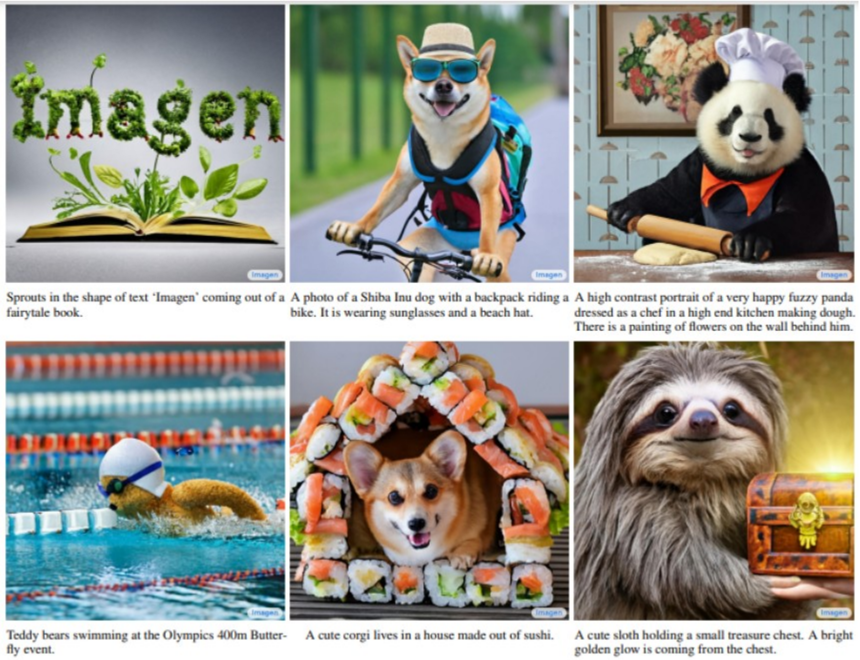
Machine learning introduction

Part II – Deep Neural Networks

3 – Text-2-Image networks

Simon Gay

- DALLE-2 (OpenAI)
- Imagen (Google)
- Midjourney
- Stable-Diffusion



Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

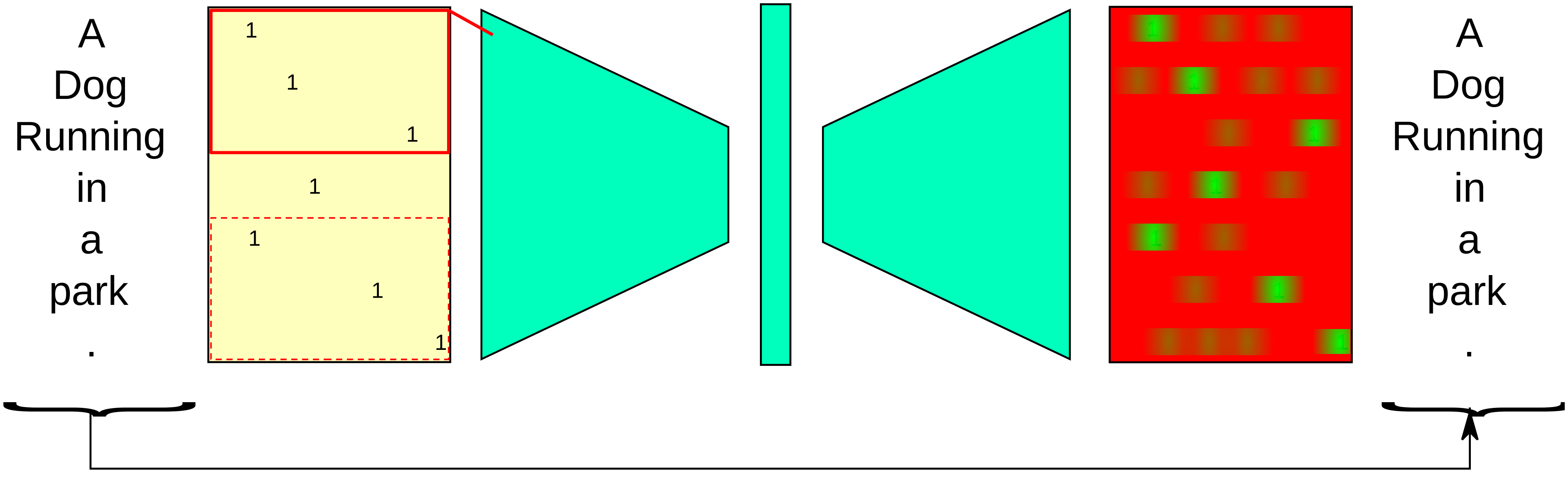
Teddy bears swimming at the Olympics 400m Butterfly event.

A cute corgi lives in a house made out of sushi.

A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

Part II-3 : Text-2-Image

- **Text2text**
  - A word is represented through a vector with the size of the dictionary (‘dog’ = [0,0,0,0,….,0,0,1,0,0,…,0,0])
  - A sentence is represented as a matrix of size n_{words} x $n_{dictionary}$
  - On first layers, the network uses 1D convolutional neurons, covering a word and its closest neighbors.



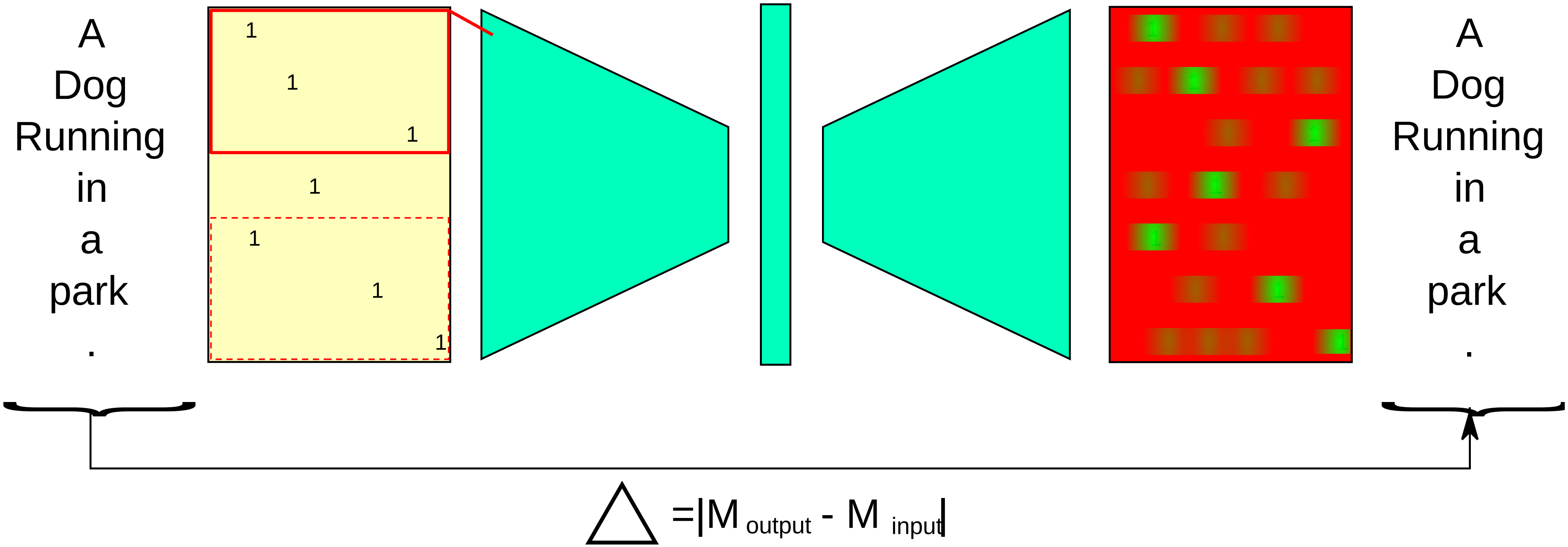
A Dog Running in a park .

A Dog Running in a park .

Δ =|M $_{output}$ - M $_{input}$|

Part II-3 : Text-2-Image

- **Text2text**
  - Consequence: a word cannot be dissociated from its context and meaning
    - Two words with similar meaning will be close in latent space
    - A word with multiple definitions will be represented by multiple distant vectors in latent space (ex : ‘close’ will be associated to different vectors if its context contains ‘travel’ or ‘door’)



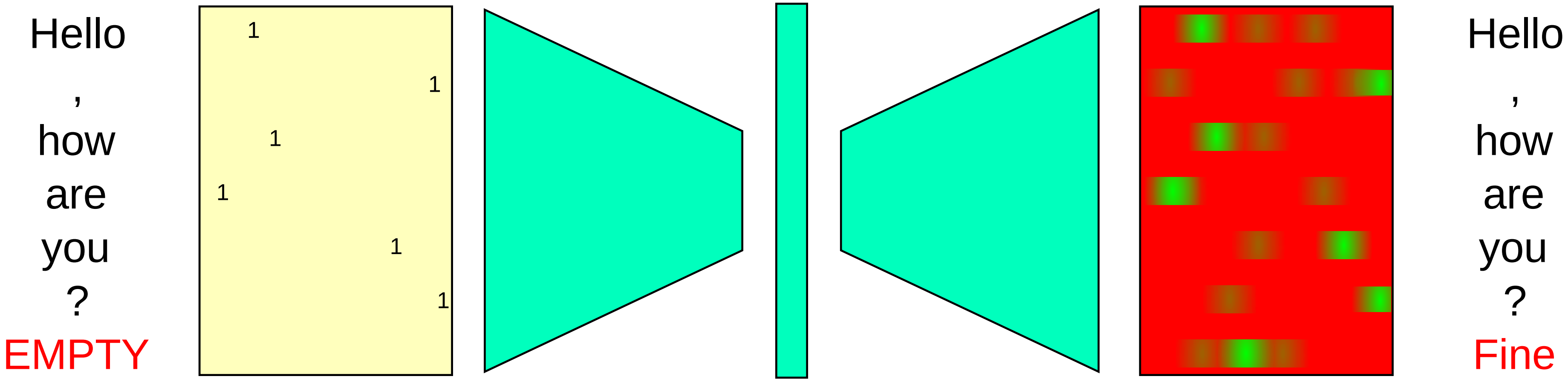
A
Dog
Running
in
a
park
.

A
Dog
Running
in
a
park
.

$\Delta = |M_{output} - M_{input}|$

Part II-3 : Text-2-Image

- **Text2text**
  - Exploitation : a word is masked, the network learns to ‘guess’ the missing word by using its context (BERT, GPT)
  - ChatGPT 3 is a very elaborated version of this principle
    - Defining the next most probable word

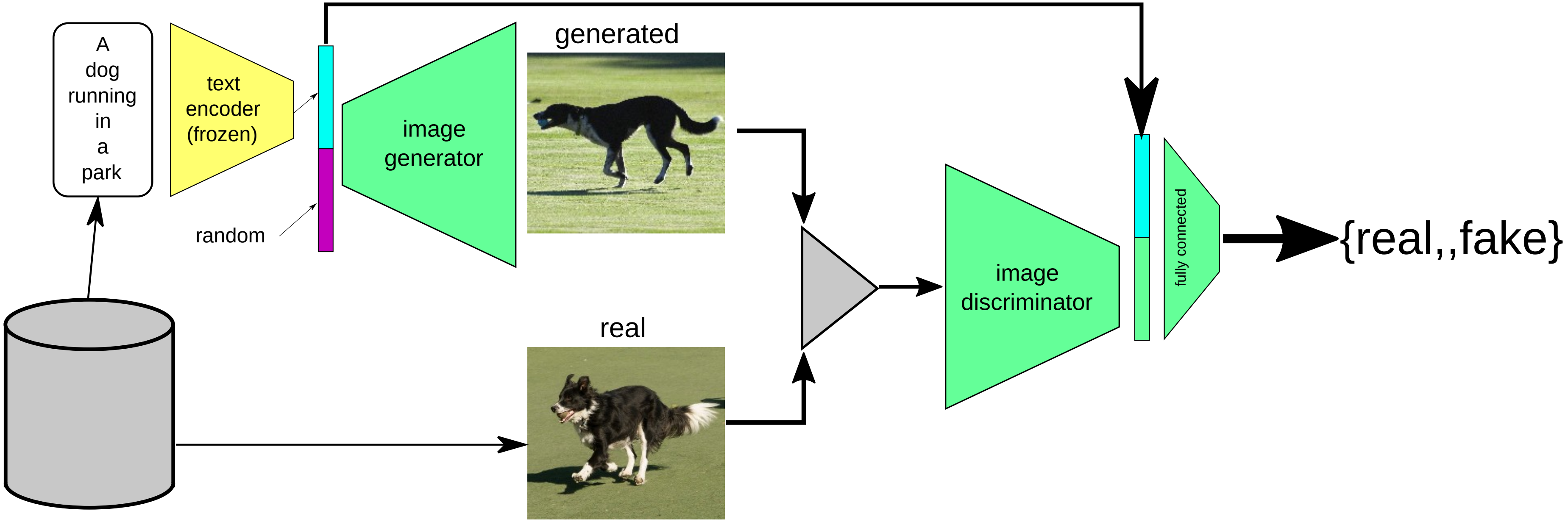


Hello
,
how
are
you
?
EMPTY

Hello
,
how
are
you
?
Fine

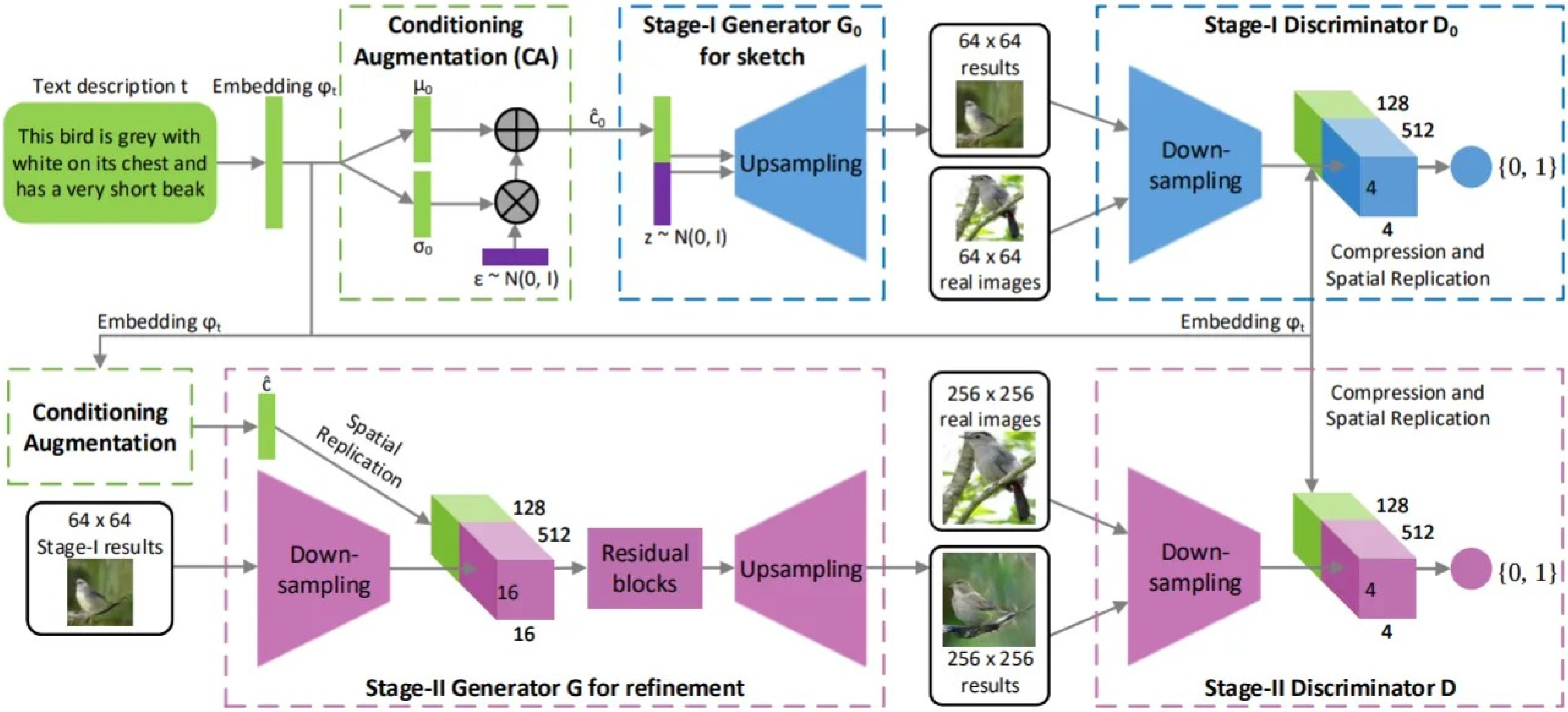
Part II-3 : Text-2-Image

- **Text2image: conditional GAN approach**
  - Encoder network (pre-trained) for text: condition vector
  - Conditional GAN uses both random vector (seed) and condition vector as input
  - Discriminator: comparison of image-condition vector pairs



Part II-3 : Text-2-Image

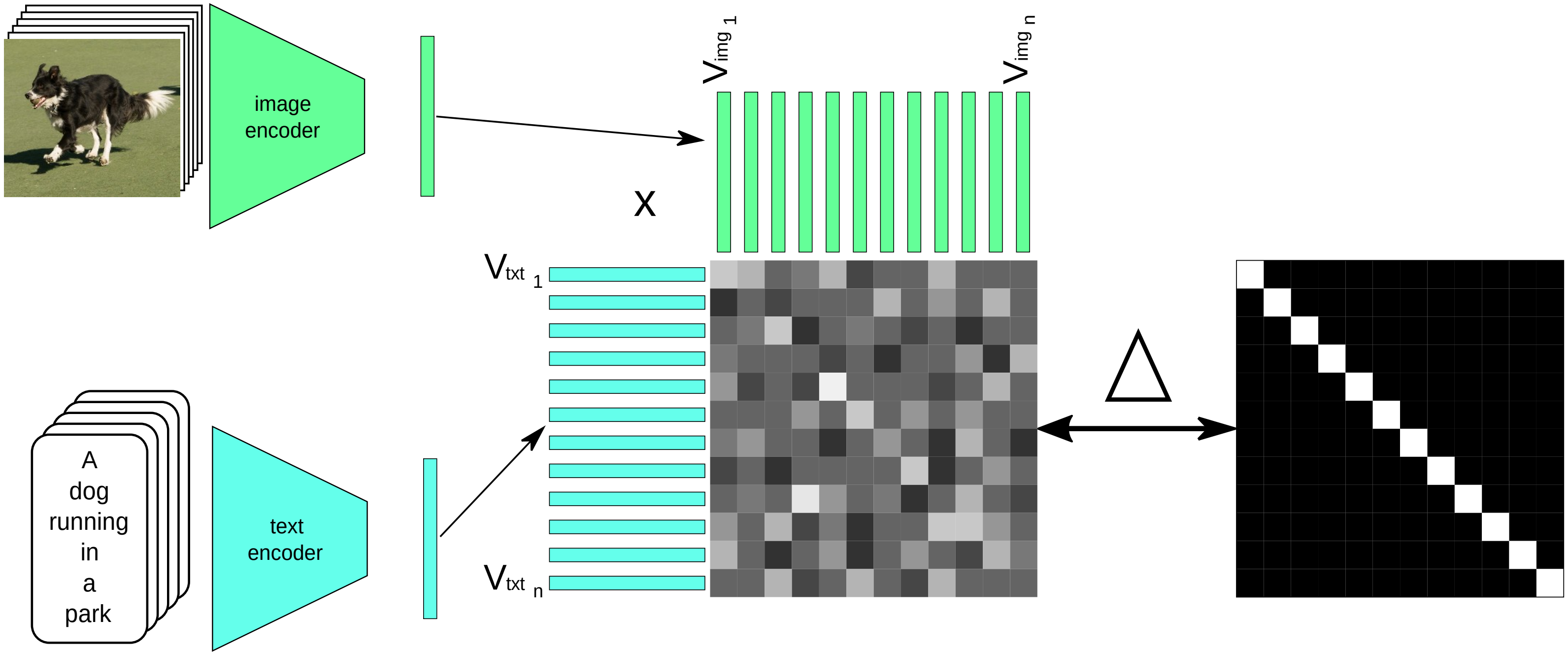
- **Text2image: conditional GAN approach**
  - StackGAN[1] use a second ‘stage’ with conditional autoencoder to increase generated image resolution



[1] https://arxiv.org/abs/1612.03242

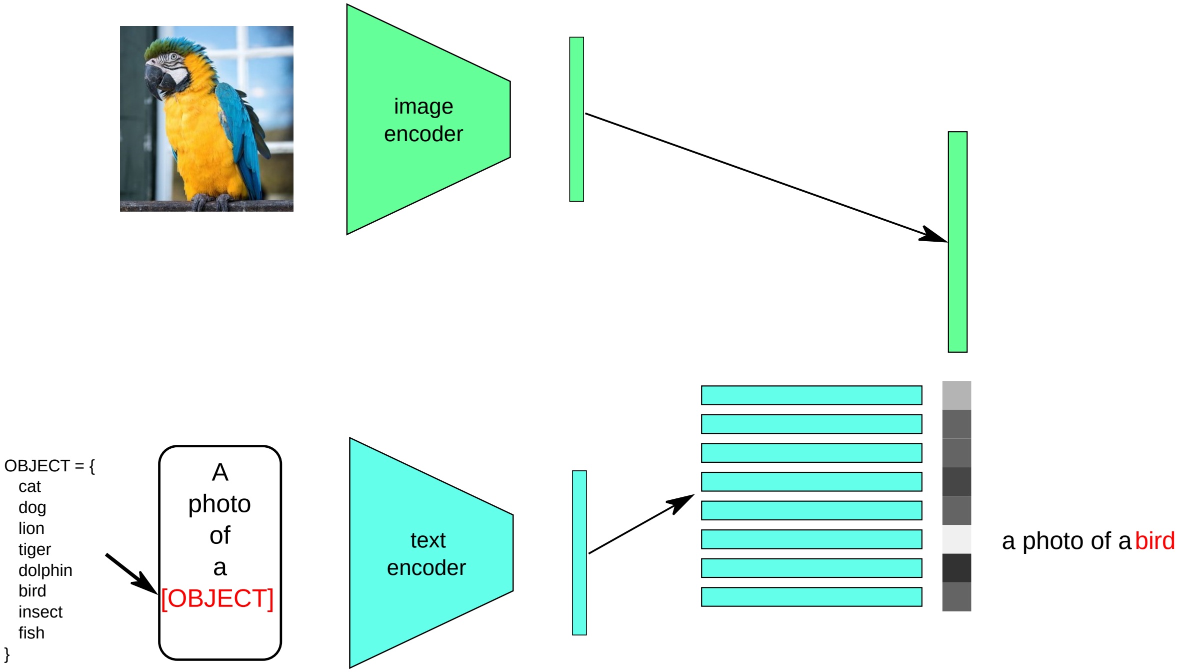
Part II-3 : Text-2-Image

- **Linking models : CLIP (Contrastive Language–Image Pre-training)**
  - Two encoder networks: one for images, one for text texte
  - Correlation matrix: scalar product between latent vectors, result must be 1 if image and text are related, 0 otherwise
  - Idea: forcing latent spaces of the two networks to converge toward the same distribution: the sentence ‘a dog running in a park’ will be represented by a similar vector than an image representing this scene



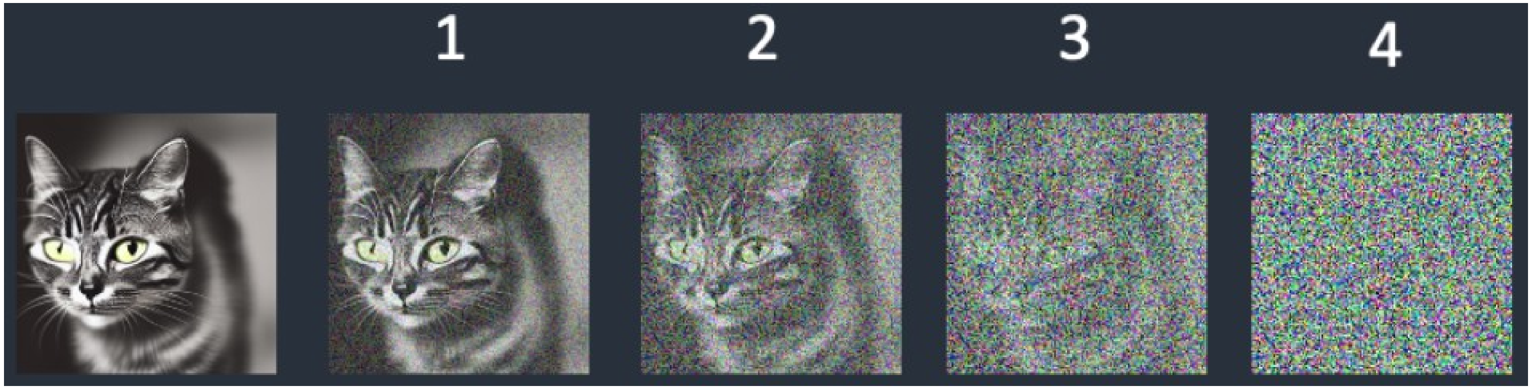
Part II-3 : Text-2-Image

- **Linking models : CLIP (Contrastive Language–Image Pre-training)**
  - A set of model sentences is created
  - An image is presented to the network
  - Test of categories: estimation of the most probable category
  - Categories can be defined after network training ! (Zero-Shot prediction)



Part II-3 : Text-2-Image

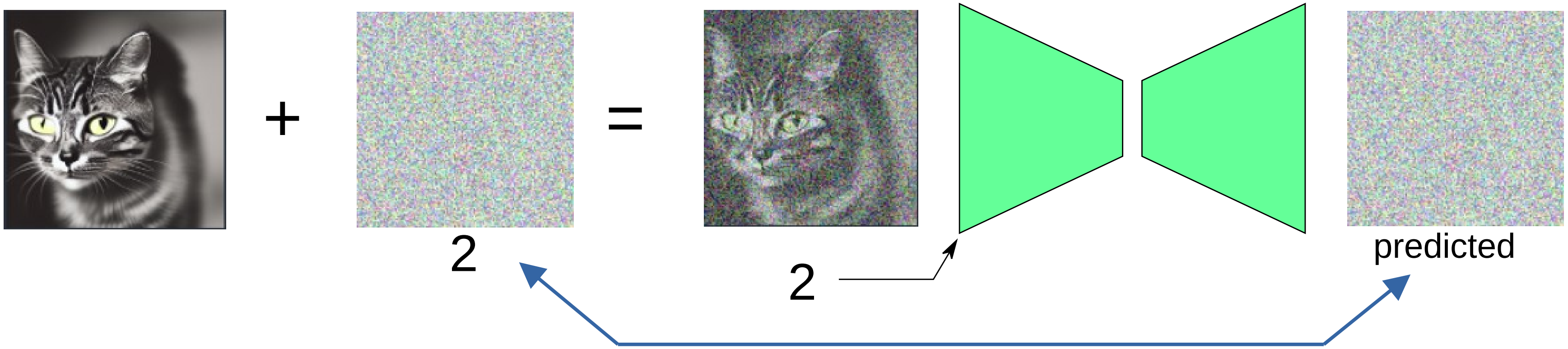
- **Diffusion**
  - Principle: noise is added to an image, the network learns to de-noise the image
  - Different level of noise are added to train images



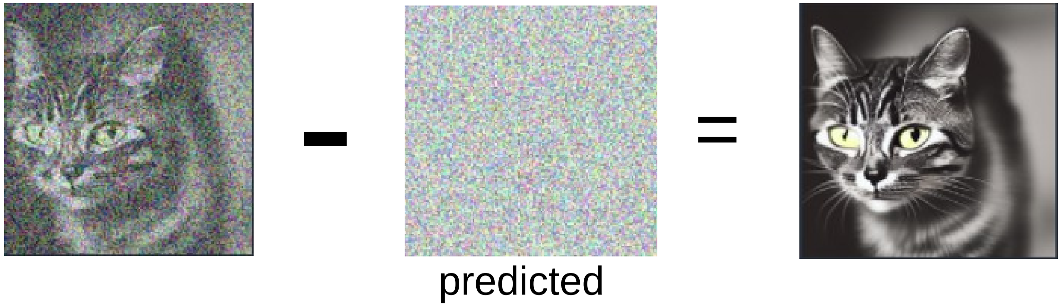
https://stable-diffusion-art.com/how-stable-diffusion-work/

Part II-3 : Text-2-Image

- **Diffusion**
  - A network learns to predict the added noise mask

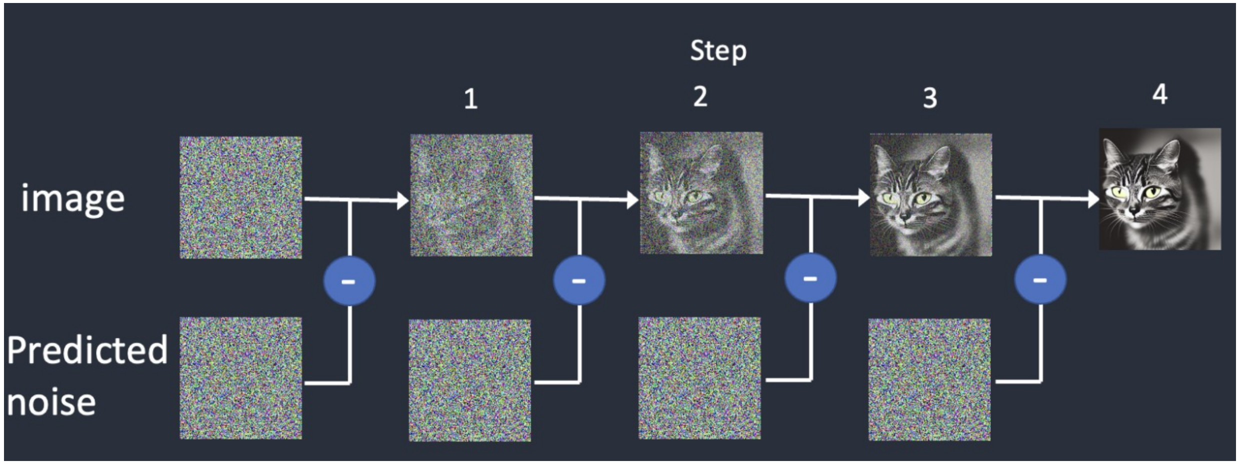


  - Theoretically, if the noise mask is subtracted, we should obtain the initial image



Part II-3 : Text-2-Image

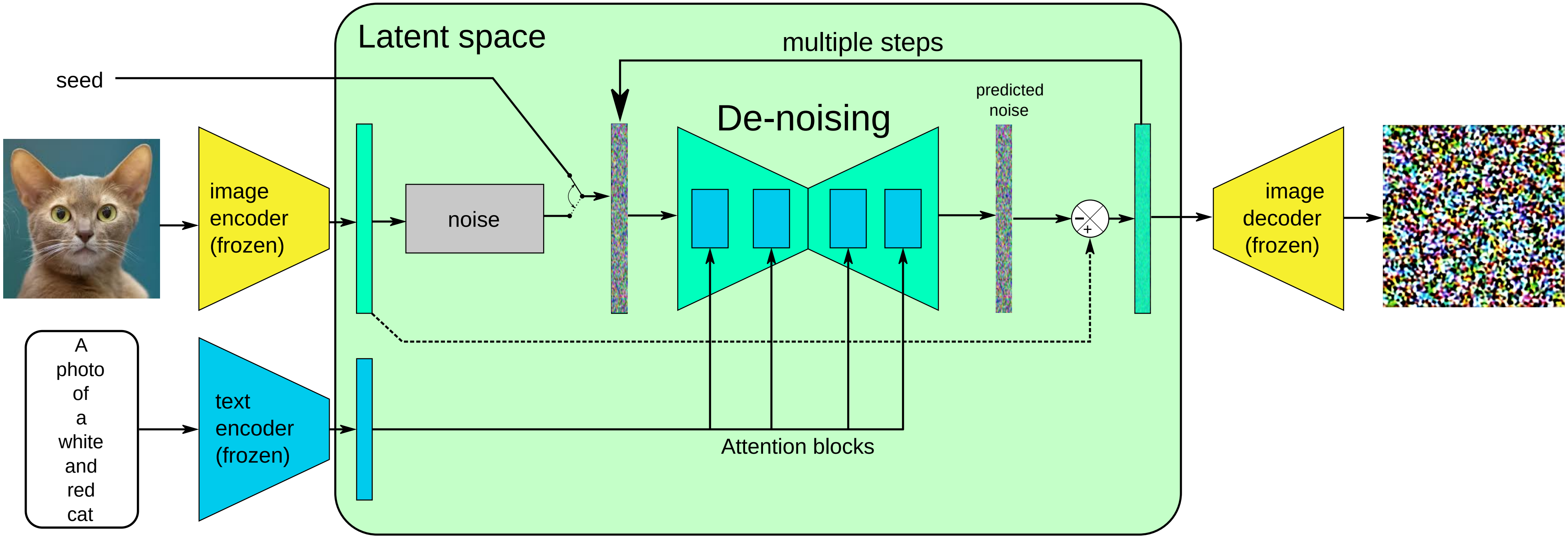
- **Diffusion**
  - If this principle is applyed to a random noise image, the network will still predict a noise mask that is subtracted from the image



  - The process can be repeated multiple times (‘diffusion’) to obtain an image
    - Form of artificial pareidolia
    - Generated images will depend on the training database !

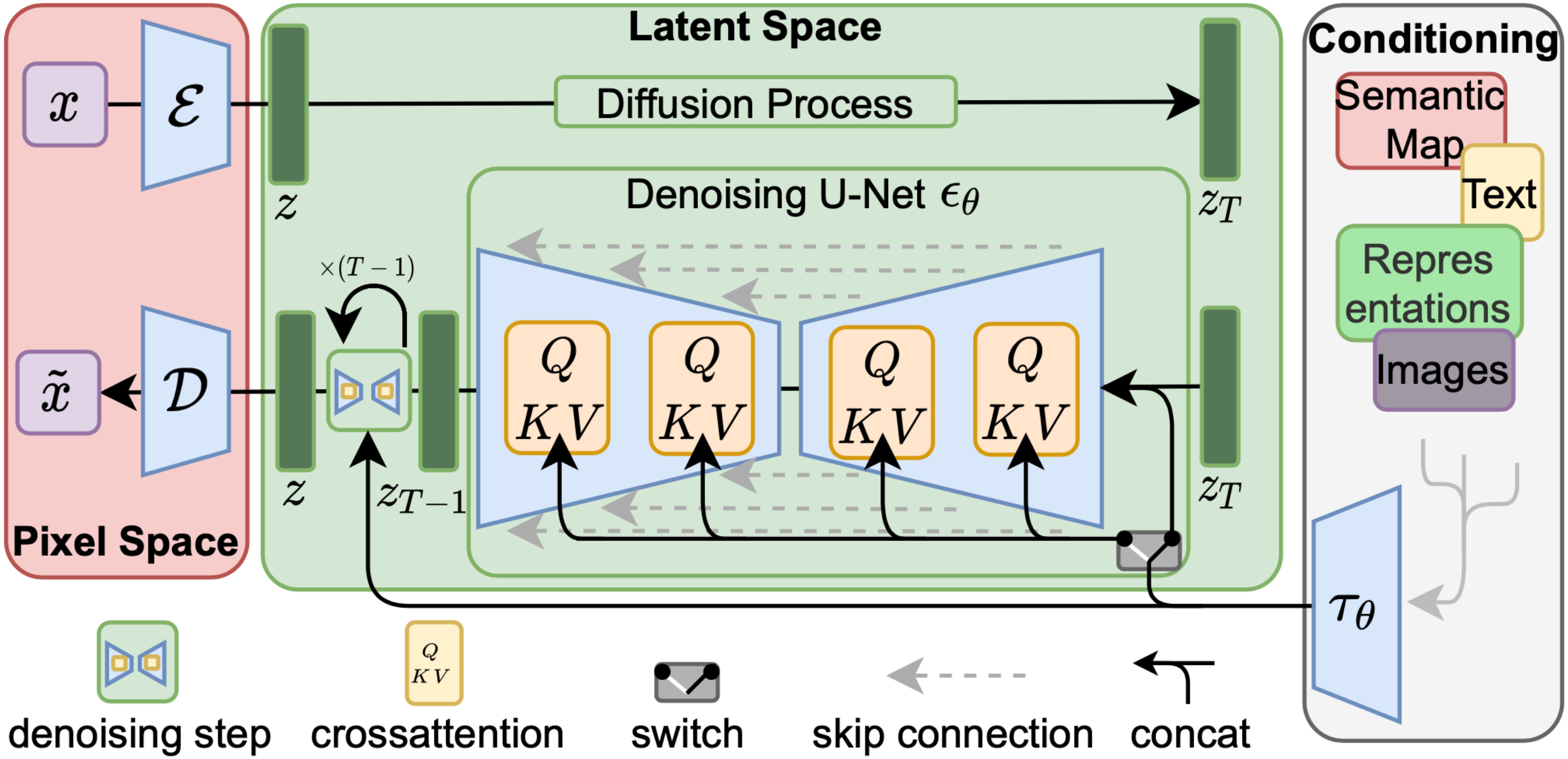
Part II-3 : Text-2-Image

- **Latent diffusion**
  - The noise is added of latent vector instead of the image → noise on high-level features
- **Conditions: attention blocks**
  - It is possible to ‘guide’ the model using conditions (such as text) → relations between text and images (CLIP), increases or decreases values on certain elements of the latent vector
  - model of cross-attention (relations between words of a sentence)



Part II-3 : Text-2-Image

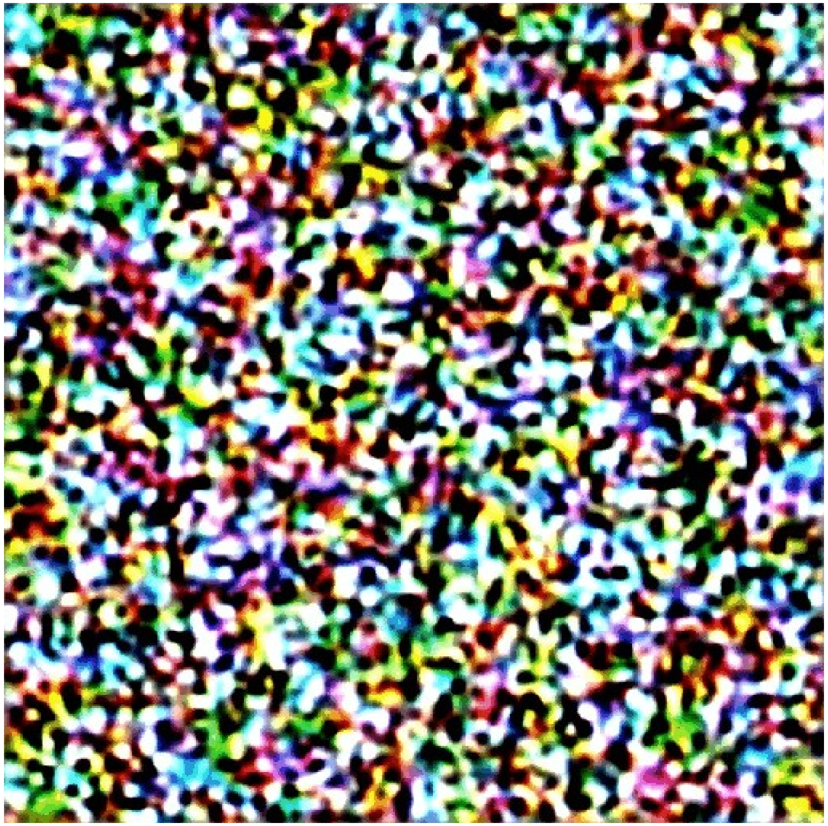
- **Stable diffusion**



Rombach & Blattmann, et al. 2022

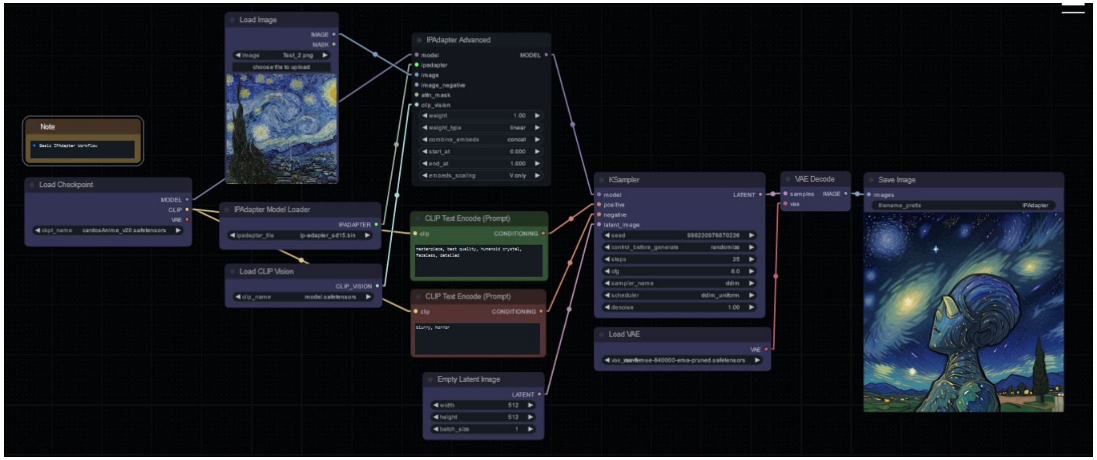
Part II-3 : Text-2-Image

- **Stable diffusion**



Part II-3 : Text-2-Image

- **Toward new assistive tools**
  - **Text assistant (ChatGPT, Co-pilot…)**
  - **Image generator (Stable diffusion, DALL-E, Midjourney…)**
    - Intuitive interfaces: ComfyUI, Stable Diffusion web UI…



Note: Basic IPAdapter Workflow

Load Checkpoint: ckpt_name cardosAnime_v20.safetensors

Load Image: image Test_2.png

IPAdapter Model Loader: ipadapter_file ip-adapter_sd15.bin

Load CLIP Vision: clip_name model.safetensors

IPAdapter Advanced: weight 1.00; weight_type linear; combine_embeds concat; start_at 0.000; end_at 1.000; embeds_scaling V only

CLIP Text Encode (Prompt): masterpiece, best quality, humanoid crystal, faceless, detailed

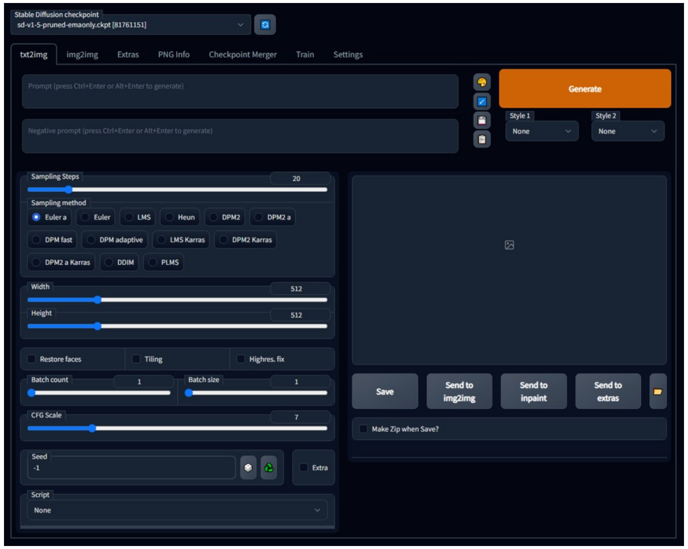
CLIP Text Encode (Prompt): blurry, horror

Empty Latent Image: width 512; height 512; batch_size 1

KSampler: seed 998220976870226; control_before_generate randomize; steps 25; cfg 6.0; sampler_name ddim; scheduler ddim_uniform; denoise 1.00

Load VAE: vae_name vae-ft-mse-840000-ema-pruned.safetensors

Save Image: filename_prefix IPAdapter



Stable Diffusion checkpoint: sd-v1-5-pruned-emaonly.ckpt [81761151]

Prompt (press Ctrl+Enter or Alt+Enter to generate)

Negative prompt (press Ctrl+Enter or Alt+Enter to generate)

Style 1: None

Style 2: None

Sampling Steps: 20

Sampling method: ● Euler a ○ Euler ○ LMS ○ Heun ○ DPM2 ○ DPM2 a ○ DPM fast ○ DPM adaptive ○ LMS Karras ○ DPM2 Karras ○ DPM2 a Karras ○ DDIM ○ PLMS

Width: 512

Height: 512

□ Restore faces □ Tiling □ Highres. fix

Batch count: 1

Batch size: 1

CFG Scale: 7

Seed: -1

□ Extra

Script: None

□ Make Zip when Save?

  - **Applications: marketing, product conception, art…**
  - **Requires experience to obtain accurate results, always verify the output results (AI models generate texts and images that <u>look</u> right)**

Conclusion

  - Deep learning is a recent domain...
  - … but already shows impressive results and achievements
  - Deep learning can still be improved on many aspects, and there is still a great margin for improvements

  - A neuronal network however remains a classifier algorithm, unable to understand or interpret data that it generates, and cannot generate more than what was in training dataset.

- There are other forms of AI, such as reinforcement learning and developmental robotics/learning, that try to interact with an environment to overcome these limitations…
  - … But this is another story !