timeline
GANs
2014
Image
Captioning
2015
AlignDraw
2016
Text
Conditional
GAN
2016
StackGAN
2016
AttnGAN
2017
GAN-
Control
2018
DALL-E
2021
Imagen
2022
MIdJourney
2022
Stable
Diffusion
2022
DALL-E 2
2022

Generative Adversarial Networks (GANs)

We will study particularly about a few GAN architectures :

- Vanilla GAN
- Cycle GAN
- Style GAN
- Text-2-Image GAN

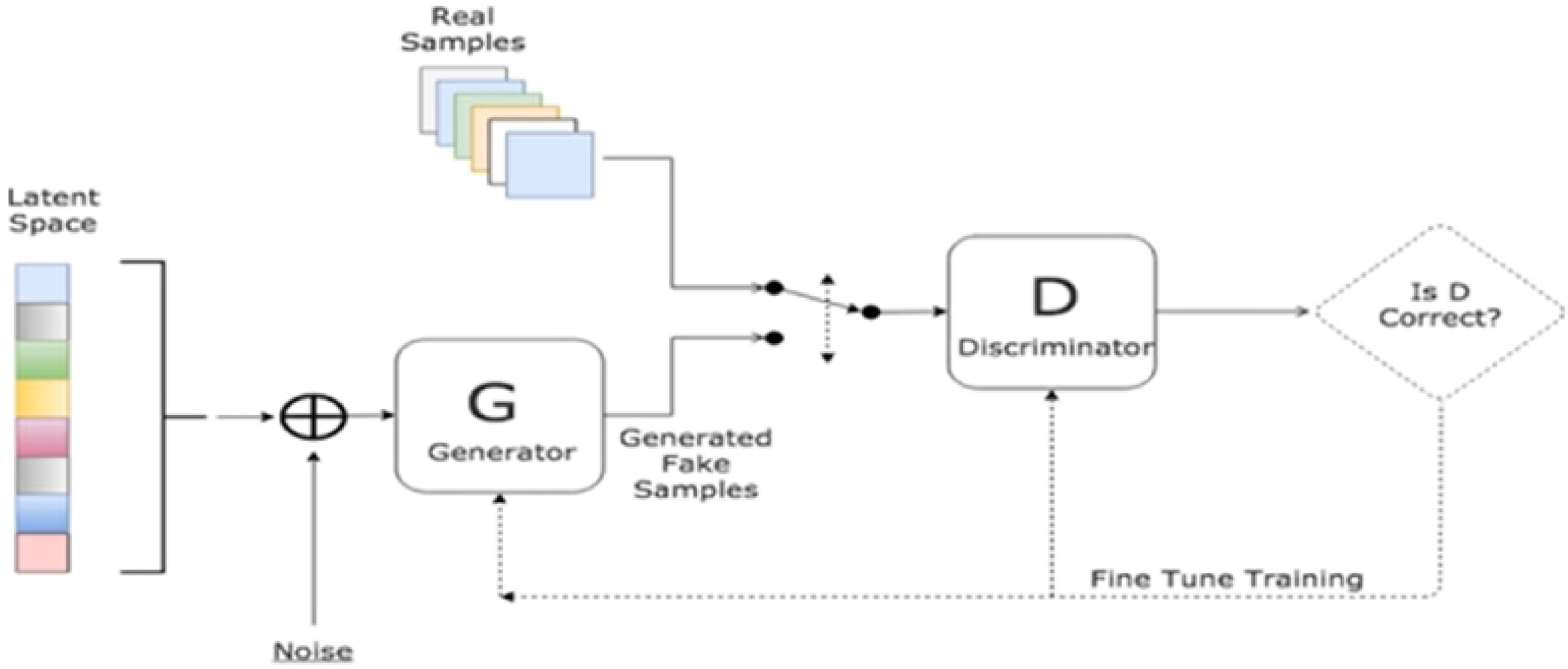
Vanilla GAN

Vanilla GAN is a term often used to refer to the original and basic form of a Generative Adversarial Network (GAN).There are 2 kinds of models in the context of Supervised Learning, Generative and Discriminative Models.

Generator : The generator's primary task is to create synthetic data that resembles real data. It takes random noise as input and generates data samples. The generator is typically implemented as a neural network that maps random noise vectors to data samples.

Discriminator : The discriminator's role is to distinguish between real data and fake data generated by the generator. It also takes the form of a neural network, but instead of generating data, it takes data samples as input and outputs a probability score indicating whether the input data is real or fake.

Vanilla GAN Architecture

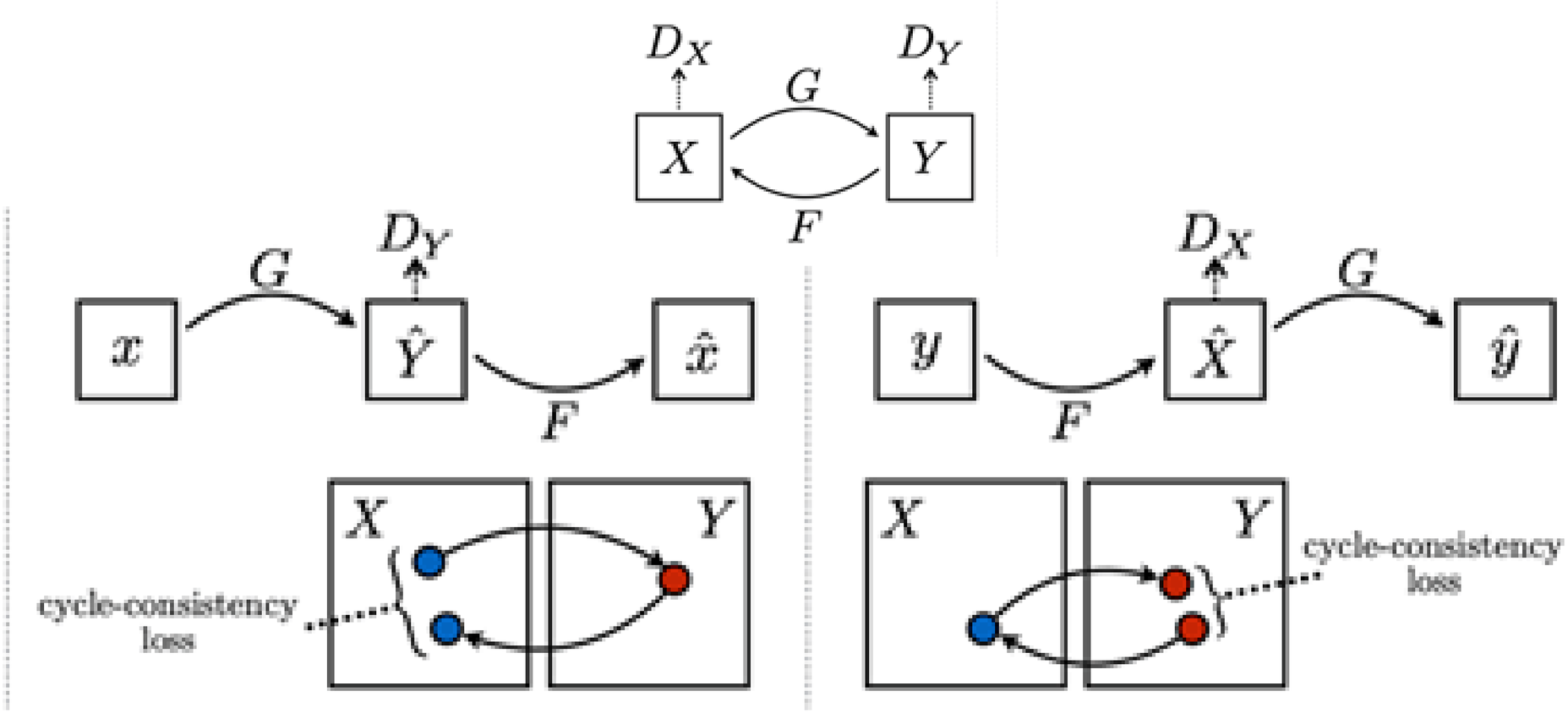


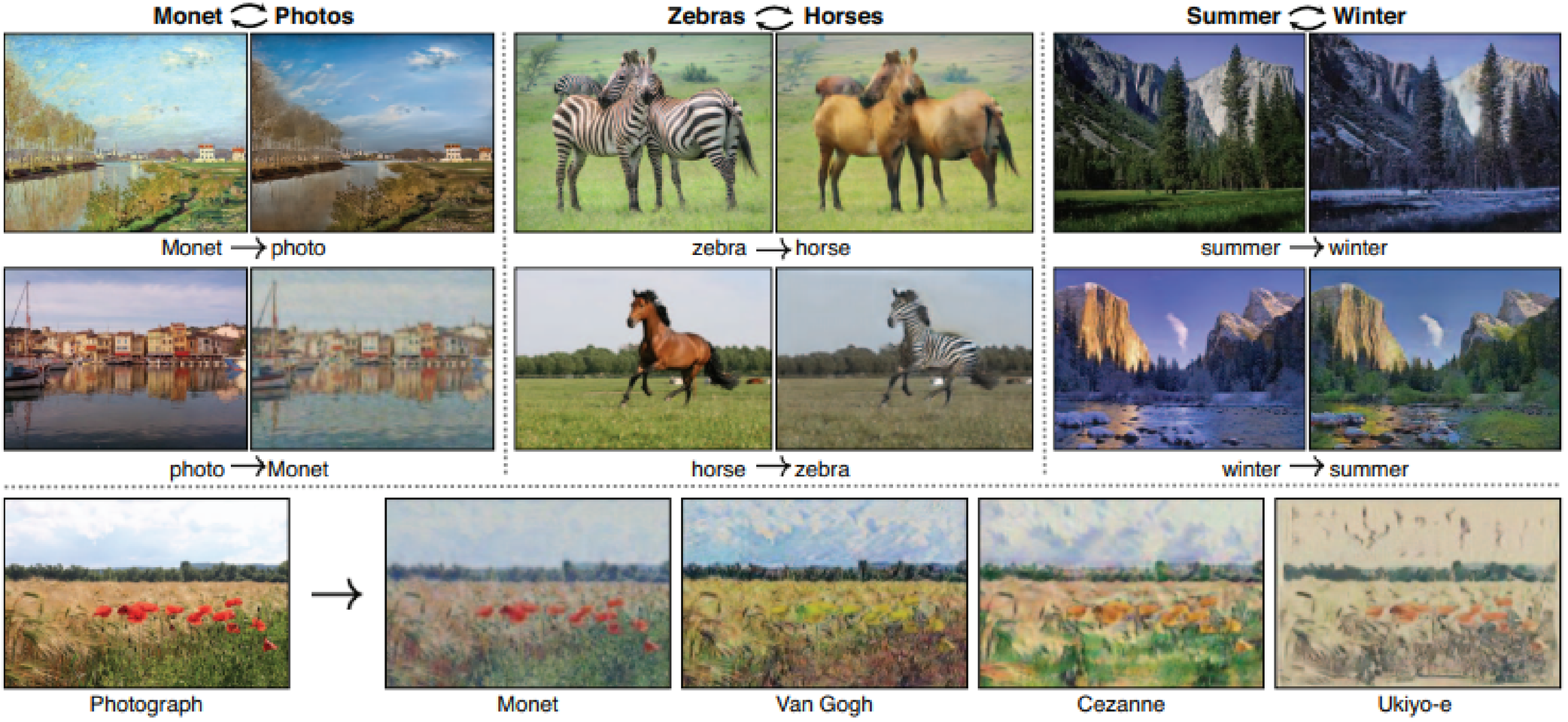
Loss function :

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

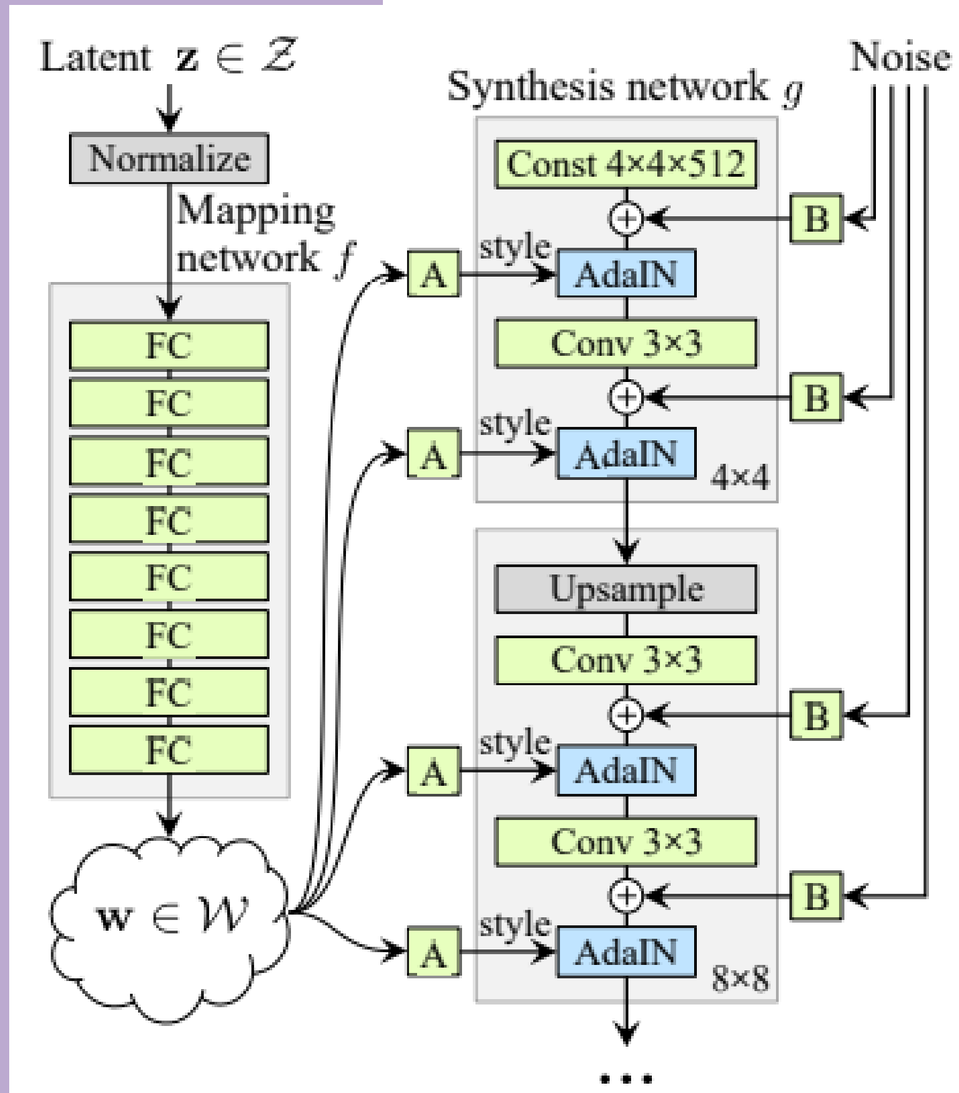
CycleGAN

CycleGAN is a type of generative adversarial network used for image-to-image translation tasks.The main idea behind CycleGAN is to learn a mapping between two different domain without requiring paired training data.



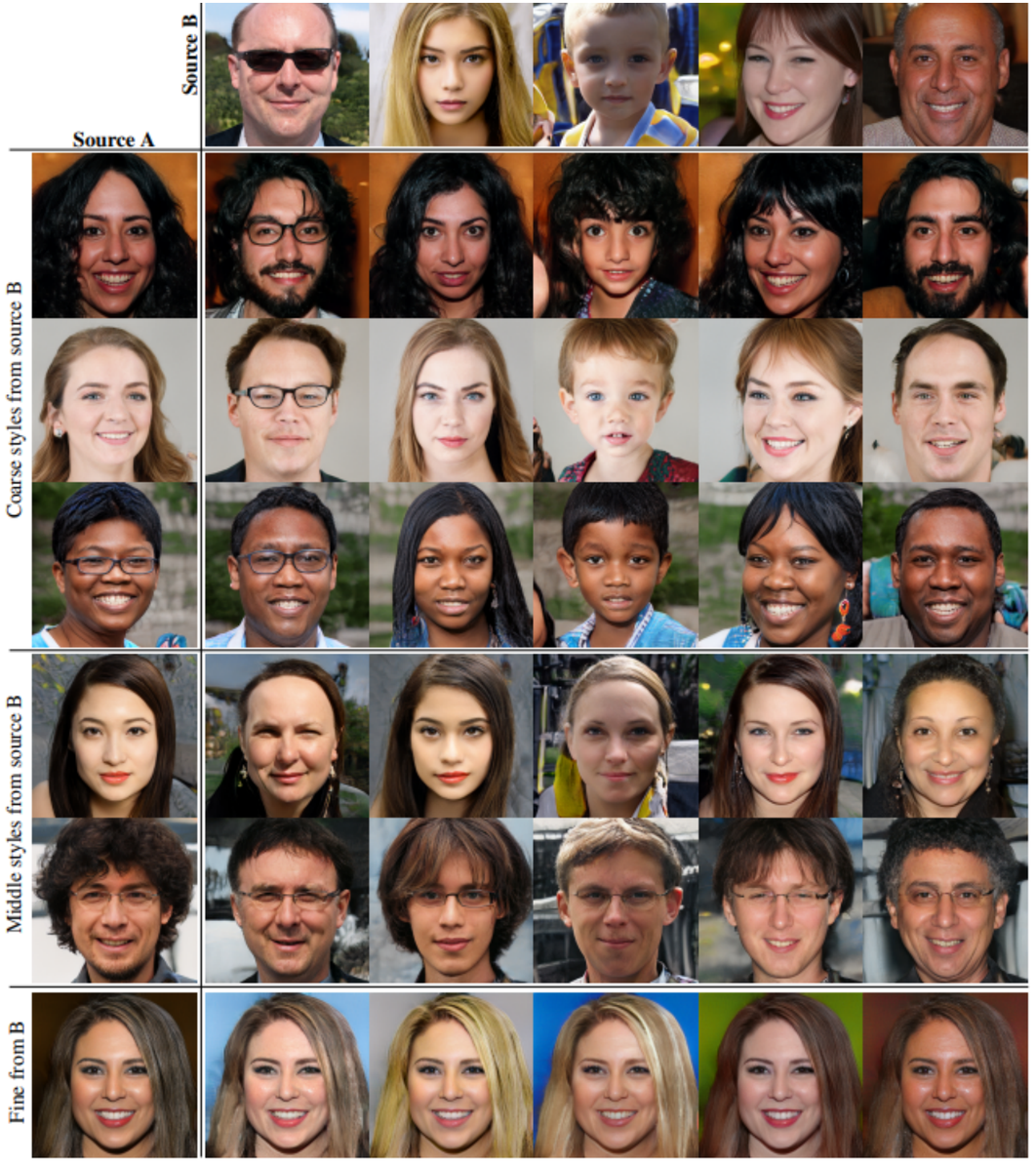

Monet ⇄ Photos
Zebras ⇄ Horses
Summer ⇄ Winter
Monet → photo
zebra → horse
summer → winter
photo → Monet
horse → zebra
winter → summer
Photograph
Monet
Van Gogh
Cezanne
Ukiyo-e

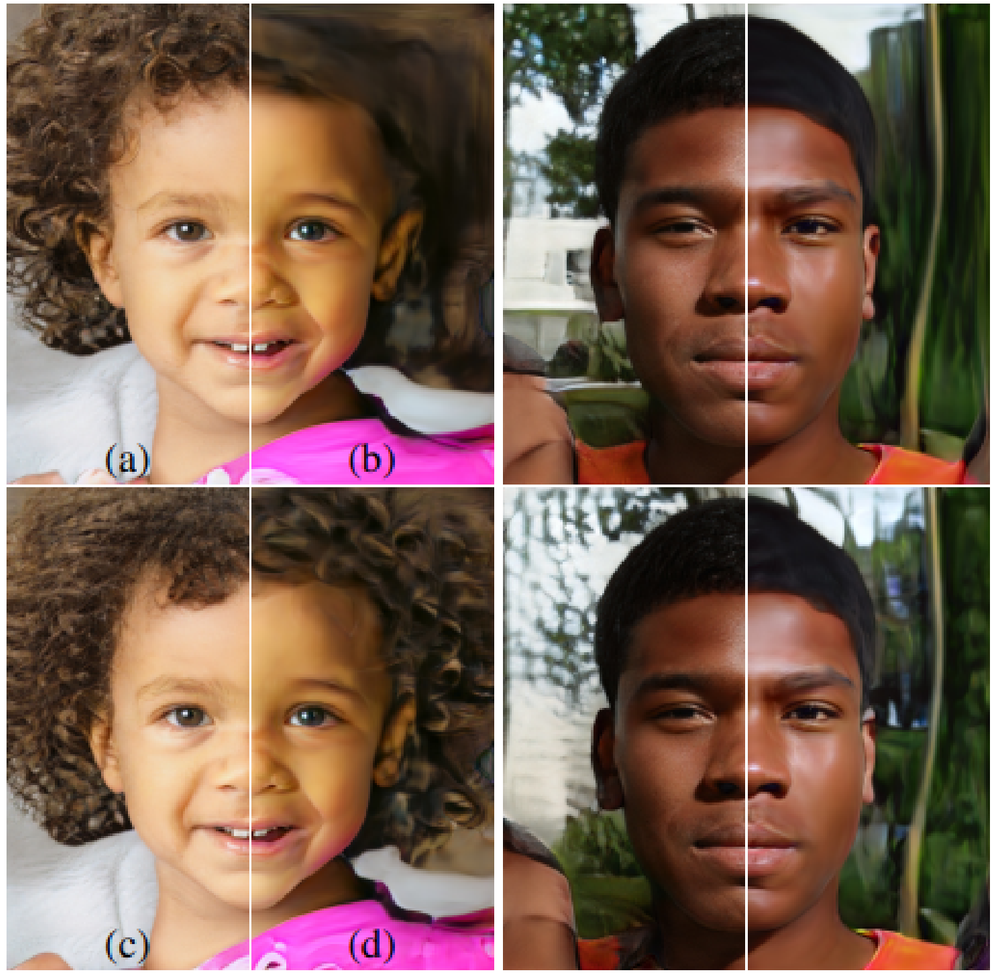
StyleGAN



- For generating high-quality images with controllable styles and attributes
- Initially latent vector was normalised using PixelNorm
- High quality dataset namely Flickr Faces- HQ (FFHQ)
- To further encourage the styles to localize, we employ mixing regularization, where a given percentage of images are generated using two random latent codes instead of one during training
- The AdaIN operation is defined as :

$$\mathrm{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i} \frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i},$$


Source B
Source A
Coarse styles from source B
Middle styles from source B
Fine from B


(a)
(b)
(c)
(d)

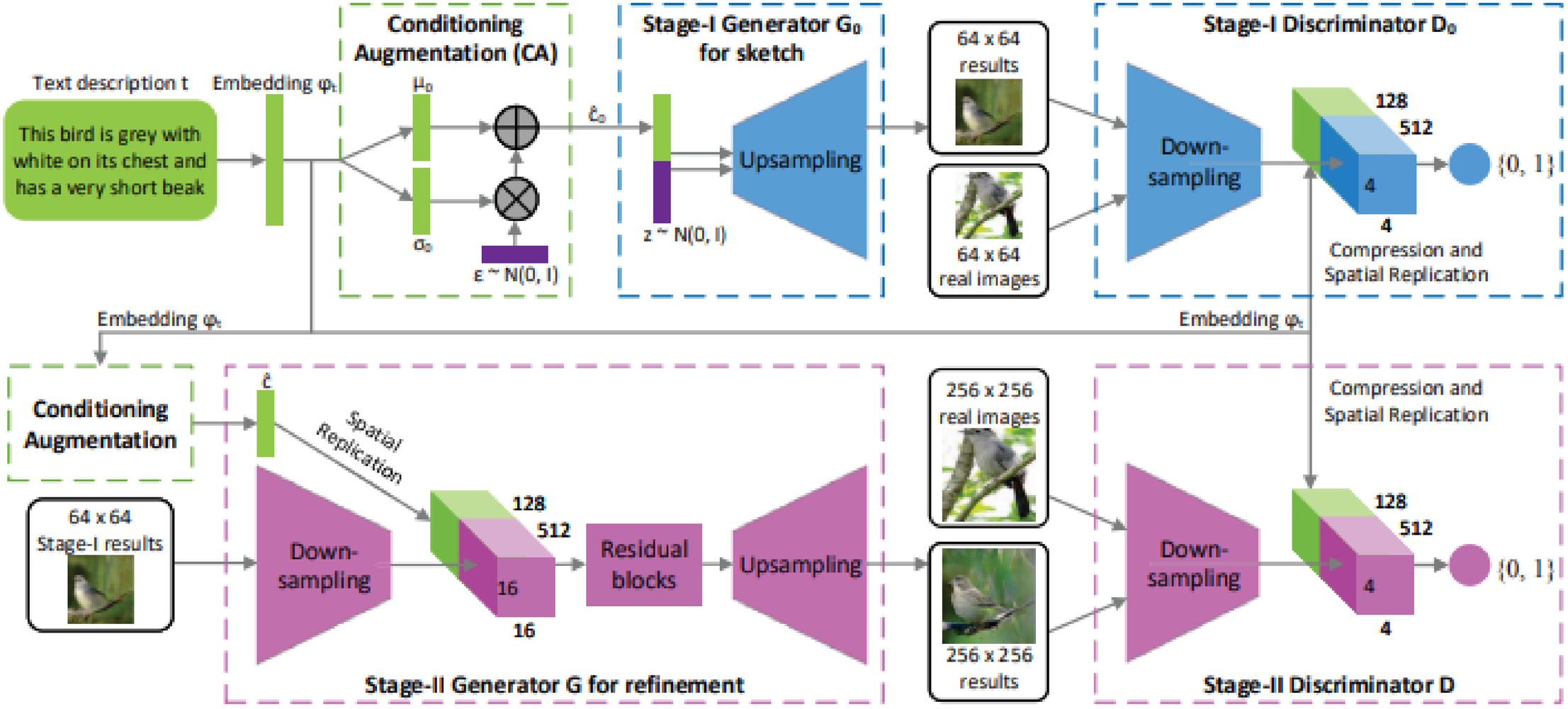
StackGAN

StackGAN, which stands for "Stacked Generative Adversarial Networks," is a generative model architecture designed for text-to-image synthesis. It is specifically developed to generate high-resolution images from textual descriptions.

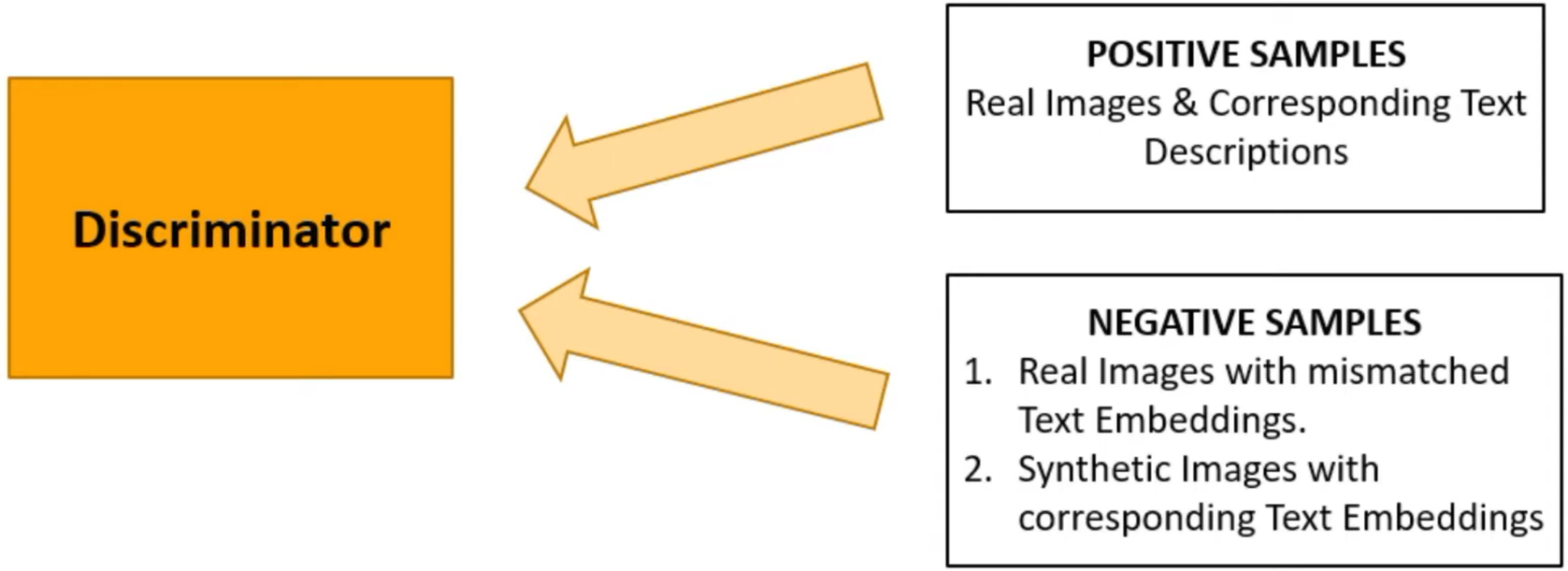
To generate high-resolution images with photo-realistic details, we propose a simple yet effective StackedGAN. It decomposes the text-to-image generative process into two stages :

Stage-I GAN: it sketches the primitive shape and basic colors of the object conditioned on the given text description, and draws the background layout from a random noise vector, yielding a low-resolution image.
Stage-II GAN: it corrects defects in the low-resolution image from Stage-I and completes details of the object by reading the text description again, producing a high resolution photo-realistic imag

StackGAN Architecture



Discriminator Model Training

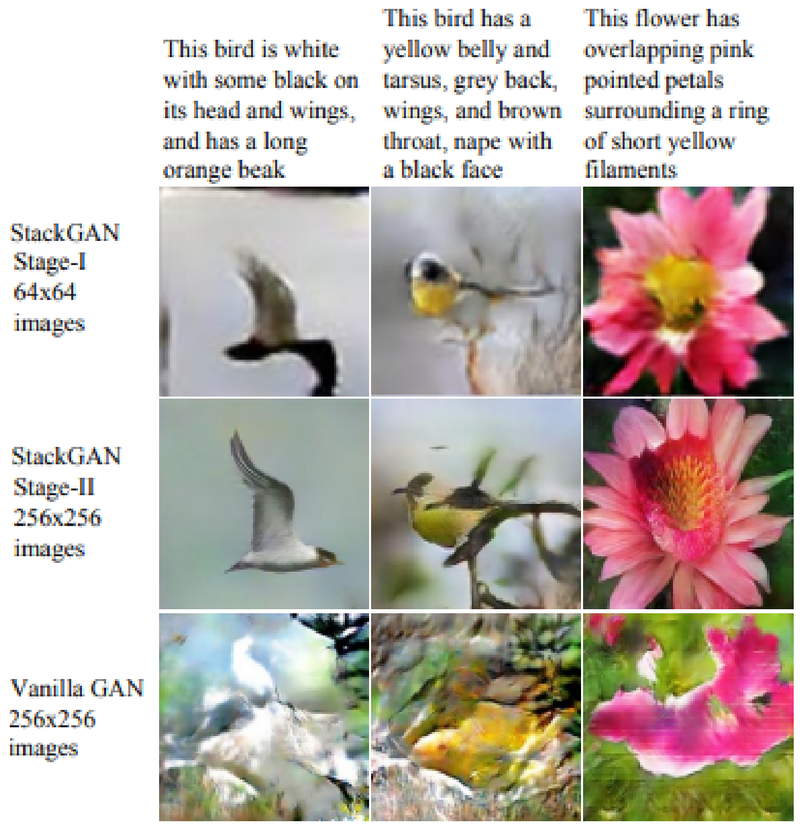


It is difficult to evaluate the per formance of generative models. We choose a recently proposed numerical assessment approach “inception score”for quantitative evaluation,

$$I = \exp(\mathbb{E}_{\boldsymbol{x}} D_{KL}(p(y|\boldsymbol{x}) \,||\, p(y)))$$

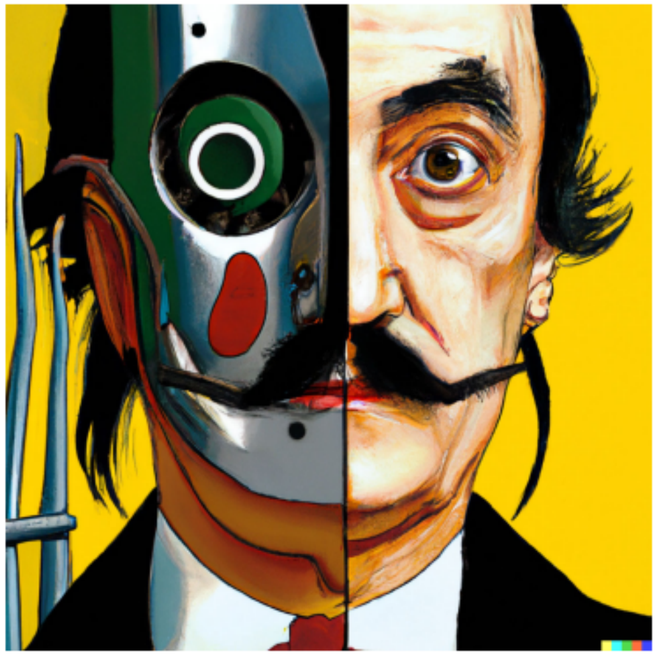
Dataset :

CUB Dataset - birds
Oxford-102 Dataset - flowers
MS COCO Dataset - backgrounds and multiple objects



Dall-E2

(Text-conditional Image generation)



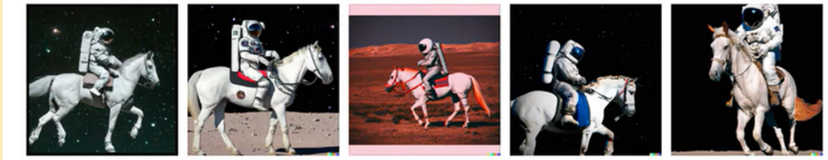
vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

Uses of Dall-E2

- Can create original realistic images and art from text description.it can combine concepts,attributes and styles.
- Generating image varaitions
- Language guided image manipulations(text-diff)
- Generating image interpolations
- Can make realistic edits to existing images from a natural language caption.It can add and remove elements while taking shadows ,reflections and textures into account.



An astronaut riding a horse in a photorealistic style

An astronaut riding a horse in the style of Andy Warhol

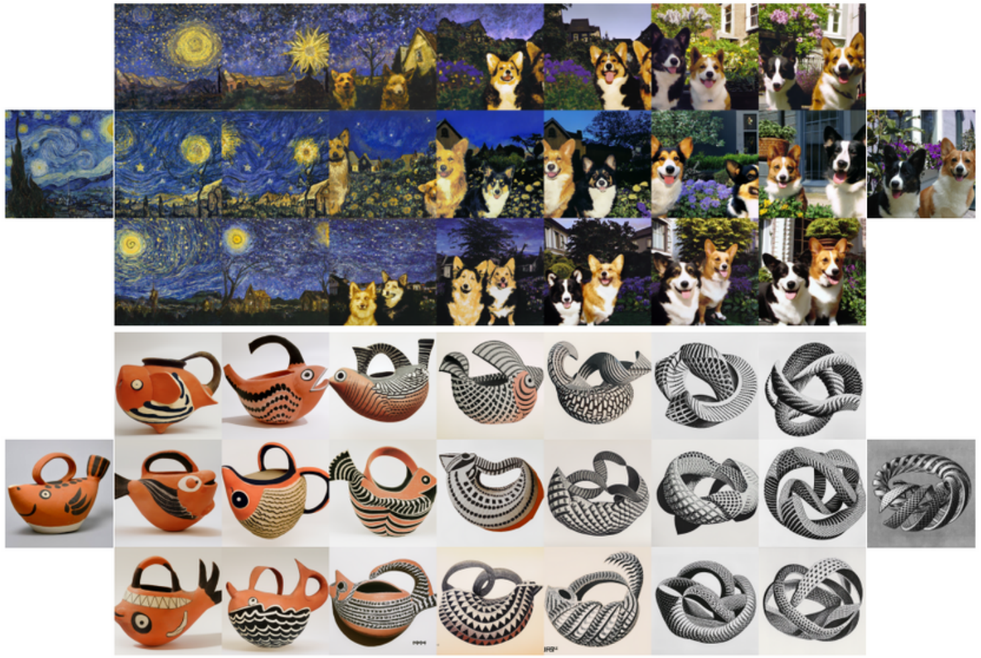
An astronaut riding a horse as a pencil drawing

a photo of a cat → an anime drawing of a super saiyan cat, artstation

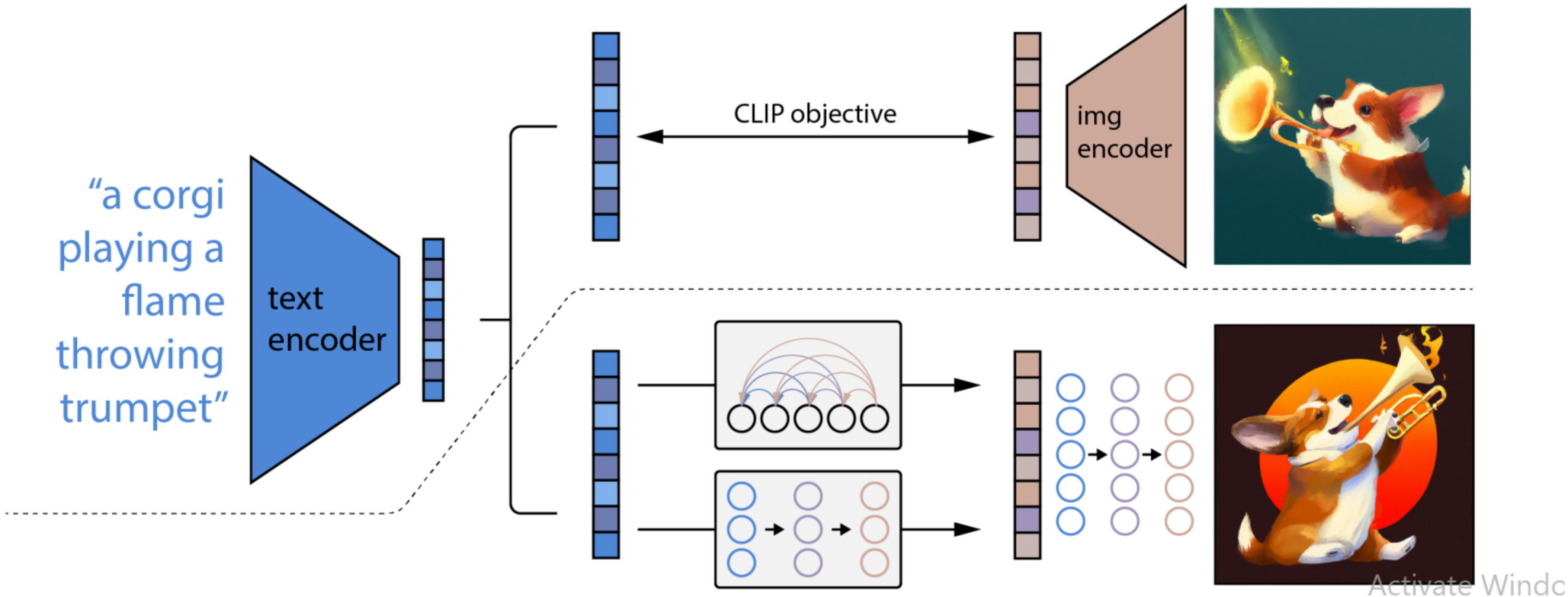
a photo of a victorian house → a photo of a modern house

a photo of an adult lion → a photo of lion cub

a photo of a landscape in winter → a photo of a landscape in fall

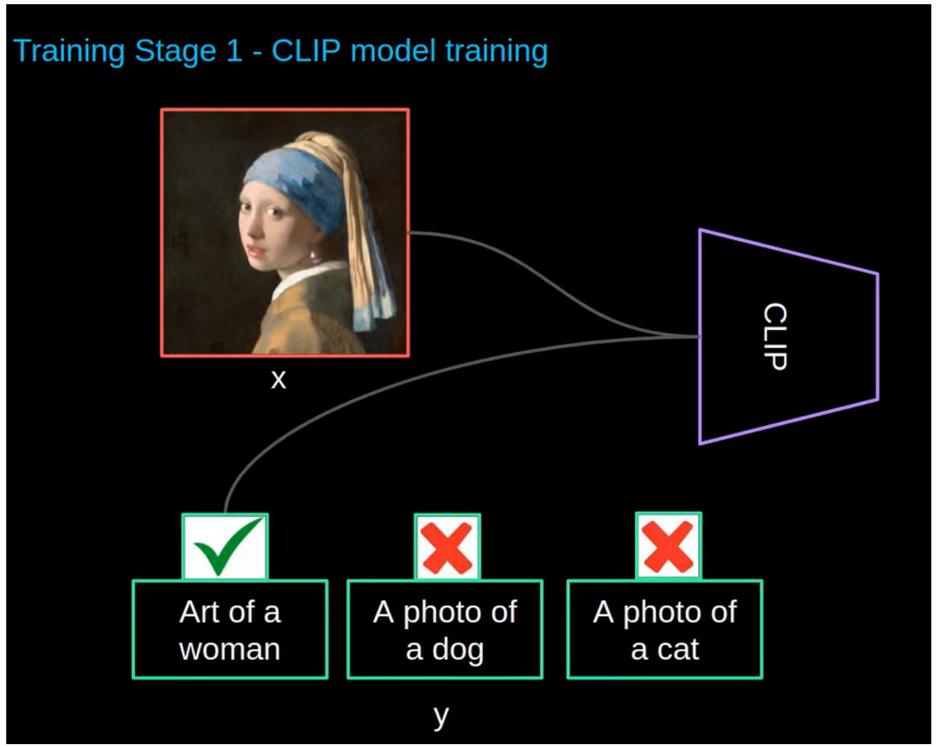


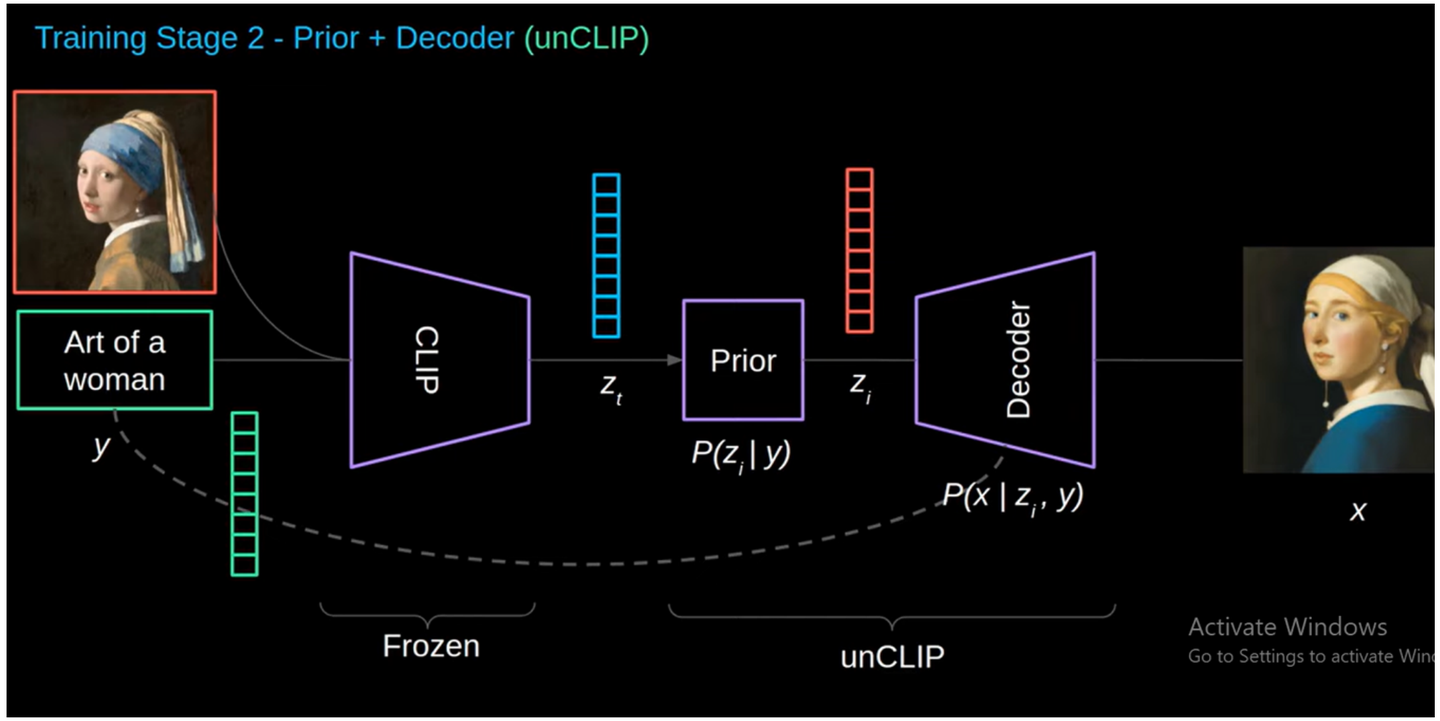
Dall-E2 Architecture



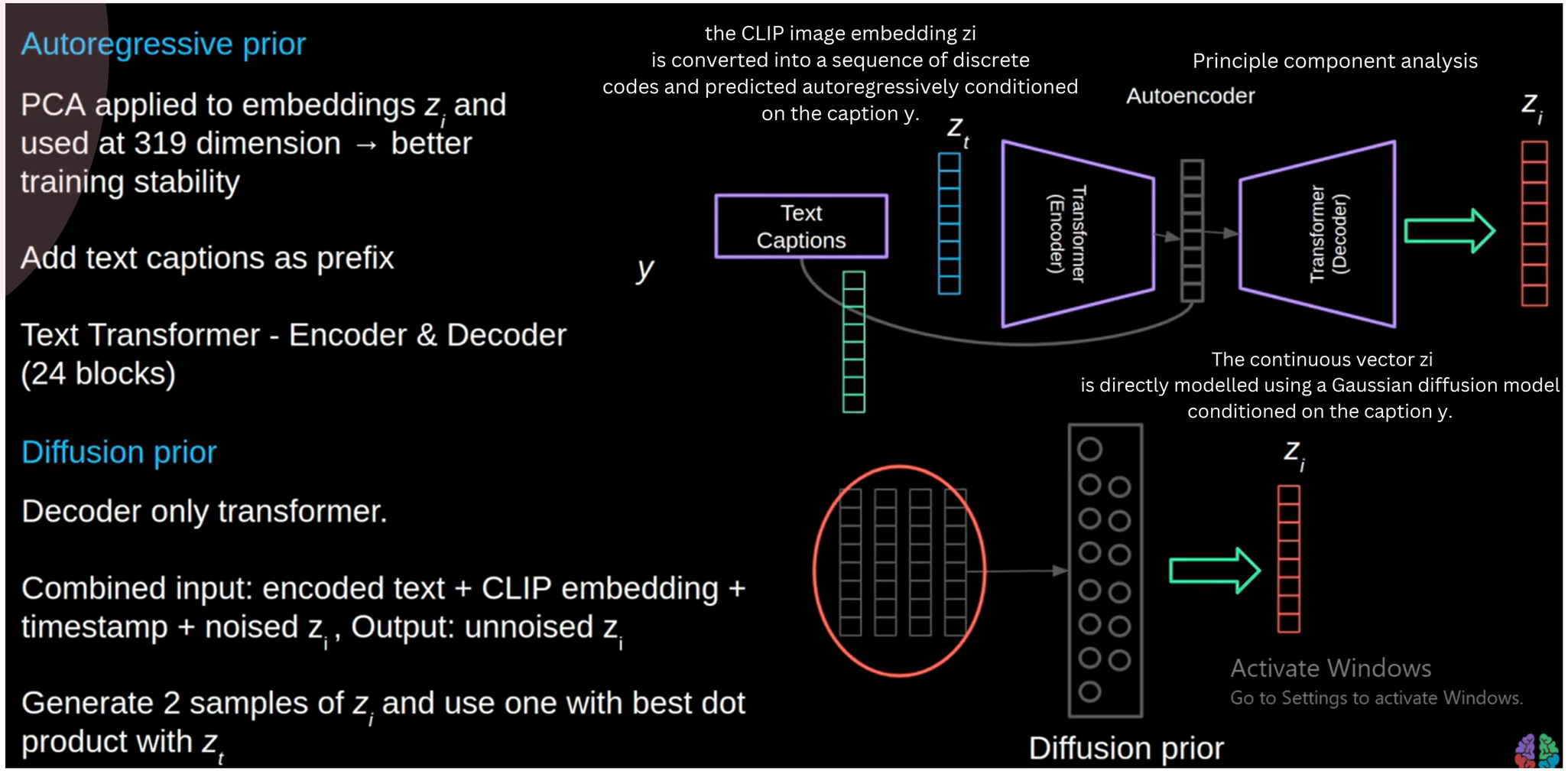
The two-stage model you mentioned consists of two key components: a prior and a decoder

Stages

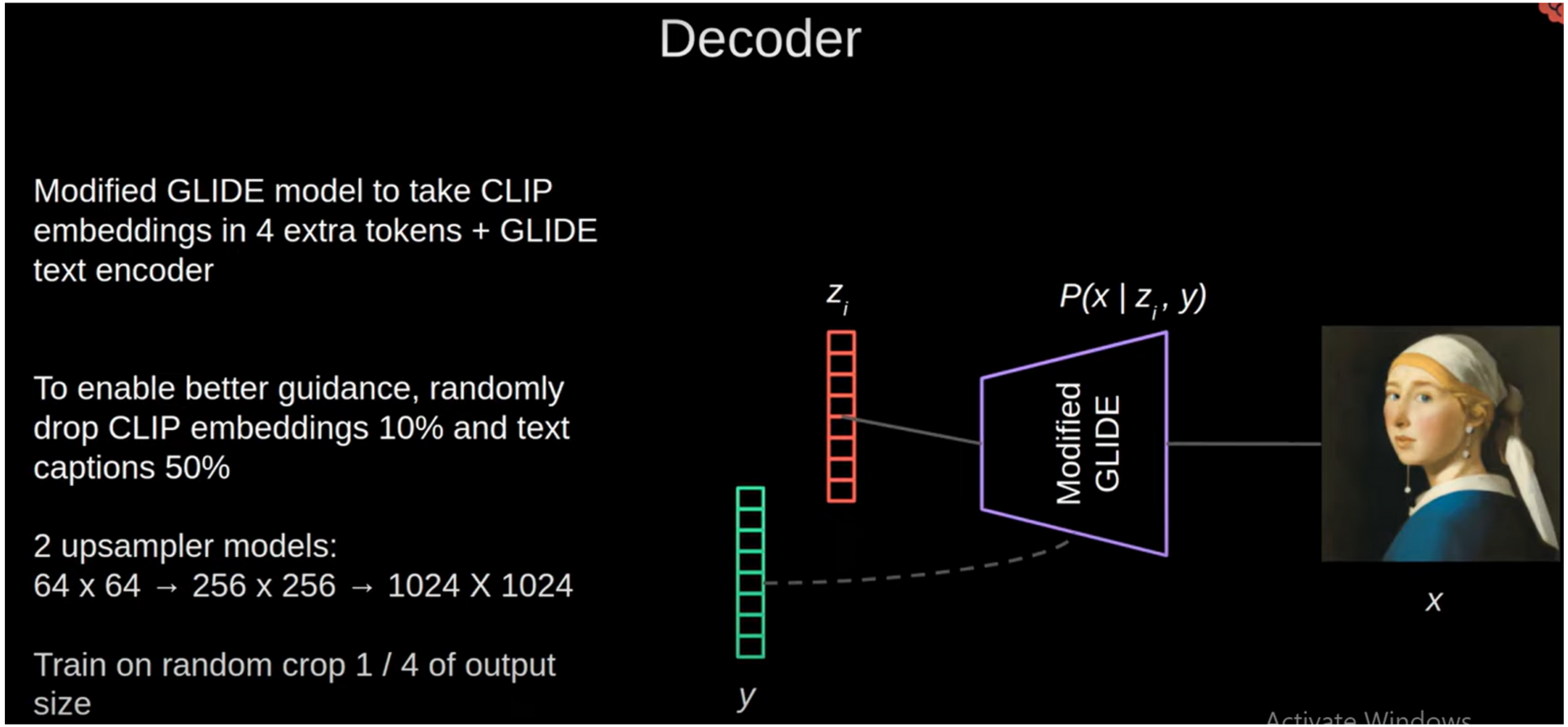

Training Stage 1 - CLIP model training
x
CLIP
Art of a woman
A photo of a dog
A photo of a cat
y


Training Stage 2 - Prior + Decoder (unCLIP)
Art of a woman
y
CLIP
z_t
Prior
$P(z_i \mid y)$
z_i
Decoder
$P(x \mid z_i, y)$
x
Frozen
unCLIP
Activate Windows

Prior Model (Autoregressive or Diffusion): The CLIP text embedding is then fed into a prior model. This prior model can either be autoregressive or diffusion-based. The prior's role is to generate an image embedding based on the CLIP text embedding. This image embedding is a numerical representation that encodes the information needed to create an image that aligns with the text's content and style.



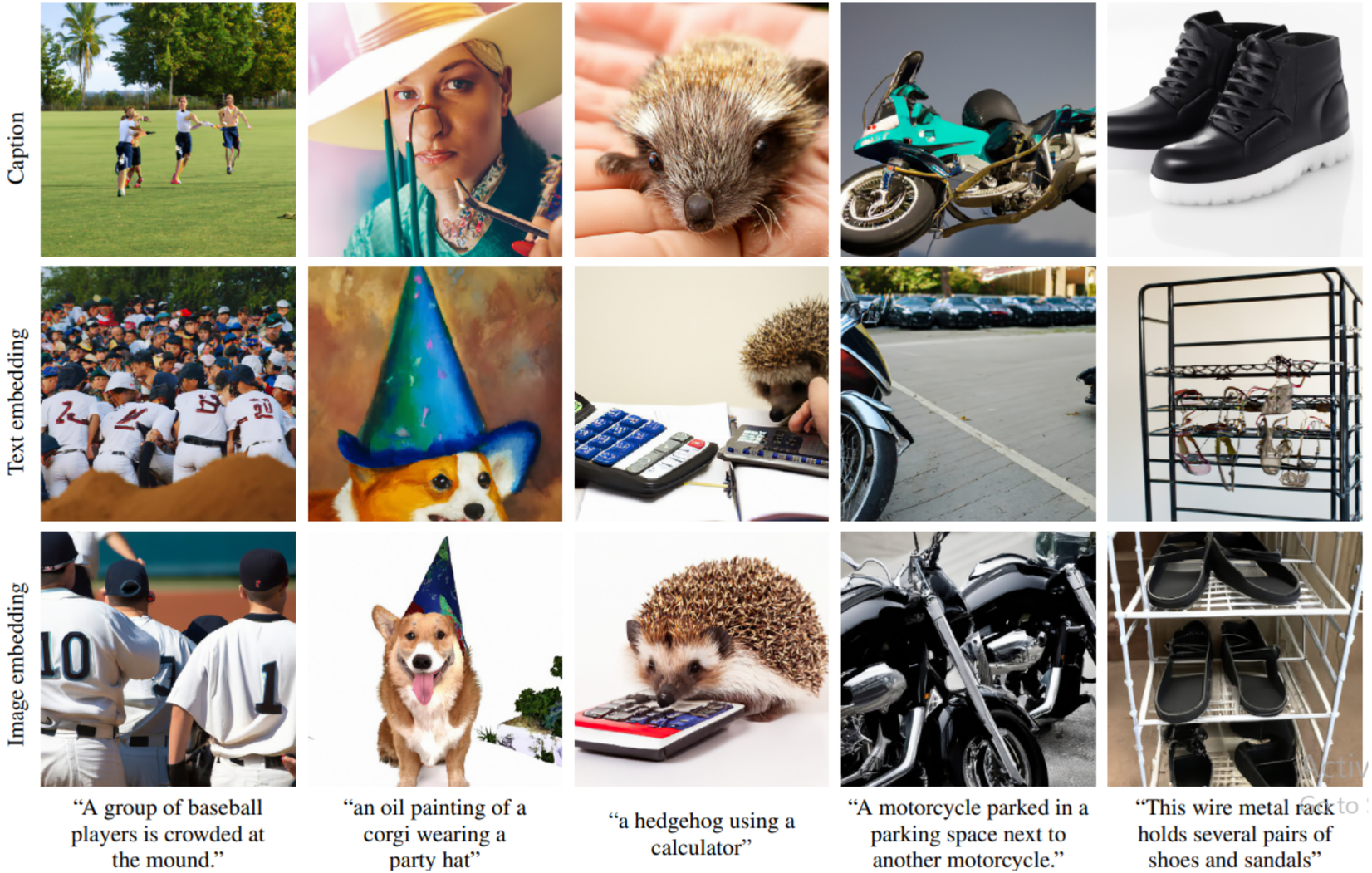
Diffusion Decoder: The image embedding is used to condition a diffusion decoder. The diffusion decoder is responsible for taking this embedding and transforming it into the final image. The decoder generates pixel-level details and overall image structure based on the information contained in the image embedding.



Variations of an input image by encoding with CLIP and then decoding with a diffusion model.The decoder allows us to invert images given their CLIP image embeddings, while the prior allows us to learn a generative model of the image embeddings themselves

$$P(x|y) = P(x, z_t|y) = P(x|z_i, y) = P(z_i|y)$$

Importance of prior



Human Evaluation

unCLIP Prior	Photorealism	Caption Similarity	Diversity
AR	$47.1\% \pm 3.1\%$	$41.1\% \pm 3.0\%$	$62.6\% \pm 3.0\%$
Diffusion	$48.9\% \pm 3.1\%$	$45.3\% \pm 3.0\%$	$70.5\% \pm 2.8\%$

Model	FID	Zero-shot FID	Zero-shot FID (filt)
AttnGAN (Xu et al., 2017)	35.49		
DM-GAN (Zhu et al., 2019)	32.64		
DF-GAN (Tao et al., 2020)	21.42		
DM-GAN + CL (Ye et al., 2021)	20.79		
XMC-GAN (Zhang et al., 2021)	9.33		
LAFITE (Zhou et al., 2021)	8.12		
Make-A-Scene (Gafni et al., 2022)	**7.55**		
DALL-E (Ramesh et al., 2021)		~ 28	
LAFITE (Zhou et al., 2021)		26.94	
GLIDE (Nichol et al., 2021)		12.24	12.89
Make-A-Scene (Gafni et al., 2022)			11.84
unCLIP (AR prior)		10.63	11.08
unCLIP (Diffusion prior)		**10.39**	**10.87**

Frechet Inception Distance (FID)
FID measures the similarity between the distribution of real images and generated images in feature space, using the Inception model's activations. Lower FID scores are indicative of better quality and diversity.

Main differences between Dall E & Dall E2

- **Clarity between visuals and texts + speedy results** : DALL-E generates realistic visuals and art from simple text.DALL-E 2 employs a technique known as “diffusion,” which begins with a pattern of random dots and gradually changes that pattern to resemble a picture when it recognizes particular characteristics of that image.

- **Realistic and high-resolution images** : The first version of DALL-E could only render AI-created images. DALL-E 2 can produce realistic images, which shows how superior it is at bringing all ideas to life.

- **Editing and retouching made simpler** : DALL-E “inpaints” or intelligently replaces specific areas in an image. DALL-E 2 has far more possibilities, including the ability to create new items.

- **Ability to produce multiple iterations of an image** : DALL-E 2 has a new feature called variations, where you provide the AI image generator with a sample image and generate as many variations as you want, ranging from near approximations to impressions. You can add another image, which will cross-pollinate the two, merging the most important parts of each.

Stable diffusion

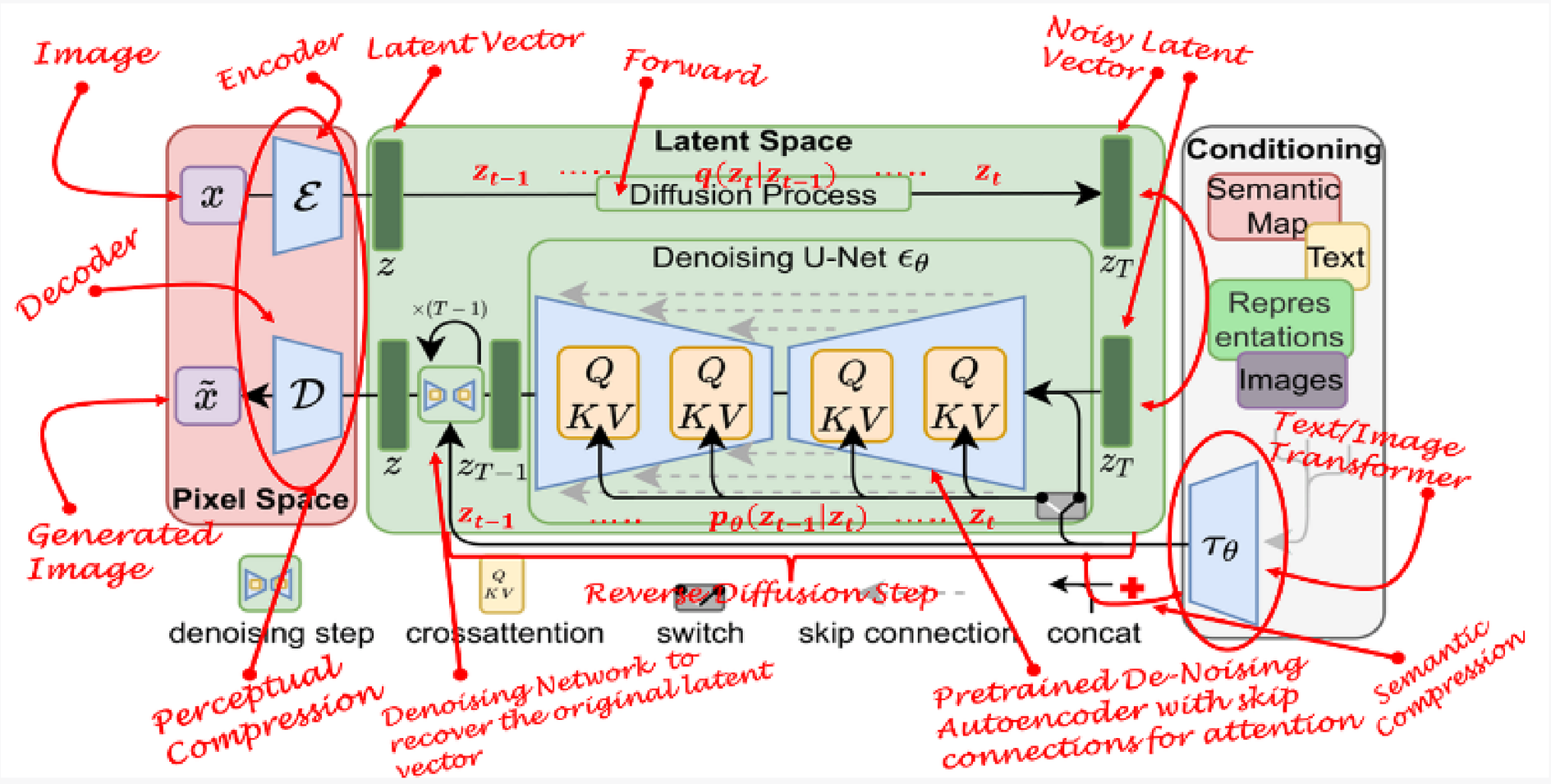
Stable Diffusion is a latent text-to-image diffusion model capable of generating photo-realistic images given any text input.

Stable Diffusion is a text-to-image model created by a collaboration between engineers and researchers from CompVis, Stability AI, and LAION.

Uses of the model & comparison with competitor models

- In contrast to previous works, stable diffusion does not require this delicate weighting of the reconstruction and generative abilities, which allows for more faithful reconstructions of images with relatively little regularization of the latent space.
- One of the key ways Stable Diffusion differs from past methodologies for diffusion modeling is the ability to scale much more easily.
- Applications in educational or creative tools
- Research on generative models

MODEL ARCHITECTURE

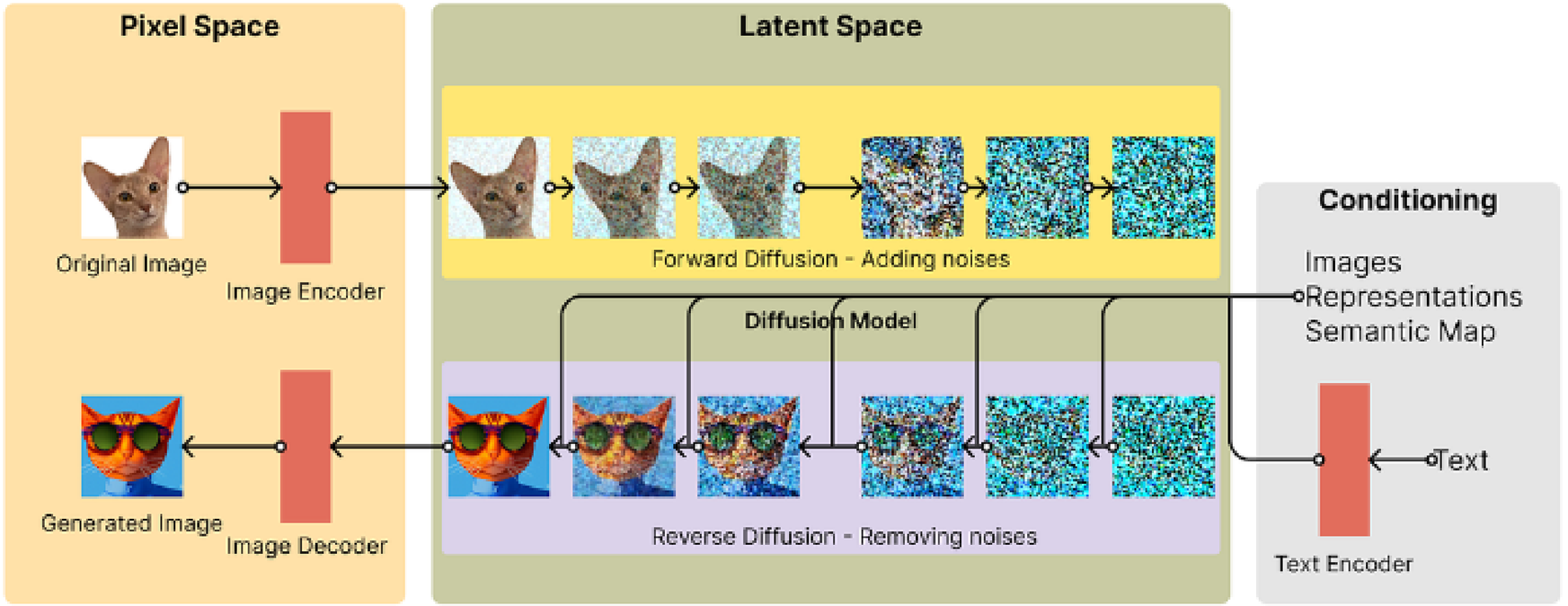


How do we go from Diffusion to Text-to-Image Generation ?

Classifier Guidance : In Stable Diffusion, we make use of something called CLIP embeddings to guide the diffusion towards the target class during the training. CLIP stands for Contrastive Loss Image Pair. The ideas behind this is to make the image and word embeddings similar in their semantics.
Classifier Free Guidance : How can it come up with images it has never seen before ? instead of one noisy image, two same images are fed to model - one without the text embeddings and one with it. The diffusion model therefore comes up with two image - one without text embeddings and one without it. Together, these two noise images* can be used to amplifiy the signal and generate images which are previously not generated.

*Rememer that the model generates noise, not the actual immage

MODEL ARCHITECTURE



What makes Stable Diffusion so special and fit for art generation, is that it was actually an LDM trained on a core dataset consisting of, we cite “**LAION-Aesthetics**, a soon to be released subset of **LAION 5B**

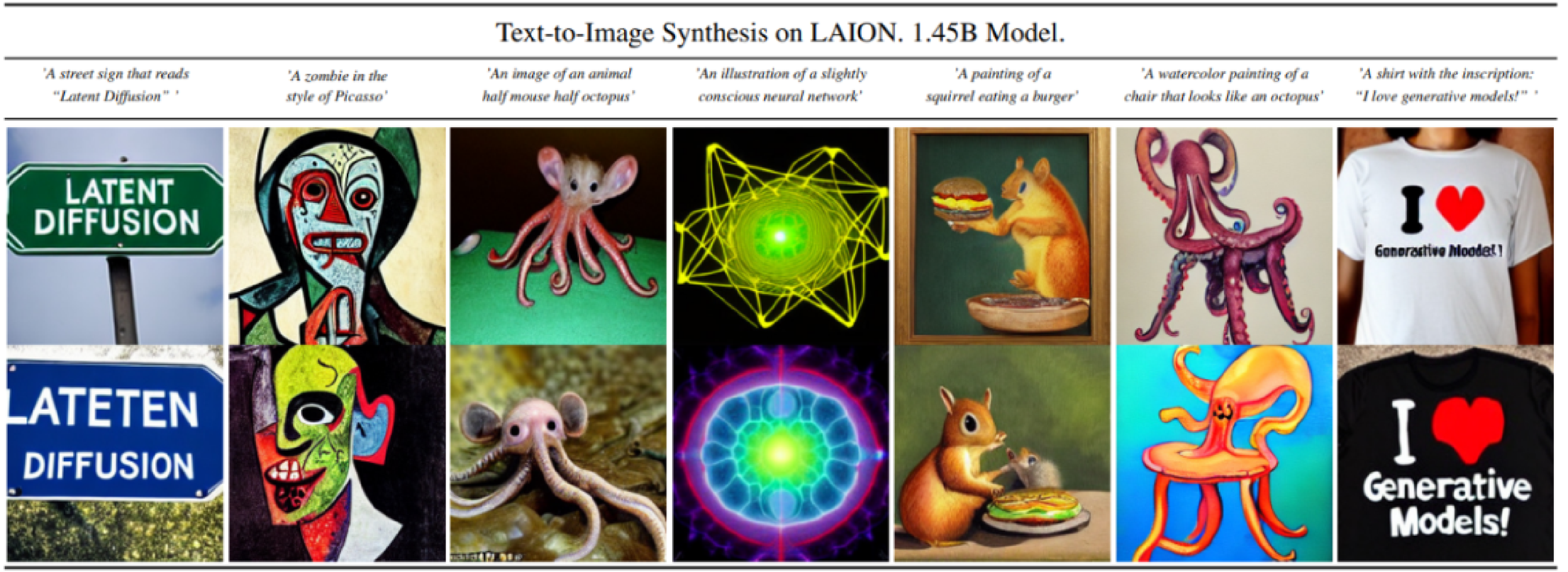


Figure 5. Samples for user-defined text prompts from our model for text-to-image synthesis, *LDM-8 (KL)*, which was trained on the LAION [78] database. Samples generated with 200 DDIM steps and $\eta = 1.0$. We use unconditional guidance [32] with $s = 10.0$.

Evaluation

Inception Score (IS): The Inception Score measures the quality and diversity of generated images. It uses a pre-trained Inception model to compute a score that balances image quality and diversity. A higher IS generally indicates better results.

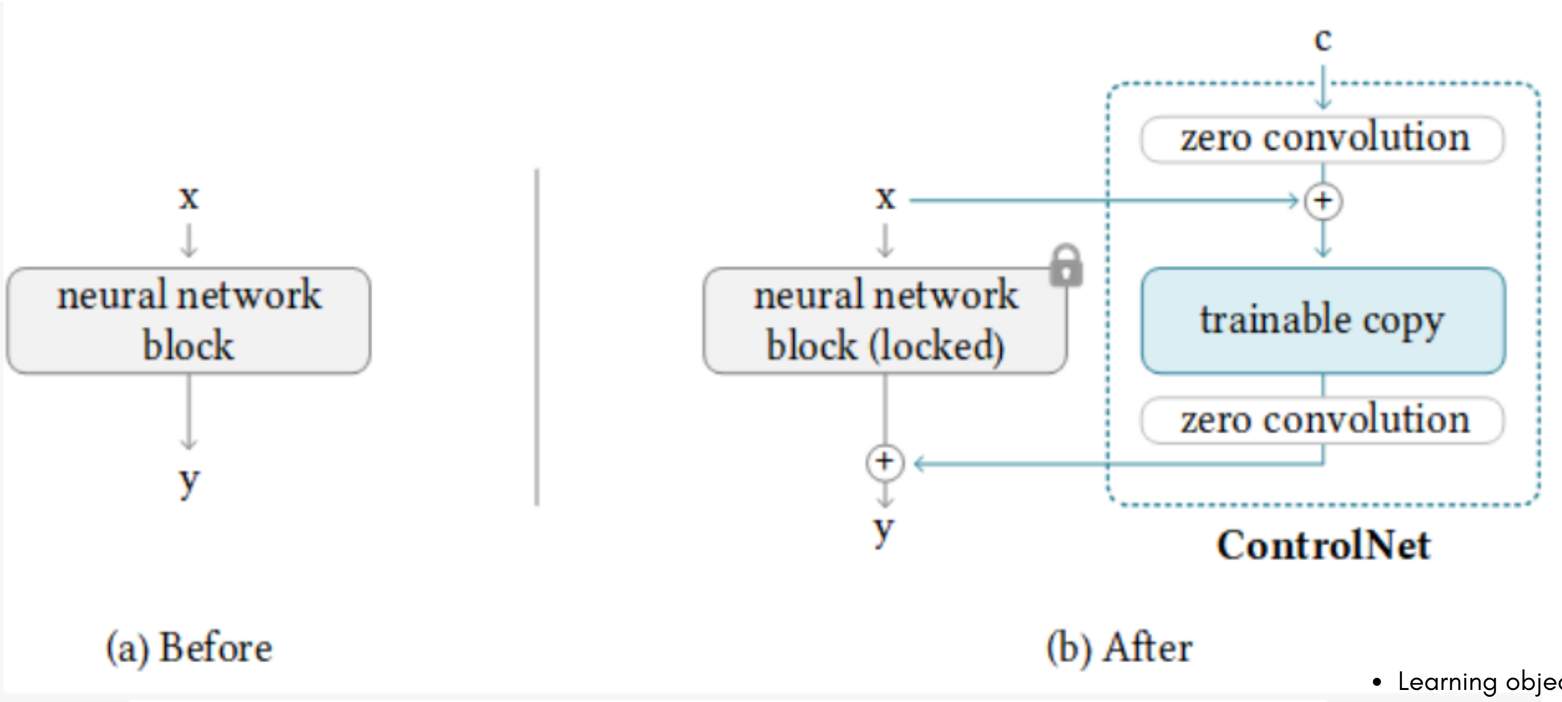
Frechet Inception Distance (FID) : FID measures the similarity between the distribution of real images and generated images in feature space, using the Inception model's activations. Lower FID scores are indicative of better quality and diversity.

Precision and Recall: Precision and recall metrics can assess the relevance of generated images to the input text. Precision measures how many generated images are relevant, while recall measures how many relevant images are generated.

LSUN-Churches 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
DDPM [30]	7.89	-	-
ImageBART [21]	7.32	-	-
PGGAN [39]	6.42	-	-
StyleGAN [41]	4.21	-	-
StyleGAN2 [42]	3.86	-	-
ProjectedGAN [76]	**1.59**	0.61	0.44
*LDM-8** (ours, 200-s)	4.02	**0.64**	**0.52**

CONTROLNET

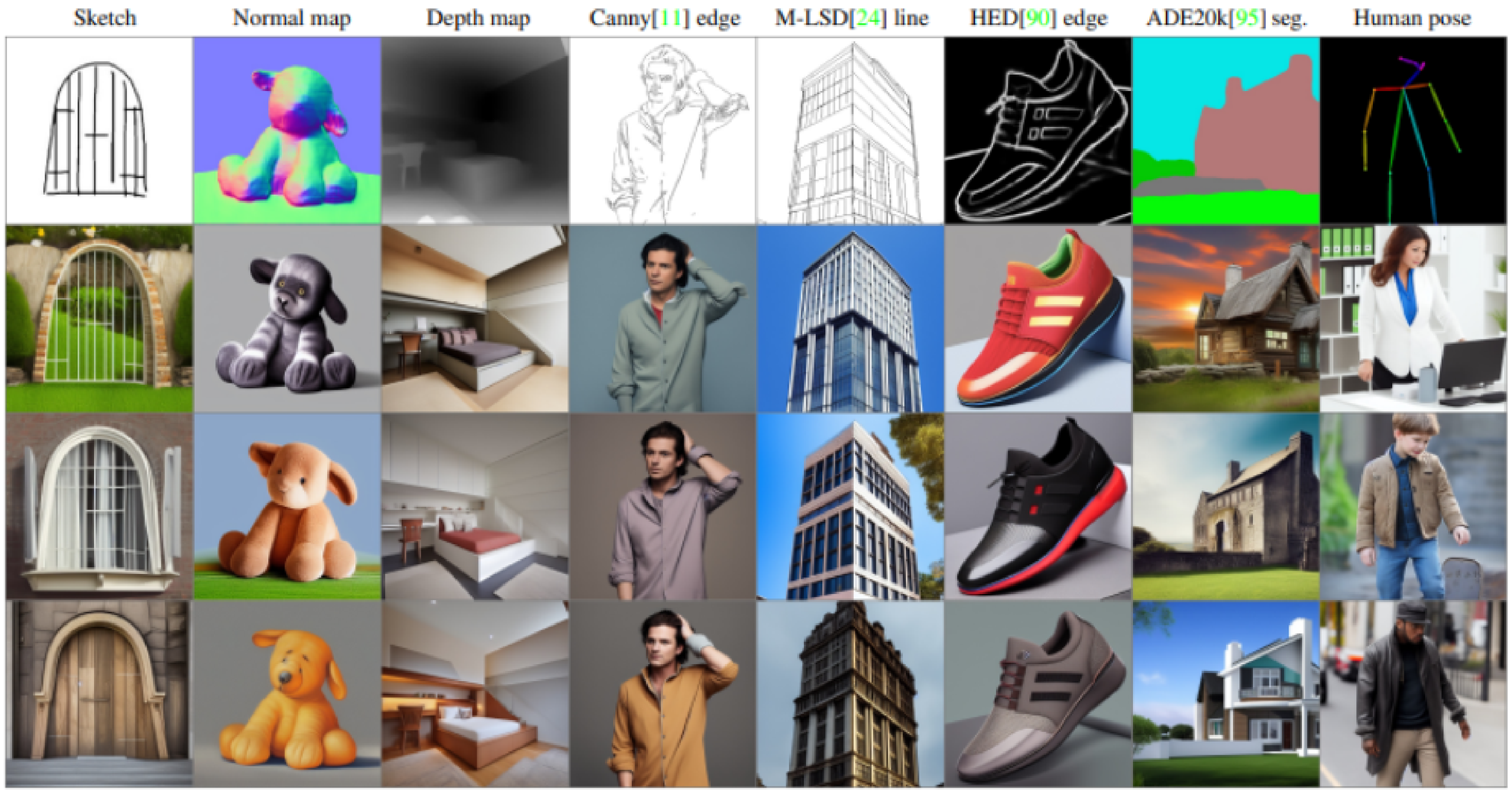
ControlNet is a neural network structure to control diffusion models by adding extra conditions.It copys the weights of neural network blocks into a "locked" copy and a "trainable" copy.
The "trainable" one learns your condition. The "locked" one preserves your model.



$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_f, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_f) \|_2^2 \right]$$

- Learning objective function
- This approach increases ControlNet ability to directly recognise semantics in the input conditioning images as a replacement for the prompts

we first convert each input conditioning image (e.g., edge, pose, depth, etc.) from an input size of 512 × 512 into a 64 × 64 feature space vector that matches the size of Stable Diffusion. We use a tiny network E(·) of four convolution layers with 4 × 4 kernels and 2 × 2 strides to encode an image-space condition ci into a feature space conditioning vector cf as, cf = E(ci). (4) The conditioning vector cf is passed into the ControlNet.



Thank You